SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 1–36

CONVERGENCE ANALYSIS OF A COLOCATED FINITE VOLUME SCHEME FOR THE INCOMPRESSIBLE NAVIER–STOKES EQUATIONS ON GENERAL 2D OR 3D MESHES*

R. EYMARD†, R. HERBIN‡, AND J. C. LATCHÉ§

Abstract. We study a colocated cell-centered finite volume method for the approximation of the incompressible Navier–Stokes equations posed on a 2D or 3D finite domain. The discrete unknowns are the components of the velocity and the pressure, all of them colocated at the center of the cells of a unique mesh; such a configuration is known to lead to stability problems, hence the need for a stabilization technique, which we choose of the Brezzi–Pitkäranta type. The scheme features two essential properties: the discrete gradient is the transpose of the divergence terms, and the discrete trilinear form associated to nonlinear advective terms vanishes on discrete divergence free velocity fields. As a consequence, the scheme is proved to be unconditionally stable and convergent for the Stokes problem and for the transient and the steady Navier–Stokes equations. In this latter case, for a given sequence of approximate solutions computed on meshes the size of which tends to zero, we prove, up to a subsequence, the L^2-convergence of the components of the velocity, and, in the steady case, the weak L^2-convergence of the pressure. The proof relies on the study of space and time translates of approximate solutions, which allows the application of Kolmogorov's theorem. The limit of this subsequence is then shown to be a weak solution of the Navier–Stokes equations. Numerical examples are performed to obtain numerical convergence rates in both the linear and nonlinear cases.

Key words. finite volume, cell-centered scheme, colocated discretizations, incompressible Navier–Stokes equations, transient Navier–Stokes equations, convergence analysis

AMS subject classifications. 15A15, 15A09, 15A23

DOI. 10.1137/040613081

1. Introduction.

In this paper we are interested in finding an approximation of the fields $\bar{u} = (\bar{u}^{(i)})_{i=1,\dots,d} : \Omega \times [0,T] \to \mathbb{R}^d$, and $\bar{p} : \Omega \times [0,T] \to \mathbb{R}$, weak solution to the incompressible Navier–Stokes equations which we write

$$\begin{aligned} &\partial_t \bar{u}^{(i)} - \nu \Delta \bar{u}^{(i)} + \partial_i \bar{p} + \sum_{j=1}^{d} \bar{u}^{(j)} \partial_j \bar{u}^{(i)} = f^{(i)} \text{ in } \Omega \times (0,T), \quad \text{for } i = 1, \dots, d, \\ &\operatorname{div} \bar{u} = \sum_{i=1}^{d} \partial_i \bar{u}^{(i)} = 0 \text{ in } \Omega \times (0,T), \end{aligned} \tag{1.1}$$

with a homogeneous Dirichlet boundary condition for $\bar{u}$ and the initial condition

$$\bar{u}^{(i)}(\cdot, 0) = \bar{u}^{(i)}_{\text{ini}} \text{ in } \Omega \text{ for } i = 1, \dots, d. \tag{1.2}$$

In the above equations, $\bar{u}^{(i)}$, $i = 1, \dots, d$, denotes the components of the velocity of a fluid which flows in a domain Ω during the time $(0,T)$, $\bar{p}$ denotes the pressure, and

*Received by the editors August 9, 2004; accepted for publication (in revised form) May 9, 2006; published electronically January 8, 2007.
http://www.siam.org/journals/sinum/45-1/61308.html
†Université de Marne-la-Vallée, Marnee-la-Vallée, F-77454, France (Robert.Eymard@univ-mlv.fr).
‡Université de Provence, Marseille, France (herbin@cmi.univ-mrs.fr).
§Institut de Radioprotection et de Sûreté Nucléaire, IRSN/CEA, Cadarache, France (jean-claude.latche@irsn.fr).

$\nu > 0$ stands for the viscosity of the fluid. We make the following assumptions:

$$\Omega \text{ is a polygonal open bounded connected subset of } \mathbb{R}^d,\ d = 2 \text{ or } 3, \tag{1.3}$$

$$T > 0 \text{ is the finite duration of the flow,} \tag{1.4}$$

$$\nu \in (0, +\infty), \tag{1.5}$$

$$\bar{u}_{\rm ini} \in L^2(\Omega)^d, \tag{1.6}$$

$$f^{(i)} \in L^2(\Omega \times (0,T)), \quad \text{for } i = 1, \ldots, d. \tag{1.7}$$

Numerical schemes for the Stokes equations and the Navier–Stokes equations have been extensively studied; see [21, 33, 34, 35, 23, 22] and the references therein. Among different schemes, finite element schemes and finite volume schemes are frequently used for mathematical or engineering studies. An advantage of finite volume schemes is that the unknowns are approximated by piecewise constant functions: this makes it easy to take into account additional nonlinear phenomena or the coupling with algebraic or differential equations, for instance in the case of reactive flows; in particular, one can find in [33, 25] the presentation of the classical finite volume scheme on rectangular meshes, which has been the basis of several industrial applications. However, the use of rectangular grids makes an important limitation to the type of domain which can be gridded and, more recently, finite volume schemes for the Navier–Stokes equations on triangular grids have been presented; see, for example, [24] where the vorticity formulation is used and [4] where primal variables are used with a Chorin-type projection method to ensure the divergence condition. Proofs of convergence for finite volume-type schemes for the Stokes and steady-state Navier–Stokes equations have recently been given for staggered grids [7, 24, 18, 3, 19, 2], following the pioneering work of Nicolaides [31] and Nicolaides and Wu [32].

In this paper, we propose the mathematical and numerical analysis of a discretization method which uses the primitive variables, that is the velocity and pressure, both approximated by piecewise constant functions on the cells of a 2D or 3D mesh. We emphasize that the approximate velocity and pressure are colocated, and therefore no dual grid is needed. The only requirement on the mesh is a geometrical assumption needed for the consistency of the approximate diffusion flux (see [13] and section 2 for a precise definition of the admissible discretizations).

As far as we know, this work is a first proof of the convergence of a finite volume scheme, which is of large interest in industry. Indeed, industrial computational fluid dynamics (CFD) codes (see, e.g., [29, 1]) use colocated cell-centered finite volume schemes; leaving aside implementation considerations, the principle of these schemes seems to differ from the present scheme only by the stabilization choice. The main reasons why this scheme is so popular in industry are

- a colocated arrangement of the unknowns,
- a very cheap assembling step (no numerical integration to perform),
- an easy coupling with other systems of equations.

The finite volume scheme studied here is based on three basic ingredients. First, a stabilization technique à la Brezzi–Pitkäranta [6] is used to cope with the instability of colocated velocity/pressure approximation spaces. Second, the discretization of the pressure gradient in the momentum balance equation is performed to ensure, by construction, that it is the transpose of the divergence term of the continuity constraint. Finally, the contribution of the discrete nonlinear advection term to the kinetic energy balance vanishes for discrete divergence free velocity fields, as in the continuous case. These features appear to be essential in the proof of convergence.

We are then able to prove the stability of the scheme and the convergence of discrete solutions towards a solution of the continuous problem when the size of the mesh tends to zero, for the steady linear case (generalized Stokes problem) and the stationary and transient Navier–Stokes equations, in two and three dimensions. Our results are valid for general meshes and do not require any assumption on the regularity of the continuous solution nor, in the nonlinear case, any small data condition. We emphasize that the convergence of the fully discrete (time and space) approximation is proven here, using an original estimate for the time translates, which yields, combined with a classical estimate on the space translates, a sufficient relative compactness property.

An error analysis is performed in the steady linear case, under regularity assumptions on the solution. An error bound of order 0.5 with respect to the step size is obtained in the discrete H^1 norm and the L^2 norm for, respectively, the velocity and the pressure. Of course, this is probably not a sharp estimate, as can be seen from the numerical results shown in section 5. Indeed, a better rate of convergence can be proven under additional assumptions on the mesh [20].

This paper is organized as follows. In section 2, we introduce the discretization tools together with some discrete functional analysis tools. Section 3 is devoted to the linear steady problem (Stokes problem), for which the finite volume scheme is given and convergence analysis and error estimates are detailed. The complete finite volume scheme for the nonlinear case is presented in section 4, in both the steady and transient cases. We then develop the analysis of its convergence to a weak solution of the continuous problem. We give some numerical results in section 5, and finally conclude with some remarks on open problems in section 6.

2. Spatial discretization and discrete functional analysis.

2.1. Admissible discretization of Ω. We first recall the notion of admissible discretization for a finite volume method, which is given in [13].

DEFINITION 2.1 (admissible discretization, steady case). *Let Ω be an open bounded polygonal (polyhedral if $d = 3$) subset of $\mathbb{R}^d$, and $\partial\Omega = \overline{\Omega}\backslash\Omega$ its boundary. An admissible finite volume discretization of Ω, denoted by $\mathcal{D}$, is given by $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$, where*

- *$\mathcal{M}$ is a finite family of nonempty open polygonal convex disjoint subsets of Ω (the "control volumes") such that $\overline{\Omega} = \cup_{K\in\mathcal{M}}\overline{K}$. For any $K \in \mathcal{M}$, let $\partial K = \overline{K} \setminus K$ be the boundary of K and $\mathrm{m}_K > 0$ denote the area of K.*
- *$\mathcal{E}$ is a finite family of disjoint subsets of $\overline{\Omega}$ (the "edges" of the mesh) such that, for all $\sigma \in \mathcal{E}$, there exists a hyperplane E of $\mathbb{R}^d$, $K \in \mathcal{M}$ with $\overline{\sigma} = \partial K \cap E$, and σ is a nonempty open subset of E. We then denote by $\mathrm{m}_\sigma > 0$ the $(d-1)$-dimensional measure of σ. We assume that, for all $K \in \mathcal{M}$, there exists a subset $\mathcal{E}_K$ of $\mathcal{E}$ such that $\partial K = \cup_{\sigma\in\mathcal{E}_K}\overline{\sigma}$. It then results from the previous hypotheses that, for all $\sigma \in \mathcal{E}$, either $\sigma \subset \partial\Omega$ or there exists $(K, L) \in \mathcal{M}^2$ with $K \neq L$ such that $\overline{K} \cap \overline{L} = \overline{\sigma}$; we denote in the latter case $\sigma = K|L$.*
- *$\mathcal{P}$ is a family of points of Ω indexed by $\mathcal{M}$, denoted by $\mathcal{P} = (x_K)_{K\in\mathcal{M}}$. The coordinates of x_K are denoted by $x_K^{(i)}$, $i = 1, \ldots, d$. The family $\mathcal{P}$ is such that, for all $K \in \mathcal{M}$, $x_K \in K$. Furthermore, for all $\sigma \in \mathcal{E}$ such that there exists $(K, L) \in \mathcal{M}^2$ with $\sigma = K|L$, it is assumed that the straight line (x_K, x_L) going through x_K and x_L is orthogonal to $K|L$. For all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{E}_K$, let z_σ be the orthogonal projection of x_K on σ. We suppose that $z_\sigma \in \sigma$.*

An example of two neighboring control volumes K and L of $\mathcal{M}$ is depicted in

Figure 2.1.

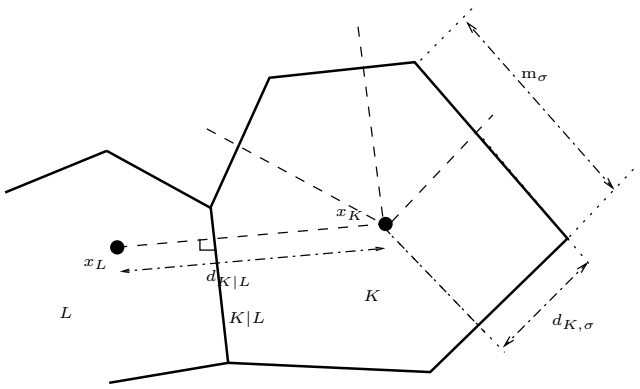


FIG. 2.1. *Notations for an admissible mesh.*

The following notations are used. The size of the discretization is defined by

$$\text{size}(\mathcal{D}) = \sup\{\text{diam}(K), K \in \mathcal{M}\}.$$

For all $K \in \mathcal{M}$ and $\sigma \in \mathcal{E}_K$, we denote by $\mathbf{n}_{K,\sigma}$ the unit vector normal to σ outward to K. We denote by $d_{K,\sigma}$ the Euclidean distance between x_K and σ. The set of interior (resp., boundary) edges is denoted by $\mathcal{E}_{\text{int}}$ (resp., $\mathcal{E}_{\text{ext}}$), that is, $\mathcal{E}_{\text{int}} = \{\sigma \in \mathcal{E};\ \sigma \not\subset \partial\Omega\}$ (resp., $\mathcal{E}_{\text{ext}} = \{\sigma \in \mathcal{E};\ \sigma \subset \partial\Omega\}$). For all $K \in \mathcal{M}$, we denote by $\mathcal{N}_K$ the subset of $\mathcal{M}$ of the neighboring control volumes. For all $K \in \mathcal{M}$ and $L \in \mathcal{N}_K$, we set $\mathbf{n}_{KL} = \mathbf{n}_{K,K|L}$, and we denote by $d_{K|L}$ the Euclidean distance between x_K and x_L.

We shall measure the regularity of the mesh through the function regul($\mathcal{D}$) defined by

$$\begin{aligned}\text{regul}(\mathcal{D}) = \inf&\left\{\frac{d_{K,\sigma}}{\text{diam}(K)},\ K \in \mathcal{M},\ \sigma \in \mathcal{E}_K\right\}\\ &\cup\left\{\frac{d_{K,K|L}}{d_{K|L}},\ L \in \mathcal{N}_K\right\} \cup \left\{\frac{1}{\text{card}(\mathcal{E}_K)},\ K \in \mathcal{M}\right\}.\end{aligned} \tag{2.1}$$

2.2. Discrete functional properties. Finite volume schemes are discrete balance equations with an adequate approximation of the fluxes; see, e.g., [13]. Recent works dealing with cell-centered finite volume methods for elliptic problems [16, 14, 19] introduce an equivalent variational formulation in adequate functional spaces. Here we shall follow this latter path, also introducing discrete analogues of the continuous Laplace, gradient, divergence, and transport operators, each of them featuring properties similar to their continuous counterparts.

DEFINITION 2.2. *Let Ω be an open bounded polygonal subset of $\mathbb{R}^d$, with $d \in \mathbb{N}^*$. Let $\mathcal{D} = (\mathcal{M}, \mathcal{E}, \mathcal{P})$ be an admissible finite volume discretization of Ω in the sense of Definition* 2.1. *We denote by $H_{\mathcal{D}}(\Omega) \subset L^2(\Omega)$ the space of functions which are piecewise constant on each control volume $K \in \mathcal{M}$. For all $w \in H_{\mathcal{D}}(\Omega)$ and for all $K \in \mathcal{M}$, we denote by w_K the constant value of w in K. The space $H_{\mathcal{D}}(\Omega)$ is equipped with the following Euclidean structure. For $(v, w) \in (H_{\mathcal{D}}(\Omega))^2$, we first define the following inner product (corresponding to Neumann boundary conditions):*

$$\langle v, w\rangle_{\mathcal{D}} = \frac{1}{2}\sum_{K\in\mathcal{M}}\sum_{L\in\mathcal{N}_K}\frac{\text{m}_{K|L}}{d_{K|L}}(v_L - v_K)(w_L - w_K). \tag{2.2}$$

We then define another inner product (corresponding to Dirichlet boundary conditions),

$$[v,w]_{\mathcal{D}} = \langle v,w\rangle_{\mathcal{D}} + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} v_K w_K. \tag{2.3}$$

Next, we define a seminorm and a norm in $H_{\mathcal{D}}(\Omega)$ (thanks to the discrete Poincaré inequality (2.4) *given below) by*

$$|w|_{\mathcal{D}} = \left(\langle w,w\rangle_{\mathcal{D}}\right)^{1/2}, \qquad \|w\|_{\mathcal{D}} = \left([w,w]_{\mathcal{D}}\right)^{1/2}.$$

We define the interpolation operator $P_{\mathcal{D}} : C(\Omega)\to H_{\mathcal{D}}(\Omega)$ by $(P_{\mathcal{D}}\varphi)_K = \varphi(x_K)$ for all $K\in\mathcal{M}$ and for all $\varphi\in C(\Omega)$.

Similarly, for $u=(u^{(i)})_{i=1,\dots,d}\in H_{\mathcal{D}}(\Omega)^d$, $v=(v^{(i)})_{i=1,\dots,d}\in H_{\mathcal{D}}(\Omega)^d$, and $w=(w^{(i)})_{i=1,\dots,d}\in H_{\mathcal{D}}(\Omega)^d$, we define

$$\|u\|_{\mathcal{D}} = \left(\sum_{i=1}^d [u^{(i)},u^{(i)}]_{\mathcal{D}}\right)^{1/2}, \qquad [v,w]_{\mathcal{D}} = \sum_{i=1}^d [v^{(i)},w^{(i)}]_{\mathcal{D}},$$

and $P_{\mathcal{D}} : C(\Omega)^d\to H_{\mathcal{D}}(\Omega)^d$ by $(P_{\mathcal{D}}\varphi)_K = \varphi(x_K)$ for all $K\in\mathcal{M}$ and for all $\varphi\in C(\Omega)^d$.

The following discrete Poincaré inequalities (see [13]) hold,

$$\|w\|_{L^2(\Omega)} \le \mathrm{diam}(\Omega)\|w\|_{\mathcal{D}} \ \forall w\in H_{\mathcal{D}}(\Omega), \tag{2.4}$$

and there exists $C_\Omega>0$, depending only on Ω, such that

$$\|w\|^2_{L^2(\Omega)} \le C_\Omega |w|^2_{\mathcal{D}} \ \forall w\in H_{\mathcal{D}}(\Omega) \text{ with } \int_\Omega w(x)\mathrm{d}x = 0. \tag{2.5}$$

Remark 2.1 (on the choice of the inner product). Before we go on with the definition of the discrete divergence and gradient operators, let us explain on the simple Laplace equation why the inner product defined by (2.3) is adequate for the approximation of the diffusion term. The discretization of the Laplace equation using finite volume methods is now classical (see [26, 13]) and is usually written in terms of fluxes; more recently [15, 27], a weak form of the scheme was introduced, which leads to more compact notations. For the sake of completeness, let us recall these two formulations for the Laplace equation $-\Delta w = f$ with homogeneous Dirichlet boundary conditions on Ω. Integrating this equation on a control volume K yields $\int_{\partial K} -\nabla w\cdot\mathbf{n} = \mathrm{m}_K f_K$. Decomposing the boundary of K into edges and approximating the diffusive flux through an edge $\sigma = K|L$ by a two-point finite difference scheme yields

$$\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}\mathrm{m}_K f_K, \tag{2.6}$$

with $F_{K,\sigma} = \frac{\mathrm{m}_\sigma}{d_{K|L}}(w_K - w_L)$ if $\sigma = K|L$ is an internal edge separating the control volumes K and L, and $F_{K,\sigma} = \frac{\mathrm{m}_\sigma}{d_{K,\sigma}} w_K$ if σ is a boundary edge. Let $\varphi\in H_{\mathcal{D}}(\Omega)$, $\varphi = \sum_{K\in\mathcal{M}}\varphi_K 1_K$. Multiplying (2.6) by φ_K, summing over K, and reordering the summations yields

$$[w,\varphi]_{\mathcal{D}} = \int_\Omega f(x)\varphi(x)\mathrm{d}x. \tag{2.7}$$

Conversely, taking $\varphi = 1_K$ in (2.7) yields (2.6). The flux form (2.6) of the scheme is therefore equivalent to the weak form (2.7) featuring the inner product (2.3).

Let us also remark that a given function $u \in H_{\mathcal{D}}(\Omega)$, considered only as an element of $L^2(\Omega)$, does not have a trace on $\partial\Omega$. However, since u is constant per control volume, one may define $u_\sigma = u_K$ for any edge $\sigma \in \mathcal{E}_{\rm ext}$, where K is the unique cell of which σ is an edge. One can then immediately deduce from the definition (2.3) of the inner product that $\sum_{\sigma\in\mathcal{E}_{\rm ext}} {\rm m}_\sigma u_\sigma^2 \le h_{\mathcal{D}}\|u\|_{\mathcal{D}}^2$. Therefore, if $(\mathcal{D}_n, u_{\mathcal{D}_n})_{n\in\mathbb{N}}$ is a sequence such that $h_{\mathcal{D}_n} \to 0$ as $n \to +\infty$, and $u_{\mathcal{D}_n} \in H_{\mathcal{D}_n}(\Omega)$ is such that $\|u_{\mathcal{D}_n}\|_{\mathcal{D}_n}$ remains bounded, then $\sum_{\sigma\in\mathcal{E}_{\rm ext}} {\rm m}_\sigma (u_{\mathcal{D}_n})_\sigma^2 \to 0$ as $n \to +\infty$. We then recover, "at the limit," the homogeneous Dirichlet boundary condition.

We define a discrete divergence operator ${\rm div}_{\mathcal{D}} : H_{\mathcal{D}}(\Omega)^d \to H_{\mathcal{D}}(\Omega)$, by

$$\text{(2.8)}\qquad {\rm div}_{\mathcal{D}}(u)(x) = \frac{1}{{\rm m}_K} \sum_{L\in\mathcal{N}_K} A_{KL}\cdot(u_K + u_L), \quad \text{for a.e. } x \in K\ \forall K \in \mathcal{M},$$

with

$$\text{(2.9)}\qquad A_{KL} = \frac{{\rm m}_{K|L}}{d_{K|L}}\frac{x_L - x_K}{2} = \frac{1}{2}\ {\rm m}_{K|L}\ \mathbf{n}_{KL}\ \forall K \in \mathcal{M} \text{ and } \forall L \in \mathcal{N}_K.$$

We then set $E_{\mathcal{D}}(\Omega) = \{u \in H_{\mathcal{D}}(\Omega)^d, {\rm div}_{\mathcal{D}}(u) = 0\}$.

Remark 2.2. Thanks to the conservative formulation (2.8), the function ${\rm div}_{\mathcal{D}}(u)$ satisfies $\int_\Omega {\rm div}_{\mathcal{D}}(u)(x){\rm d}x = 0$.

Remark 2.3. Any definition of A_{KL} such that $A_{KL} = {\rm m}_{K|L} a_{KL}\mathbf{n}_{KL}$ with $a_{KL} \ge 0$ and $a_{KL} + a_{LK} = 1$, combined with the definition ${\rm div}_{\mathcal{D}}(u)(x) = \frac{1}{{\rm m}_K}\sum_{L\in\mathcal{N}_K}(A_{KL}\cdot u_K - A_{LK}\cdot u_L)$, yields a consistent approximation of the normal fluxes, and thus the same results of convergence as those which are proven in this paper, namely an order $h^{1/2}$ error estimate. On particular meshes, one can prove a better error estimate by choosing $a_{KL} = d(x_L, K|L)/d_{KL}$ (see [20]), which yields an order 2 consistent approximation of the normal flux, and therefore an order h error estimate. Nevertheless, in the general framework of this paper, there is no specific choice which improves the convergence result nor the error estimate. Therefore, we set in this paper $a_{KL} = 1/2$, which corresponds to (2.9). The advantage of this choice is that it leads to simpler notations and shorter equations.

The adjoint of this discrete divergence defines a discrete gradient $\nabla_{\mathcal{D}} : H_{\mathcal{D}}(\Omega) \to H_{\mathcal{D}}(\Omega)^d$:

$$\text{(2.10)}\qquad (\nabla_{\mathcal{D}}u)_K = \frac{1}{{\rm m}_K}\sum_{L\in\mathcal{N}_K} A_{KL}(u_L - u_K)\ \forall K \in \mathcal{M} \text{ and } \forall u \in H_{\mathcal{D}}(\Omega).$$

This operator $\nabla_{\mathcal{D}}$ then satisfies the following property.

LEMMA 2.3. *Let $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ be a sequence of admissible discretizations of Ω in the sense of Definition* 2.1*, such that $\lim_{m\to\infty} {\rm size}(\mathcal{D}^{(m)}) = 0$. Let us assume that there exists $C > 0$ and $\alpha \in [0, 2)$, a sequence $(u^{(m)})_{m\in\mathbb{N}}$ such that $u^{(m)} \in H_{\mathcal{D}^{(m)}}(\Omega)$, and $|u^{(m)}|_{\mathcal{D}_m}^2 \le C\ {\rm size}(\mathcal{D}^{(m)})^{-\alpha}$ for all $m \in \mathbb{N}$.*

Then the following property holds:

$$\text{(2.11)}\qquad \lim_{m\to+\infty}\int_\Omega \Big(P_{\mathcal{D}_m}\varphi(x)\nabla_{\mathcal{D}_m}u^{(m)}(x) + u^{(m)}(x)\nabla\varphi(x)\Big)\,{\rm d}x = 0\ \forall\varphi \in C_c^\infty(\Omega),$$

and therefore

$$\text{(2.12)}\qquad \lim_{m\to+\infty}\int_\Omega \nabla_{\mathcal{D}_m}u^{(m)}(x)\cdot P_{\mathcal{D}_m}\psi(x){\rm d}x = 0\ \forall\psi \in C_c^\infty(\Omega)^d \cap E(\Omega),$$

where $E(\Omega)$ is defined by (3.5).

Proof. Under the hypotheses stated in Lemma 2.3, let $i = 1, \dots, d$ and $\varphi \in C_c^\infty(\Omega)$ be given. Let us study, for $m \in \mathbb{N}$, the term

$$T_1^{(m)} = \int_\Omega \left(P_{\mathcal{D}_m}\varphi(x) \nabla_{\mathcal{D}_m} u^{(m)}(x) + u^{(m)}(x) \nabla\varphi(x) \right) dx.$$

From (2.9) and (2.10), we get that

$$T_1^{(m)} = \sum_{\sigma \in \mathcal{E}_{\rm int}, \sigma = K|L} (u_L^{(m)} - u_K^{(m)}) \mathrm{m}_{K|L} \, R_{KL}^{(m)},$$

where

$$R_{KL}^{(m)} = \left(\frac{1}{2}(\varphi(x_K) + \varphi(x_L)) - \frac{1}{\mathrm{m}_{K|L}} \int_{K|L} \varphi(x) \mathrm{d}\gamma(x) \right) \mathbf{n}_{KL}.$$

Thanks to the Cauchy–Schwarz inequality,

$$|T_1^{(m)}|^2 \le |u^{(m)}|_{\mathcal{D}_m}^2 \sum_{\sigma \in \mathcal{E}_{\rm int}, \sigma = K|L} \left| R_{KL}^{(m)} \right|^2 \mathrm{m}_{K|L} d_{KL}.$$

One has $\sum_{\sigma \in \mathcal{E}_{\rm int}, \sigma = K|L} \mathrm{m}_{K|L} d_{KL} \le d \ \mathrm{m}(\Omega)$. Thanks to the existence of $C_\varphi > 0$, which depends only on φ such that $|R_{KL}^{(m)}| \le C_\varphi \mathrm{size}(\mathcal{D}^{(m)})$, and since $\alpha < 2$, we then get that

$$\lim_{m \to \infty} T_1^{(m)} = 0,$$

which yields (2.11).

Consider now $\psi \in C_c^\infty(\Omega)^d \cap E(\Omega)$. One may write (2.11) componentwise and take $\varphi = \psi_i$ in the ith equation. Summing the d resulting equations and using the fact that $\psi \in E(\Omega)$ yields (2.12). ∎

Lemma 2.4 (discrete Rellich theorem). *Let $(\mathcal{D}^{(m)})_{m \in \mathbb{N}}$ be a sequence of admissible discretizations of Ω in the sense of Definition* 2.1*, such that $\lim_{m \to \infty} \mathrm{size}(\mathcal{D}^{(m)}) = 0$. Let us assume that there exist $C > 0$ and a sequence $(u^{(m)})_{m \in \mathbb{N}}$ such that $u^{(m)} \in H_{\mathcal{D}^{(m)}}(\Omega)$ and $\|u^{(m)}\|_{\mathcal{D}_m} \le C$ for all $m \in \mathbb{N}$.*

Then, there exist $\bar{u} \in H_0^1(\Omega)$ and a subsequence of $(u^{(m)})_{m \in \mathbb{N}}$, again denoted by $(u^{(m)})_{m \in \mathbb{N}}$, such that

1. *the sequence $(u^{(m)})_{m \in \mathbb{N}}$ converges in $L^2(\Omega)$ to $\bar{u}$ as $m \to +\infty$,*
2. *for all $\varphi \in C_c^\infty(\Omega)$, we have*

$$\lim_{m \to +\infty} [u^{(m)}, P_{\mathcal{D}_m}\varphi]_{\mathcal{D}_m} = \int_\Omega \nabla\bar{u}(x) \cdot \nabla\varphi(x) \mathrm{d}x, \tag{2.13}$$

3. *$\nabla_{\mathcal{D}_m} u^{(m)}$ weakly converges to $\nabla\bar{u}$ in $L^2(\Omega)^d$ as $m \to +\infty$ and* (2.11) *holds.*

Proof. The proof of the first two items is given in [13, proof of Theorem 9.1, pp. 773–774]. Since we have $|u^{(m)}|_{\mathcal{D}_m} \le \|u^{(m)}\|_{\mathcal{D}_m}$, we can apply Lemma 2.3, which gives the third item. ∎

Remark 2.4. Following [10], if we denote

$$\mathcal{D}_{K,\sigma} = \{ tx_K + (1-t)y, \ t \in (0,1), \ y \in \sigma \} \ \forall K \in \mathcal{M} \text{ and } \forall \sigma \in \mathcal{E}_K,$$

we may alternatively define a discrete gradient $\tilde{\nabla}_{\mathcal{D}} : H_{\mathcal{D}}(\Omega) \to L^2(\Omega)^d$ by

$$\forall\, K \in \mathcal{M},$$

$$\tilde{\nabla}_{\mathcal{D}} u(x) = \frac{d}{d_{KL}}(u_L - u_K)\mathbf{n}_{KL}, \text{ for a.e. } x \in \mathcal{D}_{K,K|L} \cup \mathcal{D}_{L,K|L}\ \forall L \in \mathcal{N}_K,$$
$$\tilde{\nabla}_{\mathcal{D}} u(x) = \frac{d}{d_{K,\sigma}}(0 - u_K)\mathbf{n}_{K,\sigma}, \text{ for a.e. } x \in \mathcal{D}_{K,\sigma}\ \forall \sigma \in \mathcal{E}_K \cap \mathcal{E}_{\text{ext}}.$$

A result similar to that of Lemma 2.4 holds with this definition of a discrete gradient, and in fact it can be shown that the weak convergence of $\tilde{\nabla}_{\mathcal{D}_m} u^{(m)}$ is equivalent to the weak convergence of $\nabla_{\mathcal{D}_m} u^{(m)}$.

LEMMA 2.5. *Let $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ be a sequence of admissible discretizations of Ω in the sense of Definition* 2.1*, such that* $\lim_{m\to\infty} \operatorname{size}(\mathcal{D}^{(m)}) = 0$.

Let us assume that there exist two sequences, $(u^{(m)})_{m\in\mathbb{N}}$ and $(v^{(m)})_{m\in\mathbb{N}}$, such that

1. *for all $m \in \mathbb{N}$, $u^{(m)}$ belongs to $H_{\mathcal{D}^{(m)}}(\Omega)$ and there exists $\bar{u} \in H_0^1(\Omega)$ such that the sequence $(u^{(m)})_{m\in\mathbb{N}}$ converges in $L^2(\Omega)$ to $\bar{u}$ as $m \to +\infty$ and*

$$\lim_{m\to\infty} \|u^{(m)}\|^2_{\mathcal{D}^{(m)}} = \|\nabla \bar{u}\|^2_{L^2(\Omega)^d}; \tag{2.14}$$

2. *for all $m \in \mathbb{N}$, $v^{(m)}$ belongs to $H_{\mathcal{D}^{(m)}}(\Omega)$ and there exists $C > 0$ and $\bar{v} \in H_0^1(\Omega)$ such that $\|v^{(m)}\|_{\mathcal{D}^{(m)}} \le C$ for all $m \in \mathbb{N}$ and the sequence $(v^{(m)})_{m\in\mathbb{N}}$ converges in $L^2(\Omega)$ to $\bar{v}$ as $m \to +\infty$.*

Then the following convergence result holds:

$$\lim_{m\to+\infty} [u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} = \int_\Omega \nabla \bar{u}(x) \cdot \nabla \bar{v}(x) \mathrm{d}x. \tag{2.15}$$

Proof. Under the assumptions of the lemma, let $\varphi \in C_c^\infty(\Omega)$. We have

$$[u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} = [u^{(m)} - P_{\mathcal{D}^{(m)}}\varphi, v^{(m)}]_{\mathcal{D}^{(m)}} + [P_{\mathcal{D}^{(m)}}\varphi, v^{(m)}]_{\mathcal{D}^{(m)}},$$

and therefore, thanks to the Cauchy–Schwarz inequality, we get

$$[u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} \ge [P_{\mathcal{D}_m}\varphi, v^{(m)}]_{\mathcal{D}^{(m)}} - \|u^{(m)} - P_{\mathcal{D}^{(m)}}\varphi\|_{\mathcal{D}^{(m)}} \|v^{(m)}\|_{\mathcal{D}^{(m)}}$$

and

$$[u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} \le [P_{\mathcal{D}_m}\varphi, v^{(m)}]_{\mathcal{D}^{(m)}} + \|u^{(m)} - P_{\mathcal{D}^{(m)}}\varphi\|_{\mathcal{D}^{(m)}} \|v^{(m)}\|_{\mathcal{D}^{(m)}}.$$

From (2.14) and thanks to Lemma 2.4, we get that

$$\lim_{m\to+\infty} \|u^{(m)} - P_{\mathcal{D}^{(m)}}\varphi\|^2_{\mathcal{D}^{(m)}} = \|\nabla \bar{u} - \nabla\varphi\|^2_{L^2(\Omega)^d},$$

and thus, passing to the limit $m \to \infty$ in the two above inequalities, we get that

$$\liminf_{m\to\infty} [u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} \ge \int_\Omega \nabla\varphi(x) \cdot \nabla \bar{v}(x) \mathrm{d}x - C\|\nabla \bar{u} - \nabla\varphi\|_{L^2(\Omega)^d}$$

and

$$\limsup_{m\to\infty} [u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} \le \int_\Omega \nabla\varphi(x) \cdot \nabla \bar{v}(x) \mathrm{d}x + C\|\nabla \bar{u} - \nabla\varphi\|_{L^2(\Omega)^d}.$$

Letting φ tend to $\bar{u}$ in $H_0^1(\Omega)$ in the two above inequalities completes the proof. ∎

3. Approximation of the linear steady problem.

3.1. The Stokes problem. We first study the following linear steady problem: Find an approximation of $\bar{u}$ and $\bar{p}$, weak solution to the generalized Stokes equations with homogeneous boundary conditions on $\partial\Omega$, which read

$$\begin{aligned} &\eta\bar{u} - \nu\Delta\bar{u} + \nabla\bar{p} = f \text{ in } \Omega, \\ &\mathrm{div}\bar{u} = 0 \text{ in } \Omega, \\ &\bar{u} = 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.1}$$

For this problem, the following assumptions are made:

$$\Omega \text{ is a polygonal open bounded connected subset of } \mathbb{R}^d,\ d = 2 \text{ or } 3, \tag{3.2}$$

$$\nu \in (0, +\infty),\ \eta \in [0, +\infty), \tag{3.3}$$

$$f \in L^2(\Omega)^d. \tag{3.4}$$

We then consider the following weak sense for problem (3.1).

Definition 3.1 (weak solution to the steady Stokes equations). *Under hypotheses* (3.2)–(3.4)*, let* $E(\Omega)$ *be defined by*

$$E(\Omega) := \{\bar{v} = (\bar{v}^{(i)})_{i=1,\dots,d} \in H_0^1(\Omega)^d, \mathrm{div}\bar{v} = 0 \text{ a.e. in } \Omega\}. \tag{3.5}$$

Then $(\bar{u}, \bar{p})$ *is called a weak solution of* (3.1) *(see, e.g.,* [36] *or* [5]*) if*

$$\left\{ \begin{array}{l} \bar{u} \in E(\Omega),\ \bar{p} \in L^2(\Omega) \text{ with } \displaystyle\int_\Omega \bar{p}(x)\mathrm{d}x = 0, \\ \eta \displaystyle\int_\Omega \bar{u}(x)\cdot\bar{v}(x)\mathrm{d}x + \nu \int_\Omega \nabla\bar{u}(x) : \nabla\bar{v}(x)\mathrm{d}x \\ \qquad\qquad - \displaystyle\int_\Omega \bar{p}(x)\mathrm{div}\bar{v}(x)\mathrm{d}x = \int_\Omega f(x)\cdot\bar{v}(x)\mathrm{d}x \quad \forall \bar{v} \in H_0^1(\Omega)^d, \end{array} \right. \tag{3.6}$$

where, for all $\bar{u}, \bar{v} \in H_0^1(\Omega)^d$ *and for a.e.* $x \in \Omega$*, we use the following notation:*

$$\nabla\bar{u}(x) : \nabla\bar{v}(x) = \sum_{i=1}^{d} \nabla\bar{u}^{(i)}(x)\cdot\nabla\bar{v}^{(i)}(x).$$

The existence and uniqueness of the weak solution of (3.1) in the sense of the above definition is a classical result (see, e.g., [36] or [5]).

3.2. The finite volume scheme. Under hypotheses (3.2)–(3.4), let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition 2.1. It is then natural to write an approximate problem to the Stokes problem (3.6) in the following way (recall that $E_\mathcal{D}(\Omega) = \{u \in H_\mathcal{D}(\Omega)^d, \mathrm{div}_\mathcal{D}(u) = 0\}$):

$$\left\{ \begin{array}{l} u \in E_\mathcal{D}(\Omega),\ p \in H_\mathcal{D}(\Omega) \text{ with } \displaystyle\int_\Omega p(x)\mathrm{d}x = 0, \\ \eta \displaystyle\int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu[u, v]_\mathcal{D} \\ \qquad\qquad - \displaystyle\int_\Omega p(x)\mathrm{div}_\mathcal{D}(v)(x)\mathrm{d}x = \int_\Omega f(x)\cdot v(x)\mathrm{d}x \qquad \forall v \in H_\mathcal{D}(\Omega)^d. \end{array} \right. \tag{3.7}$$

As we use a colocated approximation for the velocity and the pressure fields, the scheme must be stabilized. Using a nonconsistent stabilization à la Brezzi–Pitkäranta [6], we then look for (u,p) such that

$$
\left\{
\begin{array}{l}
(u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega) \text{ with } \displaystyle\int_\Omega p(x)\mathrm{d}x = 0, \\
\eta \displaystyle\int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu [u,v]_{\mathcal{D}} \\
\qquad\qquad - \displaystyle\int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x = \int_\Omega f(x)\cdot v(x)\mathrm{d}x \qquad\qquad \forall v \in H_{\mathcal{D}}(\Omega)^d, \\
\displaystyle\int_\Omega \mathrm{div}_{\mathcal{D}}(u)(x) q(x)\mathrm{d}x = -\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p,q\rangle_{\mathcal{D}} \qquad\qquad \forall q \in H_{\mathcal{D}}(\Omega),
\end{array}
\right. \tag{3.8}
$$

where $\lambda > 0$ and $\alpha \in (0,2)$ are adjustable parameters of the scheme which will have to be tuned in order to make a balance between accuracy and stability.

System (3.8) is equivalent to finding the family of vectors of $\mathbb{R}^d$, $(u_K)_{K\in\mathcal{M}}$, and scalars, $(p_K)_{K\in\mathcal{M}}$, solution to the system of equations obtained by writing for each control volume K of $\mathcal{M}$

$$
\left\{
\begin{array}{l}
\eta\ \mathrm{m}_K\ u_K - \nu \displaystyle\sum_{L\in\mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}}(u_L - u_K) - \nu \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}}(0-u_K) \\
\qquad\qquad + \displaystyle\sum_{L\in\mathcal{N}_K} A_{KL}\ (p_L - p_K) = \int_K f(x)\mathrm{d}x, \\
\displaystyle\sum_{L\in\mathcal{N}_K} A_{KL}\cdot(u_K+u_L) - \lambda\ \mathrm{size}(\mathcal{D})^\alpha \sum_{L\in\mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}}(p_L - p_K) = 0,
\end{array}
\right. \tag{3.9}
$$

supplemented by the relation

$$
\sum_{K\in\mathcal{M}} \mathrm{m}_K\ p_K = 0. \tag{3.10}
$$

Defining $p_\sigma = (p_K + p_L)/2$ if $\sigma = K|L$, and $p_\sigma = p_K$ if $\sigma \in \mathcal{E}_{\mathrm{ext}} \cap \mathcal{E}_K$, and using the fact that $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma \mathbf{n}_{K,\sigma} = 0$, one notices that $\sum_{L\in\mathcal{N}_K} A_{KL}\ (p_L - p_K)$ is in fact equal to $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma p_\sigma \mathbf{n}_{K,\sigma}$, thus yielding a conservative form, which shows that (3.9) is indeed a finite volume scheme.

The existence of a solution to (3.8) will be proven below.

3.3. Study of the scheme in the linear case. We first prove a stability estimate for the velocity.

LEMMA 3.2 (discrete H_0^1 estimate for the velocity). *Under hypotheses* (3.2)–(3.4)*, let* $\mathcal{D}$ *be an admissible discretization of* Ω *in the sense of Definition* 2.1*. Let* $\lambda \in (0,+\infty)$ *and* $\alpha \in (0,2)$ *be given. Let* $(u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$ *be a solution to* (3.8)*. Then the following inequalities hold:*

$$
\nu \|u\|_{\mathcal{D}} \le \mathrm{diam}(\Omega) \|f\|_{L^2(\Omega)^d} \tag{3.11}
$$

and

$$
\nu\ \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ |p|_{\mathcal{D}}^2 \le \mathrm{diam}(\Omega)^2 \|f\|_{L^2(\Omega)^d}^2. \tag{3.12}
$$

Proof. Setting $v = u$ and $q = p$ in (3.8), we get

$$
\eta \int_\Omega u(x)^2 \mathrm{d}x + \nu \|u\|_{\mathcal{D}}^2 - \int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(u)(x)\mathrm{d}x = \int_\Omega f(x)\cdot u(x)\mathrm{d}x.
$$

Since $\eta \geq 0$, the second equation of (3.8) with $q = p$ and Young's inequality yields

$$\begin{aligned}&\eta \int_\Omega u(x)^2 \mathrm{d}x + \nu \|u\|_{\mathcal{D}}^2 + \lambda \ \mathrm{size}(\mathcal{D})^\alpha \ |p|_{\mathcal{D}}^2 \\ &\leq \frac{\mathrm{diam}(\Omega)^2}{2\nu} \|f\|_{L^2(\Omega)^d}^2 + \frac{\nu}{2 \, \mathrm{diam}(\Omega)^2} \|u\|_{L^2(\Omega)^d}^2.\end{aligned}$$

Using the Poincaré inequality (2.4) gives

$$\nu \|u\|_{\mathcal{D}}^2 + \lambda \ \mathrm{size}(\mathcal{D})^\alpha \ |p|_{\mathcal{D}}^2 \leq \frac{\mathrm{diam}(\Omega)^2}{2\nu} \|f\|_{L^2(\Omega)^d}^2 + \frac{\nu}{2} \|u\|_{\mathcal{D}}^2,$$

which leads to (3.11) and (3.12). ∎

We can now state the existence and uniqueness of a discrete solution to (3.8).

Corollary 3.3 (existence and uniqueness of a solution to the finite volume scheme). *Under hypotheses* (3.2)–(3.4)*, let* $\mathcal{D}$ *be an admissible discretization of* Ω *in the sense of Definition* 2.1*. Let* $\lambda \in (0, +\infty)$ *and* $\alpha \in (0, 2)$ *be given. Then there exists a unique solution to* (3.8)*.*

Proof. Let us define the finite dimensional vector space

$$V = \left\{ (u, p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega), \ \int_\Omega p(x) \ \mathrm{d}x = 0 \right\}.$$

Let $(u, p) \in V$ be given, and let us define $(\widetilde{u}, \widetilde{p}) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$ by

$$\left\{ \begin{array}{l} \displaystyle \int_\Omega \widetilde{u}(x) \cdot v(x) \mathrm{d}x = \eta \int_\Omega u(x) \cdot v(x) \mathrm{d}x + \nu [u, v]_{\mathcal{D}} - \int_\Omega p(x) \mathrm{div}_{\mathcal{D}}(v)(x) \mathrm{d}x \\ \qquad \forall v \in H_{\mathcal{D}}(\Omega)^d, \\ \displaystyle \int_\Omega \widetilde{p}(x) q(x) \mathrm{d}x = \int_\Omega \mathrm{div}_{\mathcal{D}}(u)(x) q(x) \mathrm{d}x + \lambda \ \mathrm{size}(\mathcal{D})^\alpha \ \langle p, q \rangle_{\mathcal{D}} \ \forall q \in H_{\mathcal{D}}(\Omega). \end{array} \right.$$

Taking $q = 1_\Omega$ shows that $\int_\Omega \widetilde{p}(x) \ \mathrm{d}x = 0$ (see Remark 2.2), and therefore $(\widetilde{u}, \widetilde{p}) \in V$. Hence we define the linear mapping $\Psi : V \to V$ such that $\Psi(u, p) = (\widetilde{u}, \widetilde{p})$. From Lemma 3.2, we get that $\Psi(u, p) = 0$ implies $u = 0$ and $p = 0$. This proves that $\Psi(\cdot)$ is one to one. This concludes the proof of existence and uniqueness of the (u, p) solution to (3.8), since $\widetilde{u}_K = \frac{1}{\mathrm{m}_K} \int_K f(x) \mathrm{d}x$ and $\widetilde{p}_K = 0$ for all $K \in \mathcal{M}$ obviously define an element of V. ∎

We then prove the following strong estimate for the pressure.

Lemma 3.4 (L^2 estimate for the pressure). *Under hypotheses* (3.2)–(3.4)*, let* $\mathcal{D}$ *be an admissible discretization of* Ω *in the sense of Definition* 2.1 *and let* $\zeta > 0$ *be such that* $\mathrm{regul}(\mathcal{D}) > \zeta$*. Let* $\lambda \in (0, +\infty)$ *and* $\alpha \in (0, 2)$ *be given. Let* $(u, p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$ *be a solution to* (3.8)*. Then there exists* C_1 *, depending only on* d*,* Ω*,* η*,* ν*,* λ*,* α*, and* ζ*, and not on* $\mathrm{size}(\mathcal{D})$*, such that the following inequality holds:*

$$\|p\|_{L^2(\Omega)} \leq C_1 \, \|f\|_{L^2(\Omega)^d}. \tag{3.13}$$

Proof. We first apply a result by Nečas [30]: thanks to $\int_\Omega p(x) \mathrm{d}x = 0$, there exists $C_2 > 0$, which depends only on d and Ω, and $\bar{v} \in H_0^1(\Omega)^d$ such that $\mathrm{div} \bar{v}(x) = p(x)$ for a.e. $x \in \Omega$ and

$$\|\bar{v}\|_{H_0^1(\Omega)^d} \leq C_2 \, \|p\|_{L^2(\Omega)}. \tag{3.14}$$

We then set

$$v_\sigma = \frac{1}{\mathrm{m}_\sigma}\int_\sigma \bar{v}(x)\mathrm{d}\gamma(x)\ \forall \sigma \in \mathcal{E}$$

(note that $v_\sigma = 0$ for all $\sigma \in \mathcal{E}_{\rm ext}$) and define $v \in H_{\mathcal{D}}(\Omega)^d$ by

$$v_K = \frac{1}{\mathrm{m}_K}\int_K \bar{v}(x)\mathrm{d}x\ \forall K \in \mathcal{M}. \tag{3.15}$$

Applying the results given in [13, p. 777], we get that there exists $C_3 > 0$, depending only on d and ζ, such that

$$(v_K^{(i)} - v_\sigma^{(i)})^2 \leq C_3\,\frac{\mathrm{diam}(K)}{\mathrm{m}_\sigma}\int_K (\nabla \bar{v}^{(i)}(x))^2\mathrm{d}x\ \forall i = 1,\ldots,d, \tag{3.16}$$

and

$$\|v\|_{\mathcal{D}} \leq C_3\,\|\bar{v}\|_{H_0^1(\Omega)^d} \leq C_3\,C_2\,\|p\|_{L^2(\Omega)}. \tag{3.17}$$

We then have

$$\int_\Omega p(x)\mathrm{div}_{\mathcal{D}}v(x)\mathrm{d}x = \sum_{K\in\mathcal{M}} p_K \sum_{L\in\mathcal{N}_K} A_{KL}\cdot(v_K + v_L) = T_2 + T_3,$$

where

$$\begin{aligned} T_2 &= \sum_{K\in\mathcal{M}} p_K \sum_{L\in\mathcal{N}_K} 2A_{KL}\cdot v_{K|L} \\ &= \sum_{K\in\mathcal{M}} p_K \sum_{L\in\mathcal{N}_K} \int_{K|L} \bar{v}(x)\cdot \mathbf{n}_{KL}\mathrm{d}\gamma(x) \\ &= \int_\Omega p(x)\mathrm{div}\bar{v}(x)\mathrm{d}x = \|p\|^2_{L^2(\Omega)}, \end{aligned}$$

and

$$\begin{aligned} T_3 &= \sum_{K\in\mathcal{M}} p_K \sum_{L\in\mathcal{N}_K} \mathrm{m}_{K|L}\left(\frac{1}{2}(v_K + v_L) - v_{K|L}\right)\cdot \mathbf{n}_{KL} \\ &= \sum_{\sigma = K|L\in\mathcal{E}_{\rm int}} \mathrm{m}_{K|L}(p_K - p_L)\left(\frac{1}{2}(v_K + v_L) - v_{K|L}\right)\cdot \mathbf{n}_{KL}. \end{aligned}$$

We then have, thanks to the Cauchy–Schwarz inequality,

$$T_3^2 \leq |p|^2_{\mathcal{D}} \sum_{\sigma = K|L\in\mathcal{E}_{\rm int}} \mathrm{m}_{K|L} d_{KL}\left(\frac{1}{2}(v_K + v_L) - v_{K|L}\right)^2.$$

Applying inequality (3.16) and thanks to $(\frac{1}{2}(v_K + v_L) - v_{K|L})^2 \leq \frac{1}{2}((v_K - v_{K|L})^2 + (v_L - v_{K|L})^2)$, we get that

$$T_3^2 \leq \frac{1}{2}\,|p|^2_{\mathcal{D}} \sum_{\sigma = K|L\in\mathcal{E}_{\rm int}} d_{KL}C_3\,\mathrm{size}(\mathcal{D})\int_{K\cup L}\sum_{i=1}^d (\nabla \bar{v}^{(i)}(x))^2\mathrm{d}x.$$

This in turn implies the existence of $C_4 > 0$, depending only on d and ζ, such that

$$T_3^2 \leq C_4 \,\mathrm{size}(\mathcal{D})^2 |p|_{\mathcal{D}}^2 \|\bar{v}\|_{H_0^1(\Omega)^d}^2.$$

Thanks to (3.14), we then get, by gathering the previous results,

$$\int_\Omega p(x)\mathrm{div}_{\mathcal{D}} v(x)\mathrm{d}x \geq \|p\|_{L^2(\Omega)}^2 - C_4 \,\mathrm{size}(\mathcal{D})|p|_{\mathcal{D}} C_2 \,\|p\|_{L^2(\Omega)}. \tag{3.18}$$

We then introduce v as a test function in (3.8). We get

$$\int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x = \eta \int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu[u,v]_{\mathcal{D}} - \int_\Omega f(x)\cdot v(x)\mathrm{d}x. \tag{3.19}$$

Applying the discrete Poincaré inequality, (3.17), and (3.18), we get the existence of C_5 , depending only on d, Ω, f, η, ν, and ζ, such that

$$\|p\|_{L^2(\Omega)}^2 - C_4 \,\mathrm{size}(\mathcal{D})|p|_{\mathcal{D}} C_2 \,\|p\|_{L^2(\Omega)} \leq C_5 \,\left(\|u\|_{\mathcal{D}} + \|f\|_{L^2(\Omega)^d}\right) \|p\|_{L^2(\Omega)}.$$

We now apply (3.11) and (3.12). Since $\mathrm{size}(\mathcal{D})^2 \leq \mathrm{size}(\mathcal{D})^\alpha \mathrm{diam}(\Omega)^{2-\alpha}$, the condition $\alpha \leq 2$ is sufficient to yield (3.13) from the above inequality, a factor $1/\lambda$ being introduced in the expression of C_1 (it is therefore not possible to let λ tend to 0 in (3.13)). □

We then have the following result, which states the convergence of the scheme (3.8).

Lemma 3.5 (convergence in the linear case). *Under hypotheses* (3.2)–(3.4)*, let* $(\bar{u},\bar{p})$ *be the unique weak solution of the Stokes problem* (3.1) *in the sense of Definition* 3.1*. Let* $\lambda \in (0,+\infty)$*,* $\alpha \in (0,2)$*, and* $\zeta > 0$ *be given, and let* $\mathcal{D}$ *be an admissible discretization of* Ω *in the sense of Definition* 2.1 *such that* $\mathrm{regul}(\mathcal{D}) \geq \zeta$*. Let* $(u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$ *be the unique solution to* (3.8)*.*

Then u *converges to* $\bar{u}$ *in* $L^2(\Omega)^d$*,* $\|u^{(i)}\|_{\mathcal{D}}$ *converges to* $\|\nabla \bar{u}^{(i)}\|_{L^2(\Omega)^d}$ *for all* $i = 1,\ldots,d$*, and* p *converges to* $\bar{p}$ *in* $L^2(\Omega)$ *as* $\mathrm{size}(\mathcal{D})$ *tends to* 0*.*

Remark 3.1. The convergence of $\|u^{(i)}\|_{\mathcal{D}}$ to $\|\nabla \bar{u}^{(i)}\|_{L^2(\Omega)^{d\times d}}$ as $\mathrm{size}(\mathcal{D})$ tends to 0 is sufficient to prove the convergence of some discrete gradient of $u^{(i)}$ to $\nabla \bar{u}^{(i)}$ in $L^2(\Omega)^d$ (see [15]).

Proof. Under the hypotheses of the above lemma, let $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ be a sequence of admissible discretizations of Ω in the sense of Definition 2.1 such that

$$\lim_{m\to\infty} \mathrm{size}(\mathcal{D}^{(m)}) = 0$$

and such that $\mathrm{regul}(\mathcal{D}^{(m)}) \geq \zeta$ for all $m \in \mathbb{N}$.

Let $(u^{(m)},p^{(m)}) \in H_{\mathcal{D}^{(m)}}(\Omega)^d \times H_{\mathcal{D}^{(m)}}(\Omega)$ be given by (3.8) for all $m \in \mathbb{N}$. We shall prove in a first step the existence of a subsequence of $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ such that the corresponding sequence $(u^{(m)})_{m\in\mathbb{N}}$ converges in $L^2(\Omega)^d$ to some function $\bar{u}$ and the sequence $(p^{(m)})_{m\in\mathbb{N}}$ weakly converges in $L^2(\Omega)^d$ to some function $\bar{p}$, as $m \to \infty$. We then show that $(\bar{u},\bar{p})$ is the solution of (3.8). Then, in a second step, following some ideas of [15], we again extract a subsequence such that $\|u^{(m)}\|_{\mathcal{D}^{(m)}}$ converges to $\|\nabla \bar{u}\|_{L^2(\Omega)^{d\times d}}$ and $p^{(m)}$ strongly converges to $\bar{p}$ in $L^2(\Omega)$ as $m \to \infty$. The proof is then complete since the solution $(\bar{u},\bar{p})$ of (3.8) is unique and therefore the convergence property holds for the whole sequence.

Step 1. Convergence of the velocity.

Using (3.11), we obtain (see [12, 13]) an estimate for the translates of $u^{(m)}$: for all $m \in \mathbb{N}$, there exists $C_6 > 0$, depending only on Ω, ν, f, and g, such that

$$\int_\Omega (u^{(m,k)}(x+\xi) - u^{(m,k)}(x))^2 \mathrm{d}x \leq C_6\, |\xi|(|\xi| + 4\,\mathrm{size}(\mathcal{D}^{(m)})), \quad \text{for } k = 1, \ldots, d\ \forall \xi \in \mathbb{R}^d, \tag{3.20}$$

where $u^{(m,k)}$ denotes the kth component of $u^{(m)}$. We may then apply Kolmogorov's theorem and obtain the existence of a subsequence of $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ and of $\bar{u} \in H_0^1(\Omega)^d$ such that $(u^{(m)})_{m\in\mathbb{N}}$ converges to $\bar{u}$ in $L^2(\Omega)^d$. Thanks to Lemma 3.4, we extract from this subsequence another one (still denoted by $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$) such that $(p^{(m)})_{m\in\mathbb{N}}$ weakly converges to some function $\bar{p}$ in $L^2(\Omega)$. In order to conclude the proof of the convergence of the scheme, there only remains to prove that $(\bar{u}, \bar{p})$ is the solution of (3.6).

Let $\varphi \in C_c^\infty(\Omega)^d$. For m large enough, and thus size$(\mathcal{D}^{(m)})$ small enough, $\partial K \cap \partial\Omega = \emptyset$ holds for all $K \in \mathcal{M}$ such that $K \cap \mathrm{support}(\varphi) \neq \emptyset$. Let us take $v = P_{\mathcal{D}^{(m)}}\varphi$ in (3.8). Applying Lemma 2.4, we get

$$\lim_{m\to\infty} [u^{(m)}, P_{\mathcal{D}^{(m)}}\varphi]_{\mathcal{D}^{(m)}} \int_\Omega \nabla\bar{u}(x) : \nabla\varphi(x)\mathrm{d}x.$$

Moreover, it is clear that

$$\lim_{m\to\infty} \int_\Omega f(x) \cdot P_{\mathcal{D}^{(m)}}\varphi(x)\mathrm{d}x \int_\Omega f(x)\cdot\varphi(x)\mathrm{d}x,$$

and

$$\lim_{m\to\infty} \eta \int_\Omega u^{(m)}(x) \cdot P_{\mathcal{D}^{(m)}}\varphi(x)\mathrm{d}x \eta \int_\Omega \bar{u}(x)\cdot\varphi(x)\mathrm{d}x.$$

Thanks to the weak convergence of the sequence of approximate pressures, to (3.12) and to the hypothesis $\alpha < 2$ we now apply Lemma 2.3, which gives

$$\lim_{m\to\infty} \int_\Omega p^{(m)}(x)\mathrm{div}_{\mathcal{D}^{(m)}}(P_{\mathcal{D}^{(m)}}\varphi)(x)\mathrm{d}x = \int_\Omega \bar{p}(x)\mathrm{div}\varphi(x)\mathrm{d}x. \tag{3.21}$$

Let us now prove that $\mathrm{div}(\bar{u}) = 0$ almost everywhere in Ω. Let $\varphi \in C_c^\infty(\Omega)$ be given and let us take $q = P_{\mathcal{D}^{(m)}}\varphi$ in (3.8). We get $T_4^{(m)} = -T_5^{(m)}$, where

$$T_4^{(m)} = \int_\Omega \mathrm{div}_{\mathcal{D}^{(m)}}(x)(u^{(m)})P_{\mathcal{D}^{(m)}}\varphi(x)\mathrm{d}x$$

and

$$T_5^{(m)} = \lambda\ \mathrm{size}(\mathcal{D}^{(m)})^\alpha \langle p^{(m)}, P_{\mathcal{D}^{(m)}}\varphi\rangle_{\mathcal{D}}.$$

On one hand, the third item of Lemma 2.4 yields

$$\lim_{m\to\infty} T_4^{(m)} = \sum_{i=1}^d \int_\Omega \varphi(x)\partial_i\bar{u}^{(i)}\mathrm{d}x.$$

On the other hand, using the Cauchy–Schwarz inequality, we get

$$T_5^{(m)} \leq \lambda \mathrm{size}(\mathcal{D}^{(m)})^\alpha |p^{(m)}|_{\mathcal{D}^{(m)}} |P_{\mathcal{D}^{(m)}}\varphi|_{\mathcal{D}^{(m)}}.$$

Therefore, thanks to (3.12) and to the regularity of φ (that implies that $|P_{\mathcal{D}^{(m)}}\varphi|_{\mathcal{D}}$ remains bounded independently of size$(\mathcal{D}^{(m)})$) we obtain $\lim_{m\to\infty} T_5^{(m)} = 0$. This in turn implies that

$$\sum_{i=1}^{d} \int_\Omega \varphi(x)\partial_i \bar{u}^{(i)}(x)\mathrm{d}x = 0 \ \forall\ \varphi \in C_c^\infty(\Omega), \tag{3.22}$$

which proves that $\bar{u} \in E(\Omega)$.

Step 2. Strong convergence of the pressure.

As in the proof of Lemma 3.2, we set $v = u^{(m)}$ and $q = p^{(m)}$ in (3.8). We get

$$\eta \int_\Omega u^{(m)}(x)^2\mathrm{d}x + \nu\|u^{(m)}\|^2_{\mathcal{D}^{(m)}} + \lambda\ \mathrm{size}(\mathcal{D}^{(m)})^\alpha\ |p|^2_{\mathcal{D}^{(m)}} = \int_\Omega f(x)\cdot u^{(m)}(x)\mathrm{d}x.$$

Passing to the limit in the above equation provides

$$\eta \int_\Omega \bar{u}(x)^2\mathrm{d}x + \nu \limsup_{m\to\infty} \|u^{(m)}\|^2_{\mathcal{D}^{(m)}} \le \int_\Omega f(x)\cdot \bar{u}(x)\mathrm{d}x,$$

and therefore, since $\bar{u}$ is the solution of the continuous problem (3.6), we have

$$\limsup_{m\to\infty} \|u^{(m)}\|^2_{\mathcal{D}^{(m)}} \le \int_\Omega (\nabla \bar{u}(x))^2\mathrm{d}x.$$

Thanks to the fact that

$$\liminf_{m\to\infty} \|u^{(i,m)}\|^2_{\mathcal{D}^{(m)}} \ge \int_\Omega (\nabla \bar{u}^{(i)}(x))^2\mathrm{d}x\ \forall i = 1,\dots,d,$$

proved in [28, Lemma 2.2], we get that

$$\lim_{m\to\infty} \|u^{(i,m)}\|^2_{\mathcal{D}^{(m)}} = \int_\Omega (\nabla \bar{u}^{(i)}(x))^2\mathrm{d}x\ \forall i = 1,\dots,d. \tag{3.23}$$

Following the same line as in the proof of Lemma 3.4, we now consider a function $\bar{v}^{(m)} \in H_0^1(\Omega)^d$ such that $\mathrm{div}\bar{v}^{(m)}(x) = p^{(m)}(x)$ for a.e. $x \in \Omega$ and

$$\|\bar{v}^{(m)}\|_{H_0^1(\Omega)^d} \le C_2\, \|p^{(m)}\|_{L^2(\Omega)}.$$

We then define $v^{(m)} \in H_{\mathcal{D}^{(m)}}(\Omega)^d$ by

$$v_K^{(m)} = \frac{1}{\mathrm{m}_K} \int_K \bar{v}^{(m)}(x)\mathrm{d}x\ \forall K \in \mathcal{M}.$$

Thanks to (3.13) and applying Lemma 2.4, we get the existence of a subsequence of $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$, and of $\bar{v} \in H_0^1(\Omega)^d$ such that $(v^{(m)})_{m\in\mathbb{N}}$ converges to $\bar{v}$ in $L^2(\Omega)^d$. Passing to the limit $m \to \infty$ shows that $\mathrm{div}\bar{v}(x) = \bar{p}(x)$ for a.e. $x \in \Omega$.

Using the relations (3.18) and (3.19) obtained in the proof of Lemma 3.4, we get

$$\begin{aligned}\|p^{(m)}\|^2_{L^2(\Omega)} &- C_4\,\mathrm{size}(\mathcal{D}^{(m)})|p^{(m)}|_{\mathcal{D}^{(m)}} C_2\, \|p^{(m)}\|_{L^2(\Omega)} \\ &\le \eta \int_\Omega u^{(m)}(x)\cdot v^{(m)}(x)\mathrm{d}x + \nu[u^{(m)}, v^{(m)}]_{\mathcal{D}^{(m)}} - \int_\Omega f(x)\cdot v^{(m)}(x)\mathrm{d}x.\end{aligned}$$

Thanks to (3.12) and (3.13), we have

$$\lim_{m\to\infty} C_4\,\mathrm{size}(\mathcal{D}^{(m)})|p^{(m)}|_{\mathcal{D}^{(m)}} C_2\,\|p^{(m)}\|_{L^2(\Omega)} = 0.$$

In addition, using (3.23), we have from Lemma 2.5 that

$$\lim_{m\to+\infty}[u^{(m)},v^{(m)}]_{\mathcal{D}^{(m)}} = \int_\Omega \nabla\bar{u}(x):\nabla\bar{v}(x)\mathrm{d}x.$$

Therefore, passing to the limit $m\to\infty$ in the above inequality yields

$$\limsup_{m\to\infty}\|p^{(m)}\|^2_{L^2(\Omega)} \le \eta\int_\Omega \bar{u}(x)\cdot\bar{v}(x)\mathrm{d}x + \nu\int_\Omega \nabla\bar{u}(x):\nabla\bar{v}(x)\mathrm{d}x - \int_\Omega f(x)\cdot\bar{v}(x)\mathrm{d}x.$$

Taking $\bar{v}$ as test function in (3.6) gives

$$\eta\int_\Omega \bar{u}(x)\cdot\bar{v}(x)\mathrm{d}x + \nu\int_\Omega \nabla\bar{u}(x):\nabla\bar{v}(x)\mathrm{d}x - \int_\Omega \bar{p}(x)^2\mathrm{d}x = \int_\Omega f(x)\cdot\bar{v}(x)\mathrm{d}x,$$

and therefore $\limsup_{m\to\infty}\|p^{(m)}\|^2_{L^2(\Omega)} \le \|\bar{p}\|^2_{L^2(\Omega)}$. Since, classically, we get from the weak convergence of $p^{(m)}$ to $\bar{p}$ in $L^2(\Omega)$ that $\liminf_{m\to\infty}\|p^{(m)}\|^2_{L^2(\Omega)} \ge \|\bar{p}\|^2_{L^2(\Omega)}$, we thus obtain that $\lim_{m\to\infty}\|p^{(m)}\|^2_{L^2(\Omega)} = \|\bar{p}\|^2_{L^2(\Omega)}$. This completes the proof of the strong convergence of $p^{(m)}$ to $\bar{p}$ in $L^2(\Omega)$. □

3.4. An error estimate. The following result states an error estimate for the scheme (3.8).

LEMMA 3.6 (error estimate in the linear case). *Under hypotheses* (3.2)–(3.4)*, we assume that the weak solution* $(\bar{u},\bar{p})$ *of the Stokes problem* (3.1) *in the sense of Definition* 3.1 *is such that* $(\bar{u},\bar{p})\in H^2(\Omega)^d\times H^1(\Omega)$*. Let* $\lambda\in(0,+\infty)$ *and* $\alpha\in(0,2)$ *be given, let* $\mathcal{D}$ *be an admissible discretization of* Ω *in the sense of Definition* 2.1*, and let* $\zeta>0$ *such that* $\mathrm{regul}(\mathcal{D}^{(m)})\ge\zeta$*. Let* $(u,p)\in H_{\mathcal{D}}(\Omega)^d\times H_{\mathcal{D}}(\Omega)$ *be the solution to* (3.8)*. Then there exists* C_7*, which depends only on* d*,* Ω*,* ν*,* η*, and* ζ *such that*

$$\|u-P_{\mathcal{D}}\bar{u}\|^2_{\mathcal{D}} \le C_7\,\varepsilon(\lambda,\mathrm{size}(\mathcal{D}),\bar{p},\bar{u}), \tag{3.24}$$

$$\|u-\bar{u}\|^2_{L^2(\Omega)} \le C_7\,\varepsilon(\lambda,\mathrm{size}(\mathcal{D}),\bar{p},\bar{u}), \tag{3.25}$$

$$\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ |p|^2_{\mathcal{D}} \le C_7\,\varepsilon(\lambda,\mathrm{size}(\mathcal{D}),\bar{p},\bar{u}), \tag{3.26}$$

$$\|p-\bar{p}\|^2_{L^2(\Omega)} \le C_7\,\varepsilon(\lambda,\mathrm{size}(\mathcal{D}),\bar{p},\bar{u}), \tag{3.27}$$

where

$$\varepsilon(\lambda,\mathrm{size}(\mathcal{D}),\bar{p},\bar{u}) = \max\left(\lambda\mathrm{size}(\mathcal{D})^\alpha, \frac{1}{\lambda}\mathrm{size}(\mathcal{D})^{2-\alpha}\right)\left(\|\bar{p}\|^2_{H^1(\Omega)}+\|\bar{u}\|^2_{H^2(\Omega)}\right). \tag{3.28}$$

Proof. The proof is divided into three steps: we first state the equation controlling the errors, then we prove the estimates (3.24)–(3.26) and, finally, (3.27).

Step 1. Statement of the control equation for the errors.

We define $(\widehat{u},\widehat{p})\in H_{\mathcal{D}}(\Omega)^d\times H_{\mathcal{D}}(\Omega)$ by $\widehat{u}=P_{\mathcal{D}}\bar{u}$, which means $\widehat{u}_K=\bar{u}(x_K)$ for all $K\in\mathcal{M}$, and $\widehat{p}_K=\frac{1}{\mathrm{m}_K}\int_K\bar{p}(x)\mathrm{d}x$ for all $K\in\mathcal{M}$.

Integrating the first equation of (3.1) on $K \in \mathcal{M}$ gives

$$\eta \int_K \bar{u}(x)\mathrm{d}x + \sum_{\sigma\in\mathcal{E}_K} \left(-\nu \int_\sigma \nabla\bar{u}(x)\cdot\mathbf{n}_{K,\sigma}\mathrm{d}\gamma(x) + \int_\sigma \bar{p}(x)\mathbf{n}_{K,\sigma}\mathrm{d}\gamma(x)\right) = \int_K f(x)\mathrm{d}x. \tag{3.29}$$

We introduce for $K \in \mathcal{M}$ the following consistency residuals:

$$\begin{aligned}
&R_{\mathrm{o},K} = \widehat{u}_K - \frac{1}{\mathrm{m}_K}\int_K \bar{u}(x)\mathrm{d}x,\\
&\text{for } L\in\mathcal{N}_K, \quad R_{\Delta,K|L} = \frac{1}{d_{K|L}}(\widehat{u}_L - \widehat{u}_K) - \frac{1}{\mathrm{m}_{K|L}}\int_{K|L}\nabla\bar{u}(x)\cdot\mathbf{n}_{K,K|L}\mathrm{d}\gamma(x),\\
&\text{for } \sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}, \quad R_{\Delta,\sigma} = \frac{1}{d_{K,\sigma}}(0 - \widehat{u}_K) - \frac{1}{\mathrm{m}_\sigma}\int_\sigma \nabla\bar{u}(x)\cdot\mathbf{n}_{K,\sigma}\mathrm{d}\gamma(x),\\
&\text{for } L\in\mathcal{N}_K, \quad R_{\nabla,K|L} = \frac{1}{2}(\widehat{p}_K + \widehat{p}_L) - \frac{1}{\mathrm{m}_{K|L}}\int_{K|L}\bar{p}(x)\mathrm{d}\gamma(x),\\
&\text{for } \sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}, \quad R_{\nabla,\sigma} = \widehat{p}_K - \frac{1}{\mathrm{m}_\sigma}\int_\sigma \bar{p}(x)\mathrm{d}\gamma(x).
\end{aligned}$$

Using these notations and the relation $\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\mathbf{n}_{K,\sigma} = 0$, we get from (3.29)

$$\eta\ \mathrm{m}_K\widehat{u}_K - \nu\left(\sum_{L\in\mathcal{N}_K}\frac{\mathrm{m}_{K|L}}{d_{K|L}}(\widehat{u}_L - \widehat{u}_K) + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}}\frac{\mathrm{m}_\sigma}{d_{K,\sigma}}(0-\widehat{u}_K)\right) + \sum_{L\in\mathcal{N}_K} A_{KL}\ (\widehat{p}_L - \widehat{p}_K) = \int_K f(x)\mathrm{d}x + \mathrm{m}_K R_K,$$

with

$$R_K = \eta\ R_{\mathrm{o},K} - \nu\frac{1}{\mathrm{m}_K}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma R_{\Delta,\sigma} + \frac{1}{\mathrm{m}_K}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\ R_{\nabla,\sigma}\mathbf{n}_{K,\sigma}.$$

We set $\delta u = \widehat{u} - u$ and $\delta p = \widehat{p} - p$. We then get, subtracting the first relation of the scheme (3.9) from the above equation, for all $v \in H_{\mathcal{D}}(\Omega)^d$,

$$\eta\int_\Omega \delta u(x)v(x)\mathrm{d}x + \nu[\delta u, v]_{\mathcal{D}} - \int_\Omega \delta p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x = \int_\Omega R(x)v\mathrm{d}x, \tag{3.30}$$

and, setting $v = \delta u$ in this relation,

$$\eta\int_\Omega \delta u(x)^2\mathrm{d}x + \nu\|\delta u\|_{\mathcal{D}}^2 - \int_\Omega \delta p(x)\mathrm{div}_{\mathcal{D}}(\delta u)(x)\mathrm{d}x = \int_\Omega R(x)\delta u(x)\mathrm{d}x.$$

We now integrate the second equation of (3.1) on $K \in \mathcal{M}$. This gives

$$\sum_{\sigma\in\mathcal{E}_K}\int_\sigma \bar{u}(x)\cdot\mathbf{n}_{K,\sigma}\mathrm{d}\gamma(x) = 0.$$

Since $\bar{u}$ vanishes on the boundary of Ω, we then obtain

$$\sum_{L\in\mathcal{N}_K} A_{KL}\cdot(\widehat{u}_K + \widehat{u}_L) = \sum_{L\in\mathcal{N}_K}\mathrm{m}_{K|L}R_{\mathrm{div},K|L}\ \forall K\in\mathcal{M}$$

with

$$R_{\mathrm{div},K|L} = \left(\frac{1}{2}(\widehat{u}_K + \widehat{u}_L) - \frac{1}{\mathrm{m}_{K|L}}\int_{K|L}\bar{u}(x)\mathrm{d}\gamma(x)\right)\cdot\mathbf{n}_{KL}\ \forall K\in\mathcal{M}\ \text{and}\,\forall L\in\mathcal{N}_K.$$

We then have, subtracting the second relation of the scheme (3.9) from the above equation,

$$\int_\Omega \mathrm{div}_\mathcal{D}(\delta u)(x)\delta p(x)\mathrm{d}x = \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p,\widehat{p}-p\rangle_\mathcal{D} + \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\mathrm{m}_\sigma R_{\mathrm{div},\sigma}(\delta p_K - \delta p_L).$$

Gathering the above results, we get

$$\begin{aligned}&\eta\int_\Omega \delta u(x)^2\mathrm{d}x + \nu\|\delta u\|_\mathcal{D}^2 + \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ |p|_\mathcal{D}^2\\ &= \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p,\widehat{p}\rangle_\mathcal{D} + \int_\Omega R(x)\cdot\delta u(x)\mathrm{d}x + \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\mathrm{m}_\sigma R_{\mathrm{div},\sigma}(\delta p_K - \delta p_L).\end{aligned}\tag{3.31}$$

Step 2. Proof of the bounds (3.24)–(3.26).

Let us study the terms at the right-hand side of the above equation. We have, using the Young inequality,

$$\langle p,\widehat{p}\rangle_\mathcal{D} \le \frac{1}{4}|p|_\mathcal{D}^2 + |\widehat{p}|_\mathcal{D}^2 \le \frac{1}{4}|p|_\mathcal{D}^2 + C_8\,\|\bar{p}\|_{H^1(\Omega)}^2.\tag{3.32}$$

We then decompose $\int_\Omega R(x)\cdot\delta u(x)\mathrm{d}x$ as $\int_\Omega R(x)\cdot\delta u(x)\mathrm{d}x = T_6+T_7+T_8$, with

$$\begin{aligned}T_6 &= \eta\ \int_\Omega R_{\mathrm{o}}(x)\cdot\delta u(x)\mathrm{d}x,\\ T_7 &= \nu\sum_{K\in\mathcal{M}}\left(\sum_{L\in\mathcal{N}_K}\mathrm{m}_{K|L}R_{\Delta,K|L} + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{ext}}}\mathrm{m}_\sigma R_{\Delta,\sigma}\right)\cdot\delta u_K,\\ T_8 &= \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\ R_{\nabla,\sigma}\mathbf{n}_{K,\sigma}\cdot\delta u_K.\end{aligned}$$

Thanks to interpolation results proven in [13] and to (2.4) (see also [20] for a comprehensive exposition of consistency estimates, although with slightly different projection operators), we obtain

$$\begin{aligned}T_6 &\le C_9\,\mathrm{size}(\mathcal{D})^2\|\bar{u}\|_{H^2(\Omega)^d}^2 + \frac{\nu}{4}\|\delta u\|_\mathcal{D}^2,\\ T_7 &\le C_{10}\,\mathrm{size}(\mathcal{D})^2\|\bar{u}\|_{H^2(\Omega)^d}^2 + \frac{\nu}{4}\|\delta u\|_\mathcal{D}^2,\\ T_8 &\le C_{11}\,\mathrm{size}(\mathcal{D})^2\|\bar{p}\|_{H^1(\Omega)}^2 + \frac{\nu}{4}\|\delta u\|_\mathcal{D}^2.\end{aligned}\tag{3.33}$$

We then have

$$\sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\mathrm{m}_\sigma R_{\mathrm{div},\sigma}(\delta p_K - \delta p_L) = T_9 - T_{10}$$

with

$$T_9 = \sum_{\sigma\in\mathcal{E}_{\mathrm{int}}}\mathrm{m}_\sigma R_{\mathrm{div},\sigma}(\widehat{p}_K - \widehat{p}_L) \le C_{12}\,\mathrm{size}(\mathcal{D})\left(\|\bar{p}\|_{H^1(\Omega)}^2 + \|\bar{u}\|_{H^2(\Omega)^d}^2\right),\tag{3.34}$$

and

$$
\begin{aligned}
T_{10} &= \sum_{\sigma\in\mathcal{E}_{\rm int}} \mathrm{m}_\sigma R_{{\rm div},\sigma}(p_K - p_L) \\
&\leq \frac{1}{4}\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ |p|_{\mathcal{D}}^2 + C_{13}\,\frac{1}{\lambda}\mathrm{size}(\mathcal{D})^{2-\alpha}\|\bar{u}\|_{H^2(\Omega)^d}^2.
\end{aligned} \tag{3.35}
$$

Gathering equations (3.31)–(3.35) gives

$$
\|\delta u\|_{\mathcal{D}}^2 + \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ |p|_{\mathcal{D}}^2 \leq C_{14}\,\varepsilon(\lambda, \mathrm{size}(\mathcal{D}), \bar{p}, \bar{u}),
$$

where $\varepsilon(\lambda, \mathrm{size}(\mathcal{D}), \bar{p}, \bar{u})$ is defined by (3.28). This in turn yields (3.24) (and thus (3.25) thanks to the Poincaré inequality) and (3.26).

Step 3. Proof of the bound (3.27).

We then again follow the method used in the proof of Lemma 3.4.

Using $\int_\Omega \widehat{p}(x)\mathrm{d}x = 0$, and therefore $\int_\Omega \delta p(x)\mathrm{d}x = 0$, let $\bar{v} \in H_0^1(\Omega)^d$ be given such that $\mathrm{div}\bar{v}(x) = \delta p(x)$ for a.e. $x \in \Omega$ and

$$
\|\bar{v}\|_{H_0^1(\Omega)^d} \leq C_2\,\|\delta p\|_{L^2(\Omega)}. \tag{3.36}
$$

We again set

$$
v_\sigma^{(i)} = \frac{1}{\mathrm{m}_\sigma}\int_\sigma \bar{v}^{(i)}(x)\mathrm{d}\gamma(x)\ \forall\sigma\in\mathcal{E} \text{ and } \forall i = 1,\dots,d,
$$

and we define $v \in H_{\mathcal{D}}(\Omega)^d$ by

$$
v_K^{(i)} = \frac{1}{\mathrm{m}_K}\int_K \bar{v}^{(i)}(x)\mathrm{d}x\ \forall K\in\mathcal{M} \text{ and } \forall i = 1,\dots,d.
$$

The same method gives

$$
\begin{aligned}
\|\delta p\|_{L^2(\Omega)}^2 &\leq \int_\Omega \delta p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x + C_4\,\mathrm{size}(\mathcal{D})|p|_{\mathcal{D}}\|\bar{v}\|_{H_0^1(\Omega)^d} \\
&\leq \int_\Omega \delta p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x + C_{15}\,\mathrm{size}(\mathcal{D})^2|p|_{\mathcal{D}}^2 + \frac{1}{4}\|\delta p\|_{L^2(\Omega)}^2.
\end{aligned}
$$

We now use v as test function in (3.30). We get

$$
\int_\Omega \delta p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x = \eta\int_\Omega \delta u(x)v(x)\mathrm{d}x + \nu[\delta u, v]_{\mathcal{D}} + \int_\Omega R(x)v\mathrm{d}x.
$$

Gathering the two above relations, (3.33) and (3.36) yield

$$
\begin{aligned}
\|\delta p\|_{L^2(\Omega)}^2 \leq \frac{1}{2}\|\delta p\|_{L^2(\Omega)}^2 &+ C_{16}\,\mathrm{size}(\mathcal{D})^2\left(\|\bar{p}\|_{H^1(\Omega)}^2 + \|\bar{u}\|_{H^2(\Omega)^d}^2\right) \\
&+ C_{17}\,\|\delta u\|_{\mathcal{D}}^2 + C_{15}\,\mathrm{size}(\mathcal{D})^2|p|_{\mathcal{D}}^2.
\end{aligned}
$$

Applying (3.25) and (3.26) then gives (3.27). □

Remark 3.2. In the above result, letting $\alpha = 1$, we get an order 1/2 for the convergence of the scheme. We recall that this result is not sharp, and that the numerical results show a much better order of convergence.

4. The finite volume scheme for the Navier–Stokes equations. Before handling the transient nonlinear case, we first address in the following section the steady-state case.

4.1. The steady-state case. For the continuous equations

$$\begin{aligned} &\eta\bar{u}^{(i)} - \nu\Delta\bar{u}^{(i)} + \partial_i\bar{p} + \sum_{j=1}^{d}\bar{u}^{(j)}\partial_j\bar{u}^{(i)} = f^{(i)} \text{ in } \Omega, \quad \text{for } i = 1,\dots,d, \\ &\mathrm{div}\bar{u} = \sum_{i=1}^{d}\partial_i\bar{u}^{(i)} = 0 \text{ in } \Omega \end{aligned} \tag{4.1}$$

with homogeneous Dirichlet boundary conditions for the velocity, we define the following weak sense.

DEFINITION 4.1 (weak solution to the steady Navier–Stokes equations). *Under hypotheses* (3.2)–(3.4)*, let* $E(\Omega)$ *be defined by* (3.5)*. Then* $(\bar{u},\bar{p})$ *is called a weak solution of* (4.1) *if*

$$\left\{\begin{array}{l} \bar{u}\in E(\Omega),\ \bar{p}\in L^2(\Omega) \text{ with } \displaystyle\int_\Omega \bar{p}(x)\mathrm{d}x = 0, \\ \eta\displaystyle\int_\Omega \bar{u}(x)\cdot\bar{v}(x)\mathrm{d}x + \nu\int_\Omega \nabla\bar{u}(x):\nabla\bar{v}(x)\mathrm{d}x \\ \quad -\displaystyle\int_\Omega \bar{p}(x)\mathrm{div}\bar{v}(x)\mathrm{d}x + b(\bar{u},\bar{u},\bar{v}) = \int_\Omega f(x)\cdot\bar{v}(x)\mathrm{d}x \qquad \forall\bar{v}\in H_0^1(\Omega)^d, \end{array}\right. \tag{4.2}$$

where the trilinear form $b(\cdot,\cdot,\cdot)$ *is defined, for all* $\bar{u},\bar{v},\bar{w}\in H_0^1(\Omega)^d$*, by*

$$b(\bar{u},\bar{v},\bar{w}) = \sum_{k=1}^{d}\sum_{i=1}^{d}\int_\Omega \bar{u}^{(i)}(x)\partial_i\bar{v}^{(k)}(x)\bar{w}^{(k)}(x)\,\mathrm{d}x.$$

The existence of a weak solution of (4.2) in the sense of the above definition, in two or three dimensions, is a classical result (again, see, e.g., [36]). Note that the uniqueness of the solution holds only under small data conditions.

We now give the finite volume scheme for this problem. Under hypotheses (3.2)–(3.4), let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition 2.1. We introduce Bernoulli's pressure $p+\frac{1}{2}u^2$ instead of p, again denoted by p, and for any real value $\lambda>0$ and $\alpha\in(0,2)$ we look for (u,p) such that

$$\left\{\begin{array}{l} (u,p)\in H_{\mathcal{D}}(\Omega)^d\times H_{\mathcal{D}}(\Omega) \text{ with } \displaystyle\int_\Omega p(x)\mathrm{d}x = 0, \\ \eta\displaystyle\int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu[u,v]_{\mathcal{D}} + \frac{1}{2}\int_\Omega u(x)^2\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x \\ \quad -\displaystyle\int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x + b_{\mathcal{D}}(u,u,v) = \int_\Omega f(x)\cdot v(x)\mathrm{d}x \quad \forall v\in H_{\mathcal{D}}(\Omega)^d, \\ \displaystyle\int_\Omega \mathrm{div}_{\mathcal{D}}(u)(x)q(x)\mathrm{d}x = -\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p,q\rangle_{\mathcal{D}} \qquad \forall q\in H_{\mathcal{D}}(\Omega), \end{array}\right. \tag{4.3}$$

where, for $u,v,w\in H_{\mathcal{D}}(\Omega)^d$, we define the following approximation for $b(u,v,w)$:

$$b_{\mathcal{D}}(u,v,w) = \frac{1}{2}\sum_{K\in\mathcal{M}}\sum_{L\in\mathcal{N}_K}(A_{KL}\cdot(u_K+u_L))\,((v_L-v_K)\cdot w_K). \tag{4.4}$$

System (4.3) is equivalent to finding the family of vectors of $\mathbb{R}^d$, $(u_K)_{K\in\mathcal{M}}$, and scalars, $(p_K)_{K\in\mathcal{M}}$, solution to the system of equations obtained by writing for each control volume K of $\mathcal{M}$

$$
\left\{\begin{array}{l}
\eta\ \mathrm{m}_K\ u_K - \nu \displaystyle\sum_{L\in\mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}}(u_L-u_K) - \nu \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm ext}} \frac{\mathrm{m}_\sigma}{d_{K,\sigma}}(0-u_K) \\
+ \displaystyle\sum_{L\in\mathcal{N}_K} \left(A_{KL}\cdot\left(\frac12(u_K+u_L)\right)\right)\ (u_L-u_K) \\
+ \displaystyle\sum_{L\in\mathcal{N}_K} A_{KL}\ (p_L-p_K) - \frac12 \sum_{L\in\mathcal{N}_K} A_{KL}\ (u_L^2-u_K^2) = \int_K f(x)\mathrm{d}x, \\
\displaystyle\sum_{L\in\mathcal{N}_K} A_{KL}\cdot(u_K+u_L) - \lambda\ \mathrm{size}(\mathcal{D})^\alpha \sum_{L\in\mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}}(p_L-p_K) = 0,
\end{array}\right. \tag{4.5}
$$

supplemented by the relation

$$
\sum_{K\in\mathcal{M}} \mathrm{m}_K\ p_K = 0.
$$

Defining $\tilde p_K = p_K - u_K^2/2$ and $\tilde p_\sigma = (\tilde p_K + \tilde p_L)/2$ if $\sigma = K|L$, $\tilde p_\sigma = \tilde p_K$ if $\sigma\in\mathcal{E}_{\rm ext}\cap\mathcal{E}_K$, and using the fact that $\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\mathbf{n}_{K,\sigma}=0$, one again notices that $\sum_{L\in\mathcal{N}_K} A_{KL}\ (\tilde p_L-\tilde p_K)$ is in fact equal to $\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\tilde p_\sigma\mathbf{n}_{K,\sigma}$, thus yielding a conservative form for the fifth and sixth terms of the left-hand side of the discrete momentum equation in (4.5). Defining $u_\sigma=(u_K+u_L)/2$ if $\sigma=K|L$, $u_\sigma=0$ if $\sigma\in\mathcal{E}_{\rm ext}\cap\mathcal{E}_K$, one obtains that the nonlinear advective term $\sum_{L\in\mathcal{N}_K}(A_{KL}\cdot(\frac12(u_K+u_L)))\ (u_L-u_K)$ is equal to $\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma(\mathbf{n}_{K,\sigma}\cdot u_\sigma)u_\sigma - \mathrm{m}_K u_K(\mathrm{div}_\mathcal{D}u)_K$; one may note that $(\mathrm{div}_\mathcal{D}u)_K = \sum_{\sigma\in\mathcal{E}_K}\mathrm{m}_\sigma\mathbf{n}_{K,\sigma}\cdot u_\sigma$. Hence the nonlinear advective term is the sum of a conservative form and a source term due to the stabilization (this source term vanishes for a discrete divergence free function u).

Let us then study some properties of the trilinear form $b_\mathcal{D}(\cdot,\cdot,\cdot)$. First note that the quantity $b_\mathcal{D}(u,v,w)$ also states that

$$
b_\mathcal{D}(u,v,w) = \frac12 \sum_{K|L\in\mathcal{E}_{\rm int}} (A_{KL}\cdot(u_K+u_L))\ ((v_L-v_K)\cdot(w_L+w_K)). \tag{4.6}
$$

We thus get that, for all $u,v\in H_\mathcal{D}(\Omega)^d$,

$$
\begin{aligned}
b_\mathcal{D}(u,v,v) &= \frac12 \sum_{K|L\in\mathcal{E}_{\rm int}} (A_{KL}\cdot(u_K+u_L))((v_L)^2-(v_K)^2) \\
&= -\frac12 \int_\Omega v(x)^2\ \mathrm{div}_\mathcal{D}(u)(x)\ \mathrm{d}x.
\end{aligned} \tag{4.7}
$$

We get, in particular, that, for all $u\in E_\mathcal{D}(\Omega)$, $b_\mathcal{D}(u,u,u)=0$, which is the discrete equivalent of the continuous property.

Remark 4.1 (upstream weighting versions of the scheme). The results of this paper are still valid setting $F_{KL}(u)=A_{KL}\cdot(u_K+u_L)$ and considering, for $u,v,w\in H_\mathcal{D}(\Omega)^d$,

$$
b_\mathcal{D}^{\rm ups}(u,v,w) = b_\mathcal{D}(u,v,w) + \frac12 \sum_{K|L\in\mathcal{E}_{\rm int}} \Theta_{KL}|F_{KL}(u)|\ (v_L-v_K)\cdot(w_L-w_K),
$$

with, for example, $\Theta_{KL} = \max(1 - 2\nu \frac{\mathrm{m}_{K|L}}{d_{K|L}} |F_{KL}(u)|, 0)$.

We then get, for all $u, v \in H_{\mathcal{D}}(\Omega)^d$, the inequality

$$b_{\mathcal{D}}^{\mathrm{ups}}(u, v, v) \geq -\frac{1}{2} \int_\Omega v(x)^2 \mathrm{div}_{\mathcal{D}}(u)(x) \mathrm{d}x,$$

which is sufficient to get all the estimates of this paper, together with the convergence properties of the scheme. The use of such a local upwinding technique may help avoid the development of nonphysical oscillations only when meshes are too coarse.

The following technical estimates are crucial to prove the convergence properties of the scheme.

LEMMA 4.2 (estimates on $b_{\mathcal{D}}(\cdot,\cdot,\cdot)$ by discrete Sobolev norms). *Under hypotheses* (1.3)–(1.7)*, let* $\mathcal{D}$ *be an admissible discretization of* Ω *in the sense of Definition* 2.1*, and* $\zeta > 0$ *such that* $\mathrm{regul}(\mathcal{D}) \geq \zeta$*. Then there exist* $C_{18} > 0$ *and* $C_{19} > 0$*, depending only on* d*,* ζ*, and* Ω*, such that*

$$b_{\mathcal{D}}(u, v, w) \leq C_{18} \ \|u\|_{L^4(\Omega)^d} \ \|v\|_{\mathcal{D}} \ \|w\|_{L^4(\Omega)^d} \leq C_{19} \ \|u\|_{\mathcal{D}} \ \|v\|_{\mathcal{D}} \ \|w\|_{\mathcal{D}}. \tag{4.8}$$

Proof. The quantity $b_{\mathcal{D}}(u, v, w)$ reads

$$b_{\mathcal{D}}(u, v, w) = \frac{1}{4} \sum_{K \in \mathcal{M}} \sum_{L \in \mathcal{N}_K} (w_K \cdot (v_L - v_K)) \ \frac{\mathrm{m}_{K|L}}{d_{K|L}} \ ((x_L - x_K) \cdot (u_K + u_L)).$$

Applying the Cauchy–Schwarz inequality twice and using the fact that $(x_L - x_K)^2 = d_{KL}^2$, and that, for any admissible discretization, $\sum_{L \in \mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}} \ d_{KL}^2 \leq d \ \frac{m_K}{\zeta}$, yields

$$\begin{aligned} b_{\mathcal{D}}(u, v, w)^2 \quad &\leq C_{20} \quad \left(\sum_{K \in \mathcal{M}} \sum_{L \in \mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}} (w_K)^2 (x_L - x_K)^2 (2(u_K)^2 + 2(u_L)^2) \right) \\ &\qquad \left(\sum_{K \in \mathcal{M}} \sum_{L \in \mathcal{N}_K} \frac{\mathrm{m}_{K|L}}{d_{K|L}} (v_L - v_K)^2 \right) \\ &\leq C_{21} \quad \left(\sum_{K \in \mathcal{M}} m_K |w_K|^4 \right)^{1/2} \left(\sum_{K \in \mathcal{M}} m_K |u_K|^4 \right)^{1/2} \|v\|_{\mathcal{D}}^2. \end{aligned}$$

The inequality (4.8) is now a straightforward consequence of the following discrete Sobolev inequality, which holds under the same regularity assumptions on the mesh (see proof in [8] or [13, pp. 790–791]):

$$\|u\|_{L^4(\Omega)} \leq C_{22} \ \|u\|_{\mathcal{D}}. \tag{4.9}$$ ∎

Remark 4.2 (2D case). In the case $d = 2$, it may be proven setting $\alpha = 2, p = p' = 2$ in the proof of [13, p. 791] that

$$\|u\|_{L^4(\Omega)^d} \leq C_{23} \|u\|_{L^2(\Omega)^d}^{1/2} \|u\|_{\mathcal{D}}^{1/2}$$

and therefore that there exists $C_{24} > 0$, depending only on d and Ω, such that

$$b_{\mathcal{D}}(u, v, w) \leq C_{24} \|v\|_{\mathcal{D}} \left(\|u\|_{\mathcal{D}} \ \|u\|_{L^2(\Omega)^d} \ \|w\|_{\mathcal{D}} \ \|w\|_{L^2(\Omega)^d} \right)^{1/2}.$$

This is a discrete analogue to the classical continuous estimate for the trilinear form.

The existence of a solution to the scheme (4.3) is obtained through a so-called "topological degree" argument. For the sake of completeness, we recall this argument (which was first used for numerical schemes in [11]) in the finite dimensional case in the following theorem and refer to [9] for the general case.

THEOREM 4.3 (application of the topological degree, finite dimensional case). *Let V be a finite dimensional vector space on $\mathbb{R}$, and g be a continuous function from V to V. Let us assume that there exists a continuous function F from $V \times [0,1]$ to V satisfying the following:*

1. *$F(\cdot,1) = g$, $F(\cdot,0)$ is an affine function.*
2. *There exists $R > 0$ such that, for any $(v,\rho) \in V \times [0,1]$, if $F(v,\rho) = 0$, then $\|v\|_V \neq R$.*
3. *The equation $F(v,0) = 0$ has a solution $v \in V$ such that $\|v\|_V < R$.*

Then there exists at least a solution $v \in V$ such that $g(v) = 0$ and $\|v\|_V < R$.

Here $g(v) = 0$ represents the nonlinear system (4.3), and we are now going to construct the function F and show the required estimates. Note that here the use of Bernouilli's pressure leads to simpler calculations.

LEMMA 4.4 (discrete H_0^1 estimate for the velocity). *Under hypotheses* (3.2)–(3.4)*, let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition* 2.1*. Let $\lambda \in (0,+\infty)$ and $\alpha \in (0,2)$ be given. Let $\rho \in [0,1]$ be given, and let $(u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$ be a solution to the following system of equations (which reduces to* (4.3) *as $\rho = 1$ and to* (3.8) *as $\rho = 0$):*

$$\left\{\begin{array}{ll} (u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega) \text{ with } \displaystyle\int_\Omega p(x)\mathrm{d}x = 0, & \\ \eta \displaystyle\int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu[u,v]_{\mathcal{D}} + \frac{\rho}{2}\int_\Omega u(x)^2 \mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x & \\ + \rho\ b_{\mathcal{D}}(u,u,v) - \displaystyle\int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x = \int_\Omega f(x)\cdot v(x)\mathrm{d}x & \forall v \in H_{\mathcal{D}}(\Omega)^d, \\ \displaystyle\int_\Omega \mathrm{div}_{\mathcal{D}}(u)(x)q(x)\mathrm{d}x = -\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p,q\rangle_{\mathcal{D}} & \forall q \in H_{\mathcal{D}}(\Omega). \end{array}\right. \tag{4.10}$$

Then u and p satisfy the following estimates, which are the same inequalities as obtained in the linear case (inequalities (3.11) *and* (3.12)*):*

$$\nu\|u\|_{\mathcal{D}} \leq \mathrm{diam}(\Omega)\|f\|_{L^2(\Omega)^d},$$
$$\nu\ \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ |p|_{\mathcal{D}}^2 \leq \mathrm{diam}(\Omega)^2\|f\|_{L^2(\Omega)^d}^2.$$

Proof. The proof is similar to that of Lemma 3.2, using the property (4.7) on the discrete trilinear form. □

We are now in position to prove the existence of at least one solution to scheme (4.3).

LEMMA 4.5 (existence of a discrete solution). *Under hypotheses* (3.2)–(3.4)*, let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition* 2.1*. Let $\lambda \in (0,+\infty)$ and $\alpha \in (0,2)$ be given. Then there exists at least one $(u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$, solution to* (4.3)*.*

Proof. Let us define $V = \{(u,p) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega)$ s.t. $\int_\Omega p(x)\mathrm{d}x = 0\}$. Consider the continuous mapping $F : V \times [0,1] \to V$ such that, for a given $(u,p) \in V$ and

$\rho \in [0,1]$, $(\widehat{u},\widehat{p}) = F(u,p,\rho)$ is defined by

$$\begin{aligned}\int_\Omega \widehat{u}(x)\cdot v(x)\mathrm{d}x &= \eta\int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu[u,v]_{\mathcal{D}} - \int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x\\ &\quad+\rho\left(\frac{1}{2}\int_\Omega u(x)^2\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x + b_{\mathcal{D}}(u,u,v)\right)\\ &\quad-\int_\Omega f(x)\cdot v(x)\mathrm{d}x \qquad\qquad \forall v\in H_{\mathcal{D}}(\Omega)^d,\\ \int_\Omega \widehat{p}(x)\cdot q(x)\mathrm{d}x &= \int_\Omega \mathrm{div}_{\mathcal{D}}(u)(x)q(x)\mathrm{d}x + \lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p,q\rangle_{\mathcal{D}} \qquad \forall q\in H_{\mathcal{D}}(\Omega).\end{aligned}$$

It is easily checked that the two above relations uniquely define the function $F(\cdot,\cdot,\cdot)$. Indeed, the value of $\widehat{u}_K^{(i)}$ and $\widehat{p}_K$ for a given $K\in\mathcal{M}$ and $i=1,\dots,d$ are readily obtained by setting $v^{(i)} = 1_K$, $v^{(j)}=0$ for $j\neq i$, and $q = 1_K$.

The mapping $F(.,.,.)$ is continuous, and, for a given (u,p) such that $F(u,p,\rho) = (0,0)$, we can apply Lemma 4.4, which proves that (u,p) is bounded independently on ρ. Since $F(u,p,0)$ is an affine function of (u,p) and $F(u,p,0)=0$ admits one solution (see Corollary 3.3), we may apply Theorem 4.3 and conclude the existence of at least one solution (u,p) to (4.3). ∎

We then have the following strong estimate for the pressure.

LEMMA 4.6 ($L^2(\Omega)$ estimate for the pressure). *Under hypotheses* (3.2)–(3.4)*, let $\mathcal{D}$ be an admissible discretization of Ω in the sense of Definition* 2.1*, and let $\zeta>0$ such that* $\mathrm{regul}(\mathcal{D})>\zeta$*. Let $\lambda\in(0,+\infty)$ and $\alpha\in(0,2)$ be given. Let $(u,p)\in H_{\mathcal{D}}(\Omega)^d\times H_{\mathcal{D}}(\Omega)$ be a solution to* (4.3)*. Then there exists C_{25}, depending only on d, Ω, η, ν, λ, α, and ζ, and not on* $\mathrm{size}(\mathcal{D})$*, such that the following inequality holds:*

$$\|p\|_{L^2(\Omega)} \le C_{25}\left(\|f\|_{L^2(\Omega)^d} + \left(\|f\|_{L^2(\Omega)^d}\right)^2\right). \tag{4.11}$$

Proof. We may follow the proof of Lemma 3.4 until (3.19), which is changed to

$$\begin{aligned}\int_\Omega p(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x &= \eta\int_\Omega u(x)\cdot v(x)\mathrm{d}x + \nu[u,v]_{\mathcal{D}} - \int_\Omega f(x)\cdot v(x)\mathrm{d}x\\ &\quad+\frac{1}{2}\int_\Omega u(x)^2\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x + b_{\mathcal{D}}(u,u,v).\end{aligned} \tag{4.12}$$

We again apply the discrete Poincaré inequality (2.4), (3.17), (3.18), and we use (4.8). We get the existence of C_{26}, depending only on d, Ω, f, η, ν, λ, and ζ, such that

$$\begin{aligned}\|p\|_{L^2(\Omega)}^2 - C_4\,\mathrm{size}(\mathcal{D})|p|_{\mathcal{D}}C_2\,\|p\|_{L^2(\Omega)}&\\ \le C_{26}\left(\|u\|_{\mathcal{D}} + \|f\|_{L^2(\Omega)^d} + \|u\|_{\mathcal{D}}^2\right)\|p\|_{L^2(\Omega)}.&\end{aligned}$$

We now apply (3.11) and (3.12), which yields the conclusion. ∎

We now can state the convergence of scheme (4.3).

THEOREM 4.7 (convergence of the scheme). *Under hypotheses* (3.2)–(3.4)*, let $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ be a sequence of admissible discretizations of Ω in the sense of Definition* 2.1*, such that* $\mathrm{size}(\mathcal{D}^{(m)})$ *tends to* 0 *as $m\to\infty$ and such that there exists $\zeta>0$ with* $\mathrm{regul}(\mathcal{D}^{(m)})\ge\zeta$ *for all $m\in\mathbb{N}$. Let $\lambda\in(0,+\infty)$ and $\alpha\in(0,2)$ be given. Let, for all*

$m \in \mathbb{N}$, $(u^{(m)}, p^{(m)}) \in H_{\mathcal{D}^{(m)}}(\Omega)^d \times H_{\mathcal{D}^{(m)}}(\Omega)$ be a solution to (4.3) *with $\mathcal{D} = \mathcal{D}^{(m)}$. Then there exists a weak solution $(\bar{u}, \bar{p})$ of* (4.1) *in the sense of Definition* 4.1 *and a subsequence of $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$, again denoted by $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$, such that the corresponding subsequence of solutions $(u^{(m)})_{m\in\mathbb{N}}$ converges to $\bar{u}$ in $L^2(\Omega)$ and $(p^{(m)} - \frac{1}{2}(u^{(m)})^2)_{m\in\mathbb{N}}$ weakly converges to $\bar{p}$ in $L^2(\Omega)$.*

Proof. Since the same estimates as in the linear case are available in the steady nonlinear case, the first step of the proof of Lemma 3.5 holds for all the terms of (4.2), which are present in (3.6). We have to prove only that for a given $\varphi \in C_c^\infty(\Omega)^d$, as $m \to +\infty$,

$$T_{11}^{(m)} = \int_\Omega u^{(m)}(x)^2 \mathrm{div}_{\mathcal{D}^{(m)}}(P_{\mathcal{D}^{(m)}}\varphi)(x)\mathrm{d}x \quad \text{tends to} \quad \int_\Omega \bar{u}(x)^2 \mathrm{div}\varphi(x)\mathrm{d}x$$

and

$$T_{12}^{(m)} = b_{\mathcal{D}}(u^{(m)}, u^{(m)}, P_{\mathcal{D}^{(m)}}\varphi) \quad \text{tends to} \quad b(\bar{u}, \bar{u}, \varphi).$$

Thanks to the convergence in $L^2(\Omega)^d$ of $(u^{(m)})_{m\in\mathbb{N}}$ to $\bar{u}$ and to the discrete Sobolev inequalities $\|v\|_{L^q(\Omega)} \le C_{27}\|v\|_{\mathcal{D}^{(m)}}$ for all $v \in H_{\mathcal{D}^{(m)}}(\Omega)$ and all $q \le 6$ (see [13, p. 790]), we get, using the first stability estimate of Lemma 4.4, the convergence in $L^2(\Omega)$ of $((u^{(m)})^2)_{m\in\mathbb{N}}$ to $\bar{u}^2$. We now remark that for $i = 1, \dots, d$ the sequence $(P_{\mathcal{D}^{(m)}}\varphi^{(i)})_{m\in\mathbb{N}}$ satisfies the hypotheses of Lemma 2.4. Hence, $\nabla_{\mathcal{D}^{(m)}} P_{\mathcal{D}^{(m)}}\varphi^{(i)}$ weakly converges to $\nabla\varphi^{(i)}$ in $L^2(\Omega)^d$. One has $\mathrm{div}_{\mathcal{D}} u = \sum_{i=1}^d \nabla_{\mathcal{D}}^{(i)} u^{(i)}$ for all $u \in H_{\mathcal{D}}(\Omega)^d$ such that $u_K = 0$ if $\mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}} \neq \emptyset$. Hence $\mathrm{div}_{\mathcal{D}^{(m)}}(P_{\mathcal{D}^{(m)}}\varphi)$ weakly converges to $\mathrm{div}\varphi$ in $L^2(\Omega)$, thus providing the limit of $T_{11}^{(m)}$.

Thanks to (4.6), dropping for short some indices (m), we have

$$b_{\mathcal{D}}(u, u, P_{\mathcal{D}}\varphi) = T_{13}^{(m)} - T_{14}^{(m)},$$

with

$$\begin{aligned} T_{13}^{(m)} &= \sum_{K\in\mathcal{M}} \sum_{L\in\mathcal{N}_K} (A_{KL}\cdot u_K)((u_L - u_K)\cdot\varphi(x_K)) \\ &= \sum_{k=1}^d \sum_{i=1}^d \int_\Omega u^{(i)}(x)\nabla_{\mathcal{D}}^{(i)}(u^{(k)})(x) P_{\mathcal{D}}\varphi^{(k)}(x)\mathrm{d}x, \\ T_{14}^{(m)} &= \frac{1}{2} \sum_{K|L\in\mathcal{E}_{\mathrm{int}}} (A_{KL}\cdot(u_L - u_K))((u_L - u_K)\cdot(\varphi(x_K) - \varphi(x_L))). \end{aligned}$$

Thanks to the convergence in $L^2(\Omega)$ of $(u^{(m)(i)} P_{\mathcal{D}^{(m)}}\varphi^{(j)})_{m\in\mathbb{N}}$ to $\bar{u}^{(i)}\varphi^{(j)}$, $i, j = 1, \dots, d$, we get from Lemma 2.4 that

$$\lim_{m\to\infty} T_{13}^{(m)} = \sum_{k=1}^d \sum_{i=1}^d \int_\Omega \bar{u}^{(i)}(x)\partial_i\bar{u}^{(k)}(x)\bar{\varphi}^{(k)}(x)\mathrm{d}x = b(\bar{u}, \bar{u}, \varphi).$$

We have

$$T_{14}^{(m)} = \frac{1}{4} \sum_{K|L\in\mathcal{E}_{\mathrm{int}}} d_{KL}\left(\frac{\mathrm{m}_{K|L}}{d_{K|L}}\mathbf{n}_{KL}\cdot(u_L - u_K)\right)((u_L - u_K)\cdot(\varphi(x_K) - \varphi(x_L))),$$

and therefore, since $|\varphi(x_K)-\varphi(x_L)| \le d_{KL} C_\varphi \mathrm{size}(\mathcal{D})$, where C_φ is a bound of $\nabla\varphi$ in $L^\infty(\Omega)^{d\times d}$, and since $d_{KL} \le 2\ \mathrm{size}(\mathcal{D})$, the following estimate holds:

$$|T_{14}^{(m)}| \le 4\ \mathrm{size}(\mathcal{D})^2 C_\varphi \|u\|_{\mathcal{D}}^2.$$

Therefore, the second estimate of Lemma 4.4 yields

$$\lim_{m\to\infty} T_{14}^{(m)} = 0,$$

which concludes the proof of convergence. □

4.2. The transient case. We now turn to the study of the finite volume scheme for the transient Navier–Stokes equations, the weak formulation of which is in the following definition.

DEFINITION 4.8 (weak solution to the transient Navier–Stokes equations). *Under hypotheses* (1.3)–(1.7)*, let* $E(\Omega)$ *be defined by* (3.5)*. Then* $\bar u$ *is called a weak solution of* (1.1)–(1.2) *if* $\bar u \in L^2(0,T;E(\Omega)) \cap L^\infty(0,T;L^2(\Omega)^d)$ *and*

$$\left\{\begin{array}{l} \forall \varphi \in L^2(0,T;E(\Omega)) \cap C_c^\infty(\Omega\times(-\infty,T))^d, \\ \\ \displaystyle -\int_0^T\int_\Omega \bar u(x,t)\cdot\partial_t\varphi(x,t)\,\mathrm{d}x\,\mathrm{d}t - \int_\Omega \bar u_{\mathrm{ini}}(x)\cdot\varphi(x,0)\,\mathrm{d}x \\ \displaystyle +\nu\int_0^T\int_\Omega \nabla\bar u(x,t):\nabla\varphi(x,t)\,\mathrm{d}x\,\mathrm{d}t \ + \int_0^T b(\bar u(\cdot,t),\bar u(\cdot,t),\varphi(\cdot,t))\,\mathrm{d}t \\ \displaystyle = \int_0^T\int_\Omega f(x)\cdot\varphi(x,t)\,\mathrm{d}x\,\mathrm{d}t. \end{array}\right. \tag{4.13}$$

The existence of a weak solution of (4.13) in the sense of the above definition, in two or three dimensions, is a classical result (again, see, e.g., [36] or [5]). Note that the uniqueness of the solution holds in two dimensions, and that it has only been proven in three dimensions under small data conditions.

Remark 4.3. From (4.13), we get that a weak solution u of (1.1)–(1.2) in the sense of Definition 4.8 satisfies $\partial_t\bar u \in L^{4/d}(0,T;E(\Omega)')$ and is therefore a weak solution in the classical sense, such that $\bar u(\cdot,0)$ is the orthogonal L^2-projection of $\bar u_{\mathrm{ini}}$ on $\{\bar v \in L^2(\Omega)^d, \mathrm{div}\bar v = 0, \mathrm{trace}(\bar v\cdot n_{\partial\Omega},\partial\Omega)=0\}$ (see, for example, [36] or [5]).

We first give the definition of an admissible discretization for a space-time domain.

DEFINITION 4.9 (admissible discretization, transient case). *Let* Ω *be an open bounded polygonal (polyhedral if* $d=3$*) subset of* $\mathbb{R}^d$*, and* $\partial\Omega = \overline{\Omega}\setminus\Omega$ *its boundary, and let* $T>0$*. An admissible finite volume discretization of* $\Omega\times(0,T)$*, denoted by* $\mathcal{D}$*, is given by* $\mathcal{D}=(\mathcal{M},\mathcal{E},\mathcal{P},N)$*, where* $(\mathcal{M},\mathcal{E},\mathcal{P})$ *is an admissible discretization of* Ω *in the sense of Definition* 2.1 *and* $N\in\mathbb{N}^*$ *is given. We then define* $\delta t = T/N$*, and we denote by* $\mathrm{size}(\mathcal{D}) = \max(\mathrm{size}(\mathcal{M},\mathcal{E},\mathcal{P}),\delta t)$ *and* $\mathrm{regul}(\mathcal{D}) = \mathrm{regul}(\mathcal{M},\mathcal{E},\mathcal{P})$.

Under hypotheses (1.3)–(1.7), let $\mathcal{D}$ be an admissible discretization of $\Omega\times(0,T)$ in the sense of Definition 4.9 and let $\lambda\in(0,+\infty)$ and $\alpha\in(0,2)$ be given. We write a Crank–Nicolson scheme for the time discretization, and follow the nonlinear steady-state case for the space discretization; the finite volume scheme for the approximation of the solution (1.1)–(1.2) is then

$$\begin{array}{l} u_0 \in H_{\mathcal{D}}(\Omega)^d, \\ \displaystyle u_{0,K} = \frac{1}{\mathrm{m}_K}\int_K u_{\mathrm{ini}}(x)\mathrm{d}x\ \forall K\in\mathcal{M}, \end{array} \tag{4.14}$$

and, again using Bernoulli's pressure $p + \frac{1}{2}u^2$ instead of p, again still denoted by p,

$$\begin{aligned}
&\text{for } n = 0, \dots, N-1, \text{ find } (u_{n+1}, p_{n+\frac{1}{2}}) \in H_{\mathcal{D}}(\Omega)^d \times H_{\mathcal{D}}(\Omega),\\
&\text{such that } \int_\Omega p_{n+\frac{1}{2}}(x)\mathrm{d}x = 0 \text{ and } \forall v \in H_{\mathcal{D}}(\Omega)^d, \forall q \in H_{\mathcal{D}}(\Omega),
\end{aligned}$$

$$\left\{\begin{aligned}
&\int_\Omega (u_{n+1}(x) - u_n(x)) \cdot v(x)\mathrm{d}x + \nu\delta t [u_{n+\frac{1}{2}}, v]_{\mathcal{D}}\\
&\qquad - \delta t \int_\Omega p_{n+\frac{1}{2}}(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x + \frac{\delta t}{2}\int_\Omega u_{n+\frac{1}{2}}(x)^2 \mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x\\
&\qquad + \delta t\ b_{\mathcal{D}}(u_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}, v) = \int_{n\delta t}^{(n+1)\delta t}\int_\Omega f(x,t)\cdot v(x)\mathrm{d}x\mathrm{d}t,\\
&\int_\Omega \mathrm{div}_{\mathcal{D}}(u_{n+\frac{1}{2}})(x) q(x)\mathrm{d}x = -\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p_{n+\frac{1}{2}}, q\rangle_{\mathcal{D}},
\end{aligned}\right. \tag{4.15}$$

where $u_{n+\frac{1}{2}}$ stands for $\frac{1}{2}(u_{n+1} + u_n)$.

In (4.15), we consider the approximation of $b_{\mathcal{D}}(\cdot,\cdot,\cdot)$ given by (4.4). We then define the set $H_{\mathcal{D}}(\Omega\times(0,T))$ of piecewise constant functions in each $K\times(n\delta t,(n+1)\delta t)$, $K \in \mathcal{M}$, $n = 0, \dots, N-1$, and we define $(u,p) \in H_{\mathcal{D}}(\Omega\times(0,T))^d \times H_{\mathcal{D}}(\Omega\times(0,T))$ by, for $n = 1, \dots, N-1$,

$$\left\{\begin{aligned}
&u(x,t) = u_{n+\frac{1}{2}}(x),\\
&p(x,t) = p_{n+\frac{1}{2}}(x),
\end{aligned}\right. \quad \text{for a.e. } (x,t) \in \Omega \times (n\delta t, (n+1)\delta t). \tag{4.16}$$

Remark 4.4 (time discretization). If, instead of the Crank–Nicolson scheme, we use the θ scheme, $u_{n+\frac{1}{2}} = \theta u_{n+1} + (1-\theta)u_n$, with $\theta \in [1/2, 1]$, the convergence proof which follows applies with a few minor changes. However, this is not so if θ is smaller than $1/2$; in particular, the estimate of Lemma 4.10 does not seem to be obtained easily in this case. Note that variable time steps may also be considered.

LEMMA 4.10 (existence of a discrete solution). *Under hypotheses* (1.3)–(1.7)*, let $\mathcal{D}$ be an admissible discretization of $\Omega \times (0,T)$ in the sense of Definition* 4.9*. Let $\lambda \in (0, +\infty)$ and $\alpha \in (0,2)$ be given. Then there exists at least one $(u,p) \in H_{\mathcal{D}}(\Omega\times(0,T))^d \times H_{\mathcal{D}}(\Omega\times(0,T))$, solution to* (4.14)–(4.16)*.*

Proof. We remark that, for a given $n = 0, \dots, N-1$, taking as unknown $u_{n+\frac{1}{2}}$, and noting that $u_{n+1} = 2u_{n+\frac{1}{2}} - u_n$, scheme (4.15) is under the same form as scheme (4.3), with $\eta = 2/\delta t$ and with a term involving u_n included in the right-hand side. Therefore the existence of at least one solution follows from Lemma 4.5. ∎

LEMMA 4.11 (estimates for the velocity). *Under hypotheses* (1.3)–(1.7)*, let $\mathcal{D}$ be an admissible discretization of $\Omega \times (0,T)$ in the sense of Definition* 4.9*. Let $\lambda \in (0, +\infty)$ and $\alpha \in (0,2)$. Let $(u,p) \in H_{\mathcal{D}}(\Omega\times(0,T))^d \times H_{\mathcal{D}}(\Omega\times(0,T))$ be a solution to* (4.14)–(4.16)*. Then there exists $C_{28} > 0$, depending only on d, Ω, ν, u_0, f, T, such that the following inequalities hold:*

$$\|u\|_{L^\infty(0,T;L^2(\Omega)^d)} \le C_{28}\,, \tag{4.17}$$

$$\|u\|_{L^2(0,T;H_{\mathcal{D}}(\Omega)^d)} \le C_{28}\,, \tag{4.18}$$

and

$$\lambda\ \mathrm{size}(\mathcal{D})^\alpha \sum_{n=0}^{N-1} \delta t |p_{n+\frac{1}{2}}|_{\mathcal{D}}^2 = \lambda\ \mathrm{size}(\mathcal{D})^\alpha \int_0^T |p(\cdot,t)|_{\mathcal{D}}^2 \mathrm{d}t \le C_{28}\,. \tag{4.19}$$

Proof. Let $k = 1, \ldots, N$. Setting $v = u_{n+\frac{1}{2}}$ in the first equation of (4.15), and summing on $K \in \mathcal{M}$ and $n = 0, \ldots, k-1$ in the first equation of (4.15), and using property (4.7), we get.

$$\frac{1}{2}\sum_{n=0}^{k-1}\int_\Omega (u_{n+1}(x)^2 - u_n(x)^2)\mathrm{d}x + \nu\sum_{n=0}^{k-1}\delta t[u_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}]_{\mathcal{D}}$$
$$-\sum_{n=0}^{k-1}\delta t\int_\Omega p_{n+\frac{1}{2}}(x)\mathrm{div}_{\mathcal{D}}(u_{n+\frac{1}{2}})(x)dx = \sum_{n=0}^{k-1}\int_{n\delta t}^{(n+1)\delta t}\int_\Omega f(x,t)\cdot u_{n+\frac{1}{2}}(x)\mathrm{d}x\mathrm{d}t.$$

This leads, setting $q = p_{n+\frac{1}{2}}$ in the second equation of (4.15), to

$$\begin{aligned}&\frac{1}{2}\int_\Omega (u_k(x)^2 - u_0(x)^2)\mathrm{d}x + \nu\sum_{n=0}^{k-1}\delta t[u_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}]_{\mathcal{D}}\\ &+\lambda\ \mathrm{size}(\mathcal{D})^\alpha\sum_{n=0}^{k-1}\delta t|p_{n+\frac{1}{2}}|_{\mathcal{D}}^2 = \int_0^{k\delta t}\int_\Omega f(x,t)\cdot u(x,t)\mathrm{d}x\mathrm{d}t.\end{aligned} \tag{4.20}$$

Setting $k = N$ in (4.20) gives (4.18) and (4.19). The discrete Poincaré inequality (2.4) and the inequality $\|u_0\|_{L^2(\Omega)^d} \le \|u_{\mathrm{ini}}\|_{L^2(\Omega)^d}$ give

$$\|u_k\|_{L^2(\Omega)^d}^2 \le \frac{\mathrm{diam}(\Omega)^2}{2\nu}\|f\|_{L^2(\Omega\times(0,T))^d}^2 + \|u_{\mathrm{ini}}\|_{L^2(\Omega)^d}^2\ \forall k = 1, \ldots, N,$$

which proves (4.17), since $\|u_{n+\frac{1}{2}}\|_{L^2(\Omega)^d} \le \frac{1}{2}(\|u_n\|_{L^2(\Omega)^d} + \|u_{n+1}\|_{L^2(\Omega)^d})$ for all $n = 0, \ldots, N-1$. ∎

LEMMA 4.12 (space and time velocity translate estimates). *Under hypotheses* (1.3)–(1.7), *let* $\mathcal{D}$ *be an admissible discretization of* $\Omega\times(0,T)$ *in the sense of Definition* 4.9. *Let* $\lambda \in (0,+\infty)$, $\alpha \in (0,2)$, *and* $\zeta > 0$, *such that* $\mathrm{regul}(\mathcal{D}) \ge \zeta$. *Let* $(u,p) \in H_{\mathcal{D}}(\Omega\times(0,T))^d \times H_{\mathcal{D}}(\Omega\times(0,T))$ *be a solution to* (4.14)–(4.16). *We still denote by* u *the extension in* $\mathbb{R}^d\times\mathbb{R}$ *of* u *by* 0 *outside of* $\Omega\times(0,T)$. *Then there exists* $C_{29} > 0$ *and* $C_{30} > 0$, *depending only on* d, Ω, ν, λ, α, u_0, f, ζ, *and* T, *such that the following inequalities hold:*

$$\|u(\cdot+\xi,\cdot) - u\|_{L^2(\mathbb{R}^d\times\mathbb{R})^d}^2 \le C_{29}\,|\xi|(|\xi| + 4\,\mathrm{size}(\mathcal{M}))\ \forall \xi \in \mathbb{R}^d \tag{4.21}$$

and

$$\|u(\cdot,\cdot+\tau) - u\|_{L^1(\mathbb{R};L^2(\mathbb{R}^d)^d)} \le C_{30}\,|\tau|^{1/2}\ \forall \tau \in \mathbb{R}. \tag{4.22}$$

Proof. In the following proof, we denote by C_i, where i is an integer, various positive real numbers which can depend only on d, Ω, ν, λ, α, u_0, f, ζ, and T. Inequality (4.21) is obtained from (4.18) (see [13]). Let us prove (4.22). Let $\tau \in (0,T)$ be given. We define the following norms on $H_{\mathcal{D}}(\Omega)^d$,

$$\forall\ w \in H_{\mathcal{D}}(\Omega)^d,$$

$$\begin{aligned}\|w\|_{\mathcal{D},\lambda}^2 &= \|w\|_{\mathcal{D}}^2\\ &+\frac{1}{\lambda\mathrm{size}(\mathcal{D})^\alpha}\left(\sup\left\{\int_\Omega \mathrm{div}_{\mathcal{D}}(w)(x)q(x)\mathrm{d}x,\ q \in H_{\mathcal{D}}(\Omega), |q|_{\mathcal{D}} = 1\right\}\right)^2\end{aligned} \tag{4.23}$$

and

$$\begin{aligned}&\forall\ w \in H_{\mathcal{D}}(\Omega)^d,\\ &\|w\|_{\star,\mathcal{D},\lambda} = \sup\left\{\int_\Omega w(x)\cdot v(x)dx, v \in H_{\mathcal{D}}(\Omega)^d, \|v\|_{\mathcal{D},\lambda} = 1\right\}.\end{aligned} \tag{4.24}$$

We then have, for a.e. $t \in (0,T)$,

$$\|u(\cdot,t+\tau)-u(\cdot,t)\|^2_{L^2(\Omega)^d} \le \|u(\cdot,t+\tau)-u(\cdot,t)\|_{\mathcal{D},\lambda}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\star,\mathcal{D},\lambda},$$

and therefore, thanks to Young's formula,

$$\begin{aligned}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{L^2(\Omega)^d} \le \frac{\sqrt{\tau}}{2}&\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\mathcal{D},\lambda}\\ &+\frac{1}{2\sqrt{\tau}}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\star,\mathcal{D},\lambda}.\end{aligned} \tag{4.25}$$

We get, from (4.15), for all $q \in H_{\mathcal{D}}(\Omega)$ and for a.e. $t \in (0,T)$,

$$\int_\Omega \mathrm{div}_{\mathcal{D}}(u(\cdot,t))(x)q(x)\mathrm{d}x = -\lambda\ \mathrm{size}(\mathcal{D})^\alpha\ \langle p(\cdot,t),q\rangle_{\mathcal{D}},$$

which proves, using (4.23), that

$$\|u(\cdot,t)\|^2_{\mathcal{D},\lambda} \le \|u(\cdot,t)\|^2_{\mathcal{D}} + \lambda\ \mathrm{size}(\mathcal{D})^\alpha |p(\cdot,t)|^2_{\mathcal{D}}.$$

Using the Cauchy–Schwarz inequality, we have

$$\left(\int_0^{T-\tau}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\mathcal{D},\lambda}\mathrm{d}t\right)^2 \le 4T\int_0^T \|u(\cdot,t)\|^2_{\mathcal{D},\lambda}\mathrm{d}t,$$

and therefore, using (4.18) and (4.19),

$$\int_0^{T-\tau}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\mathcal{D},\lambda}\mathrm{d}t \le C_{31}\,. \tag{4.26}$$

We now study $\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\star,\mathcal{D},\lambda}$. We can write, for a.e. $t \in (0,T-\tau)$ and $x \in \Omega$,

$$u(x,t+\tau)-u(x,t) = \frac{1}{2}\sum_{n=0}^{N-1}(\chi_n(t,\tau)+\chi_{n+1}(t,\tau))(u_{n+1}(x)-u_n(x)),$$

where, for all $n \in \mathbb{N}$ and $t \in (0,T)$, $\chi_n(t,\tau) = 1$ if $n\delta t \in [t,t+\tau[$, and $\chi_n(t,\tau) = 0$ otherwise. This implies

$$\begin{aligned}&\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\star,\mathcal{D},\lambda}\\ &\qquad\le \frac{1}{2}\sum_{n=0}^{N-1}(\chi_n(t,\tau)+\chi_{n+1}(t,\tau))\|u_{n+1}-u_n\|_{\star,\mathcal{D},\lambda}.\end{aligned} \tag{4.27}$$

Let us then obtain a bound for $\|u_{n+1}-u_n\|_{\star,\mathcal{D},\lambda}$. Using the definition of the scheme (4.15), we get that, for all $v \in H_{\mathcal{D}}(\Omega)^d$,

$$\begin{aligned}\int_\Omega (u_{n+1}(x)-u_n(x))\cdot v(x)\mathrm{d}x = &\int_{n\delta t}^{(n+1)\delta t}\int_\Omega f(x,t)\cdot v(x)\mathrm{d}x\mathrm{d}t\\ &-\nu\delta t[u_{n+\frac12},v]_{\mathcal{D}} + \delta t\int_\Omega p_{n+\frac12}(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x\\ &-\frac{\delta t}{2}\int_\Omega u^2_{n+\frac12}\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x - \delta t\ b_{\mathcal{D}}(u_{n+\frac12},u_{n+\frac12},v).\end{aligned} \tag{4.28}$$

Using the definition of $\mathrm{div}_{\mathcal{D}}$, the fact that $\sum_{\sigma\in\mathcal{E}_K} \mathrm{m}_\sigma \mathbf{n}_{K,\sigma} = 0$, and the Cauchy–Schwarz inequality, there exists C_{32} such that

$$\int_\Omega u^2_{n+\frac12}(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x \le C_{32}\,\|u^2_{n+\frac12}\|_{L^2(\Omega)}\ \|v\|_{\mathcal{D}}.$$

The discrete Sobolev inequality (4.9) leads to

$$\|u^2_{n+\frac12}\|_{L^2(\Omega)} \le \sum_{i=1}^d \|(u^{(i)}_{n+\frac12})^2\|_{L^2(\Omega)} = \sum_{i=1}^d \|u^{(i)}_{n+\frac12}\|^2_{L^4(\Omega)} \le C_{33}\,\|u_{n+\frac12}\|^2_{\mathcal{D}}.$$

We take $\|v\|_{\mathcal{D},\lambda} = 1$ and note that, from definition (4.23), we obtain that $\|v\|_{\mathcal{D}} \le 1$ and that

$$\int_\Omega p_{n+\frac12}(x)\mathrm{div}_{\mathcal{D}}(v)(x)\mathrm{d}x \le (\lambda\ \mathrm{size}(\mathcal{D})^\alpha)^{1/2}\ \ |p_{n+\frac12}|_{\mathcal{D}}.$$

We then take the supremum in (4.28). Using the Cauchy–Schwarz inequality, the discrete Poincaré inequality, and (4.8), this yields

$$\begin{aligned}\|u_{n+1}-u_n\|_{\star,\mathcal{D},\lambda} \ &\le \sqrt{\delta t}\,\mathrm{diam}(\Omega)\|f\|_{L^2((n\delta t,(n+1)\delta t);L^2(\Omega)^d)}\delta t\\ &+\delta t\,\nu\,\|u_{n+\frac12}\|_{\mathcal{D}} + \ (\lambda\ \mathrm{size}(\mathcal{D})^\alpha)^{1/2}\ \ |p_{n+\frac12}|_{\mathcal{D}}\\ &+\delta t\left(\frac12 C_{32}\,C_{33}\ + C_{19}\right)\|u_{n+\frac12}\|^2_{\mathcal{D}}.\end{aligned}$$

Summing the above equation for $n = 0$ to $N-1$, applying the Cauchy–Schwarz inequality to all terms of the right-hand side except the last one, and using (4.18) and (4.19), we get that there exists C_{34} such that

$$\sum_{n=0}^{N-1}\|u_{n+1}-u_n\|_{\star,\mathcal{D},\lambda} \le C_{34}\,.$$

Hence, noting that for all $n = 0,\dots,N$, $\int_0^{T-\tau}\chi_n(t,\tau)\mathrm{d}t \le \tau$, we have

$$\frac12\int_0^{T-\tau}\sum_{n=0}^{N-1}(\chi_n(t,\tau)+\chi_{n+1}(t,\tau))\|u_{n+1}-u_n\|_{\star,\mathcal{D},\lambda}\mathrm{d}t \le C_{34}\,\tau,$$

which proves, using (4.27),

$$\int_0^{T-\tau}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{\star,\mathcal{D},\lambda}\mathrm{d}t \le C_{34}\,\tau. \tag{4.29}$$

Thanks to (4.25), (4.26), and (4.29), we obtain that

$$\int_0^{T-\tau}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{L^2(\Omega)^d}\mathrm{d}t \le C_{35}\,\sqrt{\tau}.$$

Using (4.17), we have

$$\int_{T-\tau}^{T}\|u(\cdot,t+\tau)-u(\cdot,t)\|_{L^2(\Omega)^d}\mathrm{d}t = \int_{T-\tau}^{T}\|-u(\cdot,t)\|_{L^2(\Omega)^d}\mathrm{d}t \le C_{28}\,\tau \le \sqrt{\tau}\sqrt{T}C_{28}\,,$$

and a similar inequality holds for $\int_{-\tau}^{0} \|u(\cdot,t+\tau)-u(\cdot,t)\|_{L^2(\Omega)^d}\mathrm{d}t$. This thus gives (4.22), for any $\tau \in (0,T)$. The case $\tau \geq T$ is obtained again using (4.17), and the case $\tau \leq 0$ is obtained from $\tau \geq 0$ by the change of variable $s = t+\tau$. This completes the proof of (4.22). □

THEOREM 4.13 (convergence of the scheme). *Under hypotheses* (1.3)–(1.7)*, let* $\zeta > 0$ *be given and let* $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$ *be a sequence of admissible discretizations of* $\Omega \times (0,T)$ *in the sense of Definition* 4.9*, such that* $\mathrm{regul}(\mathcal{D}^{(m)}) \geq \zeta$ *and* $\mathrm{size}(\mathcal{D}^{(m)})$ *tends to* 0 *as* $m \to \infty$*. Let* $\lambda \in (0,+\infty)$ *and* $\alpha \in (0,2)$ *be given. Let, for all* $m \in \mathbb{N}$*,* $(u^{(m)}, p^{(m)}) \in H_{\mathcal{D}^{(m)}}(\Omega\times(0,T))^d \times H_{\mathcal{D}^{(m)}}(\Omega\times(0,T))$ *be a solution to* (4.14)–(4.16) *with* $\mathcal{D} = \mathcal{D}^{(m)}$*. Then there exists a subsequence of* $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$*, again denoted by* $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$*, such that the corresponding subsequence of solutions* $(u^{(m)})_{m\in\mathbb{N}}$ *converges in* $L^2(0,T;L^2(\Omega)^d)$ *to a weak solution* $\bar{u}$ *of* (1.1)–(1.2) *in the sense of Definition* 4.8*.*

Proof. Let us assume that the assumptions of the theorem hold. Using translates estimates (4.21) and (4.22) in the space $L^1(\mathbb{R};L^1(\mathbb{R}^d)^d)$, we can apply Kolmogorov's theorem. We get that there exist $\bar{u} \in L^1(0,T;L^1(\Omega)^d)$ and a subsequence of $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$, again denoted by $(\mathcal{D}^{(m)})_{m\in\mathbb{N}}$, such that the corresponding subsequence of solutions $(u^{(m)})_{m\in\mathbb{N}}$ converges in $L^1(0,T;L^1(\Omega)^d)$ to $\bar{u}$ as $m \to \infty$. Using (4.18), we get $\|u^{(m)}\|_{L^2(0,T;H_{\mathcal{D}_m}(\Omega)^d)} \leq C_{28}$ for all $m \in \mathbb{N}$, which gives, using the discrete Sobolev inequalities, $\|u^{(m)}\|_{L^1(0,T;L^4(\Omega)^d)} \leq C_{36}$ for all $m \in \mathbb{N}$. Using a classical result on spaces $L^p(0,T;L^q(\Omega))$, we get that $(u^{(m)})_{m\in\mathbb{N}}$ converges in $L^1(0,T;L^2(\Omega)^d)$ to $\bar{u}$ as $m \to \infty$. Thanks to (4.17), we have $\|u^{(m)}\|_{L^\infty(0,T;L^2(\Omega)^d)} \leq C_{28}$ for all $m \in \mathbb{N}$. The same result on spaces $L^p(0,T;L^q(\Omega))$ implies that $(u^{(m)})_{m\in\mathbb{N}}$ converges in $L^2(0,T;L^2(\Omega)^d)$ to $\bar{u}$ as $m \to \infty$. We can therefore pass to the limit in (4.21). The resulting inequality implies $\bar{u} \in L^2(0,T;H_0^1(\Omega)^d)$ (see [13]). Passing to the limit in (4.17) leads to $\bar{u} \in L^\infty(0,T;L^2(\Omega)^d)$.

Let us now prove that $\bar{u}$ is a weak solution of (1.1)–(1.2) in the sense of Definition 4.8.

Let $\varphi \in C_c^\infty(\Omega\times(-\infty,T))^d$ be given, with $\mathrm{div}\varphi(x,t) = 0$ for all $(x,t) \in \Omega\times(-\infty,T)$. Let $\mathcal{D}^{(m)}$ be a given admissible discretization extracted from the considered subsequence. Omitting some of the indices m for the simplicity of notation, we then set $v = P_{\mathcal{D}}\varphi(\cdot,n\delta t)$ in (4.15), and we sum for $n = 0,\dots,N-1$. We thus get

$$T_{15}^{(m)} + T_{16}^{(m)} + T_{17}^{(m)} + T_{18}^{(m)} + T_{19}^{(m)} = T_{20}^{(m)}, \tag{4.30}$$

with

$$T_{15}^{(m)} = \sum_{n=0}^{N-1} \int_\Omega (u_{n+1}(x) - u_n(x)) \cdot P_{\mathcal{D}}\varphi(x,n\delta t)\mathrm{d}x,$$

$$T_{16}^{(m)} = \sum_{n=0}^{N-1} \delta t [u_{n+\frac{1}{2}}, P_{\mathcal{D}}\varphi(\cdot,n\delta t)]_{\mathcal{D}},$$

$$T_{17}^{(m)} = -\sum_{n=0}^{N-1} \delta t \int_\Omega p_{n+\frac{1}{2}}(x)\mathrm{div}_{\mathcal{D}}(P_{\mathcal{D}}\varphi(\cdot,n\delta t))(x)\mathrm{d}x,$$

$$T_{18}^{(m)} = \frac{1}{2}\sum_{n=0}^{N-1} \delta t \int_\Omega u_{n+\frac{1}{2}}(x)^2\mathrm{div}_{\mathcal{D}}(P_{\mathcal{D}}\varphi(\cdot,n\delta t))(x)\mathrm{d}x,$$

$$T_{19}^{(m)} = \sum_{n=0}^{N-1} \delta t b_{\mathcal{D}}(u_{n+\frac{1}{2}}, u_{n+\frac{1}{2}}, P_{\mathcal{D}}\varphi(\cdot, n\delta t)),$$

and

$$T_{20}^{(m)} = \sum_{n=0}^{N-1} \int_{n\delta t}^{(n+1)\delta t} \int_{\Omega} f(x,t) \cdot P_{\mathcal{D}}\varphi(x, n\delta t) \mathrm{d}x\mathrm{d}t.$$

In the following, we denote by C_i various positive reals which can depend only on d, Ω, T, u_{ini}, f, ν, ζ, and λ. We first start with the study of T_{16}. We classically have (see [13])

$$\lim_{m\to\infty} T_{16}^{(m)} = \int_0^T \int_{\Omega} \nabla \bar{u}(x,t) : \nabla \varphi(x,t) \mathrm{d}x\mathrm{d}t. \tag{4.31}$$

The proof that

$$\lim_{m\to\infty} T_{17}^{(m)} = 0 \tag{4.32}$$

is a consequence of (4.19) and of a direct adaptation of Lemma 2.3 to time-dependent functions. Let us now prove that

$$\lim_{m\to\infty} T_{18}^{(m)} = 0. \tag{4.33}$$

Since $(u^{(m)})^2$ tend to $\bar{u}^2$ as $m \to \infty$ in $L^1(\Omega \times (0,T))$, the same argument as in the steady-state case (see proof of Theorem 4.7) provides (4.33).

We now turn to the study of T_{19}. Following the proof of Lemma 4.7, the proof that

$$\lim_{m\to\infty} T_{19}^{(m)} = \int_0^T b(\bar{u}(\cdot,t), \bar{u}(\cdot,t), \varphi(\cdot,t)) \mathrm{d}t \tag{4.34}$$

is a direct consequence of the convergence of u to $\bar{u}$ in $L^2(\Omega \times (0,T))^d$ and Lemma 2.3. The study of T_{20} is classical, and we have

$$\lim_{m\to\infty} T_{20}^{(m)} = \int_0^T \int_{\Omega} f(x,t) \cdot \varphi(x,t) \mathrm{d}x\mathrm{d}t. \tag{4.35}$$

Let us now prove that

$$\lim_{m\to\infty} T_{15}^{(m)} = -\int_0^T \int_{\Omega} \bar{u}(x,t) \partial_t \varphi(x,t) \mathrm{d}x\mathrm{d}t - \int_{\Omega} u_{\mathrm{ini}}(x) \varphi(x,0) \mathrm{d}x. \tag{4.36}$$

Indeed, we have

$$T_{15}^{(m)} = -\int_{\Omega} u_0(x) \cdot P_{\mathcal{D}}\varphi(x,0) \mathrm{d}x - T_{21}^{(m)} - \frac{1}{2} T_{22}^{(m)}$$

with

$$T_{21}^{(m)} = \sum_{n=0}^{N-1} \int_{\Omega} u_{n+\frac{1}{2}}(x) \cdot (P_{\mathcal{D}}\varphi(x,(n+1)\delta t) - P_{\mathcal{D}}\varphi(x, n\delta t)) \mathrm{d}x$$

and

$$T_{22}^{(m)} = \sum_{n=0}^{N-1} \int_\Omega (u_{n+1}(x) - u_n(x)) \cdot (P_{\mathcal{D}}\varphi(x,(n+1)\delta t) - P_{\mathcal{D}}\varphi(x,n\delta t))\mathrm{d}x.$$

We classically have

$$\lim_{m\to\infty} \int_\Omega u_0(x) \cdot P_{\mathcal{D}}\varphi(x,0)\mathrm{d}x = \int_\Omega u_{\mathrm{ini}}(x)\varphi(x,0)\mathrm{d}x.$$

We also easily have, thanks to the convergence properties of $u^{(m)}$, that

$$\lim_{m\to\infty} T_{21}^{(m)} = \int_0^T \int_\Omega \bar{u}(x,t)\partial_t\varphi(x,t)\mathrm{d}x\mathrm{d}t.$$

Let us prove that the term $T_{22}^{(m)}$ tends to 0 as $m \to \infty$. We have $T_{22}^{(m)} = T_{23}^{(m)} - T_{15}^{(m)}$, with

$$T_{23}^{(m)} = \sum_{n=0}^{N-1} \int_\Omega (u_{n+1}(x) - u_n(x)) \cdot P_{\mathcal{D}}\varphi(x,(n+1)\delta t)\mathrm{d}x.$$

Thanks to the limits given by (4.31), (4.32), (4.33), (4.34), and (4.35), and thanks to (4.30), we obtain that $\lim_{m\to\infty} T_{15}^{(m)} = T_{24}$, with

$$\begin{aligned} T_{24} = \; & -\nu \sum_{i=1}^{d} \int_0^T \int_\Omega \nabla u^{(i)}(x,t) \cdot \nabla \varphi^{(i)}(x,t)\mathrm{d}x\mathrm{d}t - \int_0^T b(u(\cdot,t), u(\cdot,t), \varphi(\cdot,t))\mathrm{d}t \\ & + \int_0^T \int_\Omega f(x) \cdot \varphi(x,t)\mathrm{d}x\mathrm{d}t. \end{aligned}$$

Since (4.31), (4.32), (4.33), (4.34), and (4.35) are available as well, replacing $P_{\mathcal{D}}\varphi(\cdot,n\delta t)$ by $P_{\mathcal{D}}\varphi(\cdot,(n+1)\delta t)$ in T_{16}, T_{17}, T_{18}, T_{19}, and T_{20}, we also get using (4.15) with $v = P_{\mathcal{D}}\varphi(\cdot,(n+1)\delta t)$, that $\lim_{m\to\infty} T_{23}^{(m)} = T_{24}$. Thus we get that $\lim_{m\to\infty} T_{22}^{(m)} = 0$, which concludes the proof of (4.36). Thanks to (4.30), (4.36), (4.31), (4.32), (4.33), (4.34), and (4.35), we thus obtain (4.13), provided that we can prove that

$$\mathrm{div}\bar{u}(x,t) = 0 \quad \text{for a.e. } (x,t) \in \Omega \times (0,T).$$

This last relation can be shown by following the proof of (3.22). This completes the proof of the above theorem. □

Remark 4.5. Using the above proof of convergence, we get the energy inequality for $d = 2$ or 3 from inequality (4.20), since we have the property

$$\int_0^T \int_\Omega (\nabla u^{(i)}(x,t))^2 \mathrm{d}x\mathrm{d}t \le \liminf_{m\to\infty} \sum_{n=0}^{N^{(m)}-1} \delta t \left[u_{n+\frac{1}{2}}^{(m,i)}, u_{n+\frac{1}{2}}^{(m,i)} \right]_{\mathcal{D}^{(m)}}.$$

5. Numerical results. Some simple numerical experiments are described here to observe the convergence rate of schemes (3.8) and (4.14)–(4.15) with respect to the space and time discretizations. To that purpose, we use a prototype code where the nonlinear equations are solved by an underrelaxed Newton method, and the linear

systems by a direct band Gaussian elimination solver. This code handles Stokes or Navier–Stokes problems with various boundary conditions, using nonuniform rectangular or triangular meshes on general 2D polygonal domains.

The linear Stokes equations are first considered in the case $d = 2$, $\Omega = (0,1) \times (0,1)$, $\nu = 1$, and f is taken to satisfy (3.1) with a solution equal to

$$\begin{aligned} \bar{u}^{(1)}(x^{(1)}, x^{(2)}) &= -\partial^{(2)}\Psi(x^{(1)}, x^{(2)}), \\ \bar{u}^{(2)}(x^{(1)}, x^{(2)}) &= \partial^{(1)}\Psi(x^{(1)}, x^{(2)}), \\ \bar{p}(x^{(1)}, x^{(2)}) &= 100 \ \left((x^{(1)})^2 + (x^{(2)})^2\right), \end{aligned}$$

denoting by $\Psi(x^{(1)}, x^{(2)}) = 1000\ [x^{(1)}(1 - x^{(1)})x^{(2)}(1 - x^{(2)})]^2$. The approximate solution (u, p) is computed with the scheme (3.8). The observed numerical order of convergence, considering the norms $\|u - P_{\mathcal{D}}\bar{u}\|_{L^2(\Omega)^d}$ and $\|p - P_{\mathcal{D}}\bar{p}\|_{L^2(\Omega)}$, is equal to 2 for the velocity components, and to 1 for the pressure in the cases of nonuniform rectangular and square meshes (from 400 to 6400 control volumes). Note that in these cases, there is apparently no need for a significant positive value of the stabilization coefficient λ. The observed numerical order of convergence is similar in the case of triangular meshes (from 1400 to 5600 control volumes), but values such as $\lambda = 10^{-4}$, $\alpha = 1$ have to be used in order to avoid oscillations in the pressure field. This confirms that in the case of triangles, the approximate pressure space is too large to avoid stabilization. In fact, other tests were performed (e.g., the classical backward step) which show that stabilization is also needed in the case of rectangles when more severe problems are considered. Note that in industrial implementations, stabilization may be performed by other means; see [29, 1] (see also [4] in the triangular case).

We then proceed to a similar comparison in the case of transient nonlinear problems. Considering a transient adaptation of the above steady-state analytical solution, the continuous problem is then defined by zero initial and boundary conditions, $T = 0.1$, and the function f is taken to satisfy (1.1) with a solution equal to

$$\begin{aligned} \bar{u}^{(1)}(x^{(1)}, x^{(2)}, t) &= -t\ \partial^{(2)}\Psi(x^{(1)}, x^{(2)}), \\ \bar{u}^{(2)}(x^{(1)}, x^{(2)}, t) &= t\ \partial^{(1)}\Psi(x^{(1)}, x^{(2)}), \\ \bar{p}(x^{(1)}, x^{(2)}, t) &= 100\ t\ \left((x^{(1)})^2 + (x^{(2)})^2\right), \end{aligned}$$

with the same function Ψ as above. We again observe an order 2 of convergence of the approximate solution at times $t = .05$ and $t = .1$, when the space and time discretizations are simultaneously modified with the same ratio (from $\delta t = 0.01$ to $\delta t = 0.0025$ as the size of the mesh is divided by 4). Similar observations are still valid for the classical Green–Taylor example.

6. Conclusions. The above numerical results show that the theoretical error estimate, which is proven in section 3 for the linear Stokes equations, is nonoptimal; a sharper estimate is currently being written [20] under more regularity assumptions on the mesh.

The proof of convergence of the full space-time discrete approximation of (1.1) given by (4.15) uses estimates on the time translates, which were introduced in the $L^2(\Omega \times (0,T))$ framework for the proof of convergence of the finite volume method for degenerate parabolic equations [17, 13] and used for several other cases; see, e.g., [16]. A major difficulty which arises here is the handling on the nonlinear advective term, as in the continuous case, which leads us to establish an estimate for the time translates in $L^1(0,T;L^2(\Omega))$. This new technique may be used for parabolic problems with other types of nonlinearities.

We remarked that industrial codes use other types of stabilizations than the one used here. Further works will be devoted to the mathematical study of such stabilizations, for which, to our knowledge, no proof of convergence is known up to now.

Finally, let us also mention undergoing work on a generalization of the scheme studied here to the full transient Navier–Stokes equations including the energy balance under the Boussinesq approximation.

REFERENCES

[1] F. Archambeau, N. Mehitoua, and M. Sakiz, *Code Saturne: A finite volume code for turbulent flows*, Int. J. Finite Volumes, http://averoes.math.univ-paris13.fr/JOURNAL/IJFV/, 2004.

[2] M. Benartzi, J. P. Croisille, D. Fishelov, and S. Trachtenberg, *A pure-compact scheme for the streamfunction formulation of Navier–Stokes equations*, J. Comput. Phys., 205 (2005), pp. 640–664.

[3] Ph. Blanc, R. Eymard, and R. Herbin, *A staggered finite volume scheme on general meshes for the generalized Stokes problem in two space dimensions*, Int. J. Finite Volumes, http://averoes.math.univ-paris13.fr/JOURNAL/IJFV/, 2005.

[4] S. Boivin, F. Cayré, and J. M. Hérard, *A finite volume method to solve the Navier–Stokes equations for incompressible flows on unstructured meshes*, Int. J. Therm. Sci., 38 (2000), pp. 806–825.

[5] F. Boyer and P. Fabrie, *Eléments d'Analyse pour l'Étude de Quelques Modèles d'Écoulements de Fluides Visqueux Incompressibles*, Math. Appl. 52, Springer-Verlag, 2006.

[6] F. Brezzi and J. Pitkäranta, *On the stabilization of finite element approximations of the Stokes equations*, in Efficient Solutions of Elliptic Systems, Kiel, 1984, Notes Numer. Fluid Mech. 10, Vieweg, Braunschweig, Germany, 1984, pp. 11–19.

[7] S. C. Chou, *Analysis and convergence of a covolume method for the generalized Stokes problem*, Math. Comp., 66 (1997), pp. 85–104.

[8] Y. Coudière, T. Gallouët, and R. Herbin, *Discrete Sobolev Inequalities and L^p Error Estimates for Finite Volume Solutions of Convection Diffusion Equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 67–778.

[9] K. Deimling, *Nonlinear Functional Analysis*, Springer-Verlag, Berlin, 1985.

[10] R. Eymard and T. Gallouët, *H-convergence and numerical schemes for elliptic equations*, SIAM J. Numer. Anal., 41 (2000), pp. 539–562.

[11] R. Eymard, T. Gallouët, M. Ghilani, and R. Herbin, *Error estimates for the appropriate solutions of a nonlinear hyperbolic equation given by finite volume schemes*, IMA J. Numer. Anal., 18 (1998), pp. 563–594.

[12] R. Eymard, T. Gallouët, and R. Herbin, *Convergence of finite volume approximations to the solutions of semilinear convection diffusion reaction equations*, Numer. Math., 82 (1999), pp. 91–116.

[13] R. Eymard, T. Gallouët, and R. Herbin, *Finite Volume Methods*, Handbook of Numerical Analysis, Vol. VII, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2000, pp. 713–1020.

[14] R. Eymard, T. Gallouët, and R. Herbin, *A finite volume scheme for anisotropic diffusion problems*, C. R. Math. Acad. Sci. Paris, 339 (2004), pp. 299–302.

[15] R. Eymard, T. Gallouët, and R. Herbin, *A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension*, IMA J. Numer. Anal., 26 (2006), pp. 326–353.

[16] R. Eymard, T. Gallouët, R. Herbin, and A. Michel, *Convergence of a finite volume scheme for nonlinear degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 41–82.

[17] R. Eymard, T. Gallouët, D. Hilhorst, and Y. Naït Slimane, *Finite volumes and nonlinear diffusion equations*, M2AN Math. Model. Numer. Anal., 32 (1998), pp. 747–761.

[18] R. Eymard and R. Herbin, *A cell-centered finite volume scheme on general meshes for the Stokes equations in two dimensions*, C. R. Math. Acad. Sci. Paris, 337 (2003), pp. 125–128.

[19] R. Eymard and R. Herbin, *A staggered finite volume scheme on general meshes for the Navier–Stokes equations in two space dimensions*, Int. J. Finite Volumes, http://averoes.math.univ-paris13.fr/JOURNAL/IJFV/, 2005.

[20] R. Eymard, R. Herbin, and J. C. Latché, *On a stabilized colocated finite volume scheme for the Stokes problem*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 501–527.

[21] V. Girault and P.-A. Raviart, *Finite Element Methods for the Navier–Stokes Equations. Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[22] R. Glowinski, *Numerical Methods for Fluids, (Part 3)*, Handbook of Numerical Analysis, Vol. IX, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 2003.
[23] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows—A Guide to Theory, Practice, and Algorithms*, Computer Science and Scientific Computing, Academic Press, Boston, 1989.
[24] M. D. Gunzburger and R. A. Nicolaïdes, *Incompressible Computational Fluid Dynamics,* Cambridge University Press, New York, 1993.
[25] F. H. Harlow and J. E. Welch, *Numerical calculation of time dependent viscous incompressible flow of fluids with free surface*, Phys. Fluids, 8 (1965), pp. 2182–2189.
[26] R. Herbin, *An error estimate for a finite volume scheme for a convection-diffusion equation on a triangular mesh*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 165–173.
[27] R. Herbin, *Analysis of Cell-Centered Finite Volume Methods for Incompressible Fluid Flows*, Ecole de Printemps de Mécanique des Fluides Numérique, Roscoff, France, http://www.cmi.univ-mrs.fr/~herbin/PUBLI/roscoff.ps, 2005.
[28] R. Herbin and E. Marchand, *Finite volume approximation of a class of variational inequalities*, IMA J. Numer. Anal., 21 (2001), pp. 553–585.
[29] S. R. Mathur and J. Y. Murthy, *Pressure boundary conditions for incompressible flow using unstructured meshes*, Numer. Heat Transfer, Part B, 32 (1997), pp. 283–298.
[30] J. Nečas, *Equations aux Dérivées Partielles*, Presses de l'Université de Montréal, Montreal, Canada, 1965.
[31] R. A. Nicolaïdes, *Analysis and convergence of the MAC scheme.* I. *The linear problem*, SIAM J. Numer. Anal., 29 (1992), pp. 1579–1591.
[32] R. A. Nicolaïdes and X. Wu, *Analysis and convergence of the MAC scheme.* II*., Navier–Stokes equations*, Math. Comp., 65 (1996), pp. 29–44.
[33] S. V. Patankar, *Numerical Heat Transfer and Fluid Flow*, Series in Computational Methods in Mechanics and Thermal Sciences, M. A. Philips and E. M. Millman, eds., McGraw–Hill, Washington, DC, 1980.
[34] R. Peyret and T. Taylor, *Computational Methods for Fluid Flow*, Springer-Verlag, New York, 1983.
[35] O. Pironneau, *Finite Element Methods for Fluids*, John Wiley & Sons, Chichester, UK, 1989.
[36] R. Temam, *Navier–Stokes Equations, Theory of Numerical Analysis*, Stud. Math. Appl., J. L. Lions, G. Papanicolaou, and R. T. Rockafellar, eds., North–Holland, Amsterdam, 1977.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 37–53

A REFINED GALERKIN ERROR AND STABILITY ANALYSIS FOR HIGHLY INDEFINITE VARIATIONAL PROBLEMS*

L. BANJAI† AND S. SAUTER†

Abstract. Recently, a refined finite element analysis for highly indefinite Helmholtz problems was introduced by the second author. We generalize the analysis to the Galerkin method applied to an *abstract* highly indefinite variational problem. In the refined analysis, the condition for stability and a quasi-optimal error estimate are expressed in terms of approximation properties $\mathcal{T}(S) \approx S$ and $\mathcal{T}(u+S) \approx S$. Here, u is the solution of the original variational problem, $\mathcal{T}$ is a certain continuous solution operator, and S is the finite dimensional test and trial space. The abstract analysis can be applied to both finite and boundary element solutions of high-frequency Helmholtz problems. We apply the analysis to investigate the properties of the Brakhage–Werner boundary integral formulation of the Helmholtz problem, discretized by a standard Galerkin boundary element method. In the case of scattering by the unit sphere, we derive the explicit dependence of the error and of the stability condition on the wave number k. We show that $hk \lesssim 1$ is a sufficient condition for stability and a quasi-optimal error estimate. Further, we show that the constant of quasioptimality is independent of k, which is an improvement over previously available results. Thus, the boundary element method does not suffer from the *pollution effect*.

Key words. indefinite problems, Helmholtz equation, finite and boundary element methods

AMS subject classifications. 65N30, 65N38, 65R20

DOI. 10.1137/060654177

1. Introduction.

The numerical solution of high-frequency Helmholtz problems has attracted much interest in recent years; see, for example, [3, 4, 7, 10, 11, 12, 17, 28, 29]. The main aim of this paper is to develop a refined analysis for the error and the stability of the Galerkin discretization of high-frequency Helmholtz problems. The analysis should be general enough to include both boundary and finite element methods and allow for discussion of standard and special finite/boundary elements such as the ones used in [23, 27, 29]. Most importantly, it should be possible to obtain optimal results on the dependence of the error bounds and the stability condition on the wave number k. The explicit dependence on k is rarely given in existing literature; for exceptions, see [8, 11, 13].

It is well known that the Galerkin finite element method with standard piecewise polynomial basis functions suffers from the so-called *pollution effect* [3]. If piecewise linear basis functions are used, the stability condition in the mesh width h is very strong: $hk^2 \lesssim 1$. In [3], a generalized finite element method was presented in one dimension, with the stability condition reduced to $hk \lesssim 1$; see also [17]. The proofs rely on explicit knowledge of the Green's function and, hence, do not carry over to higher dimensions. Further, the general stability and convergence analysis given in [23] does not yield the improved stability condition. With this in mind, in [29] a refined finite element analysis was developed that gives improved stability and error estimates.

In this paper, we generalize the results of [29] to an abstract theory applicable to a general indefinite variational problem. We prove that the condition $\mathcal{T}(S) \approx S$,

*Received by the editors March 16, 2006; accepted for publication (in revised form) July 5, 2006; published electronically January 8, 2007.
http://www.siam.org/journals/sinum/45-1/65417.html
†Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland (lehelb@math.unizh.ch, stas@math.unizh.ch).

of approximate invariance of the test and trial space S under a certain continuous solution operator $\mathcal{T}$, is sufficient for stability. The quasi-optimal error estimate is proved under a similar condition $\mathcal{T}(u+S) \approx S$, where u is the solution of the continuous variational problem. This new concept is the crux of the abstract analysis we develop. We describe how the abstract analysis can be used to prove the results of [29] for the finite element method. As a further example of its use, we consider the boundary element method for the solution of high-frequency Helmholtz problems using the Brakhage–Werner boundary integral formulation. This problem has already been considered in [13] and recently in [11]. There, the stability condition $hk \lesssim 1$ and a quasi-optimal error estimate, with the constant of quasi-optimality proportional to $k^{1/3}$, was proved for the case of the unit sphere. In [11], the authors consider the problem of high-frequency scattering by a convex object in two dimensions. Known asymptotics of the scattered wave were used to reduce the problem to the computation of unknown amplitudes, which are less oscillatory than the original scattered wave. These were then computed using a Galerkin method for which the quasi-optimal error with constant of $\mathcal{O}(k^{1/3})$ was proved in the case of the unit disk and sphere.

We obtain a sharper error estimate, with the quasi-optimality constant independent of k. More importantly, our paper provides a framework in which to investigate the properties of boundary element methods with special basis elements such as plane waves [27]. For special finite element methods, it was already shown in [29] that the refined analysis obtains results outside the reach of standard analyses. We give reasons to expect the same to be true for boundary element methods. Further, the condition of the approximability of $\mathcal{T}(S)$ and $\mathcal{T}(u+S)$ by the boundary element space can give guidelines for the construction of special boundary elements.

2. A highly indefinite variational problem. Let H and V be Hilbert spaces such that H is continuously imbedded in V and, hence, V' is continuously imbedded in H', where V' and H' are the dual spaces; see [33]. Denote by $(\cdot,\cdot)_H$ and $(\cdot,\cdot)_V$ the respective inner products, and by $\|\cdot\|_H$ and $\|\cdot\|_V$ the induced norms.

We are interested in the following abstract variational problem: Given $f \in H'$, find $u \in H$ such that

$$a(u,v) = \langle f, v\rangle \quad \text{for all } v \in H, \tag{2.1}$$

where $a(\cdot,\cdot) : H \times H \to \mathbb{C}$ and we have written $\langle f, v\rangle = f(v)$ for the value of the functional f at v.

Naturally, we need to place some conditions on the above problem.

Assumptions.

A1: $a(\cdot,\cdot) : H \times H \to \mathbb{C}$ is a bounded sesquilinear form. Thus, $a(u,v)$ is linear in u, conjugate linear in v, and

$$|a(u,v)| \le C_{\mathrm{c}} \|u\|_H \|v\|_H.$$

A2: There exist bounded sesquilinear forms $a_H(\cdot,\cdot) : H \times H \to \mathbb{C}$ and $a_V(\cdot,\cdot) : V \times V \to \mathbb{C}$ such that

$$a(u,v) = a_H(u,v) + a_V(u,v)$$

and

$$|a_H(u,u)| \ge \alpha_H \|u\|_H^2, \quad |a_V(u,v)| \le C_V \|u\|_V \|v\|_V \quad \text{for any } u, v \in H.$$

A3: Problem (2.1) and its adjoint have a unique solution $u \in H$. Further,

$$\|u\|_H \leq C_{\mathrm{reg}} \|f\|_{H'}.$$

The sesquilinear forms $a(\cdot,\cdot)$, $a_H(\cdot,\cdot)$, and $a_V(\cdot,\cdot)$ define the corresponding bounded linear operators:

$$A : H \to H', \quad A_H : H \to H', \quad \text{and } A_V : V \to V'. \tag{2.2}$$

In view of A3, the inverses of A and the adjoint A^* are also bounded linear operators:

$$A^{-1} : H' \to H \quad \text{and} \quad A^{*-1} : H' \to H. \tag{2.3}$$

We now investigate the properties of the Galerkin discretization of (2.1).

2.1. Abstract stability and convergence analysis of the Galerkin method. Let $S \subset H$ be a finite dimensional subspace of H. We wish to consider the Galerkin discretization of problem (2.1): Given $f \in H'$, find $u_S \in S$ such that

$$a(u_S, v) = \langle f, v \rangle \quad \text{for all } v \in S. \tag{2.4}$$

We now derive a condition on S that guarantees the existence and uniqueness of u_S and a quasi-optimal error estimate.

2.1.1. Stability and convergence. For our analysis of the stability and convergence of (2.4), the following continuous dual problem will be crucial: Given $w \in H$, let $z \in H$ be such that

$$a(v, z) = -a_V(w, v) \quad \text{for all } v \in H.$$

From (A2) it follows that $a_V(w,\cdot)$ defines a bounded linear functional on V. Since H is continuously imbedded in V, i.e., the identity mapping $I : H \to V$ is continuous, $a_V(w,\cdot)$ defines also a bounded linear functional on H. Therefore, we can apply (A3) to obtain that the solution $z \in H$ of the above adjoint problem exists and is unique. Consequently, we can define a solution operator by $\mathcal{T} w := z$. Using again the fact that H is continuously imbedded in V and the properties of the operators in (2.2) and (2.3), we conclude that the solution operator $\mathcal{T} = -A^{*-1} A_V$ is a bounded linear operator mapping from H to H. Hence, there exists a constant $C_{\mathcal{T}}$ such that

$$\|\mathcal{T} u\|_H \leq C_{\mathcal{T}} \|u\|_H \quad \text{for all } u \in H. \tag{2.5}$$

Remark 1. *In applications, the operator $\mathcal{T}$ will be a compact operator. Usually it is also a smoothening operator; see Remark* 5 *and* [29].

Let us now define a measure of approximability in the space S. This measure depends on some subset $\widetilde{H} \subseteq H$, which satisfies $S \subset \widetilde{H}$ and $u + S \subset \widetilde{H}$, where u is the exact solution of (2.1). The measure is defined by

$$\eta(S) := \sup_{w \in \widetilde{H} \setminus \{0\}} \inf_{v \in S} \frac{\|\mathcal{T} w - v\|_H}{\|w\|_H}. \tag{2.6}$$

Remark 2.

1. *For a dense sequence $(S_l)_{l \geq 1}$ of spaces, i.e., $\overline{\cup_l S_l}^{\|\cdot\|_H} = H$, we have* $\lim_{l \to \infty} \eta(S_l) = 0$.

2. *We will prove stability of* (2.4) *and a quasi-optimal error estimate, under the condition that* $\eta(S)$ *is small enough.*
3. *Note that the choice* $\widetilde{H} = H$ *is always possible. However, a choice of a smaller set* $\widetilde{H} \subsetneq H$ *might result in a smaller value of* $\eta(S)$ *and a less restrictive stability condition.*

THEOREM 2.1. *Let* S *be such that*

$$\eta(S) \leq \frac{\alpha_H}{2C_c}, \tag{2.7}$$

and let $u \in H$ *be the solution of* (2.1)*. Then there exists a unique solution* $u_S \in S$ *of the discrete problem* (2.4)*. Moreover,*

$$\|u - u_S\|_H \leq \frac{2C_c}{\alpha_H} \inf_{v \in S} \|u - v\|_H.$$

Proof. Since S is finite dimensional, it suffices to prove uniqueness. Given $w_S \in S$, let z_S be the best approximation to $z = \mathcal{T} w_S$ with respect to the H-norm. Then,

$$\begin{aligned} |a(w_S, w_S + z_S)| &= |a_H(w_S, w_S) - a(w_S, z - z_S)| \geq \alpha_H \|w_S\|_H^2 - C_{\mathrm{c}} \|w_S\|_H \|z - z_S\|_H \\ &\geq \alpha_H \|w_S\|_H^2 - C_{\mathrm{c}} \eta(S) \|w_S\|_H^2. \end{aligned}$$

From (2.5) we have that

$$\|z\|_H \leq C_{\mathcal{T}} \|w_S\|_H$$

and hence

$$\|w_S + z_S\|_H \leq \|w_S\|_H + \|z\|_H + \|z - z_S\|_H \leq (1 + C_{\mathcal{T}} + \eta(S)) \|w_S\|_H.$$

Using (2.7), we have that

$$|a(w_S, w_S + z_S)| \geq \frac{\alpha_H}{2} \|w_S\|_H^2 \geq \frac{\alpha_H}{2 + 2C_{\mathcal{T}} + 2\eta(S)} \|w_S\|_H \|w_S + z_S\|_H.$$

Hence, we have the discrete inf-sup condition

$$\inf_{u \in S \setminus \{0\}} \sup_{v \in S \setminus \{0\}} \frac{|a(u, v)|}{\|u\|_H \|v\|_H} \geq \frac{\alpha_H}{2 + 2C_{\mathcal{T}} + 2\eta(S)} > 0,$$

and we have proved that the discrete solution u_S exists and is unique.

Next, let $z' = \mathcal{T} e$, where $e = u - u_S$, and again let z'_S be the best approximation to z' in the H-norm. Then,

$$|a_V(e, e)| = |a(e, z')| = |a(e, z' - z'_S)| \leq C_{\mathrm{c}} \eta(S) \|e\|_H^2.$$

Hence, for any $v \in S$,

$$\begin{aligned} \alpha_H \|e\|_H^2 &\leq |a_H(e, e)| = |a(e, e) - a_V(e, e)| = |a(e, u - v) - a_V(e, e)| \\ &\leq C_{\mathrm{c}} \|e\|_H \|u - v\|_H + C_{\mathrm{c}} \eta(S) \|e\|_H^2. \end{aligned}$$

Therefore, using (2.7),

$$\|e\|_H \leq \frac{2C_{\mathrm{c}}}{\alpha_H} \|u - v\|_H \quad \text{for any } v \in S.$$

Thus, we have also proved the quasioptimality of the Galerkin method. ∎

REMARK 3. *A result on the stability and convergence of the Galerkin finite element method applied to an indefinite PDE can be found in Theorem* 5.7.6 *of* [6]. *The same constant of quasioptimality* $2C_c/\alpha_H$, *as above, is also given in* [6]*; this is an improvement over the usual estimate given by Céa's lemma; see Remark* 6. *The essential novelty of our concept is that for stability and convergence it is sufficient to have* $\mathcal{T}(S) \approx S$ *and* $\mathcal{T}(u+S) \approx S$. *In contrast, the approach taken in* [6] *requires that the adjoint problem have full regularity. Theorem* 2.1 *is a stronger result, which implies the result of* [6]. *In particular, the kind of condition given in* [6] *does not allow for improved stability estimates of* [29]*; for details see* [29].

2.1.2. Error estimate in the V-norm. By using the Aubin–Nitsche technique, we can bound the V-norm of the error by the H-norm of the error. Let $\psi \in H$ be such that

$$a(v,\psi) = (e,v)_V \quad \text{for all } v \in H.$$

Let $\mathcal{S} : H \to H$ be the solution operator defined by $\mathcal{S}e := \psi$, and let

$$\mu(S) := \sup_{w \in \widetilde{H}\setminus\{0\}} \inf_{v\in S} \frac{\|\mathcal{S}w - v\|_H}{\|w\|_V}.$$

If ψ_S is the best approximation to ψ with respect to the H-norm, then

$$\|e\|_V^2 = a(e,\psi) = a(e,\psi-\psi_S) \le C_c\mu(S)\|e\|_H\|e\|_V. \tag{2.8}$$

Hence, we have an estimate of the V-norm of the error in terms of the H-norm of the error. We proceed now to obtain an alternative condition to that given in Theorem 2.1 for the existence of a quasi-optimal error estimate. For any $v \in H$,

$$\begin{aligned}\alpha_H\|e\|_H^2 &\le |a_H(e,e)| = |a(e,e) - a_V(e,e)| \le C_c\|e\|_H\|u-v\|_H + C_V\|e\|_V^2\\ &\le C_c\|e\|_H\|u-v\|_H + C_V\left(C_c\mu(S)\right)^2\|e\|_H^2.\end{aligned}$$

Hence, under the alternative condition

$$C_V\left(C_c\mu(S)\right)^2 < \alpha_H/2,$$

we have obtained the same quasi-optimal estimate as before. The results are collected in the following theorem.

THEOREM 2.2. *Let* $u \in H$ *be the solution of* (2.1) *and* $u_S \in S$ *be a solution of* (2.4). *Then*

$$\|u-u_S\|_V \le C_c\mu(S)\|u-u_S\|_H.$$

Further, if S *is such that* $C_V\left(C_c\mu(S)\right)^2 < \alpha_H/2$, *then*

$$\|u-u_S\|_H \le \frac{2C_c}{\alpha_H}\inf_{v\in S}\|u-v\|_H.$$

REMARK 4. *An abstract indefinite problem similar to the one we investigate here has been considered by Schatz in* [31]. *As an assumption of the abstract problem, Schatz imposes a condition of the type* (2.8) *with* $\mu(S) \to 0$ *for* $\dim(S) \to \infty$*; see* [31, (12)]. *This is not possible if* $V = H$, *which is the case of the boundary integral equation considered in section* 3*; hence the results of* [31] *do not apply, and Theorem* 2.1 *needs to be used. Further in* [31] *the constant of quasioptimality is not investigated.*

2.2. An example application in a finite element setting. The abstract analysis given here is a generalization of the finite element analysis for highly indefinite Helmholtz problems introduced in [29]. The appropriate choice of spaces H and V for the finite element method in [29] is

$$H = H^1(\Omega), \quad V = L^2(\Omega),$$

where the space H is equipped with a weighted norm (cf. [23]):

$$\|u\|_{\mathcal{H}} := \left(|u|_{1,\Omega}^2 + k^2\|u\|_{0,\Omega}^2\right)^{1/2}.$$

With this choice of spaces, the assumptions A1–A3 are proved in [29]. Theorems 2.2 and 2.5 of [29] are then implied by Theorems 2.1 and 2.2, respectively. For details we refer the reader to [29].

We now turn to another case to which the abstract theory can be applied. Namely, we consider the solution of a Helmholtz problem by a Galerkin boundary element method.

3. A Helmholtz scattering problem. Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 2, 3$, with a smooth boundary Γ. We consider the following exterior Helmholtz problem: Given $g \in H^{1/2}(\Gamma)$, find $u \in H^1_{\mathrm{loc}}(\Omega^c)$ such that

$$\begin{aligned} -\Delta u - k^2 u &= 0 && \text{in } \Omega^c, \\ u &= g && \text{on } \Gamma, \\ \lim_{r\to\infty} r^{(d-1)/2}\left(\frac{\partial u}{\partial r} - iku\right) &= 0, && \text{where } r := \|x\|, \end{aligned} \tag{3.1}$$

is satisfied in a weak sense. The equation governs the process of acoustic scattering by a sound soft object; see [25].

Let $G_k(\cdot)$ be the fundamental solution of the Helmholtz equation:

$$G_k(r) = \frac{i}{4}H_0(kr), \quad \text{for } d = 2,$$
$$G_k(r) = \frac{1}{4\pi}\frac{e^{ikr}}{r}, \quad \text{for } d = 3,$$

with $r > 0$. Throughout the paper H_ν is the Hankel function of the first kind of order ν defined by

$$H_\nu(x) := J_\nu(x) + i\,Y_\nu(x), \quad x > 0,$$

where J_ν and Y_ν are the Bessel functions of the first and second kind. Employing the fundamental solution, we define, respectively, the single layer and the double layer integral operators:

$$(S_k\varphi)(x) := \int_\Gamma G_k(\|x-y\|)\varphi(y)d\Gamma_y, \quad x \in \mathbb{R}^d \setminus \Gamma, \tag{3.2}$$

$$(D_k\varphi)(x) := \int_\Gamma \frac{\partial}{\partial n_y}G_k(\|x-y\|)\varphi(y)d\Gamma_y, \quad x \in \mathbb{R}^d \setminus \Gamma, \tag{3.3}$$

where n_y is the unit normal to the surface Γ at the point $y \in \Gamma$. The corresponding boundary integral operators are defined by

$$(V_k\varphi)(x) := \int_\Gamma G_k(\|x-y\|)\varphi(y)d\Gamma_y, \quad x \in \Gamma, \tag{3.4}$$

$$(K_k\varphi)(x) := \int_\Gamma \frac{\partial}{\partial n_y}G_k(\|x-y\|)\varphi(y)d\Gamma_y, \quad x \in \Gamma. \tag{3.5}$$

We now state the well-known mapping properties of the above operators; see [9, 30].

PROPOSITION 3.1. *Let $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, be a bounded domain with smooth boundary Γ. Then for any $s \in \mathbb{R}$ the following are bounded linear operators:*

(a) $V_k : H^s(\Gamma) \to H^{s+1}(\Gamma)$,

(b) $K_k : H^s(\Gamma) \to H^{s+1}(\Gamma)$.

It is well known that every solution $\varphi \in H^{-1/2}(\Gamma)$ of $V_k\varphi = g$ has the property that $u = S_k\varphi$ satisfies the exterior Helmholtz problem (3.1). However, for countably many wave numbers k the operator V_k is not injective. To avoid this problem Brakhage and Werner [5], Leis [22], and Panič [26], independently suggested representing the solution as a combination of the single and double layer potentials,

$$u = D_k\varphi - i\alpha S_k\varphi, \tag{3.6}$$

for some coupling parameter $\alpha > 0$. The unknown density φ in (3.6) satisfies the boundary integral equation

$$g = \left(\frac{1}{2}I + K_k - i\alpha V_k\right)\varphi, \tag{3.7}$$

where I is the identity operator. We denote by $(\cdot,\cdot)_0$ the $L^2(\Gamma)$ inner product, and by $\|\cdot\|_0$ the corresponding norm, and define

$$a(\varphi, v) := (R_k\varphi, v)_0, \quad \text{where } R_k := \frac{1}{2}I + K_k - i\alpha V_k. \tag{3.8}$$

To be able to apply the abstract theory developed in section 2, we need to prove that the assumptions A1–A3 hold in this case. Proposition 3.1 implies that the condition A1 is satisfied with the choice $H = L^2(\Gamma)$. We can then define

$$a_{_H}(\varphi, v) := \frac{1}{2}(I\varphi, v)_0 \quad \text{and} \quad a_{_V}(\varphi, v) := (\widetilde{R}_k\varphi, v)_0, \qquad \text{where } \widetilde{R}_k := K_k - i\alpha V_k.$$

Therefore, $A := R_k$, $A_{_H} := \frac{1}{2}I$, and $A_{_V} := \widetilde{R}_k$. Again by Proposition 3.1, it follows that the condition A2 holds with the choice $V = L^2(\Gamma)$; trivially, V is then continuously imbedded in H. Furthermore, we can clearly set $\alpha_{_H} = 1/2$. The following proposition deals with assumption A3.

PROPOSITION 3.2. *Let $\Omega \subset \mathbb{R}^d$ be a bounded domain with smooth boundary Γ. Then, for any $g \in L^2(\Gamma)$ there exists a unique $\varphi \in L^2(\Gamma)$ such that*

$$a(\varphi, v) = (g, v)_0 \quad \textit{for all } v \in L^2(\Gamma), \tag{3.9}$$

and there exists a constant $C_{reg} > 0$, which depends on both k and Ω, such that

$$\|\varphi\|_0 \le C_{reg}\|g\|_0.$$

Moreover,

$$u = (D_k\varphi) - i\alpha\,(S_k\varphi)$$

is the solution of the Helmholtz problem (3.1).

Proof. In the original paper of Brakhage and Werner [5], the existence and uniqueness were proved for the classical formulation. To extend the proof to the variational

formulation we proceed as in [13].[1] Since $\widetilde{R}_k$ is a continuous operator from $L^2(\Gamma)$ to $H^1(\Gamma)$, and $H^1(\Gamma)$ is compactly imbedded in $L^2(\Gamma)$, we have that $\widetilde{R}_k$ is a compact operator from $L^2(\Gamma)$ to $L^2(\Gamma)$. Therefore we can apply the Fredholm–Riesz–Schauder theory to the operator $R_k = I/2 + \widetilde{R}_k$, which implies that to prove invertibility it suffices to prove injectivity; i.e., it suffices to prove that $\operatorname{Ker} R_k = \{0\}$.

Let $R_k\varphi = 0$; then $\varphi = -2\widetilde{R}_k\varphi$. Applying the mapping property $\widetilde{R}_k : H^s(\Gamma) \to H^{s+1}(\Gamma)$ twice, we have that $\varphi \in H^2(\Gamma)$ and is hence continuous. For continuous functions the proof of uniqueness given in [5] is applicable, therefore $\varphi = 0$. ∎

To find an approximation to the solution φ numerically, we use the Galerkin discretization. Let S be a finite dimensional subset of $L^2(\Gamma)$. Then, find a $\varphi_S \in S$ such that

$$a(\varphi_S, v) = (g, v)_0 \quad \text{for all } v \in S. \tag{3.10}$$

Since we have checked that all the assumptions of the abstract theory hold, from Theorem 2.1 we immediately obtain the following result.

COROLLARY 3.3. *Let S be such that $C_c\eta(S) \le 1/4$. Then* (3.10) *has a unique solution $\varphi_S \in L^2(\Gamma)$ and*

$$\|\varphi - \varphi_S\|_0 \le 4C_c \inf_{v\in S} \|\varphi - v\|_0,$$

where $\varphi \in L^2(\Gamma)$ is the solution of (3.9).

REMARK 5. *Recall the definition of $\mathcal{T}$ from the previous section. Since $\mathcal{T} = R_k^{*-1}\widetilde{R}_k$, from Proposition* 3.1 *we have that $\mathcal{T} : L^2(\Gamma) \to H^1(\Gamma)$; therefore, $\mathcal{T}$ is a smoothening operator. To emphasize the dependence of $\mathcal{T}$ on k, for the rest of the paper we denote it by $\mathcal{T}_k := \mathcal{T}$.*

We will later show that for the case of $\Omega = \mathbb{S}^2$ and a particular choice of the coupling parameter α, the constant C_c is independent of k. The result of Theorem 2.2 brings little new in this setting, since $V = H$. For the finite element method of [29], Theorem 2.2 is of more interest.

So far we have made no specification for the set S except that it is a finite dimensional subspace of $L^2(\Gamma)$. Next, we consider the special case of the usual piecewise polynomial boundary elements.

3.1. Piecewise polynomial boundary elements. Let $\mathcal{G}$ be a shape-regular triangulation of Γ. We assume that no approximation of the boundary occurs; namely,

$$\Gamma = \bigcup_{\tau\in\mathcal{G}} \tau.$$

The mesh width h is defined to be

$$h := \max\{h_\tau : \tau \in \mathcal{G}\}, \quad \text{where } h_\tau := \sup_{x,y\in\tau} \|x - y\|.$$

The set S is then defined to be a space of piecewise polynomial functions on the triangulation $\mathcal{G}$. In particular we are interested in the space $\mathcal{S}^{0,-1}_{\mathcal{G},h}$ of functions constant on each triangle $\tau \in \mathcal{G}$.

Next we give the well-known approximation property of the piecewise-constant finite element spaces.

[1] In [13] a weaker assumption is made on the smoothness of Γ but stronger on the spaces: $\Gamma \in C^{2,\lambda}$, $0 < \lambda < 1$, and $u, f \in H^{1/2}(\Gamma)$.

THEOREM 3.4. *Let $\varphi \in H^1(\Gamma)$ and $S = \mathcal{S}^{0,-1}_{\mathcal{G},h}$. There exists a constant C_A, which depends only on the minimal angle of the triangulation $\mathcal{G}$, such that*

$$\inf_{v\in S} \|\varphi - v\|_0 \le C_A h \|\varphi\|_1.$$

We now proceed to investigate the dependence of the stability and the Galerkin error on the wave number. To do this, we make the assumption that the derivatives of the solution grow proportionally with the wave number k.

DEFINITION 3.5. *For a given $\rho > 0$, the set $\mathcal{O}_{\rho,k,l}$ contains functions $\varphi \in H^l(\Gamma)$ such that*

$$\|\varphi\|_l \le \rho k^l \|\varphi\|_0.$$

The conditions under which the solution of (3.9) belongs to a class $\mathcal{O}_{\rho,k,l}$ are discussed in [8].

COROLLARY 3.6. *Let $S = \mathcal{S}^{0,-1}_{\mathcal{G},h}$, and let $\varphi \in L^2(\Gamma)$ be the solution of (3.9). If*

$$C_c C_A h \|\mathcal{T}_k\|_{H^1(\Gamma)\leftarrow L^2(\Gamma)} < 1/4,$$

the discrete problem (3.10) has a unique solution $\varphi_S \in S$. If, further, $\varphi \in \mathcal{O}_{\rho,k,1}$ and $\varphi \neq 0$, then the relative error is bounded as

$$\frac{\|\varphi - \varphi_S\|_0}{\|\varphi\|_0} \le 4 C_c C_A h k.$$

Proof. Using the approximation property of the piecewise-constant space and choosing $\widetilde{H} = H = L^2(\Gamma)$, we have that

$$\eta(S) = \sup_{\varphi\in L^2(\Gamma)\setminus\{0\}} \inf_{v\in S} \frac{\|\mathcal{T}_k\varphi - v\|_0}{\|\varphi\|_0} \le C_A \sup_{\varphi\in L^2(\Gamma)\setminus\{0\}} \frac{h\|\mathcal{T}_k\varphi\|_1}{\|\varphi\|_0} \le \frac{1}{4C_c}.$$

Hence, by Corollary 3.3, we have the required stability condition.

Let us now assume that $\varphi \in \mathcal{O}_{\rho,k,l}$. Using Corollary 3.3 again,

$$\|\varphi - \varphi_s\|_0 \le 4C_{\mathrm{c}} \inf_{v\in S} \|\varphi - v\|_0 \le 4C_c C_A h \|\varphi\|_1 \le 4C_c C_A h k \|\varphi\|_0. \qquad \square$$

In the next section we investigate the dependence of C_{c} and of $\|\mathcal{T}_k\|_{H^1(\Gamma)\leftarrow L^2(\Gamma)}$ on the wave number k. Our goal is to state the dependence on k of all the constants in Corollary 3.6 in the case of the sphere.

3.2. The special case of the unit sphere. In this section we restrict our discussion to the case $\Gamma = \mathbb{S}^2$. This case was investigated by Giebermann in [13] and by Domínguez, Graham, and Smyshlyaev in [11]. Our final result will be a slight improvement on the results of [13] and [11]. The improvement is in part due to the abstract theory developed at the start of the paper and in part due to some stronger bounds on the eigenvalues that we prove; the details are stated in Remark 6.

The Fourier coefficients of a function $f \in L^2(\mathbb{S}^2)$ are defined by

$$f_n^m := \int_{\mathbb{S}^2} Y_n^m(\hat{x})\overline{f(\hat{x})} ds_x, \tag{3.11}$$

where Y_n^m are the spherical harmonics; see [1]. Spaces equivalent to the usual Sobolev spaces on $\mathbb{S}^2$ can be defined through the Fourier coefficients.

DEFINITION 3.7. *For any $s \geq 0$, let $\mathcal{H}^s(\mathbb{S}^2)$ be the space containing all functions $f \in L^2(\mathbb{S}^2)$ whose Fourier coefficients satisfy*

$$\sum_{n=0}^{\infty} \sum_{m=-n}^{n} |f_n^m|^2 (1+n^2)^s < \infty.$$

The inner product is defined by

$$\langle f, g \rangle_s := \sum_{n=0}^{\infty} (1+n^2)^s \sum_{m=-n}^{n} f_n^m \overline{g_n^m}.$$

For negative s, $\mathcal{H}^s(\mathbb{S}^2)$ is the dual space of $\mathcal{H}^{-s}(\mathbb{S}^2)$.

In the following, j_n, y_n, and $h_n^{(1)}$ are spherical Bessel functions of the first, second, and third kind, respectively; see [1]. These can be defined through the Bessel functions

$$\begin{aligned} j_n(x) &:= \sqrt{\frac{\pi}{2x}} J_{n+\frac{1}{2}}(x), \\ y_n(x) &:= \sqrt{\frac{\pi}{2x}} Y_{n+\frac{1}{2}}(x), \\ h_n^{(1)}(x) &:= j_n(x) + i y_n(x) = \sqrt{\frac{\pi}{2x}} H_{n+\frac{1}{2}}(x). \end{aligned} \tag{3.12}$$

LEMMA 3.8.

(a) *The space $\mathcal{H}^s(\mathbb{S}^2)$ is a Hilbert space and is equivalent to $H^s(\mathbb{S}^2)$. Namely, the norms induced by the inner products are equivalent, and the sets $\mathcal{H}^s(\mathbb{S}^2)$ and $H^s(\mathbb{S}^2)$ coincide.*

(b) *The spherical harmonics form a complete orthogonal system in $\mathcal{H}^s(\mathbb{S}^2)$ and are the eigenfunctions of operators V_k, K_k, R_k, and $\mathcal{T}_k$. We have that*

$$\begin{aligned} V_k Y_n^m &= \lambda_{n,k}^{(V)} Y_n^m, \quad \text{with } \lambda_{n,k}^{(V)} := 2ikh_n^{(1)}(k) j_n(k), \\ K_k Y_n^m &= \lambda_{n,k}^{(K)} Y_n^m, \quad \text{with } \lambda_{n,k}^{(K)} := -1/2 + ik^2 h_n^{(1)}(k) j_n'(k), \\ R_k Y_n^m &= \lambda_{n,k}^{(R)} Y_n^m, \quad \text{with } \lambda_{n,k}^{(R)} := 1/2 + \lambda_{n,k}^{(K)} - i\alpha \lambda_{n,k}^{(V)} \\ &\qquad\qquad = ik^2 h_n^{(1)}(k) j_n'(k) + 2\alpha k h_n^{(1)}(k) j_n(k). \\ \mathcal{T}_k Y_m^n &= R_k^{*-1} \widetilde{R}_k Y_m^n = \lambda_{n,k}^{(\mathcal{T})} Y_n^m, \quad \text{with } \lambda_{n,k}^{(\mathcal{T})} := \frac{\lambda_{n,k}^{(K)} - i\alpha \lambda_{n,k}^{(V)}}{1/2 + \overline{\lambda_{n,k}^{(K)}} + i\alpha \overline{\lambda_{n,k}^{(V)}}}. \end{aligned}$$

(c) *For $s \geq 0$,*

$$\|R_k\|_{\mathcal{H}^s(\mathbb{S}^2) \leftarrow \mathcal{H}^s(\mathbb{S}^2)} = \sup_{n \in \mathbb{N}_0} |\lambda_{n,k}^{(R)}|, \quad \|\mathcal{T}_k\|_{\mathcal{H}^{s+1}(\mathbb{S}^2) \leftarrow \mathcal{H}^s(\mathbb{S}^2)} = \sup_{n \in \mathbb{N}_0} \sqrt{1+n^2} |\lambda_{n,k}^{(\mathcal{T})}|.$$

Proof. For the proof of (a) see [24]. The eigenvalues of the operators V_k and K_k are given in [19]. From these it is easy to derive the eigenvalues of the remaining two operators. A proof of (c) can be found in [13]; see also [24]. ∎

The above result justifies our writing $H^s(\mathbb{S}^2)$ for both $H^s(\mathbb{S}^2)$ and $\mathcal{H}^s(\mathbb{S}^2)$. We now prove some results on the Bessel functions that, in view of (3.12) and Lemma 3.8, have direct use in bounding eigenvalues $\lambda_{n,k}^{(R)}$. Recall that the Bessel functions $J_\nu(x)$ and $Y_\nu(x)$ are real valued for $\nu \in \mathbb{R}$ and $x \geq 0$.

LEMMA 3.9.
(a) $J_\nu(x), J'_\nu(x), Y'_\nu(x) > 0$, $Y_\nu(x) < 0$, *for* $0 < x < \nu$,
(b) $J_\nu(x)$ *and* $xJ'_\nu(x)$ *are positive increasing functions of* x, *for* $0 < x < \nu$,
(c) *for* $x > 0$ *the product* $x\left[J_\nu^2(x) + Y_\nu^2(x)\right]$, *as a function of* x, *decreases monotonically if* $\nu > 1/2$, *and increases monotonically if* $\nu < 1/2$.

Proof. Parts (a) and (b) are proved in Watson [32, section 15.3]. A proof of part (c) can also be found in Watson [32, section 13.74]. ∎

PROPOSITION 3.10. *There exists a constant* $C > 0$ *such that for any* $x \geq 1$ *and* $\nu \in [1/2, \infty) \cup \{0\}$,
(a) $|J_\nu(x)H_\nu(x)| \leq Cx^{-2/3}$,
(b) $|xJ'_\nu(x)H_\nu(x)| \leq C$.

Proof. A proof of part (a) for $\nu > 1/2$ is given in [13] and [11], where also a bound that is less sharp than what we prove here is given for part (b).

In the proof we make use of the following asymptotic expansions [1, (9.3.31)–(9.3.34)]:

$$\begin{aligned} J_\nu(\nu) &= a\nu^{-1/3} + O(\nu^{-5/3}), \\ Y_\nu(\nu) &= -\sqrt{3}a\nu^{-1/3} + O(\nu^{-5/3}), \\ J'_\nu(\nu) &= b\nu^{-2/3} - c\nu^{-4/3} + O(\nu^{-8/3}), \\ Y'_\nu(\nu) &= \sqrt{3}\big(b\nu^{-2/3} + c\nu^{-4/3}\big) + O(\nu^{-8/3}), \end{aligned} \tag{3.13}$$

where a, b, and c are certain positive constants.

We divide the proof into two cases, as follows.

Case 1: $\nu > x \geq 0$. Using the identity $J_\nu(x)Y'_\nu(x) - J'_\nu(x)Y_\nu(x) = 2/(\pi x)$ [1, (9.1.16)], we have that

$$0 \overset{\text{Lemma 3.9(a)}}{\leq} J_\nu(x)Y'_\nu(x) \overset{[1,\ (9.1.16)]}{=} \frac{2}{\pi x} + J'_\nu(x)Y_\nu(x).$$

Therefore,

$$|xJ'_\nu(x)Y_\nu(x)| \overset{\text{Lemma 3.9(a)}}{=} -xJ'_\nu(x)Y_\nu(x) \leq \frac{2}{\pi}.$$

Also,

$$|xJ'_\nu(x)J_\nu(x)| = xJ'_\nu(x)J_\nu(x) \overset{\text{Lemma 3.9(b)}}{\leq} \nu J'_\nu(\nu)J_\nu(\nu) \overset{(3.13)}{\leq} C,$$

where C is independent of x and ν. Combining the last two results, we have that

$$|xJ'_\nu(x)H_\nu(x)| \leq |xJ'_\nu(x)J_\nu(x)| + |xJ'_\nu(x)Y_\nu(x)| \leq C + \frac{2}{\pi} \quad \text{for } x < \nu. \tag{3.14}$$

Case 2: $1/2 < \nu \leq x$. We use the following definitions:

$$M_\nu(x) := |H_\nu(x)| \quad \text{and} \quad N_\nu(x) := |H'_\nu(x)|.$$

We have that

$$x^2|J'_\nu(x)H_\nu(x)|^2 \leq x^2N_\nu^2(x)M_\nu^2(x) \overset{[1,\ (9.2.22)]}{=} x^2{M'_\nu}^2(x)M_\nu^2(x) + \frac{4}{\pi}. \tag{3.15}$$

Next,

$$\begin{aligned} x\frac{d}{dx}\{-xM_\nu'(x)\} &\overset{[1,\ (9.2.25)]}{=} (x^2-\nu^2)M_\nu(x) - \frac{4}{\pi^2}\frac{1}{M_\nu^3(x)} = M_\nu(x)\left(x^2-\nu^2-\frac{4}{\pi^2}M_\nu^{-4}(x)\right) \\ &\overset{[14,\ (8.479)]}{\leq} M_\nu(x)\left(x^2-\nu^2-x^2\right) \leq 0. \end{aligned}$$

Hence, $-xM_\nu'(x)$ is a monotonically decreasing function. From Lemma 3.9(c) we have that, for $\nu > 1/2$, $xM_\nu^2(x)$ is monotonically decreasing, and hence $M_\nu'(x) \leq 0$. It is now not difficult to see that $xM_\nu'^2(x)$ is also a monotonically decreasing function. Therefore,

$$xM_\nu'^2(x)xM_\nu^2(x) \leq \nu^2 M_\nu'^2(\nu)M_\nu(\nu)^2 \overset{(3.13)}{\leq} C \quad \text{for } x \geq \nu > \frac{1}{2}. \tag{3.16}$$

Combining this last result with (3.14) and (3.15) gives the required bound for $\nu > 1/2$. The result for $\nu = 1/2$ is obtained by the continuity of Bessel functions in the argument ν.

Finally we prove (a) and (b) for $\nu = 0$.

$$|J_0(k)H_0(k)| \leq \frac{1}{k}kM_0^2(k) \overset{\text{Lemma 3.9(c)}}{\leq} \frac{1}{k}\lim_{k\to\infty} kM_0^2(k) \overset{[1,\ (9.2.3)]}{\leq} C\frac{1}{k} \leq Ck^{-2/3}.$$

Similarly,

$$\begin{aligned} k|J_0'(k)H_0(k)| = k|J_1(k)H_0(k)| &\leq \sqrt{k}M_1(k)\sqrt{k}M_0(k) \\ &\overset{\text{Lemma 3.9(c)}}{\leq} M_1(1)\lim_{k\to\infty}\sqrt{k}M_0(k) \overset{[1,\ (9.2.3)]}{\leq} C. \qquad \square \end{aligned}$$

COROLLARY 3.11. *Let $R_k : L^2(\mathbb{S}^2) \to L^2(\mathbb{S}^2)$ be the operator defined, as in* (3.8)*, by*

$$R_k = I/2 + K_k - i\alpha V_k.$$

Then R_k is bounded, and there exists a constant $C > 0$ independent of k such that

$$\|R_k\|_{L^2(\mathbb{S}^2)\leftarrow L^2(\mathbb{S}^2)} \leq C(1+\alpha k^{-2/3}).$$

Proof. In view of Lemma 3.8, to prove the statement we need to find bounds on the eigenvalues of the operator R_k. Using the definition of spherical Bessel functions (3.12) and Proposition 3.10, we have that

$$\left|\lambda_{n,k}^{(V)}\right| = \left|2kh_n^{(1)}(k)j_n(k)\right| = \left|\pi H_{n+\frac{1}{2}}(k)J_{n+\frac{1}{2}}(k)\right| \leq Ck^{-2/3},$$

and

$$\begin{aligned} \left|\frac{1}{2}+\lambda_{n,k}^{(K)}\right| = \left|k^2h_n^{(1)}(k)j_n'(k)\right| &= \left|\frac{\pi}{2}kH_{n+\frac{1}{2}}(k)\left(J_{n+\frac{1}{2}}'(k)+\frac{1}{2k}J_{n+\frac{1}{2}}(k)\right)\right| \\ &\leq \left|\frac{\pi}{2}kH_{n+\frac{1}{2}}(k)J_{n+\frac{1}{2}}'(k)\right| + \left|\frac{\pi}{4}H_{n+\frac{1}{2}}(k)J_{n+\frac{1}{2}}(k)\right| \leq C(1+k^{-2/3}). \end{aligned}$$

The result now follows from the identity

$$\|R_k\|_{\mathcal{H}^s(\mathbb{S}^2)\leftarrow\mathcal{H}^s(\mathbb{S}^2)} = \sup_{n\in\mathbb{N}_0}\left|\lambda_{n,k}^{(R)}\right| = \sup_{n\in\mathbb{N}_0}\left|1/2+\lambda_{n,k}^{(K)}-i\alpha\lambda_{n,k}^{(V)}\right|. \qquad \square$$

Note that for $\alpha \le k^{2/3}$, $\|R_k\|_{L^2(\mathbb{S}^2)\leftarrow L^2(\mathbb{S}^2)}$ is bounded by a constant independent of k. Numerical experiments suggest $C_c = \|R_k\|_{L^2(\mathbb{S}^2)\leftarrow L^2(\mathbb{S}^2)} \le 1.76$, for $\alpha = k^{2/3}$.

DEFINITION 3.12. *Let $\alpha := k^{2/3}$ in the definition of R_k; see* (3.8).

REMARK 6. *The choice $\alpha \propto k$ is prevalent in the literature; see* [2, 11, 13, 21]. *In* [2] *and* [21] *the choice was made to minimize the condition number of the matrices arising from the discretization of boundary integral operators in the case of the unit sphere and the unit disk. The same choice maximizes the inf-sup constant and hence optimizes the error estimate given by Céa's lemma; see* [13]. *The error estimate in Corollary* 3.3 *is not affected by the inf-sup constant, and with the choice $\alpha = k^{2/3}$ the constant of quasioptimality C_c is independent of k. Céa's lemma gives a more pessimistic bound, with the quasioptimality constant growing as $k^{1/3}$; see* [11, 13].

It remains now to find the dependence on k of the continuity constant of the operator $\mathcal{T}_k = R_k^{*-1}\widetilde{R}$. From Lemma 3.8 we have that

$$\|\mathcal{T}_k\|_{H^1(\mathbb{S}^2)\leftarrow L^2(\mathbb{S}^2)} = \sup_n \sqrt{1+n^2}|\lambda_{n,k}^{\mathcal{T}}| = \sup_n \sqrt{1+n^2}\left|\frac{\lambda_{n,k}^{(K)}-i\alpha\lambda_{n,k}^{(V)}}{1/2+\overline{\lambda_{n,k}^{(K)}}+i\alpha\overline{\lambda_{n,k}^{(V)}}}\right|.$$

By taking into account the properties of the zeros of Bessel functions (see [1, (9.5)]), it can be seen that the denominator in the above expression is never zero; however, a proof of a useful upper bound for the whole expression is beyond the scope of this paper. Instead, we consider the three asymptotic cases: k fixed and $n\to\infty$, $n\approx k$, and n fixed and $k\to\infty$.

PROPOSITION 3.13.

(a) *For fixed ν and $k\to\infty$ we have, for $\alpha \le k$,*

$$\left|\lambda_{\nu,k}^{\mathcal{T}}\right| = \left|1-\frac{1}{2e^{i\chi}\left(-\frac{2\alpha}{k}\cos\chi + i\sin\chi\right)}+O(k^{-1})\right|,$$

where $\chi = k-\nu\pi/2-\pi/2$.

(b) *For $\nu+1/2 = k$ and $\alpha \le k^{4/3}$ we have*

$$|\lambda_{\nu,k}^{\mathcal{T}}| = 1+\left|i\pi ab(1+\sqrt{3}i)+2\pi a^2(1-\sqrt{3}i)\alpha k^{-2/3}+O(k^{-2/3})\right|^{-1},$$

where a and b are constants from the asymptotic expansions (3.13).

(c) *For fixed k and $\nu\to\infty$ we have*

$$\lambda_{\nu,k}^{\mathcal{T}} = O(\nu^{-1}).$$

Proof. Part (a). We first use the definition of spherical functions to write the eigenvalues in terms of Bessel functions and then make use of asymptotic expansions given in [1, (9.2)]. From (3.12), as in proof of Corollary 3.11, we have for ν fixed and

$k \to \infty$ that

$$\begin{aligned}
\left|\lambda_{\nu,k}^{\mathcal{T}}\right| &= \left|\frac{-\frac{1}{2} + i\frac{\pi}{2}kH_{\nu+\frac{1}{2}}(k)J'_{\nu+\frac{1}{2}}(k) - \frac{\pi}{2}(\frac{i}{2} - 2\alpha)H_{\nu+\frac{1}{2}}(k)J_{\nu+\frac{1}{2}}(k)}{i\frac{\pi}{2}kH_{\nu+\frac{1}{2}}(k)J'_{\nu+\frac{1}{2}}(k) - \frac{\pi}{2}(\frac{i}{2} - 2\alpha)H_{\nu+\frac{1}{2}}(k)J_{\nu+\frac{1}{2}}(k)}\right| \\
&= \left|1 - \frac{1}{i\pi kH_{\nu+\frac{1}{2}}(k)J'_{\nu+\frac{1}{2}}(k) - \pi(\frac{i}{2} - 2\alpha)H_{\nu+\frac{1}{2}}(k)J_{\nu+\frac{1}{2}}(k)}\right| \\
&\overset{[1,\ (9.2)]}{=} \left|1 - \frac{1}{2e^{i\chi}\left(-\frac{2\alpha}{k}\cos\chi + i\sin\chi\right) - \frac{\alpha}{k}O(k^{-1}) + O(k^{-1})}\right|,
\end{aligned}$$

where $\chi = k - (\nu + 1/2)\pi/2 - \pi/4 = k - \nu\pi/2 - \pi/2$. The result now follows from the assumption $\alpha \le k$.

Part (b). Using the asymptotic expansions (3.13), we obtain that

$$\begin{aligned}
\lambda_{\nu,k}^{\mathcal{T}} = 1 + \Big|& i\pi kab\left((1+\sqrt{3}i)k^{-1} + O(k^{-5/3})\right) \\
&- \pi a^2(i/2 - 2\alpha)\left((1-\sqrt{3}i)k^{-2/3} + O(k^{-2})\right)\Big|^{-1} \\
= 1 + \Big|& i\pi ab(1+\sqrt{3}i) + 2\pi a^2(1-\sqrt{3}i)\alpha k^{-2/3} + O(k^{-2/3}) + \alpha O(k^{-2})\Big|^{-1}.
\end{aligned}$$

Part (c). For the proof, we use the asymptotic expansions given in [1, (9.3)]:

$$J_\nu(k)H_\nu(k) \overset{[1,\ (9.3.1)]}{\sim} \frac{1}{2\pi\nu}\left(\frac{ek}{2\nu}\right)^{2\nu} - i\frac{1}{\pi\nu} = O(\nu^{-1}). \tag{3.17}$$

We also make use of Stirling's approximation to the Gamma function [1, (6.1.39)]:

$$\begin{aligned}
J'_\nu(k) &\overset{[1,\ (9.1.10)]}{=} \nu\frac{(\frac{1}{2}k)^\nu}{\Gamma(\nu+1)}\left(\frac{1}{k} - \frac{2+\nu}{2\nu}\left(\frac{1}{2}k\right)\frac{1}{\nu+1} + \cdots\right) \\
&\overset{[1,\ (6.1.39)]}{\sim} \sqrt{\frac{\nu}{2\pi}}\left(\frac{ke}{2\nu}\right)^\nu\left(\frac{1}{k} + O(\nu^{-1})\right).
\end{aligned}$$

Hence,

$$J'_\nu(k)H_\nu(k) \overset{[1,\ (9.3.1)]}{\sim} -i\frac{1}{\pi k} + O(\nu^{-1}). \tag{3.18}$$

Finally,

$$\lambda_{\nu,k}^{\mathcal{T}} \overset{(3.17),\,(3.18)}{\sim} \frac{-1/2 + 1/2 + O(\nu^{-1})}{1/2 + O(\nu^{-1})} = O(\nu^{-1}). \qquad \square$$

Part (c) in the above proposition merely confirms that $\mathcal{T}_k$ is a pseudodifferential operator of order -1. From part (b) we conclude that for $n + 1/2 = k$,

$$\sqrt{1+n^2}|\lambda_{n,k}^{\mathcal{T}}| \sim O(k). \tag{3.19}$$

The denominator in the expression of part (a) is clearly never 0; however, it becomes arbitrarily close to zero for certain large enough values of k and for $\alpha < k$. Nevertheless, note that $|-\frac{2\alpha}{k}\cos\chi + i\sin\chi| \ge 2\alpha/k$, for $k > 2\alpha$. Thus,

$$|\lambda_{\nu,k}^{\mathcal{T}}| = O(k/\alpha) \quad \text{for } k > 2\alpha.$$

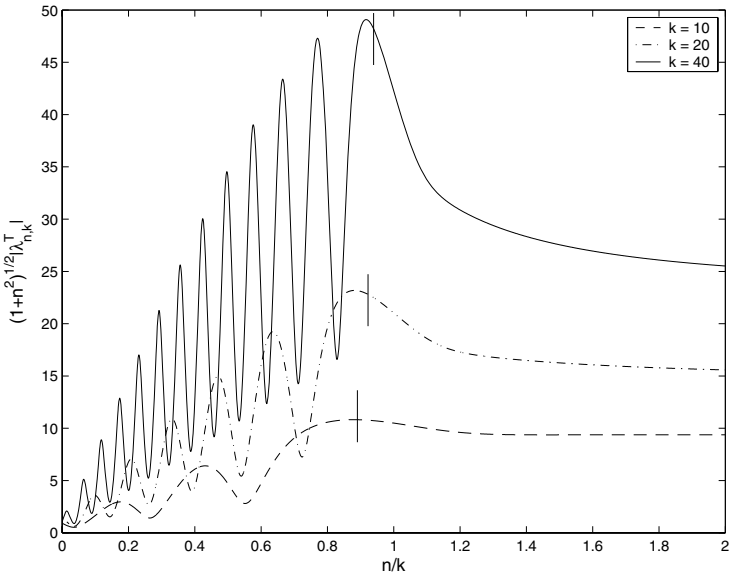


FIG. 3.1. *Plot of* $\sqrt{1+n^2}|\lambda^{\mathcal{T}}_{n,k}|$ *for different values of* n *and* k. *The vertical lines denote the positions at which* $n + 1/2 = k$.

Since $\alpha = k^{2/3}$, the condition $k > 2\alpha$ is equivalent to $k > 8$.

To see how relevant these asymptotic cases are for estimating the continuity constant, in Figure 3.1 we plot $\sqrt{1+n^2}|\lambda^{\mathcal{T}}_{n,k}|$ for different values of k and n. The picture suggests that the maximum occurs for $n + 1/2 \approx k$. Hence, in view of (3.19), we are lead to the following heuristic:

$$\|\mathcal{T}_k\|_{H^1(\mathbb{S}^2)\leftarrow L^2(\mathbb{S}^2)} \leq C_X k \tag{3.20}$$

for some constant $C_X > 0$ independent of k. Numerical experiments suggest that $C_X \leq 1.7$. In [11], it was proved that, in two dimensions with the coupling parameter $\alpha = k$ and large enough k, the eigenvalues of R_k are bounded below by $1/2$. This further supports our claim (3.20).

Now we are in a position to give estimates of the dependence on k of the stability and the accuracy of the boundary element method.

3.2.1. Piecewise-constant Galerkin boundary element method.

PROPOSITION 3.14. *Let* $\Gamma = \mathbb{S}^2$, $S = \mathcal{S}^{0,-1}_{\mathcal{G},h}$, $\varphi \in L^2(\Gamma)$ *be the solution of* (3.9), *and let* (3.20) *hold. There exists a constant* c *independent of* k *such that if* $hk < c$, *the discrete problem* (3.10) *has a unique solution* $\varphi_S \in S$. *If, further,* $\varphi \in \mathcal{O}_{\rho,k,1}$, *then there exists a constant* C *independent of* k *such that*

$$\|\varphi - \varphi_S\|_0 \leq Chk\|\varphi\|_0.$$

Therefore, the boundary element method does not suffer from the pollution effect, and a condition $hk \lesssim 1$ is sufficient to guarantee stability and a quasi-optimal error estimate.

REMARK 7. *Let us consider the two dimensional case,* $\Gamma = \{x \in \mathbb{R}^2 : \|x\| = 1\}$. *The Sobolev space* $H^s(\Gamma)$ *can be identified with the space* $H^s([0, 2\pi])$ *of* 2π *periodic*

distributions; see [2, 20]. *Periodic functions,* $e^{\pm in\theta}$, $n \in \mathbb{N}_0$, *are then the eigenfunctions of the operators* V_k *and* K_k *with eigenvalues given by*

$$\lambda_{n,k}^{(V)} = \frac{i\pi}{2} J_n(k) H_n(k), \qquad \lambda_{n,k}^{(K)} = -\frac{1}{2} + \frac{i\pi}{2} k J_n'(k) H_n(k).$$

Comparing these with the case of the sphere, it is clear that the analogous analysis of this section holds for the two dimensional case as well. In particular, the statement of Proposition 3.14 *also holds for the case of the unit ball in two dimensions.*

3.2.2. The *h*-*p* version of the Galerkin method. Just as in the finite element method [17, 18], the use of higher order polynomials improves the stability condition of the boundary element method. Let $S = \mathcal{S}_{\mathcal{G},h}^{p,1}$ be the usual boundary element space of continuous piecewise polynomial functions of order p. Using the approximation properties of such spaces proved in [15, 16, 17, 18], we proceed as in the case of piecewise-constant basis functions. Assuming that $\widetilde{H} = \mathcal{O}_{\rho,k,l}$, where $1 \le l \le p$, we obtain the estimate

$$\begin{aligned}
\eta(S) &= \sup_{\psi \in \widetilde{H}\setminus\{0\}} \inf_{v \in S} \frac{\|\mathcal{T}_k\psi - v\|_0}{\|\psi\|_0} \overset{[15,\,17]}{\le} C_A(l) \sup_{\psi \in \widetilde{H}\setminus\{0\}} \frac{\|\mathcal{T}_k\psi\|_{l+1}}{\|\psi\|_0} \left(\frac{h}{2p}\right)^{l+1} \\
&\le C_A(l) C_X k \sup_{\psi \in \widetilde{H}\setminus\{0\}} \frac{\|\psi\|_l}{\|\psi\|_0} \left(\frac{h}{2p}\right)^{l+1} \\
&\le \rho C_A(l) C_X \left(\frac{kh}{2p}\right)^{l+1},
\end{aligned}$$

where $C(l)$ is a constant depending only on l. Therefore, the condition for stability and the quasi-optimal error estimate reduces to $hk \lesssim 2p$. Thus, higher order elements allow for a coarser mesh and the following error estimate:

$$\|\varphi - \varphi_s\|_0 \le C \left(\frac{kh}{2p}\right)^{l+1} \|\varphi\|_0.$$

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover Publications, New York, 1992.

[2] S. Amini and N. D. Maines, *Preconditioned Krylov subspace methods for boundary element solution of the Helmholtz equation*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 875–898.

[3] I. M. Babuška and S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM Rev., 42 (2000), pp. 451–484.

[4] L. Banjai and W. Hackbusch, *$\mathcal{H}$- and $\mathcal{H}^2$-Matrices for Low and High Frequency Helmholtz Equation*, Technical Report 17/2005, Max Planck Institute for Mathematics in the Sciences, Leipzig, Germany, 2005.

[5] H. Brakhage and P. Werner, *Über das Dirichletsche Außenraumproblem für die Helmholtzsche Schwingungsgleichung*, Arch. Math., 16 (1965), pp. 325–329.

[6] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts in Appl. Math. 15, Springer-Verlag, New York, 2002.

[7] O. P. Bruno, C. A. Geuzaine, J. A. Monro, Jr., and F. Reitich, *Prescribed error tolerances within fixed computational times for scattering problems of arbitrarily high frequency: The convex case*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 629–645.

[8] A. Buffa and S. Sauter, *On the acoustic single layer potential: Stabilization and Fourier Analysis*, SIAM J. Sci. Comput., to appear.

[9] G. Chen and J. Zhou, *Boundary Element Methods*, Comput. Math. Appl., Academic Press, London, 1992.
[10] E. Darrigrand, *Coupling of fast multipole method and microlocal discretization for the 3-D Helmholtz equation*, J. Comput. Phys., 181 (2002), pp. 126–154.
[11] V. Domínguez, I. G. Graham, and V. P. Smyshlyaev, *A Hybrid Numerical-Asymptotic Boundary Integral Method for High-Frequency Acoustic Scattering*, Bath Institute for Complex Systems Preprint 1/06, University of Bath, Bath, UK, 2006.
[12] B. Engquist and O. Runborg, *Computational high frequency wave propagation*, Acta Numer., 12 (2003), pp. 181–266.
[13] K. Giebermann, *Schnelle Summationsverfahren zur numerischen Lösung von Integralgleichungen für Streuprobleme im* $\mathbb{R}^3$, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, Germany, 1997.
[14] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, Academic Press, San Diego, CA, 2000.
[15] P. Q. Guo and I. Babuška, *The hp-version of the finite element method* I. *The basic approximation results*, Comput. Mech., 1 (1986), pp. 21–41.
[16] P. Q. Guo and I. Babuška, *The hp-version of the finite element method* II. *General results and applications*, Comput. Mech., 1 (1986), pp. 203–226.
[17] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Appl. Math. Sci. 132, Springer-Verlag, New York, 1998.
[18] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number. Part* II. *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.
[19] R. Kress, *Minimizing the condition number of boundary integral operators in acoustic and electromagnetic scattering*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 323–341.
[20] R. Kress, *Linear Integral Equations*, Appl. Math. Sci. 82, Springer-Verlag, Berlin, 1989.
[21] R. Kress and W. T. Spassov, *On the condition number of boundary integral operators for the exterior Dirichlet problem for the Helmholtz equation*, Numer. Math., 42 (1983), pp. 77–95.
[22] R. Leis, *Zur Dirichletschen Randwertaufgabe des Aussenraumes der Schwingungsgleichung*, Math. Z., 90 (1965), pp. 205–211.
[23] J. M. Melenk, *On Generalised Finite Element Methods*, Ph.D. thesis, University of Maryland at College Park, College Park, MD, 1995.
[24] S. G. Mikhlin and S. Prössdorf, *Singular Integral Operators*, Springer-Verlag, Berlin, 1986.
[25] J.-C. Nédélec, *Acoustic and Electromagnetic Equations. Integral Representations for Harmonic Problems*, Appl. Math. Sci. 144, Springer-Verlag, New York, 2001.
[26] O. I. Panič, *On the solvability of exterior boundary-value problems for the wave equation and for a system of Maxwell's equations*, Uspehi Mat. Nauk, 20 (1965), pp. 221–226.
[27] E. Perrey-Debain, O. Laghrouche, P. Bettess, and J. Trevelyan, *Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 561–577.
[28] V. Rokhlin, *Rapid solution of integral equations of scattering theory in two dimensions*, J. Comput. Phys., 86 (1990), pp. 414–439.
[29] S. Sauter, *A refined finite element convergence theory for highly indefinite Helmholtz problems*, Computing, 78 (2006), pp. 101–115.
[30] S. Sauter and C. Schwab, *Randelementmethoden*, Teubner, Leipzig, 2004.
[31] A. H. Schatz, *An observation concerning Ritz-Galerkin methods with indefinite bilinear forms*, Math. Comp., 28 (1974), pp. 959–962.
[32] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, Cambridge University Press, Cambridge, England, 1944.
[33] K. Yosida, *Functional Analysis*, Grundlehren Math. Wiss. 123, Academic Press, New York, 1965.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 54–67

TIME SPLITTING ERROR IN DSMC SCHEMES FOR THE SPATIALLY HOMOGENEOUS INELASTIC BOLTZMANN EQUATION*

SERGEJ RJASANOW† AND WOLFGANG WAGNER‡

Abstract. The paper is concerned with the numerical treatment of the uniformly heated inelastic Boltzmann equation by the direct simulation Monte Carlo (DSMC) method. This technique is presently the most widely used numerical method in kinetic theory. We consider three modifications of the DSMC method and study them with respect to their efficiency and convergence properties. Convergence is investigated with respect to both the number of particles and the time step. The main issue of interest is the time step discretization error due to various splitting strategies. A scheme based on the Strang-splitting strategy is shown to be of second order with respect to time step, while there is only first order for the commonly used Euler-splitting scheme. On the other hand, a no-splitting scheme based on appropriate Markov jump processes does not produce any time step error. It is established in numerical examples that the no-splitting scheme is about two orders of magnitude more efficient than the Euler-splitting scheme. The Strang-splitting scheme reaches almost the same level of efficiency as that of the no-splitting scheme, since the deterministic time step error vanishes sufficiently fast.

Key words. granular matter, Boltzmann equation, stochastic numerics

AMS subject classifications. 82C40, 82C80, 65R20

DOI. 10.1137/050643842

1. Introduction.

A basic tool for modeling low-density flows of granular materials is the inelastic Boltzmann equation. We refer to the conference proceedings [17], [16] and to the monograph [5] for details concerning applications and an appropriate physical justification. In this paper we consider the spatially homogeneous uniformly heated inelastic Boltzmann equation

$$\partial_t f - \beta \, \Delta_v \, f = Q_\alpha(f, f) \tag{1.1}$$

with initial condition

$$f(0, v) = f_0(v). \tag{1.2}$$

Equation (1.1) describes the time evolution of a function $f(t, v)$ representing the average number of particles at time t having a velocity close to v. The symbol Δ denotes the Laplace operator and the parameter $\beta > 0$ determines the strength of the random forcing. The collision integral is most conveniently written in the weak form

$$\begin{aligned} &\int_{\mathbb{R}^3} Q_\alpha(f, f)(v) \, \varphi(v) \, dv \\ &\quad = \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \int_{S^2} B(v, w, e) \, [\varphi(v'_\alpha) + \varphi(w'_\alpha) - \varphi(v) - \varphi(w)] \, f(v) \, f(w) \, de \, dw \, dv, \end{aligned} \tag{1.3}$$

*Received by the editors October 31, 2005; accepted for publication (in revised form) July 13, 2006; published electronically January 8, 2007.
http://www.siam.org/journals/sinum/45-1/64384.html
†Fachrichtung 6.1 – Mathematik, Universität des Saarlandes, Postfach 15 11 50, 66041 Saarbrücken, Germany (rjasanow@num.uni-sb.de).
‡Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstraße 39, D-10117 Berlin, Germany (wagner@wias-berlin.de).

where φ is a test function and S^2 denotes the unit sphere in the Euclidean space $\mathbb{R}^3$. The function B is called the collision kernel. The postcollisional velocities are defined by

$$
\begin{aligned}
v'_\alpha = v'_\alpha(v,w,e) &= \frac{1}{2}(v+w) + \frac{1-\alpha}{4}(v-w) + \frac{1+\alpha}{4}|v-w|\,e,\\
w'_\alpha = w'_\alpha(v,w,e) &= \frac{1}{2}(v+w) - \frac{1-\alpha}{4}(v-w) - \frac{1+\alpha}{4}|v-w|\,e.
\end{aligned}
\tag{1.4}
$$

The parameter $0 < \alpha \le 1$ is called the restitution coefficient. For $\alpha = 1$ the collisions are elastic and $Q_1(f,f)$ coincides with the classical Boltzmann collision operator. A discussion of the relevance of (1.1), as well as more references, can be found in [10].

In this paper we address the issue of the numerical treatment of (1.1) by the direct simulation Monte Carlo (DSMC) method. This technique is presently the most widely used numerical method in kinetic theory (cf. [2], [6]). It is based on a system of particles performing a random evolution that imitates the behavior of the underlying physical model. As to inelastic collisions, the homogeneous cooling state of a low-density granular flow was studied by the DSMC method in [4]. A DSMC method for uniformly heated granular fluids, described by (1.1), was introduced in [13]. Related studies were performed in [1], [21]. We refer to [8] for an account of deterministic numerical methods for the elastic Boltzmann equation and to [9] concerning a deterministic numerical approach to (1.1).

The purpose of this paper is to study three modifications of the DSMC method for the uniformly heated inelastic Boltzmann equation with respect to their efficiency and convergence properties. The main issue of interest is the time step discretization error due to various splitting strategies. The first method, from [13], implements a straightforward Euler-type splitting in analogy with the classical Bird scheme for the spatially inhomogeneous elastic Boltzmann equation. The second method follows the Strang-splitting strategy (cf. [20]). The third method, previously used in [11], avoids any time step discretization error, since no splitting is used. Convergence is studied with respect to both the number of particles and the time step. All methods are of first order with respect to the inverse number of particles. The Strang-splitting scheme is of second order with respect to the time step, while the Euler-splitting scheme is of first order. In the numerical examples, the no-splitting scheme is about two orders of magnitude more efficient than the Euler-splitting scheme. It is observed that the Strang-splitting scheme reaches almost the same level of efficiency compared to the no-splitting scheme, since the deterministic time step error vanishes sufficiently fast.

The paper is organized as follows. In section 2 we describe a Markovian particle system approximating (1.1). In section 3 we define the three DSMC algorithms mentioned above. In section 4 we introduce a test example and present the results of numerical experiments. Here we study efficiency and convergence properties of the algorithms both in the transient and in the steady state cases. Section 5 contains some concluding remarks.

2. The direct simulation process. Here we describe the time evolution of a Markovian particle system

$$
\Big(v_1(t),\dots,v_n(t)\Big), \qquad t \ge 0, \tag{2.1}
$$

where each particle is characterized by its velocity. The process (2.1) corresponds to (1.1) in the sense that the family of empirical measures

$$\nu^{(n)}(t, dv) = \frac{1}{n} \sum_{j=1}^{n} \delta_{v_j(t)}(dv) \tag{2.2}$$

converges (as $n \to \infty$) to the measures $f(t, v)\, dv$. We refer to [14], [12] concerning rigorous convergence results for a wide class of Boltzmann-type models (see also [19, section 2.3.3]).

Roughly speaking, the system interacts through binary inelastic collisions. In addition, the particles continuously gain kinetic energy due to Gaussian white noise forcing. More precisely, we assume

$$\int_{S^2} B(v, w, e)\, de \le B_{max} \qquad \forall\, v, w \in \mathbb{R}^3. \tag{2.3}$$

Then the evolution of system (2.1) is determined via the following steps.

Initial measure. System (2.1) at $t = 0$ is chosen in such a way that the empirical measure $\nu^{(n)}(0, dv)$ (cf. (2.2)) approximates the initial measure $f_0(v)\, dv$ (cf. (1.2)).

Time counter. Given system (2.1) at time t, the next interaction (collision) takes place at a random time $t + \tau$, where

$$\text{Prob}\,\{\tau \ge s\} = \exp\left(-\frac{n-1}{2} B_{max}\, s\right), \qquad s \ge 0. \tag{2.4}$$

Brownian motion. The particle velocities perform individual Brownian motions between the collisions. After some time τ without collisions, the particle velocities are given by

$$v_j(t + \tau) = v_j(t) + \sqrt{2\,\beta\,\tau}\;\xi_j, \qquad j = 1, \ldots, n, \tag{2.5}$$

where $\xi_j \in \mathbb{R}^3$ are independent standard Gaussian random variables.

Collision partners. The indices i and j of the collision partners are chosen uniformly on the set $\{1 \le i \ne j \le n\}$.

Fictitious collision. Given i and j, the collision is fictitious (the system does not change) with probability

$$1 - \frac{\int_{S^2} B(v_i, v_j, e)\, de}{B_{max}}. \tag{2.6}$$

Collision. With the remaining probability, a direction vector e is generated according to the density

$$\frac{B(v_i, v_j, e)}{\int_{S^2} B(v_i, v_j, e)\, de}, \qquad e \in S^2, \tag{2.7}$$

and the postcollisional velocities

$$v'_\alpha(v_i, v_j, e), \qquad w'_\alpha(v_i, v_j, e) \tag{2.8}$$

are computed according to the collision transformation (1.4).

3. DSMC algorithms. Here we describe three algorithms based on the Markov process introduced in the previous section. They differ in the way time splitting is carried out.

The algorithms perform the time evolution of a particle system $(v_1, \dots, v_n)$. At some observation points

$$s_m, \qquad m = 0, 1 \dots, M, \tag{3.1}$$

functionals of the system

$$\xi^{(n)} = \frac{1}{n} \sum_{j=1}^{n} \varphi(v_j) \tag{3.2}$$

are computed, where φ is an appropriate test function. The random variable (3.2) approximates the functional

$$\int_{\mathbb{R}^3} \varphi(v)\, f(s_m, v)\, dv \tag{3.3}$$

of the solution of (1.1).

In order to reduce the random fluctuations of the estimator (3.2), a number N of independent ensembles of particles is generated. The corresponding values of the random variable are denoted by $\xi_1^{(n)}, \dots, \xi_N^{(n)}$. The empirical mean value of the random variable (3.2)

$$\eta^{(n,N)} = \frac{1}{N} \sum_{j=1}^{N} \xi_j^{(n)} \tag{3.4}$$

is used as an approximation to the functional (3.3). The independent ensembles of particles are also used to estimate the random fluctuations by means of confidence intervals. For details we refer, e.g., to [19, section 3.1.4].

3.1. Euler-splitting scheme. First we describe the DSMC method introduced in [13]. It implements the idea of standard (elastic) DSMC, where the free flow and collision simulation are separated (cf. [2]). The simulation of random "kicking" and collisions is split over a time step Δt. The state of the particle system is calculated at the discrete time points

$$t_k = k\, \Delta t, \qquad k = 0, 1, \dots, \tag{3.5}$$

until all observation points (3.1) (assumed to be multiples of the time step) are reached.

ALGORITHM 3.1.

1. Initialization
   - 1.1 set system time $t = t_0$
   - 1.2 generate v_j, $j = 1, \dots, n$, according to $f_0(v)$
2. Simulation (for $k = 1, 2, \dots$)
   - 2.1 Collision step of length Δt
     - 2.1.1 compute τ according to (2.4)
     - 2.1.2 update the system time $t := t + \tau$
     - 2.1.3 if $t \geq t_k$ then go to Step 2.2

2.1.4 generate the indices $i \neq j$
2.1.5 go to Step 2.1.1 with probability (2.6)
2.1.6 generate e according to (2.7)
2.1.7 replace v_i and v_j according to (2.8)
2.2 Kicking step of length Δt
2.2.1 update all velocities (cf. (2.5))

$$v_i := v_i + \sqrt{2\beta\,\Delta t}\,\xi_i, \qquad i = 1, \dots, n$$

2.2.2 set system time $t = t_k$
3. Compute functional (3.2) at all s_m

3.2. Strang-splitting scheme. Next we describe a modification of the algorithm from the previous section. We apply the idea of the Strang splitting. This has been introduced in the context of the elastic Boltzmann equation in [15]. Its application to equations with rather general operators was studied in [3].

ALGORITHM 3.2.
1. Initialization
2. Simulation (for $k = 1, 2, \dots$)
2.1 Collision step of length $\Delta t/2$
2.2 Kicking step of length Δt
2.3 Collision step of length $\Delta t/2$
3. Computation of functionals

3.3. No-splitting scheme. Finally we recall a DSMC algorithm that was introduced in [11]. The symbols σ_j denote the last time, at which the particle j was kicked.

ALGORITHM 3.3.
1. Initialization
1.1 set system time $t = 0$
1.2 generate v_j, $j = 1, \dots, n$, according to $f_0(v)$
1.3 set $\sigma_j = 0$, $j = 1, \dots, n$
2. Simulation (for $m = 0, 1, \dots, M$)
2.1 compute τ according to (2.4)
2.2 update the system time $t := t + \tau$
2.3 if $t \geq s_m$ then go to Step 3
2.4 generate the indices $i \neq j$
2.5 update the velocities v_i and v_j (cf. (2.5))

$$v_i := v_i + \sqrt{2\beta\,(t - \sigma_i)}\,\xi_i, \quad v_j := v_j + \sqrt{2\beta\,(t - \sigma_j)}\,\xi_j$$

2.6 update the times of last kicking $\sigma_i = \sigma_j := t$
2.7 go to Step 2.1 with probability (2.6)
2.8 generate e according to (2.7)
2.9 replace v_i and v_j according to (2.8) and go to Step 2.1
3. Calculation of functionals
3.1 update the velocities of all particles (cf. (2.5))

$$v_i := v_i + \sqrt{2\beta\,(s_m - \sigma_i)}\,\xi_i, \qquad i = 1, \dots, n$$

3.2 compute (3.2)
3.3 set system time $t = s_m$ and go to Step 2.1

3.4. Comments. In Algorithms 3.1 and 3.2, the kicking step is computed accurately, i.e., without any further time discretization. Particles are just moved according to Brownian motion. The collision step contains the random interaction times distributed according to (2.4). Its accuracy depends on the number of particles. In Algorithm 3.3 particles perform Brownian motion between collisions so that any splitting errors are avoided.

One might use deterministic interaction times obtained as the expectation of the distribution (2.4). This would introduce another error, which is small for large n. We refer to [19, section 3.5.2] concerning a discussion of various time counting procedures.

Unbounded collision kernels. The *variable hard sphere model*

$$B(v,w,e) = C_\lambda\,|v-w|^\lambda, \qquad 0 \le \lambda \le 1, \tag{3.6}$$

is widely used in applications (cf. [2, Chapter 2]). Particular cases are the models of *hard spheres* ($\lambda = 1$) and of *pseudo-Maxwell molecules* ($\lambda = 0$). The kernel (3.6) does not satisfy condition (2.3), unless $\lambda = 0$. In order to fit into the framework of section 2, one has to truncate the kernel using some maximal relative velocity U_{max}. The truncated kernel

$$\hat{B}(v,w,e) = \begin{cases} B(v,w,e) & \text{if} \quad |v-w| \le U_{max}, \\ C_\lambda\, U_{max}^\lambda & \text{otherwise} \end{cases} \tag{3.7}$$

satisfies (2.3) with $B_{max} = 4\pi\, C_\lambda\, U_{max}^\lambda$. Correspondingly, the parameter of the waiting time distribution (2.4) takes the form

$$2\pi\,(n-1)\,C_\lambda\,U_{max}^\lambda. \tag{3.8}$$

The probability of a fictitious collision (2.6) is

$$1 - \left(\frac{|v_i - v_j|}{U_{max}}\right)^\lambda.$$

The density (2.7) is constant so that the vector e is distributed uniformly on the unit sphere.

Adapting majorants. There are two aspects related to the choice of the truncation parameter U_{max}. If it is small, then the solution of (1.1) for the kernel (3.7) will significantly differ from the solution for the original kernel B. If, on the other hand, the parameter U_{max} is big, then the time steps between collisions are small (inverse of parameter (3.8)) and the algorithms are time consuming. Therefore, the parameter U_{max} is usually derived from the particle system used in the simulation.

In the classical (elastic) DSMC algorithm (cf. [2]) the starting value of U_{max} is based on the temperature of the initial particle system. Then this value is adapted during the process of calculation each time the relative velocity of a pair of particles exceeds the stored quantity. This procedure works well in steady state calculations. The error related to this procedure in transient calculations was studied in [18].

A problem with finding the maximum relative velocity in a particle system is related to the fact that the effort is quadratic in the number of particles. However, this can easily be reduced to a linear effort by using the estimate

$$\max_{i,j} |v_i - v_j| \le \max_{i,j} \Big(|v_i - V| + |V - v_j| \Big) = 2 \max_i |v_i - V|,$$

where V is any fixed vector. A particular choice is the numerical bulk velocity

$$V = \frac{1}{n} \sum_{j=1}^{n} v_j.$$

Thus, one may start with $U_{max} = 2 \max_i |v_i - V|$ and update the majorant after each collision

$$U_{max} := \max \Big\{ U_{max}, |v_i - V|, |v_j - V| \Big\}. \tag{3.9}$$

This procedure is used in Algorithms 3.1 and 3.2. The situation in Algorithm 3.3 is slightly more difficult, since particle velocities change continuously as a result of the kicking process. Here we implemented the above procedure of adapting the majorant, but in addition the quantity U_{max} was updated according to (3.9) after each Step 2.5. The error caused by this truncation does not seem to be significant, as shown by the very precise tail calculations in [11].

4. Numerical examples. Here we test the algorithms introduced in the previous section with respect to their convergence properties and efficiency.

We consider the case of a constant collision kernel, namely, (3.6), with

$$\lambda = 0, \qquad C_0 = \frac{1}{\pi}. \tag{4.1}$$

Note that other values of the constant C_0 can be handled by an appropriate time scaling, since the function $f(c\,t, v)$ solves (1.1) with diffusion coefficient $c\,\beta$ and collision kernel $c\,B$, where $c > 0$ is some constant. Furthermore, we assume

$$\int_{\mathbb{R}^3} f_0(v)\, dv = 1, \qquad \int_{\mathbb{R}^3} v\, f_0(v)\, dv = 0.$$

Note the conservation properties

$$\int_{\mathbb{R}^3} f(t,v)\, dv = \int_{\mathbb{R}^3} f_0(v)\, dv, \qquad \int_{\mathbb{R}^3} v\, f(t,v)\, dv = \int_{\mathbb{R}^3} v\, f_0(v)\, dv,$$

which can be derived easily from the weak form of the equation (cf. (1.3)).

In this case the relaxation of the temperature

$$T(t) = \frac{1}{3} \int_{\mathbb{R}^3} |v|^2 f(t,v)\, dv \tag{4.2}$$

is known analytically. Assuming $0 < \alpha < 1$, one obtains (cf. [11])

$$\begin{aligned} T(t) = T_{\alpha,\beta}(t) &= T_0\, e^{-(1-\alpha^2)\,t} + T_{\alpha,\beta}(\infty) \left(1 - e^{-(1-\alpha^2)\,t} \right) \\ &= T_{\alpha,\beta}(\infty) + [T_0 - T_{\alpha,\beta}(\infty)]\, e^{-(1-\alpha^2)\,t}, \end{aligned} \tag{4.3}$$

where

$$T_0 = \frac{1}{3} \int_{\mathbb{R}^3} |v|^2 f_0(v)\, dv$$

and

$$T_{\alpha,\beta}(\infty) = \frac{2\,\beta}{1-\alpha^2}. \tag{4.4}$$

Note that

$$\lim_{\alpha\to 1} T_{\alpha,\beta}(t) = T_0 + 2\beta\, t.$$

According to (4.1), the collision kernel is bounded so that the only sources of error are the number of particles n, the time step Δt (in Algorithms 3.1 and 3.2), and the number of independent samples N (cf. (3.4)). First order of convergence with respect to n has been established under rather general assumptions (cf. [14], [12] concerning the transient case and [7] concerning the steady state case). Convergence with respect to Δt for Euler splitting (first order) and Strang splitting (second order) was studied in [3] in the context of rather general operator equations. We refer to [19, section 3.5.5] for more details.

4.1. Approximation on a finite time interval (transient case). Here we use the Maxwell distribution

$$f_0(v) = \frac{1}{(2\,\pi)^{3/2}}\, e^{-\frac{|v|^2}{2}} \tag{4.5}$$

as the initial condition. For the parameters

$$\alpha = \frac{1}{2}, \qquad \beta = 1, \tag{4.6}$$

one obtains from (4.3), (4.4)

$$T(t) = e^{-\frac{3}{4}\,t} + \frac{8}{3}\left(1 - e^{-\frac{3}{4}\,t}\right). \tag{4.7}$$

We approximate the evolution of the temperature (4.2) on the time interval $[0, 8.0]$, using (3.2) with $\varphi(v) = \frac{1}{3}\,|v|^2$ (cf. (3.3)). The time step in the splitting schemes (cf. (3.5)) is chosen in the form

$$\Delta t = \frac{8}{K}, \qquad K \geq 4.$$

Unless indicated otherwise, the results are averaged over $N = 10\,000$ independent ensembles (cf. (3.4)).

Particle number convergence. Figure 4.1 illustrates the approximation of the analytical solution (4.7) (dashed line) by confidence bands (solid lines) computed using the no-splitting scheme with two different values of n. A "zoom" on the time interval $[5.0, 8.0]$ of the no-splitting scheme for two higher values of n is shown in Figure 4.2. The analytical solution (4.7) is mostly covered by the confidence interval for $n = 4\,096$ and completely covered for $n = 65\,536$. More detailed results are given in Table 4.1. The errors are computed as

$$E_{end} = \left|\frac{T(t_K) - T_K}{T(t_K)}\right|, \qquad E_{max} = \max_{0\leq k\leq K}\left|\frac{T(t_k) - T_k}{T(t_k)}\right|,$$

where $T(t_k)$ are the exact values (4.7) of the temperature at time point t_k and T_k is the computed temperature. The convergence factors (quotients of subsequent values) are denoted by CF. The results in Table 4.1 clearly indicate the expected convergence order $O(n^{-1})$ of the error. Note that the width Conf_{end} of the confidence interval at t_K is proportional to $\frac{1}{\sqrt{n\,N}}$, since the variance of the estimator (3.2) has the order $\frac{1}{n}$.

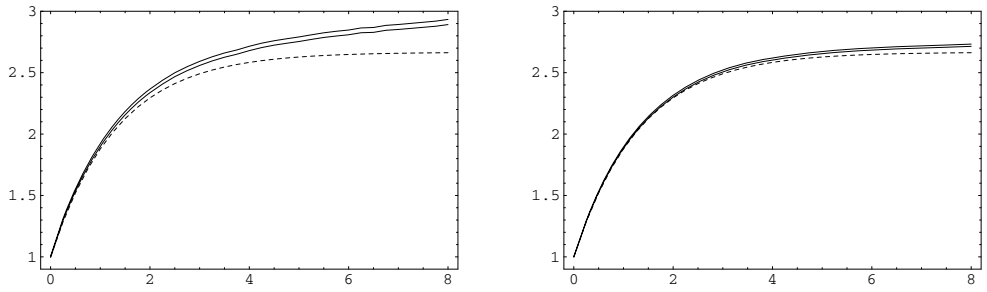


FIG. 4.1. *No-splitting scheme for $n = 64$ (left) and $n = 256$ (right).*

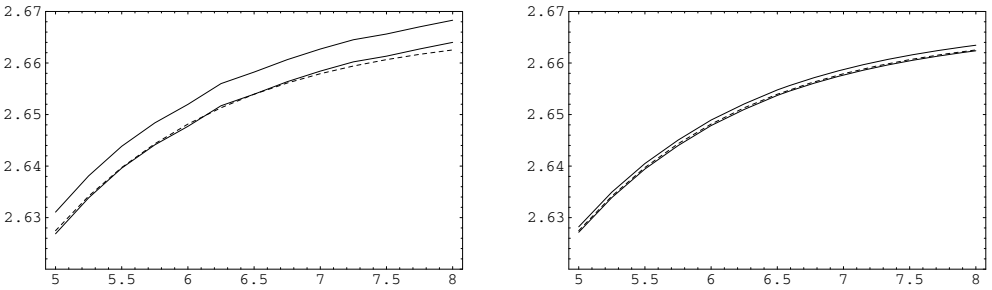


FIG. 4.2. *No-splitting scheme for $n = 4\,096$ (left) and $n = 65\,536$ (right).*

TABLE 4.1
No-splitting scheme for different n.

n	E_{end}	CF	Conf_{end}	CF	E_{max}	CF
16	0.359 E-00	-	0.119 E-00	-	0.359 E-00	-
64	0.939 E-01	3.82	0.420 E-01	2.83	0.939 E-01	3.82
256	0.229 E-01	4.10	0.183 E-01	2.30	0.229 E-01	4.10
1 024	0.584 E-02	3.92	0.865 E-02	2.11	0.584 E-02	3.92
4 096	0.136 E-02	4.29	0.430 E-02	2.01	0.136 E-02	4.29
16 382	0.293 E-03	4.64	0.215 E-02	2.00	0.332 E-03	4.10
65 536	0.141 E-03	2.08	0.108 E-02	1.99	0.141 E-03	2.35

TABLE 4.2
Euler- and Strang-splitting schemes for $n = 4\,096$.

K	E_{max}^{Euler}	CF	E_{max}^{Strang}	CF
4	0.931 E-00	-	0.863 E-01	-
8	0.421 E-00	2.21	0.195 E-01	4.42
16	0.200 E-00	2.11	0.501 E-02	3.89
32	0.977 E-01	2.05	0.997 E-03	5.03

Time step convergence. Similar convergence behavior with respect to the particle number is observed for the Euler- and Strang-splitting schemes, except that the n-limits contain an error depending on Δt. The corresponding numerical results are collected in Table 4.2. The linear convergence of the Euler-splitting scheme as well as the quadratic convergence of the Strang-splitting scheme are clearly indicated.

For $n = 4\,096$ and $K = 32$ the error of the Strang-splitting scheme is comparable to that of the no-splitting scheme, while the error of the Euler-splitting scheme is about 70 times larger. We note that for these parameters the numerical work of all

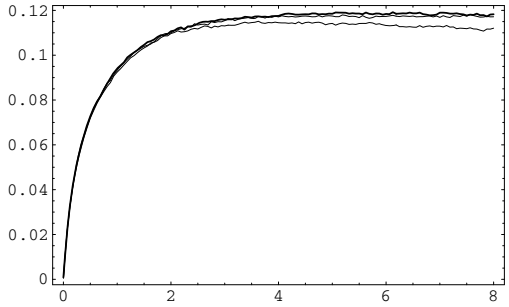


FIG. 4.3. *No-splitting scheme for* $n = 256$, 1 024, *and* 4 096 *(from below).*

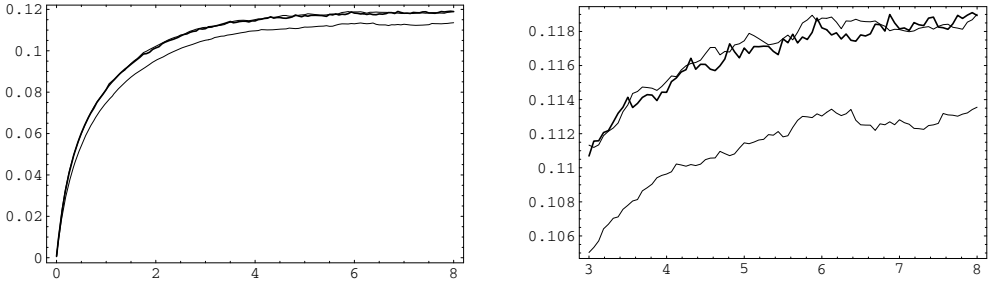


FIG. 4.4. *No-splitting scheme (thick line), Strang-splitting scheme (upper thin line), and Euler-splitting scheme (lower thin line) for* $n = 4\,096$.

schemes is roughly the same. A more detailed study of the efficiency will be made in section 4.2.

Deviation from the Maxwellian state. An interesting feature of the inelastic Boltzmann equation (1.1) is a non-Maxwellian steady state. A specific "criterion" for detecting deviations from the Maxwellian state has the form (cf. [19, section 1.8])

$$Crit(t) = \frac{1}{T(t)} \left(\frac{1}{2} \|P(t) - p(t)\, I\|_F^2 + \frac{2}{5\,T(t)} \|q(t)\|^2 + \frac{1}{120\,T(t)^2} \gamma(t)^2 \right)^{1/2}, \tag{4.8}$$

where $P(t)$ is the pressure tensor, $p(t)$ is the scalar pressure, $q(t)$ is the heat flux vector, I denotes the identity matrix, $\|A\|_F$ denotes the Frobenius norm of a matrix A, and

$$\gamma(t) = \int_{\mathcal{R}^3} \|v\|^4 f(t,v)\, dv - 15\, T(t)^2$$

is a fourth moment of the distribution function.

In Figure 4.3 we show the criterion (4.8) obtained by the no-splitting scheme with different values of n. The curves were calculated using, respectively, $N = 65\,536$, 16 192, and 4 096 repetitions. The functional starts from zero, according to (4.5), and tends to a strictly positive stationary value as $t \to \infty$. So it allows one to quantify the deviation from the Maxwellian state.

For comparison the corresponding curves for the splitting schemes with $K = 32$, $n = 4\,096$, and $N = 4\,096$ are provided in Figure 4.4. Similar to what was observed

TABLE 4.3
Euler-splitting scheme for $n = 65\,536$.

K	T_∞	Error	CF	Conf	CPU	CF
32	2.9226	0.2559	-	0.0046	1.09	-
256	2.6975	0.0308	8.3	0.0041	6.57	6.03
512	2.6817	0.0150	2.1	0.0040	12.8	1.95
2 048	2.6719	0.0052	2.9	0.0039	51.1	3.99
4 094	2.6710	0.0043	1.2	0.0053	100.4	1.96

TABLE 4.4
Strang-splitting scheme for $n = 65\,536$.

K	T_∞	Error	CF	Conf	CPU
8	2.6055	0.0612	-	0.0040	0.61
16	2.6520	0.0147	4.2	0.0047	0.86
32	2.6620	0.0047	3.1	0.0038	1.07
64	2.6655	0.0012	3.9	0.0045	1.87

TABLE 4.5
No-splitting scheme for $n = 65\,536$.

T_∞	Error	Conf	CPU
2.6684	0.0017	0.0040	1.0

for the temperature, the Strang-splitting scheme gives basically the same accuracy as the no-splitting scheme, while the error of the Euler-splitting scheme is significantly larger.

4.2. Approximation of the stationary value (steady state case). Here we consider the same example as in section 4.1 (cf. (4.7)). Figures 4.1 and 4.2 show that the stationary value $T(\infty) \sim 2.6667$ has almost been reached. So we start averaging at $s_0 = 8$ (cf. (3.1)). The quantity of interest is measured after each $\Delta t_{\max} = 2$ so that

$$s_m = s_0 + m\,\Delta t_{\max}, \qquad m = 1, \ldots, M.$$

This time step (corresponding to $K = 4$ in the context of the previous subsection) is big enough to assure almost independent observations. Confidence intervals are constructed over $M = 64$ observation points, so that $s_M = 136$.

Time step error. We choose $n = 65\,536$ so that the particle number error is negligible. Results for the different methods are given in Tables 4.3–4.5. As above, Conf denotes the width of the confidence interval and CF denotes the convergence factors (quotients of subsequent values). The CPU times for the splitting schemes are measured relative to the CPU time for the no-splitting scheme.

First order of the time step error is observed for the Euler-splitting scheme, while the Strang-splitting scheme provides second order. The no-splitting scheme avoids any time step error. Note that 0.0267 would be a 1% error.

Efficiency. The effort is roughly determined by the sum of the mean number of collisions N_{coll} and the mean number of kicks N_{kick}. These quantities can be predicted rather accurately.

The mean number of collisions is (cf. (3.8), (4.1))

$$N_{\mathrm{coll}}(n) = 2\,(n-1)\,s_M.$$

TABLE 4.6
Example (4.9) *for* $n = 65\,536$.

Method	T_∞	Error	Conf
Euler	109.596	9.596	0.1711
Strang	99.739	0.261	0.1425
no-split	99.981	0.019	0.1877

This quantity does not depend on the particular splitting procedure. The number of kicks is easily calculated from the other parameters. In the no-splitting scheme one obtains

$$N_{\text{kick}}^{\text{nosplit}}(n) = 2\,N_{\text{coll}}(n) + n\,M, \qquad M = \frac{s_M - s_0}{\Delta t_{\max}},$$

since at each collision both partners are kicked and, in addition, all particles are kicked before making a measurement. In the other methods one obtains

$$N_{\text{kick}}^{\text{split}}(n) = \frac{s_M}{\Delta t}\,n,$$

independently of the particular way of splitting. Accordingly, the effort is roughly the same for the splitting and no-splitting schemes if

$$4\,(n-1)\,s_M + n\,\frac{s_M - s_0}{\Delta t_{\max}} \;\sim\; \frac{s_M}{\Delta t}\,n$$

or

$$\frac{1}{\Delta t} \sim 4 + \frac{1}{\Delta t_{\max}}.$$

Thus, all methods have a similar effort for $\Delta t \sim \frac{1}{4}$, i.e., $K \sim 32$. In general, the effort for the splitting schemes increases inversely proportional to the time step. Note that these predictions are confirmed by the CPU measurements in Tables 4.3–4.5.

In conclusion, the Euler-splitting scheme needs running time about 100 times longer than the no-splitting scheme to cover the correct temperature by the confidence interval. The Strang-splitting scheme needs running time only about two times longer. Alternatively, with the same effort, the error for the Euler-splitting scheme is 100 times bigger than that for the no-splitting scheme, while the error for the Strang-splitting scheme is two times bigger.

These conclusions are qualitatively confirmed by a rough test for another parameter configuration, namely,

$$\alpha = 0.5, \qquad \beta = 37.5 \tag{4.9}$$

instead of (4.6). In this case the exact asymptotic value of the temperature is $T(\infty) = 100$ (cf. (4.4)). All other parameters are as above, in particular, $K = 32$, so that all three methods have approximately the same effort. The results are given in Table 4.6.

5. Concluding remarks. The direct simulation Monte Carlo (DSMC) method is one of the basic tools for the numerical treatment of nonlinear kinetic equations so that improvements of its efficiency are of significant practical importance. In this paper we considered a particular application, namely, the uniformly heated inelastic Boltzmann equation. We investigated the performance of two new DSMC algorithms

compared to a commonly used procedure. The order of convergence with respect to the numerical parameters (number of particles, time step) as well as the computational efficiency of the algorithms were studied both in the transient case (approximation of the solution on a finite time interval) and in the steady state case (approximation of the stationary solution). One scheme uses the Strang-splitting strategy instead of the Euler-splitting scheme. It provides second order time step convergence instead of first order. The other scheme is based on an appropriate Markov process avoiding any splitting procedure. It can be considered as providing infinite order time step convergence. All schemes are of first order with respect to the inverse particle number. In our particular numerical test cases, both the Strang-splitting scheme and the no-splitting scheme were up to two orders of magnitude more efficient than the Euler-splitting scheme.

Here we comment on the relevance of the results for more general applications. The first direction of generalization concerns the type of the driving force in (1.1). The adaptation of the schemes to other mechanisms instead of Brownian motion, e.g., to deterministic force terms as in [13], is rather obvious. We expect that the main messages of the paper concerning "Strang versus Euler" and "no-splitting scheme" remain valid.

It should be emphasized that in the case of inelastic collisions the spatially homogeneous situation is of independent interest, since there is some "nontrivial" behavior as, for example, a non-Maxwellian steady state. This issue has been intensively studied in recent years. The no-splitting scheme is very useful for investigating "fine" properties of the solution, as higher moments or tails of the steady state distribution. It is remarkable that the no-splitting scheme not only avoids the time step discretization error, but also is usually even more efficient than the other schemes. The quantitative value of the efficiency gain depends on the concrete example, in particular, on the level of "acceptable" time step error: if big time steps are sufficient, there is less or no efficiency gain; if small time steps are required, the efficiency gain may be rather significant.

Another interesting aspect of the present study is that it throws some additional light on the controversial issue about the order of the time step error in the elastic case ($\beta = 0$, $\alpha = 1$). Without going into detail, we refer to [19, section 3.5.5] concerning a discussion of this matter for the DSMC method in rarefied gas dynamics. Since temperature is not conserved, it provides a simple nontrivial test example in the inelastic situation. This is in contrast to the elastic case, where the time step issue can be studied only in spatially inhomogeneous examples.

A second direction of further study concerns the spatially inhomogeneous situation. In this case a term (v, ∇_x) is added to (1.1) and the solution $f(t, x, v)$ depends on three more variables (position coordinates x). The direct simulation process introduced in section 2 can be adapted to this situation. In addition to being accelerated by a random force, particles change their positions between collisions. However, DSMC algorithms in engineering applications [6] are based on splitting. The point is that it would be computationally too expensive to take into account the relative positions for the whole system at all times. The splitting should be performed by the Strang strategy, moreover, since the Strang and Euler schemes have basically the same effort per trajectory. The no-splitting approach provides alternatives for the splitting procedure in the spatially inhomogeneous situation. For example, the acceleration term might be combined either with the motion term or with the collision term. Additional time step errors should be avoided, whenever this is possible, and the no-splitting idea may help to do so.

REFERENCES

[1] A. Barrat, T. Biben, Z. Rácz, E. Trizac, and F. van Wijland, *On the velocity distributions of the one-dimensional inelastic gas*, J. Phys. A, 35 (2002), pp. 463–480.

[2] G. A. Bird, *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*, Clarendon Press, Oxford, UK, 1994.

[3] A. V. Bobylev and T. Ohwada, *On the generalization of Strang's splitting scheme*, Riv. Mat. Univ. Parma (6), 2* (1999), pp. 235–243.

[4] J. J. Brey, M. J. Ruiz-Montero, and D. Cubero, *Homogeneous cooling state of a low-density granular flow*, Phys. Rev. E (3), 54 (1996), pp. 3664–3671.

[5] N. V. Brilliantov and T. Pöschel, *Kinetic Theory of Granular Gases*, Oxford University Press, Oxford, UK, 2004.

[6] M. Capitelli, ed., *Rarefied Gas Dynamics*, Proceedings of the 24th International Symposium (Bari, Italy, 2004), AIP Conf. Proc. 762, AIP, New York, 2005.

[7] S. Caprino, M. Pulvirenti, and W. Wagner, *Stationary particle systems approximating stationary solutions to the Boltzmann equation*, SIAM J. Math. Anal., 29 (1998), pp. 913–934.

[8] F. Filbet and G. Russo, *Accurate numerical methods for the Boltzmann equation*, in Modeling and Computational Methods for Kinetic Equations, Model. Simul. Sci. Eng. Technol., Birkhäuser Boston, Boston, MA, 2004, pp. 117–145.

[9] F. Filbet, L. Pareschi, and G. Toscani, *Accurate numerical methods for the collisional motion of (heated) granular flows*, J. Comput. Phys., 202 (2005), pp. 216–235.

[10] I. M. Gamba, V. Panferov, and C. Villani, *On the Boltzmann equation for diffusively excited granular media*, Comm. Math. Phys., 246 (2004), pp. 503–541.

[11] I. M. Gamba, S. Rjasanow, and W. Wagner, *Direct simulation of the uniformly heated granular Boltzmann equation*, Math. Comput. Modelling, 42 (2005), pp. 683–700.

[12] C. Graham and S. Méléard, *Stochastic particle approximations for generalized Boltzmann models and convergence estimates*, Ann. Probab., 25 (1997), pp. 115–132.

[13] J. M. Montanero and A. Santos, *Computer simulations of uniformly heated granular fluids*, Granular Matter, 2 (2000), pp. 53–64.

[14] V. V. Nekrutkin and N. I. Tur, *On the justification of a scheme of direct modelling of flows of rarefied gases*, Zh. Vychisl. Mat. i Mat. Fiz., 29 (1989), pp. 1380–1392 (in Russian).

[15] T. Ohwada, *Higher order approximation methods for the Boltzmann equation*, J. Comput. Phys., 139 (1998), pp. 1–14.

[16] T. Pöschel and N. Brilliantov, eds., *Granular Gas Dynamics*, Lecture Notes in Phys. 624, Springer-Verlag, Berlin, New York, 2003.

[17] T. Pöschel and S. Luding, eds., *Granular Gases*, Lecture Notes in Phys. 564, Springer-Verlag, Berlin, New York, 2001.

[18] S. Rjasanow and W. Wagner, *On time counting procedures in the DSMC method for rarefied gases*, Math. Comput. Simulation, 48 (1998), pp. 151–176.

[19] S. Rjasanow and W. Wagner, *Stochastic Numerics for the Boltzmann Equation*, Springer Ser. Comput. Math. 37, Springer-Verlag, Berlin, 2005.

[20] G. Strang, *On the construction and comparison of difference schemes*, SIAM J. Numer. Anal., 5 (1968), pp. 506–517.

[21] J. S. van Zon and F. C. MacKintosh, *Velocity distributions in dissipative granular gases*, Phys. Rev. Lett., 93 (2004), article 038001.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 68–82

FRAMEWORK FOR THE A POSTERIORI ERROR ANALYSIS OF NONCONFORMING FINITE ELEMENTS*

CARSTEN CARSTENSEN†, JUN HU‡, AND ANTONIO ORLANDO§

Abstract. This paper establishes a unified framework for the a posteriori error analysis of a large class of nonconforming finite element methods. The theory assures reliability and efficiency of explicit residual error estimates up to data oscillations under the conditions (H1)–(H2) and applies to several nonconforming finite elements: the Crouzeix–Raviart triangle element, the Han parallelogram element, the nonconforming rotated (NR) parallelogram element of Rannacher and Turek, the constrained NR parallelogram element of Hu and Shi, the P_1 element on parallelograms due to Park and Sheen, and the DSSY parallelogram element. The theory is extended to include 1-irregular meshes with at most one hanging node per edge.

Key words. nonconforming quadrilateral finite elements, a posteriori error analysis

AMS subject classifications. 65N30, 65R20, 73C50

DOI. 10.1137/050628854

1. Introduction.

Nonconforming finite element methods are very appealing for the numerical approximation of partial differential equations, for they enjoy better stability properties compared to the conforming finite elements. While the study of the approximation properties of nonconforming triangular and quadrilateral elements has reached a certain level of maturity [3, 18, 27], the a posteriori error analysis of nonconforming quadrilateral finite element approximations is still in its infancy.

Following the contribution of [16, 15] the a posteriori error analysis for the L^2 norm of the piecewise gradient of the error, $\|\nabla_h e\|_{L^2(\Omega)}$, has been carried out successfully for triangular elements [9, 1] on the basis of two arguments: (a) the Helmholtz decomposition of $\nabla_h e$, and (b) some orthogonality with respect to some conforming finite element space V_h^c. Condition (b) fails for some quadrilateral nonconforming finite elements, e.g., the nonconforming rotated quadrilateral element of Rannacher and Turek, referred to as the NR element [25]. As a result, the a posteriori error analysis of $\|\nabla_h e\|_{L^2(\Omega)}$ for nonconforming quadrilateral elements appears as a minefield. For the NR element, for instance, the work [23] bypasses condition (b) by some enlargement of V_h^{nc} with local bubble trial functions, but their analysis applies only to goal-oriented error control and cannot be extended to the control of $\|\nabla_h e\|_{L^2(\Omega)}$. Another inherent mathematical difficulty for the NR element functions results from the nonequivalence of the continuity at midpoints *and* the equality of integral averages along edges. This makes the operator Π in [2] *not* well defined (while correct for all triangular elements of [1]).

*Received by the editors April 11, 2005; accepted for publication (in revised form) June 30, 2006; published electronically January 12, 2007. This work was supported by the DFG Research Center *MATHEON* "Mathematics for Key Technologies" in Berlin.

http://www.siam.org/journals/sinum/45-1/62885.html

†Institut für Mathematik, Humboldt-Universität zu Berlin, Unter den Linden 6, D-12489 Berlin, Germany (cc@math.hu-berlin.de).

‡LMAM and School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (hujun@math.pku.edu.cn). The research of this author was supported by the Alexander von Humboldt Foundation through the Alexander von Humboldt Fellowship.

§School of Engineering, Swansea University, Singleton Park, Swansea SA2 8PP, UK (a.orlando@swansea.ac.uk). The research of this author was supported by DFG Schwerpunktprogram 1095.

This paper aims to clarify and develop a unified framework for the a posteriori error analysis of nonconforming finite element methods based on properties for meshes obtained through affine mappings. The resulting framework is exemplified in the two-dimensional elliptic model problem

$$\operatorname{div} \nabla u = f \text{ in } \Omega, \quad u = u_D \text{ on } \Gamma_D, \quad \nabla u \cdot \nu = g \text{ on } \Gamma_N \tag{1.1}$$

on some Lipschitz domain $\Omega \subset \mathbb{R}^2$ with the outward unit normal ν along $\partial\Omega := \Gamma_D \cup \Gamma_N$. Let $V := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$ denote the space of the test functions approximated by conforming, $V_{h,0}^c$, and nonconforming, $V_{h,0}^{nc}$, finite element spaces associated with a shape regular triangulation $\mathcal{T}$, with $\mathcal{E}$ the set of the edges and $\mathcal{E}(\Omega)$ and $\mathcal{E}(\Gamma_D)$ the interior and boundary edges, respectively. Also, define $[v_h]$ as the jump across $E \in \mathcal{E}(\Omega)$ of the general discontinuous $v_h \in V_h^{nc}$ and $\mathbb{P}_k(\omega)$ the polynomials of total degree k on the domain ω. Throughout the paper, the hypotheses (H1)–(H2) characterize some class of nonconforming finite elements allowing for efficient and reliable error control.

(H1) For all $v_h \in V_h^{nc}$ there holds

$$\int_E [v_h]\, ds = 0 \text{ for } E \in \mathcal{E}(\Omega) \quad \text{and} \quad \int_E (v_h - u_D)\, ds = 0 \text{ for } E \in \mathcal{E}(\Gamma_D). \tag{1.2}$$

(H2) There exists some bounded, linear operator $\Pi : V \mapsto V_{h,0}^{nc}$ and some mesh size independent constant C with the properties (1.3)–(1.5) for every $v_h \in V_{h,0}^c$, $K \in \mathcal{T}$, and $E \in \mathcal{E}$,

$$\int_K \nabla w_h \cdot \nabla (v_h - \Pi v_h)\, dx = 0 \quad \text{for all } w_h \in V_h^{nc}; \tag{1.3}$$

$$\int_K (v_h - \Pi v_h)\, dx = 0; \quad \int_E (v_h - \Pi v_h)\, ds = 0; \tag{1.4}$$

$$\|\nabla \Pi v_h\|_{L^2(K)} \le C \|\nabla v_h\|_{L^2(K)}. \tag{1.5}$$

The main result of the paper (Theorem 3.1 below) establishes the reliability of

$$\eta^2 := \sum_{K \in \mathcal{T}} \eta_K^2 + \sum_{E \in \mathcal{E}} \eta_E^2, \quad \text{with} \tag{1.6}$$

$$\eta_K^2 := h_K^2 \|f + \operatorname{div} \nabla u_h\|_{L^2(K)}^2 \text{ for } K \in \mathcal{T}; \tag{1.7}$$

$$\eta_E^2 := h_E \big(\|J_{E,\nu}\|_{L^2(E)}^2 + \|J_{E,\tau}\|_{L^2(E)}^2 \big) \text{ for } E \in \mathcal{E}, \tag{1.8}$$

up to the data oscillations $\operatorname{osc}(f)$ and $\operatorname{osc}(g)$ (see section 2.5 below):

$$\|\nabla_h (u - u_h)\|_{L^2(\Omega)} \le C(\eta + \operatorname{osc}(f) + \operatorname{osc}(g)), \tag{1.9}$$

with $J_{E,\nu}$ and $J_{E,\tau}$ defined by (2.9) and (2.10), respectively.

The weak continuity condition (H1) is met by quite a large class of nonconforming finite elements proposed in the literature [14, 19, 25, 17, 24, 21]. However, there are also elements that fail the above condition, for instance, the version of the Rannacher–Turek element [25] with local degree of freedom equal to the value of the function at the midside nodes of each edge, and the nonconforming quadrilateral element of Wilson et al. [29]. Both elements are therefore ruled out by the present analysis.

Condition (H2) represents a key assumption of the theory. It weakens the orthogonality condition (b) mentioned above (see Lemma 3.3 below) by means of an estimate depending on data oscillations and allows the analysis of nonconforming finite elements obtained through affine mappings.

The efficiency of η in the sense that there exists a mesh size–independent constant C such that

$$\eta \leq C(\|\nabla_h e\|_{L^2(\Omega)} + \operatorname{osc}(f) + \operatorname{osc}(u_D) + \operatorname{osc}(g)), \tag{1.10}$$

with $\operatorname{osc}(u_D)$ defined in section 2.5, can be proved by adapting the arguments from [28, pp. 15–18] and [16, 9].

An outline of the remaining parts of the paper is as follows. Section 2 displays the setup of the model problem (1.1), and introduces the conforming and nonconforming finite element spaces as well as the a posteriori error estimate (1.6) and the data oscillations in (1.9). Theorem 3.1 shows that the abstract conditions (H1)–(H2) imply the reliability in the sense of (1.9). This is stated and proved in section 3 in the abstract framework, while the relevant examples follow in section 4. Namely, applications of the theory are given for the Crouzeix–Raviart element, the Han element [19], the NR element [25] with local degrees of freedom equal to the average value over the edges, the constrained NR element of Hu and Shi [21], the P_1 quadrilateral element of Park and Sheen [24], and the DSSY element [17]. Section 4 concludes with a discussion of the applicability of the theory to 1-irregular meshes, with at most one hanging node per edge, and its generalization to elliptic systems. Section 5 describes an adaptive finite element method and a numerical example for the NR element with hanging nodes.

2. Notation and preliminaries.

2.1. Model problem. Let Ω be a polygonal domain in $\mathbb{R}^2$ with boundary $\Gamma := \partial\Omega$ split into a closed Dirichlet boundary $\Gamma_D \subseteq \Gamma$ with positive surface measure and the remaining Neumann boundary $\Gamma_N := \Gamma \setminus \Gamma_D$. Given $f \in L^2(\Omega)$, $g \in L^2(\Gamma_N)$, $u_D \in H^{1/2}(\Gamma_D)$, and $V := \{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$, the solution of (1.1) satisfies

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx + \int_{\Gamma_N} g v \, ds \quad \text{for every } v \in V, \tag{2.1}$$

where the symbol $\cdot$ is the scalar product in the Euclidean space $\mathbb{R}^2$. Furthermore, we denote by L^2 the Lebesgue space of square integrable functions, and by H^s with $s > 0$ the Sobolev space defined in the usual way [18]. For the corresponding norm we use the symbols $\|\cdot\|_{L^2}$ and $\|\cdot\|_{H^s}$, respectively, with explicit indication of the domain of integration. With Ω an open set of $\mathbb{R}^2$, and $\varphi \in H^1(\Omega)$, the curl and gradient operators are given as

$$\operatorname{curl}\varphi = (-\partial\varphi/\partial x_2, \, \partial\varphi/\partial x_1), \quad \nabla\varphi = (\partial\varphi/\partial x_1, \, \partial\varphi/\partial x_2), \tag{2.2}$$

whereas for an $\mathbb{R}^2$-valued function $v = (v_1, v_2)$ the divergence is

$$\operatorname{div} v = \partial v_1/\partial x_1 + \partial v_2/\partial x_2. \tag{2.3}$$

Throughout the paper, the letter C denotes a generic constant, not necessarily the same at each occurrence.

2.2. Conforming finite element spaces. For approximating (2.1) by the finite element method, we introduce a regular triangulation $\mathcal{T}$ of $\bar{\Omega} \subset \mathbb{R}^2$ in the sense of Ciarlet [12, 6] into closed triangles, and/or convex quadrilaterals, such that $\bigcup_{K\in\mathcal{T}} K = \bar{\Omega}$, two distinct elements K and K' in $\mathcal{T}$ are either disjoint, or share the common edge E, or a common vertex; that is, hanging nodes at this stage are not allowed, and we refer to section 4.6 and [11] for further discussion. Let $\mathcal{E}$ denote the set of all edges in $\mathcal{T}$, $\mathcal{N}$ the set of vertices of the elements $K \in \mathcal{T}$, and $\mathcal{N}_m$ the set of the midside nodes m_E of the edges $E \in \mathcal{E}$. The set of interior edges of Ω are denoted by $\mathcal{E}(\Omega)$, the set of edges of the element K by $\mathcal{E}(K)$, whereas those that belong to the Dirichlet and Neumann boundary are denoted by $\mathcal{E}(\Gamma_D)$ and $\mathcal{E}(\Gamma_N)$, respectively. For the set of midpoints of the edges $E \in \mathcal{E}(\Gamma_D)$ we use the notation $\mathcal{N}_m(\Gamma_D)$. By h_K and h_E we denote the diameter of the element $K \in \mathcal{T}$ and of the edge $E \in \mathcal{E}$, respectively. Also, we denote by ω_K the patch of elements $K' \in \mathcal{T}$ that share an edge with K, and by ω_E the patch of elements having in common the edge E. Given any edge $E \in \mathcal{E}$ we assign one fixed unit normal ν_E; if (ν_1, ν_2) are its components, τ_E denotes the orthogonal vector of components $(-\nu_2, \nu_1)$. For $E \in \mathcal{E}(\Gamma_D) \cup \mathcal{E}(\Gamma_N)$ on the boundary we choose $\nu_E = \nu$, the unit outward normal to Ω, and concordantly the unit tangent vector τ. Once ν_E and τ_E have been fixed on E, in relation to ν_E one defines the elements $K_{in} \in \mathcal{T}$ and $K_{out} \in \mathcal{T}$, with $E = K_{out} \cap K_{in}$, as depicted in Figure 1.

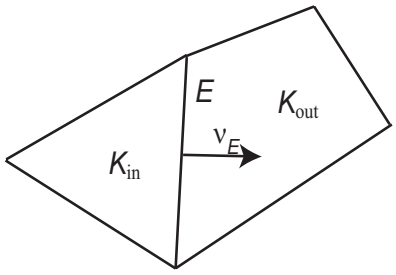


FIG. 1. *Definition of the elements K_{in} and K_{out} in relation to ν_E.*

Given $E \in \mathcal{E}(\Omega)$ and an $\mathbb{R}^d$-valued function v defined in Ω, with $d = 1, 2$, we denote by $[v]_E$ the jump of v across E, that is,

$$[v]_E(x) = (v|_{K_{out}}(x) - v|_{K_{in}}(x)) \quad \text{for } x \in E = K_{in} \cap K_{out};$$

the subscript E will be omitted whenever it is clear from the context.

With the triangulation $\mathcal{T}$ we associate, moreover, the space $H^1(\mathcal{T})$ defined as

$$H^1(\mathcal{T}) = \{v \in L^2(\Omega) : \forall K \in \mathcal{T},\, v|_K \in H^1(K)\},$$

and for $v \in H^1(\mathcal{T})$, we denote by $\nabla_h v$ the gradient operator defined piecewise with respect to $\mathcal{T}$, i.e.,

$$\nabla_h v|_K := \nabla(v|_K).$$

Whenever it is clear from the context that we are considering the restriction of v to an element $K \in \mathcal{T}$, then we clearly write only ∇v in lieu of $\nabla_h v$.

For a nonnegative integer k the space $Q_k(\omega)$ consists of polynomials of total degree at most k defined over ω in the case in which $\omega = K$ is a triangle, whereas it denotes polynomials of degree at most k in each variable in the case in which K is a

quadrilateral. For this presentation it will suffice to assume $k = 1$. The corresponding conforming space will be denoted by

$$V_h^c := \{v \in H^1(\Omega) : v|_K \in Q_1(K)\} \quad \text{and} \quad V_{h,0}^c := \{v \in V_h^c : v = 0 \text{ on } \Gamma_D\}.$$

Throughout the paper, for triangular elements, $V_{h,0}^c$ stands for the conforming space of P_1 elements, whereas for quadrilateral elements it denotes the conforming space of bilinear elements.

Given the conforming finite element space $V_{h,0}^c$, we consider the Clément interpolation operator or any other regularized conforming finite element approximation operator $\mathcal{J} : H^1(\Omega) \mapsto V_h^c$ with the property

$$\|\nabla \mathcal{J}\varphi\|_{L^2(K)} + \|h_K^{-1}(\varphi - \mathcal{J}\varphi)\|_{L^2(K)} \le C\|\nabla\varphi\|_{L^2(\omega_K)}, \tag{2.4}$$

$$\|h_E^{-1/2}(\varphi - \mathcal{J}\varphi)\|_{L^2(E)} \le C\|\nabla\varphi\|_{L^2(\omega_E)} \tag{2.5}$$

for all $K \in \mathcal{T}$, $E \in \mathcal{E}$, and $\varphi \in H^1(\Omega)$. The existence of such operators is guaranteed, for instance, in [13, 26, 7, 5].

2.3. Nonconforming finite element spaces and a posteriori error estimator. A nonconforming finite element approximation is defined by a finite-dimensional trial space $V_h^{nc} \subset H^1(\mathcal{T})$ along with the test space $V_{h,0}^{nc}$ corresponding to the discrete homogeneous Dirichlet boundary conditions. The nonconforming finite element approximation $u_h \in V_h^{nc}$ of (2.1) then satisfies

$$\int_\Omega \nabla_h u_h \cdot \nabla_h v_h \, dx = \int_\Omega f v_h \, dx + \int_{\Gamma_N} g v_h \, ds \quad \text{for every } v_h \in V_{h,0}^{nc}. \tag{2.6}$$

The Helmholtz decomposition is a well-established tool in the a posteriori error analysis of nonconforming finite element methods [16, 9].

LEMMA 2.1. *Given any $e \in V + V_h^{nc}$ such that $\nabla_h e \in L^2(\Omega;\mathbb{R}^2)$ there exist w, $\varphi \in H^1(\Omega)$ with $w = 0$ on Γ_D, and $\nabla\varphi \cdot \tau = \operatorname{curl}\varphi \cdot \nu = 0$ on Γ_N such that*

$$\nabla_h e = \nabla w + \operatorname{curl}\varphi, \tag{2.7}$$

$$\|\nabla_h e\|_{L^2(\Omega)}^2 = \|\nabla w\|_{L^2(\Omega)}^2 + \|\operatorname{curl}\varphi\|_{L^2(\Omega)}^2. \tag{2.8}$$

2.4. A posteriori error estimator. For each edge $E \in \mathcal{E}$, define $J_{E,\nu}$ the jump of $\nabla_h u_h$ across E in direction ν_E, i.e.,

$$J_{E,\nu} := \begin{cases} [\nabla_h u_h]_E \cdot \nu_E & \text{if } E \in \mathcal{E}_\Omega, \\ g - \nabla u_h \cdot \nu & \text{if } E \in \mathcal{E}_N, \\ 0 & \text{if } E \in \mathcal{E}_D, \end{cases} \tag{2.9}$$

and $J_{E,\tau}$ the jump of $\nabla_h u_h$ across E in direction t_E, i.e.,

$$J_{E,\tau} := \begin{cases} [\nabla_h u_h]_E \cdot \tau_E & \text{if } E \in \mathcal{E}_\Omega, \\ 0 & \text{if } E \in \mathcal{E}_N, \\ (\nabla u_D - \nabla u_h) \cdot \tau & \text{if } E \in \mathcal{E}_D, \end{cases} \tag{2.10}$$

and recall η from (1.6) with the local contributions η_K (1.7) and η_E (1.8) for each $K \in \mathcal{T}$ and $E \in \mathcal{E}$, respectively.

2.5. Data oscillations. For $f \in L^2(\Omega)$ and its piecewise constant approximation f_h with respect to $\mathcal{T}$, we refer to $\operatorname{osc}(f)$ as the oscillation of f [28],

$$\operatorname{osc}^2(f) := \sum_{K\in\mathcal{T}} h_K^2 \|f - f_h\|^2_{L^2(K)}, \tag{2.11}$$

with $\operatorname{osc}(f)$ being a higher order term if $f \in H^1(\Omega)$. Similar definitions hold for the oscillations $\operatorname{osc}(u_D)$ and $\operatorname{osc}(g)$ of the Dirichlet and Neumann boundary data, $u_D \in H^{1/2}(\Gamma_D)$ and $g \in L^2(\Gamma_N)$, and their piecewise affine and constant approximations $u_{D,h}$ and g_h, respectively, as [28, 8]

$$\operatorname{osc}^2(u_D) := \sum_{E\in\mathcal{E}(\Gamma_D)} h_E \left\| \frac{\partial}{\partial s}(u_D - u_{D,h}) \right\|^2_{L^2(E)},$$

$$\operatorname{osc}^2(g) := \sum_{E\in\mathcal{E}(\Gamma_N)} h_E \|g - g_h\|^2_{L^2(E)}.$$

3. Reliability of η. This section presents the main result of this paper, that is, (H1)–(H2) imply the reliability of η. Throughout this section, let u solve (2.1), let u_h solve (2.6), and set $e := u - u_h$.

THEOREM 3.1. *Assume that the space V_h^{nc} along with the corresponding $V_{h,0}^{nc}$ satisfy* (H1)–(H2). *Then there exists a positive constant C depending only on the minimum angle of $\mathcal{T}$ such that η is reliable in the sense that*

$$\|\nabla_h e\|_{L^2(\Omega)} \le C\big(\eta + \operatorname{osc}(f) + \operatorname{osc}(g)\big). \tag{3.1}$$

The remainder of this section is devoted to the proof of Theorem 3.1.

We establish first some interpolation error estimates for the operator Π in (H2).

LEMMA 3.2. *Given the operator Π meeting* (H2), *there then exists some mesh size–independent constant C such that there holds*

$$\begin{aligned} h_K^{-1}\|v_h - \Pi v_h\|_{L^2(K)} + \|\nabla(v_h - \Pi v_h)\|_{L^2(K)} &\le C\|\nabla v_h\|_{L^2(K)}, \\ h_E^{-1/2}\|v_h - \Pi v_h\|_{L^2(E)} &\le C\|\nabla v_h\|_{L^2(\omega_E)}. \end{aligned} \tag{3.2}$$

Proof. Let Π_0^K denote the mean average operator over K. Using condition $(1.4)_1$ with $\Pi_0^K v_h = \Pi_0^K \Pi\, v_h$, the triangular inequality, and (1.5), one obtains

$$\begin{aligned} \|v_h - \Pi v_h\|_{L^2(K)} &\le \|v_h - \Pi_0^K v_h\|_{L^2(K)} + \|\Pi_0^K \Pi v_h - \Pi v_h\|_{L^2(K)} \\ &\le C(h_K\|\nabla v_h\|_{L^2(K)} + h_K\|\nabla v_h\|_{L^2(K)}). \end{aligned} \tag{3.3}$$

A triangular inequality and (1.5) also gives

$$\|\nabla v_h - \nabla\,\Pi v_h\|_{L^2(K)} \le C\|\nabla v_h\|_{L^2(K)},$$

which, combined with (3.3), finally yields $(3.2)_1$. Arguing in a similar way and using the trace theorem [6, 12] one obtains $(3.2)_2$. ∎

Here and throughout, f_h and g_h denote piecewise constant approximations of f and g, respectively. From (H2) and for every $v_h \in V_{h,0}^c$, the following holds:

$$\int_\Omega \nabla_h u_h \cdot \nabla v_h\, dx = \int_\Omega f\,\Pi v_h\, dx + \int_{\Gamma_N} g\,\Pi v_h\, ds. \tag{3.4}$$

LEMMA 3.3. *There exists a mesh size–independent constant C such that, for every $v_h \in V_{h,0}^c$, the following holds:*

$$\int_\Omega \nabla_h e \cdot \nabla v_h \, dx \le C(\mathrm{osc}(f) + \mathrm{osc}(g)) \, \|\nabla v_h\|_{L^2(\Omega)}. \tag{3.5}$$

Proof. From (2.1) and (3.4), for every $v_h \in V_{h,0}^c$ it follows that

$$\int_\Omega \nabla_h e \cdot \nabla v_h \, dx = \sum_{K\in\mathcal{T}} \Big(\int_K (f - f_h)(v_h - \Pi v_h) \, dx + \int_K f_h(v_h - \Pi v_h) \, dx \Big) + \sum_{E\in\mathcal{E}(\Gamma_N)} \Big(\int_E (g - g_h)(v_h - \Pi v_h) \, ds + \int_E g_h(v_h - \Pi v_h) \, ds \Big).$$

Since (1.4), this equals

$$\int_\Omega (f - f_h)(v_h - \Pi v_h) \, dx + \int_{\Gamma_N} (g - g_h)(v_h - \Pi v_h) \, ds.$$

The combination of Cauchy inequalities with (3.2) yields its upper bound:

$$C\bigg(\Big(\sum_{K\in\mathcal{T}} h_K^2 \|f - f_h\|_{L^2(K)}^2 \Big)^{1/2} + \Big(\sum_{E\in\mathcal{E}(\Gamma_N)} h_E \|g - g_h\|_{L^2(E)}^2 \Big)^{1/2} \bigg) \|\nabla v_h\|_{L^2(\Omega)}. \qquad \square$$

Remark 1. If $V_{h,0}^c$ is a subspace of $V_{h,0}^{nc}$, then (H1)–(H2) hold for $\Pi = I$ and (3.5) recovers the L^2-orthogonality of $\nabla_h e$ and ∇v_h for every $v_h \in V_{h,0}^c$ (because $C = 0$ in (3.2)).

The following orthogonality condition (3.6) is well established in the literature on a posteriori error estimates for nonconforming finite element schemes.

LEMMA 3.4. *For every $v_h \in V_h^c$ such that $\partial v_h/\partial s = 0$ on Γ_N, it holds that*

$$\int_\Omega \nabla_h e \cdot \mathrm{curl}\, v_h \, dx = 0. \tag{3.6}$$

Proof. The proof is along the lines of [16, eqn. (3.4)] for the Crouzeix–Raviart element. An integration by parts over each element gives

$$\int_\Omega \nabla_h e \cdot \mathrm{curl}\, v_h \, dx = \sum_{E\in\mathcal{E}} \int_E [u - u_h] \frac{\partial v_h}{\partial s} \, ds. \tag{3.7}$$

Since for $v_h \in V_h^c$, $\partial v_h/\partial s$ is constant over each edge $E \in \mathcal{E}(\Omega) \cup \mathcal{E}(\Gamma_D)$, or is zero on $E \in \mathcal{E}(\Gamma_N)$, accounting for (H1), one obtains (3.6). $\square$

The proof of (3.1) starts with the decomposition (2.7), the interpolation operator $\mathcal{J}$ of Clément, and Lemma 3.4. Without loss of generality one can choose φ in (2.7) to be equal to a constant on Γ_N, and $\mathcal{J}\varphi|_{\Gamma_N} = \varphi|_{\Gamma_N}$. Then it follows that

$$\|\nabla_h e\|_{L^2(\Omega)}^2 = \int_\Omega \nabla_h e \cdot (\nabla w + \mathrm{curl}\, \varphi) \, dx = \int_\Omega \nabla_h e \cdot \nabla(w - \mathcal{J}w) \, dx + \int_\Omega \nabla_h e \cdot \mathrm{curl}(\varphi - \mathcal{J}\varphi) \, dx + \int_\Omega \nabla_h e \cdot \nabla \mathcal{J} w \, dx.$$

From Lemma 3.3 and the estimate (2.4), one obtains

$$\begin{aligned}\int_\Omega \nabla_h e \cdot \nabla \mathcal{J} w \, dx &\le C\big(\operatorname{osc}(f) + \operatorname{osc}(g)\big)\|\nabla \mathcal{J} w\|_{L^2(\Omega)} \\ &\le C\big(\operatorname{osc}(f) + \operatorname{osc}(g)\big)\|\nabla w\|_{L^2(\Omega)}.\end{aligned} \tag{3.8}$$

Since $(w - \mathcal{J}w)$ and $(\varphi - \mathcal{J}\varphi)$ belong to $H^1(\Omega)$, the use of the Stokes theorem and Green's formula over each element gives, after some rearrangements,

$$\begin{aligned}&\int_\Omega \nabla_h e \cdot \nabla(w - \mathcal{J}w)\, dx + \int_\Omega \nabla_h e \cdot \operatorname{curl}(\varphi - \mathcal{J}\varphi)\, dx \\ &= \sum_{E\in\mathcal{E}} \left(\int_E J_{E,\tau}(\varphi - \mathcal{J}\varphi)\, ds + \int_E J_{E,\nu}(w - \mathcal{J}w)\, ds \right) \\ &+ \sum_{K\in\mathcal{T}} \int_K (f + \operatorname{div} \nabla u_h)\,(w - \mathcal{J}w)\, dx.\end{aligned}$$

It is a standard argument with Cauchy inequalities and (2.4)–(2.5) to bound this by

$$C\eta\big(\|\nabla w\|_{L^2(\Omega)} + \|\nabla \varphi\|_{L^2(\Omega)}\big),$$

with η from (1.6). The combination of the aforementioned estimates with (2.8) concludes the proof of (3.1).

4. Examples. In this section, we verify (H1)–(H2) for several nonconforming finite elements proposed in the literature and discuss the applicability of the theory to 1-irregular meshes and to elliptic systems in divergence form. For the following examples, the operator Π that enters (H2) is the interpolation operator of V associated with $V_{h,0}^{nc}$.

4.1. The Crouzeix–Raviart element. The nonconforming finite element space associated with the Crouzeix–Raviart element [14] reads

$$\begin{aligned}V_h^{nc} := \big\{ v_h \in H^1(\mathcal{T}) :\ & v_h|_K \in P_1(K) \quad \forall K \in \mathcal{T},\ v_h \text{ is continuous at each} \\ & m_E \in \mathcal{N}_m \setminus \mathcal{N}_m(\Gamma_D),\ \text{and } v_h(m_E) = u_D(m_E) \text{ for } m_E \in \mathcal{N}_m(\Gamma_D)\big\},\end{aligned} \tag{4.1}$$

and $V_{h,0}^{nc}$ denotes the space corresponding to the discrete homogeneous Dirichlet boundary conditions. For this element, it is trivial to check that the space V_h^{nc} meets (H1). Furthermore, since $V_{h,0}^c \subset V_{h,0}^{nc}$, (H2) follows immediately (see Remark 1) and Theorem 3.1 recovers the results of [16, 9].

4.2. The Han element. With respect to the global coordinate system (x_1, x_2), the nonparametric formulation of rectangular and parallelogram elements proposed by Han in [19] is obtained by introducing the local space

$$\mathcal{Q}_{\mathcal{H}}^{nc} = \operatorname{span}\left\{ 1,\, x_1,\, x_2,\, x_1^2 - \frac{5}{3}x_1^4,\, x_2^2 - \frac{5}{3}x_2^4 \right\}, \tag{4.2}$$

and the $\mathcal{Q}_{\mathcal{H}}^{nc}$-unisolvent set of linearly independent linear forms [12, 19] reads

$$\mathcal{F}_E(v) = \frac{1}{h_E}\int_E v\, ds, \quad \mathcal{F}_K(v) = \frac{1}{|K|}\int_K v\, dx \text{ with } E \in \mathcal{E}(K), \quad K \in \mathcal{T}. \tag{4.3}$$

This defines the five degrees of freedom for the Han element. In (4.3), $|K|$ denotes the area of the element. Recall from [12] that, given $E = K \cap K'$ for K, $K' \in \mathcal{T}$, and $v \in H^1(\mathcal{T})$ such that $v|_K \in \mathcal{Q}^{nc}_{\mathcal{H}}(K)$ and $v|_{K'} \in \mathcal{Q}^{nc}_{\mathcal{H}}(K')$, we say that v is continuous with respect to $\mathcal{F}_E$ if $\mathcal{F}_E(v|_K) = \mathcal{F}_E(v|_{K'})$. The nonconforming finite element space V_h^{nc} is then defined as

(4.4)
$$V_h^{nc} := \big\{ v \in H^1(\mathcal{T}) : v|_K \in \mathcal{Q}^{nc}_{\mathcal{H}}(K) \text{ for each } K \in \mathcal{T},\ v \text{ continuous with respect to } \mathcal{F}_E\ \forall\ E \in \mathcal{E}(\Omega), \text{ and } \mathcal{F}_E(v) = \mathcal{F}_E(u_D)\ \forall\ E \in \mathcal{E}(\Gamma_D) \big\},$$

whereas $V_{h,0}^{nc}$ denotes the space corresponding to the discrete homogeneous Dirichlet boundary conditions in (4.4). For $v_h \in V_h^{nc}$, the definition (4.4) of V_h^{nc} and (4.3) yield

(4.5)
$$\int_E [v_h]\, ds = 0 \ \text{ for all } E \in \mathcal{E}(\Omega) \quad \text{and} \quad \int_E (v_h - u_D)\, ds = 0 \ \text{ for all } E \in \mathcal{E}(\Gamma_D),$$

and so V_h^{nc} verifies (H1). Let V_h^c be the conforming space of the bilinear elements constructed from the local spaces $\mathcal{Q}^c(K) = \text{span}\{1, x_1, x_2, x_1x_2\}$. Consider then the interpolation operator $\Pi : V \mapsto V_{h,0}^{nc}$ defined by the following conditions: For all $E \in \mathcal{E}(K)$ and $K \in \mathcal{T}$,

$$\Pi v \in V_{h,0}^{nc}, \quad \mathcal{F}_E(\Pi v|_K) = \mathcal{F}_E(v|_K), \quad \mathcal{F}_K(\Pi v|_K) = \mathcal{F}_K(v|_K). \tag{4.6}$$

Given $v \in V_{h,0}^c$, the restriction of v to $K \in \mathcal{T}$ has the following representation:

$$v = a_0 + a_1x_1 + a_2x_2 + a_3x_1x_2 \tag{4.7}$$

for some interpolation constants a_i, $i = 0, \ldots, 3$. Since the degrees of freedom (4.3) vanish over the nonconforming bubble function $x_1x_2 \in \mathcal{Q}^c(K)$, it follows that the restriction of Π to $V_{h,0}^c$ yields [21]

$$\Pi v|_K = a_0 + a_1x_1 + a_2x_2. \tag{4.8}$$

By a scaling argument, one can verify that Π meets (1.5) and therefore the estimates (3.2). Furthermore, for every $v_h \in V_{h,0}^c$ a direct evaluation of the integrals shows (1.3)–(1.4) over rectangular and parallelogram element domains, i.e., the space V_h^{nc} meets (H2).

4.3. The quadrilateral rotated nonconforming element. In [25] Rannacher and Turek introduced two types of quadrilateral nonconforming elements referred to as NR elements. The corresponding local finite element spaces are obtained by rotating the mixed term of the bilinear element, and assuming as local degree of freedom either the average of the function over the edge or its value at the midside node. In this section we consider the nonparametric formulation for rectangular and parallelogram elements with the first choice of degree of freedom. More precisely, for each element $K \in \mathcal{T}$ and with respect to the global coordinate system (x_1, x_2), we set [25]

$$\mathcal{Q}^{nc}_{\mathcal{R}} = \text{span}\{\, 1,\, x_1,\, x_2,\, x_1^2 - x_2^2 \,\} \tag{4.9}$$

and introduce the four degrees of freedom as

$$\mathcal{F}_E(v) = \frac{1}{h_E} \int_E v\, ds \text{ with } E \in \mathcal{E}(K). \tag{4.10}$$

With the corresponding nonconforming finite element space defined as in (4.4) and concordantly $V_{h,0}^{nc}$, it follows that V_h^{nc} meets (H1).

For any $v \in V$, the interpolation operator $\Pi v \in V_{h,0}^{nc}$ is defined as in [25, 21]: For all $E \in \mathcal{E}(K)$ and $K \in \mathcal{T}$,

$$\Pi v \in V_{h,0}^{nc} \quad \text{and} \quad \mathcal{F}_E(\Pi v|_K) = \mathcal{F}_E(v|_K), \tag{4.11}$$

and, hence, as with the Han element, since $\mathcal{F}_E$ vanishes over the nonconforming bubble function $x_1x_2 \in \mathcal{Q}^c(K)$, the restriction of Π to $V_{h,0}^c \subset V$ is represented locally by (4.8) [21]. Therefore, the above arguments verify (H2).

Remark 2. For the version of the NR element with function evaluation at the midpoints as degree of freedom, (H1) is not satisfied and we refer to section 4.5 for a modification of the NR element.

Remark 3. The proof of Lemma 3.4 for the NR element can be found in [20, 22].

Remark 4. The interpolation operator $\Pi_{\mathcal{P}}$ defined in [2, eqn. (6)] does not, in general, map into the space $X_{\mathcal{P},E}$ of the NR element functions continuous at the midside nodes [2, p. 4]. This results in a gap in the analysis of [2] for this finite element; the remaining assertions in [2] seem to be correct.

Remark 5. The present analysis shows that the augmentation of V_h^{nc} with local bubble trial functions proposed in [23] is not necessary for the error control of $\|\nabla_h e\|$.

Remark 6. The flux $\nabla_h u|_K \cdot \nu_E$ is not required to be constant over each edge E with normal ν_E as in [2]. The latter hypothesis would in fact restrict the analysis to only rectangular meshes.

4.4. The constrained NR element and the P_1-quadrilateral element. The constrained NR finite element (referred to as the CNR element) introduced in [20, 21] is obtained by enforcing a constraint on the degree of freedom of the NR element described in section 4.3. With $\mathcal{Q}_{\mathcal{R}}^{nc}$ denoting here the space of the global trial functions defined over Ω and corresponding to the NR element, the space of the CNR element is then defined as follows:

$$\begin{aligned}\mathcal{Q}_{\mathcal{J}}^{nc} := \Bigg\{ v \in \mathcal{Q}_{\mathcal{R}}^{nc} : \forall K \in \mathcal{T} \int_{E_1} v\,ds + \int_{E_3} v\,ds = \int_{E_2} v\,ds + \int_{E_4} v\,ds \\ \text{with } E_i, \quad 1 \le i \le 4, \text{ edges of } K \in \mathcal{T} \text{ numbered counterclockwise} \Bigg\}.\end{aligned} \tag{4.12}$$

For rectangular and parallelogram element domains, considered here, the element is equivalent to the P_1-quadrilateral element of [24]. For homogeneous Dirichlet boundary conditions, it is trivial to check that the space V_h^{nc} meets (H1) for being the CNR space, a subspace of NR. Furthermore, in [20, 21] it is also proved that on the generic element $K \in \mathcal{T}$ with vertices 1, 2, 3, 4 labeled counterclockwise, the interpolation $\Pi v \in V_{h,0}^{nc}$ defined as in (4.11) and for $v \in V_{h,0}^c$ has the representation

$$\Pi v|_K = v_1\phi_1 + v_2\phi_2 + v_3\phi_3 + v_4\phi_4, \tag{4.13}$$

with v_i nodal value of $v \in V_{h,0}^c$ and

$$\begin{aligned}\phi_1(x_1,x_2) &= \frac{1}{4}(1 - x_1 - x_2), \quad \phi_2(x_1,x_2) = \frac{1}{4}(1 - x_1 + x_2), \\ \phi_3(x_1,x_2) &= \frac{1}{4}(1 + x_1 + x_2), \quad \phi_4(x_1,x_2) = \frac{1}{4}(1 - x_1 + x_2)\end{aligned} \tag{4.14}$$

associated with each of such vertices. The arguments of section 4.3 finally show (H2).

4.5. The DSSY element. The main motivation for the definition of this element is to obtain a quadrilateral element with approximation properties similar to those of the Crouzeix–Raviart element. For parallelogram elements these properties were identified in [17] by (i) continuity at the midpoints of each edge, (ii) value of the function at these points as degrees of freedom, and (iii) validity of the orthogonality condition [17, eqn. (6.1)]: For all $v_h \in V_{h,0}^{nc}$ there holds

$$\int_E [v_h]\, ds = 0 \quad \text{for } E \in \mathcal{E}(\Omega). \tag{4.15}$$

The latter condition plays a crucial role in the proof of optimal error estimates as realized in [17], for instance, by two spaces of local basis obtained by an ad hoc modification of the local basis of the Rannacher–Turek element. Set

$$\theta_\ell(t) = \begin{cases} t^2 - \frac{5}{3}t^4 & \text{for } \ell = 1, \\ t^2 - \frac{25}{6}t^4 + \frac{7}{2}t^6 & \text{for } \ell = 2. \end{cases} \tag{4.16}$$

Then the local space reads

$$\mathcal{Q}_{\mathcal{D}}^{nc} = \operatorname{span}\{1,\, x_1,\, x_2,\, \theta_\ell(x_1) - \theta_\ell(x_2)\} \quad \text{for } \ell = 1, 2, \tag{4.17}$$

and the $\mathcal{Q}_{\mathcal{D}}^{nc}$-unisolvent linear forms read

$$\mathcal{F}_{E_i}(v_h|_K) = v_h|_K(m_{E_i}) \quad \text{for } E_i \in \mathcal{E}(K),\ 1 \le i \le 4,\ v_h \in \mathcal{Q}_{\mathcal{D}}^{nc}, \tag{4.18}$$

with m_{E_i} midside nodes of the edge E_i. The nonconforming finite element spaces V_h^{nc} and $V_{h,0}^{nc}$ are then defined as in (4.4) with $\mathcal{Q}_{\mathcal{H}}^{nc}$ replaced by $\mathcal{Q}_{\mathcal{D}}^{nc}$. Following [17], one can show that (H1) holds. Furthermore, with the interpolation operator $\Pi : V \mapsto V_{h,0}^{nc}$ defined as in (4.11), one obtains

$$\Pi v \in V_{h,0}^{nc}, \quad \Pi v|_K(m_E) = \frac{1}{h_E}\int_E v\, ds \quad \text{for each edge } E \in \mathcal{E}(K), \quad K \in \mathcal{T}, \tag{4.19}$$

with the restriction of Π to the space $V_{h,0}^c$ having the local representation (4.8) that implies (H2).

4.6. Hanging nodes. This section discusses 1-irregular meshes and refers to [11] for further details and technicalities. Given an initial regular mesh $\mathcal{T}_0$ of Ω in the sense of Ciarlet [12, 6], a 1-irregular mesh $\mathcal{T}_\ell$ is obtained from $\mathcal{T}_{\ell-1}$ by refining some elements K into four congruent elements by connecting the midside points of the edges of K [4].

Let $\mathcal{N}_H$ denote the set of hanging nodes, $\mathcal{N}_E$ the set of the endpoints of the edges containing one hanging node, $\mathcal{E}_C$ the set of edges with one endpoint in $\mathcal{N}_H$, and $\mathcal{E}_H$ the set of edges containing one hanging node, hereafter referred to as hanging edges. We define the set $\mathcal{N}_R$ of regular nodes as $\mathcal{N}_R = \mathcal{N} \setminus (\mathcal{N}_H \cup \mathcal{N}_E)$ and the set $\mathcal{E}_R$ of regular edges as $\mathcal{E}_R = (\mathcal{E}(\Omega) \setminus \mathcal{E}_C) \cup \mathcal{E}(\Gamma_D)$. It is then possible to construct a partition of unity $(\varphi_z)_{z\in\mathcal{N}_E\cup\mathcal{N}_R}$ on Ω that forms a basis for $V_{h,0}^c$ and define a regularized operator $\mathcal{J} : H^1(\Omega) \mapsto V_h^c$ meeting (2.4)–(2.5) [11].

Under proper constraints for the degrees of freedom for the hanging edges we have the following result that controls the nonconforming part of the error [11]:

$$\min_{\substack{v\in H^1(\Omega)\\ v=u_D \text{ on } \Gamma_D}} \|\nabla_h(u_h - v)\|_{L^2(\Omega)} \le C\left(\sum_{E\in\mathcal{E}_R} h_E \|J_{E,\tau}\|_{L^2(E)}^2\right)^{1/2} + C\operatorname{osc}(u_D). \tag{4.20}$$

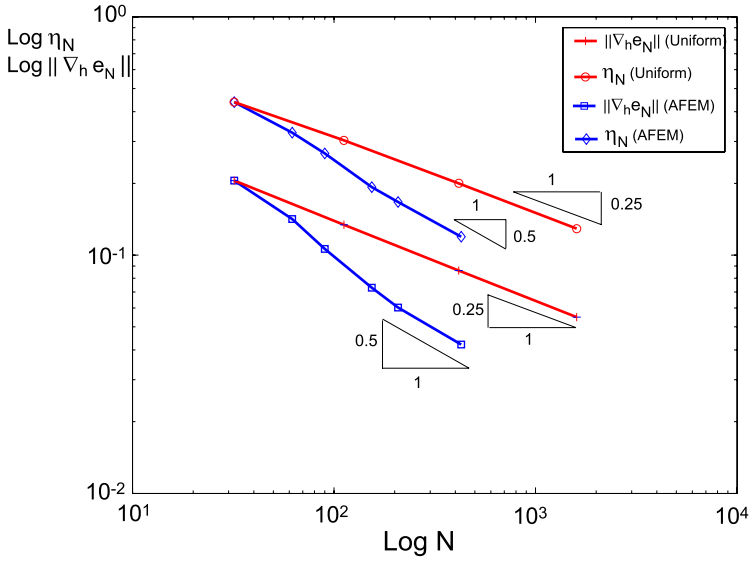


FIG. 2. *Experimental convergence rate of η_N and the exact error $\|\nabla_h e_N\|$ with respect to the number N of degrees of freedom for the adaptive and uniform refinement based on η_N and with the NR finite element. The displayed results show $2.13 \le \eta_N/\|\nabla_h e_N\| \le 2.83$ for adaptive and $2.13 \le \eta_N/\|\nabla_h e_N\| \le 2.35$ for uniform mesh refinement.*

An integration by parts, use of Young's inequality, the properties of the operator $\mathcal{J}$, and (4.20) finally prove (3.1) with $\eta + \operatorname{osc}(f) + \operatorname{osc}(g) + \operatorname{osc}(u_D)$ and corresponding modifications for the contribution to η from the hanging edges [11].

4.7. Generalizations. If $A \in L^\infty(\Omega; \mathbb{R}^{2\times 2})$ denotes a symmetric positive definite matrix piecewise constant with respect to $\mathcal{T}$, then Theorem 3.1 with corresponding modifications for the definition of η applies also to the elliptic PDE $\operatorname{div} A\nabla u = f$ with boundary conditions $u = u_D$ on Γ_D and $(A\nabla u)\cdot\nu = g$ on Γ_N.

5. Numerical experiment. This section concludes the paper with an example of an adaptive finite element model for the Poisson problem.

5.1. Adaptive finite element method. By rewriting η from (1.6) as $\eta^2 = \sum_{K\in\mathcal{T}} \eta_K^2$, with

$$\eta_K^2 := h_K^2 \|f + \operatorname{div}\nabla u_h\|_{L^2(K)}^2 + \frac{1}{2}\sum_{E\in\mathcal{E}(K)} h_E\big(\|J_{E,\nu}\|_{L^2(E)}^2 + \|J_{E,\tau}\|_{L^2(E)}^2\big),$$

the estimate η and the elemental contributions η_K can be used to generate the triangulations $\{\mathcal{T}_\ell\}_{\ell\in\mathbb{N}}$ in an adaptive way using the following algorithm.

ALGORITHM 1. *Input a coarse mesh $\mathcal{T}_0$ with rectangular and/or triangular elements, and set $\ell = 0$.*

(a) *Solve the discrete problem on $\mathcal{T}_\ell$ with N degrees of freedom.*
(b) *Compute η_K for all $K \in \mathcal{T}_\ell$ and $\eta_N := (\sum_{K\in\mathcal{T}} \eta_K^2)^{1/2}$.*
(c) *Mark $K \in \mathcal{M} \subset \mathcal{T}_\ell$ for refinement into four congruent elements by connecting the midside points of its edges if $\theta \max_{T\in\mathcal{T}_\ell} \eta_T \le \eta_K$.*
(d) *Mark further elements to ensure at most one hanging node per edge. Define the resulting mesh as the actual mesh $\mathcal{T}_{\ell+1}$, update ℓ, and go to* (a).

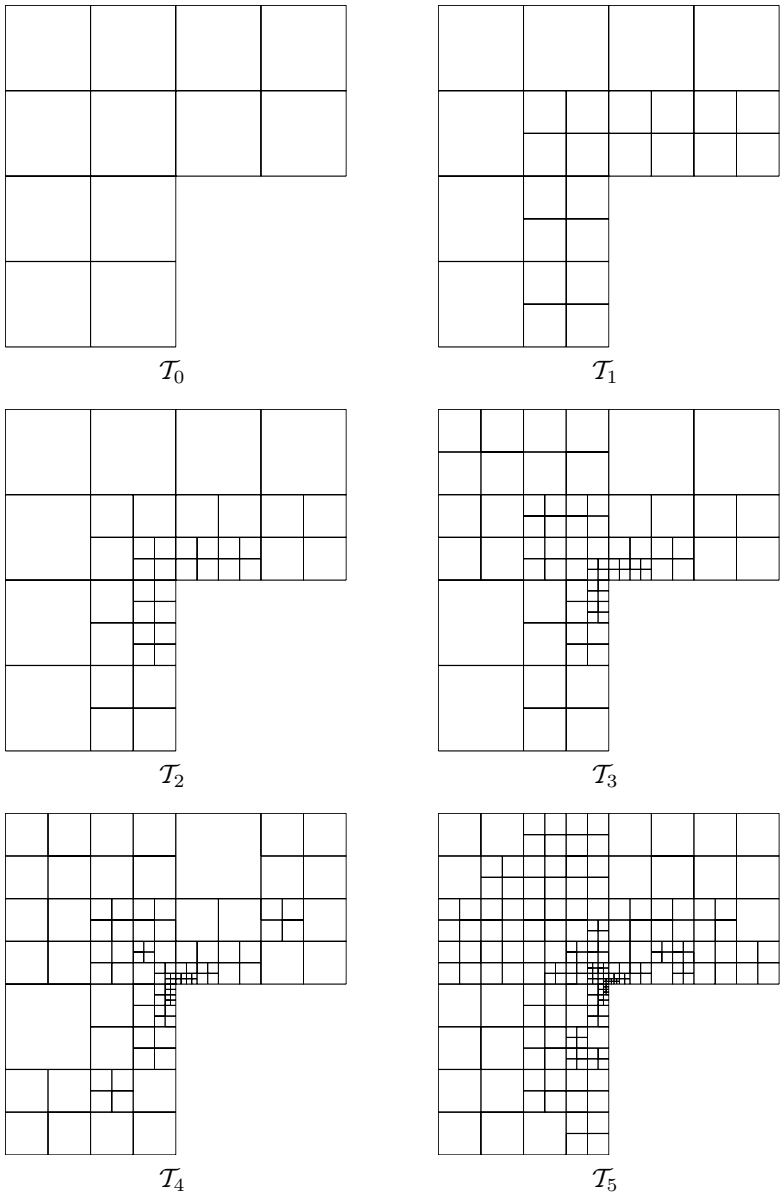


FIG. 3. *Adapted triangulations $\mathcal{T}_0, \ldots, \mathcal{T}_5$ generated with Algorithm 1 with $\theta = 1/2$. Notice a local higher refinement towards the reentrant corner.*

The triangulations $\mathcal{T}$ generated by Algorithm 1 are 1-irregular meshes. Error reduction and convergence of the adaptive finite element method based on the bulk criterion has been established in [10] for the Crouzeix–Raviart element.

5.2. Numerical example. On the L-shaped domain $\Omega = [0, 1]^2 \setminus [0.5, 1.0]^2$, we use the NR element defined in section 4.3 to approximate the Poisson problem (1.1) with $f \equiv 0$, $\Gamma_D = \partial\Omega$, $\Gamma_N = \emptyset$, and u_D a smooth function such that in polar coordinates

$$u(r, \theta) = r^{2/3} \sin\left(\frac{2}{3}\theta\right)$$

is the exact solution of (1.1). Figure 2 displays experimental convergence rates for the exact error and the estimate η_N for uniform and adaptive refinement with the corresponding triangulations depicted in Figure 3. The adaptive refinement improves the convergence rate of uniform refinement to the optimal one, $O(N^{-1/2})$, with respect to the number of degrees of freedom, and the convergence rate of the estimate mirrors that of the exact error for both uniform and adaptive refinement.

REFERENCES

[1] M. Ainsworth, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[2] M. Ainsworth, *A posteriori error estimation for nonconforming quadrilateral finite elements*, Int. J. Numer. Anal. Model., 2 (2005), pp. 1–18.

[3] D. N. Arnold, D. Boffi, and R. S. Falk, *Approximation by quadrilateral finite elements*, Math. Comp., 71 (2002), pp. 909–922.

[4] I. Babuska and A. Miller, *A feedback finite element method with a posteriori error estimation: Part* I. *The finite element method and some basic properties of the a posteriori error estimator*, Comput. Methods Appl. Mech. Engrg., 61 (1987), pp. 1–40.

[5] C. Bernardi and V. Girault, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.

[6] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer-Verlag, New York, 2002.

[7] C. Carstensen, *Quasi-interpolation and a posteriori error analysis in finite element methods*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 1187–1202.

[8] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part* I: *Low order conforming, nonconforming, and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.

[9] C. Carstensen, S. Bartels, and S. Jansche, *A posteriori error estimates for nonconforming finite element methods*, Numer. Math., 92 (2002), pp. 233–256.

[10] C. Carstensen and R. H. W. Hoppe, *Convergence analysis of an adaptive nonconforming finite element method*, Numer. Math., 103 (2006), pp. 251–266.

[11] C. Carstensen and J. Hu, *A Priori and A Posteriori Error Estimates for Nonconforming Finite Element Methods with Hanging Nodes*, Preprint, Humboldt-Universität zu Berlin, 2005.

[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978; reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[13] P. Clément, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.

[14] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Anal. Numér., 7 (1973), pp. 33–76.

[15] E. Dari, R. Duran, and C. Padra, *Error estimators for nonconforming finite element approximations of the Stokes problem*, Math. Comp., 64 (1995), pp. 1017–1033.

[16] E. Dari, R. Duran, C. Padra, and V. Vampa, *A posteriori error estimators for nonconforming finite element methods*, M2AN Math. Model. Numer. Anal., 30 (1996), pp. 385–400.

[17] J. Douglas Jr., J. E. Santos, D. Sheen, and X. Ye, *Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 747–770.

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

[19] H.-D. Han, *Nonconforming elements in the mixed finite element method*, J. Comput. Math., 2 (1984), pp. 223–233.

[20] J. Hu, *Quadrilateral Locking Free Elements in Elasticity*, Doctorate Dissertation, Institute of Computational Mathematics, Chinese Academy of Science, 2004 (in Chinese).

[21] J. Hu and Z.-C. Shi, *Constrained nonconforming quadrilateral rotated Q_1-element*, J. Comput. Math., 23 (2005), pp. 561–586.

[22] J. Hu and Z.-C. Shi, *Analysis of Nonconforming-Nonconforming Quadrilateral Rotated Q_1 Element for Reissner–Mindlin Plate*, Preprint 2003–10, Institute of Computational Mathematics, Chinese Academy of Science, 2003.

[23] G. Kanschat and F.-T. Suttmeier, *A posteriori error estimates for nonconforming finite element schemes*, Calcolo, 36 (1999), pp. 129–141.

[24] C. Park and D. Sheen, *P1-nonconforming quadrilateral finite element methods for second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 624–640.
[25] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.
[26] L. R. Scott and S. Shang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[27] Z.-C. Shi, *The F-E-M test for convergence of nonconforming finite elements*, Math. Comp., 49 (1987), pp. 391–405.
[28] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley–Teubner, Bath, UK, 1996.
[29] E. L. Wilson, R. L. Taylor, W. Doherty, and J. Ghaboussi, *Incompatible displacement models*, in Numerical and Computer Methods in Structural Mechanics, S. J. Fenves, N. Perrone, A. R. Robinson, and W. C. Schnobrich, eds., Academic Press, New York, 1973, pp. 43–57.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 83–107

FAST SWEEPING METHODS FOR EIKONAL EQUATIONS ON TRIANGULAR MESHES*

JIANLIANG QIAN†, YONG-TAO ZHANG‡, AND HONG-KAI ZHAO‡

Abstract. The original fast sweeping method, which is an efficient iterative method for stationary Hamilton–Jacobi equations, relies on natural ordering provided by a rectangular mesh. We propose novel ordering strategies so that the fast sweeping method can be extended efficiently and easily to any unstructured mesh. To that end we introduce multiple reference points and order all the nodes according to their l^p-metrics to those reference points. We show that these orderings satisfy the two most important properties underlying the fast sweeping method: (1) these orderings can cover all directions of information propagating efficiently; (2) any characteristic can be decomposed into a finite number of pieces and each piece can be covered by one of the orderings. We prove the convergence of the new algorithm. The computational complexity of the algorithm is nearly optimal in the sense that the total computational cost consists of $O(M)$ flops for iteration steps and $O(M\log M)$ flops for sorting at the predetermined initialization step which can be efficiently optimized by adopting a linear time sorting method, where M is the total number of mesh points. Extensive numerical examples demonstrate that the new algorithm converges in a finite number of iterations independent of mesh size.

Key words. eikonal equations, fast sweeping, Hamilton–Jacobi, viscosity solution

AMS subject classifications. Primary, 54C40, 14E20; Secondary, 46E25, 20C20

DOI. 10.1137/050627083

1. Introduction.

The eikonal equation in its simplest form says that the magnitude of the gradient of the eikonal is constant: $|\nabla T| = 1$, where T is the so-called eikonal. Because it appears in a variety of applications, it is essential to develop fast and efficient numerical methods to solve such an equation. In this work, we design a class of fast sweeping methods on triangulated domains for an eikonal equation of the following form:

$$\begin{cases} |\nabla T(\mathbf{x})| = f(\mathbf{x}), & \mathbf{x} \in \Omega \setminus \Gamma, \\ T(\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \Gamma \subset \Omega, \end{cases} \tag{1.1}$$

where $f(\mathbf{x})$ is a nonnegative function, Ω is an open, bounded polygonal domain in R^d, and Γ is a subset of Ω.

Two key points in designing an efficient numerical algorithm for solving such a nonlinear boundary value problem of hyperbolic type are (1) a numerical discretization that is both consistent with the causality of the PDE and able to deal with singularities in the solution gradient, and (2) a fast algorithm to solve the resulting large nonlinear system of equations. There are usually two types of methods for solving the nonlinear system: time marching methods and direct methods. Time marching methods add to the equation a pseudo–time variable which transforms the problem into

*Received by the editors March 17, 2005; accepted for publication (in revised form) July 6, 2006; published electronically January 12, 2007.

http://www.siam.org/journals/sinum/45-1/62708.html

†Department of Mathematics and Statistics, Wichita State University, Wichita, KS 67260-0033 (qian@math.wichita.edu, qian@math.ucla.edu). The research of this author was supported by NSF grant DMS-0542174.

‡Department of Mathematics, University of California, Irvine, CA 92697-3875 (zyt@math.uci.edu, zhao@math.uci.edu). The research of the third author was partially supported by ONR grant N00014-02-1-0090, DARPA grant N00014-02-1-0603, and the Sloan Foundation Fellowship.

a time dependent one and evolve the solution to the steady state. Due to the finite speed of propagation and the Courant–Friedrichs–Lewy (CFL) condition for stability, many iterations are needed to reach the steady state solution. The last two decades have witnessed much effort towards solving the eikonal equation directly: upwinding schemes [32, 31], dynamic programming sweeping methods [27], Jacobi iterations [26], semi-Lagrangian schemes [8], fast marching-type methods [30, 10, 28, 13], down-n-out approaches [7], wavefront expanding methods [23], adaptive upwinding methods [19, 21], fast sweeping methods [2, 37, 29, 35, 12, 34, 11, 36, 33]; see also the references therein. Accuracy of numerical solutions is determined by the discretization scheme. For example, if a first-order monotone scheme is used, in general only the $h^{1/2}$ convergence rate can be shown [6] and the $h \log h$ convergence rate is optimal for the eikonal equation [35].

Among all these methods, both the fast marching method and the fast sweeping method are designed to solve the nonlinear discretized system directly and efficiently by exploiting causality of the underlying PDE. In terms of complexity, the fast marching method [30, 10, 28, 13] has the complexity of $O(M \log M)$, where M is the total number of mesh points and the $\log M$ factor comes from the heapsort algorithm needed for sorting out the causality order at each step, while the fast sweeping method has the complexity of $O(M)$, where the constant in O depends on the equation, and this was proved in [35] for eikonal equations on rectangular grids. For a particular problem on a fixed grid, one method could be faster than the other. When the grid is more refined the fast sweeping method will be faster eventually. In [9], concrete and detailed comparisons are presented for various numerical examples. In terms of accuracy there is no difference since they are two different ways of solving the same nonlinear discretized equation. The main difference between these two methods lies in the use of causality. The fast marching method enforces the causality sequentially and on the fly during each update step; that is why a heapsort algorithm is needed to order all possible candidates and pick up the correct one by the causality at each step; once a point is accepted it cannot be revisited and its value cannot be changed afterwards. On the other hand the fast sweeping method is an iterative method of Gauss–Seidel type which is extremely simple to implement; such a simple iterative method for a nonlinear problem is able to achieve an optimal complexity because it can capture the causality of the PDE in a parallel way, as shown in [35]. Since it is an iterative method by nature the fast sweeping method is applicable to other situations such as higher order schemes with ease [34, 33], nonconvex Hamiltonians [12], and parallel implementation [36].

On the other hand, most of these methods are based on rectangular meshes. However, it is important to design fast methods on triangulated meshes as well. For example, in seismics a subsurface velocity model usually consists of several irregular interfaces, and in robotic path planning an obstacle may have an irregular boundary. Thus, for applications involving irregular boundaries or interfaces, it is much desired to triangulate a computational domain into irregular meshes to fit with boundaries or interfaces. Kimmel and Sethian [13] extended the fast marching method to triangulated domains to compute geodesics on manifolds.

In this work, we extend the fast sweeping method to triangulated domains by introducing novel ordering processes into the sweeping strategy. The resulting method is proved to be convergent, and numerical examples demonstrate that the method converges in a finite number of iterations independent of mesh size. The computational complexity of the new algorithm is nearly optimal in the sense that the total

computational cost consists of $O(M)$ flops for iteration steps and $O(M\log M)$ flops for sorting at the predetermined initialization step, which can be efficiently optimized by adopting a linear time sorting method.

An essential property of the eikonal equation is that it is hyperbolic, and a stable scheme must look for information by following characteristics in an upwind fashion, which is equivalent to the simple causality for the eikonal equation in that its solution is always increasing (or decreasing) along a characteristic. To satisfy such a property, it is crucial for a scheme of computing viscosity solutions to be based on a monotone numerical Hamiltonian [1, 17]. Once we have in place such a discretization for eikonal equations, the problem reduces to one of solving the resulting nonlinear system efficiently; the fast sweeping method is designed to do exactly that. The original fast sweeping method was inspired by the work in [2]. The fast sweeping method uses Gauss–Seidel iterations with alternate sweeping orderings to solve the nonlinear system. The fact that the iterative algorithm for a nonlinear system can converge in a finite number of iterations independent of mesh size is quite remarkable; even for a linear system, such as the discretized system for the Laplace equation, this is not true.

The crucial idea underlying the fast sweeping method is the following [35]: all directions of characteristics can be divided into a finite number of groups; any characteristic can be decomposed into a finite number of pieces that belong to one of the above groups; there are systematic orderings that can follow the causality of each group of directions simultaneously.

On a rectangular grid there are natural orderings of all grid points. For example, in the two-dimensional (2-D) case, all directions of the characteristics can be partitioned into four groups, up-right, up-left, down-right, and down-left, and it is very natural to order all the nodes according to their indexes in ascent or descent orders [2, 37, 29, 11, 35, 12, 34], which yields four possible orderings to cover all those four directions of characteristics.

However, on an unstructured mesh, only local connection of the nodes is available and natural ordering no longer exists. To overcome these difficulties we propose general ordering strategies by introducing multiple reference points and ordering all the nodes according to their l^p-distances to those reference points. For example, information is propagated as plane waves in different directions when the l^1-metric is used or as spherical waves with different centers when the l^2-metric is used. We show that these orderings satisfy the key properties essential for the fast sweeping method to converge and numerically demonstrate that the fast sweeping method converges in a finite number of iterations independent of mesh size. Although it may still cost $O(M\log M)$ by a comparison-based sorting method, the ordering step in our algorithm may be made to be $O(M)$ by a linear time sorting method since we know the distribution of nodes at the initial step. For example, the radix sorting method [4] may be used for such a purpose. Moreover this initial ordering is done for a fixed mesh once and for all. This is different from other methods based on heap sorting to maintain a dynamic data structure. Therefore the methods proposed here are very efficient and extremely easy to write in any number of dimensions.

The rest of the paper is organized as follows. In section 2, we construct local solvers at each node on a triangulated mesh, propose novel ordering strategies, and detail fast sweeping algorithms. In section 3, we analyze the new algorithm and prove convergence results. In section 4, we present various numerical examples to illustrate the efficiency and the accuracy of the new method. We conclude the paper in section 5.

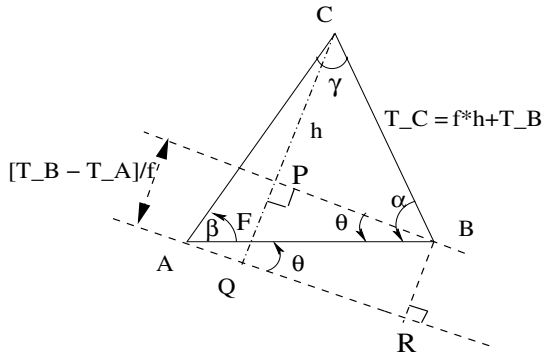


FIG. 2.1. *Update the value at C in a triangle when causality is satisfied.*

2. Fast sweeping methods on unstructured meshes.

2.1. 2-D local solvers. Take $d = 2$ in (1.1):

$$\begin{cases} \sqrt{T_x^2 + T_y^2} = f(x,y), & (x,y) \in \Omega \subset R^2, \\ T(x,y) = g(x,y), & (x,y) \in \Gamma \subset \Omega, \end{cases} \tag{2.1}$$

where $f(\mathbf{x})$ is a nonnegative function, Ω is an open, bounded polygonal domain in R^d, and Γ is a subset of Ω.

We consider a triangulation $\mathcal{T}_h$ of Ω which consists of nonoverlapping, nonempty, and closed triangles $\mathcal{T}$, with diameter $h_{\mathcal{T}}$, such that $\bar{\Omega} = \cup_{\mathcal{T} \in \mathcal{T}_h} \mathcal{T}$. We assume that $\mathcal{T}_h$ satisfies the following conditions:

- No more than μ triangles have a common vertex; $h = \sup_{\mathcal{T} \in \mathcal{T}_h} h_{\mathcal{T}} < 1$.
- $\mathcal{T}_h$ is regular; there exists a constant ω_0 independent of h such that if $\rho_{\mathcal{T}}$ is the diameter of the largest ball $B \subset \mathcal{T}$, then for all $\mathcal{T} \in \mathcal{T}_h$, $h_{\mathcal{T}} \le \omega_0 \rho_{\mathcal{T}}$.

For a given triangle $\triangle ABC$, we denote $\angle A = \beta$, $\angle B = \alpha$, and $\angle C = \gamma$; $\overline{AB} = c$, $\overline{AC} = b$, and $\overline{BC} = a$ are the lengths of the edges AB, AC, and BC, respectively.

During the solution process we need a local solver at vertex C for each triangle; see Figure 2.1. Given the values T_A and T_B at A and B of triangle $\triangle ABC$, we want to calculate the value T_C at C.

To make the description specific, we introduce the definition of causality.

DEFINITION 2.1. *Under the above regular triangulation we consider a local scheme based on piecewise linear reconstructions. By the causality condition of isotropic wave propagation for updating the travel-time at the node C from travel-times T_A and T_B, we mean that the ray which is orthogonal to the wavefront and passes through C must fall inside the triangle $\triangle ABC$.*

We notice that in isotropic wave propagation the ray direction is the same as the gradient direction of the travel-time field and thus it is the same as the outward normal of the wavefront.

First we assume that $\triangle ABC$ is acute. To construct a first-order scheme we determine a planar wavefront from the known values T_A and T_B. Suppose that the angle is θ between the incoming wavefront and the edge AB.

Without loss of generality, we further assume that $T_B \ge T_A$. If T_C is determined by both T_A and T_B, then by the Huygens principle the wavefront must first pass through the vertex A, then B, and finally C. To guarantee this, the following conditions must be satisfied:

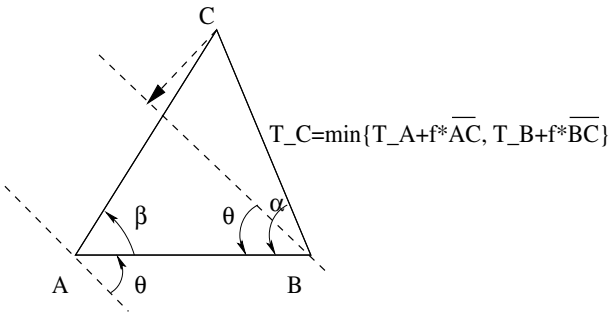


FIG. 2.2. *Update the value at vertex C in a triangle when causality is not satisfied.*

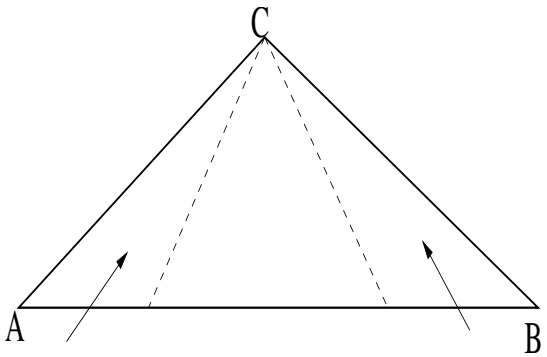


FIG. 2.3. *C and its obtuse triangle.*

- $[T_B - T_A]/f_C \le \overline{AB} = c$; i.e., it is possible for the wavefront to propagate from A to B with the given speed, where f_C is the value of $f(C)$, the inverse of the speed at C.
- $\theta \le \alpha$ so that the wavefront passes through B first rather than C.
- $\theta + \beta < \frac{\pi}{2}$; otherwise the causality is violated since the vertical line from C to the wavefront does *not* fall inside the triangle; see Figure 2.2.

If all n triangles $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_n$ around C are acute, the wavefront can be captured well in one of these triangles, no matter which direction the wave comes from. However, if one of the triangles is obtuse and the wavefront comes in just from this obtuse angle, then the situation is different; there are two possible cases: (i) if the normal of the wavefront is contained between those two dotted lines in Figure 2.3, then the value at C can be updated using values at A and B even though the accuracy will be degraded; (ii) otherwise, the value at C cannot be updated by A and B correctly [25]. These will be shown in numerical examples in section 4.

In order to treat obtuse triangles, we adopt the strategy used in [25]. As illustrated in Figure 2.4, if $\angle C$ is obtuse, then we connect C to a vertex D of a neighboring triangle to cut the obtuse angle into two smaller angles. If these two angles are both acute, then we are done, as shown in Figure 2.4(a); otherwise, if one of the smaller angles is still obtuse, then we keep connecting C to the vertexes of the neighboring triangles of the next level until all new angles at C are acute, as shown in Figure 2.4(b). All these added edges are "virtual"; i.e., they exist only when the value at C is updated. Because such a treatment depends on a given mesh, we only need to do that once before the iteration in the algorithm begins; the resulting algorithm is

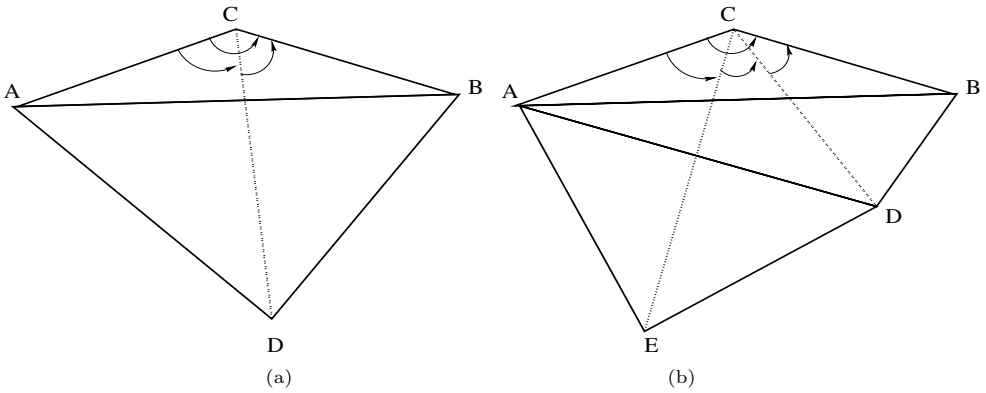


FIG. 2.4. *A strategy to treat obtuse angles.*

simple with almost no extra computational cost, as shown by numerical examples in section 4. This construction is different from the one used in [13].

We first give a geometric version of our local solvers.

A 2-D LOCAL SOLVER (Version 1: given $T_A \le T_B$, determine $T_C = T_C(T_A, T_B)$).

1. If $[T_B - T_A] \le c\, f_C$, then

$$\theta = \arcsin\left(\frac{[T_B - T_A]}{c\, f_C}\right);$$

   (a) if $\max(0, \alpha - \frac{\pi}{2}) \le \theta \le \frac{\pi}{2} - \beta$, then

$$h = \overline{CP} = a\sin(\alpha - \theta); T_C = \min\{T_C, h\, f_C + T_B\};$$

   (b) else

$$T_C = \min\{T_C, T_A + b\, f_C, T_B + a\, f_C\};$$

2. else

$$T_C = \min\{T_C, T_A + b\, f_C, T_B + a\, f_C\}.$$

The angle condition,

$$\max\left(0, \alpha - \frac{\pi}{2}\right) \le \theta \le \frac{\pi}{2} - \beta,$$

can be obtained in the following way:

1. If $\beta > \frac{\pi}{2}$, then the causality condition is not valid.
2. If $\beta < \frac{\pi}{2}$, then we must have $\theta \le \frac{\pi}{2} - \beta$; otherwise, the causality is violated since the vertical line from C to the wavefront does *not* fall inside the triangle. Furthermore,
   (a) from this condition we can directly deduce that $\alpha \ge \theta$, since $\angle C = \gamma < \frac{\pi}{2}$ by construction;
   (b) if $\alpha \ge \frac{\pi}{2}$, then we must have $\alpha - \theta \le \frac{\pi}{2}$ so that the ray from C reaching the wavefront is located inside the triangle.

The following algorithm unifies all the cases into one.

A 2-D LOCAL SOLVER (Version 2: given T_A and T_B, determine $T_C = T_C(T_A, T_B)$).

1. If $|T_B - T_A| \le c\, f_C$, then

$$\theta = \arcsin\left(\frac{[T_B - T_A]}{c\, f_C}\right);$$

(a) if $\max(0, \alpha - \frac{\pi}{2}) \le \theta \le \frac{\pi}{2} - \beta$ or $\alpha - \frac{\pi}{2} \le \theta \le \min(0, \frac{\pi}{2} - \beta)$, then

$$h = \overline{CP} = a\, \sin(\alpha - \theta); H = \overline{CQ} = b\, \sin(\beta + \theta);$$

$$T_C = \min\{T_C, 0.5\, (h\, f_C + T_B) + 0.5\, (H\, f_C + T_A)\};$$

(b) else

$$T_C = \min\{T_C, T_A + b\, f_C, T_B + a\, f_C\};$$

2. else

$$T_C = \min\{T_C, T_A + b\, f_C, T_B + a\, f_C\}.$$

In the special case that a given mesh is rectangular and $\alpha = \beta = \frac{\pi}{4}$, it is straightforward to verify that the above local solver reduces to the one given in [35]. Therefore, the local solver is consistent with the one on rectangular meshes.

If a triangle is acute, then the angle conditions in Version 2 reduce to one condition:

$$\alpha - \frac{\pi}{2} \le \theta \le \frac{\pi}{2} - \beta;$$

otherwise, the two angle conditions cannot be combined into one, since there are gaps corresponding to one of the angles α or β being obtuse. See Figures 2.1 and 2.2 for illustrations.

We emphasize that both updating algorithms require that $\angle C = \gamma < \frac{\pi}{2}$, but one of the other two angles may be obtuse.

2.2. A 3-D local solver. A local solver in three dimensions can be derived similarly. Take $d = 3$ in (1.1):

$$\begin{cases} \sqrt{T_x^2 + T_y^2 + T_z^2} = f(x, y, z), & (x, y, z) \in \Omega \subset R^3, \\ T(x, y, z) = g(x, y, z), & (x, y, z) \in \Gamma \subset \Omega. \end{cases} \tag{2.2}$$

Equation (2.2) is solved in the domain Ω, which has a triangulation $\mathcal{T}_h$ consisting of tetrahedrons. We consider every vertex and all tetrahedrons which are associated to this vertex. Again the question reduces to one of calculating the numerical solution at the current central vertex for each tetrahedron; see Figure 2.5.

Given the values T_A, T_B, and T_C at A, B, and C of the tetrahedron $ABCD$, we need to calculate the value T_D at the current central vertex D. The key is to determine the normal direction $\vec{\mathbf{n}}$ of the wavefront and determine whether the causality condition is satisfied or not. Analogous to Definition 2.1, the ray which has direction $\vec{\mathbf{n}}$ and passes through D must fall inside the tetrahedron $ABCD$ so as to satisfy the causality condition. To check the causality condition numerically, we first compute the coordinates of the point E at which the ray passing through D with direction $\vec{\mathbf{n}}$ intersects the plane spanned by A, B, and C; afterwards, we check to see whether E is inside $\triangle ABC$ or not.

Without loss of generality, we assume that $T_A = \min\{T_A, T_B, T_C\}$.

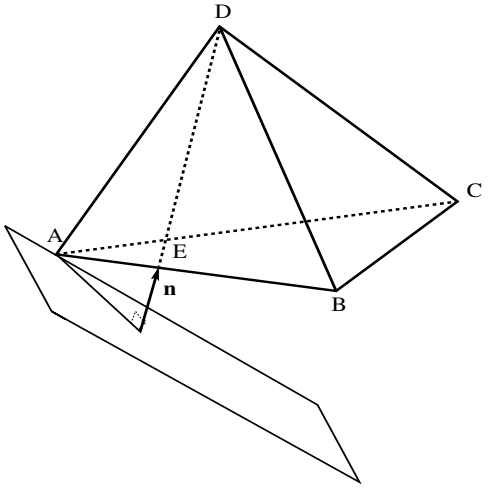


FIG. 2.5. *A 3-D local solver.*

A 3-D LOCAL SOLVER (given T_A, T_B, and T_C, determine $T_D = T_D(T_A, T_B, T_C)$).

1. If $[T_B - T_A] \le \overline{AB} \cdot f_D$ and $[T_C - T_A] \le \overline{AC} \cdot f_D$, then we solve the quadratic equation for the normal direction $\vec{\mathbf{n}}$ of the wavefront:

$$\begin{cases} \overrightarrow{AB} \cdot \vec{\mathbf{n}} = [T_B - T_A]/f_D, \\ \overrightarrow{AC} \cdot \vec{\mathbf{n}} = [T_C - T_A]/f_D, \\ \quad |\vec{\mathbf{n}}| = 1; \end{cases} \tag{2.3}$$

   (a) if there exist solutions $\vec{\mathbf{n}}^{(i)}, i = 1, 2$, for the quadratic equation (2.3) and the area $|\triangle EAB| + |\triangle EAC| + |\triangle EBC| = |\triangle ABC|$ for an $\vec{\mathbf{n}}^{(i)}$, then

$$T_D = \min\{T_D, T_A + (|\overrightarrow{AD} \cdot \vec{\mathbf{n}}^{(i)}|) \cdot f_D\};$$

   (b) else, apply the 2-D local solver on surfaces $\triangle ABD, \triangle ACD$, and $\triangle BCD$ and take the minimal one;

2. else, apply the 2-D local solver on surfaces $\triangle ABD, \triangle ACD$, and $\triangle BCD$ and take the minimal one.

2.3. Sweeping orders and a complete algorithm. An essential ingredient for making the fast sweeping method [35] successful is a systematic ordering that covers all directions of characteristics efficiently. With a causality preserving discretization in place, information along characteristics of certain directions is captured simultaneously in each sweeping ordering. Moreover, once the solution at a node gets its correct value, i.e., the smallest possible value, it will not change in later iterations. There are natural orderings on rectangular meshes. For example, in 2-D cases [35], all directions can be divided into four groups, up-right, up-left, down-left, and down-right, which can be covered by the orderings $i = 1 : I, j = 1 : J$; $i = 1 : I, j = J : 1$; $i = I : 1, j = 1 : J$; $i = I : 1, j = J : 1$, respectively, where i and j are the running indexes in the x- and y-directions, respectively. However, such natural orderings no longer exist on an unstructured mesh.

To devise efficient fast sweeping methods on unstructured meshes, we propose systematic orderings by introducing multiple reference points and sorting all the nodes according to their l^p-distances to each individual reference point. In this paper we

focus on $p = 1$ and 2 and give explicit geometric interpretation. The argument works for all other p's.

The l^p-metric for a vector $\mathbf{x} = (x_1, x_2, \ldots, x_n) \in R^n$ is defined as $|\mathbf{x}|_p = (\sum_{j=1}^{n} |x_j|^p)^{1/p}$. Without abuse of notation we also use $|\mathbf{x}|$ to denote the 2-norm of a vector $\mathbf{x}$. For example, in two dimensions, we first fix a reference point $\mathbf{x}_{\text{ref}}$; if we sweep through all nodes according to $|\mathbf{x} - \mathbf{x}_{\text{ref}}|_1$ in the ascent (or descent) order, then the sweeping wavefront is an outgoing (or incoming) plane wave since the unit ball of the l^1-metric is a tilted square. If we use $|\mathbf{x} - \mathbf{x}_{\text{ref}}|_2$ to order all nodes, then the sweeping wavefront is an outgoing (or incoming) spherical wave.

Next we address the following questions:

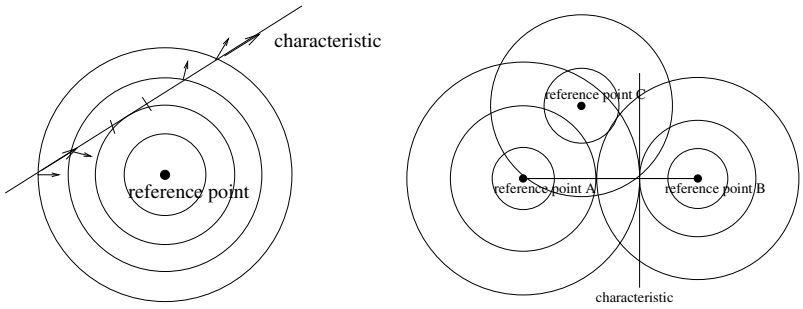
1. How many references points are needed in a systematic ordering that can cover all directions of information propagating?
2. How many iterations are needed for the algorithm to converge?

To address the first question, we have to understand the directional relation between a sweeping wavefront and a characteristic. In the continuous case the following is a basic fact: if the propagating direction of the sweeping wavefront forms an acute angle with the direction of the characteristic, then the causality along this characteristic can be captured in this ordering. As illustrated in Figure 2.6, if we use the l^2-metric, i.e., with a spherical sweeping wavefront, a straight characteristic in any direction can be partitioned into two pieces by the tangent point to a particular spherical sweeping wavefront, and each piece forms an acute angle to the outgoing or incoming sweeping wavefront. If all characteristics are straight lines, which is the case when the right-hand side of the eikonal equation is constant, we cover almost all characteristics by sweeping all nodes according to the l^2-distance to a single reference point in both ascent and descent orders alternately. However, for all characteristics at the tangent point, the normal of the sweeping wavefront is orthogonal to the direction of characteristics. So information will not propagate across the tangent point from one piece to other pieces effectively. To remedy this problem we introduce another reference point. Now all directions of characteristics can be covered effectively by the four orderings except one direction, which is orthogonal to the line connecting these two reference points, as shown in Figure 2.6. Therefore we need at least three noncollinear reference points and we sweep through all the nodes according to their l^2-distances to these reference points in ascent and descent orderings; a total of six orderings cover all directions of information propagating along characteristics. It can be easily seen that four noncoplanar reference points are needed in three dimensions.

If we use the l^1-metric, the sweeping wavefront is a tilted square. For each reference point, as shown in Figure 2.7, the whole plane can be divided into four quadrants, and each quadrant can be covered by one planar sweeping wavefront. If we choose two reference points such that the computational domain lies in different quadrants of these two reference points, all directions of characteristics can be covered by the four orderings corresponding to the ascent and descent sorting according to the l^1-metric; see Figure 2.7.

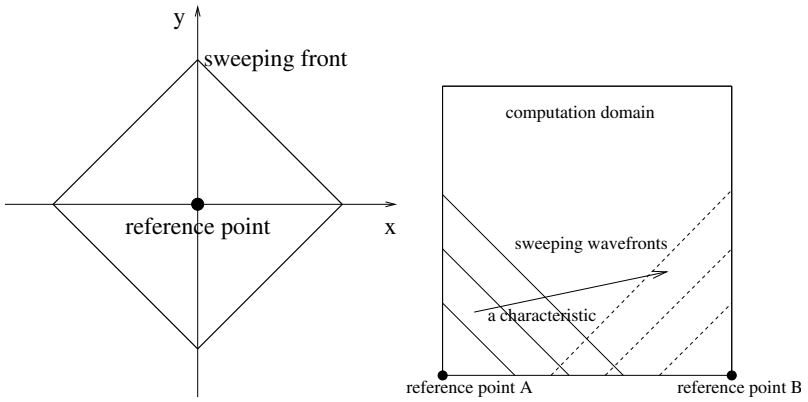
When characteristics are not straight lines, any characteristic can be divided into a finite number of pieces so that each piece can be covered effectively by one of the orderings, as shown in [35]. The total number of sweepings is increased due to curved characteristics, but it is still finite. The number of iterations will be estimated in section 3.

In terms of numerical implementation on a particular mesh some remarks are in order.



(a) one reference point (b) three reference points

FIG. 2.6. *Reference points and sweeping wavefronts for the l^2-metric.*



(a) one reference point (b) two reference points

FIG. 2.7. *Reference points and sweeping wavefronts for the l^1-metric.*

The domain of dependence for a node in the discrete case is a region rather than only the characteristic that passes through the node in the continuous case. On a triangular mesh, the propagating direction of a sweeping wavefront has to fall into the triangle which satisfies the causality criterion in Definition 2.1 so that the two neighbors that determine the current vertex have already been updated in the current sweeping. Numerically this means that the normal of the sweeping wavefront has to make an acute angle with the characteristic that passes through this vertex.

The criterion for an optimal choice of reference points and their locations on a triangular mesh is that all directions of characteristics should be covered with minimal redundancy. In practice, it is better if these reference points are evenly spaced both spatially and angularly with respect to the data set or boundary where the solution is prescribed. In our numerical tests we use the corners as reference points if the computational domain is rectangular. Other points, such as the center point of the domain or middle points of each edge, can be used as well. The number of iterations needed for convergence may be different for different choices of reference points but it will be finite.

If we have only a point source as the boundary condition on a rectangular mesh and we use that point as the single reference point, then the square wave sweeping accesses nodes in the ascent order in the same way that the down-n-out model does

[32, 7], and the spherical wave sweeping shares some similarities with the expanding wavefront model proposed in [31, 23]. However, we are not aware of any work accessing the nodes in the way similar to the plane wave sweeping proposed here.

The above isotropic metrics are suitable for ordering nodes in solving isotropic eikonal equations. For general anisotropic eikonal equations considered in [24, 18, 20], we may introduce anisotropic Riemannian metrics [5] to sort all the nodes, derive a local solver to update solutions at each node by using phase velocity and group velocity, as illustrated in [24, 18], and design fast sweeping methods accordingly; see [22] for a recent work along this direction.

Now we summarize local solvers and sweeping orderings into a complete algorithm.

THE FAST SWEEPING ALGORITHM ON A TRIANGULAR MESH.

1. Initialization:
   (a) Triangulate the computational domain Ω. Add virtual edges to cut obtuse angles if there are any.
   (b) Choose multiple reference points: $\mathbf{x}_{\text{ref}}^i$, $i = 1, \ldots, R$.
   (c) Sort all nodes according to their l^p-distances to the reference points in ascent and descent orders, and put them into arrays:

$$\begin{aligned} &S_i^+ : \text{ascent order}, i = 1, 2, \ldots, R; \\ &S_i^- : \text{descent order}, i = 1, 2, \ldots, R. \end{aligned} \tag{2.4}$$

   (d) Assign exact values or interpolated values $T^{(0)}$ at vertexes on or near the given boundary Γ, and keep these values fixed during the iterations. At all other vertexes, assign large positive values N to $T^{(0)}$, where N should be larger than the maximum of the true solution, and these values will be updated in later iterations.
2. Gauss–Seidel iteration for $k = 0, 1, \ldots$:
   (a) For $i = 1, \ldots, R$:
      i. For $j = +, -$:
         A. To every vertex $C \in S_i^j$ and every triangle associated with C, f_C=$f(C)$, apply the local solver;
         B. Convergence test: $\|T^{(k+1)} - T^{(k)}\| \le \epsilon$ for $\epsilon > 0$ given, where $\|\cdot\|$ is some specified norm.

We remark that during the Gauss–Seidel iteration the numerical solution at C is calculated using the current values of its neighbors in every triangle. The smallest one will be taken as the possible new value. If this smallest new value is smaller than the current value at C, then the numerical solution at C is updated to be the smallest new value.

In passing we point out that the sorting procedure in the above algorithm can cost $O(M\log M)$ flops if a comparison-based sorting method is used; however, to achieve an optimal $O(M)$ complexity for the algorithm, we may use a radix sorting method [4] in that we know the distribution of nodes. Radix sorting runs an $O(M)$ counting sort on each digit of the key, starting with the least significant and working for bounded integers. For general distances computed in the above algorithm, we argue that a fixed number of digits is sufficient because in some sense the order of updates does not matter too much for two nodes sufficiently close to each other. Moreover, this initial ordering is done for a fixed mesh once and for all.

3. Convergence results. In this section we prove convergence of the fast sweeping algorithm on triangular meshes. In the following analysis, we consider a regular

triangulation $\mathcal{T}_h$ of Ω with the property that all the inner angles of the triangles in $\mathcal{T}_h$ satisfy $\leq \frac{\pi}{2}$.

Considering a triangle $\triangle ABC$ in which T_A and T_B are given, we update the travel-time T_C at the vertex C. Denoting

$$p_1 = \frac{T_C - T_A}{b}, \;\; p_2 = \frac{T_C - T_B}{a}, \;\; p_3 = \frac{T_B - T_A}{c},$$

we adopt the framework in [3] to show consistency and monotonicity of the Godunov numerical Hamiltonian resulting from the local solver introduced in section 2.

LEMMA 3.1 (Godunov numerical Hamiltonian). *Assuming that the causality condition holds, the updating formula for the local solver is one of the solutions for the following equations:*

$$\text{(3.1)} \qquad \begin{cases} \frac{(T_C-T_A)^2}{b^2} - 2\frac{(T_C-T_A)(T_C-T_B)}{a\,b}\cos\gamma + \frac{(T_C-T_B)^2}{a^2} = f_C^2\,\sin^2\gamma \\ \qquad\qquad \textit{if } |p_3| \leq f_C \textit{ and } \alpha - \frac{\pi}{2} \leq \arcsin(\frac{p_3}{f_C}) \leq \frac{\pi}{2} - \beta; \\ \max\left(\frac{T_C-T_A}{b}, \frac{T_C-T_B}{a}\right) = f_C \quad \textit{otherwise.} \end{cases}$$

Here $\angle C = \gamma$, $\angle A = \beta$, $\angle B = \alpha$, and $f_C = f(C)$. This discretization for the eikonal equation is based on the Godunov numerical Hamiltonian:

$$\text{(3.2)} \qquad \hat{H}_C\left(\frac{T_C - T_A}{b}, \frac{T_C - T_B}{a}\right) = f_C,$$

where

$$\text{(3.3)} \quad \hat{H}_C(p_1, p_2) = \begin{cases} \frac{1}{\sin\gamma}\sqrt{p_1^2 - 2p_1\,p_2\cos\gamma + p_2^2} \\ \qquad\qquad \textit{if } |p_3| \leq f_C \textit{ and } \alpha - \frac{\pi}{2} \leq \arcsin(\frac{p_3}{f_C}) \leq \frac{\pi}{2} - \beta; \\ \max(p_1, p_2) \quad \textit{otherwise.} \end{cases}$$

Proof. By Version 2 of the local solver, we have

$$\text{(3.4)} \quad T_C = \begin{cases} \frac{1}{2}(T_A + T_B) + \frac{\sin(\alpha-\beta)}{2\sin\gamma}(T_B - T_A) + \frac{\sin\alpha\,\sin\beta}{\sin\gamma}\sqrt{c^2 f_C^2 - (T_B - T_A)^2} \\ \qquad \text{if } |p_3| \leq f_C \text{ and } \alpha - \frac{\pi}{2} \leq \arcsin(\frac{p_3}{f_C}) \leq \frac{\pi}{2} - \beta; \\ \min(T_A + bf_C, T_B + af_C) \quad \text{otherwise.} \end{cases}$$

By solving (3.1), we have

$$\text{(3.5)} \quad T_C = \begin{cases} \frac{1}{2}(T_A + T_B) + \frac{b^2-a^2}{2c^2}(T_B - T_A) \pm \frac{a\,b\sin\gamma}{c^2}\sqrt{c^2 f_C^2 - (T_B - T_A)^2} \\ \qquad \text{if } |p_3| \leq f_C \text{ and } \alpha - \frac{\pi}{2} \leq \arcsin(\frac{p_3}{f_C}) \leq \frac{\pi}{2} - \beta; \\ \min(T_A + bf_C, T_B + af_C) \quad \text{otherwise;} \end{cases}$$

one of the two roots corresponds to (3.4).

Next we derive the numerical Hamiltonian. Denote $A : (x_A, y_A)$, $B : (x_B, y_B)$, and $C : (x_C, y_C)$. Since the causality condition holds, we have

$$\text{(3.6)} \qquad \frac{T_C - T_A}{b} = \nabla T(C) \cdot \left(\frac{x_C - x_A}{b}, \frac{y_C - y_A}{b}\right)^t + o(h^2),$$

$$\text{(3.7)} \qquad \frac{T_C - T_B}{a} = \nabla T(C) \cdot \left(\frac{x_C - x_B}{a}, \frac{y_C - y_B}{a}\right)^t + o(h^2),$$

where t denotes the transpose of vectors. Furthermore we have

$$\begin{pmatrix} \frac{T_C - T_A}{b} \\ \frac{T_C - T_B}{a} \end{pmatrix} = \mathbf{P}\nabla T(C) + o(h^2), \tag{3.8}$$

where

$$\mathbf{P} = \begin{pmatrix} \frac{x_C - x_A}{b} & \frac{y_C - y_A}{b} \\ \frac{x_C - x_B}{a} & \frac{y_C - y_B}{a} \end{pmatrix}.$$

Ignoring higher-order terms and solving for ∇T_C, we have

$$|\nabla T(C)| \approx \begin{cases} \frac{1}{\sin\gamma}\sqrt{\frac{(T_C - T_A)^2}{b^2} - 2\frac{(T_C - T_A)(T_C - T_B)}{a\,b}\cos\gamma + \frac{(T_C - T_B)^2}{a^2}} \\ \qquad \text{if } |p_3| \le f_C \text{ and } \alpha - \frac{\pi}{2} \le \arcsin(\frac{p_3}{f_C}) \le \frac{\pi}{2} - \beta; \\ \max\left(\frac{T_C - T_A}{b}, \frac{T_C - T_B}{a}\right) \quad \text{otherwise;} \end{cases} \tag{3.9}$$

this is the Godunov numerical Hamiltonian for the eikonal equation. ∎

LEMMA 3.2 (consistency and causality). *The Godunov numerical Hamiltonian*

$$\hat{H}_C(p_1, p_2) = \begin{cases} \frac{1}{\sin\gamma}\sqrt{p_1^2 - 2p_1\,p_2\cos\gamma + p_2^2} \\ \qquad \textit{if } |p_3| \le f_C \textit{ and } \alpha - \frac{\pi}{2} \le \arcsin(\frac{p_3}{f_C}) \le \frac{\pi}{2} - \beta; \\ \max(p_1, p_2) \quad \textit{otherwise} \end{cases} \tag{3.10}$$

is consistent; namely,

$$\hat{H}_C\left(\frac{T_C - T_A}{b}, \frac{T_C - T_B}{a}\right) = |\mathbf{p}| \tag{3.11}$$

if $\nabla T_h = \mathbf{p} \in \mathcal{R}^2$. *It is monotone if the causality condition holds:* $0 \le \gamma_1 \le \gamma$, *where* γ_1 *is the angle from the edge CA to the ray (i.e., the vertical line to the wavefront) CQ counterclockwise; see Figure* 2.1.

Proof. By $\nabla T_h = \mathbf{p} \in \mathcal{R}^2$, we have

$$\begin{pmatrix} \frac{T_C - T_A}{b} \\ \frac{T_C - T_B}{a} \end{pmatrix} = \mathbf{P}\mathbf{p}. \tag{3.12}$$

Inserting this into the numerical Hamiltonian, we have (3.11).

Differentiating $\hat{H}_C(p_1, p_2)$ with respect to p_1 and p_2, the monotonicity of the Hamiltonian requires

$$\frac{\partial \hat{H}_C}{\partial p_1} \ge 0, \qquad \frac{\partial \hat{H}_C}{\partial p_2} \ge 0; \tag{3.13}$$

these can be satisfied if and only if $\cos\gamma \le \frac{p_2}{p_1} \le \frac{1}{\cos\gamma}$. By

$$p_1 = \frac{T_C - T_A}{b} = f_C \sin(\beta + \theta), \tag{3.14}$$

$$p_2 = \frac{T_C - T_B}{a} = f_C \sin(\alpha - \theta), \tag{3.15}$$

where $\theta = \arcsin(\frac{p_3}{f_C})$, we have

$$\cos\gamma \le \frac{\sin(\beta + \theta)}{\sin(\alpha - \theta)} \le \frac{1}{\cos\gamma}, \tag{3.16}$$

which is equivalent to the causality condition $0 \le \gamma_1 \le \gamma$, since $\gamma_1 = \frac{\pi}{2} - (\beta + \theta)$ and $\gamma_1 = (\gamma + \alpha - \theta) - \frac{\pi}{2}$. ∎

LEMMA 3.3 (monotonicity). *The fast sweeping algorithm is monotone and Lipschitz continuous, i.e.,*

$$1 \ge \frac{\partial T_C}{\partial T_B} \ge 0, \quad 1 \ge \frac{\partial T_C}{\partial T_A} \ge 0, \tag{3.17}$$

and

$$\frac{\partial T_C}{\partial T_B} + \frac{\partial T_C}{\partial T_A} = 1. \tag{3.18}$$

Proof. Consider the case that $T_A \le T_B$. We need only verify that the above inequalities hold when T_C is updated by

$$T_C = h\, f_C + T_B, \tag{3.19}$$

which is the case that the causality condition is satisfied. From Version 1 of the local solver we have

$$\frac{\partial T_C}{\partial T_B} = 1 + a f_C \cos(\alpha - \theta) \left(-\frac{\partial \theta}{\partial T_B} \right) \tag{3.20}$$

$$= 1 - \frac{a \cos(\alpha - \theta)}{c \cos\theta}; \tag{3.21}$$

$$\frac{\partial T_C}{\partial T_A} = a f_C \cos(\alpha - \theta) \left(-\frac{\partial \theta}{\partial T_A} \right) \tag{3.22}$$

$$= \frac{a \cos(\alpha - \theta)}{c \cos\theta}. \tag{3.23}$$

From Figure 2.1, we have $a\cos(\alpha - \theta) = \overline{PB}$, $c\cos(\theta) = \overline{AR}$, and $\overline{PB} \le \overline{AR}$; therefore, $1 \ge \frac{\partial T_C}{\partial T_B} \ge 0$, $1 \ge \frac{\partial T_C}{\partial T_A} \ge 0$, and $\frac{\partial T_C}{\partial T_B} + \frac{\partial T_C}{\partial T_A} = 1$. ∎

LEMMA 3.4 (maximum change principle). *In the Gauss–Seidel iteration for the fast sweeping algorithm, the maximum change of T_h at any vertex is less than or equal to the maximum change of T_h at its neighboring points.*

Proof. This follows from the above monotonicity property proved in Lemma 3.3. ∎

LEMMA 3.5 (order preserving). *The fast sweeping algorithm is monotone in the initial data.*

Proof. By the monotonicity property of the solution, if $T_h(C) \le R_h(C)$ at all vertexes initially, then $T_h(C) \le R_h(C)$ at all vertexes after any number of Gauss–Seidel iterations. ∎

LEMMA 3.6 (nonincreasing). *The solution of the fast sweeping algorithm is nonincreasing with each Gauss–Seidel iteration.*

Proof. This is evident from the updating formula, which updates the current value only if it is larger than the newly computed value during the Gauss–Seidel iteration. ∎

LEMMA 3.7 (l^∞-contraction). *Let $T^{(k)}$ and $R^{(k)}$ be two numerical solutions at the kth iteration of the fast sweeping algorithm. Let $\|\cdot\|_\infty$ be the maximum norm. Then*

$$\|T^{(k)} - R^{(k)}\|_\infty \le \|T^{(k-1)} - R^{(k-1)}\|_\infty; \tag{3.24}$$

$$0 \le \max_C \left\{ T_C^{(k)} - T_C^{(k+1)} \right\} \le \max_C \left\{ T_C^{(k-1)} - T_C^{(k)} \right\}. \tag{3.25}$$

Proof. Assume that the first update at the kth iteration is at C,

$$T_C^{(k)} = \min\{T_C^{(k-1)}, \bar{T}\},$$

where $\bar{T}$ is the solution computed from its neighbors $T_A^{(k-1)}$ and $T_B^{(k-1)}$. The same is true for $R_C^{(k)}$. By the maximum change principle, we have

$$|T_C^{(k)} - R_C^{(k)}| \le \|T^{(k-1)} - R^{(k-1)}\|_\infty. \tag{3.26}$$

For an update at any other node later in the iteration, the neighboring values used for the update are either from the previous iteration or from an earlier update in the current iteration, both of which satisfy the above bound. By induction, we have l^∞-contraction (3.24). By the monotonicity of the fast sweeping algorithm and (3.24), setting $R^{(k)} = T^{(k-1)}$ we conclude (3.25). ∎

THEOREM 3.8 (convergence). *The solution of the fast sweeping algorithm converges monotonically to the solution of the discretized system.*

Proof. Denote the numerical solution after the kth iteration by $T_C^{(k)}$. Since $T_C^{(k)}$ is bounded below by 0 and is nonincreasing with Gauss–Seidel iterations, $T_C^{(k)}$ is convergent for all C. After each sweep for each C at each triangle, we have by the monotonicity of the numerical Hamiltonian

$$\frac{(T_C^{(k)} - T_A^{(k)})^2}{b^2 \sin^2\gamma} - 2\frac{(T_C^{(k)} - T_A^{(k)})(T_C^{(k)} - T_B^{(k)})}{a\, b \sin^2\gamma}\cos\gamma + \frac{(T_C^{(k)} - T_B^{(k)})^2}{a^2 \sin^2\gamma} \ge f_C^2 \tag{3.27}$$

because any later update of neighbors of $T_C^{(k)}$ in the same iteration is nonincreasing. Moreover, it is easy to see that after $T_C^{(k)}$ is updated, the function

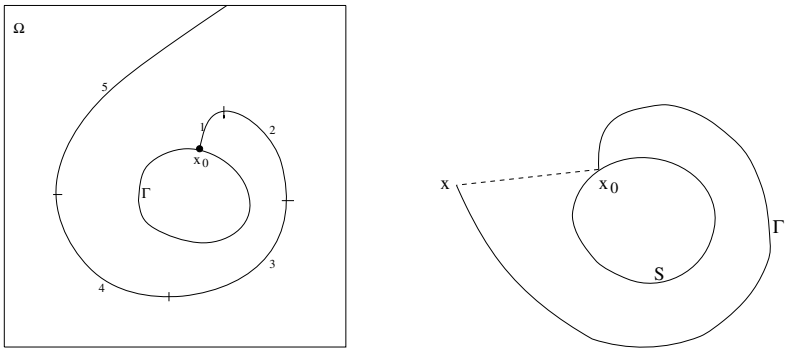
$$\begin{aligned} F(T_A^{(k)}, T_B^{(k)}) = {} & \frac{(T_C^{(k)} - T_A^{(k)})^2}{b^2 \sin^2\gamma} - 2\frac{(T_C^{(k)} - T_A^{(k)})(T_C^{(k)} - T_B^{(k)})}{a\, b \sin^2\gamma}\cos\gamma \\ & + \frac{(T_C^{(k)} - T_B^{(k)})^2}{a^2 \sin^2\gamma} - f_C^2 \end{aligned} \tag{3.28}$$

is Lipschitz continuous in $T_A^{(k)}$ and $T_B^{(k)}$, and the Lipschitz constant is bounded by

$$2\max\left\{ \frac{|T_C^{(k)} - T_A^{(k)}|}{b^2 \sin^2\gamma} + \frac{|T_C^{(k)} - T_B^{(k)}|}{a\, b \sin^2\gamma}\cos\gamma, \frac{|T_C^{(k)} - T_B^{(k)}|}{a^2 \sin^2\gamma} + \frac{|T_C^{(k)} - T_A^{(k)}|}{a\, b \sin^2\gamma}\cos\gamma \right\}.$$

Since $T_C^{(k)}$ is monotonically convergent for all C, we can have an upper bound $Z > 0$ for the Lipschitz constant. Let $\delta^{(k)} = \max\{T_C^{(k-1)} - T_C^{(k)}\}$ be the maximum change at all grid points during the kth iteration. By the l^∞-contraction property and the convergence property of $T_C^{(k)}$, $\delta^{(k)}$ converges monotonically to zero. After the kth iteration, we have

$$\begin{aligned} 0 \le {} & \frac{(T_C^{(k)} - T_A^{(k)})^2}{b^2 \sin^2\gamma} - 2\frac{(T_C^{(k)} - T_A^{(k)})(T_C^{(k)} - T_B^{(k)})}{a\, b \sin^2\gamma}\cos\gamma + \frac{(T_C^{(k)} - T_B^{(k)})^2}{a^2 \sin^2\gamma} - f_C^2 \\ & \le Z\delta^{(k)}. \end{aligned} \tag{3.29}$$



(a) (b)

FIG. 3.1. *Partitioning of a characteristic.*

Thus $T^{(k)}$ converges to the solution to (3.1). □

Note that the monotone convergence is very important during iterations. Once the solution at a node reaches the minimal value that it can get, it is the correct value at that node and will not change in later iterations.

Next we show the estimate for the total number of iterations needed for convergence. As pointed out above, given a systematic ordering, any characteristic can be partitioned into a finite number of pieces and each piece will be covered correctly by one of the sweeping orderings, as shown in Figure 3.1(a). Since these pieces have to be captured sequentially the total number of iterations needed is proportional to the number of pieces. Finally the number of pieces needed to partition a characteristic is related to the directional change of the characteristic. We now give an estimate on the total variation of the tangent direction of any characteristic in a fixed domain Ω.

Denote $H(\mathbf{p}, \mathbf{x}) = |\mathbf{p}| - f(\mathbf{x})$, where $\mathbf{p} = \nabla T$. The characteristic equation for the eikonal equation is

$$\begin{cases} \dot{\mathbf{x}} = \nabla_{\mathbf{p}} H = \frac{\nabla T}{f(\mathbf{x})}, \\ \dot{\mathbf{p}} = -\nabla_{\mathbf{x}} H = \nabla f(\mathbf{x}), \\ \dot{T} = \nabla T \cdot \dot{\mathbf{x}} = f(\mathbf{x}), \end{cases}$$

where $\dot{}$ denotes the derivative along characteristics parametrized by the arc length s.

Since $|\dot{\mathbf{x}}| = 1$, it was shown in [35] that the curvature bound along a characteristic is

$$|\ddot{\mathbf{x}}| \le \left| \frac{\nabla f(\mathbf{x})}{f(\mathbf{x})} \right|. \tag{3.30}$$

LEMMA 3.9. *Assuming that $f(\mathbf{x})$ is strictly positive and C^1 in Ω, the total variation of the tangent direction of the shortest characteristic L from an initial point $\mathbf{x}_0 \in \Gamma$ to a point $\mathbf{x} \in \Omega$ is bounded by*

$$\int_L |\ddot{\mathbf{x}}| ds \le \frac{DKf_M}{f_m}, \tag{3.31}$$

where s is the arc-length along the characteristic L, D is the diameter of domain Ω, and

$$K = \sup_{\mathbf{x}\in\Omega} \left| \frac{\nabla f(\mathbf{x})}{f(\mathbf{x})} \right|, \qquad f_M = \sup_{\mathbf{x}\in\Omega} f(\mathbf{x}), \qquad f_m = \inf_{\mathbf{x}\in\Omega} f(\mathbf{x}).$$

Proof. The existence of the shortest path L, yielding the first-arrival travel-time from an initial point $\mathbf{x}_0 \in \Gamma$ to a point $\mathbf{x} \in \Omega$, is guaranteed by the results in [14, 16]. If L is a single characteristic curve, then from (3.30) we have

$$\int_L |\ddot{\mathbf{x}}| ds \leq \int_L \frac{|\nabla f(\mathbf{x})|}{f(\mathbf{x})} ds \leq K \int_L ds, \tag{3.32}$$

where s is the arc-length; see Figure 3.1(b). The travel-time at $\mathbf{x}$ is $T(\mathbf{x}) = \int_L f(s)ds$. This travel-time, which is the first arrival time at $\mathbf{x}$, is smaller than the travel-time along the direct path from $\mathbf{x}_0$ to $\mathbf{x}$. So we have

$$f_m \int_L ds \leq \int_L f(s)ds = T(\mathbf{x}) \leq \int_{\mathbf{x}_0}^{\mathbf{x}} f(s)ds \leq f_M |\mathbf{x} - \mathbf{x}_0|. \tag{3.33}$$

Hence

$$\text{length}(L) = \int_L ds \leq \frac{D f_M}{f_m}. \tag{3.34}$$

Together with (3.32) we finish the proof. In general L may be composed of several pieces of characteristic curves. The above integral may be broken into several parts accordingly, but the same proof goes through. □

According to the above lemma the maximal number of sweeping needed to cover all characteristics can be bounded by $C \times \frac{DKf_M}{f_m}$, where the constant C may depend on the number of reference points and orderings.

Here is a discrete version of the above argument [36]. For an appropriate upwind scheme the corresponding discretized nonlinear system of equations has a solution (see Theorem 3.8). We can classify all nodes into a few groups according to the solution. All nodes in each group have a dependence pattern similar to their neighbors. For example, on a rectangular grid in two dimensions, almost all grid points can be divided into simply connected regions. In each region the value at a grid point depends on two of its neighbors in the following ways: (1) left and down neighbors; (2) left and up neighbors; (3) right and down neighbors; (4) right and up neighbors. By the Gauss–Seidel iteration each connected region can be covered by one of the orderings simultaneously when the ordering is in the upwind direction of the dependence pattern. The number of connected regions is proportional to the number of directional changes of characteristics which is bounded above. This relates the number of iterations for the fast sweeping method to the above bound. On a triangular mesh, because an arbitrary unstructured mesh may accommodate much more information flowing directions than a rectangular mesh, the situation is more complicated. However, given a triangulation and a choice of the reference points, all nodes can be partitioned into a finite number of connected regions. In each region the nodal dependence follows one of the orderings according to the increase/decrease of the distance to the reference points. For example, all those connected nodes, whose values depend on neighboring nodes that are closer to one of the reference points, belong to one region. The number of regions is proportional to the bound above. Although the triangulation and the choice of the reference points may affect the number of iterations, it is finite for a fixed setup.

4. Numerical examples. Now we show numerical examples in both two and three dimensions to illustrate the efficiency and the accuracy of our algorithm. In all the examples we have used the quick-sort method to order the nodes, though a radix sorting method may be implemented as well.

Our computational experience indicates that for an acute triangulation, using four corners in 2-D rectangular domains or eight corners in 3-D rectangular domains as the reference points is sufficient for the algorithm to converge in a finite number of iterations. For a triangulation with some obtuse triangles, more reference points may be needed. However, if the virtual splitting of obtuse angles as described in section 2.1 is used, then no extra reference point is needed; the results in convergence and accuracy are similar to those with all triangles being acute.

In all the presented examples the number of iterations is independent of the mesh size. The convergence of iteration is measured as full convergence in terms of the l^∞-norm; i.e., the iteration stops when the successive error reaches machine zero. On the other hand, the convergence order of the method is measured in the l^1-norm, as advocated by Lin and Tadmor [15].

We note that in our implementation, the convergence test is checked for every sweeping; here one sweeping is defined as passing through each node once according to a given ordering of nodes. So the iteration numbers reported in numerical examples are, in fact, the sweeping numbers needed for the algorithm to converge.

4.1. 2-D acute triangulation. We first triangulate the computational domain into acute triangles, then we refine the mesh uniformly by cutting each triangle in the coarse mesh into four smaller similar ones. We have chosen the four corners as the reference points in Examples 1, 2, and 3, with both the l^1- and l^2-metric–based sortings.

Example 1 (two-circle problem). The eikonal equation (2.1) with $f(x, y) = 1$. The computational domain is $\Omega = [-2, 2] \times [-2, 2]$; Γ consists of two circles of equal radius 0.5 with centers located at $(-1, 0)$ and $(\sqrt{1.5}, 0)$, respectively. The exact solution is the distance function to Γ. An acute triangulation is used in the computation. The solution is shown in Figure 4.1(a).

Example 2 (shape-from-shading). This example is taken from [26], in which

$$f(x, y) = 2\pi\sqrt{[\cos(2\pi x)\sin(2\pi y)]^2 + [\sin(2\pi x)\cos(2\pi y)]^2}. \tag{4.1}$$

$\Gamma = \{(\frac{1}{4}, \frac{1}{4}),\ (\frac{3}{4}, \frac{3}{4}),\ (\frac{1}{4}, \frac{3}{4}),\ (\frac{3}{4}, \frac{1}{4}),\ (\frac{1}{2}, \frac{1}{2})\}$, consisting of five isolated points. The computational domain is $\Omega = [0, 1] \times [0, 1]$. $T(x, y) = 0$ is prescribed on the boundary of the unit square. The solution to this problem is the shape function, which has the brightness $I(x, y) = 1/\sqrt{1 + f(x, y)^2}$ under vertical lighting. We have used acute triangulations for the following two cases.

Case a.

$$g\left(\frac{1}{4},\ \frac{1}{4}\right) = g\left(\frac{3}{4},\ \frac{3}{4}\right) = 1, \quad g\left(\frac{1}{4},\ \frac{3}{4}\right) = g\left(\frac{3}{4},\ \frac{1}{4}\right) = -1, \quad g\left(\frac{1}{2},\ \frac{1}{2}\right) = 0.$$

The exact solution for this case is smooth,

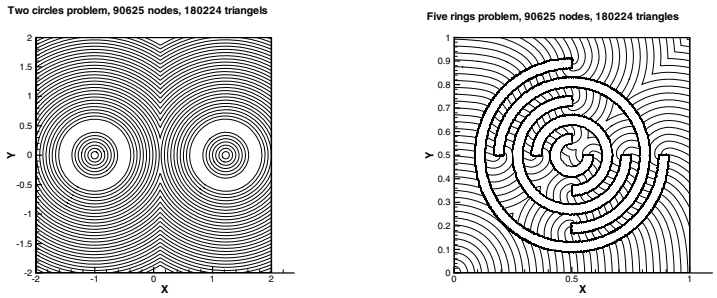
$$T(x, y) = \sin(2\pi x)\sin(2\pi y).$$

Case b.

$$g\left(\frac{1}{4},\ \frac{1}{4}\right) = g\left(\frac{3}{4},\ \frac{3}{4}\right) = g\left(\frac{1}{4},\ \frac{3}{4}\right) = g\left(\frac{3}{4},\ \frac{1}{4}\right) = 1, \quad g\left(\frac{1}{2},\ \frac{1}{2}\right) = 2.$$

The exact solution for this case is nonsmooth,

$$T(x, y) = \begin{cases} \max(|\sin(2\pi x)\sin(2\pi y)|,\ 1 + \cos(2\pi x)\cos(2\pi y)) \\ \qquad\qquad \text{if } |x + y - 1| < \frac{1}{2} \text{ and } |x - y| < \frac{1}{2}; \\ |\sin(2\pi x)\sin(2\pi y)| \qquad \text{otherwise.} \end{cases}$$



(a) (b)

FIG. 4.1. (a) *Example* 1: *two-circle problem.* (b) *Example* 3: *five-ring problem.*

TABLE 4.1
Accuracy tests for Examples 1 *and* 2. *Acute triangulation. Four corners as the reference points.*

		Two-circle		Shape (Case a)		Shape (Case b)	
Nodes	Elements	L^1 error	Order	L^1 error	Order	L^1 error	Order
1473	2816	7.71E-3	–	4.54E-2	–	2.83E-2	–
5716	11264	4.21E-3	0.87	2.54E-2	0.84	1.62E-2	0.81
22785	45056	2.18E-3	0.95	1.34E-2	0.92	8.76E-3	0.89
90625	180224	1.11E-3	0.97	6.90E-3	0.96	4.60E-3	0.93

TABLE 4.2
Iteration numbers for Examples $1, 2,$ *and* 3. *Acute triangulation. Spherical wave sweeping based on the* l^2*-metric ordering. Four corners as the reference points.*

Nodes	Elements	Two-circle	Shape (Case a)	Shape (Case b)	Five-ring
1473	2816	6	9	9	19
5716	11264	6	13	13	20
22785	45056	8	11	13	21
90625	180224	8	11	13	21

Example 3 (five-ring). The computational domain is $\Omega = [0,1] \times [0,1]$, Γ is a point source at $(0,0)$, and a five-ring obstacle is placed in the computational domain. This example is borrowed from [9]. Here we also use an acute triangulation. The solution is shown in Figure 4.1(b).

From Table 4.1, we can see that the accuracy of the algorithm for Examples 1 and 2 is first order. Although the same discretized nonlinear system is solved, no matter which ordering metric is used, different ordering strategies may result in different numbers of iterations, as illustrated in Tables 4.2 and 4.3, where we have applied orderings based on l^1- and l^2-metrics, respectively. Certainly, the two tables also indicate that the iteration number does not depend on the mesh size as the mesh is refined.

Table 4.4 shows the number of iterations needed using the l^1-metric with only two reference points. The two reference points are two corners that are not diagonal to each other.

On the other hand, Table 4.5 shows that a simple extension of the ordering strategy used for rectangular meshes, i.e., sorting all vertexes according to the ascent and descent orders of their x- and y-coordinates, may result in more iterations.

TABLE 4.3
Iteration numbers for Examples 1, 2, *and* 3. *Acute triangulation. Planar wave sweeping based on the* l^1*-metric ordering. Four corners as the reference points.*

Nodes	Elements	Two-circle	Shape (Case a)	Shape (Case b)	Five-ring
1473	2816	7	12	9	26
5716	11264	7	12	9	27
22785	45056	7	16	9	27
90625	180224	7	15	9	27

TABLE 4.4
Iteration numbers for Examples 1, 2, *and* 3. *Acute triangulation. Planar wave sweeping based on the* l^1*-metric ordering using only two reference points.*

Nodes	Elements	Two-circle	Shape (Case a)	Shape (Case b)	Five-ring
1473	2816	6	12	8	16
5716	11264	6	12	9	25
22785	45056	7	17	9	29
90625	180224	7	14	10	29

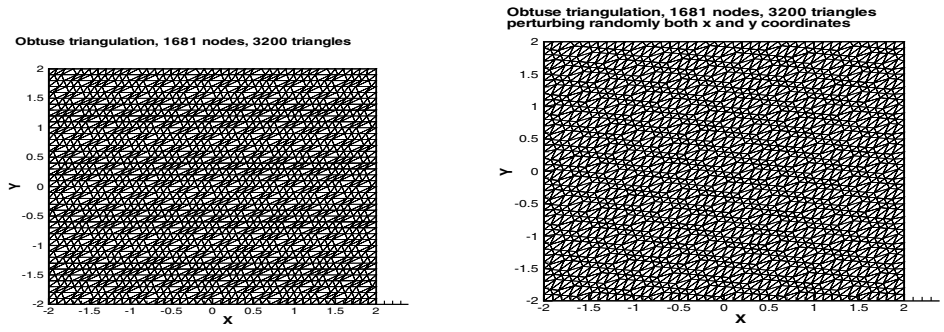
TABLE 4.5
Iteration numbers for Examples 1, 2, *and* 3. *Acute triangulation. Nodes are sorted by* x*- and* y*-coordinates. Four corners as the reference points.*

Nodes	Elements	Two-circle	Shape (Case a)	Shape (Case b)	Five-ring
1473	2816	9	9	9	22
5716	11264	9	10	14	26
22785	45056	13	18	15	33
90625	180224	13	13	15	33

4.2. 2-D obtuse triangulation. We test our strategy for treating a triangulation which has obtuse angles. The obtuse triangulations are constructed by perturbing randomly the x-coordinates of vertexes (Figure 4.2(a)) or perturbing randomly both the x-coordinates and the y-coordinates of vertexes (Figure 4.2(b)) in a uniform triangulation. The uniform triangulation, in turn, is obtained by connecting the diagonal line in every rectangle of a rectangular mesh and cutting every rectangle into two equivalent isosceles triangles. The perturbation range is $[-0.5h, 0.5h]$, where h is the length of an isosceles triangle. We use Example 1 in section 4.1 as a test example and apply spherical-wave sweepings.

As a first test, we use the obtuse triangulation as in Figure 4.2(a), choose four corners of the computational domain as the reference points, and sweep through all the nodes according to both ascent and descent sortings. The accuracy and the number of iterations for the algorithm without and with the obtuse-angle treatment are listed in Table 4.6.

As a second test, we use eight reference points which include both the four corners and four middle points of the four edges of the computational domain, and we use only ascent orders. The accuracy and the number of iterations for the algorithm without and with the obtuse-angle treatment are listed in Table 4.7 for the obtuse triangulation as in Figure 4.2(a) and in Table 4.8 for the obtuse triangulation as in Figure 4.2(b). Comparing Tables 4.6, 4.7, and 4.8, we can see that more reference points may help us reduce the number of sweepings needed in the algorithm. Roughly speaking, for different meshes the errors from the algorithm with the obtuse-angle treatment are decreased $2 \sim 4$ times in comparison to the errors from the algorithm without such a treatment, as shown in both Table 4.7 and Table 4.8. The first-order



(a) (b)

FIG. 4.2. *Obtuse triangulations.* (a) *Perturbing randomly the x-coordinate of vertexes in a uniform triangulation;* (b) *perturbing randomly x- and y-coordinates of vertexes.*

TABLE 4.6

Two-circle problem. Obtuse triangulation (Figure 4.2(a)*). Spherical wave sweepings:* 4 *reference points (*4 *corners of computational domain). Both ascent and descent orderings.*

			Before treatment			After treatment		
Elements	Obtuse elements	Max obtu	L^1 error	Order	Iter.	L^1 error	Order	Iter.
200	78	120°	6.70E-2	–	6	4.26E-2	–	5
800	528	115°	2.49E-2	1.43	8	1.71E-2	1.32	6
3200	958	125°	2.90E-2	−0.22	15	9.71E-3	0.81	12
12800	5890	118°	1.98E-2	0.55	34	4.60E-3	1.08	18
51200	40558	116°	4.71E-3	2.07	44	2.31E-3	0.99	24

TABLE 4.7

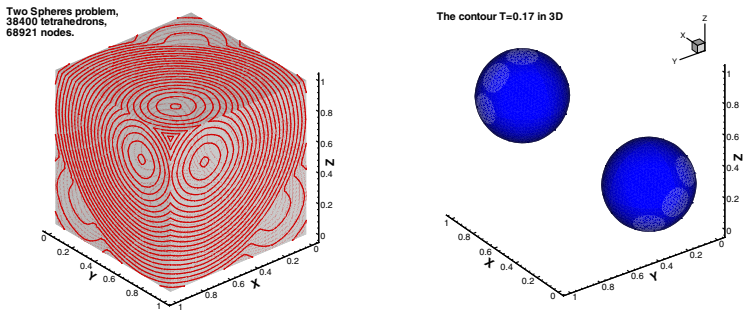
Two-circle problem. Obtuse triangulation (Figure 4.2(a)*). Spherical wave sweepings:* 8 *reference points (*4 *corners and* 4 *middle points of the* 4 *edges). Only ascent ordering.*

			Before treatment			After treatment		
Elements	Obtuse elements	Max obtu	L^1 error	Order	Iter.	L^1 error	Order	Iter.
200	78	120°	6.70E-2	–	4	4.26E-2	–	4
800	528	115°	2.49E-2	1.43	8	1.71E-2	1.32	6
3200	958	125°	2.91E-2	−0.22	8	9.71E-3	0.81	8
12800	5890	118°	1.98E-2	0.55	8	4.60E-3	1.08	9
51200	40558	116°	4.72E-3	2.07	13	2.31E-3	0.99	11

TABLE 4.8

Two-circle problem. Obtuse triangulation (Figure 4.2(b)*). Spherical wave sweepings:* 8 *reference points (*4 *corners and* 4 *middle points of the* 4 *edges). Only ascent ordering.*

			Before treatment			After treatment		
Elements	Obtuse elements	Max obtu	L^1 error	Order	Iter.	L^1 error	Order	Iter.
200	81	106°	3.55E-2	–	4	3.08E-2	–	4
800	727	108°	2.30E-2	0.63	8	1.70E-2	0.86	4
3200	1344	106°	1.32E-2	0.80	8	8.04E-3	1.08	6
12800	5909	106°	7.73E-3	0.77	11	4.66E-3	0.79	10
51200	50560	108°	3.88E-3	0.99	14	1.89E-3	1.31	14



(a) (b)

FIG. 4.3. *Two-sphere problem. Use a tetrahedral mesh.* (a) *The surface contour,* 30 *equally spaced contour lines from* $T = 0$ *to* $T = 0.742402$ *(produced automatically by the plotting software);* (b) *the contour plot of* $T = 0.17$ *in the* 3*-D case.*

TABLE 4.9
Two-sphere problem. Comparison between tetrahedral meshes and rectangular meshes. Spherical wave sweepings: 8 *corners as reference points. Both ascent and descent orderings.*

		Unstructured mesh			Structured mesh		
Nodes	Elements	L^1 error	Order	Iter.	L^1 error	Order	Iter.
9261	48000	1.25E-2	–	12	1.77E-2	–	15
68921	384000	7.17E-3	0.81	12	1.02E-2	0.80	15
531441	3072000	3.79E-3	0.92	12	5.41E-3	0.91	16

accuracy with the obtuse-angle treatment is more regular than that without the treatment. Moreover, comparing the errors in Table 4.6 with those in Table 4.7, we can observe that without the obtuse-angle treatment different sweeping ordering strategies yield slightly different numerical solutions, and with the obtuse-angle treatment different sweeping ordering strategies yield the same solutions up to machine zero. This indicates that the causality of PDEs may *not* be captured accurately if obtuse angles are *not* treated.

4.3. A 3-D example. We test our 3-D fast sweeping methods on tetrahedral meshes. We use a two-sphere problem as an example: the eikonal equation (2.3) with $f(x, y, z) = 1$.

The computational domain is $\Omega = [0, 1] \times [0, 1] \times [0, 1]$; Γ consists of two spheres of equal radius 0.1 with centers located at $(0.25, 0.25, 0.25)$ and $(0.75, 0.75, 0.75)$, respectively. The exact solution is the distance function to Γ.

We first partition the computational domain into identical rectangular cubes. Then a tetrahedral mesh is obtained by cutting each cube into six tetrahedrons.

The results in Figure 4.3 are obtained by using a tetrahedral mesh which consists of $40 \times 40 \times 40 \times 6 = 384000$ tetrahedrons. We choose the eight corners of the computational domain as the reference points. Both ascent and descent orderings are used, and the ordering strategy is based on the l^2-metric. Figure 4.3(a) shows contours of the solution on the surface of the domain, and Figure 4.3(b) shows 3-D plots of the contour $T = 0.17$.

In Table 4.9, we present the accuracy and numbers of iterations when the tetrahedral mesh is refined. To calibrate the result, we apply the same sweeping ordering to the rectangular mesh from which the tetrahedral mesh is obtained. For the rectangular mesh we use the local solver for rectangular grids as in [35]. Although the

TABLE 4.10

Two-circle problem. Comparison between triangular meshes and rectangular meshes. Spherical wave sweepings: 4 corners as reference points.

		Unstructured mesh			Structured mesh		
Nodes	Elements	L^1 error	Order	Iter.	L^1 error	Order	Iter.
1681	3200	9.85E-3	–	5	1.46E-2	–	5
6561	12800	5.30E-3	0.89	5	7.91E-3	0.88	5
25921	51200	2.74E-3	0.95	5	4.13E-3	0.94	5
103041	204800	1.39E-3	0.98	5	2.10E-3	0.97	5

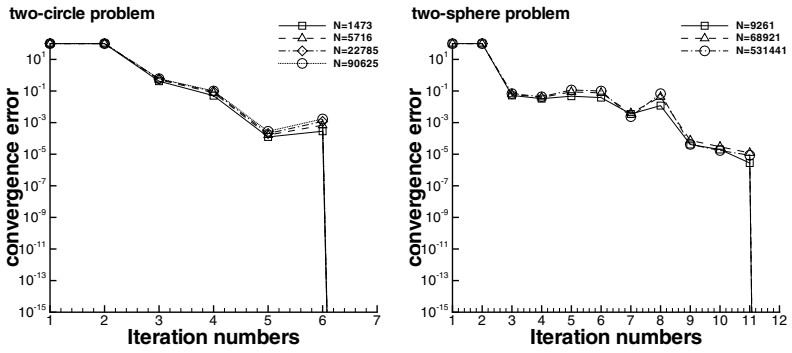


FIG. 4.4. log *plot of convergence error for* 2-*D and* 3-*D examples.*

nodes are the same, the local solvers at each node are different so that the discretized nonlinear systems of the equation are different. The comparison results are also shown in Table 4.9. It is obvious from the table that the local solver on unstructured meshes can achieve higher accuracy than that on structured meshes since the former uses more neighboring points at each node and captures directions of characteristics more accurately than the latter.

Also we can see from Table 4.9 that if the l^2-metric is used for ordering, the number of iterations on an unstructured mesh can be less than that on a structured one. However, the local solver at each node for an unstructured mesh is more expensive than that for a rectangular mesh. Most importantly, we see that both iteration numbers do not change as the mesh is refined. So our ordering strategy works for both cases.

A similar comparison for a 2-D case, Example 1 of section 4.1, is presented in Table 4.10; again the local solver on unstructured meshes achieves higher accuracy than that on structured meshes.

4.4. Typical convergence behavior. Figure 4.4 shows the typical behavior of convergence error of the fast sweeping method in terms of the difference between two consecutive iterations in maximum norm. It demonstrates that the exact solution (up to machine error) to the discretized system is achieved in a finite number of iterations independent of mesh size.

5. Conclusion. We propose novel ordering strategies to extend the fast sweeping method to unstructured meshes. To that end we introduce multiple reference points and order all the nodes according to their l^p-metrics to those reference points. Information propagating along all characteristics can be covered efficiently by the

systematic orderings. We prove that the new algorithm converges and numerical examples demonstrate that the algorithm converges in a finite number of iterations independent of mesh size. The computational complexity of the new algorithm is nearly optimal in the sense that the total computational cost consists of $O(M)$ flops for iteration steps and $O(M\log M)$ flops for sorting at the predetermined initialization step, which can be efficiently optimized by adopting a linear time sorting method, where M is the total number of mesh points. Extensive numerical examples demonstrate the accuracy and the efficiency of the new fast sweeping method.

Acknowledgment. Qian thanks Profs. Stan Osher, William W. Symes, and Eitan Tadmor for encouragement in this project. Qian also thanks Prof. Ian Mitchell for comments related to sorting algorithms. The authors would like to thank the anonymous referees for constructive comments on the paper.

REFERENCES

[1] M. Bardi and S. Osher, *The nonconvex multi-dimensional Riemann problem for Hamilton-Jacobi equations*, SIAM J. Math. Anal., 22 (1991), pp. 344–351.
[2] M. Boué and P. Dupuis, *Markov chain approximations for deterministic control problems with affine dynamics and quadratic costs in the control*, SIAM J. Numer. Anal., 36 (1999), pp. 667–695.
[3] B. Cockburn and J. Qian, *Continuous dependence results for Hamilton–Jacobi equations*, in Collected Lectures on the Preservation of Stability under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002, pp. 67–90.
[4] T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms*, MIT Press, Cambridge, MA, 1990.
[5] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, John Wiley and Sons, New York, 1962.
[6] M. G. Crandall and P. L. Lions, *Two approximations of solutions of Hamilton–Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.
[7] J. Dellinger and W. W. Symes, *Anisotropic finite-difference traveltimes using a Hamilton–Jacobi solver*, in Proceedings of the 67th Annual International Meeting, Soc. Expl. Geophys., Expanded Abstracts, Soc. Expl. Geophys., Tulsa, OK, 1997, pp. 1786–1789.
[8] M. Falcone and R. Ferretti, *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.
[9] P. A. Gremaud and C. M. Kuster, *Computational study of fast methods for the eikonal equations*, SIAM J. Sci. Comput., 27 (2006), pp. 1803–1816.
[10] J. Helmsen, E. Puckett, P. Colella, and M. Dorr, *Two new methods for simulating photolithography development in 3-D*, Proc. SPIE, 2726 (1996), pp. 253–261.
[11] C.-Y. Kao, S. Osher, and Y.-H. Tsai, *Fast sweeping methods for static Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 42 (2005), pp. 2612–2632.
[12] C. Y. Kao, S. J. Osher, and J. Qian, *Lax–Friedrichs sweeping schemes for static Hamilton–Jacobi equations*, J. Comput. Phys., 196 (2004), pp. 367–391.
[13] R. Kimmel and J. A. Sethian, *Computing geodesic paths on manifolds*, in Proc. Natl. Acad. Sci., 95 (1998), pp. 8431–8435.
[14] S. N. Kruzkov, *Generalized solutions of nonlinear first order equations with several independent variables.* II, Math. USSR-Sb., 1 (1967), pp. 93–116.
[15] C.-T. Lin and E. Tadmor, *High-resolution nonoscillatory central schemes for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2163–2186.
[16] P. L. Lions, *Generalized Solutions of Hamilton-Jacobi equations*, Pitman, Boston, MA, 1982.
[17] S. Osher and C.-W. Shu, *High-order essentially nonoscillatory schemes for Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 28 (1991), pp. 907–922.
[18] J. Qian and W. W. Symes, *Paraxial eikonal solvers for anisotropic quasi-P traveltimes*, J. Comput. Phys., 173 (2001), pp. 1–23.
[19] J. Qian and W. W. Symes, *Adaptive finite difference method for traveltime and amplitude*, Geophys., 67 (2002), pp. 167–176.
[20] J. Qian and W. W. Symes, *Finite-difference quasi-P traveltimes for anisotropic media*, Geophys., 67 (2002), pp. 147–155.
[21] J. Qian and W. W. Symes, *Paraxial geometrical optics for quasi-P waves: Theories and*

numerical methods, Wave Motion, 35 (2002), pp. 205–221.
[22] J. Qian, Y.-T. Zhang, and H. K. Zhao, *Fast sweeping methods for static Hamilton–Jacobi equations on triangular meshes*, J. Sci. Comput., submitted.
[23] F. Qin, Y. Luo, K. B. Olsen, W. Cai, and G. T. Schuster, *Finite difference solution of the eikonal equation along expanding wavefronts*, Geophys., 57 (1992), pp. 478–487.
[24] F. Qin and G. T. Schuster, *First-arrival traveltime calculation for anisotropic media*, Geophys., 58 (1993), pp. 1349–1358.
[25] R. Rawlinson and M. Sambridge, *Wave front evolution in strongly heterogeneous layered media using the fast marching method*, Geophys. J. Internat., 156 (2004), pp. 631–647.
[26] E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from-shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.
[27] W. A. Schneider, Jr., K. Ranzinger, A. Balch, and C. Kruse, *A dynamic programming approach to first arrival traveltime computation in media with arbitrarily distributed velocities*, Geophys., 57 (1992), pp. 39–50.
[28] J. A. Sethian, *Level Set Methods*, Cambridge University Press, Cambridge, UK, 1996.
[29] Y.-H. R. Tsai, L.-T. Cheng, S. Osher, and H.-K. Zhao, *Fast sweeping algorithms for a class of Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 673–694.
[30] J. N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Tran. Automat. Control, 40 (1995), pp. 1528–1538.
[31] J. van Trier and W. W. Symes, *Upwind finite-difference calculation of traveltimes*, Geophys., 56 (1991), pp. 812–821.
[32] J. Vidale, *Finite-difference calculation of travel times*, Bull. Seismol. Soc. Amer. 78 (1988), 2062–2076.
[33] Y.-T. Zhang, H. K. Zhao, and S. Chen, *Fixed-point iterative sweeping methods for static Hamilton-Jacobi equations*, Methods Appl. Anal., to appear.
[34] Y.-T. Zhang, H. K. Zhao, and J. Qian, *High order fast sweeping methods for static Hamilton-Jacobi equations*, J. Sci. Comput., 29 (2006), pp. 25–56.
[35] H. K. Zhao, *Fast sweeping method for eikonal equations*, Math. Comp., 74 (2005), pp. 603–627.
[36] H. K. Zhao, *Parallel Implementations of the Fast Sweeping Method*, research report, UCLA CAM 06-13, University of California, Los Angeles, 2006.
[37] H. K. Zhao, S. Osher, B. Merriman, and M. Kang, *Implicit and nonparametric shape reconstruction from unorganized points using variational level set method*, Comput. Vision and Image Understanding, 80 (2000), pp. 295–319.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 108–132

APPROXIMATION AND RECONSTRUCTION FROM ATTENUATED RADON PROJECTIONS*

YUAN XU†, OLEG TISCHENKO‡, AND CHRISTOPH HOESCHEN‡

Abstract. Attenuated Radon projections with respect to the weight function $W_\mu(x,y) = (1 - x^2 - y^2)^{\mu-1/2}$ are shown to be closely related to the orthogonal expansion in two variables with respect to W_μ. This leads to an algorithm for reconstructing two-dimensional functions (images) from attenuated Radon projections. Similar results are established for reconstructing functions on the sphere from projections described by integrals over circles on the sphere, and for reconstructing functions on the three-dimensional ball and cylinder domains.

Key words. approximation, reconstruction of images, Radon projections, polynomials of several variables, algorithms

AMS subject classifications. 42A38, 42B08, 42B15

DOI. 10.1137/05064388X

1. Introduction.

Computer tomography (CT) offers a noninvasive method for two-dimensional (2D) cross-sectional or three-dimensional (3D) imaging of an object. In a typical CT application, the distribution of the attenuation coefficient through a body from measurements of x-ray transmission is estimated and used to reconstruct an image of the object. The mathematical foundation of CT is the Radon transform. Let f be a function defined on the unit disk B^2 of the $\mathbb{R}^2$ plane. A Radon transform of f is a line integral,

$$\mathcal{R}_\theta(f;t) := \int_{I(\theta,t)} f(x,y)dx\,dy, \qquad 0 \le \theta \le 2\pi, \quad -1 \le t \le 1, \tag{1.1}$$

where $I(\theta,t) = \{(x,y) : x\cos\theta + y\sin\theta = t\} \cap B^2$ is a line segment inside B^2. An essential problem in CT is to reconstruct the function f from its Radon projections. An algorithm amounts to an approximation to f that uses values of $\mathcal{R}_\theta(f;t)$ from a finite set of parameters (θ,t).

The attenuation of an x-ray beam is dependent on the energy of each photon. A line integral as defined in (1.1) represents a monochromatic x-ray. In practice, however, an x-ray is usually polychromatic, meaning that it consists of photons with different energies. This could lead to artifacts in the reconstruction; see, for example, [4, Chap. 4]. A polychromatic x-ray is represented by the so-called attenuated Radon projections for which the integral is taken against $\exp\{-\alpha_\theta(x,y)\}dx\,dy$, where $\alpha_\theta(x,y)$ is a given function, instead of $dx\,dy$. The attenuated Radon transform appears in, for example, emission tomography [7]. The reconstruction algorithms for attenuated Radon data have been derived from Novikov's inversion formula; see [10] and [8]. See also the recent survey in [3] in this direction.

In the present paper we consider the special case that $\exp\{-\alpha_\theta(x,y)\}$ is given, or

*Received by the editors October 31, 2005; accepted for publication (in revised form) July 5, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sinum/45-1/64388.html
†Department of Mathematics, University of Oregon, Eugene, OR 97403-1222 (yuan@math.uoregon.edu).
‡Institute of Radiation Protection, GSF-National Research Center for Environment and Health, D-85764 Neuherberg, Germany (oleg.tischenko@gsf.de, christoph.hoeschen@gsf.de).

can be approximated, by the function

$$W_\mu(x,y) = (1-x^2-y^2)^{\mu-1/2}, \qquad (x,y)\in B^2, \tag{1.2}$$

where $\mu \ge 0$; in other words, $\alpha_\theta(x,y) = -(\mu-1/2)\log(1-x^2-y^2)$. The attenuated Radon transform, denoted by $\mathcal{R}_\theta^\mu$, then takes the form

$$\mathcal{R}_\theta^\mu(f;t) := \int_{I(\theta,t)} f(x,y)W_\mu(x,y)dx\,dy, \qquad 0\le\theta\le 2\pi, \quad -1\le t\le 1. \tag{1.3}$$

Clearly this is just a special case of the attenuated Radon transform. This case, however, appears to be useful in understanding the effect of monochromatic and polychromatic x-rays. In this regard let us mention the classical example of the water phantom in a skull in [4, p. 121], which demonstrated that beam hardening causes an elevation in CT numbers for tissues close to the skull bone. The attenuated Radon transform defined in (1.3) models the boundary behavior of the x-rays differently.

Our approach is based on orthogonal polynomial expansions on B^2. Let $\mathcal{V}_n^2(W_\mu)$ denote the space of orthogonal polynomials with respect to the weight function W_μ on B^2. It is well known that

$$L^2(B^2,W_\mu) = \sum_{k=0}^\infty \oplus \mathcal{V}_k^2 : \qquad f = \sum_{k=1}^\infty \operatorname{proj}_k^\mu f,$$

where $\operatorname{proj}_k^\mu f$ is the projection of f on $\mathcal{V}_k^2(W_\mu)$. The infinite series holds in the sense that the sequence of the partial sums

$$S_n^\mu(f;x,y) := \sum_{k=0}^n \operatorname{proj}_k^\mu f(x,y), \qquad n\ge 0,$$

converges to f as $n\to\infty$ in $L^2(B^2,W_\mu)$ norm. The partial sum S_nf provides a natural approximation to f. It turns out that there is a remarkable connection between $S_n^\mu f$ and the attenuated Radon transforms, which states that

$$S_{2m}f(x,y) = \sum_{\nu=0}^{2m}\int_{-1}^1 \mathcal{R}_{\phi_\nu}^\mu(f;t)\Phi_\nu(t;x,y)dt, \qquad \phi_\nu = \frac{2\nu\pi}{2m+1}, \tag{1.4}$$

where Φ_ν are polynomials of two variables given by explicit formulas. This representation provides a simple and direct access to attenuated Radon data. For the ordinary Radon transforms ($\mu = 1/2$), this was discovered recently in [16]. Applying an appropriate quadrature formula to the integrals in the expression leads to an approximation to f that uses discrete attenuated Radon projections. One important feature of the algorithm is that polynomials up to a certain degree are reconstructed exactly, which guarantees that the algorithm has a fast rate of convergence. Such an algorithm can be easily implemented numerically. For the ordinary Radon transforms, the algorithm is named OPED (orthogonal polynomial expansion on the disk) and has proved to be a highly effective method [17, 18].

There are other expressions in the spirit of (1.4). In order to prove them, we need to study orthogonal expansions in terms of orthogonal polynomials with respect to $W_\mu(x,y)$ on B^2. The case $\mu = 1/2$ is easier since an orthonormal basis for $\mathcal{V}_k^2(W_{1/2})$ is known to be $U_k(x\cos\frac{j\pi}{k+1} + y\sin\frac{j\pi}{k+1})$, $0\le j\le k$. No such convenient orthonormal basis is available for $\mu\ne 1/2$.

There is another advantage for considering the attenuated Radon transform $\mathcal{R}_\theta^\mu(f;t)$. It is known that there is a close relation between orthogonal polynomials on the unit ball and those on the unit sphere, which allows us to establish analogous results on the unit sphere S^2. In particular, the case $\mu = 0$ on B^2 can be used to show that we can reconstruct a function f from its integral projections:

$$Qf(\zeta;t) = \int_{\langle \mathbf{x},\zeta\rangle = t} f(\mathbf{x}) d\omega(\mathbf{x}), \qquad 0 \neq \zeta \in S^2, \quad -1 \le t \le 1, \tag{1.5}$$

where $\mathbf{x} = (x_1, x_2, x_3)$ and $d\omega$ is the surface measure on S^2. Reconstruction from such spherical transforms has been studied in the literature; see [9].

From the disk B^2 we can also extend the results to the unit ball B^3 and to cylinder domains in $\mathbb{R}^3$, taking Radon projections on parallel disks in each case. It turns out, however, that there is an important difference between the ball and the cylinder. For the cylinder domain, all results obtained in the disk can be extended without problem. For the unit ball, however, we still have an analogue of (1.4), but the reconstruction algorithm may no longer work as efficiently as in the cylinder case. The problem is that the operator produced by the algorithm no longer preserves polynomials.

For the algorithm on B^2, we provide a numerical example in section 2, which reconstructs a 2D phantom image for three different values of μ. For the transform on the sphere and the 3D transforms on the ball and on the cylinder domain, we will be content with deriving the algorithms and will not discuss convergence or the performance of the algorithms at this time.

The paper is organized as follows. In the following section we consider the reconstruction and approximation on the unit disk B^2 from attenuated Radon projections. This section is divided into several subsections, the last one including the numerical example. In section 3 the results on B^2 are transplanted to those on the surface S^2, while the attenuated Radon projections become weighted spherical transforms. The analogous results are then established for the unit ball B^3 in section 4 and for the cylinder domain in section 5.

2. Reconstruction and approximation on the unit disk. Let Π^d denote the space of polynomials of d variables and let Π_n^d denote the subspace of polynomials of total degree n in Π^d, which has dimension $\dim \Pi_n^d = \binom{n+d}{d}$. We set $\Pi_n := \Pi_n^1$. In this section we mainly work with the case $d = 2$.

2.1. Orthogonal polynomials on the unit disk. Let W_μ be the weight function defined in (1.2). Let $\mathcal{V}_k^2(W_\mu)$ denote the space of orthogonal polynomials of degree k on B^2 with respect to the inner product

$$\langle P, Q\rangle_\mu = a_\mu \int_{B^2} P(x,y)Q(x,y)W_\mu(x,y)dx\,dy, \qquad a_\mu = (\mu + 1/2)/\pi,$$

where a_μ is the normalization constant of W_μ, $a_\mu = 1/\int_{B^2} W_\mu(x)dx$. Thus, $P \in \mathcal{V}_k^2(W_\mu)$ if P is of degree k and $\langle P, Q\rangle_\mu = 0$ for all $Q \in \Pi_{k-1}^2$. We note that elements in a basis for $\mathcal{V}_k^2(W_\mu)$ may not be orthogonal with respect to each other according to our definition. A basis for $\mathcal{V}_k^2(W_\mu)$ is called orthonormal if the elements in the basis are mutually orthogonal and $\langle P, P\rangle_\mu = 1$.

The reproducing kernel of the space $\mathcal{V}_k^2(W_\mu)$ plays an important role in our development. In terms of an orthonormal basis $\{P_j^k : 0 \le j \le k\}$ of $\mathcal{V}_k^2(W_\mu)$, the

reproducing kernel satisfies

$$P_k(W_\mu; \mathbf{x}, \mathbf{y}) = \sum_{j=0}^{k} P_j^k(\mathbf{x}) P_j^k(\mathbf{y}). \tag{2.1}$$

The kernel is independent of the choice of the bases of $\mathcal{V}_k^2(W_\mu)$. In fact, a compact formula for this kernel can be given in terms of the Gegenbauer polynomial [13],

$$\begin{aligned} P_k(W_\mu; \mathbf{x}, \mathbf{y}) = \frac{k+\mu+1/2}{\mu+1/2} b_{\mu-1} \int_{-1}^{1} C_k^{\mu+1/2}\Big(\langle \mathbf{x}, \mathbf{y}\rangle \\ + \sqrt{1-\|\mathbf{x}\|^2}\sqrt{1-\|\mathbf{y}\|^2}\, t\Big)(1-t^2)^{\mu-1} dt \end{aligned} \tag{2.2}$$

for $\mu > 0$; the formula also holds for $\mu = 0$ upon taking limit $\mu \to 0$. Here and in the following, the Gegenbauer polynomials $C_k^\lambda(s)$ are orthogonal with respect to $(1-s^2)^{\lambda-1/2}$ on $[-1,1]$,

$$c_{\lambda-1/2} \int_{-1}^{1} C_k^\lambda(s) C_l^\lambda(s)(1-s^2)^{\lambda-1/2} ds = \frac{\lambda(2\lambda)_k}{(k+\lambda)k!}\delta_{k,l} := h_k \delta_{k,l}, \tag{2.3}$$

where $c_{\lambda-1/2} := \Gamma(\lambda+1)/(\sqrt{\pi}\Gamma(\lambda+1/2))$ is the normalization constant of the weight function $(1-s^2)^{\lambda-1/2}$ on $[-1,1]$, and $(a)_k := a(a+1)\cdots(a+k-1)$. For $\mu = 1/2$, $C_k^{\mu+1/2}(s) = U_k(s)$ is the Chebyshev polynomial of the second kind.

For the weight function $W_{1/2}(x) = 1$, it is known [5] that the set

$$\{U_k(x\cos\theta_{j,k} + y\sin\theta_{j,k}) : 0 \le j \le k\}$$

forms an orthonormal basis of $\mathcal{V}_k^2(W_{1/2})$. The elements of this basis are the so-called ridge functions. In general, given an angle ϕ and a polynomial $p \in \Pi_k := \Pi_k^1$, a ridge polynomial is defined by

$$p(\phi; x, y) := p(x\cos\phi + y\sin\phi), \qquad \phi \in [0, 2\pi].$$

It is easy to see that $p(\phi; x, y)$ is a polynomial in Π_k^2 as well. The functions $\{C_k^{\mu+1/2}(\theta_{j,k}; x, y) : 0 \le j \le k\}$, where $\theta_{j,k} = j\pi/(k+1)$, form a basis for $\mathcal{V}_k^2(W_\mu)$, albeit not a mutually orthogonal one (see, for example, [14]). The lack of an orthonormal ridge basis in the case of $\mu \ne 1/2$ makes the results for the attenuated Radon transform more difficult, as we shall see below.

We call a polynomial $P \in \Pi_k$ of one variable *symmetric* with respect to the origin if P is even when k is even, and if P is odd when k is odd. It is known that $C_k^{\mu+1/2}(t)$ is symmetric with respect to the origin. The ridge polynomials arising from such a polynomial turn out to satisfy a remarkable relation.

PROPOSITION 2.1. *For $n \ge 0$ and $k \le n$, the identity*

$$\frac{1}{n+1}\sum_{\nu=0}^{n} U_k\left(\frac{\nu\pi}{n+1}; \cos\theta, \sin\theta\right) P_k\left(\frac{\nu\pi}{n+1}; x, y\right) = P_k(\theta; x, y) \tag{2.4}$$

holds for all polynomials $P_k \in \Pi_k$ that are symmetric with respect to the origin.

Proof. The proof uses the elementary trigonometric identities

$$\sum_{\nu=0}^{n} \sin k\frac{2\nu\pi}{n+1} = 0 \quad \text{and} \quad \sum_{\nu=0}^{n} \cos k\frac{2\nu\pi}{n+1} = \begin{cases} n+1 & \text{if } k = 0 \mod n+1, \\ 0 & \text{otherwise,} \end{cases} \tag{2.5}$$

which hold for all nonnegative integers k. Let us prove the case $k = 2l$. We follow the proof of Proposition 2.3 in [16]. The polynomial P_k can be written as a linear combination of U_{k-2j} for $0 \le 2j \le k$. Consequently, we can write $P_{2l}(\theta; x, y)$ as

$$P_{2l}(\theta; x, y) = P_{2l}(r\cos(\theta - \phi)) = \sum_{j=0}^{l} b_j(r) \cos 2j(\theta - \phi) \tag{2.6}$$

in polar coordinates $x = r\cos\phi$ and $y = r\sin\phi$, where $b_j(r)$ is a polynomial of degree $2j$ in r. Furthermore, we know that

$$U_{2l}(\theta; \cos\phi, \sin\phi) = U_{2l}(\cos(\theta - \phi)) = \sum_{j=0}^{l} d_j \cos 2j(\theta - \phi),$$

where $d_0 = 1$ and $d_j = 2$ for $j \ge 1$. The identities (2.5) and the product formula of the cosine function show that

$$\frac{1}{n+1} \sum_{\nu=0}^{n} \cos 2i\left(\theta - \frac{\nu\pi}{n+1}\right) \cos 2j\left(\phi - \frac{\nu\pi}{n+1}\right) = \begin{cases} 0 & \text{if } i \ne j, \\ \frac{1}{2}\cos 2j(\theta - \phi) & \text{if } 0 < i = j \le n, \\ 1 & \text{if } i = j = 0. \end{cases}$$

Let us denote by I_k the left-hand side of (2.4). The above trigonometric identity implies immediately that, for $0 \le 2l \le n$,

$$\begin{aligned} I_{2l} &= \sum_{i=0}^{l} d_i \sum_{j=0}^{l} b_j(r) \frac{1}{n+1} \sum_{\nu=0}^{n} \cos 2i\left(\theta - \frac{\nu\pi}{n+1}\right) \cos 2j\left(\phi - \frac{\nu\pi}{n+1}\right) \\ &= \sum_{j=0}^{l} b_j(r) \cos 2j(\theta - \phi) = P_{2l}(r\cos(\theta - \phi)) = P_{2l}(\theta; x, y). \end{aligned}$$

This completes the proof for the case $k = 2l \le n$. The case $k = 2l - 1$ is similar. □

In (2.4) the summation is over angles, $\nu\pi/(n+1)$, that are equally spaced in the interval $[0, \pi)$. In the case that n is even, the angles can be arranged as equally spaced angles in $[0, 2\pi]$ by using the fact that

$$\cos\frac{(2k+1)\pi}{2m+1} = -\cos\frac{(2m+2k)\pi}{2m+1} \quad \text{and} \quad \sin\frac{(2k+1)\pi}{2m+1} = -\sin\frac{(2m+2k)\pi}{2m+1}. \tag{2.7}$$

The result is the following proposition proved in [16] for P_k being the Chebyshev polynomial of the second kind.

PROPOSITION 2.2. *For $m \ge 0$ and $k \le 2m$, the identity*

$$\frac{1}{2m+1} \sum_{\nu=0}^{2m} U_k\left(\frac{2\nu\pi}{2m+1}; \cos\theta, \sin\theta\right) P_k\left(\frac{2\nu\pi}{2m+1}; x, y\right) = P_k(\theta; x, y) \tag{2.8}$$

holds for all polynomials $P_k \in \Pi_k$ that are symmetric with respect to the origin.

There are many orthonormal bases of $\mathcal{V}_k^2(W_\mu)$ that are known explicitly (see [2]). One that is particularly useful for us is given in terms of the polar coordinates

$$x = r\cos\phi, y = r\sin\phi, \qquad 0 \le r \le 1, \quad 0 \le \phi \le 2\pi,$$

and Jacobi polynomials [2, Prop. 2.3.1]. Let $p_n^{(\alpha,\beta)}(t)$ denote the orthonormal Jacobi polynomials, that is,

$$c_{\alpha,\beta}\int_{-1}^{1} p_n^{(\alpha,\beta)}(t)p_m^{(\alpha,\beta)}(t)(1-t)^\alpha(1+t)^\beta dt = \delta_{m,n}, \quad m,n = 0,1,2,\ldots,$$

where $c_{\alpha,\beta}$ is the normalized constant so that $c_{\alpha,\beta}\int_{-1}^{1}(1-t)^\alpha(1+t)^\beta dt = 1$.

PROPOSITION 2.3. *For $\varepsilon = 0$ or 1, define the polynomials $P_{l,\varepsilon}^k$ by*

$$P_{l,\varepsilon}^k(x,y) = h_{l,k}p_l^{(\mu-\frac{1}{2},k-2l)}(2r^2-1)r^{k-2l}S_{k-2l,\varepsilon}(\phi), \tag{2.9}$$

where

$$\begin{aligned} S_{k-2l,0}(\phi) &= \cos(k-2l)\phi \quad \text{for } 0 \le 2l \le k,\\ S_{k-2l,1}(\phi) &= \sin(k-2l)\phi \quad \text{for } 0 \le 2l \le k-1, \end{aligned}$$

and

$$[h_{l,k}]^2 := \frac{\Gamma(k-2l+\mu+3/2)}{\Gamma(\mu+3/2)\Gamma(k-2l+1)}.$$

Then these polynomials form an orthonormal basis for $\mathcal{V}_k^2(W_\mu)$.

By the definition of the reproducing kernel (2.1) and formula (2.2), it follows that the above orthonormal basis satisfies

$$\sum_{\varepsilon=0,1}\sum_{0\le 2l\le k} P_{l,\varepsilon}^k(x,y)P_{l,\varepsilon}^k(\cos\phi,\sin\phi) = \frac{k+\lambda}{\lambda}C_k^\lambda(\phi;x,y), \tag{2.10}$$

where $\lambda = \mu + 1/2$. This formula will play an important role below. It shows, in particular, the expansion of $C_k^{\mu+1/2}(\phi;x,y)$ in terms of our orthonormal basis. The following lemma shows the converse.

LEMMA 2.4. *Let $\theta_{j,k} = j\pi/(k+1)$. Then for $0 \le 2l \le k$ if $\varepsilon = 0$ and $0 \le 2l \le k-1$ if $\varepsilon = 1$,*

$$\frac{1}{k+1}\sum_{j=0}^{k} S_{k-2l,\varepsilon}(\theta_{j,k})C_k^{\mu+1/2}(\theta_{j,k};x,y) = \frac{\mu+\frac{1}{2}}{k+\mu+\frac{1}{2}}H_{l,k}^\mu d_{l,k}P_{l,\varepsilon}^k(x,y),$$

where $d_{l,k} = 1/2$ if $2l < k$ and $d_{l,k} = 1$ if $2l = k$, $H_{l,k}^\mu := h_{l,k}^\mu p_l^{(\mu+1/2,k-2l)}(1)$ and

$$\left[H_{l,k}^\mu\right]^2 = \frac{(\mu+\frac{1}{2})_l(\mu+\frac{3}{2})_{k-l}(k+\mu+\frac{3}{2})}{l!(k-l)!(k-l+\mu+\frac{3}{2})}.$$

Proof. Using the identities (2.5) it is easy to verify that

$$\frac{1}{k+1}\sum_{j=0}^{k} S_{k-2l,\varepsilon}(\theta_{j,k})S_{k-2l',\varepsilon}(\theta_{j,k}) = d_{l,k}\delta_{l,l'}. \tag{2.11}$$

Using (2.9) and the fact that $P^k_{l,\varepsilon}(\cos\theta_{j,k}, \sin\theta_{l,k}) = H^\mu_{l,k} S_{k-2l,\varepsilon}(\theta_{j,k})$, we obtain

$$\begin{aligned}
&\frac{1}{k+1}\sum_{j=0}^{k} S_{k-2l,\varepsilon}(\theta_{j,k}) C_k^{\mu+1/2}(\theta_{j,k}; x, y) \\
&\quad = \frac{\mu+\frac12}{k+\mu+\frac12} \sum_{0\le l\le 2k} P^k_{l,\varepsilon}(x,y) \frac{1}{k+1}\sum_{l=0}^{k} P^k_{l,\varepsilon}(\cos\theta_{j,k}, \sin\theta_{j,k}) S_{k-2l,\varepsilon}(\theta_{j,k}) \\
&\quad = \frac{\mu+\frac12}{k+\mu+\frac12} H^\mu_{l,k} d_{l,k} P^k_l(x,y)
\end{aligned}$$

upon using (2.11). Finally, the expression of $[H^\mu_{l,k}]^2$ is derived from the well-known formula of $p_l^{\alpha,\beta}(1)$ (see [11]) and the formula of $h^\mu_{l,k}$. □

LEMMA 2.5. *Let $\theta_{j,k}$ be as above. Then*

$$\frac{1}{k+1}\sum_{j=0}^{k} S_{k-2l,\varepsilon}(\theta_{j,k}) U_k(\theta_{j,k}; \cos\phi, \sin\phi) = d_{l,k} S_{k-2l,\varepsilon}(\phi).$$

Proof. Using (2.6) and the analogous formula for U_{2l-1}, the identity is an easy consequence of (2.11). □

2.2. Attenuated Radon transforms. Let θ be an angle measured counterclockwise from the positive x-axis. Let ℓ denote the line perpendicular to the direction $(\cos\theta, \sin\theta)$ and passing through the point $(t\cos\theta, t\sin\theta)$. The equation of the line is $\ell(\theta, t) = \{(x, y) : x\cos\theta + y\sin\theta = t\}$ for $-1 \le t \le 1$. We use

$$I(\theta, t) = \ell(\theta, t) \cap B^2, \qquad 0 \le \theta < 2\pi, \quad -1 \le t \le 1, \tag{2.12}$$

to denote the line segment of ℓ inside B^2. Let W_μ be the weight function defined in (1.2). The attenuated Radon projection of a function f, with respect to W_μ, in the direction θ with parameter $t \in [-1, 1]$ is defined in (1.3). It can be written as

$$\mathcal{R}^\mu_\theta(f; t) = \int_{-\sqrt{1-t^2}}^{\sqrt{1-t^2}} f(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta) W_\mu(s, t) ds, \tag{2.13}$$

using the fact that the mapping $(s, t) \mapsto (x, y)$ defined by $x = t\cos\theta - s\sin\theta$ and $y = t\sin\theta + s\cos\theta$ amounts to a rotation. When $\mu = 1/2$, this is the usual Radon projection, which is also called an x-ray transform. The definition (1.3) or (2.13) shows that $\mathcal{R}^\mu_\theta(f; t) = \mathcal{R}^\mu_{\pi+\theta}(f; -t)$.

The ridge polynomials are particularly useful for studying Radon transforms, as seen in the following result.

PROPOSITION 2.6. *For $f \in L^1(B^2)$ and $p \in \Pi_k$,*

$$\int_{B^2} f(x, y) p(\phi; x, y) W_\mu(x, y) dx\, dy = \int_{-1}^{1} \mathcal{R}^\mu_\phi(f; t) p(t) dt. \tag{2.14}$$

Proof. Since the change of variables $t = x\cos\phi + y\sin\phi$ and $s = -x\sin\phi + y\cos\phi$

amounts to a rotation, we have

$$\begin{aligned}\int_{B^2} & f(x,y)p_k(\phi;x,y)W_\mu(x,y)dx\,dy \\ &= \int_{B^2} f(t\cos\phi - s\sin\phi, t\sin\phi + s\cos\phi)p_k(t)W_\mu(t,s)dt\,ds \\ &= \int_{-1}^{1}\int_{-\sqrt{1-t^2}}^{\sqrt{1-t^2}} f(t\cos\phi - s\sin\phi, t\sin\phi + s\cos\phi)W_\mu(t,s)ds p_k(t)dt,\end{aligned}$$

and the inner integral is exactly $\mathcal{R}^\mu_\phi(f;t)$ by (2.13). ∎

In particular, attenuated Radon transforms of the orthogonal polynomials in $\mathcal{V}^2_n(W_\mu)$ can be explicitly computed.

LEMMA 2.7. *If $P \in \mathcal{V}^2_k(W_\mu)$, then for each $t \in (-1,1)$, $0 \le \theta \le 2\pi$,*

$$\mathcal{R}^\mu_\theta(P;t) = b_\mu(1-t^2)^\mu \frac{C_k^{\mu+1/2}(t)}{C_k^{\mu+1/2}(1)} P(\cos\theta,\sin\theta), \tag{2.15}$$

where $b_\mu = c_\mu^{-1}$ for c_μ defined in (2.3).

Proof. Changing variables in (2.13) shows that

$$\begin{aligned} Q(t) &:= (1-t^2)^{-\mu}\mathcal{R}^\mu_\theta(P;t) \\ &= \int_{-1}^{1} P\left(t\cos\theta - s\sqrt{1-t^2}\sin\theta, t\sin\theta + s\sqrt{1-t^2}\cos\theta\right)(1-s^2)^{\mu-1/2}ds.\end{aligned}$$

Since an odd power of $\sqrt{1-t}$ in the integrand is always attached to an odd power of s, which has integral zero, $Q(t)$ is a polynomial of t of degree at most k. Furthermore, the integral shows that $Q(1) = b_\mu P(\cos\theta,\sin\theta)$. Equation (2.14) in Proposition 2.6 shows that

$$\int_{-1}^{1} \frac{\mathcal{R}^\mu_\theta(P;t)}{(1-t^2)^\mu} C_j^{\mu+1/2}(t)(1-t^2)^\mu dt = \int_{B^2} P(x,y)C_j^{\mu+1/2}(\theta;x,y)dx\,dy = 0$$

for $j = 0,1,\ldots,k-1$, since $P \in \mathcal{V}_k(B^2)$. In particular, this shows that $Q(t)$ is in fact orthogonal to all polynomials in Π_{k-1} with respect to the weight function $(1-t^2)^\mu$ on $[-1,1]$. Since Q is of degree k, it must be an orthogonal polynomial of degree k with respect to this weight function. Hence, we conclude that $Q(t) = cC_k^{\mu+1/2}(t)$ for some constant c independent of t. Setting $t = 1$ shows that $c = b_\mu P(\cos\theta,\sin\theta)/C_k^{\mu+1/2}(1)$. ∎

In the case of $\mu = 1/2$, the above lemma appeared first in [6].

2.3. Orthogonal expansion and attenuated Radon projections. The standard Hilbert space theory shows that any function in $L^2(W_\mu;B^2)$ can be expanded as a Fourier orthogonal series in terms of $\mathcal{V}^2_n(W_\mu)$. More precisely,

$$L^2(W_\mu;B^2) = \sum_{k=1}^{\infty}\oplus \mathcal{V}^2_k(W_\mu): \qquad f = \sum_{k=1}^{\infty} \operatorname{proj}^\mu_k f, \tag{2.16}$$

where $\operatorname{proj}^\mu_k f$ is the orthogonal projection of f from $L^2(W_\mu;B^2)$ onto the subspace

$\mathcal{V}_k^2(W_\mu)$. It is well known that $\operatorname{proj}_k^\mu f$ can be written as an integral operator in terms of the reproducing kernel $P_k(W_\mu;\cdot,\cdot)$ of $\mathcal{V}_k(B^2)$ in $L^2(B^2)$; that is,

$$\operatorname{proj}_k^\mu f(\mathbf{x}) = \int_{B^2} P_k(W_\mu;\mathbf{x},\mathbf{y}) f(\mathbf{y}) W_\mu(\mathbf{y}) d\mathbf{y}, \tag{2.17}$$

where $\mathbf{x} = (x_1, x_2)$ and $\mathbf{y} = (y_1, y_2)$.

This formula plays an essential role in studying the convergence behavior of the orthogonal expansions; see, for example, [13, 15]. For our purposes, we need a different expression for $\operatorname{proj}_k f$. This is the following remarkable formula that relates $\operatorname{proj}_k f$ to the attenuated Radon transforms of f directly. Let

$$\xi_\nu = \frac{\nu\pi}{n+1}, \qquad 0 \le \nu \le n.$$

THEOREM 2.8. *For $n \ge 0$ and $k \le n$, the operator $\operatorname{proj}_k^\mu f$ can be written as*

$$\operatorname{proj}_k^\mu f(x,y) = \frac{1}{n+1} \sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}^\mu(f;t) D_k^{\mu+1/2}(\xi_\nu, t; x, y) dt \tag{2.18}$$

$$= \frac{1}{2n+2} \sum_{\nu=0}^{2n+1} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}^\mu(f;t) D_k^{\mu+1/2}(\xi_\nu, t; x, y) dt, \tag{2.19}$$

where

$$D_k^{\mu+1/2}(\xi, t; x, y) = \frac{k+\mu+1/2}{\mu+1/2} C_k^{\mu+1/2}(t) D_k^{\mu+1/2}(\xi; x, y) \tag{2.20}$$

with $\lambda_{l,k}^\mu = [H_{l,k}^\mu]^{-2}$ and

$$D_k^{\mu+1/2}(\xi_\nu; x, y) := \sum_{l=0}^{k} \lambda_{l,k}^\mu P_l^k(\cos\xi_\nu, \sin\xi_\nu) P_l^k(x,y).$$

Proof. Since $C_k^{\mu+1/2}$ is symmetric with respect to the origin, using Propositions 2.1 and 2.6, we have

$$\begin{aligned}
a_\mu &\int_{B^2} f(x,y) C_k^{\mu+1/2}(\theta_{j,k}; x, y) W_\mu(x,y) dx\, dy \\
&= \frac{1}{n+1} \sum_{\nu=0}^{n} U_k(\xi_\nu; \cos\theta_{j,k}, \sin\theta_{j,k}) \\
&\quad \times a_\mu \int_{B^2} f(x,y) C_k^{\mu+1/2}(\xi_\nu; x, y) W_\mu(x,y) dx\, dy \\
&= \frac{1}{n+1} \sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}(f;t) C_k^{\mu+1/2}(t) dt U_k(\xi_\nu; \cos\theta_{j,k}, \sin\theta_{j,k}).
\end{aligned}$$

Using Lemmas 2.4 and 2.5 we conclude that

$$\begin{aligned}
a_\mu &\int_{B^2} f(x,y) P_{l,\varepsilon}^k(x,y) W_\mu(x,y)\,dx\,dy \\
&= \frac{1}{n+1}\sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}(f;t) C_k^{\mu+1/2}(t)\,dt \frac{k+\mu+\frac{1}{2}}{\mu+\frac{1}{2}} [H_{l,k}^\mu]^{-1} \\
&\quad \times d_{l,k}^{-1} \frac{1}{k+1}\sum_{j=0}^{k} S_{k-2l,\varepsilon}(\theta_{j,k}) U_k(\xi_\nu; \cos\theta_{j,k}, \sin\theta_{j,k}) \\
&= \frac{1}{n+1}\sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}(f;t) C_k^{\mu+1/2}(t)\,dt \frac{k+\mu+\frac{1}{2}}{\mu+\frac{1}{2}} [H_{l,k}^\mu]^{-1} S_{k-2l,\varepsilon}(\xi_\nu).
\end{aligned}$$

Multiplying by $P_{l,\varepsilon}^k(x,y)$ and summing up, it follows from the definition of the reproducing kernel that

$$\begin{aligned}
\operatorname{proj}_k^\mu f(x,y) &= \frac{1}{n+1}\sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}(f;t) C_k^{\mu+1/2}(t)\,dt \frac{k+\mu+\frac{1}{2}}{\mu+\frac{1}{2}} \\
&\quad \times \sum_{l=0}^{k} [H_{l,k}^\mu]^{-1} \left[S_{k-2l,0}(\xi_\nu) P_{l,0}^k(x,y) + S_{k-2l,1}(\xi_\nu) P_{l,1}^k(x,y) \right] \\
&= \frac{1}{n+1}\sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_\nu}(f;t) C_k^{\mu+1/2}(t)\,dt \frac{k+\mu+\frac{1}{2}}{\mu+\frac{1}{2}} D_k^{\mu+1/2}(\xi_\mu; x, y),
\end{aligned}$$

since $P_{l,\varepsilon}^k(\cos\xi_\nu, \sin\xi_\nu) = H_{l,k}^\mu S_{k-2l,\varepsilon}(\xi_\nu)$ and $\lambda_{l,k}^\mu = [H_{l,k}^\mu]^{-2}$. This proves the first identity.

We now prove the second equation, (2.19). Using the fact that $\xi_{n+\nu+1} = \xi_\nu + \pi$,

$$\cos(k-2l)\xi_{n+\nu+1} = (-1)^k \cos(k-2l)\xi_\nu, \quad \sin(k-2l)\xi_{n+\nu+1} = (-1)^k \sin(k-2l)\xi_\nu,$$

we conclude that $D_k^{\mu+1/2}(\xi_\nu; x, y) = (-1)^k C_k^{\mu+1/2}(\xi_{n+1+\nu}; x, y)$. Hence, using the fact that $\mathcal{R}_{\xi_\nu+\pi}^\mu(f;t) = \mathcal{R}_{\xi_\nu}^\mu(f;-t)$, we conclude that

$$\operatorname{proj}_k^\mu f(x,y) = \frac{1}{n+1}\sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} \mathcal{R}_{\xi_{n+1+\nu}}^\mu(f;t) D_k^{\mu+1/2}(\xi_{n+1+\nu}, t; x, y)\,dt.$$

Adding this equation and the first equation of (2.18) and dividing the result by 2, we then have (2.19). □

In the case of $\mu = 1/2$, it is easy to see that $\lambda_{l,k}^{\frac{1}{2}} = 1/(k+1)$, independent of l. Hence, for $\mu = 1/2$, (2.10) shows that

$$D_k^{\mu+1/2}(\xi_\nu; x, y) = \frac{1}{k+1}(k+1) C_k^1(\xi_\nu; x, y) = U_k(\xi_\nu; x, y),$$

and formulas (2.18) and (2.19) are of a particular simple form. This case was studied in [16].

The two expressions of $\operatorname{proj}_k f$ look similar but are different in one important point: the first expression consists of Radon projections in equally spaced directions along half of the circumference of the circle, while the second expression uses Radon

projections in equally spaced directions along the entire circumference of the circle. This distinction is meaningful for reconstruction algorithms for Radon data.

If n is even, then we can use Proposition 2.2 instead of Proposition 2.1 in the proof. The result is another identity that uses Radon projections over equally spaced angles in $[0, 2\pi]$. Let

$$\phi_\nu = \frac{2\nu\pi}{2m+1}, \qquad 0 \le \nu \le 2m.$$

THEOREM 2.9. *For $m \ge 0$ and $k \le 2m$, the operator $\operatorname{proj}_k^\mu f$ can be written as*

$$\operatorname{proj}_k^\mu f(x,y) = \frac{1}{2m+1} \sum_{\nu=0}^{2m} a_\mu \int_{-1}^1 \mathcal{R}_{\phi_\nu}^\mu(f;t) D_k^{\mu+1/2}(\phi_\nu, t; x, y)dt. \tag{2.21}$$

This expression of $\operatorname{proj}_k f$ is not a special case of (2.19), even though both use equally spaced angles. In fact, setting $n = 2m$ shows that (2.19) uses exactly twice as many Radon projections in equally spaced directions. For $\mu = 1/2$ the identity (2.21) has appeared in [16]. Equation (2.21) can be deduced from (2.18) by using the fact that $\mathcal{R}_{\phi+\pi} f(t) = \mathcal{R}_\phi(f; -t)$ and changing variable $t \mapsto -t$ in the integral whenever $\phi = \xi_{2\nu-1}$ in (2.18), then making use of the equations in (2.7) and the fact that the Gegenbauer polynomial is symmetric.

Let $S_n^\mu f$ denote the nth partial sum of the expansion (2.16); that is,

$$S_n^\mu(f; x, y) = \sum_{k=0}^n \operatorname{proj}_k^\mu f(x,y).$$

The operator S_n^μ is a projection operator from $L^2(W_\mu; B^2)$ onto Π_n^2. An immediate consequence of Theorem 2.8 is the following corollary.

COROLLARY 2.10. *For $n \ge 0$, the partial sum operator $S_n^\mu f$ can be written as*

$$\begin{aligned} S_n^\mu(f; x, y) &= \frac{1}{n+1} \sum_{\nu=0}^n a_\mu \int_{-1}^1 \mathcal{R}_{\xi_\nu}^\mu(f;t) \Phi_n^\mu(\xi_\nu, t; x, y)dt \\ &= \frac{1}{2n+2} \sum_{\nu=0}^{2n+1} a_\mu \int_{-1}^1 \mathcal{R}_{\xi_\nu}^\mu(f;t) \Phi_n^\mu(\xi_\nu, t; x, y)dt, \end{aligned} \tag{2.22}$$

where

$$\Phi_n^\mu(\xi, t; x, y) = \sum_{k=0}^n \frac{k+\mu+1/2}{\mu+1/2} C_k^{\mu+1/2}(t) D_k^{\mu+1/2}(\xi; x, y). \tag{2.23}$$

Likewise, an immediate consequence of Theorem 2.9 is the following corollary.

COROLLARY 2.11. *For $m \ge 0$, the partial sum operator $S_{2m}^\mu f$ can be written as*

$$S_{2m}^\mu(f; x, y) = \frac{1}{2m+1} \sum_{\nu=0}^{2m} a_\mu \int_{-1}^1 \mathcal{R}_{\phi_\nu}^\mu(f;t) \Phi_{2m}^\mu(\phi_\nu, t; x, y)dt. \tag{2.24}$$

2.4. Discretization and reconstruction algorithm. Equation (2.22) expresses the partial sum of the Fourier orthogonal expansion as the integrals of attenuated Radon projections in the equally spaced directions. In order to derive an

algorithm that uses only values of attenuated Radon projections on a set of finite line segments, we approximate the integrals by a quadrature formula. If f is a polynomial, then $\mathcal{R}^\mu_\phi(f;t)/(1-t^2)^\mu$ is a polynomial of the same degree by Lemma 2.7, which shows that we should use a quadrature formula with respect to the weight function $(1-t^2)^\mu$; that is,

$$\int_{-1}^{1} g(t)(1-t^2)^\mu dt \approx \sum_{j=1}^{N} \lambda_j g(t_j),$$

where t_j are real numbers and λ_j are chosen so that the quadrature produces exact values of the integrals for polynomials of degree at least M. Such a quadrature is said to be of N points and of precision M. A Gaussian quadrature of N points has the highest precision $M = 2N-1$ among all quadrature formulas of N points.

For our purpose we are interested in quadrature formulas of precision $2n$ that use $n+1$ points. A class of such formulas is given in the following proposition, which is based on the zeros of the quasi-orthogonal polynomial $C_{n+1}^{\mu+1/2}(t) + aC_n^{\mu+1/2}(t)$, where a is a real number [11]. For a certain range of a, such a polynomial has $n+1$ real distinct zeros in the interval $[-1,1]$.

PROPOSITION 2.12. *Let $t_{j,n}$, $0 \le j \le n$, be the distinct zeros of a quasi-orthogonal polynomial $C_{n+1}^{\mu+1/2}(t) + aC_n^{\mu+1/2}(t)$. Then there are positive numbers $\lambda_{j,n}$ such that the quadrature*

$$\int_{-1}^{1} g(t)(1-t^2)^\mu dt \approx \sum_{j=0}^{n} \lambda_{j,n} g(t_{j,n}) := \mathcal{I}_n^\mu(g) \tag{2.25}$$

has precision $2n$ if $a \neq 0$. If $a = 0$, then the quadrature has precision $2n+1$.

Using an appropriate quadrature on the integrals in (2.22) we obtain a reconstruction algorithm for the attenuated Radon data. We state such an algorithm only in the case of the quadrature formula in (2.25).

ALGORITHM 2.13. *Let $\mu \ge 0$ and $n \ge 0$. Let $t_{j,n}$ and $\lambda_{j,n}$ be as in* (2.25). *For $(x,y) \in B^2$ define*

$$\mathcal{A}_n^\mu(f;x,y) = \sum_{\nu=0}^{n} \sum_{j=0}^{n} \mathcal{R}^\mu_{\xi_\nu}(f;t_{j,n}) T^\mu_{j,\nu}(x,y), \tag{2.26}$$

where

$$T^\mu_{j,\nu}(x,y) = \frac{a_\mu \lambda_{j,n}}{n+1} (1-t_{j,n}^2)^{-\mu} \Phi_n^\mu(\xi_\nu, t_{j,n}; x, y).$$

For a given f, the approximation process $\mathcal{A}_n^\mu f$ uses attenuated Radon data

$$\left\{ \mathcal{R}^\mu_{\xi_\nu}(f;t_{j,n}) : 0 \le \nu \le n, \quad 0 \le j \le n \right\}$$

of f. The data consist of Radon projections on $n+1$ equally spaced directions (specified by ξ_ν) along the circumference of a half circle, and there are $n+1$ parallel lines (specified by $t_{j,n}$) in each direction. The algorithm produces a polynomial $\mathcal{A}_n^\mu f$ which is an approximation to f. In the case of $\mu = 1/2$ the algorithm (2.13) appeared earlier in [1]; the connection to the orthogonal partial sums, however, was neither established nor used there.

THEOREM 2.14. *The operator $\mathcal{A}_n^\mu$ is a projection operator on Π_n^2. In other words, $\mathcal{A}_n^\mu f \in \Pi_n^2$ and $\mathcal{A}_n^\mu P = P$ for $P \in \Pi_n^2$.*

Proof. The function $\Phi^\mu(\xi_\nu, t_{j,n}; x, y)$ is evidently an element in Π_n^2. It follows immediately that $\mathcal{A}_n^\mu f \in \Pi_n^2$. By definition, S_n^μ is a projection operator on Π_n^2. The operator $\mathcal{A}_n^\mu f$ is obtained from $S_n^\mu f$ by applying the quadrature (2.25), exactly for polynomials in Π_{2n}^2, on $(1-t^2)^{-\mu}\mathcal{R}_{\xi_\nu}^\mu(f;t)\Phi_n^\mu(\xi_\nu, t; \cdot)$, which is a polynomial of degree $2n$ in t variable by Lemma 2.7 and (2.23) whenever $f \in \Pi_n^2$. Hence, the quadrature (2.25) is exact. Thus, $\mathcal{A}_n^\mu f = S_n^\mu f = f$ if $f \in \Pi_n^2$. ∎

Alternatively, we can use a quadrature formula of proper order on the second expression of (2.22) to derive an algorithm that uses Radon projections on $2n+2$ directions equally distributed along the circumference of the entire circle. Instead of stating such an algorithm we consider the case of $n = 2m$ and use the expression (2.24). This leads to an algorithm that sums over $2m+1$ angles that are equally spaced over $[0, 2\pi]$, as we shall discuss in the following subsection.

2.5. Reconstruction algorithm using attenuated Radon projections. For practical applications in CT, the discretization described in Algorithm 2.13 needs to be further specified or simplified. In fact, one has to take into consideration what scan geometry is used in practice. For example, the zeros of quasi-orthogonal polynomials will not coincide with the discrete measurement of the attenuated Radon projections in the usual scan geometry. If these points were used, then it would be necessary to introduce an interpolation process, which would introduce new errors. As an alternative, we suggest using a different discretization, which amounts to using a different quadrature formula.

For the ordinary Radon projections ($\mu = 1/2$), Gaussian quadrature formulas for the weight function $\sqrt{1-x^2}$ are used for the integrals in (2.24) to generate algorithms. For practical implementation in CT, the quadrature formula

$$\frac{1}{\pi}\int_{-1}^{1} f(t)\frac{dt}{\sqrt{1-t^2}} = \frac{1}{n+1}\sum_{j=0}^{n} f\left(\cos\frac{(2j+1)\pi}{2n+2}\right), \tag{2.27}$$

based on zeros of $T_{n+1}(x) = \cos(n+1)\theta$, $x = \cos\theta$, is used [17]. The reason for such a choice lies in the scanning geometry of the input data. It turns out that for $n = 2m$, such a choice allows us to adopt fan beam geometry and use it as parallel geometry in a straightforward way.

It is possible to use the quadrature formula (2.27) for attenuated Radon transforms $\mathcal{R}_\phi^\mu(f;t)$, especially when μ is a half integer. The resulting $\mathcal{A}_{2m}$ will no longer be a projection operator, but it still reproduces polynomials of degree slightly less than n when μ is a half integer.

ALGORITHM 2.15. *For $m \ge 0$, $(x, y) \in B^2$,*

$$\mathcal{A}_{2m}^\mu(f; x, y) = \sum_{\nu=0}^{2m}\sum_{j=0}^{2m} \mathcal{R}_{\phi_\nu}^\mu\left(f; \cos\psi_j\right) T_{j,\nu}^\mu(x, y), \tag{2.28}$$

where

$$T_{j,\nu}^\mu(x, y) = \frac{\mu + 1/2}{(2m+1)^2}\sin\psi_j \Phi_{2m}^\mu(\phi_\nu, \cos\psi_j; x, y), \qquad \psi_j = \frac{(2j+1)\pi}{4m+2}.$$

The constant $\mu + 1/2$ in $T_{j,\nu}^\mu$ comes from the fact that $a_\mu = (\mu + 1/2)/\pi$.

This algorithm provides an approximation for the reconstruction of a function $f(x, y)$ from a set of attenuated Radon projections

$$\big\{\mathcal{R}^\mu_{\phi_\nu}(f; \cos\psi_j), \quad 0 \le \nu \le 2m, \quad 1 \le j \le 2m\big\}.$$

The set $\{\phi_\nu : 0 \le \nu \le 2m\}$ consists of equally spaced angles along the circumference of the disk. For $\mu = 1/2$ it has appeared in [16]. The advantage of this algorithm lies in the fact that it can be used with attenuated Radon data obtained from the fan beam geometry directly; see the discussion in [17]. The operator, however, reproduces polynomials up to a lower degree.

THEOREM 2.16. *Let μ be a half integer, $\mu + 1/2 \in \mathbb{N}$. Then the operator $\mathcal{A}^\mu_{2m}$ in Algorithm* 2.15 *preserves polynomials of degree $2m - 2\mu$; that is, $\mathcal{A}^\mu_{2m} P = P$ for $P \in \Pi^2_{2m-2\mu}$.*

Proof. The algorithm is obtained by using the Gaussian quadrature formula (2.27) to discretize the integrals in (2.24); that is,

$$\begin{aligned}\int_{-1}^1 \mathcal{R}^\mu_{\phi_\nu}(f;t) C_k^{\mu+1/2}(t)dt &= \int_{-1}^1 \frac{\mathcal{R}^\mu_{\phi_\nu}(f;t)}{\sqrt{1-t^2}} C_k^{\mu+1/2}(t)\sqrt{1-t^2}dt \\ &\approx \frac{\pi}{2m+1}\sum_{k=0}^{2m} \sin\psi_j \mathcal{R}^\mu_{\phi_\nu}(f;\cos\psi_j) C_k^{\mu+1/2}(\cos\psi_j).\end{aligned}$$

If $f \in \Pi^2_{2m-2\mu}$, then using the fact that $\mathcal{R}^\mu_\phi(f;t)/(1-t^2)^\mu$ is a polynomial of degree $2m - 2\mu$, the assumption that μ is a half integer shows that

$$\mathcal{R}_{\phi_\nu}(f;t)/\sqrt{1-t^2} = (1-t^2)^{\mu-1/2}\mathcal{R}_{\phi_\nu}(f;t)/(1-t^2)^\mu$$

is a polynomial of $2\mu - 1 + 2m - 2\mu = 2m - 1$. Since $\Phi^\mu_{2m}(\xi_\nu, t; \cdot)$ is of degree $2m$ and the quadrature (2.27) is of precision $4m - 1$, the discretization becomes exact in this case and we conclude that $\mathcal{A}^\mu_{2m} f = f$ if $f \in \Pi^2_{2m-2\mu}$. ∎

Let $C(B^2)$ denote the space of continuous function on B^2 with the uniform norm $\|\cdot\|_\infty$ and let $\|\mathcal{A}^\mu_n\|$ denote the operator norm of $\mathcal{A}^\mu_n$ under the uniform norm. By $A \sim B$ we mean that there are two constants c_1 and c_2 such that $c_1 A \le B \le c_2 A$. Evidently the convergence of the algorithm depends on $\|\mathcal{A}^\mu_n\|$. In fact, since $\mathcal{A}^\mu_n$ in Algorithm 2.13 preserves Π_n, it is easy to see that

$$\|f - \mathcal{A}^\mu_n f\| \le c_f\,(1 + \|\mathcal{A}^\mu_n\|)\,E_n(f),$$

where $E_n(f) := \inf\{\|f - P\| : P \in \Pi^2_n\}$ is the error of the best approximation of f by polynomials on B^2. If f has rth order continuous derivatives, then $E_n(f) \le c_f n^{-r}$, in which c_f depends on the norm of the rth derivatives of f. The same applies to $\mathcal{A}^\mu_{2m}$ in Algorithm 2.15, which preserves $\Pi_{2m-2\mu}$. Using the formula in (2.13), the proof of Proposition 5.1 of [16] gives the following formula of the norm of $\mathcal{A}^\mu_{2m}$ in Algorithm 2.15.

PROPOSITION 2.17. *The operator norm $\|\mathcal{A}^\mu_{2m}\|$ of $C(B^2)$ to $C(B^2)$ is given by*

$$\|\mathcal{A}^\mu_{2m}\| = \max_{(x,y)\in B^2} \Lambda_m(x,y), \qquad \Lambda_m(x,y) := \sum_{\nu=0}^{2m}\sum_{j=0}^{2m} (\sin\theta_{j,m})^\mu |T^\mu_{j,\nu}(x,y)|.$$

As $m \to \infty$, the norm grows in an essentially polynomial order of m. Hence, the algorithm converges uniformly if f is sufficiently smooth. Estimating the exact order

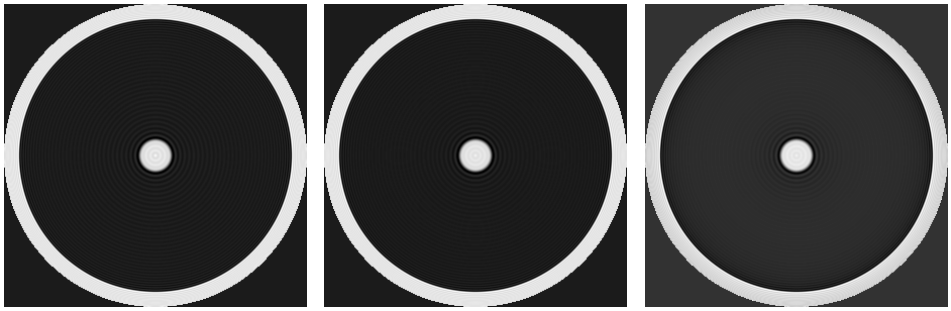


FIG. 1. *From left to right,* $\mu = 0, 1/2, 3/2$.

of $\mathcal{A}_{2m}^{\mu}$ is difficult. In the case of $\mu = 1/2$, it is carried out in [16] and the order is $\|\mathcal{A}_{2m}\| \sim m \log(m+1)$. Based on this fact, we conjecture that the operator norm of $\mathcal{A}_{2m}^{\mu}$ is of the the order

$$\|\mathcal{A}_{2m}^{\mu}\| \sim m^{\mu+1/2} \log(m+1) \qquad \text{as } m \to \infty,$$

which is only slightly worse than the norm $\|S_n^{\mu}\| \sim n^{\mu+1/2}$ (see [15]). If the conjecture holds, then the algorithm will converge uniformly for smooth $f \in C^r(B^2)$ with $r > \mu + 1/2$. In most applications, however, the function or image could have jumps; that is, there is not even continuity. The numerical tests in the case of ordinary Radon data shows that the algorithm is stable and yields fairly accurate results even when the data are highly singular (see [17]). See also the example given in the following subsection.

2.6. Numerical example. For the numerical examples we use Algorithm 2.15, for which the scan geometry is easy to implement. The data required are $g_{j,\nu} := \mathcal{R}_{\phi_\nu}^{\mu}(f; \cos\psi_j)$, where $\phi_\nu = 2\nu\pi/(2m+1)$ stands for the $2m+1$ views equally spaced along the circumference of the region to be reconstructed and $\psi_j = (2j+1)/(4m+2)$ means that the x-rays in each view are distributed according to the zeros of the Chebyshev polynomial T_{2m+1}. In this case the fan data can be resorted into parallel data (see [17]).

We reconstruct a simple analytical phantom defined by the function

$$f(x,y) = \begin{cases} 1 & \text{if } 0.9 \le r \le 1 \ \text{ or } \ 0 \le r \le 0.1, \\ 0 & \text{if } 0.1 < r < 0.9, \end{cases}$$

where $r = \sqrt{x^2+y^2}$ on the unit disk. This phantom contains strong singularity along the circles $r = 0.9$ and $r = 0.1$. The rotationally invariant nature of the function allows certain simplification of the algorithm.

For the reconstruction, we choose three values of the parameter μ, $\mu = 0, 1/2, 3/2$. The case $\mu = 1/2$ means the ordinary Radon transform. The case $\mu = 0$ means that the Radon transform is attenuated by the weight function $(1-x^2-y^2)^{-1/2}$, which is infinity at the boundary of the disk. The case $\mu = 3/2$ means that the Radon transform is attenuated by the weight function $1-x^2-y^2$, which is zero at the boundary. In each case, the Radon data are computed analytically.

For each of the three values of μ, we use Algorithm 2.15 for the reconstruction with a moderate $m = 100$. The reconstructed image is evaluated on a 300×300 grid. The result is shown in Figure 1.

These images show that the function is reconstructed rather faithfully in each of the three cases, even though the function has strong singularity. The case $\mu = 1/2$ has been tested extensively and compared with the FBP (filtered back-projection) algorithm (see [17, 18]). The above is our first attempt to test the algorithm for attenuated Radon transforms.

3. Reconstruction and approximation on the unit sphere. It is known that orthogonal polynomials on the unit ball and on the unit sphere are closely related (see [12]). Since the approximation and the reconstruction in the previous section are based on orthogonal expansions on the unit disk, the relation suggests analogous results on the unit sphere $S^2 = \{(x, y, z) : x^2 + y^2 + z^2 = 1\}$, which we explore in this section.

On the sphere we consider the attenuated spherical transform defined by

$$Q^\mu f(\zeta; t) = \int_{\langle \mathbf{x}, \zeta \rangle = t} f(\mathbf{x}) |x_3|^{2\mu} d\omega,$$

where $\mathbf{x} = (x_1, x_2, x_3) \in S^2$, $\zeta \in \mathbb{R}^3$, and $\xi \neq 0$, and $d\omega$ is the measure on the subset $\{x \in S^2 : \langle \mathbf{x}, \zeta \rangle = t\}$, which is the circle on the sphere. When $\mu = 0$, this is the usual spherical transform (1.5); see, for example, [9, p. 33]. We will mainly work with the case that $\zeta_3 = 0$. We say that a function is even in x_3 if $f(x_1, x_2, x_3) = f(x_1, x_2, -x_3)$.

PROPOSITION 3.1. *Let f be even in x_3. If $\zeta = (\cos\theta, \sin\theta, 0)$, then*

$$Q^\mu f(\zeta; t) = \mathcal{R}^\mu_\theta(F; t), \qquad F(x_1, x_2) = f\Big(x_1, x_2, \sqrt{1 - x_1^2 - x_2^2}\Big). \tag{3.1}$$

Proof. Since f is even in x_3 we have $f(\mathbf{x}) = F(x_1, x_2)$ for $\mathbf{x} \in S^2$. The definition of ζ shows that $\langle \mathbf{x}, \zeta \rangle = x_1 \cos\theta + x_2 \sin\theta = I(\theta, t)$. In terms of x_1 and x_2, $d\omega = dx_1 dx_2 / \sqrt{1 - x_1^2 - x_2^2}$. Thus,

$$Q^\mu f(\zeta; t) = \int_{x_1 \cos\theta + x_2 \sin\theta = t} F(x_1, x_2) \left(1 - x_1^2 - x_2^2\right)^\mu \frac{dx_1 dx_2}{\sqrt{1 - x_1^2 - x_2^2}},$$

which is precisely $\mathcal{R}^\mu_\theta(F; t)$. ∎

Let $H_\mu(\mathbf{x}) = |x_3|^\mu$. The space $L^2(H_\mu; S^2)$ has an orthogonal decomposition

$$L^2(H_\mu; S^2) = \sum_{k=0}^{\infty} \oplus \mathcal{H}^\mu_k, \tag{3.2}$$

where the subspace $\mathcal{H}^\mu_k$ contains homogeneous polynomials of degree k that are orthogonal to lower degree polynomials with respect to $H_\mu d\omega$ on S^2. For $\mu = 0$, $\mathcal{H}^0_k$ is the space of ordinary spherical harmonics. Let

$$\mathrm{proj}_{\mathcal{H}^\mu_k} f : L^2(H_\mu; S^2) \mapsto \mathcal{H}^\mu_k$$

be the orthogonal projection from $L^2(H_\mu; S^2)$ onto $\mathcal{H}^\mu_k$. The space $\mathcal{H}^\mu_k$ is closely related to the space $\mathcal{V}^2_k(W_\mu)$ discussed in the previous section (see [12]). For our purpose, we only need the following relation on the orthogonal projections: if f is even in x_3, then

$$\mathrm{proj}_{\mathcal{H}^\mu_n} f(\mathbf{x}) = \mathrm{proj}^\mu_n F(x_1, x_2), \tag{3.3}$$

where F is the function defined in (3.1). This relation, together with (3.1), allows us to express the projection operator on the sphere in terms of spherical transforms. Using these relations and Theorem 2.8 we obtain the following result.

THEOREM 3.2. *Let f be even in x_3. For $n \ge 0$ and $k \le n$, the operator* $\mathrm{proj}_{\mathcal{H}_k^\mu}$ *can be written as*

$$\mathrm{proj}_{\mathcal{H}_k^\mu} f(\mathbf{x}) = \frac{1}{n+1} \sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} Q^\mu f(\zeta_\nu; t) D_k^{\mu+1/2}(\xi_\nu, t; x_1, x_2) dt, \tag{3.4}$$

where $\xi_\nu = \frac{\nu\pi}{n+1}$, $\zeta_\nu = (\cos\xi_\nu, \sin\xi_\nu, 0)$, and $D_k^{\mu+1/2}(\xi, t; x, y)$ is defined in (2.20).

Let $Y_n^\mu f$ denote the nth partial sum of the expansion (3.2); that is,

$$Y_n^\mu(f; \mathbf{x}) = \sum_{k=0}^{n} \mathrm{proj}_{\mathcal{H}_k^\mu} f(x_1, x_2).$$

The operator Y_n^μ is a projection operator from $L^2(H_\mu; S^2)$ onto $\Pi_n(S^2)$, the restriction of Π_n^3 on S^2. An immediate consequence of Theorem 3.2 is the following.

COROLLARY 3.3. *Let f be even in x_3. For $n \ge 0$, the partial sum operator $Y_n^\mu f$ can be written as*

$$\begin{aligned} Y_n^\mu(f; \mathbf{x}) &= \frac{1}{n+1} \sum_{\nu=0}^{n} a_\mu \int_{-1}^{1} Q^\mu f(\zeta_\nu; t) \Phi_n^\mu(\xi_\nu, t; x_1, x_2) dt \\ &= \frac{1}{2n+2} \sum_{\nu=0}^{2n+1} a_\mu \int_{-1}^{1} Q^\mu f(\zeta_\nu; t) \Phi_n^\mu(\xi_\nu, t; x_1, x_2) dt, \end{aligned} \tag{3.5}$$

where Φ_n^μ is the function defined in (2.23).

For $n = 2m$ we can also use Theorem 2.9 to derive an expression for $Y_{2m}^\mu(f)$, which leads to the following corollary.

COROLLARY 3.4. *Let f be even in x_3. For $m \ge 0$, the partial sum operator $Y_{2m}^\mu f$ can be written as*

$$Y_{2m}^\mu(f; \mathbf{x}) = \frac{1}{2m+1} \sum_{\nu=0}^{2m} a_\mu \int_{-1}^{1} Q^\mu f(\zeta_\nu; t) \Phi_{2m}^\mu(\phi_\nu, t; x_1, x_2) dt, \tag{3.6}$$

where $\phi_\nu = \frac{2\nu\pi}{2m+1}$, $\zeta_\nu = (\cos\phi_\nu, \sin\phi_\nu, 0)$, and Φ_{2m}^μ is the function defined in (2.23).

In the case of $\mu = 0$, equations (3.5) and (3.6) are representations of the partial sums of ordinary spherical harmonic expansions, which are expressed in terms of the Legendre polynomial $P_k(t) = C_k^{1/2}(t)$.

Just like the case of orthogonal expansions on the unit disk, we can use a quadrature formula to obtain a reconstruction algorithm using spherical transforms. For example, using the quadrature formula with respect to $(1-t^2)^\mu$ in Proposition 2.12 as in the case of Algorithm 2.13, we get the following result.

ALGORITHM 3.5. *Let f be even in x_3. Let $\mu \ge 0$. For $n \ge 0$, $\mathbf{x} \in S^2$,*

$$\mathcal{S}_n^\mu(f; \mathbf{x}) = \sum_{\nu=0}^{n} \sum_{j=0}^{n} Q^\mu f(\zeta_\nu; t_{j,n}) T_{j,\nu}^\mu(x_1, x_2), \tag{3.7}$$

where $t_{j,n}$ are as in the quadrature (2.25) *and $T_{j,\nu}^\mu$ are defined in Algorithm* 2.13.

This algorithm reconstructs a function $f(\mathbf{x})$ from a set of spherical transforms

$$\{Q^\mu f(\zeta_\nu; t_j), \quad \zeta_\nu = (\cos\xi_\nu, \sin\xi_\mu, 0), \quad 0 \le \nu \le 2m, \quad 1 \le j \le 2m\},$$

which consists of integrals over a number of circles on the sphere. These circles lie on planes that are parallel to the x_3-axis. The circles intersect the circumference of a disk perpendicular to the x_3-axis at equally spaced angles. The distance between these parallel circles depends on the values of $t_{j,n}$. In the case $\mu = 0$, the algorithm provides an approximation to the function based on ordinary spherical transforms. The assumption that f is even in x_3 implies that we can use the algorithm to reconstruct a function defined on the upper hemisphere from spherical transforms that are integrals over half circles parallel to the x_3-axis on the upper hemisphere.

If μ is a half integer, we can also state an algorithm using the quadrature (2.27), as in Algorithm 2.15, so that $t_{j,n} = \cos j\pi/(2m+1)$. However, in the most interesting case of $\mu = 0$, we do not have such a somewhat simplified algorithm.

4. Reconstruction and approximation on the unit ball. In this section we consider reconstruction of functions on a unit ball B^3 in $\mathbb{R}^3$ based on the attenuated Radon projections.

4.1. Radon projections and orthogonal polynomials. We will work with attenuated Radon projections that are integrals on line segments inside B^3 with respect to the weight function

$$W_\mu(\mathbf{x}) = (1 - \|\mathbf{x}\|^2)^{\mu - 1/2}, \quad \mathbf{x} = (x_1, x_2, x_3) \in B^3, \quad \mu \ge 0.$$

For our purpose, however, we will consider only those lines lying on the planes that are perpendicular to the x_3-axis. Let $x_3 = w$ be such a plane. Its intersection with the unit ball B^3 is a disk $\{\mathbf{x} : x_1^2 + x_2^2 \le \sqrt{1-w^2}, x_3 = w\}$. A line on this disk is given by the equation

$$\ell : \quad x\cos\theta + y\sin\theta = t\sqrt{1-w^2}, \qquad -1 \le t \le 1.$$

Let $I(\theta, w; t)$ denote the intersection of ℓ with B^3. The attenuated Radon projection on such a line is then defined by

$$\mathcal{R}_\theta^\mu(f; t, w) := \int_{I(\theta,w;t)} f(\mathbf{x}) W_\mu(\mathbf{x}) d\mathbf{x}. \tag{4.1}$$

The case $\mu = 1/2$ again corresponds to the usual Radon projection.

LEMMA 4.1. *For $f \in L^1(W_\mu; B^3)$ and for a fixed $w \in [-1, 1]$, define a function g_w on B^2 by*

$$g_w(x, y) = f\left(\sqrt{1-w^2}\, x, \sqrt{1-w^2}\, y, w\right).$$

The x-ray transform (4.1) *is related to the 2D Radon transform* (1.3) *by*

$$\mathcal{R}_\theta^\mu(f; t, w) = (1-w^2)^\mu \mathcal{R}_\theta^\mu(g_w; t). \tag{4.2}$$

Proof. Since $I(\theta, w; t)$ can be represented by

$$x_1 = \sqrt{1-w^2}(t\cos\theta - s\sin\theta), \quad x_2 = \sqrt{1-w^2}(t\sin\theta + t\cos\theta), \quad x_3 = w$$

for $s \in [-\sqrt{1-t^2}, \sqrt{1-t^2}]$, which is a rotation around the x_3-axis on the plane defined by $x_3 = w$, we have

$$\mathcal{R}_\theta(f;t,w) = (1-w^2)^\mu \int_{-\sqrt{1-t^2}}^{\sqrt{1-t^2}} g_w(t\cos\theta - s\sin\theta, t\sin\theta + s\cos\theta) W_\mu(s,t) ds.$$

The integral is precisely $\mathcal{R}_\theta^\mu(g_w;t)$ by (2.13). □

Let $\mathcal{V}_n^3(W_\mu)$ denote the space of orthogonal polynomials with respect to W_μ on B^3, which contains polynomials of degree n that are orthogonal to polynomials of lower degrees with respect to the inner product

$$\langle P, Q\rangle = a_{\mu,3} \int_{B^3} P(\mathbf{x}) Q(\mathbf{x}) W_\mu(x) d\mathbf{x}, \qquad a_{\mu,3} = \frac{\Gamma(\mu+2)}{\pi^{3/2}\Gamma(\mu+1/2)},$$

where $a_{\mu,3}$ is the normalization constant of W_μ. We derive a basis for $\mathcal{V}_n^3(W_\mu)$, making use of an orthogonal basis for $\mathcal{V}_n^2(W_\mu)$. We note that the W_μ in these two notations are different: the first one is on B^3 and the second one is on B^2. We denote by $\widetilde{C}_j^\lambda$ the orthonormal Gegenbauer polynomial, which is equal to $C_n^\lambda/\sqrt{h_n}$ by (2.3).

PROPOSITION 4.2. *Let $\{P_j^k : 0 \le j \le k\}$ be an orthonormal basis for $\mathcal{V}_k^2(W_\mu)$. Then the polynomials*

$$Q_{l,k,j}(x,y,z) = h_k (1-z^2)^{k/2} P_j^k\left(\frac{x}{\sqrt{1-z^2}}, \frac{y}{\sqrt{1-z^2}}\right) \widetilde{C}_{l-k}^{k+\mu+1}(z) \tag{4.3}$$

for $0 \le j \le k \le l$, where $h_k^2 = (\mu+2)_k/(\mu+3/2)_k$, form an orthonormal basis for $\mathcal{V}_l^3(W_\mu)$.

Proof. From Proposition 2.3, it is easy to see that P_j^k is a sum of even powers of homogeneous polynomials when k is even, and a sum of odd powers of homogeneous polynomials when k is odd. Thus, it follows that $Q_{l,k,j} \in \Pi_l^3$. Using the fact that P_j^k is orthonormal, it follows from the integral relation

$$\int_{B^3} f(\mathbf{x}) d\mathbf{x} = \int_{-1}^1 \int_{B^2} f\left(x_1\sqrt{1-x_3^2}, x_2\sqrt{1-x_3^2}, x_3\right) dx_1 dx_2 (1-x_3^2) dx_3 \tag{4.4}$$

and the fact that $a_{\mu,3} = a_\mu c_{\mu+1/2}$, where a_μ is the normalization of W_μ on B^2 and c_μ is defined in (2.3), that

$$\begin{aligned} &a_{\mu,3} \int_{B^3} Q_{l,k,j}(\mathbf{x}) Q_{l',k',j'}(\mathbf{x}) W_\mu(\mathbf{x}) d\mathbf{x} \\ &\quad = h_k^2 c_{\mu+1/2} \int_{-1}^1 \widetilde{C}_{l-k}^{k+\mu+1}(t) \widetilde{C}_{l'-k}^{k+\mu+1}(t) (1-t^2)^{k+\mu+1/2} dt \delta_{k,k'} \delta_{j,j'} \\ &\quad = h_k^2 \frac{c_{\mu+1/2}}{c_{k+\mu+1/2}} \delta_{l,l'} \delta_{k,k'} \delta_{j,j'}. \end{aligned}$$

It follows from the definition of c_μ that $c_{\mu+1/2}/c_{k+\mu+1/2} = (\mu+3/2)_k/(\mu+2)_k$, which completes the proof. □

The attenuated Radon transforms of this basis can be computed explicitly.

PROPOSITION 4.3. *Let $\mu \ge 0$ and let $Q_{l,k,j}$ be defined by (4.3). Then*

$$\frac{\mathcal{R}_\phi^\mu(Q_{l,k,j};t,w)}{(1-t^2)^\mu(1-w^2)^\mu} = b_\mu \frac{C_k^{\mu+1/2}(t)}{C_k^{\mu+1/2}(1)} Q_{l,k,j}\left(\sqrt{1-w^2}\cos\phi, \sqrt{1-w^2}\sin\phi, w\right). \tag{4.5}$$

Proof. By Lemma 4.1 and the definition of $Q_{l,k,j}$ we have

$$\mathcal{R}_\phi^\mu(Q_{l,k,j};t,w) = (1-w^2)^\mu \mathcal{R}_\phi^\mu(g_w;t),$$

where $g_w(x,y) = h_k P_j^k(x,y)(1-w^2)^{k/2}\widetilde{C}_{l-k}^{k+\mu+1}(w)$. By Lemma 2.7, it follows that

$$\begin{aligned}\mathcal{R}_\phi^\mu(g_w;t) &= h_k(1-w^2)^{k/2}\widetilde{C}_{l-k}^{k+\mu+1}(w)\mathcal{R}_\phi^\mu(P_j^k;t)\\ &= b_\mu h_k(1-w^2)^{k/2}\widetilde{C}_{l-k}^{k+\mu+1}(w)(1-t^2)^\mu \frac{C_k^{\mu+1/2}(t)}{C_k^{\mu+1/2}(1)} P_j^k(\cos\phi,\sin\phi)\\ &= b_\mu(1-t^2)^\mu \frac{C_k^{\mu+1/2}(t)}{C_k^{\mu+1/2}(1)} Q_{l,k,j}\left(\sqrt{1-w^2}\cos\phi, \sqrt{1-w^2}\sin\phi, w\right)\end{aligned}$$

by the definition of $Q_{l,k,j}$. Putting these equations together completes the proof. ∎

Let $\mathrm{proj}_{l,3}^\mu$ denote the projection operator from $L^2(W_\mu;B^3)$ onto the space $\mathcal{V}_l^3(W_\mu)$. Again we have the decomposition

$$L^2(W_\mu;B^3) = \sum_{k=0}^\infty \oplus \mathcal{V}_k^3(W_\mu): \qquad f = \sum_{k=0}^\infty \mathrm{proj}_{k,3}^\mu f. \tag{4.6}$$

Proposition 4.4. *For $n \ge 0$ and $0 \le l \le n$,*

$$\begin{aligned}\mathrm{proj}_{l,3}^\mu f(\mathbf{x}) &= \frac{1}{n+1}\sum_{\nu=0}^n \int_{-1}^1\int_{-1}^1 \mathcal{R}_{\xi_\nu}^\mu(f;t,w)G_l(\xi_\nu,t,w;\mathbf{x})dt\sqrt{1-w^2}dw\\ &= \frac{1}{2n+2}\sum_{\nu=0}^{2n+1} \int_{-1}^1\int_{-1}^1 \mathcal{R}_{\xi_\nu}^\mu(f;t,w)G_l(\xi_\nu,t,w;\mathbf{x})dt\sqrt{1-w^2}dw,\end{aligned} \tag{4.7}$$

where

$$\begin{aligned}G_l(\xi,t,w;\mathbf{x}) = a_{\mu,3}\sum_{k=0}^l h_k^2 D_k^{\mu+1/2}\left(\xi,t;\frac{x_1}{\sqrt{1-x_3^2}},\frac{x_2}{\sqrt{1-x_3^2}}\right)\\ \times (1-w^2)^{k/2}(1-x_3^2)^{k/2}\widetilde{C}_{l-k}^{k+\mu+1}(w)\widetilde{C}_{l-k}^{k+\mu+1}(x_3).\end{aligned}$$

Proof. The projection operator has an integral expression just as that of (2.17). Furthermore, the kernel function $P(W_\mu;\mathbf{x},\mathbf{y})$ can be written as a sum of an orthonormal basis. In particular,

$$\mathrm{proj}_{l,3}^\mu f(\mathbf{x}) = \sum_{k=0}^l\sum_{j=0}^k \widehat{f}_{l,k,j}Q_{l,k,j}(\mathbf{x}),$$

where $Q_{l,k,j}$ is the orthonormal basis for $\mathcal{V}_l^3(W_\mu)$ defined in (4.3) and

$$\widehat{f}_{l,k,j} = a_{\mu,3}\int_{B^3} f(\mathbf{y})Q_{l,k,j}(\mathbf{y})W_\mu(\mathbf{y})d\mathbf{y}.$$

Using (4.4), the definition of $Q_{k,l,j}$, and the fact that $a_{\mu,3} = a_\mu c_{\mu+1/2}$, we have

$$\begin{aligned}\widehat{f}_{l,k,j} = c_{\mu+1/2}\int_{-1}^1 \left[a_\mu\int_{B^2} g_w(u,v)P_j^k(u,v)W_\mu(u,v)dudv\right]\\ \times h_k\widetilde{C}_{l-k}^{k+\mu+1}(w)(1-w^2)^{k/2+\mu+1/2}dw,\end{aligned}$$

where g_w is defined as in Lemma 4.1. Hence, it follows from (2.17) and (2.1) that

$$\begin{aligned}\operatorname{proj}_{l,3} f(\mathbf{x}) = &\sum_{k=0}^{l} h_k^2 \widetilde{C}_{l-k}^{k+\mu+1}(x_3)(1-x_3^2)^{k/2} c_{\mu+1/2} \\ &\times \int_{-1}^{1} \operatorname{proj}_k^{\mu} g_w\left(\frac{x_1}{\sqrt{1-x_3^2}}, \frac{x_2}{\sqrt{1-x_3^2}}\right) \widetilde{C}_{l-k}^{k+\mu+1}(w)(1-w^2)^{(k+1)/2+\mu} dw.\end{aligned}$$

The identity (4.7) follows from the above equation upon using (2.18) and (4.2). ∎

Let us denote by $S_{n,3}^{\mu} f$ the nth partial sum of the orthogonal expansion (4.6),

$$S_{n,3}^{\mu} f(\mathbf{x}) = \sum_{l=0}^{n} \operatorname{proj}_{l,3}^{\mu} f(\mathbf{x}).$$

As an immediate consequence of Proposition 4.4 we have the following corollary.

COROLLARY 4.5. *For $n \ge 0$,*

$$\begin{aligned}S_{n,3}^{\mu} f(\mathbf{x}) &= \frac{1}{n+1} \sum_{\nu=0}^{n} \int_{-1}^{1} \int_{-1}^{1} \mathcal{R}_{\xi_\nu}^{\mu}(f;t,w) \Phi_n^{\mu}(\xi_\nu, t, w; \mathbf{x}) dt \sqrt{1-w^2} dw \\ &= \frac{1}{2n+2} \sum_{\nu=0}^{2n+1} \int_{-1}^{1} \int_{-1}^{1} \mathcal{R}_{\xi_\nu}^{\mu}(f;t,w) \Phi_n^{\mu}(\xi_\nu, t, w; \mathbf{x}) dt \sqrt{1-w^2} dw,\end{aligned} \tag{4.8}$$

where

$$\Phi_n^{\mu}(\xi, t, w; \mathbf{x}) = \sum_{l=0}^{n} G_l(\xi, t, w; \mathbf{x}).$$

In the case of $n = 2m$ we can use (2.21) instead of (2.18) in the last step of the proof of Proposition 4.4 to get an expression for $\operatorname{proj}_{l,3}^{\mu} f$. The corresponding expression for the partial sum is the following result.

PROPOSITION 4.6. *For $m \ge 0$,*

$$S_{2m,3}^{\mu} f(\mathbf{x}) = \frac{1}{2m+1} \sum_{\nu=0}^{2m} \int_{-1}^{1} \int_{-1}^{1} \mathcal{R}_{\phi_\nu}^{\mu}(f;t,w) \Phi_{2m}(\phi_\nu, t, w; \mathbf{x}) dt \sqrt{1-w^2} dw.$$

From such an expression of $S_{n,3}^{\mu}$ we naturally want to derive an algorithm as in the 2D case. However, there is a problem when we use a quadrature formula. Indeed, in order to obtain an algorithm, we need to discretize the integrals

$$\int_{-1}^{1} \int_{-1}^{1} \mathcal{R}_{\xi_\nu}^{\mu}(f;t,w) \Phi_n^{\mu}(\xi_\nu, t, w; \mathbf{x}) dt \sqrt{1-w^2} dw \tag{4.9}$$

in $S_{n,3}^{\mu} f$ by a quadrature formula. We can use, for example, the quadrature (2.25) of precision $2n$, which we denote by

$$\int_{-1}^{1} f(t)(1-t^2)^{\alpha} dt \approx \sum_{k=0}^{n} \lambda_{k,n}^{\alpha} f(t_{k,n}^{\alpha})$$

to emphasis the dependence of $t_{k,n}$ and $\lambda_{k,n}$ on the weight function. If we follow the 2D case, then (4.5) indicates that we should apply the quadrature with respect to

$(1-t^2)^\mu$ in the t variable, and apply the quadrature with respect to $(1-w^2)^{\mu+1/2}$ in the w variable. The result of using these quadrature formulas gives us the following.

ALGORITHM 4.7. *Let $\mu \ge 0$. For $n \ge 0$, $\mathbf{x} = (x_1, x_2, x_3) \in B^3$,*

$$\mathcal{B}_n^\mu(f;\mathbf{x}) = \sum_{\nu=0}^{n} \sum_{j=0}^{n} \sum_{k=0}^{n} \mathcal{R}_{\phi_\nu}^\mu(f; t_{j,n}^\mu, t_{k,n}^{\mu+1/2}) T_{j,k,\nu}(\mathbf{x}),$$

where

$$T_{j,k,\nu}^\mu(\mathbf{x}) = \frac{\lambda_j^\mu \lambda_k^{\mu+1/2}}{n+1} \Phi_n^\mu(\xi_\nu, t_{j,n}^\mu, t_{k,n}^{\mu+1/2}; \mathbf{x}).$$

However, this is not likely an accurate algorithm. The problem is that the operator $\mathcal{B}_n^\mu$ does not preserve polynomials of degree n. In fact, in order that $\mathcal{B}_n P = P$ for $P \in \Pi_n^3$, we need the discretization of the integrals (4.9) to be exact whenever f is a polynomial of degree at most n. The function

$$F_\mu(t,w) := (1-t^2)^{-\mu}(1-w^2)^{-\mu} \mathcal{R}_{\xi_\nu}^\mu(f;t,w) \Phi_n^\mu(\xi_\nu, t, w; \mathbf{x})$$

is a polynomial of degree $2n$ in variable t whenever f is a polynomial of degree n by the definition of Φ_n^μ and Proposition 4.3, so that the discretization in the t variable is exact. However, the function $F_\mu(t,w)$ is not a polynomial in the w variable. By the definition of $Q_{l,k,j}$ in (4.3), equation (4.5) shows that $F_\mu(t,w)$ with $f = Q_{l,k,j}$ contains $(1-w)^{k/2} \widetilde{C}_{l-k}^{k+\mu+1}(w)$, which is not a polynomial in the w variable if k is odd. The formula of $\Phi_n^\mu(\xi_\nu, t, w; \mathbf{x})$ shows that it is a sum of functions, which is also not a polynomial. This means that the quadrature will not be exact and polynomials are not preserved by $\mathcal{B}_n^\mu$.

An algorithm should have high convergence order if it preserves polynomials up to certain degrees. The fact that $B_n^\mu f$ does not preserve polynomials means that the convergence of the algorithm may not be as desirable.

5. Reconstruction and approximation on the cylinder domain. In contrast to the unit ball in $\mathbb{R}^3$, the reconstruction algorithm on a cylinder domain works well. Let $L > 0$ and let B_L be the cylinder domain defined by

$$B_L = B^2 \times [0, L] = \{(x,y,z) : (x,y) \in B^2, 0 \le z \le L\}.$$

We will show that the partial sum operator of the orthogonal expansions on W_L admits an expression that relates to Radon data, and we will use it to get a reconstruction algorithm.

Let W_μ be defined as in (1.2). Let $W_{\mu,L}$ be the weight function

$$W_{\mu,L}(x,y,z) = W_\mu(x,y) W_L(z), \qquad (x,y,z) \in B_L.$$

We retain the notation $\mathcal{R}_\phi^\mu(g;t)$ for the attenuated Radon projection of a function $g : B^2 \mapsto \mathbb{R}$, as defined in (1.3). For a fixed z in $[0, L]$, we define

$$\mathcal{R}_\phi^\mu(f(\cdot,\cdot,z);t) := \int_{I(\phi,t)} f(x,y,z) W_\mu(x,y)\, dx\, dy, \tag{5.1}$$

which is the attenuated Radon projection of f in a disk that is perpendicular to the z-axis.

We consider the orthogonal polynomials with respect to the inner product

$$\langle f, g\rangle_{B_L} = \frac{1}{\pi}\int_{B_L} f(x,y,z)g(x,y,z)W_{\mu,L}(x,y,z)\,dx\,dy\,dz. \tag{5.2}$$

Let $\mathcal{V}_n^3(W_{\mu,L})$ denote the subspace of orthogonal polynomials of degree n on B_L with respect to the inner product (5.2); that is, $P \in \mathcal{V}_n^3(W_{\mu,L})$ if $\langle P, Q\rangle_{B_L} = 0$ for all polynomial $Q \in \Pi_{n-1}^3$.

Let p_k be the orthonormal polynomial of degree n with respect to W_L on $[0, L]$ and let $\{P_j^k(x,y) : 0 \le j \le k\}$ denote an orthonormal basis of $\mathcal{V}_k^2(W_\mu)$. Since $W_{\mu,L}$ is a product on a product domain, the following proposition is obvious.

PROPOSITION 5.1. *An orthonormal basis for $\mathcal{V}_l^3(W_{\mu,L})$ is given by*

$$\mathbb{P}_l = \left\{P_{l,k,j}^\mu : 0 \le j \le k \le n\right\}, \qquad P_{n,k,j}^\mu(x,y,z) = P_j^k(x,y)p_{n-k}(z).$$

In particular, the set $\{\mathbb{P}_l : 0 \le l \le n\}$ is an orthonormal basis for Π_n^3.

For $f \in L^2(W_{\mu,L}; B_L)$, the Fourier coefficients of f with respect to the orthonormal system $\{\mathbb{P}_l : l \ge 0\}$ are given by

$$\widehat{f}_{l,k,j}^\mu = a_\mu \int_{B_L} f(\mathbf{x})P_{l,k,j}^\mu(\mathbf{x})W_{\mu,L}(\mathbf{x})d\mathbf{x}, \quad 0 \le j \le k \le l.$$

Let $S_{n,L}^\mu f$ denote the Fourier partial sum operator,

$$S_{n,L}^\mu f(\mathbf{x}) = \sum_{l=0}^n \sum_{k=0}^l \sum_{j=0}^k \widehat{f}_{l,k,j}^\mu P_{l,k,j}^\mu(\mathbf{x}).$$

Just like its counterpart in two variables, this is a projection operator. The following is an analogue of Theorem 2.10 for the cylinder domain B_L.

THEOREM 5.2. *For $n \ge 0$,*

$$S_{n,L}^\mu f(\mathbf{x}) = \frac{1}{n+1}\sum_{\nu=0}^n a_\mu \int_{-1}^1 \int_0^L \mathcal{R}_{\xi_\nu}^\mu(f(\cdot,\cdot,w);t)\Phi_n^\mu(\xi_\nu, w, t; \mathbf{x})W_L(w)dw\,dt, \tag{5.3}$$

where

$$\Phi_n^\mu(\xi, w, t; \mathbf{x}) = \sum_{k=0}^n \frac{k+\mu+1/2}{\mu+1/2} D_k^{\mu+1/2}(\xi, t; x_1, x_2)\sum_{l=0}^{n-k} p_l(w)p_l(x_3). \tag{5.4}$$

Proof. By the definition of $\widehat{f}_{l,k,j}^\mu$ we can write

$$\widehat{f}_{l,k,j}^\mu = a_\mu \int_{B^2} f_{l-k}(x,y)P_j^k(x,y)W_\mu(x,y)dx\,dy,$$

where

$$f_{l-k}(x,y) := \int_0^L f(x,y,w)p_{l-k}(w)W_L(w)dw, \qquad l \ge k \ge 0.$$

Consequently, by the definition of proj_k^μ in (2.17), it follows that

$$S_{n,L}^\mu f(\mathbf{x}) = \sum_{l=0}^n \sum_{k=0}^l \mathrm{proj}_k^\mu(f_{l-k}; x_1, x_2)p_{l-k}(x_3).$$

We can then use the expression (2.18) for $\text{proj}_k^\mu f$ and the fact that

$$\mathcal{R}_\xi^\mu(f_{l-k};t) = \int_0^L \mathcal{R}_\xi^\mu(f(\cdot,\cdot,w);t)p_{l-k}(w)W_L(w)dw$$

to complete the proof. □

In the case of $n = 2m$, we can use (2.21) in place of (2.18) in the proof. The result is the following proposition, which has appeared in [16] when $\mu = 1/2$.

PROPOSITION 5.3. *For $m \ge 0$,*

$$S_{2m,L}^\mu f(\mathbf{x}) = \frac{1}{2m+1}\sum_{\nu=0}^{2m} a_\mu \int_{-1}^1 \int_0^L \mathcal{R}_{\xi_\nu}^\mu(f(\cdot,\cdot,w);t)\Phi_{2m}^\mu(\phi_\nu,w,t;\mathbf{x})W_L(w)dw\,dt. \tag{5.5}$$

From the expression (5.3) or (5.5) of $S_{n,L}^\mu f$, we can apply a quadrature formula to get a reconstruction algorithm on B_l for the attenuated Radon data. In [16] the weight function W_L is chosen to be the Chebyshev weight function

$$W_L(z) = \frac{1}{\pi}\frac{1}{\sqrt{z(L-z)}}, \qquad z \in [0,L],$$

normalized to have integral 1 on $[0,L]$. The reason for this choice is that the Gaussian quadrature formula takes a simple form

$$\int_0^L g(z)W_L(z)dz \approx \frac{1}{n+1}\sum_{j=0}^n g(z_i), \qquad z_i = \frac{1}{2}\left(1+\cos\frac{2j+1}{2n+2}\right), \tag{5.6}$$

which is of precision $2n+1$. We can apply this quadrature for the integral with respect to w and use the quadrature (2.25) for the integral with respect to t in (5.3) or (5.5). The result is the following algorithm.

ALGORITHM 5.4. *Let $\mu \ge 0$ and let $\gamma_{\mu,j,i} = \mathcal{R}_{\xi_\nu}^\mu(f(\cdot,\cdot,z_i);t_{j,n})$. For $n \ge 0$*

$$\mathcal{B}_{n,L}^\mu(f;\mathbf{x}) = \sum_{\nu=0}^n\sum_{j=0}^n\sum_{i=0}^n \gamma_{\nu,j,i}T_{\nu,j,i}(\mathbf{x}), \tag{5.7}$$

where

$$T_{\nu,j,i}(\mathbf{x}) = \frac{a_\mu\lambda_{j,n}}{n+1}(1-t_{j,n}^2)^{-\mu}\Phi_n^\mu(\xi_\nu,z_i,t_{j,n};\mathbf{x}).$$

Like the algorithms in the previous sections, this algorithm produces a polynomial as an approximation to the function. It does preserve polynomials of lower degrees.

THEOREM 5.5. *The operator $\mathcal{B}_{n,L}^\mu$ is a projection operator on Π_n^3. In other words, $\mathcal{B}_n f \in \Pi_n^3$ and $\mathcal{B}_{n,L}(f) = f$ if $f \in \Pi_n^3$.*

Proof. Let $P_{n,k,j}^\mu$ be defined as in Proposition 5.1. It follows from the definition in (5.1) that $\mathcal{R}_\phi^\mu(P_{l,k,j}^\mu(\cdot,\cdot,w);t) = \mathcal{R}_\phi^\mu(P_j^k;t)p_{l-k}(w)$. Consequently, it follows from (2.15) that $\mathcal{R}_\phi^\mu(P(\cdot,\cdot,w);t)/(1-t^2)^\mu$ is a polynomial of degree n in both the t variable and the w variable whenever $P \in \Pi_n^3$. By its definition in (5.4), the function $\Phi^\mu(\xi,w,t;\mathbf{x})$ is evidently a polynomial of degree n in both t and w variables. Hence,

we can apply (5.6) for the w variable and apply the quadrature (2.25) of precision $2n$ to the t variable, which are exact on $(1-t^2)^{-\mu}\mathcal{R}^\mu_\phi(P(\cdot,\cdot,w);t)\Phi^\mu(\xi,w,t;\cdot)$. □

The approximation process in Algorithm 5.4 uses the attenuated Radon data

$$\big\{\mathcal{R}^\mu_{\xi_\nu}(f(\cdot,\cdot,z_i);t_{j,n}) : 0 \le \nu \le n,\ 0 \le j \le n,\ 0 \le i \le n\big\},$$

which consists of Radon projections on $n+1$ disks that are parallel to the z-axis. In other words, it consists of reconstructions of the function on $n+1$ planes.

In the case in which $n = 2m$ and μ is a half integer, we can also use the quadrature (2.27) to derive a more explicit algorithm as in Algorithm 2.15. Such an algorithm is given in [17] for $\mu = 1/2$. We shall not elaborate further.

REFERENCES

[1] T. Bortfeld and U. Oelfke, *Fast and exact 2D image reconstruction by means of Chebyshev decomposition and backprojection*, Phys. Med. Biol., 44 (1999), pp. 1105–1120.

[2] C. F. Dunkl and Y. Xu, *Orthogonal Polynomials of Several Variables*, Cambridge University Press, Cambridge, UK, 2001.

[3] D. Finch, *The attenuated x-ray transform: Recent developments*, in Inside Out: Inverse Problems and Applications, Math. Sci. Res. Inst. Publ. 47, Cambridge University Press, Cambridge, UK, 2003, pp. 47–66.

[4] A. C. Kak and M. Slaney, *Principles of Computerized Tomographic Imaging*, IEEE Press, New York, 1988; reprinted as Classics Appl. Math. 33, SIAM, Philadelphia, 2001.

[5] B. Logan and L. Shepp, *Optimal reconstruction of a function from its projections*, Duke Math. J., 42 (1975), pp. 645–659.

[6] R. Marr, *On the reconstruction of a function on a circular domain from a sampling of its line integrals*, J. Math. Anal. Appl., 45 (1974), pp. 357–374.

[7] F. Natterer, *The Mathematics of Computerized Tomography*, B. G. Teubner, Stuttgart, John Wiley, New York, 1986; reprinted as Classics Appl. Math. 32, SIAM, Philadelphia, 2001.

[8] F. Natterer, *Inversion of the attenuated Radon transform*, Inverse Problems, 17 (2001), pp. 113–119.

[9] F. Natterer and F. Wübbeling, *Mathematical Methods in Image Reconstruction*, SIAM, Philadelphia, 2001.

[10] R. G. Novikov, *An inversion formula for the attenuated X-ray transformation*, Ark. Mat., 40 (2002), pp. 145–167.

[11] G. Szegő, *Orthogonal Polynomials*, 4th ed., Amer. Math. Soc. Colloq. Publ. 23, AMS, Providence, 1975.

[12] O. Tischenko, Y. Xu, and C. Hoeschen, *An alternative tomographic reconstruction algorithm*, submitted, 2005.

[13] Y. Xu, *Orthogonal polynomials and cubature formulae on spheres and on balls*, SIAM J. Math. Anal., 29 (1998), pp. 779–793.

[14] Y. Xu, *Summability of Fourier orthogonal series for Jacobi weight on a ball in $\mathbb{R}^d$*, Trans. Amer. Math. Soc., 351 (1999), pp. 2439–2458.

[15] Y. Xu, *Funk-Hecke formula for orthogonal polynomials on spheres and on balls*, Bull. London Math. Soc., 32 (2000), pp. 447–457.

[16] Y. Xu, *Representation of reproducing kernels and the Lebesgue constants on the ball*, J. Approx. Theory, 112 (2001), pp. 295–310.

[17] Y. Xu, *A new approach to the reconstruction of images from Radon projections*, Adv. in Appl. Math., 36 (2006), pp. 388–420.

[18] Y. Xu, O. Tischenko, and C. Hoeschen, *A new reconstruction algorithm for Radon data*, in Medical Imaging 2006: Physics of Medical Imaging, Proc. SPIE 6142, SPIE, Bellingham, WA, 2006, pp. 791–798.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 133–162

COMPOSITE WAVELET BASES WITH EXTENDED STABILITY AND CANCELLATION PROPERTIES*

ROB STEVENSON†

Abstract. The efficient solution of operator equations using wavelets requires that they generate a Riesz basis for the underlying Sobolev space and that they have cancellation properties of a sufficiently high order. Suitable biorthogonal wavelets were constructed on reference domains as the n-cube. Via a domain decomposition approach, these bases have been used as building blocks to construct biorthogonal wavelets on general domains or manifolds, where, in order to end up with local wavelets, biorthogonality was realized with respect to a modified L_2-scalar product. The use of this modified scalar product restricts the application of these so-called composite wavelets to problems of orders strictly larger than -1. Moreover, those wavelets with supports that extend to more than one patch generally have no cancellation properties. In this paper, we construct local, composite wavelets that are close to being biorthogonal with respect to the standard L_2-scalar product. As a consequence, they generate Riesz bases for the Sobolev spaces H^s for the *full range of* s allowed by the continuous gluing of functions over the patch interfaces, the properties of the primal and dual approximation spaces on the reference domain, and, in the manifold case, by the regularity of the manifold. Moreover, all these wavelets have *cancellation properties of the full order* induced by the approximation properties of the dual spaces on the reference domain. We illustrate our findings by a concrete realization of wavelets on a perturbed sphere.

Key words. wavelets, Riesz bases, cancellation properties, domain decomposition, boundary integral equations

AMS subject classifications. 46B15, 46E35, 65N55, 65T60

DOI. 10.1137/060651021

1. Introduction.

The use of wavelet bases for solving operator equations, as *partial differential equations* or *(boundary) integral equations*, has a number of advantages; cf. [9, 3]. Let us assume that the operator is symmetric, and, for H being some Hilbert space, H-bounded and H-coercive, and that the infinite collection of properly scaled wavelets generates a *Riesz basis* for H. Then the stiffness matrix in wavelet coordinates resulting from a Ritz–Galerkin discretization is well conditioned uniformly in its size, guaranteeing a uniform rate of convergence of an iterative method. In case of a differential operator, this stiffness matrix is not truly sparse, but has the well-known "finger structure." For multiplying with this matrix, however, one may switch to a single-scale basis, with respect to which the stiffness matrix is sparse.

For integral operators, the stiffness matrix with respect to both single-scale and wavelet basis is densely populated. Here the second important property of wavelets—that of having vanishing moments or, more generally, *cancellation properties*, meaning that the integral of a wavelet against a smooth function vanishes with a certain order of the length scale of the wavelet—can be exploited. If, depending on the order of the operator and the order of approximation, this order of the cancellation properties is sufficiently large, then the stiffness matrix with respect to the wavelet basis can be a priorily compressed to a sparse one without reducing the order of convergence. With this a method of linear complexity is obtained for solving integral equations [18, 10].

*Received by the editors January 27, 2006; accepted for publication (in revised form) July 11, 2006; published electronically January 12, 2007. This work was supported by the Netherlands Organization for Scientific Research and by the EC-IHP project "Breaking Complexity."

http://www.siam.org/journals/sinum/45-1/65102.html

†Department of Mathematics, Utrecht University, P.O. Box 80.010, NL-3508 TA Utrecht, The Netherlands (stevenson@math.uu.nl).

Instead of projecting the operator equation onto a fixed finite dimensional space, and then solving the resulting matrix-vector problem with an iterative method, the availability of a *Riesz basis* for H opens an attractive alternative for approximating the solution by adaptive wavelet methods [4, 5]. By writing this unknown solution in terms of this basis and testing the equation for all basis functions, one obtains an infinite dimensional matrix-vector problem. This problem is *equivalent* to the operator equation, and it is well-posed in ℓ_2-metric, meaning that it can be solved using an iterative method. In each iteration of such a method, the application of the infinite stiffness matrix to the current approximation vector has to be approximated. Here the concept of *adaptivity* enters; the accuracy with which a column is approximated grows with the modulus of the corresponding entry of the vector. The resulting method, extended with a so-called coarsening routine to remove small entries from the approximation vector, can be proven to be optimal in the following sense. Whenever, for a certain range of s, the solution is in a class of functions for which the error of the best N-term approximations from the wavelet basis decays like N^{-s}, the sequence of approximations produced by this adaptive method has the same rate of convergence, whereas the computational cost is equivalent to their support sizes. A necessary condition for this statement to be true is that the stiffness matrix is sufficiently close to a sparse matrix, which depends on the smoothness of the wavelets and, again, on the *order of the cancellation properties* [21]. Recently, it has been shown that an optimal adaptive wavelet method can even be obtained without coarsening [15].

Aiming at the aforementioned applications, this paper deals with *the construction on general n-dimensional domains or manifolds of wavelets that, properly scaled, generate Riesz bases for a range of Sobolev spaces, and satisfy cancellation properties of any required order*. To be able to choose this order independently from the order of approximation, we will consider biorthogonal wavelets. Their construction starts with two nested sequences of approximation spaces that both satisfy Jackson and Bernstein estimates ("multiresolution analyses"). Then the primal and dual wavelets are sought as bases of the biorthogonal complements of successive approximation spaces at primal and dual side, respectively. In case the primal and dual approximation spaces can be equipped with bases of local, biorthogonal scaling functions, local primal wavelets are found by applying the biorthogonal projector onto a local basis of some complement space of two successive primal approximation spaces. In this case, under some mild additional condition, the corresponding dual wavelets are also local. Actually, for constructing only local primal wavelets, a reduced set of assumptions already suffices, which for simplicity we will ignore in this introduction. Note that in algorithms for solving operator equations, usually dual wavelets do not play any role.

Biorthogonal scaling functions have been constructed on the real line [6] and, as adaptations of these, on the interval [11]. By taking tensor products, one obtains biorthogonal scaling functions on the n-dimensional unit cube. To construct biorthogonal scaling functions and wavelets on general domains and manifolds, a domain decomposition approach has been developed by Dahmen and Schneider in [12] (see [1, 7] for related approaches). The domain or manifold of interest is written as a disjoint union of smooth parametric images of the unit cube. The biorthogonal scaling functions on the cube are lifted to the patches, and, assuming that the decomposition satisfies some matching condition, they are continuously connected over the interfaces. With respect to a *modified L_2-scalar product*, defined by ignoring the Jacobian determinants of the parametrizations in the definition of the canonical L_2-scalar product, the resulting collections of scaling functions are biorthogonal. Wavelets, in this setting called *composite wavelets*, can now be constructed using the biorthogonal

projector. There are, however, two principal limitations related to the realization of biorthogonality with respect to the modified L_2-scalar product. First, wavelets with supports that extend to more than one patch generally have *no cancellation properties* with respect to the canonical L_2-scalar product. So results concerning matrix compression do not apply to entries involving such wavelets. Second, with respect to the interpretation of a wavelet as a functional using the duality pairing in terms of the canonical L_2-scalar product, generally the resulting wavelets *cannot* generate a *Riesz basis for* H^s *for* $s \leq -\frac{1}{2}$. So for operators of order $2s \leq -1$, like the single-layer potential operator, neither are the optimal preconditioning results valid, nor can the adaptive wavelet method be applied.

These limitations were already recognized by the authors in [12]. In [13], they developed an elegant approach to construct wavelets on general domains or manifolds that, properly scaled, generate Riesz bases for H^s for in principal any s, and that have cancellation properties of any required order. Unfortunately, so far with this approach it seems not easy to construct wavelets that have competitive quantitative properties. A recent investigation of this approach was made in [16].

In this paper, we reconsider the approach from [12], except that, in view of the aforementioned limitations, we make use of the canonical L_2-scalar product. Although, generally, the lifted and connected scaling functions are not biorthogonal with respect to this scalar product, we can derive a general formula for the corresponding biorthogonal wavelets. Since this formula, however, involves the inverse of the matrix consisting of the L_2-scalar products between all primal and dual scaling functions (this matrix is thus generally not diagonal), these wavelets have global supports. On the other hand, this matrix is nearly diagonal, so that its inverse can be well approximated by sparse matrices, which gives rise to local, approximate wavelets. We derive general conditions under which, properly scaled, such approximate wavelets generate a Riesz basis for H^s for the *full range of* s allowed by the continuous gluing of the scaling functions over the interfaces, by the properties of the primal and dual approximation spaces on the cube, and, in the manifold case, by the regularity of the manifold. We give three possibilities for the construction of approximate wavelets that are local and generate Riesz bases for H^s for the aforementioned full range of s, and *all* have *cancellation properties of the full order* induced by the approximation properties of the dual spaces on the unit cube. First, we show that the approximation of the inverse of the matrix of L_2-scalar products of primal and dual scaling functions by a suitable, fixed number of Jacobi iterations yields such approximate wavelets. In view of the relatively large supports of these wavelets, second, we show that away from the patch interfaces they can be replaced by the wavelets one gets by ignoring the Jacobian determinants, which are the wavelets from [12]. Third, we show that also along the patch interfaces suitable approximate wavelets with smaller supports can be constructed, which, however, will involve solving some local systems. Although several proofs will be quite involved, we emphasize that the implementation of the approximate wavelets is relatively straightforward.

In [14, 20], we constructed wavelet bases for Lagrange finite element spaces based on a subdivision of polygonal domains into n-simplices. In this paper, we include the option that these finite element wavelets, or more precisely the underlying scaling functions, are used as building blocks for wavelets on general (nonpolygonal) domains or manifolds, where then the unit n-cube as reference domain should be replaced by some reference n-simplex.

This paper is organized as follows. In the remainder of this section we fix a few notations. In section 2, we specify the type of domains and manifolds and their

parametrizations that we will consider. We recall the definition of the Sobolev spaces, that may involve zero order Dirichlet boundary conditions, which we are going to equip with Riesz bases. In section 3, we collect all assumptions on the multiresolution analyses on the reference domain. The induced, continuous, multiresolution analyses on the target domain or manifold are defined in section 4. Although put here into a more general framework, the main construction principles from sections 2–4 originate from [12]. Biorthogonal space decompositions and the, generally, globally supported biorthogonal wavelets are constructed in section 5. Sections 6 and 7, which form the main part of this paper, are devoted to the construction of local, approximate wavelets. Finally, in section 8 we show examples of approximate wavelets on a perturbed sphere and give some numerically computed condition numbers.

In order to limit the size of this paper, for some technical (parts of) proofs we will refer to the extended preprint version [22].

In order to avoid the repeated use of generic but unspecified constants, in this paper by $C \lesssim D$ we mean that C can be bounded by a multiple of D, independently of parameters on which C and D may depend. Obviously, $C \gtrsim D$ is defined as $D \lesssim C$, and $C \eqsim D$ as $C \lesssim D$ and $C \gtrsim D$.

Let H be a separable Hilbert space with scalar product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|$. For a countable collection Σ of functions in H, which we formally view as a (column) vector, and for $\mathbf{c} = (c_\sigma)_{\sigma\in\Sigma}$ a vector of scalars, with $\mathbf{c}^T\Sigma$ we will mean the expansion $\sum_{\sigma\in\Sigma} c_\sigma\sigma$. The span of Σ will be denoted as $\mathcal{S}(\Sigma)$. For $x \in H$, by $\langle\Sigma, x\rangle$ and $\langle x, \Sigma\rangle$ we will mean the column- and row-vectors with coefficients $\langle\sigma, x\rangle$ and $\langle x, \sigma\rangle$, $\sigma \in \Sigma$. When $\tilde\Sigma$ is another countable collection in H, with $\langle\Sigma, \tilde\Sigma\rangle$ we denote the matrix $(\langle\sigma, \tilde\sigma\rangle)_{\sigma\in\Sigma,\tilde\sigma\in\tilde\Sigma}$. For $V \subset H$ being a dense, continuously embedded Banach space, as usual we will use $\langle\cdot,\cdot\rangle$ sometimes also to denote the duality pairing $\langle\cdot,\cdot\rangle_{V\times V'}$, which, with the aforementioned meaning, can also be applied to collections from V and/or V'.

On the spaces of (possibly infinite) scalar vectors or matrices, we will exclusively use the ℓ_2-scalar product, ℓ_2-norm, or the resulting operator norm, that we therefore simply denote by $\langle\cdot,\cdot\rangle$ or $\|\cdot\|$, respectively. A collection Σ is called a *Riesz system* when $\|\mathbf{c}^T\Sigma\| \eqsim \|\mathbf{c}\|$, i.e., when $\langle\Sigma,\Sigma\rangle$ is boundedly invertible, and Σ is called a *Riesz basis* when it is in addition a basis for H. When Σ depends on a parameter, we will speak about *uniform Riesz systems* (or bases) when the above equivalence holds uniformly over the values this parameter may attain. We set $\|\Sigma\| = \|\langle\Sigma,\Sigma\rangle\|^{\frac12}$ and collect a few properties related to this definition.

PROPOSITION 1.1.
(i) $\sup_{\mathbf{c}\neq0} \frac{\|\mathbf{c}^T\Sigma\|}{\|\mathbf{c}\|} = \|\Sigma\|$,
(ii) $\|\langle\Sigma,\tilde\Sigma\rangle\| \le \|\Sigma\|\|\tilde\Sigma\|$,
(iii) $\|\Sigma+\tilde\Sigma\| \le \|\Sigma\| + \|\tilde\Sigma\|$,
(iv) *for a matrix* $\mathbf{A}$, $\|\mathbf{A}\Sigma\| \le \|\mathbf{A}\|\|\Sigma\|$.

Proof. For (i), use $\|\mathbf{c}^T\Sigma\|^2 = \langle\langle\Sigma,\Sigma\rangle\mathbf{c},\mathbf{c}\rangle$. Part (ii) follows from $|\langle\langle\Sigma,\tilde\Sigma\rangle\mathbf{c},\tilde{\mathbf{c}}\rangle| = |\langle\mathbf{c}^T\Sigma,\tilde{\mathbf{c}}^T\tilde\Sigma\rangle| \le \|\mathbf{c}\|\|\tilde{\mathbf{c}}\|\|\Sigma\|\|\tilde\Sigma\|$ because of (i). Part (iii) follows easily from (ii). For (iv), use $\langle\mathbf{A}\Sigma,\mathbf{A}\Sigma\rangle = \mathbf{A}\langle\Sigma,\Sigma\rangle\mathbf{A}^*$. ∎

2. Domains and function spaces. For some $n' \ge n \ge 1$, let Γ be an n-dimensional bounded manifold in $\mathbb{R}^{n'}$, with or without a boundary. For $\square$ denoting the interior either of the n-cube $[0,1]^n$ or, despite its notation, of some reference n-simplex, we assume that Γ is given as

$$\overline{\Gamma} = \cup_{q=1}^M \overline{\Gamma_q},\ \text{with } \Gamma_q \cap \Gamma_{q'} = \emptyset \text{ when } q \neq q',\ \text{and } \Gamma_q = \kappa_q(\square),$$

where $\kappa_q : \mathbb{R}^n \to \mathbb{R}^{n'}$ are some smooth, regular parametrizations. With Π we will denote the collection of all affine mappings from $\square$ onto $\square$. So in case $\square$ is the interior of an n-simplex, this collection consists of the permutations of the $n+1$ barycentric coordinates, and otherwise it consists of the compositions of any permutation of the n Cartesian coordinates and reflections of type $y \mapsto (y_1, \ldots, y_{i-1}, 1-y_i, y_{i+1}, \ldots, y_n)$ $(1 \le i \le n)$. We assume that the splitting of Γ into the patches Γ_q is conforming in the sense that for any $q \neq q'$, either $\overline{\Gamma_q} \cap \overline{\Gamma_{q'}}$ is empty or

$$\kappa_q^{-1}(\overline{\Gamma_q} \cap \overline{\Gamma_{q'}}) \text{ is a face of } \square.$$

In addition, we assume that the parametrizations can be chosen such that the following matching condition is satisfied: There exists a $\pi \in \Pi$ with

$$\kappa_{q'} \circ \pi \circ \kappa_q^{-1} = Id \quad \text{on } \overline{\Gamma_q} \cap \overline{\Gamma_{q'}}. \tag{$\mathcal{M}$}$$

Here and in the remainder of this paper, by a "face" of $\square$, we mean a (complete, closed) face of any dimension $0 \le k \le n-1$; i.e., for $n=3$, it is a vertex, an edge, or a facet. Note that our setting allows Γ to be a bounded domain in $\mathbb{R}^n$, as well as an open or closed bounded manifold in $\mathbb{R}^{n'}$ for some $n' > n$.

We include the possibility that homogeneous, zero order Dirichlet boundary conditions are prescribed on some part $\partial\Gamma_D \subset \overline{\Gamma}\backslash\Gamma$, for which, for all $1 \le q \le M$,

$$\kappa_q^{-1}(\partial\Gamma_D \cap \overline{\Gamma}_q) \text{ is a, possibly empty, union of faces of } \square; \tag{2.1}$$

see Figure 2.1.

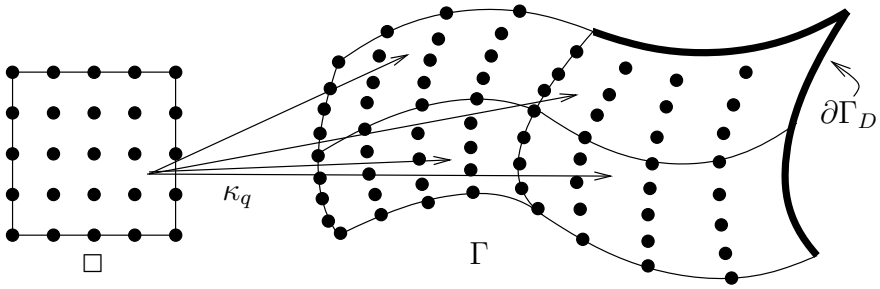


FIG. 2.1. *Illustration of the domain decomposition approach.*

For some $s_\Gamma > 0$, we assume that, globally,

$$\Gamma \in C^{s_\Gamma} \text{ when } s_\Gamma \notin \mathbb{N}, \text{ or } \Gamma \in C^{s_\Gamma - 1,1} \text{ when } s_\Gamma \in \mathbb{N}.$$

This means that for $0 \le s < s_\Gamma \notin \mathbb{N}$, or $0 \le s \le s_\Gamma \in \mathbb{N}$, the Sobolev spaces

$$\mathcal{H}^s(\Gamma) := \begin{cases} H^s_{0,\partial\Gamma_D}(\Gamma) & \text{when } s \le 1, \\ H^s(\Gamma) \cap H^1_{0,\partial\Gamma_D}(\Gamma) & \text{when } s > 1 \end{cases}$$

can be defined in the usual way using a partition of unity relative to some atlas. For $s > 0$ in the above range, $\mathcal{H}^{-s}(\Gamma)$ will be understood as being the dual of $\mathcal{H}^s(\Gamma)$.

With μ being the induced Lebesgue measure on Γ, the inner product on $L_2(\Gamma)$ is given by

$$\langle u, v\rangle_{L_2(\Gamma)} = \int_\Gamma u\overline{v}d\mu = \sum_{q=1}^{M} \langle u \circ \kappa_q, v \circ \kappa_q\rangle_{L_2(\square),|\partial\kappa_q|}. \tag{2.2}$$

Here, for $w \in L^\infty(\square)$ with $w > 0$ a.e., $\langle f, g\rangle_{L_2(\square),w} := \int_\square f(y)\overline{g(y)}w(y)dy$, and $|\partial\kappa_q| : z \mapsto |\partial\kappa_q(z)|$ are the Jacobian determinants of the parametrizations. We will also make use of a modified inner product

$$\langle\!\langle u, v\rangle\!\rangle_0 := \sum_{q=1}^M \langle u\circ\kappa_q, v\circ\kappa_q\rangle_{L_2(\square)}, \tag{2.3}$$

which is the inner product one gets by ignoring the Jacobian determinants. It is equivalent to $\langle\,,\,\rangle_{L_2(\Gamma)}$ in the sense that $|\!|\!|\cdot|\!|\!|_0 := \langle\!\langle\cdot,\cdot\rangle\!\rangle_0^{\frac12} \eqsim \|\cdot\|_{L_2(\Gamma)}$. More generally, for any $s \geq 0$, we define

$$\langle\!\langle u, v\rangle\!\rangle_s = \sum_{q=1}^M \langle u\circ\kappa_q, v\circ\kappa_q\rangle_{H^s(\square)},$$

and let $\mathcal{H}_s(\Gamma)$ denote the closure with respect to $|\!|\!|\cdot|\!|\!|_s := \langle\!\langle\cdot,\cdot\rangle\!\rangle_s^{\frac12}$ of the set all *globally continuous*, and with respect to the subdivision $\overline{\Gamma} = \cup_{q=1}^M \overline{\Gamma_q}$, *piecewise* C^∞ functions on Γ that are *zero on* $\partial\Gamma_D$. For $s > 0$, we define $\mathcal{H}_{-s}(\Gamma) = (\mathcal{H}_s(\Gamma))'$. For $0 \leq s < s_\Gamma \notin \mathbb{N}$ or $0 \leq s \leq s_\Gamma \in \mathbb{N}$, it holds that $\|\cdot\|_{H^s(\Gamma)} \eqsim |\!|\!|\cdot|\!|\!|_s$ on $H^s(\Gamma)$. Furthermore, if $s < \frac32$, then the functions in the aforementioned set generate a dense subset in $\mathcal{H}^s(\Gamma)$. Using in addition duality, we infer that

$$\mathcal{H}^s(\Gamma) \asymp \mathcal{H}_s(\Gamma) \quad (|s| < \tfrac32 \text{ with } |s| < s_\Gamma \notin \mathbb{N} \text{ or } |s| \leq s_\Gamma \in \mathbb{N}), \tag{2.4}$$

meaning that both spaces agree as sets and have equivalent norms. The spaces $\mathcal{H}_s(\Gamma)$ will serve only as auxiliary spaces to prove that the wavelets we are going to construct generate, when properly scaled, a Riesz basis for $\mathcal{H}^s$ for the full range of s; in case this range is limited by the regularity of Γ to a *closed* range $[-s_\Gamma, s_\Gamma]$.

3. Multiresolution analyses on the reference domain. On the reference domain, we will need two nested sequences of approximation spaces (multiresolution analyses) that satisfy Jackson and Bernstein estimates. We will assume that these spaces are equipped with single-scale bases that satisfy certain conditions concerning their supports and symmetry (cf. assumptions ($\mathcal{L}$), ($\mathcal{V}$), ($\mathcal{S}$)), so that after their lifting to the patches, they can be continuously connected over the interfaces. Furthermore, we will assume that the rate of best approximation from these sequences is realized by some concrete projector (cf. ($\mathcal{J}$) and Proposition 3.1), with which it will be shown that the induced approximation spaces on Γ are nested and have the same rate of approximation. We will make some assumptions (($\mathcal{J}1$) and ($\mathcal{J}2$)) connecting primal and dual multiresolution analyses to ensure the existence and uniform boundedness of the biorthogonal projector (cf. Proposition 5.2). Finally, we will assume the existence of a suitable "initial stable completion."

For $j \in \mathbb{N}_0$, let $I_j^\square \subset \overline{\square}$ be some index set with

$$\pi(I_j^\square) = I_j^\square \quad (\pi \in \Pi), \quad \sup_{y\in\square} \#(I_j^\square \cap B(y; 2^{-j})) \lesssim 1$$

(see Figure 2.1). For completeness, for $A \subset \mathbb{R}^n$ and $\delta \geq 0$, by $B(A;\delta)$ we mean $\{y \in \mathbb{R}^n : \mathrm{dist}(y, A) \leq \delta\}$, and $B(\emptyset;\delta) := \emptyset$. For $j \in \mathbb{N}_0$, we assume a collection $\Phi_j^\square = (\phi_{j,x}^\square)_{x\in I_j^\square} \subset C(\overline{\square})$, usually referred to as the set of *scaling functions*, such that the following hold:

($\mathcal{L}$) $\exists$ constant $\varepsilon > 0$, $\operatorname{supp} \phi^{\square}_{j,x} \subset B(x; \varepsilon 2^{-j})$.
($\mathcal{V}$) $\phi^{\square}_{j,x}$ vanishes on any face of $\square$ that does not contain x.
($\mathcal{S}$) $\phi^{\square}_{j,x} = \phi^{\square}_{j,\pi(x)} \circ \pi \quad (\pi \in \Pi)$.
($\mathcal{R}$) $\Phi^{\square}_j$ is a uniform $L_2(\square)$-Riesz system.
($\mathcal{J}$) There exists a collection of functionals $\Lambda^{\square}_j = (\lambda^{\square}_{j,x})_{x \in I^{\square}_j} \subset C(\overline{\square})'$ such that
  (i) $\exists$ constant $\vartheta > 0$, $\operatorname{supp} \lambda^{\square}_{j,x} \subset B(x; \vartheta 2^{-j})$.
  (ii) If $x \in \partial\square$, then $\operatorname{supp} \lambda^{\square}_{j,x}$ is contained in the lowest dimensional face of $\square$ that contains x.
  (iii) $\langle u, \lambda^{\square}_{j,\pi(x)} \rangle_{L_2(\square)} = \langle u \circ \pi, \lambda^{\square}_{j,x} \rangle_{L_2(\square)} \quad (\pi \in \Pi)$.
  (iv) $|\langle u, \lambda^{\square}_{j,x} \rangle_{L_2(\square)}| \lesssim 2^{-jn/2} \|u\|_{L_\infty(\operatorname{supp} \lambda^{\square}_{j,x})}$.
  (v) $\langle \Phi^{\square}_j, \Lambda^{\square}_j \rangle_{L_2(\square)} = Id$.
  (vi) For some $\frac{n}{2} < d \in \mathbb{N}$, $P_{d-1}(\square) \subset \mathcal{S}(\Phi^{\square}_j)$.
($\mathcal{N}$) $\mathcal{S}(\Phi^{\square}_j) \subset \mathcal{S}(\Phi^{\square}_{j+1})$.
($\mathcal{B}$) For some $\gamma > 0$, and any $s \in [0, \gamma)$, it holds that

$$\|u_j\|_{H^s(\square)} \lesssim 2^{sj} \|u_j\|_{L_2(\square)} \quad (u_j \in \mathcal{S}(\Phi^{\square}_j)).$$

Note that, in particular, ($\mathcal{J}$)(ii) implies that for x being a vertex of $\square$, $\langle u, \lambda^{\square}_{j,x} \rangle_{L_2(\square)}$ is a multiple of $u(x)$. Examples of such collections will be given at the end of this section.

PROPOSITION 3.1. *For the projector $P^{\square}_j : u \mapsto \langle u, \Lambda^{\square}_j \rangle_{L_2(\square)} \Phi^{\square}_j$ onto $\mathcal{S}(\Phi^{\square}_j)$, we have*

$$\|u - P^{\square}_j u\|_{L_2(\diamond)} \lesssim 2^{-dj} |u|_{H^d(B(\diamond; (\vartheta + 3\varepsilon) 2^{-j}) \cap \square)} \quad (\diamond \subset \square, u \in H^d(\square)).$$

The proof given in [22] follows standard lines.

Apart from the above collection $\Phi^{\square}_j$ of *primal* scaling functions, for $j \in \mathbb{N}_0$ we assume the existence of a collection $\tilde{\Phi}^{\square}_j = (\tilde{\phi}^{\square}_{j,x})_{x \in I^{\square}_j} \subset C(\overline{\square})$ of *dual* scaling functions. This collection should also satisfy all of ($\mathcal{L}$)–($\mathcal{B}$) with the same index set $I^{\square}_j$, but with generally different parameters and functionals in ($\mathcal{B}$) and ($\mathcal{J}$) that we will denote as $\tilde{\gamma} > 0$, $\tilde{d} > \frac{n}{2}$, $\tilde{\Lambda}^{\square}_j$, and $\tilde{\varepsilon}, \tilde{\vartheta} > 0$. The resulting projector $\tilde{P}^{\square}_j : u \mapsto \langle u, \tilde{\Lambda}^{\square}_j \rangle_{L_2(\square)} \tilde{\Phi}^{\square}_j$ satisfies the analogue of Proposition 3.1 with $(d, \vartheta, \varepsilon)$ replaced by $(\tilde{d}, \tilde{\vartheta}, \tilde{\varepsilon})$.

Since $\Phi^{\square}_j$ and $\tilde{\Phi}^{\square}_j$ are uniform $L_2(\square)$-Riesz systems, the matrix $\langle \Phi^{\square}_j, \tilde{\Phi}^{\square}_j \rangle_{L_2(\square)}$ defines a uniformly bounded linear operator on $\ell_2(I^{\square}_j)$. A relation between $\mathcal{S}(\Phi^{\square}_j)$ and $\mathcal{S}(\tilde{\Phi}^{\square}_j)$ is established by assuming that its real part satisfies

$$\Re \langle \Phi^{\square}_j, \tilde{\Phi}^{\square}_j \rangle_{L_2(\square)} \gtrsim Id. \tag{$\mathcal{J}1$}$$

Finally, for $j \in \mathbb{N}_0$, let $J^{\square}_j \subset \overline{\square}$ be some index set with $\pi(J^{\square}_j) = J^{\square}_j$ $(\pi \in \Pi)$, $\sup_{y \in \square} \#(J^{\square}_j \cap B(y; 2^{-j})) \lesssim 1$, and for e being either $\overline{\square}$ or any face of $\square$, $\#((I^{\square}_j \cup J^{\square}_j) \cap e) = \#(I^{\square}_{j+1} \cap e)$. In case $I^{\square}_j \subset I^{\square}_{j+1}$, a natural candidate is $J^{\square}_j = I^{\square}_{j+1} \backslash I^{\square}_j$. We assume the existence of collections $\Theta^{\square}_j = (\theta^{\square}_{j,x})_{x \in I^{\square}_j}$ with

$$\langle \Theta^{\square}_j, \tilde{\Phi}^{\square}_j \rangle_{L_2(\square)} = Id, \tag{$\mathcal{J}2$}$$

and $\Xi^{\square}_j = (\xi^{\square}_{j,x})_{x \in J^{\square}_j}$ such that the union $\Upsilon^{\square}_{j+1} := \left[(\Theta^{\square}_j)^T \quad (\Xi^{\square}_j)^T \right]^T$ satisfies ($\mathcal{L}$)–($\mathcal{R}$), and $\mathcal{S}(\Upsilon^{\square}_{j+1}) = \mathcal{S}(\Phi^{\square}_{j+1})$.

Remark 3.2. "Classical" wavelet constructions start with assuming biorthogonal scaling functions, i.e., $\langle \Phi_j^\square, \tilde{\Phi}_j^\square \rangle_{L_2(\square)} = Id$, in which case (J1) and (J2) are satisfied with $\Theta_j^\square = \Phi_j^\square$. When there is no need for locally supported dual wavelets, biorthogonality of the scaling functions can be relaxed to the conditions given here, with generally $\Theta_j^\square$ different from $\Phi_j^\square$, and in particular not contained in $\mathcal{S}(\Phi_j)$. For the case that $\Theta_j^\square = \Phi_j^\square$, in the literature the set $\Xi_j^\square$ is sometimes called an initial "stable" completion of $\Phi_j^\square$, that is, a completion of $\tilde{\Phi}_j^\square$ to a uniform $L_2(\square)$-Riesz basis for $\mathcal{S}(\Phi_{j+1}^\square)$. The wavelets to be constructed are then thought of being the target stable completion.

Remark 3.3. The condition (J2) can be further relaxed, which turned out to be useful in [20]. Instead of assuming that $\langle \Theta_j^\square, \tilde{\Phi}_j^\square \rangle_{L_2(\square)}$ is diagonal, more generally it is also sufficient when, for some fixed p, $I_j^\square$ is the union of disjoint sets $I_{j,1}^\square, \ldots, I_{j,p}^\square$, with $\pi(I_{j,i}^\square) = I_{j,i}^\square$ ($\pi \in \Pi$, $1 \le i \le p$), such that, with respect to this partitioning, $\langle \Theta_j^\square, \tilde{\Phi}_j^\square \rangle_{L_2(\square)}$ is a block triangular matrix, with diagonal blocks that are identity matrices. Although all results from this paper are also valid under this relaxed assumption, for ease of presentation we will stick to assumption (J2).

Because of $\mathcal{S}(\Phi_j^\square) \subset \mathcal{S}(\Phi_{j+1}^\square)$, $\mathcal{S}(\tilde{\Phi}_j^\square) \subset \mathcal{S}(\tilde{\Phi}_{j+1}^\square)$, and $\mathcal{S}(\Upsilon_{j+1}^\square) = \mathcal{S}(\Phi_{j+1}^\square)$, it holds that $\Phi_j^\square = \langle \Phi_j^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)} \Phi_{j+1}^\square$, $\tilde{\Phi}_j^\square = \langle \tilde{\Phi}_j^\square, \tilde{\Lambda}_{j+1}^\square \rangle_{L_2(\square)} \tilde{\Phi}_{j+1}^\square$, and $\Upsilon_{j+1}^\square = \langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)} \Phi_{j+1}^\square$, where $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}$ is uniformly boundedly invertible.

LEMMA 3.4. *For the matrix* $\mathbf{R}_j^\square$ *being* $\langle \Phi_j^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}$, $\langle \tilde{\Phi}_j^\square, \tilde{\Lambda}_{j+1}^\square \rangle_{L_2(\square)}$, $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}$, *or* $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}^{-1}$, *it holds that*

(a) $(\mathbf{R}_j^\square)_{\pi(x),\pi(y)} = (\mathbf{R}_j^\square)_{x,y}$ ($\pi \in \Pi$).

(b) $(\mathbf{R}_j^\square)_{x,y} = 0$ *when* y *is on a face of* $\square$ *that does not contain* x.

Proof. Part (a) follows from the assumptions ($\mathcal{S}$) or ($\mathcal{J}$)(iii) for the involved collections of functions and functionals, respectively. Similarly, for the first three matrices, part (b) follows from the assumptions ($\mathcal{V}$) or ($\mathcal{J}$)(ii). Now let e be a face of $\square$. With respect to the partitioning of the index sets for $\Upsilon_{j+1}^\square$ and $\Lambda_{j+1}^\square$ into indices on e and indices not on e, $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}$ is a 2×2 upper block triangular matrix with square diagonal blocks, and thus so is its inverse, which shows (b) also for $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}^{-1}$. ∎

By our assumptions, the matrices $\langle \Phi_j^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}$ and $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}$ are *uniformly local*, by which we mean that only entries with indices (x, y) with $|x - y| \lesssim 2^{-j}$ might be nonzero. As a consequence, for the wavelets we are going to construct, the basis transformation from wavelet to single-scale basis will be of optimal computational complexity.

For some applications, it is also essential to have a basis transformation from single-scale to wavelet basis that is of optimal computational complexity. In that case, one has to assume *both* that

$$\Theta_j^\square = \Phi_j^\square,$$

with which (J1) can be dropped since it is implied by (J2), *and* also that

$$\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}^{-1} \text{ is } \textit{uniformly local}. \tag{3.1}$$

Note that, for $\Theta_j^\square = \Phi_j^\square$, $\langle \Upsilon_{j+1}^\square, \Lambda_{j+1}^\square \rangle_{L_2(\square)}^{-1}$ is the basis transformation from $\Phi_{j+1}^\square$ to the two-level basis $\Phi_j^\square \cup \Xi_j^\square$.

All conditions imposed in this section are satisfied by the collections $\Phi_j^\square$, $\tilde{\Phi}_j^\square$, $\Theta_j^\square$, $\Xi_j^\square$ underlying the *finite element wavelets* introduced in [14]. With this construction,

$\mathcal{S}(\Phi_j^\square)$, $\mathcal{S}(\tilde{\Phi}_j^\square)$ are standard Lagrange finite element spaces, so that $\gamma = \tilde{\gamma} = \frac{3}{2}$, of orders d and $\tilde{d}$, respectively, with respect to a j times repeated uniform dyadic refinement of an initial simplicial partition of a polytope. For the present setting, we take this polytope to be a reference *n-simplex*. Thinking of $\tilde{d} \geq d$, these orders are chosen such that there is an $m \in \mathbb{N}$ with $2^m(d-1) = \tilde{d} - 1$, so that with the initial partition at the dual side being the reference simplex itself, and at the primal side being created by applying m dyadic recursive refinements to this simplex, we have $\dim \mathcal{S}(\Phi_j^\square) = \dim \mathcal{S}(\tilde{\Phi}_j^\square)$. So each "element" at the dual side is equal to a macro-element at the primal side consisting of 2^m "elements." The collections $\tilde{\Phi}_j^\square$ at the dual side, and $\Phi_j^\square$, $\Theta_j^\square$, $\Xi_j^\square$ at the primal side are now assembled in the standard finite element fashion from local collections, of a small, fixed dimension, of functions defined on the individual elements or macroelements, respectively. Each of these collections is a copy, or more precisely, a push forward using an affine bijection of such a collection created once and for all on a reference (macro)element. The functionals from $\tilde{\Lambda}_j^\square$ and $\Lambda_j^\square$ are assembled in the same manner from local collections, and are either simply scaled function evaluations in the "nodal points" $I_j^\square$ or local linear combinations of these. Actually, in the present paper, we will repeat the idea of assembling functions and functionals from collections defined on (macro)elements, which in turn are push forwards of collections defined on a reference (macro)element. Now the role of the (macro)elements will be played by the patches Γ_q, and that of the reference (macro)element by $\kappa_q^{-1}(\Gamma_q)$. A difference is that the number of patches is fixed, and that, as a consequence, the dimension of the local collections grows with the level. The major difficulty we have to deal with is that generally the κ_q are not affine, so that the Jacobian determinants are not constants.

In [20], we reconsidered the finite element wavelets and constructed collections with $\Theta_j^\square = \Phi_j^\square$, so that also the resulting dual wavelets are locally supported. In this case, the dual spaces, although consisting of continuous piecewise polynomials, are not standard finite element spaces.

Other examples of collections $\Phi_j^\square$, $\tilde{\Phi}_j^\square$, $\Theta_j^\square$, $\Xi_j^\square$ that satisfy our assumptions, with $\Theta_j^\square = \Phi_j^\square$, and now with $\square$ being the *n-cube*, are given in [12] and underlie the construction of *biorthogonal spline wavelets*. These collections are slight modifications of those developed in [11], and, for $n > 1$, they are simply generated using tensor products from univariate collections $\Phi_j^{[0,1]}$, $\tilde{\Phi}_j^{[0,1]}$, $\Xi_j^{[0,1]}$ defined on $[0,1]$. For given $\tilde{d} \geq d \geq 2$ with $d + \tilde{d}$ even, $\mathcal{S}(\Phi_j^{[0,1]})$ is the spline space of order d, so that $\gamma = d - \frac{1}{2}$, with respect to the knot sequence

$$(\underbrace{0, \ldots, 0}_{d \text{ times}}, r2^{-j}, r2^{-j} + 2^{-j}, \ldots, 1 - r2^{-j}, \underbrace{1, \ldots, 1}_{d \text{ times}}),$$

where $\mathbb{N} \ni r \geq d - 1$ is some parameter that one can choose. The collection $\tilde{\Phi}_j^{[0,1]}$ is such that $P_{\tilde{d}-1}[0,1] \subset \mathcal{S}(\tilde{\Phi}_j^{[0,1]})$ and $\langle \Phi_j^{[0,1]}, \tilde{\Phi}_j^{[0,1]} \rangle_{L_2([0,1])} = Id$, where $\tilde{\gamma}$ grows linearly with $\tilde{d}$. For x not near the endpoints 0 or 1, $\phi_{j,x}^{[0,1]} = 2^{j/2}\phi(2^j \cdot -x)$ and $\tilde{\phi}_{j,x}^{[0,1]} = 2^{j/2}\tilde{\phi}(2^j \cdot -x)$, where $(\phi, \tilde{\phi})$ is a biorthogonal pair constructed in [6]. Also the functionals from $\Lambda_j^\square$ and $\tilde{\Lambda}_j^\square$ are constructed from the collections of univariate functionals $\Lambda_j^{[0,1]}$ and $\tilde{\Lambda}_j^{[0,1]}$ using tensor products, where $\lambda_{j,x}^{[0,1]} = \tilde{\phi}_{j,x}^{[0,1]}$, $\tilde{\lambda}_{j,x}^{[0,1]} = \phi_{j,x}^{[0,1]}$ for $x \notin \{0,1\}$, and where they are simply scaled function evaluations in 0 or 1, respectively, otherwise.

4. Induced, continuous multiresolution analyses on Γ. By lifting the collections of functions on $\square$ to the patches of Γ, and by connecting those that do not vanish at the interfaces continuously with ones from other patches, we will construct nested sequences of primal and dual spaces that satisfy Jackson estimates and Bernstein inequalities.

We define the index sets $I_j \subset \overline{\Gamma}\backslash\partial\Gamma_D$, and analogously J_j, by

$$I_j = (\cup_{q=1}^M \kappa_q(I_j^\square)) \cap (\overline{\Gamma}\backslash\partial\Gamma_D)$$

(see Figure 2.1). By $(\mathcal{M})$ and $\pi(I_j^\square) = I_j^\square$ $(\pi \in \Pi)$, for any $1 \le q, q' \le M$ with $\overline{\Gamma_q} \cap \overline{\Gamma_{q'}} \neq \emptyset$, the sets $\kappa_q(I_j^\square)$ and $\kappa_{q'}(I_j^\square)$ restricted to this interface coincide. For $x \in \overline{\Gamma}$, we set $k(x) = \#\{q : x \in \overline{\Gamma_q}\}$.

For $j \in \mathbb{N}_0$, we define the collection $\Phi_j = (\phi_{j,x})_{x\in I_j} \subset C(\overline{\Gamma})$ by

$$\phi_{j,x}(y) = k(x)^{-\frac{1}{2}} \begin{cases} \phi^\square_{j,\kappa_q^{-1}(x)}(\kappa_q^{-1}(y)) & \text{when } x, y \in \overline{\Gamma_q} \text{ for some } 1 \le q \le M, \\ 0 & \text{elsewhere.} \end{cases} \tag{4.1}$$

Note that by $(\mathcal{S})$, $(\mathcal{V})$, and (2.1), $\phi_{j,x}$ is well defined and indeed continuous, and it vanishes on $\partial\Gamma_D$. By assumption $(\mathcal{L})$, the collection Φ_j is *uniformly local*, by which we mean that $x \in \operatorname{supp}\phi_{j,x}$, and that $d_\Gamma(x,y) \lesssim 2^{-j}$ for any $y \in \operatorname{supp}\phi_{j,x}$, where $d_\Gamma(x,y)$ denotes the geodesic distance of x and y over Γ, i.e., the length of the shortest curve on Γ connecting x and y.

With $\mathbf{E}_{j,q} : \ell_2(I_j^\square) \to \ell_2(I_j)$ defined by

$$(\mathbf{E}_{j,q}\mathbf{c}_j^\square)_x = k(x)^{-\frac{1}{2}} \begin{cases} \mathbf{c}^\square_{j,\kappa_q^{-1}(x)}, & x \in \overline{\Gamma_q}, \\ 0, & \text{otherwise,} \end{cases} \tag{4.2}$$

and similarly $\mathbf{F}_{j,q} : \ell_2(J_j^\square) \to \ell_2(J_j)$, we have $\sum_{q=1}^M \mathbf{E}_{j,q}\mathbf{E}_{j,q}^T = Id$ and $\sum_{q=1}^M \mathbf{F}_{j,q}\mathbf{F}_{j,q}^T = Id$. By construction of Φ_j from $\Phi_j^\square$, we have

$$\langle \Phi_j, \Phi_j\rangle_{L_2(\Gamma)} = \sum_{q=1}^M \mathbf{E}_{j,q}\langle \Phi_j^\square, \Phi_j^\square\rangle_{L_2(\square),|\partial\kappa_q|}\mathbf{E}_{j,q}^T, \tag{4.3}$$

so that, because of $\langle \Phi_j^\square, \Phi_j^\square\rangle_{L_2(\square),|\partial\kappa_q|} \eqsim \langle \Phi_j^\square, \Phi_j^\square\rangle_{L_2(\square)} \eqsim Id$ by $|\partial\kappa_q| \eqsim 1$ and $(\mathcal{R})$, Φ_j *is a uniform* $L_2(\Gamma)$*-Riesz system.*

PROPOSITION 4.1. *Setting* $\Lambda_j = (\lambda_{j,x})_{x\in I_j} \subset C(\overline{\Gamma})'$ *by*

$$\lambda_{j,x}(u) = k(x)^{\frac{1}{2}}\lambda^\square_{j,\kappa_q^{-1}(x)}(u\circ\kappa_q) \quad \textit{when } x \in \overline{\Gamma}_q,$$

we have $\langle \Phi_j, \Lambda_j\rangle_{L_2(\Gamma)} = Id$*. The projector* $P_j : u \mapsto \langle u, \Lambda_j\rangle_{L_2(\Gamma)}\Phi_j$ *onto* $\mathcal{S}(\Phi_j)$ *satisfies*

$$\|(Id-P_j)u\|_{L_2(\Omega)} \lesssim 2^{-dj}\sum_{q=1}^M |u\circ\kappa_q|_{H^d(B(\kappa_q^{-1}(\Omega\cap\Gamma_q);(\vartheta+3\varepsilon)2^{-j})\cap\square)} \quad (\Omega \subset \Gamma,\ u \in \mathcal{H}_d(\Gamma)).$$

Proof. Assumption $(\mathcal{J})$(ii) shows that $\lambda_{j,x}(u)$ is well defined for $u \in C(\overline{\Gamma})$, also when x is on an interface between patches, and, because of (4.1), that $(P_ju)\circ\kappa_q =$

$P_j^{\square}(u\circ\kappa_q)$ when u vanishes on $\partial\Gamma_D$. Condition $(\mathcal{J})$(v) shows that $\langle\Phi_j,\Lambda_j\rangle_{L_2(\Gamma)}=Id$. By Proposition 3.1, we have

$$\begin{aligned}\|(Id-P_j)u\|_{L_2(\Omega)} &\eqsim \sum_{q=1}^{M}\|((Id-P_j)u)\circ\kappa_q\|_{L_2(\kappa_q^{-1}(\Omega\cap\Gamma_q))}\\ &= \sum_{q=1}^{M}\|(Id-P_j^{\square})(u\circ\kappa_q)\|_{L_2(\kappa_q^{-1}(\Omega\cap\Gamma_q))}\\ &\lesssim 2^{-dj}\sum_{q=1}^{M}|u\circ\kappa_q|_{H^d(B(\kappa_q^{-1}(\Omega\cap\Gamma_q);(\vartheta+3\varepsilon)2^{-j})\cap\square)}. \qquad \square\end{aligned}$$

By substituting $\Omega=\Gamma$ in Proposition 4.1, we have the following *Jackson estimate*:

$$\inf_{u_j\in\mathcal{S}(\Phi_j)}\|u-u_j\|_{L_2(\Gamma)}\lesssim 2^{-dj}\|\!|u|\!\|_d \qquad (u\in\mathcal{H}_d(\Gamma)). \tag{4.4}$$

A direct consequence of $(\mathcal{B})$ is the following *Bernstein inequality*: For $s\in[0,\gamma)$,

$$\|\!|u_j|\!\|_s\lesssim 2^{sj}\|u_j\|_{L_2(\Gamma)} \quad (u_j\in\mathcal{S}(\Phi_j)). \tag{4.5}$$

Thanks to properties of a Sobolev scale, (4.5) gives rise to the following extended version that will be used in the appendix.

LEMMA 4.2. *For any $t\le s<\gamma$ with $t\le 0$,*

$$\|\!|u_j|\!\|_s\lesssim 2^{(s-t)j}\|\!|u_j|\!\|_t \quad (u_j\in\mathcal{S}(\Phi_j)).$$

The short proof of this lemma can be found in [22].

As $\Phi_j^{\square}$, via (4.1), gave rise to a uniformly local, uniform $L_2(\Gamma)$-Riesz system Φ_j, analogously the collections $\Upsilon_{j+1}^{\square}=\left[(\Theta_j^{\square})^T\quad(\Xi_j^{\square})^T\right]^T$ and $\tilde\Phi_j^{\square}$ yield *uniformly local, uniform $L_2(\Gamma)$-Riesz systems* $\Upsilon_{j+1}=\left[\Theta_j^T\quad\Xi_j^T\right]^T$ and $\tilde\Phi_j$, respectively.

We have the analogue of Proposition 4.1 at the dual side, with functionals and a projector denoted as $\tilde\Lambda_j=(\tilde\lambda_{j,x})_{x\in I_j}$ and $\tilde P_j$, respectively, and with Φ_j, d, ϑ, ε replaced by $\tilde\Phi_j$, $\tilde d$, $\tilde\vartheta$, $\tilde\varepsilon$. In particular, we have the *Jackson estimate*

$$\inf_{u_j\in\mathcal{S}(\tilde\Phi_j)}\|u-u_j\|_{L_2(\Gamma)}\lesssim 2^{-\tilde dj}\|\!|u|\!\|_{\tilde d} \qquad (u\in\mathcal{H}_{\tilde d}(\Gamma)), \tag{4.6}$$

and furthermore also the *Bernstein inequality*: For any $s\in[0,\tilde\gamma)$,

$$\|\!|u_j|\!\|_s\lesssim 2^{sj}\|u_j\|_{L_2(\Gamma)} \quad (u_j\in\mathcal{S}(\tilde\Phi_j)), \tag{4.7}$$

which can be extended analogously to Lemma 4.2.

Analogously to [12, Prop. 4.3.1], using Lemma 3.4, one may verify the following easily implementable formulas for the representations of the global embeddings in terms of corresponding representations of local embeddings.

PROPOSITION 4.3. *It holds that*

$$\Phi_j = \sum_{q=1}^{M} \mathbf{E}_{j,q} \langle \Phi_j^{\square}, \Lambda_{j+1}^{\square} \rangle_{L_2(\square)} \mathbf{E}_{j+1,q}^T \Phi_{j+1},$$

$$\tilde{\Phi}_j = \sum_{q=1}^{M} \mathbf{E}_{j,q} \langle \tilde{\Phi}_j^{\square}, \tilde{\Lambda}_{j+1}^{\square} \rangle_{L_2(\square)} \mathbf{E}_{j+1,q}^T \tilde{\Phi}_{j+1},$$

$$\Upsilon_{j+1} = \sum_{q=1}^{M} \left[\begin{array}{c|c} \mathbf{E}_{j,q} & 0 \\ \hline 0 & \mathbf{F}_{j,q} \end{array}\right] \langle \Upsilon_{j+1}^{\square}, \Lambda_{j+1}^{\square} \rangle_{L_2(\square)} \mathbf{E}_{j+1,q}^T \Phi_{j+1},$$

$$\Phi_{j+1} = \sum_{q=1}^{M} \mathbf{E}_{j+1,q} \langle \Upsilon_{j+1}^{\square}, \Lambda_{j+1}^{\square} \rangle_{L_2(\square)}^{-1} \left[\begin{array}{c|c} \mathbf{E}_{j,q}^T & 0 \\ \hline 0 & \mathbf{F}_{j,q}^T \end{array}\right] \Upsilon_{j+1}.$$

So, in particular, $\mathcal{S}(\Phi_j) \subset \mathcal{S}(\Phi_{j+1})$, $\mathcal{S}(\tilde{\Phi}_j) \subset \mathcal{S}(\tilde{\Phi}_{j+1})$, *and* $\mathcal{S}(\Upsilon_{j+1}) = \mathcal{S}(\Phi_{j+1})$.

5. Biorthogonal space decompositions and wavelets. We have constructed primal and dual sequences of nested spaces that satisfy Jackson and Bernstein estimates. To conclude existence and stability, with respect to a range of Sobolev norms, of the corresponding biorthogonal space decompositions, the only thing left to show is the existence and uniform $L_2(\Gamma)$-boundedness of the biorthogonal projector.

Results similar to the next lemma are often used in the context of saddle point problems. A proof of (the nontrivial part of) this lemma can be found in, e.g., [14, Theorem 2.1(a)].

LEMMA 5.1. *Let* V, U *be closed subspaces of a Hilbert space* H. *Then the following statements are equivalent:*

(a) $\gamma := \inf_{0 \neq u \in U} \sup_{0 \neq v \in V} \frac{|\langle u,v \rangle|}{\|u\|\|v\|} > 0$, *and for any* $v \in V$, *there exists a* $u \in U$ *with* $\langle u, v \rangle \neq 0$.

(b) *There exists a bounded projector* $Q : H \to H$ *with* $\Im(Q) = V$ *and* $\Im(I - Q) = U^{\perp}$, *which is therefore appropriately called a* biorthogonal projector.

In either case it holds that $\gamma = \|Q\|^{-1}$, *and the adjoint* Q^* *satisfies* $\Im(Q^*) = U$ *and* $\Im(I - Q^*) = V^{\perp}$.

When Σ *and* Δ *are Riesz bases for* U *and* V, *respectively, then* (a) *or* (b) *is equivalent to the existence of a bounded inverse of* $\langle \Sigma, \Delta \rangle : \ell_2(\Delta) \to \ell_2(\Sigma)$. *In that case it holds that*

$$\|\langle \Sigma, \Sigma \rangle^{-1}\|^{-\frac{1}{2}} \|\langle \Delta, \Delta \rangle^{-1}\|^{-\frac{1}{2}} \leq \frac{\|Q\|}{\|\langle \Sigma, \Delta \rangle^{-1}\|} \leq \|\Sigma\| \|\Delta\|.$$

To be able to transfer results valid on the reference parameter domain to the manifold, in particular those concerning $L_2(\square)$- or $L_2(\Gamma)$-angles between spaces, we will have to assume that the coarsest "mesh" is sufficiently fine in order to control the influence of the generally nonconstant Jacobian determinants.

PROPOSITION 5.2. *For* $j \geq j_0$ *being large enough, there exists a uniformly bounded projector* $Q_j : L_2(\Gamma) \to L_2(\Gamma)$ *with* $\Im(Q_j) = \mathcal{S}(\Phi_j)$ *and* $\Im(I - Q_j) = \mathcal{S}(\tilde{\Phi}_j)^{\perp_{L_2(\Gamma)}}$.

Proof. Setting

$$\Delta_{j,q}^{\square} = \operatorname{diag}(|\partial \kappa_q(x)|)_{x \in I_j^{\square}}, \tag{5.1}$$

by $(\mathcal{L})$ for both $\Phi_j^\square, \tilde{\Phi}_j^\square$, the smoothness of $z \mapsto |\partial\kappa_q(z)|$, and the uniform boundedness of $\|\phi_{j,x}^\square\|_{L_2(\square)}$, $\|\tilde{\phi}_{j,x}^\square\|_{L_2(\square)}$, we have

$$\|\langle \Phi_j^\square, \tilde{\Phi}_j^\square\rangle_{L_2(\square),|\partial\kappa_q|} - (\Delta_{j,q}^\square)^{\frac{1}{2}}\langle \Phi_j^\square, \tilde{\Phi}_j^\square\rangle_{L_2(\square)}(\Delta_{j,q}^\square)^{\frac{1}{2}}\| \lesssim 2^{-j}. \tag{5.2}$$

By assumption $(\mathcal{J}1)$ and $|\partial\kappa_q| \gtrsim 1$, we have

$$\Re((\Delta_{j,q}^\square)^{\frac{1}{2}}\langle \Phi_j^\square, \tilde{\Phi}_j^\square\rangle_{L_2(\square)}(\Delta_{j,q}^\square)^{\frac{1}{2}}) = (\Delta_{j,q}^\square)^{\frac{1}{2}}\Re\langle \Phi_j^\square, \tilde{\Phi}_j^\square\rangle_{L_2(\square)}(\Delta_{j,q}^\square)^{\frac{1}{2}} \gtrsim Id,$$

so that for $j \geq j_0$ large enough, $\Re\langle \Phi_j^\square, \tilde{\Phi}_j^\square\rangle_{L_2(\square),|\partial\kappa_q|} \gtrsim Id$. Similarly to (4.3), we find that

$$\Re\langle \Phi_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)} = \sum_{q=1}^M \mathbf{E}_{j,q}\Re\langle \Phi_j^\square, \tilde{\Phi}_j^\square\rangle_{L_2(\square),|\partial\kappa_q|}\mathbf{E}_{j,q}^T \gtrsim \sum_{q=1}^M \mathbf{E}_{j,q}\mathbf{E}_{j,q}^T = Id.$$

Since apparently, for $j \geq j_0$, $\langle \Phi_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is uniformly boundedly invertible, and Φ_j and $\tilde{\Phi}_j$ are uniform $L_2(\Gamma)$-Riesz systems, an application of Lemma 5.1 completes the the proof. □

For $j \geq j_0$, the nesting $\mathcal{S}(\tilde{\Phi}_j) \subset \mathcal{S}(\tilde{\Phi}_{j+1})$ gives $Q_j^* = Q_{j+1}^* Q_j^*$ or $Q_j = Q_j Q_{j+1}$, from which it follows that

$$\Im(Q_{j+1} - Q_j) = \mathcal{S}(\Phi_{j+1}) \cap \mathcal{S}(\tilde{\Phi}_j)^{\perp_{L_2(\Gamma)}}.$$

Analogously, $\mathcal{S}(\Phi_j) \subset \mathcal{S}(\Phi_{j+1})$ implies that

$$\Im(Q_{j+1}^* - Q_j^*) = \mathcal{S}(\tilde{\Phi}_{j+1}) \cap \mathcal{S}(\Phi_j)^{\perp_{L_2(\Gamma)}}.$$

From the Jackson estimates and Bernstein inequalities at primal and dual sides (4.4), (4.5), (4.6), and (4.7), and the existence and uniform $L_2(\Gamma)$-boundedness of the biorthogonal projectors Q_j from Proposition 5.2, we have the following theorem.

THEOREM 5.3 (cf., e.g., [8], [14, Theorem 2.1]). *With $Q_{j_0-1} := 0$, it holds that*

$$\Big\|\Big\| \sum_{j=j_0}^\infty w_j \Big\|\Big\|_s^2 \lesssim \sum_{j=j_0}^\infty 4^{sj}\|w_j\|_{L_2(\Gamma)}^2 \qquad (w_j \in \Im(Q_j - Q_{j-1}),\ s \in (-\tilde{d}, \gamma)), \tag{5.3}$$

and

$$\sum_{j=j_0}^\infty 4^{sj}\|(Q_j - Q_{j-1})u\|_{L_2(\Gamma)}^2 \lesssim \|\!|u|\!\|_s^2 \qquad (u \in \mathcal{H}_s(\Gamma),\ s \in (-\tilde{\gamma}, d)). \tag{5.4}$$

For $s \in (-\min\{\tilde{\gamma}, \tilde{d}\}, \min\{\gamma, d\})$, $(w_j)_{j\geq j_0} \mapsto \sum_{j=j_0}^\infty w_j$, and $u \mapsto ((Q_j - Q_{j-1})u)_{j\geq j_0}$, mappings that are bounded in the sense of (5.3) *and* (5.4) *are each others' inverse.*

Analogous results are valid with (Q_j) replaced by (Q_j^*) and with interchanged roles of (γ, d) and $(\tilde{\gamma}, \tilde{d})$.

Next, we construct a uniform $L_2(\Gamma)$-Riesz basis for $\Im(Q_{j+1} - Q_j)$, whose elements are called *wavelets*.

PROPOSITION 5.4.

(a) *For $j \geq j_0$ being large enough, there exists a uniformly bounded projector $\bar{Q}_j : L_2(\Gamma) \to L_2(\Gamma)$ with $\Im(\bar{Q}_j) = \mathcal{S}(\Theta_j)$ and $\Im(Id - \bar{Q}_j) = \mathcal{S}(\tilde{\Phi}_j)^{\perp_{L_2(\Gamma)}}$.*

(b) *This projector can be computed as* $\bar{Q}_j u = \langle u, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}\langle \Theta_j, \tilde{\Phi}_j\rangle^{-1}_{L_2(\Gamma)}\Theta_j$.

(c) *The collection of wavelets*

$$\boxed{\Psi_j := \Xi_j - \langle \Xi_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}\langle \Theta_j, \tilde{\Phi}_j\rangle^{-1}_{L_2(\Gamma)}\Theta_j} \tag{5.5}$$

is a uniform $L_2(\Gamma)$*-Riesz basis for* $\mathcal{S}(\Phi_{j+1}) \cap \mathcal{S}(\tilde{\Phi}_j)^{\perp_{L_2(\Gamma)}}$.

So by taking j_0 *to be the maximum of the values from* (a) *and that of Proposition* 5.2*, for* $s \in (-\min\{\tilde{\gamma}, \tilde{d}\}, \min\{\gamma, d\})$,

$$\Phi_{j_0} \cup \cup_{j\geq j_0} 2^{-sj}\Psi_j \text{ is a Riesz basis for } \mathcal{H}_s(\Gamma),$$

and thus, in view of (2.4)*, when in addition* $|s| < \frac{3}{2}$, $|s| < s_\Gamma \notin \mathbb{N}$, *or* $|s| \leq s_\Gamma \in \mathbb{N}$, *it is a Riesz basis for* $\mathcal{H}^s(\Gamma)$.

Proof. (a) Since Θ_j and $\tilde{\Phi}_j$ are uniform $L_2(\Gamma)$-Riesz bases, by Lemma 5.1 we have to show that, for $j \geq j_0$ large enough, $\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is uniformly boundedly invertible, which follows from (J2) similarly to the proof of Proposition 5.2.

(b) Using that $\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is uniformly boundedly invertible, one easily verifies that $\bar{Q}_j$, as given in (b), indeed has the properties listed in (a).

(c) Let $u \in \mathcal{S}(\Phi_{j+1})$; then $u = \mathbf{c}_j^T\Theta_j + \mathbf{d}_j^T\Xi_j$ with $\|u\|_{L_2(\Gamma)} \eqsim (\|\mathbf{c}_j\|^2 + \|\mathbf{d}_j\|^2)^{\frac{1}{2}}$. If, in addition, $u \in \mathcal{S}(\tilde{\Phi}_j)^{\perp_{L_2(\Gamma)}}$, then $u = (Id - \bar{Q}_j)u = (Id - \bar{Q}_j)\mathbf{d}_j^T\Xi_j = \mathbf{d}_j^T\Psi_j$, and so with $\|u\|_{L_2(\Gamma)} \lesssim (1 + \|\bar{Q}_j\|_{L_2(\Gamma)\to L_2(\Gamma)})\|\mathbf{d}_j^T\Xi_j\|_{L_2(\Gamma)} \lesssim \|\mathbf{d}_j\| \lesssim \|u\|_{L_2(\Gamma)}$. Noting that $\Psi_j \subset \mathcal{S}(\Phi_{j+1}) \cap \mathcal{S}(\tilde{\Phi}_j)^{\perp_{L_2(\Gamma)}}$, we conclude that it is a uniform $L_2(\Gamma)$-Riesz basis for this space. The last statements are now consequences of Theorem 5.3. ∎

Note that Proposition 4.1 at the dual side implies that $\Psi_j = (\psi_{j,x})_{x\in J_j}$, yielded by (5.5), satisfies

$$\begin{aligned} |\langle\, \psi_{\,j,x}, u\rangle_{L_2(\Gamma)}| &= |\langle \psi_{j,x}, (Id - \tilde{P}_j)u\rangle_{L_2(\Gamma)}| \\ &\lesssim 2^{-\tilde{d}j}\sum_{q=1}^{M} |u\circ\kappa_q|_{H^{\tilde{d}}(B(\kappa_q^{-1}(\operatorname{supp}\psi_{j,x}\cap\Gamma_q);(\tilde{\vartheta}+3\tilde{\varepsilon})2^{-j})\cap\square)} \quad (u \in \mathcal{H}_{\tilde{d}}(\Gamma)). \end{aligned} \tag{5.6}$$

This property of the collections Ψ_j, with $\tilde{\vartheta} + 3\tilde{\varepsilon}$ replaced by an arbitrary but fixed $\tilde{\eta} \geq 0$ and the seminorms $|\cdot|_{H^{\tilde{d}}(\dots)}$ replaced by the norms $\|\cdot\|_{H^{\tilde{d}}(\dots)}$, will be referred to as the *uniform cancellation property of order* $\tilde{d}$.

6. Stability of *approximate* wavelet bases. Similarly to (4.3), the definition of the collections Θ_j and $\tilde{\Phi}_j$ via (4.1) shows that

$$\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)} = \sum_{q=1}^{M} \mathbf{E}_{j,q}\langle \Theta_j^{\square}, \tilde{\Phi}_j^{\square}\rangle_{L_2(\square),|\partial\kappa_q|}\mathbf{E}_{j,q}^T. \tag{6.1}$$

So if, for each q, $z \mapsto |\partial\kappa_q(z)|$ is a *constant function*, then (J2) shows that $\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is diagonal, and the collection of wavelets Ψ_j given in (5.5) is *uniformly local*. Unfortunately, only a restricted class of manifolds can be described as the union of patches that are the images of $\square$ under parametrizations that have constant Jacobians. In case *not all Jacobians are constants*, then, generally, $\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is not diagonal and its inverse is densely populated, so that (5.5) yields wavelets Ψ_j that have *global supports*.

A possibility to circumvent this problem, pursued in [12], is to carry out the whole wavelet construction outlined so far using the modified scalar product $\langle\!\langle\, ,\, \rangle\!\rangle_0$ instead of $\langle\, ,\, \rangle_{L_2(\Gamma)}$. Indeed, $\langle\!\langle \Theta_j, \tilde{\Phi}_j \rangle\!\rangle_0 = \sum_{q=1}^M \mathbf{E}_{j,q} \langle \Theta_j^\square, \tilde{\Phi}_j^\square \rangle_{L_2(\square)} \mathbf{E}_{j,q}^T = \sum_{q=1}^M \mathbf{E}_{j,q}\ \mathbf{E}_{j,q}^T = Id$, and so uniformly local wavelets are obtained. What is more, by employing this scalar product, it is always possible to take the coarsest level $j_0 = 0$.

As was already recognized in [12], this approach, however, has two limitations: First, the obtained wavelets will be orthogonal to the constant function with respect to $\langle\!\langle\, ,\, \rangle\!\rangle_0$. As a consequence, if the function

$$J : \cup_{q=1}^M \Gamma_q \to \mathbb{R} : x \mapsto |\partial \kappa_{q'}(\kappa_{q'}^{-1}(x))| \quad \text{when } x \in \Gamma_q'$$

has discontinuities, or more precisely, cannot be extended to a continuous function on Γ, then wavelets with supports that are not contained in one patch will generally *not* have a zero mean value with respect to the canonical Lebesgue measure on Γ, meaning that they have *no cancellation property* with respect to $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$. The application of wavelets we focus on is that for the solution of differential or integral equations in variational form using the duality pairing with respect to $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$ (taking a different scalar product here yields other disadvantages; cf. [14, section 1.2]). For obtaining nearly sparse representations of these operators in wavelet coordinates, and with that algorithms of optimal computational complexity, the wavelets should have a cancellation property of sufficiently high order (cf. [10, 21] or the surveys [9, 3]) with respect thus to $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$. In a nonadaptive setting, under certain circumstances it might be possible that the fact that only wavelets along the lower dimensional patch interfaces do not have cancellation properties does not spoil optimal complexity. In an adaptive setting, however, such an argument cannot be applied.

The second limitation has to do with the interpretation for $s < 0$ of the statement that $\Phi_{j_0} \cup \cup_{j \geq j_0} 2^{-sj} \Psi_j$ is a Riesz basis for $\mathcal{H}_s(\Gamma)$, which is a consequence of Theorem 5.3. In case biorthogonality is realized with respect to $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$, then an expansion in terms of the basis $\Phi_{j_0} \cup \cup_{j \geq j_0} 2^{-sj} \Psi_j$ should be interpreted as an element of $\mathcal{H}_s(\Gamma)$, i.e., as a functional, using the embedding $L_2(\Gamma) \to \mathcal{H}_s(\Gamma) : u \mapsto (v \mapsto \langle v, u \rangle_{L_2(\Gamma)})$. Replacing $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$ by $\langle\!\langle \cdot, \cdot \rangle\!\rangle_0$ means that also the embedding should be changed into $u \mapsto (v \mapsto \langle\!\langle v, u \rangle\!\rangle_0)$. One can show (cf. [17, section 4]) that if J has discontinuities, then for $s \leq -\frac{1}{2}$ a set $\Phi_{j_0} \cup \cup_{j \geq j_0} 2^{-sj} \Psi_j$ which is a Riesz basis for $\mathcal{H}_s(\Gamma)$ using the embedding $u \mapsto (v \mapsto \langle\!\langle v, u \rangle\!\rangle_0)$ *cannot* be a Riesz basis for $\mathcal{H}_s(\Gamma)$ using the embedding $u \mapsto (v \mapsto \langle v, u \rangle_{L_2(\Gamma)})$. For $s > -\frac{1}{2}$, the property of being a Riesz basis for $\mathcal{H}_s(\Gamma)$ is the same for both embeddings. Again, since in applications duality pairing with respect to $\langle \cdot, \cdot \rangle_{L_2(\Gamma)}$ is used, in this paper the property of a collection of functions to be a Riesz basis for $\mathcal{H}_s(\Gamma)$ for $s < 0$ will always be interpreted with respect to the canonical embedding $u \mapsto (v \mapsto \langle v, u \rangle_{L_2(\Gamma)})$.

In this paper, we propose another approach to solve the problem that generally $\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}^{-1}$ is densely populated, so that the wavelets yielded by (5.5) have global supports. As we will see, $\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}^{-1}$ can be well approximated by uniformly local matrices, so that close to the collections of the wavelets Ψ_j, there are collections of uniformly local functions of which suitable ones might be applied instead. In the following main theorem of this paper we derive general criteria under which such *approximate wavelets* satisfy the *same* conditions as Ψ_j concerning *both* stability with respect to a range of Sobolev norms *and* the order of the cancellation property, where, moreover, in contrast to Ψ_j, they are uniformly local.

THEOREM 6.1. *If, for $j \geq j_0$,*

(i) $\breve{\Psi}_j = (\breve{\psi}_{j,x})_{x \in J_j} \subset \mathcal{S}(\Phi_{j+1})$ *is uniformly local,*

(ii) $\breve{\Psi}_j$ *has the uniform cancellation property of order* $\tilde{d}$ *(with respect to* $\langle\cdot,\cdot\rangle_{L_2(\Gamma)}$*),*

(iii) *for some* $\omega\in(0,1)$, $\|\Psi_j-\breve{\Psi}_j\|_{L_2(\Gamma)}\lesssim\omega^j$,

then, possibly for a larger value of j_0*, for* $s\in(-\min\{\tilde{\gamma},\tilde{d}\},\min\{\gamma,d\})$,

$$\Phi_{j_0}\cup\cup_{j\geq j_0}2^{-sj}\breve{\Psi}_j \text{ is a Riesz basis for } \mathcal{H}_s(\Gamma),$$

and thus, in view of (2.4)*, when in addition* $|s|<\frac{3}{2}$, $|s|<s_\Gamma\notin\mathbb{N}$*, or* $|s|\leq s_\Gamma\in\mathbb{N}$*, it is a Riesz basis for* $\mathcal{H}^s(\Gamma)$*.*

The rather lengthy proof of this theorem is postponed to the appendix. The new aspect of this theorem is that instead of assuming (iii) with $\omega\leq 2^{-\min\{\tilde{\gamma},\tilde{d}\}}$, which would yield the statement by "brute force" arguments, it is allowed that ω is arbitrarily close to 1 when, in addition, (ii) is valid, which property we like to have anyway. So although in two of our three constructions of approximate wavelets in the next section, ω will be equal to $\frac{1}{2}$, we nevertheless thus end up with Riesz bases for $\mathcal{H}_s(\Gamma)$ for the full range of s allowed by $\tilde{\gamma}$, $\tilde{d}$, γ, and d. It is easily seen that (ii) alone is not sufficient to guarantee that the approximate wavelets generate a Riesz basis for any $\mathcal{H}_s(\Gamma)$.

The proof of Theorem 6.1 is based on perturbation arguments making use of the fact that we know that the true $L_2(\Gamma)$-biorthogonal wavelets generate Riesz bases for the full range of Sobolev spaces. We derived this fact in section 5 by generalizing upon the well-known concept of stable completions developed in [2]. Note that in our setting we did not assume to have explicitly available $L_2(\Gamma)$-biorthogonal collections of scaling functions. Theorem 6.1 and the applications in the following sections show the value of this generalization.

Remark 6.2. The approximate wavelets $\breve{\Psi}_j$ we are going to construct will be of type $\breve{\Psi}_j=\Xi_j-\mathbf{Z}_j\Theta_j$, where $\mathbf{Z}_j$ is a uniformly local $\#J_j\times\#I_j$ matrix. Since the basis transformation $\langle\Upsilon_{j+1},\Lambda_{j+1}\rangle_{L_2(\Gamma)}^T$ from $\Upsilon_{j+1}=\begin{bmatrix}\Theta_j^T & \Xi_j^T\end{bmatrix}^T$ to Φ_{j+1} is uniformly local, so is the basis transformation from $\begin{bmatrix}\Theta_j^T & \breve{\Psi}_j^T\end{bmatrix}^T$ to Φ_{j+1}, and the transformation from the multiscale basis $\Phi_{j_0}\cup\cup_{k=j_0}^{j-1}\breve{\Psi}_k$ to the single-scale basis Φ_j has linear complexity.

In the special case that $\Theta_j=\Phi_j$ and $\langle\Upsilon_{j+1},\Lambda_{j+1}\rangle_{L_2(\Gamma)}^{-T}$ is uniformly local, which by Proposition 4.3 holds assuming (3.1), the basis transformation from Φ_{j+1} to $\begin{bmatrix}\Phi_j^T & \breve{\Psi}_j^T\end{bmatrix}^T$ will also be uniformly local, and so the inverse transformation from single-scale basis Φ_j to multiscale basis $\Phi_{j_0}\cup\cup_{k=j_0}^{j-1}\breve{\Psi}_k$ also has linear complexity. However, since $\mathcal{S}(\breve{\Psi}_j)$ is only approximately $L_2(\Gamma)$-orthogonal to $\mathcal{S}(\Phi_j)$, the corresponding dual wavelets will not be explicitly given.

7. Construction of uniformly local approximate wavelet bases.

7.1. Approximating $\langle\Theta_j,\tilde{\Phi}_j\rangle_{L_2(\Gamma)}^{-1}$ using Jacobi iteration. As we will see, for $j\to\infty$, the matrix $\langle\Theta_j,\tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is increasingly close to its diagonal, and so it makes sense to approximate its inverse by a few Jacobi iteration steps ($\tilde{d}$ steps will be sufficient). We will denote the resulting collection of approximate wavelets as Ψ_j^{Jc}, where "Jc" refers to Jacobi iteration.

THEOREM 7.1. *With* $\mathbf{D}_j:=\mathrm{diag}\langle\Theta_j,\tilde{\Phi}_j\rangle_{L_2(\Gamma)}$*, and for* $j\geq j_0$ *large enough,*

$$\Psi_j^{\mathrm{Jc}}:=\Xi_j-\langle\Xi_j,\tilde{\Phi}_j\rangle_{L_2(\Gamma)}\left[\sum_{k=0}^{\tilde{d}-1}(Id-\mathbf{D}_j^{-1}\langle\Theta_j,\tilde{\Phi}_j\rangle_{L_2(\Gamma)})^k\mathbf{D}_j^{-1}\right]\Theta_j$$

is uniformly local, and it has the uniform cancellation property of order $\tilde{d}$; and finally, for $s \in (-\min\{\tilde{\gamma}, \tilde{d}\}, \min\{\gamma, d\})$,

$$\Phi_{j_0} \cup \cup_{j \geq j_0} 2^{-sj} \Psi_j^{\mathrm{Jc}} \text{ is a Riesz basis for } \mathcal{H}_s(\Gamma).$$

In view of (2.4)*, when in addition $|s| < \frac{3}{2}$, $|s| < s_\Gamma \notin \mathbb{N}$, or $|s| \leq s_\Gamma \in \mathbb{N}$, the collection is thus a Riesz basis for $\mathcal{H}^s(\Gamma)$.*

Proof. With $\Delta_{j,q}^{\square}$ as defined in (5.1), and by using (J2), similarly to (5.2) we have $\|\langle \Theta_j^{\square}, \tilde{\Phi}_j^{\square} \rangle_{L_2(\square), |\partial \kappa_q|} - \Delta_{j,q}^{\square}\| \lesssim 2^{-j}$. Since

$$\bar{\mathbf{D}}_j := \sum_{q=1}^{M} \mathbf{E}_{j,q} \Delta_{j,q}^{\square} \mathbf{E}_{j,q}^T$$

is diagonal, by (6.1) we have

$$\begin{aligned} \|\mathbf{D}_j - \bar{\mathbf{D}}_j\| &\leq \max_{x,y \in I_j} |(\mathbf{D}_j - \bar{\mathbf{D}}_j)_{x,y}| \leq \max_{x,y \in I_j} |(\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)} - \bar{\mathbf{D}}_j)_{x,y}| \\ &\leq \|\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)} - \bar{\mathbf{D}}_j\| \lesssim 2^{-j}, \end{aligned}$$

and so $\|\mathbf{D}_j - \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}\| \leq \|\mathbf{D}_j - \bar{\mathbf{D}}_j\| + \|\bar{\mathbf{D}}_j - \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}\| \lesssim 2^{-j}$.

As we have seen in the proof of Proposition 5.4, for $j \geq j_0$ large enough, the matrix $\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}$ is uniformly boundedly invertible, and thus, possibly for a larger j_0, so is $\mathbf{D}_j$. We infer that

$$\begin{aligned} (7.1) \qquad &\left\| \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}^{-1} - \sum_{k=0}^{\tilde{d}-1} (Id - \mathbf{D}_j^{-1} \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)})^k \mathbf{D}_j^{-1} \right\| \\ &\qquad = \|\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}^{-1} ((\mathbf{D}_j - \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}) \mathbf{D}_j^{-1})^{\tilde{d}}\| \lesssim 2^{-\tilde{d}j}, \end{aligned}$$

and thus by $\|\langle \Xi_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}\| \lesssim 1$ that

$$(7.2) \qquad \|\Psi_j - \Psi_j^{\mathrm{Jc}}\|_{L_2(\Gamma)} \lesssim 2^{-\tilde{d}j} \|\Theta_j\|_{L_2(\Gamma)} \lesssim 2^{-\tilde{d}j}.$$

Since furthermore Ψ_j^{Jc} is uniformly local, in view of Theorem 6.1 the only thing left to show is that Ψ_j^{Jc} has the uniform cancellation property of order $\tilde{d}$.

Although Ψ_j has the uniform cancellation property of order $\tilde{d}$, we cannot immediately conclude this from (7.2) for Ψ_j^{Jc}. Indeed, since the wavelets from Ψ_j generally have global supports, invoking (7.2) and the cancellation property of $\psi_{j,x}$ would yield a bound for $|\langle \psi_{j,x}^{\mathrm{Jc}}, u \rangle_{L_2(\Gamma)}|$ in terms of the global H^d-norms of $u \circ \kappa_q$, whereas the definition of the cancellation property requires a bound in terms of the H^d-norms of $u \circ \kappa_q$ in a neighborhood of $\mathrm{supp}(\psi_{j,x}^{\mathrm{Jc}} \circ \kappa_q)$ with diameter of order 2^{-j}. To arrive at this result, we split u into $(Id - \tilde{P}_j)u$ and $\tilde{P}_j u$, and then replace $\tilde{P}_j u$ by a function, equal to $\tilde{P}_j u$ on $\mathrm{supp}\, \psi_{j,x}^{\mathrm{Jc}}$ and still in $\mathcal{S}(\tilde{\Phi}_j)$, that has a support with diameter of order 2^{-j}. The details can be found in [22]. ∎

Remark 7.2. The construction of the approximate wavelets Ψ_j^{Jc} in Theorem 7.1 has some similarities to the construction of approximate "prewavelets" in [23]. There the inverse of a mass matrix with respect to a standard finite element basis is approximated by a number of steps of an iterative method, as the Jacobi or symmetric Gauss–Seidel method. A difference is that in our case the matrix $\mathbf{D}_j$ *converges to*

$\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ as $j\to\infty$, allowing us to derive much stronger results concerning the generation of Riesz bases by the resulting approximate wavelets.

Compared to the approximate wavelets one gets by simply replacing $\langle\cdot,\cdot\rangle_{L_2(\Gamma)}$ by $\langle\!\langle\cdot,\cdot\rangle\!\rangle_0$ in (5.5), for $\tilde{d}\geq 2$ the approximate wavelets $\psi^{\mathrm{Jc}}_{j,x}$ have relatively large supports. Although this has not so much an effect on the multiscale to single-scale transform that can be implemented much more efficiently than is suggested by the sizes of the supports, it is a disadvantage, for example, when it concerns the compression of the stiffness matrix of an integral operator with respect to these approximate wavelets. In the following two subsections, we construct approximate wavelets with smaller supports.

As a preparation, the next proposition facilitates the verification of the third condition from Theorem 6.1, in case different constructions of approximate wavelets are used on different parts of Γ. In the proof, the problem of the generally global supports of the true biorthogonal wavelets is circumvented by approximating them by sufficiently accurate, uniformly local approximate wavelets generated by the Jacobi iteration approach.

PROPOSITION 7.3. *Let $\omega\in(0,1)$ and let $\breve{\Psi}_j=(\breve{\psi}_{j,x})_{x\in J_j}$ be uniformly local. Then $\|\Psi_j-\breve{\Psi}_j\|_{L_2(\Gamma)}\lesssim\omega^j$ if and only if $\sup_{x\in J_j}\|\psi_{j,x}-\breve{\psi}_{j,x}\|_{L_2(\Gamma)}\lesssim\omega^j$.*

Proof. Let $\sup_{x\in J_j}\|\psi_{j,x}-\breve{\psi}_{j,x}\|_{L_2(\Gamma)}\lesssim\omega^j$. Selecting $m\in\mathbb{N}$ such that $2^{-m}\leq\omega$, from the proof of Theorem 7.1 we learn that there exists a uniformly local Ψ^{Jc}_j with $\|\Psi_j-\Psi^{\mathrm{Jc}}_j\|_{L_2(\Gamma)}\lesssim 2^{-mj}\leq\omega^j$, and so $\sup_{x\in J_j}\|\psi_{j,x}-\psi^{\mathrm{Jc}}_{j,x}\|_{L_2(\Gamma)}\lesssim\omega^j$ and thus $\sup_{x\in J_j}\|\psi^{\mathrm{Jc}}_{j,x}-\breve{\psi}_{j,x}\|_{L_2(\Gamma)}\lesssim\omega^j$. Since both Ψ^{Jc}_j and $\breve{\Psi}_j$ are uniformly local, this implies $\|\Psi^{\mathrm{Jc}}_j-\breve{\Psi}_j\|_{L_2(\Gamma)}\lesssim\omega^j$ and thus that $\|\Psi_j-\breve{\Psi}_j\|_{L_2(\Gamma)}\leq\|\Psi_j-\Psi^{\mathrm{Jc}}_j\|_{L_2(\Gamma)}+\|\Psi^{\mathrm{Jc}}_j-\breve{\Psi}_j\|_{L_2(\Gamma)}\lesssim\omega^j$. The proof of the other implication is trivial. ∎

7.2. Ignoring the Jacobian determinants away from the interfaces. In this subsection, we show that *away from the patch interfaces*, we may replace the wavelets from Ψ^{Jc}_j by the corresponding ones from

$$\Psi^{(0)}_j:=\Xi_j-\langle\!\langle\Xi_j,\tilde{\Phi}_j\rangle\!\rangle_0\Theta_j.$$

This is the collection of biorthogonal wavelets one obtains when biorthogonality is realized with respect to $\langle\!\langle\,,\,\rangle\!\rangle_0$ instead of $\langle\,,\,\rangle_{L_2(\Gamma)}$, i.e., when, in the wavelet formula (5.5), all Jacobian determinants are replaced by the constant 1.

Recalling that for $x\in\overline{\Gamma}$, $k(x)=\#\{q: x\in\overline{\Gamma_q}\}$, we set $I^\circ_j=\{x\in I_j: k(x)=1\}$ and

$$J^\circ_j=\{x\in J_j: k(x)=1 \text{ and } \langle\xi_{j,x},\tilde{\phi}_{j,y}\rangle_{L_2(\Gamma)}=0 \text{ for all } y\in I_j\backslash I^\circ_j\};$$

this set is designed such that for $x\in J^\circ_j$, $\psi^{(0)}_{j,x}$ is fully supported inside one patch $\overline{\Gamma_q}$.

THEOREM 7.4. *The set $\breve{\Psi}_j=\{\breve{\psi}_{j,x}: x\in J_j\}$, defined by $\breve{\psi}_{j,x}=\psi^{(0)}_{j,x}$ when $x\in J^\circ_j$, and $\breve{\psi}_{j,x}=\psi^{\mathrm{Jc}}_{j,x}$ when $x\in J_j\backslash J^\circ_j$, is uniformly local and it has the uniform cancellation property of order $\tilde{d}$, and for any $s\in(-\min\{\tilde{\gamma},\tilde{d}\},\min\{\gamma,d\})$ and j_0 large enough,*

$$\Phi_{j_0}\cup\cup_{j\geq j_0}2^{-sj}\breve{\Psi}_j \text{ is a Riesz basis for } \mathcal{H}_s(\Gamma).$$

In view of (2.4), when in addition $|s|<\frac{3}{2}$, $|s|<s_\Gamma\notin\mathbb{N}$, or $|s|\leq s_\Gamma\in\mathbb{N}$, this collection is thus a Riesz basis for $\mathcal{H}^s(\Gamma)$.

Proof. Using that, for $j \geq j_0$ large enough, $\|\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}^{-1} - \mathbf{D}_j^{-1}\| \lesssim 2^{-j}$, which follows as a special case from (7.2), and that the mappings $z \mapsto |\partial\kappa_q(z)|$ are smooth, one easily verifies that for $x \in J_j^\circ$,

$$\|\psi_{j,x} - \psi_{j,x}^{(0)}\| \lesssim 2^{-j}.$$

Since for $x \in J_j \backslash J_j^\circ$, $\|\psi_{j,x} - \psi_{j,x}^{\mathrm{Jc}}\| \lesssim 2^{-\tilde{d}j}$, and $\breve{\Psi}_j$ is uniformly local, in view of Theorem 6.1 and Proposition 7.3 the only thing left to show is that $\breve{\Psi}_j$ has the uniform cancellation property of order $\tilde{d}$. Knowing this for Ψ_j^{Jc}, we have only to consider $\psi_{j,x}^{(0)}$ for $x \in J_j^\circ$.

Let $x \in J_j^\circ$, say, $x \in \Gamma_q$, and let u be a globally continuous, patchwise smooth function on Γ that is zero on $\partial\Gamma_D$. Let v be some arbitrary extension of the mapping $x \mapsto u(x)|\partial\kappa_q(\kappa_q^{-1}(x))|$ on Γ_q to a globally continuous, patchwise smooth function on Γ that is zero on $\partial\Gamma_D$. Since $\operatorname{supp}\psi_{j,x}^{(0)} \subset \overline{\Gamma_q}$, from Proposition 4.1 at the dual side we have

$$\begin{aligned}
|\langle \psi_{j,x}^{(0)}, u\rangle_{L_2(\Gamma)}| &= |\langle\!\langle \psi_{j,x}^{(0)}, v\rangle\!\rangle_0| = |\langle\!\langle \psi_{j,x}^{(0)}, ((Id - \tilde{P}_j)v)|_{\operatorname{supp}\psi_{j,x}^{(0)}}\rangle\!\rangle_0| \\
&\lesssim 2^{-\tilde{d}j}|v \circ \kappa_q|_{H^{\tilde{d}}(B(\kappa_q^{-1}(\operatorname{supp}\psi_{j,x}^{(0)});(\tilde{\vartheta}+3\tilde{\varepsilon})2^{-j})\cap\square)} \\
&\lesssim 2^{-\tilde{d}j}\|u \circ \kappa_q\|_{H^{\tilde{d}}(B(\kappa_q^{-1}(\operatorname{supp}\psi_{j,x}^{(0)});(\tilde{\vartheta}+3\tilde{\varepsilon})2^{-j})\cap\square)},
\end{aligned}$$

which completes the proof. ∎

Note that for $x \in J_j^\circ$, generally it holds only that $\|\psi_{j,x} - \psi_{j,x}^{(0)}\|_{L_2(\Gamma)} \eqsim 2^{-j}$. So in contrast to Ψ_j^{Jc}, for the approximate wavelets $\breve{\Psi}_j$ from this subsection, generally $\|\Psi_j - \breve{\Psi}_j\|_{L_2(\Gamma)} \not\lesssim 2^{-\tilde{d}j}$ when $\tilde{d} > 1$. The same will hold true for the collections $\breve{\Psi}_j$ that will be constructed in the next subsection. As follows from Theorem 6.1, however, this fact does not limit the range of s for which the approximate wavelets generate a Riesz basis for $\mathcal{H}^s(\Gamma)$.

7.3. Approximate wavelets with small supports near the interfaces. As we saw in the previous subsection, for $x \in J_j^\circ$ we can replace $\psi_{j,x}^{\mathrm{Jc}}$ by $\psi_{j,x}^{(0)}$, which, for $\tilde{d} \geq 2$, has a much smaller support. In this subsection, we investigate whether also near the interfaces we can find appropriate approximate wavelets $\breve{\psi}_{j,x}$ with smaller supports.

We set

$$\mathbb{P}_{\tilde{d}-2}(\Gamma) := C(\Gamma) \cap \prod_{q=1}^{M} \kappa_q(\mathbb{P}_{\tilde{d}-2}(\square)),$$

where when $\square$ is the interior of an n-simplex, $\mathbb{P}_{\tilde{d}-2}(\square) := P_{\tilde{d}-2}(\square)$, and when $\square = (0,1)^n$, $\mathbb{P}_{\tilde{d}-2}(\square) := Q_{\tilde{d}-2}(\square)$, being the tensor product space of the univariate polynomial spaces $P_{\tilde{d}-2}(0,1)$ in the n coordinate directions. In the latter case, in addition to the assumption that $P_{\tilde{d}-1}(\square) \subset \mathcal{S}(\tilde{\Phi}_j^\square)$ ((𝒥)(ii) at the dual side), in this subsection we assume that

(𝒥̃e) $$Q_{\tilde{d}-2}(\square) \subset \mathcal{S}(\tilde{\Phi}_j^\square).$$

For $z \in \Gamma$ and $\varepsilon \geq 0$, let $B_\Gamma(z;\varepsilon) = \{y \in \overline{\Gamma} : d_\Gamma(z,y) \leq \varepsilon\}$. With, for some constant $\rho \geq 0$, setting

$$\tilde{V}_{j,x,\rho} = \Big\{v \in \mathbb{P}_{\tilde{d}-2}(\Gamma) : v|_{\partial\Gamma_D \cap B_\Gamma(x;\rho 2^{-j})} = 0\Big\},$$

for $x \in J_j \backslash J_j^\circ$ we will search

$$\breve{\psi}_{j,x} \perp_{L_2(\Gamma)} \tilde{V}_{j,x,\rho} \text{ with } \|\breve{\psi}_{j,x} - \hat{\psi}_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}. \tag{7.3}$$

We note that by taking $\tilde{V}_{j,x,\rho}$ to be the smaller set $\{v \in \mathbb{P}_{\tilde{d}-2}(\Gamma) : v|_{\partial\Gamma_D} = 0\}$, $\breve{\psi}_{j,x}$ would not necessarily have the cancellation property of order $\tilde{d}$, and on the other hand, as we will see later, without incorporating boundary conditions in the definition of $\tilde{V}_{j,x,\rho}$, generally we cannot expect that $\|\breve{\psi}_{j,x} - \psi_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$. For the moment assuming that such $\breve{\psi}_{j,x}$ can be found, a topic that will be treated later in this subsection, Theorem 7.5 shows that they have the uniform cancellation property of order $\tilde{d}$, which, by Theorem 6.1, additionally yields the Riesz basis property for the full range of s. This may look surprising since, ignoring boundary conditions, the condition $\breve{\psi}_{j,x} \perp_{L_2(\Gamma)} \tilde{V}_{j,x,\rho}$ seems only to imply the uniform cancellation property of order $\tilde{d}-1$. Yet, given a $u \in \mathcal{H}_{\tilde{d}}(\Gamma)$, an interpolant $v \in \tilde{V}_{j,x,\rho}$ can be constructed such that $\|u-v\|_{L_2(\operatorname{supp}\breve{\psi}_{j,x})} \lesssim 2^{-(\tilde{d}-1)j}$ and, using $(\tilde{\partial}\mathrm{e})$, such that $(Id - \tilde{P}_j)v$ vanishes on $\operatorname{supp}\breve{\psi}_{j,x}$. Now by writing

$$\begin{aligned}\langle\breve{\psi}_{j,x}, u\rangle_{L_2(\Gamma)} &= \langle\breve{\psi}_{j,x}, u-v\rangle_{L_2(\Gamma)} = \langle\breve{\psi}_{j,x}, (Id-\tilde{P}_j)u\rangle_{L_2(\Gamma)} + \langle\breve{\psi}_{j,x}, \tilde{P}_j(u-v)\rangle_{L_2(\Gamma)}\\ &= \langle\breve{\psi}_{j,x}, (Id-\tilde{P}_j)u\rangle_{L_2(\Gamma)} + \langle\breve{\psi}_{j,x}-\psi_{j,x}, \tilde{P}_j(u-v)\rangle_{L_2(\Gamma)},\end{aligned}$$

and, for the second term, using the additional factor 2^{-j} from $\|\breve{\psi}_{j,x}-\psi_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$, the cancellation property of order $\tilde{d}$ can be shown. For a detailed proof of Theorem 7.5, we refer the reader to [22].

THEOREM 7.5. *Let $\breve{\Psi}_j = \{\breve{\psi}_{j,x} : x \in J_j\} \subset \mathcal{S}(\Phi_{j+1})$ be a uniformly local set, with $\breve{\psi}_{j,x} = \psi^{(0)}_{j,x}$ when $x \in J_j^\circ$, and such that for $x \in J_j\backslash J_j^\circ$, (7.3) is valid. Then, for $j \geq j_0$ large enough, $\breve{\Psi}_j$ has the uniform cancellation property of order $\tilde{d}$, and for $s \in (-\min\{\tilde{\gamma},\tilde{d}\}, \min\{\gamma,d\})$,*

$$\Phi_{j_0} \cup \cup_{j\geq j_0} 2^{-sj}\breve{\Psi}_j \text{ *is a Riesz basis for* } \mathcal{H}_s(\Gamma).$$

In view of (2.4)*, when in addition $|s| < \frac{3}{2}$, $|s| < s_\Gamma \notin \mathbb{N}$, or $|s| \leq s_\Gamma \in \mathbb{N}$, this collection is thus a Riesz basis for $\mathcal{H}^s(\Gamma)$.*

Next, we discuss a *construction* of $\breve{\Psi}_j$ as in Theorem 7.5. Consider, for $x \in J_j\backslash J_j^\circ$, the first order approximation $\hat{\psi}_{j,x}$ for $\psi_{j,x}$ from the collection

$$\hat{\Psi}_j := \Xi_j - \langle\Xi_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}\mathbf{D}_j^{-1}\Theta_j. \tag{7.4}$$

As a special case of (7.2), we have $\|\psi_{j,x} - \hat{\psi}_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$, where generally $\hat{\psi}_{j,x}$ only has the cancellation property of order 1. We will construct $\breve{\psi}_{j,x}$ from $\hat{\psi}_{j,x}$ by adding correction terms. In view of our requirement that $\|\psi_{j,x} - \breve{\psi}_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$, we first show that $\hat{\psi}_{j,x}$ is already nearly orthogonal to $\tilde{V}_{j,x,\rho}$, so that the correction can be small. For this to be true, the incorporation of boundary conditions in the definition of $\tilde{V}_{j,x,\rho}$ is essential.

We set $\bar{I}_j = \cup_{q=1}^{M} \kappa_q(I_j^{\square})$, i.e., without the exclusion of possible points on $\partial\Gamma_D$, and define $\tilde{\phi}_{j,y}$ and $\tilde{\lambda}_{j,y}$ similarly as in (4.1) and Proposition 4.1, respectively, but now also for $y \in \bar{I}_j \backslash I_j$.

LEMMA 7.6. *Let the constant ρ in the definition of $\tilde{V}_{j,x,\rho}$ be sufficiently large such that for all $x \in J_j$ and $y \in \bar{I}_j$ with $\operatorname{supp}\tilde{\phi}_{j,y} \cap \operatorname{supp}\hat{\psi}_{j,x} \neq \emptyset$, $\operatorname{supp}\tilde{\lambda}_{j,y} \subset B_\Gamma(x; \rho 2^{-j})$. Then*

$$|\langle \hat{\psi}_{j,x}, p\rangle_{L_2(\Gamma)}| \lesssim 2^{-j} \|p\|_{L_2(\operatorname{supp}\hat{\psi}_{j,x})} \quad (p \in \tilde{V}_{j,x,\rho}).$$

Proof. Since $p \in \mathbb{P}_{\tilde{d}-2}(\Gamma)$, by the inclusion of possible points on $\partial\Gamma_D$ and $(\tilde{\mathcal{J}}\mathrm{e})$, we have $p = \sum_{y\in\bar{I}_j} \langle p, \tilde{\lambda}_{j,y}\rangle_{L_2(\Gamma)} \tilde{\phi}_{j,y}$. Terms in this sum for $y \in \bar{I}_j \backslash I_j$ vanish on $\operatorname{supp}\hat{\psi}_{j,x}$ by $(\mathcal{J})$(ii) and because p vanishes on $\partial\Gamma_D \cap B_\Gamma(x; \rho 2^{-j})$. Setting $p_{j,x} = \sum_{\{y\in I_j : \operatorname{supp}\tilde{\phi}_{j,y} \cap \operatorname{supp}\hat{\psi}_{j,x} \neq \emptyset\}} \langle p, \tilde{\lambda}_{j,y}\rangle_{L_2(\Gamma)} \tilde{\phi}_{j,y}$, which is a function in $\mathcal{S}(\tilde{\Phi}_j)$, we find that

$$\begin{aligned} |\langle \hat{\psi}_{j,x}, p\rangle_{L_2(\Gamma)}| = |\langle \hat{\psi}_{j,x}, p_{j,x}\rangle_{L_2(\Gamma)}| &= |\langle \hat{\psi}_{j,x} - \psi_{j,x}, p_{j,x}\rangle_{L_2(\Gamma)}| \\ &\lesssim 2^{-j} \|p_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j} \|p\|_{L_2(\operatorname{supp}\hat{\psi}_{j,x})}, \end{aligned}$$

where in the last step we used $(\mathcal{J})$(iv). ∎

Let us first consider the *special case* $\tilde{d} = 2$ making the natural assumption that $|\int \xi_{j,x} d\mu| \gtrsim 2^{-jn/2}$. Let ρ be as in Lemma 7.6. If $x \in J_j \backslash J_j^\circ$ is such that $\partial\Gamma_D \cap B_\Gamma(x; \rho 2^{-j}) \neq \emptyset$, then $\tilde{V}_{j,x,\rho} = \{0\}$, and we can take $\breve{\psi}_{j,x} = \hat{\psi}_{j,x}$. Otherwise, we take $\breve{\psi}_{j,x} := \hat{\psi}_{j,x} - [\int_\Gamma \hat{\psi}_{j,x} d\mu / \int_\Gamma \xi_{j,x} d\mu]\xi_{j,x}$. Obviously $\breve{\psi}_{j,x} \perp_{L_2(\Gamma)} 1$, i.e., $\breve{\psi}_{j,x} \perp \tilde{V}_{j,x,\rho}$, and Lemma 7.6 shows that $|\int_\Gamma \hat{\psi}_{j,x} d\mu| \lesssim 2^{-j(1+n/2)}$, so that indeed $\|\hat{\psi}_{j,x} - \breve{\psi}_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$. In view of our aim to replace $\psi_{j,x}^{\mathrm{Jc}}$ for $x \in J_j \backslash J_j^\circ$ by an approximate wavelet with smaller support, note that the support of $\breve{\psi}_{j,x}$ is *equal* to that of $\hat{\psi}_{j,x}$ (which is equal to that of $\psi_{j,x}^{(0)}$).

For $\tilde{d} > 2$, generally we have to add more than one degree of freedom to find a correction of $\hat{\psi}_{j,x}$ that is orthogonal to $\tilde{V}_{j,x,\rho}$. We will search the correction from the span of $\theta_{j,y}$ with $d_\Gamma(x,y) \lesssim 2^{-j}$. Instead of adding as many degrees of freedom as $\dim(\tilde{V}_{j,x,\rho})$, generally we add more degrees of freedom, but then solve the resulting underdetermined problem in a minimal norm sense to end up with a correction term that is as small as possible. The resulting approximate wavelets will be denoted as $\psi_{j,x}^{\mathrm{ls}}$, where "ls" refers to least squares. In Theorem 7.7 it is stated that if, for sufficiently large δ, we use all $\theta_{j,y}$ for $y \in I_j$ with $d_\Gamma(x,y) \leq \delta 2^{-j}$, then the constrained minimization problem has a unique solution $\psi_{j,x}^{\mathrm{ls}}$, with $\|\psi_{j,x}^{\mathrm{ls}} - \psi_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$. Note that although in our numerical example we end up with $\psi_{j,x}^{\mathrm{ls}}$, which has the *same* support as $\hat{\psi}_{j,x}$, which thus in particular is much smaller than the support of $\psi_{j,x}^{\mathrm{Jc}}$, we cannot prove this in general.

THEOREM 7.7. *Let ρ be as in Lemma 7.6. For a sufficiently large constant $\delta > 0$, and with $\Theta_{j,x}^{\delta} := \{\theta_{j,y} : y \in I_j \cap B_\Gamma(x; \delta 2^{-j})\}$, for any $x \in J_j \backslash J_j^\circ$ the problem of determining*

$$\operatorname*{argmin}_{\psi_{j,x}^{\mathrm{ls}} \in \hat{\psi}_{j,x} + \mathcal{S}(\Theta_{j,x}^{\delta})} \{\|\psi_{j,x}^{\mathrm{ls}} - \hat{\psi}_{j,x}\|_{L_2(\Gamma)} : \psi_{j,x}^{\mathrm{ls}} \perp_{L_2(\Gamma)} \tilde{V}_{j,x,\rho}\}$$

has a unique solution with $\|\psi_{j,x}^{\mathrm{ls}} - \psi_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$, so that Theorem 7.5 applies.

For a proof of this theorem we refer the reader to [22]. From the theory of saddle point problems we know that the constrained minimization problem has a solution if and only if for each $p \in \tilde{V}_{j,x,\rho}$ there exists a $v_j \in \mathcal{S}(\Theta^{\delta}_{j,x})$ with $\langle p, v_j\rangle_{L_2(\Gamma)} > 0$. It can be observed that on a sufficiently large neighborhood of $\operatorname{supp}\hat{\psi}_{j,x}$ with diameter $\eqsim 2^{-j}$, any $p \in \mathbb{P}_{\tilde{d}-2}$ can be well approximated by a linear combination of those $\tilde{\phi}_{j,y}$ ($y \in I_j$) that are fully supported in this neighborhood. Since $\langle \Theta_j, \tilde{\Phi}_j\rangle_{L_2(\Gamma)}$ is nearly diagonal, for any function w_j in the span of these $\tilde{\phi}_{j,y}$, a v_j can be found in the span of the $\theta_{j,y}$, for the same set of y, with $\langle w_j, v_j\rangle_{L_2(\Gamma)} > 0$. By combining both properties, the existence and uniqueness of the constrained minimization problem can be inferred. The estimate $\|\psi^{\mathrm{ls}}_{j,x} - \psi_{j,x}\|_{L_2(\Gamma)} \lesssim 2^{-j}$ can be derived from the fact that $\hat{\psi}_{j,x}$ is already nearly orthogonal to $\tilde{V}_{j,x,\rho}$, as was shown in Lemma 7.6.

Remark 7.8. In case $\Theta_j \neq \Phi_j$, it is not of any interest that $\breve{\Psi}_j - \Xi_j \in \mathcal{S}(\Theta_j)$ (cf. Remark 6.2). In that case, in Theorem 7.7 we may search $\psi^{\mathrm{ls}}_{j,x}$ in the larger space $\hat{\psi}_{j,x} + \mathcal{S}(\Phi^{\delta}_{j+1,x})$, with $\Phi^{\delta}_{j+1,x} := \{\phi_{j+1,y} : x \in I_{j+1} \cap B_{\Gamma}(x; \delta 2^{-j})\}$, which opens the possibility that we may take a smaller δ, and so reduce the support of the resulting $\psi^{\mathrm{ls}}_{j,x}$.

Remark 7.9. Both the construction of $\psi^{\mathrm{Jc}}_{j,x}$ from Theorem 7.1, and that of $\psi^{\mathrm{ls}}_{j,x}$ from Theorem 7.7 requires the evaluation of $L_2(\Gamma)$-scalar products. For general parametrizations κ_q, these scalar products cannot be evaluated exactly, and therefore have to be approximated using numerical quadrature. Theorem 6.1 shows that if the quadrature is organized such that it causes an $L_2(\Gamma)$-error $\lesssim 2^{-\tilde{d}j}$ in the resulting approximate wavelet, then all results concerning cancellation properties and the generation of Riesz bases remain valid.

8. Numerical example. We consider a 2-dimensional Lipschitz manifold $\Gamma = \cup_{i=1}^{4}\overline{\Gamma_q} \subset \mathbb{R}^3$ as illustrated in Figure 8.1, which, together with its parametrization that satisfies ($\mathcal{M}$) is defined as follows. Let P be a tetrahedron in $\mathbb{R}^3$, with vertices on the unit sphere, geometric centroid in $(0,0,0)$, and one of its four facets $F_1, \ldots, F_4$, say, F_4, parallel to and below the $x_3 = 0$ plane. Let $\square$ be the interior of a reference 2-simplex in $\mathbb{R}^2$, with $\mathrm{vol}(\square) = 1$, and for $1 \leq q \leq 4$, let $B_q : \square \to F_q$ some affine bijection. The parametrizations $\kappa_q : \square \to \Gamma_q$ are defined by $\kappa_q(z) = B_q(z)/\|B_q(z)\|$ for $1 \leq q \leq 3$, and by $\kappa_4(z) = B_q(z)/\|B_q(z)\| - \frac{27}{4}(0,0,\lambda_1(z)\lambda_2(z)\lambda_3(z))$, where $(\lambda_1(z), \lambda_2(z), \lambda_3(z))$ are the barycentric coordinates of z with respect to $\square$. So without the perturbation by this cubic bubble, Γ would be the unit sphere. We added this perturbation term so that $J : x \mapsto |\partial\kappa_q(\kappa_q^{-1}(x))|$ when $x \in \Gamma_q$ cannot be extended to a continuous function

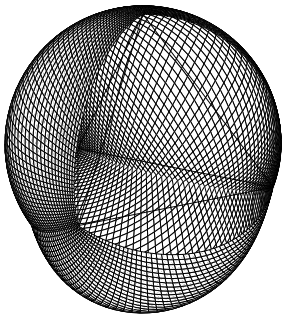


FIG. 8.1. *The manifold Γ, excluding one of the patches F_1, F_2, or F_3, and the tetrahedron P.*

on Γ. This means that constructions based on ignoring the Jacobian determinants will yield wavelets of which those that have supports that intersect an interface between Γ_4 and one of the three other patches have no cancellation properties. Furthermore, in view of our discussion at the beginning of section 6, note that $z \mapsto |\partial\kappa_q(z)|$ are not constant functions.

We consider two examples of collections $\Phi_j^\square$, $\tilde{\Phi}_j^\square$, $\Theta_j^\square$, $\Xi_j^\square$, $\Lambda_j^\square$, $\tilde{\Lambda}_j^\square$, both based on the construction of finite element wavelets from [14]. These collections satisfy all conditions formulated in section 3, where $\square$ is a reference 2-simplex, $\gamma = \tilde{\gamma} = \frac{3}{2}$, $d = 2$, and $\tilde{d} = 2$ or $\tilde{d} = 3$. With S_j being the standard Lagrange finite element space of order 2 with respect to a j times repeated uniform dyadic refinement of $\square$, for $\tilde{d} = 2$, $\mathcal{S}(\Phi_j^\square) = S_j$, and for $\tilde{d} = 3$, $\mathcal{S}(\Phi_j^\square) = S_{j+1}$. The index set $I_j^\square$ is the set of all vertices in the triangulation underlying the finite element space, and $J_j^\square = I_{j+1}^\square \backslash I_j^\square$. For more details, we refer the reader to [22].

We implemented the approximate wavelet constructions from sections 7.2 and 7.3,which away from the patch interfaces both yield the approximate wavelets from the collection $\Psi_j^{(0)} = \Xi_j - \langle\!\langle \Xi_j, \tilde{\Phi}_j \rangle\!\rangle_0 \langle\!\langle \Theta_j, \tilde{\Phi}_j \rangle\!\rangle_0^{-1} \Theta_j$ obtained by ignoring the Jacobian determinants. The pull-backs of these wavelets to the parameter domain are illustrated in Figure 8.2; the functions are continuous piecewise linear with respect to the indicated triangulation.

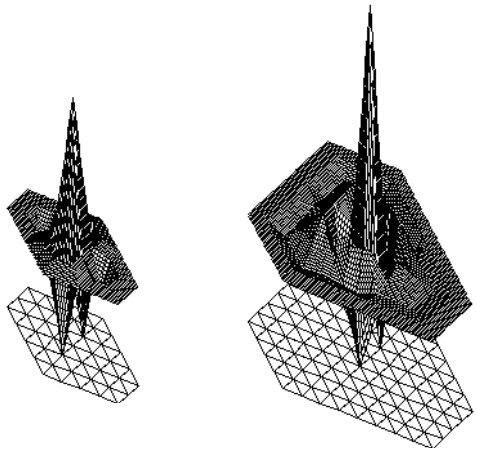


FIG. 8.2. *Wavelets $\psi_{j,x}^{(0)}$ away from the patch interfaces for $\tilde{d} = 2$ and $\tilde{d} = 3$ (one of the two different types), and their supports in terms of the underlying triangulation.*

With the approach from section 7.2, wavelets along the patch interfaces are taken from the collection $\Psi_j^{\mathrm{Jc}} := \Xi_j - \langle \Xi_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)} \big[\sum_{k=0}^{\tilde{d}-1} (Id - \mathbf{D}_j^{-1} \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)})^k \mathbf{D}_j^{-1} \big] \Theta_j$, where $\mathbf{D}_j := \mathrm{diag} \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}$. Illustrations of the naturally joined, patchwise pull-backs of these wavelets can be found in Figure 8.3. For $j = 0$, the Neumann series does not converge, and as a consequence $\sum_{k=0}^{\tilde{d}-1} (Id - \mathbf{D}_j^{-1} \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)})^k \mathbf{D}_j^{-1}$ provides a very poor approximation for $\langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}^{-1}$. We redefined $\breve{\Psi}_0 := \Xi_0 - \langle \Xi_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)} \langle \Theta_j, \tilde{\Phi}_j \rangle_{L_2(\Gamma)}^{-1} \Theta_0$.

For the construction from section 7.3, for each $x \in J_j \backslash J_j^\circ$ we have to specify a

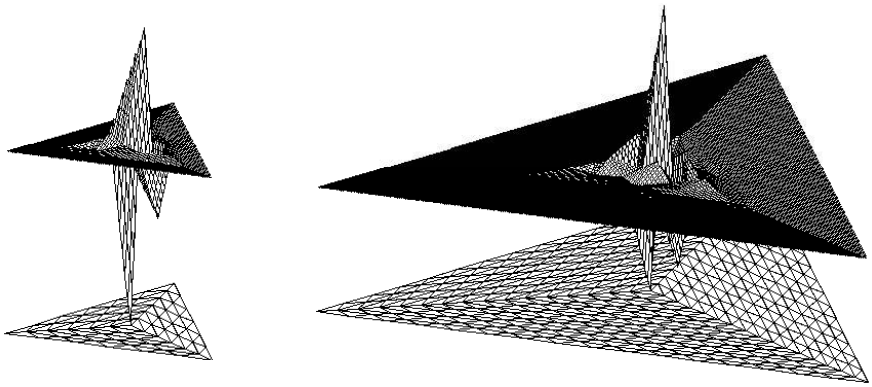


FIG. 8.3. *Wavelets $\psi^{\mathrm{Jc}}_{4,x}$ for $\tilde{d}=2$ and $\psi^{\mathrm{Jc}}_{3,x}$ for $\tilde{d}=3$ near the "north pole," and their supports in terms of the underlying triangulation.*

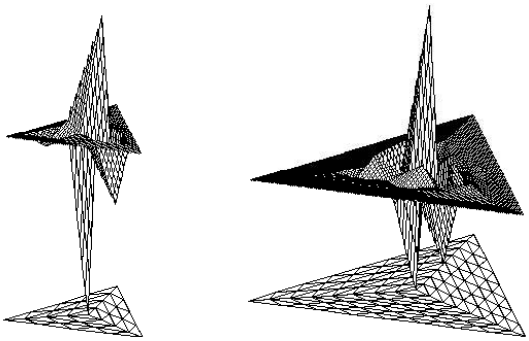


FIG. 8.4. *Wavelets $\psi^{\mathrm{ls}}_{4,x}$ for $\tilde{d}=2$ and $\psi^{\mathrm{ls}}_{3,x}$ for $\tilde{d}=3$ near the "north pole," and their supports in terms of the underlying triangulation.*

subspace $A_{j,x}\subset\mathcal{S}(\Phi_{j+1})$ that defines $\psi^{\mathrm{ls}}_{j,x}$ via

$$\operatorname*{argmin}_{\psi^{\mathrm{ls}}_{j,x}\in\hat{\psi}_{j,x}+A_{j,x}}\{\|\psi^{\mathrm{ls}}_{j,x}-\hat{\psi}_{j,x}\|_{L_2(\Gamma)}:\psi^{\mathrm{ls}}_{j,x}\perp_{L_2(\Gamma)}\mathbb{P}_{\tilde{d}-2}(\Gamma)\}.$$

For $\tilde{d}=2$, we took $A_{j,x}=\mathcal{S}(\{\xi_{j,x}\})$, so that $\psi^{\mathrm{ls}}_{j,x}=\hat{\psi}_{j,x}+\alpha\xi_{j,x}$ with α such that $\int_\Gamma\psi^{\mathrm{ls}}_{j,x}=0$. For $\tilde{d}=3$, we took $A_{j,x}=\mathcal{S}(\{\phi_{j+1,y}:\operatorname{supp}\phi_{j+1,y}\subset\operatorname{supp}\hat{\psi}_{j,x}\})$, where the space turns out to be sufficiently large so that the constrained minimization problem has a solution $\psi^{\mathrm{ls}}_{j,x}$, with $\|\psi^{\mathrm{ls}}_{j,x}-\hat{\psi}_{j,x}\|_{L_2(\Gamma)}\lesssim 2^{-j}$. The naturally joined, patchwise pull-backs of the resulting $\psi^{\mathrm{ls}}_{j,x}$ are illustrated in Figure 8.4. By definition they have the *same supports* as the corresponding $\psi^{(0)}_{j,x}$ that one obtains by ignoring the Jacobian determinants also along the interfaces. As with the Jacobi iteration approach, for $j=0$, we redefined $\breve{\Psi}_0:=\Xi_0-\langle\Xi_j,\tilde{\Phi}_j\rangle_{L_2(\Gamma)}\langle\Theta_j,\tilde{\Phi}_j\rangle^{-1}_{L_2(\Gamma)}\Theta_0$.

Let

$$\breve{\Psi}^{(j)}:=\begin{cases}\Phi_0\cup\cup_{k=0}^{j-1}\breve{\Psi}_k & \text{when } \tilde{d}=2,\\ \Phi_0\cup\cup_{k=0}^{j-2}\breve{\Psi}_k & \text{when } \tilde{d}=3,\end{cases}$$

in both cases being a basis for $C(\Gamma)\cap\prod_{q=1}^4\kappa_q(S_j)$. With

$$\kappa_{\Sigma,\|\cdot\|}:=\sup_{0\neq\mathbf{c}=(c_\sigma)_{\sigma\in\Sigma}}\frac{\left\|\sum_{\sigma\in\Sigma}c_\sigma\frac{\sigma}{\|\sigma\|}\right\|^2}{\|\mathbf{c}\|^2}\Bigg/\inf_{0\neq\mathbf{c}=(c_\sigma)_{\sigma\in\Sigma}}\frac{\left\|\sum_{\sigma\in\Sigma}c_\sigma\frac{\sigma}{\|\sigma\|}\right\|^2}{\|\mathbf{c}\|^2},$$

which, in case $\|\cdot\|$ corresponds to a scalar product, is the spectral norm of the corresponding mass matrix, for $j\leq 6$ we computed $\kappa_{\breve\Psi^{(j)},\|\cdot\|_{L_2(\Gamma)}}$, $\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_1}$, and $\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_{-1,6}}$, where on $C(\Gamma)\cap\prod_{q=1}^4\kappa_q(S_m)$, $|\!|\!|u|\!|\!|_{-1,m}:=\sup_{v\in C(\Gamma)\cap\prod_{q=1}^4\kappa_q(S_m)}\frac{|\langle u,v\rangle_{L_2(\Gamma)}|}{|\!|\!|v|\!|\!|_1}$. The uniform boundedness in $|\!|\!|\cdot|\!|\!|_1$ of the $L_2(\Gamma)$-orthogonal projector onto $C(\Gamma)\cap\prod_{q=1}^4\kappa_q(S_m)$, which is a consequence of Theorem 5.3, shows that $|\!|\!|\cdot|\!|\!|_{-1}\eqsim|\!|\!|\cdot|\!|\!|_{-1,m}$ on $C(\Gamma)\cap\prod_{q=1}^4\kappa_q(S_j)$ uniformly in m and $j\leq m$.

Since the functions from Φ_0, and for $\tilde d=2$, from $\breve\Psi_0$ and $\breve\Psi_1$, and for $\tilde d=3$, from $\breve\Psi_0$ have global supports anyway, for computing the condition numbers we replaced each of these collections by an orthonormalized version with respect to $\langle\cdot,\cdot\rangle_{L_2(\Gamma)}$, $\langle\!\langle\cdot,\cdot\rangle\!\rangle_1$, or $\langle\!\langle\cdot,\cdot\rangle\!\rangle_{-1,6}$, the latter being the scalar product corresponding to the norm $\|\cdot\|_{-1,6}$. The results for the Jacobi approximation or least squares approximation along the interfaces are presented in Tables 8.1 and 8.2, respectively. Although it turns out that unfortunately, in particular in the $\|\cdot\|_{-1,6}$-norm, the condition numbers were not completely stabilized yet, we stopped our computations at $j=6$ since mainly due to the normalization of the wavelets, in particular with respect to the $\|\cdot\|_{-1,6}$-norm, on higher levels they became too time consuming.

TABLE 8.1
Condition numbers for $\tilde d=2$.

	Jacobi approximation			Least squares approximation		
j	$\kappa_{\Psi^{(j)},\|\cdot\|_{L_2(\Gamma)}}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_1}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_{-1,6}}$	$\kappa_{\breve\Psi^{(j)},\|\cdot\|_{L_2(\Gamma)}}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_1}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_{-1,6}}$
1	1.35e0	2.30e0	3.06e0	1.35e0	2.30e0	3.09e0
2	1.45e0	7.86e0	8.65e0	1.82e0	7.64e0	1.09e1
3	1.75e1	4.82e1	2.79e1	2.42e1	6.84e1	3.41e1
4	1.79e1	6.51e1	4.53e1	2.46e1	9.06e1	5.21e1
5	1.79e1	7.66e1	6.20e1	2.46e1	1.06e2	6.69e1
6	1.79e1	8.25e1	7.34e1	2.46e1	1.14e2	7.84e1

TABLE 8.2
Condition numbers for $\tilde d=3$.

	Jacobi approximation			Least squares approximation		
j	$\kappa_{\Psi^{(j)},\|\cdot\|_{L_2(\Gamma)}}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_1}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_{-1,6}}$	$\kappa_{\breve\Psi^{(j)},\|\cdot\|_{L_2(\Gamma)}}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_1}$	$\kappa_{\breve\Psi^{(j)},|\!|\!|\cdot|\!|\!|_{-1,6}}$
2	2.80e0	1.26e1	4.18e1	2.80e0	1.26e1	4.22e0
3	1.36e1	5.47e1	3.36e1	1.85e1	8.51e1	3.82e1
4	1.81e1	7.88e1	5.76e1	2.34e1	1.19e2	6.86e1
5	1.97e1	8.83e1	8.00e1	2.44e1	1.32e2	9.65e1
6	2.03e1	9.20e1	9.64e1	2.47e1	1.36e2	1.34e2

Appendix. Proof of Theorem 6.1. The proof consists of steps (I)–(VI). Although basically (V) and (VI) can be found in [19, Theorem 3.1], which in turn was based on [24, appendix], for convenience we include a complete proof.

(I) Since, as shown in Proposition 5.4, for $j \geq j_0$ large enough, Ψ_j is a uniform $L_2(\Gamma)$-Riesz system, and by condition (iii),

$$\begin{aligned}&\|\langle\breve{\Psi}_j,\breve{\Psi}_j\rangle_{L_2(\Gamma)}-\langle\Psi_j,\Psi_j\rangle_{L_2(\Gamma)}\|\\&=\|\langle\breve{\Psi}_j-\Psi_j,\Psi_j\rangle_{L_2(\Gamma)}+\langle\Psi_j,\breve{\Psi}_j-\Psi_j\rangle_{L_2(\Gamma)}+\langle\breve{\Psi}_j-\Psi_j,\breve{\Psi}_j-\Psi_j\rangle_{L_2(\Gamma)}\|\lesssim\omega^j,\end{aligned}$$

we infer that, possibly for a larger value of j_0, for $j \geq j_0$, $\breve{\Psi}_j$ is a uniform $L_2(\Gamma)$-Riesz system.

(II) Next we investigate whether $\mathcal{S}(\Phi_j)+\mathcal{S}(\breve{\Psi}_j)$ is a uniformly $L_2(\Gamma)$-stable *two-level* decomposition of $\mathcal{S}(\Phi_{j+1})$.

PROPOSITION A.1. *Let V, W be subspaces of a Hilbert space H. Then the following are equivalent:*

(a) *$H = V + W$ and $\alpha := \sup_{0\neq v\in V,\,0\neq w\in W}\frac{|\langle v,w\rangle|}{\|v\|\|w\|}<1$.*

(b) *There exists a bounded projector $Q : H \to H$ with $\Im(Q) = V$ and $\Im(Id-Q) = W$.*

Moreover, $\|Q\| = (1-\alpha^2)^{-\frac{1}{2}}$.

Now let (a)*, or equivalently* (b)*, be satisfied, and let $\breve{W}$ be another subspace of H for which there exists a linear mapping $R : \breve{W} \to W$ with $\|Id - R\| < \frac{1-\alpha}{1+\alpha}$. Then*

$$\breve{\alpha} := \sup_{0\neq v\in V,\,0\neq\breve{w}\in\breve{W}}\frac{|\langle v,\breve{w}\rangle|}{\|v\|\|\breve{w}\|}\leq\alpha+(1+\alpha)\|Id-R\|<1.$$

With $\breve{Q} : H \to H$ being the bounded projector with $\Im(\breve{Q}) = V$ and $\Im(Id-\breve{Q}) = \breve{W}$, it holds that

$$\|Q-\breve{Q}\|\leq\frac{\|Id-R\|}{(1-\alpha^2)^{\frac{1}{2}}(1-\breve{\alpha}^2)^{\frac{1}{2}}}.$$

Proof. If (a) is valid, then $H = V \oplus W$, and so there exists a projector Q with $\Im(Q) = V$ and $\Im(Id - Q) = W$. For any $u \in H$, we have

$$\begin{aligned}\|u\|^2&=\|Qu+(Id-Q)u\|^2\geq\|Qu\|^2-2|\langle Qu,(Id-Q)u\rangle|+\|(Id-Q)u\|^2\\&\geq\|Qu\|^2-2\alpha\|Qu\|\|(Id-Q)u\|+\|(Id-Q)u\|^2\geq(1-\alpha^2)\|Qu\|^2,\end{aligned}$$

or $\|Q\| \leq (1-\alpha^2)^{-\frac{1}{2}}$.

Now let (b) be valid. Suppose there exist nonzero $v \in V$, $w \in W$ such that $\mu := \frac{|\langle v,w\rangle|}{\|v\|\|w\|} > (1-\|Q\|^{-2})^{\frac{1}{2}}$. Then there exist nonzero $v \in V$, $w \in W$ with $\langle v, w\rangle = -\mu\|v\|\|w\|$, moreover, which can be chosen such that $\|w\| = -\mu\|v\|$. From

$$\|Q\|^{-2}\|v\|^2\leq\|v+w\|^2=\|v\|^2+2\langle v,w\rangle+\|w\|^2=(1-\mu^2)\|v\|^2$$

we conclude a contradiction, so that $\sup_{0\neq v\in V,\,0\neq w\in W}\frac{|\langle v,w\rangle|}{\|v\|\|w\|}\leq(1-\|Q\|^{-2})^{\frac{1}{2}}$.

Now let (a) or, equivalently, (b) be valid, and let $\breve{W}$ be a subspace as in the proposition. For any $v \in V$, $\breve{w} \in \breve{W}$,

$$\begin{aligned}|\langle v,\breve{w}\rangle|&=|\langle v,R\breve{w}\rangle+\langle v,(Id-R)\breve{w}\rangle|\leq\alpha\|v\|\|R\breve{w}\|+\|v\|\|(Id-R)\breve{w}\|\\&\leq\alpha\|v\|\|\breve{w}\|+(1+\alpha)\|v\|\|(Id-R)\breve{w}\|\leq(\alpha+(1+\alpha)\|Id-R\|)\|v\|\|\breve{w}\|,\end{aligned}$$

showing the statement about $\breve{\alpha}$. The last statement follows from $\|Q\| = (1-\alpha^2)^{-\frac{1}{2}}$, $\|Id-\breve{Q}\| = (1-\breve{\alpha}^2)^{-\frac{1}{2}}$, and $Q-\breve{Q} = Q(Id-R)(Id-\breve{Q})$ by $QR = 0$ and $Q\breve{Q} = \breve{Q}$. ∎

Returning to the proof of Theorem 6.1, with, for $j \geq j_0$, $Q_j^{(j+1)} := Q_j|_{\mathcal{S}(\Phi_{j+1})}$, it holds that $\Im(Q_j^{(j+1)}) = \mathcal{S}(\Phi_j)$ and $\Im(Id - Q_j^{(j+1)}) = \mathcal{S}(\Psi_j)$. Setting $R_j : \mathcal{S}(\breve{\Psi}_j) \to \mathcal{S}(\Psi_j) : \mathbf{c}_j^T \breve{\Psi}_j \mapsto \mathbf{c}_j^T \Psi_j$, by condition (iii) and (I) we have

$$\|(Id - R_j)\mathbf{c}_j^T \breve{\Psi}_j\|_{L_2(\Gamma)} = \|\mathbf{c}_j^T(\breve{\Psi}_j - \Psi_j)\|_{L_2(\Gamma)} \leq \|\mathbf{c}\| \|\breve{\Psi}_j - \Psi_j\|_{L_2(\Gamma)} \lesssim \|\mathbf{c}_j^T \breve{\Psi}_j\|_{L_2(\Gamma)} \omega^j,$$

or $\|Id - R_j\|_{L_2(\Gamma)\to L_2(\Gamma)} \lesssim \omega^j$. From Proposition A.1 we conclude that, possibly for a larger value of j_0, for $j \geq j_0$ there exists a uniformly $L_2(\Gamma)$-bounded projector $\breve{Q}_j^{(j+1)} : \mathcal{S}(\Phi_{j+1}) \to \mathcal{S}(\Phi_{j+1})$ with

$$\Im(\breve{Q}_j^{(j+1)}) = \mathcal{S}(\Phi_j), \quad \Im(Id - \breve{Q}_j^{(j+1)}) = \mathcal{S}(\breve{\Psi}_j),$$

and

$$\|Q_j^{(j+1)} - \breve{Q}_j^{(j+1)}\|_{L_2(\Gamma)\to L_2(\Gamma)} \lesssim \omega^j. \tag{A.1}$$

(III) By condition (i), $\breve{\Psi}_j$ is uniformly local. Since furthermore, as shown in (I), $\breve{\Psi}_j$ is a uniform $L_2(\Gamma)$-Riesz system that, by condition (ii), has the uniform cancellation property of order $\tilde{d}$, for some $\tilde{\eta} \geq 0$ we have

$$\begin{aligned}
|\langle \mathbf{c}_j^T \breve{\Psi}_j, u\rangle_{L_2(\Gamma)}| &\leq \sum_{x\in J_j} |c_{j,x}||\langle \breve{\psi}_{j,x}, u\rangle_{L_2(\Gamma)}| \quad (\mathbf{c}_j \in \ell_2(J_j),\, u \in \mathcal{H}_{\tilde{d}}(\Gamma)) \\
&\lesssim 2^{-j\tilde{d}} \|\mathbf{c}_j\| \left[\sum_{x\in J_j} \sum_{q=1}^{M} \|u \circ \kappa_q\|^2_{H^{\tilde{d}}(B(\kappa_q^{-1}(\operatorname{supp} \breve{\psi}_{j,x} \cap \Gamma_q); \tilde{\eta} 2^{-j}) \cap \square)} \right]^{\frac{1}{2}} \\
&\lesssim 2^{-j\tilde{d}} \|\mathbf{c}_j^T \breve{\Psi}_j\|_{L_2(\Gamma)} |\!|\!| u |\!|\!|_{\tilde{d}},
\end{aligned}$$

or

$$|\!|\!| \cdot |\!|\!|_{-\tilde{d}} \lesssim 2^{-\tilde{d}j} \|\cdot\|_{L_2(\Gamma)} \quad \text{on } \mathcal{S}(\breve{\Psi}_j). \tag{A.2}$$

By the uniform $L_2(\Gamma)$-boundedness of $\breve{Q}_j^{(j+1)}$ for $j \geq j_0$, a direct consequence of the last result is that $|\!|\!| Id - \breve{Q}_j^{(j+1)} |\!|\!|_{0\to -\tilde{d}} \lesssim 2^{-\tilde{d}j}$. By the Jackson estimate at the dual side (4.6), and the uniform $L_2(\Gamma)$-boundedness of Q_j for $j \geq j_0$, we have $|\!|\!| Id - Q_j^{(j+1)} |\!|\!|_{0\to -\tilde{d}} \leq |\!|\!| Id - Q_j |\!|\!|_{0\to -\tilde{d}} \lesssim 2^{-\tilde{d}j}$, and so $|\!|\!| Q_j^{(j+1)} - \breve{Q}_j^{(j+1)} |\!|\!|_{0\to -\tilde{d}} \lesssim 2^{-\tilde{d}j}$. By the extended Bernstein inequality Lemma 4.2, $|\!|\!| Q_j^{(j+1)} - \breve{Q}_j^{(j+1)} |\!|\!|_{-\tilde{d}\to -\tilde{d}} \lesssim 1$, so that by interpolation using (A.1) we infer that

$$|\!|\!| Q_j^{(j+1)} - \breve{Q}_j^{(j+1)} |\!|\!|_{s\to s} \lesssim \omega^{(1+\frac{s}{\tilde{d}})j} \quad (s \in [-\tilde{d}, 0]). \tag{A.3}$$

(IV) Knowing (A.3), we are ready to investigate the stability of the *multilevel* decomposition defined by the collections $\breve{\Psi}_j$. We are going to construct a projector $\breve{Q}_j$ defined on $\mathcal{H}_s(\Gamma)$ for some range of s, such that, for $j \geq j_0$, $\Im(\breve{Q}_j) = \mathcal{S}(\Phi_j)$ and $\Im(\breve{Q}_{j+1} - \breve{Q}_j) = \mathcal{S}(\breve{\Psi}_j)$.

By writing $Q_j = \sum_{k=j_0}^{j} Q_k - Q_{k-1}$, for any $s \in (-\min\{\tilde{\gamma}, \tilde{d}\}, \min\{\gamma, d\})$ and $u \in \mathcal{H}_s(\Gamma)$, Theorem 5.3 shows that $|\!|\!| Q_j u |\!|\!|_s^2 \lesssim \sum_{k=j_0}^{j} 4^{sk} |\!|\!| (Q_k - Q_{k-1})u |\!|\!|_{L_2(\Gamma)}^2 \leq \sum_{k=j_0}^{\infty} 4^{sk} |\!|\!| (Q_k - Q_{k-1})u |\!|\!|_{L_2(\Gamma)}^2 \lesssim |\!|\!| u |\!|\!|_s^2$ or $|\!|\!| Q_j |\!|\!|_{s\to s} \lesssim 1$.

For $\ell \geq j \geq j_0$, we define $\breve{Q}_j^{(\ell)} : \mathcal{S}(\Phi_\ell) \to \mathcal{S}(\Phi_j)$ by $\breve{Q}_\ell^{(\ell)} = Id$ and, for $j < \ell$, by $\breve{Q}_j^{(\ell)} = \breve{Q}_j^{(j+1)}\breve{Q}_{j+1}^{(\ell)}$. For some arbitrary but fixed $t \in (-\min\{\tilde{\gamma}, \tilde{d}\}, 0]$, we set

$$\rho_j^{(\ell)} := \max_{j_0 \leq k \leq j} \|\!| Q_k \breve{Q}_j^{(\ell)} |\!\|_{t\to t}, \quad \varepsilon_j := \max_{j_0 \leq k \leq j} \|\!| Q_k(\breve{Q}_j^{(j+1)} - Q_j^{(j+1)}) |\!\|_{t\to t}.$$

From $Q_k\breve{Q}_j^{(\ell)} = Q_k(\breve{Q}_j^{(j+1)} - Q_j^{(j+1)})\breve{Q}_{j+1}^{(\ell)} + Q_k\breve{Q}_{j+1}^{(\ell)}$, we find $\rho_j^{(\ell)} \leq (\varepsilon_j + 1)\rho_{j+1}^{(\ell)}$, and so by $\rho_\ell^{(\ell)} = \max_{j_0 \leq k \leq j} \|\!| Q_k |\!\|_{t\to t} \lesssim 1$, and $\varepsilon_j \lesssim \omega^{(1+\frac{t}{d})j}$ by (A.3), we infer that

$$\|\!| \breve{Q}_j^{(\ell)} |\!\|_{t\to t} \leq \rho_j^{(\ell)} \lesssim \prod_{k=j}^{\ell-1}(\varepsilon_k + 1) \lesssim 1 + \sum_{k=j}^{\ell-1} \varepsilon_k \lesssim 1,$$

which thus holds uniformly in j and ℓ.

As a consequence of $Id = \sum_{j=j_0}^\infty (Q_j - Q_{j-1})$ on $\mathcal{H}_t(\Gamma)$ by Theorem 5.3, we have $\mathrm{clos}_{H_t(\Gamma)} \cup_{j\geq j_0} \mathcal{S}(\Phi_j) = \mathcal{H}_t(\Gamma)$. Since for any $u \in \mathcal{H}_t(\Gamma)$, $j \leq k \leq \ell$, and $u_k \in \mathcal{S}(\Phi_k)$, $\breve{Q}_j^{(\ell)}Q_\ell u = \breve{Q}_j^{(k)}u_k + \breve{Q}_j^{(\ell)}Q_\ell(u - u_k)$, from $\|\!| \breve{Q}_j^{(\ell)}Q_\ell |\!\|_{t\to t} \lesssim 1$ we infer that for any $j \geq j_0$, $(\breve{Q}_j^{(\ell)}Q_\ell u)_{\ell\geq j}$ is a Cauchy sequence in $\mathcal{H}_t(\Gamma)$, and we set $\breve{Q}_j u = \lim_{\ell\to\infty} \breve{Q}_j^{(\ell)}Q_\ell u$. We conclude that $\breve{Q}_j : \mathcal{H}_t(\Gamma) \to \mathcal{H}_t(\Gamma)$ is uniformly bounded, with $\Im(\breve{Q}_j) = \mathcal{S}(\Phi_j)$ and $\Im(\breve{Q}_{j+1} - \breve{Q}_j) = \mathcal{S}(\breve{\Psi}_j)$.

(V) Let $s \in (t, \min\{\gamma, d\})$. With $\breve{Q}_{j_0-1} := 0$, we are going to prove that

$$\sum_{j=j_0}^\infty 4^{sj} \|(\breve{Q}_j - \breve{Q}_{j-1})u\|_{L_2(\Gamma)}^2 \lesssim \|u\|_s^2 \quad (u \in \mathcal{H}_s(\Gamma)). \tag{A.4}$$

For $u \in \mathcal{H}_s(\Gamma)$, $w_\ell := (Q_\ell - Q_{\ell-1})u$, Theorem 6.1 shows that $u = \sum_{\ell=j_0}^\infty w_\ell$, $\|u\|_s^2 \eqsim \sum_{\ell=j_0}^\infty 4^{s\ell}\|w_\ell\|_{L_2(\Gamma)}^2$, and $\|w_\ell\|_t \lesssim 2^{t\ell}\|w_\ell\|_{L_2(\Gamma)}$. Since $\|\!| \breve{Q}_j |\!\|_{t\to 0} \lesssim 2^{-tj}$, which follows from $\|\!| \breve{Q}_j |\!\|_{t\to t} \lesssim 1$ and the extended Bernstein inequality Lemma 4.2, and $\breve{Q}_j - \breve{Q}_{j-1}$ vanishes on $\mathcal{S}(\Phi_{j-1})$, we arrive at

$$\begin{aligned}
\sum_{j=j_0}^\infty 4^{sj}\|(\breve{Q}_j - \breve{Q}_{j-1})u\|_{L_2(\Gamma)}^2 &= \sum_{j=j_0}^\infty \sum_{\ell,\ell'=j_0}^\infty 4^{sj}\langle(\breve{Q}_j - \breve{Q}_{j-1})w_\ell, (\breve{Q}_j - \breve{Q}_{j-1})w_{\ell'}\rangle_{L_2(\Gamma)} \\
&= \sum_{\ell,\ell'=j_0}^\infty \sum_{j=j_0}^{\min\{\ell,\ell'\}} 4^{sj}\langle(\breve{Q}_j - \breve{Q}_{j-1})w_\ell, (\breve{Q}_j - \breve{Q}_{j-1})w_{\ell'}\rangle_{L_2(\Gamma)} \\
&\lesssim \sum_{\ell,\ell'=j_0}^\infty \sum_{j=j_0}^{\min\{\ell,\ell'\}} 4^{sj}4^{-tj}\|w_\ell\|_t\|w_{\ell'}\|_t \lesssim \sum_{\ell,\ell'=j_0}^\infty 4^{(s-t)\min\{\ell,\ell'\}}\|w_\ell\|_t\|w_{\ell'}\|_t \\
&\lesssim \sum_{\ell,\ell'=j_0}^\infty 4^{(s-t)\min\{\ell,\ell'\}}2^{(t-s)(\ell+\ell')}(2^{s\ell}\|w_\ell\|_{L_2(\Gamma)})(2^{s\ell'}\|w_{\ell'}\|_{L_2(\Gamma)}) \\
&\lesssim \sum_{\ell=j_0}^\infty 4^{s\ell}\|w_\ell\|_{L_2(\Gamma)}^2 \eqsim \|u\|_s^2,
\end{aligned}$$

where we have used that the infinite matrix $[2^{(s-t)(2\min\{\ell,\ell'\}-\ell-\ell')}]_{\ell,\ell'\geq j_0}$ is bounded.

(VI) For $s \in [-\tilde{d}, \gamma)$, it holds that

$$\|\cdot\|_s \lesssim 2^{sj}\|\cdot\|_{L_2(\Gamma)} \quad \text{on } \Im(\breve{Q}_j - \breve{Q}_{j-1}).$$

Indeed for $s > 0$ this is just the Bernstein inequality (4.5), whereas for $s < 0$ it is a consequence of the extended Bernstein inequality Lemma 4.2 and (A.2). Now let $s \in (-\tilde{d}, \gamma)$, and let $\varepsilon > 0$ be such that $s \pm \varepsilon \in [-\tilde{d}, \gamma)$. Then for any $\breve{w}_j \in \Im(\breve{Q}_j - \breve{Q}_{j-1})$, with $\sum_{j=j_0}^{\infty} 4^{sj} \|\breve{w}_j\|^2_{L_2(\Gamma)} < \infty$, it holds that

$$\Big\| \sum_{j=j_0}^{\infty} \breve{w}_j \Big\|_s^2 = \sum_{j,j'=j_0}^{\infty} \langle\!\langle \breve{w}_j, \breve{w}_{j'} \rangle\!\rangle_s \lesssim \sum_{j=j_0}^{\infty} \sum_{j' \geq j} |\!|\!| \breve{w}_j |\!|\!|_{s+\varepsilon} |\!|\!| \breve{w}_{j'} |\!|\!|_{s-\varepsilon}$$
$$\lesssim \sum_{j=j_0}^{\infty} \sum_{j' \geq j} 2^{\varepsilon j} 2^{-\varepsilon j'} (2^{sj} \|\breve{w}_j\|_{L_2(\Gamma)}) (2^{sj'} \|\breve{w}_{j'}\|_{L_2(\Gamma)}) \lesssim \sum_{j=j_0}^{\infty} 4^{sj} \|\breve{w}_j\|^2_{L_2(\Gamma)}.$$

Since, when $s \in (t, \min\{\gamma, d\})$, (A.4) shows that $\sum_{j=j_0}^{\infty} 4^{sj} \|\breve{w}_j\|^2_{L_2(\Gamma)} \lesssim |\!|\!| \sum_{j=j_0}^{\infty} \breve{w}_j |\!|\!|^2_s$, and Φ_{j_0} is an $L_2(\Gamma)$-Riesz basis for $\Im(\breve{Q}_{j_0})$, and for $j > j_0$, $\breve{\Psi}_{j-1}$ is a uniform $L_2(\Gamma)$-Riesz basis for $\Im(\breve{Q}_j - \breve{Q}_{j-1})$, we conclude that $\Phi_{j_0} \cup \cup_{j \geq j_0} 2^{-sj} \breve{\Psi}_j$ is a Riesz system in $\mathcal{H}_s(\Gamma)$. Finally, since, as follows from Theorem 5.3, $\mathrm{clos}_{H_s(\Gamma)} \cup_{j \geq 0} \mathcal{S}(\Phi_j) = \mathcal{H}_s(\Gamma)$, we conclude it is even Riesz basis for this space, with which the proof of Theorem 6.1 is completed. □

Acknowledgment. The author is indebted to H. Nguyen for the computation of the numerical results.

REFERENCES

[1] C. Canuto, A. Tabacco, and K. Urban, *The wavelet element method part* I: *Construction and analysis*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 1–52.

[2] J. M. Carnicer, W. Dahmen, and J. M. Peña, *Local decomposition of refinable spaces and wavelets*, Appl. Comput. Harmon. Anal., 3 (1996), pp. 127–153.

[3] A. Cohen, *Numerical Analysis of Wavelet Methods*, Elsevier, Amsterdam, 2003.

[4] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods for elliptic operator equations—convergence rates*, Math. Comp., 70 (2001), pp. 27–75.

[5] A. Cohen, W. Dahmen, and R. DeVore, *Adaptive wavelet methods* II: *Beyond the elliptic case*, Found. Comput. Math., 2 (2002), pp. 203–245.

[6] A. Cohen, I. Daubechies, and J. C. Feauveau, *Biorthogonal bases of compactly supported wavelets*, Comm. Pure Appl. Math., 45 (1992), pp. 485–560.

[7] A. Cohen and R. Masson, *Wavelet adaptive method for second order elliptic problems: Boundary conditions and domain decomposition*, Numer. Math., 86 (2000), pp. 193–238.

[8] W. Dahmen, *Stability of multiscale transformations*, J. Fourier Anal. Appl., 4 (1996), pp. 341–362.

[9] W. Dahmen, *Wavelet and multiscale methods for operator equations*, Acta Numer., 6 (1997), pp. 55–228.

[10] W. Dahmen, H. Harbrecht, and R. Schneider, *Compression techniques for boundary integral equations—asymptotically optimal complexity estimates*, SIAM J. Numer. Anal., 43 (2006), pp. 2251–2271.

[11] W. Dahmen, A. Kunoth, and K. Urban, *Biorthogonal spline-wavelets on the interval—stability and moment conditions*, Appl. Comput. Harmon. Anal., 6 (1999), pp. 132–196.

[12] W. Dahmen and R. Schneider, *Composite wavelet bases for operator equations*, Math. Comp., 68 (1999), pp. 1533–1567.

[13] W. Dahmen and R. Schneider, *Wavelets on manifolds* I: *Construction and domain decomposition*, SIAM J. Math. Anal., 31 (1999), pp. 184–230.

[14] W. Dahmen and R. Stevenson, *Element-by-element construction of wavelets satisfying stability and moment conditions*, SIAM J. Numer. Anal., 37 (1999), pp. 319–352.

[15] T. Gantumur, H. Harbrecht, and R. Stevenson, *An Optimal Adaptive Wavelet Method without Coarsening of the Iterands*, Technical report 1325, Utrecht University, Utrecht, The Netherlands, 2005. Math. Comp., to appear.

[16] A. Kunoth and J. Sahner, *Wavelets on manifolds: An optimized construction*, Math. Comp., 75 (2006), pp. 1319–1349.

[17] H. Nguyen and R. Stevenson, *Finite element wavelets on manifolds*, IMA J. Numer. Math., 23 (2003), pp. 149–173.
[18] R. Schneider, *Multiskalen- und Wavelet-Matrixkompression: Analysisbasierte Methoden zur Lösung großer vollbesetzter Gleigungssysteme*, Advances in Numerical Mathematics, Teubner, Stuttgart, 1998.
[19] R. Stevenson, *Stable three-point wavelet bases on general meshes*, Numer. Math., 80 (1998), pp. 131–158.
[20] R. Stevenson, *Locally supported, piecewise polynomial biorthogonal wavelets on nonuniform meshes*, Constr. Approx., 19 (2003), pp. 477–508.
[21] R. Stevenson, *On the compressibility of operators in wavelet coordinates*, SIAM J. Math. Anal., 35 (2004), pp. 1110–1132.
[22] R. Stevenson, *Composite Wavelet Bases with Extended Stability and Cancellation Properties—Version Including All Proofs*, Technical report 1345, Utrecht University, Utrecht, The Netherlands, 2006; available online at http://www.math.uu.nl/people/stevenso/publ.htm.
[23] P. S. Vassilevski and J. Wang, *Stabilizing the hierarchical basis by approximate wavelets* II: *Implementation and numerical experiments*, SIAM J. Sci. Comput., 20 (1998), pp. 490–514.
[24] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 163–192

DISCONTINUOUS GALERKIN METHODS FOR QUASI-LINEAR ELLIPTIC PROBLEMS OF NONMONOTONE TYPE*

THIRUPATHI GUDI† AND AMIYA K. PANI†

Abstract. In this paper, both symmetric and nonsymmetric interior penalty discontinuous *hp*-Galerkin methods are applied to a class of quasi-linear elliptic problems which are of nonmonotone type. Using Brouwer's fixed point theorem, it is shown that the discrete problem has a solution, and then, using Lipschitz continuity of the discrete solution map, uniqueness is also proved. A priori error estimates in the broken H^1-norm, which are optimal in h and suboptimal in p, are derived. Moreover, on a regular mesh an *hp*-error estimate for the L^2-norm is also established. Finally, numerical experiments illustrating the theoretical results are provided.

Key words. *hp*-finite elements, discontinuous Galerkin methods, second order quasi-linear elliptic problems, optimal estimates, Brouwer's fixed point theorem

AMS subject classifications. 65N12, 65N15, 65N30

DOI. 10.1137/050643362

1. Introduction.

In recent years, there has been renewed interest in discontinuous Galerkin methods for the numerical solution of a wide range of partial differential equations. This is due to their flexibility in local mesh adaptivity and their flexibility in handling nonuniform degrees of approximation for solutions whose smoothness exhibits variation over the computational domain. Based on Nitsche's symmetric formulation in 1970, these methods were introduced for second order elliptic and parabolic equations by Arnold [3], Douglas and Dupont [13], and Wheeler [24] and hence are presently called symmetric interior penalty discontinuous Galerkin (SIPG) methods. It is observed that SIPG methods are adjoint consistent, but the stabilizing parameters in these methods depend on the bounds of the coefficients of the problem considered and various constants involved in inverse inequalities. Recently, Oden, Babuška, and Baumann [21] proposed another discontinuous Galerkin method, which is based on a nonsymmetric formulation. Rivière, Wheeler, and Girault [23] and Houston, Schwab, and Süli [16] introduced and analyzed the nonsymmetric interior penalty discontinuous Galerkin (NIPG) method, which is a stabilized discontinuous Galerkin method. For a review, see [22], and for variants of discontinuous formulations, see Brezzi et al. [9], Arnold et al. [4], Houston, Robson, and Süli [17], and the references therein. A significant property of an NIPG method is that it is unconditionally stable with respect to the choice of the penalty parameter. Hence, this advantage has stimulated renewed interest in applying these methods to a large class of partial differential equations.

In the literature, optimal a priori error estimates are derived in the broken H^1-norm, and numerical experiments are conducted for SIPG and NIPG methods for linear self-adjoint elliptic problems; see [4], [23]. Except for [17], there are hardly any results on discontinuous Galerkin approximation of nonlinear elliptic problems. In [17], a one-parameter family of discontinuous Galerkin methods is applied to the

*Received by the editors October 24, 2005; accepted for publication (in revised form) July 19, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sinum/45-1/64336.html

†Industrial Mathematics Group, Department of Mathematics, Indian Institute of Technology Bombay, Powai, Mumbai 400076, India (trpathi@math.iitb.ac.in, akp@math.iitb.ac.in). The second author's research was supported by the DST-DAAD (PPP-05) project.

quasi-linear elliptic problems where the differential operator is strongly monotone and Lipschitz continuous. In particular, the authors have considered a class of elliptic problems

$$-\nabla \cdot (\mu(x,\ |\nabla u|)\nabla u) = f(x)$$

subject to mixed Dirichlet–Neumann boundary conditions. Under the structural conditions on $\mu \in C(\bar{\Omega} \times [0,\infty))$,

$$m_\mu(t-s) \le \mu(x,t)t - \mu(x,s)s \le M_\mu(t-s) \quad \text{for } \ t \ge s \ge 0, \tag{1.1}$$

and for some positive constants m_μ and M_μ, it is shown that the discontinuous Galerkin formulation is monotone, and hence a priori error estimates in broken H^1-norm are derived. For nonlinear problems of the type

$$-\nabla \cdot (a(u)\nabla u) = f \quad \text{in } \Omega, \tag{1.2}$$

$$u = g \quad \text{on } \partial\Omega, \tag{1.3}$$

where $0 < \alpha \le a(u) \le M$, for some $\alpha,\ M \in \mathbb{R}^+$, the nonlinearity may not satisfy (1.1), and hence it is difficult to extend the analysis of [17]. Therefore, an attempt has been made in this paper to study discontinuous Galerkin methods for the problem (1.2)–(1.3). The results presented in this paper can be thought of an extension to discontinuous Galerkin methods of the results established for the nonlinear Dirichlet problem (1.2)–(1.3) by using a Galerkin method in [12]. Both SIPG and NIPG methods are discussed for the problem (1.2)–(1.3), and a priori error estimates are derived in the broken H^1-norm which are optimal in h. These results lead to precisely the same h-optimal and p-suboptimal rates of convergence in the broken H^1-norm as in the case of linear elliptic problems, when it is approximated by an NIPG method; see [23, Theorem 3.1].

The organization of this article is as follows. Section 1 is introductory in nature, and section 2 is devoted to notation, definitions, and preliminaries. In section 3, discontinuous Galerkin methods are discussed for linear nonselfadjoint elliptic problems, and a priori error estimates are derived in the broken H^1-norm, which are optimal in h and suboptimal in p. Section 4 is devoted to SIPG and NIPG methods for the quasi-linear elliptic problems (1.2)–(1.3). Using Brouwer's fixed point theorem, existence of a discrete solution is proved. Then a priori error estimates are derived in the broken H^1-norm, which are optimal in h and suboptimal in p. Further, an a priori error estimate in the L^2-norm is established on regular meshes for (1.2)–(1.3) with piecewise polynomial or zero Dirichlet boundary datum. In section 5, we provide some numerical experiments to illustrate the theoretical results obtained in this article. Finally, in section 6, we present a summary and some extensions.

2. Preliminaries. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with boundary $\partial\Omega$, where the boundary $\partial\Omega$ is assumed sufficiently smooth in order that a duality argument can be employed in our subsequent analysis. Let $\mathcal{T}_h = \{K_i : 1 \le i \le N_h\}$ be a shape regular finite element subdivision of Ω, where K_i is either a triangle or a rectangle (possibly curvilinear) defined as follows. Let $\hat{K}$ be a shape regular master triangle or rectangle in $\mathbb{R}^2$, and let $\{F_i\}$ be a family of invertible maps such that F_i maps from $\hat{K}$ onto K_i. For a definition of shape regularity, we refer to [10, p. 124]. Let h_i be the diameter of K_i and $h = \max\{h_i : 1 \le i \le N_h\}$. We denote the set of interior edges of $\mathcal{T}_h$ by $\Gamma_I = \{e_{ij} : e_{ij} = \partial K_i \cap \partial K_j,\ |e_{ij}| > 0\}$ and the set of

boundary edges by $\Gamma_\partial = \{e_{i\partial} : e_{i\partial} = \partial K_i \cap \partial\Omega,\ |e_{i\partial}| > 0\}$, where $|e_k|$ denotes the one-dimensional measure of e_k. Let $\Gamma = \Gamma_I \cup \Gamma_\partial$. Since for each $e_k \in \Gamma_I$ there exist $K_i, K_j \in \mathcal{T}_h$ such that $e_k = \partial K_i \cap \partial K_j$ $(i > j)$, we associate with e_k a unit normal vector ν_k which is directed outward from K_i. For $e_k \in \Gamma_\partial$, let ν_k be the unit outward normal to the boundary $\partial\Omega$. For simplicity, we denote $\nu = \nu_k$. Note that our definition of e_k also admits hanging nodes along each edge of the finite elements. On this subdivision $\mathcal{T}_h$, we define the following broken Sobolev space of composite order $\mathbf{s} = \{s_i \geq 0 : 1 \leq i \leq N_h\}$ and exponent r, with $1 \leq r \leq \infty$:

$$W_r^{\mathbf{s}}(\Omega, \mathcal{T}_h) = \{v \in L^r(\Omega) : v|_{K_i} \in W_r^{s_i}(K_i)\ \forall K_i \in \mathcal{T}_h\},$$

where $W_r^{s_i}(K_i)$ is the standard Sobolev space of order s_i with exponent r for each K_i. For $1 \leq r < \infty$, the associated broken norm and seminorm are defined, respectively, by

$$\|v\|_{W_r^{\mathbf{s}}(\Omega,\mathcal{T}_h)} = \left(\sum_{i=1}^{N_h} \|v\|^r_{W_r^{s_i}(K_i)}\right)^{1/r} \quad \text{and} \quad |v|_{W_r^{\mathbf{s}}(\Omega,\mathcal{T}_h)} = \left(\sum_{i=1}^{N_h} |v|^r_{W_r^{s_i}(K_i)}\right)^{1/r},$$

and for the case $r = \infty$, the associated broken norm and seminorm are defined, respectively, by

$$\|v\|_{W_\infty^{\mathbf{s}}(\Omega,\mathcal{T}_h)} = \max_{1\leq i\leq N_h} \|v\|_{W_\infty^{s_i}(K_i)} \quad \text{and} \quad |v|_{W_\infty^{\mathbf{s}}(\Omega,\mathcal{T}_h)} = \max_{1\leq i\leq N_h} |v|_{W_\infty^{s_i}(K_i)},$$

where $||v||_{W_r^{s_i}(K_i)}$ and $|v|_{W_r^{s_i}(K_i)}$ are the standard Sobolev norm and seminorm on K_i. When $r = 2$, we write $H^{\mathbf{s}}(\Omega, \mathcal{T}_h) = W_2^{\mathbf{s}}(\Omega, \mathcal{T}_h)$ and also write the norm and seminorm as

$$||v||_{\mathbf{s},h} = \|v\|_{W_r^{\mathbf{s}}(\Omega,\mathcal{T}_h)} \quad \text{and} \quad |v|_{\mathbf{s},h} = |v|_{W_r^{\mathbf{s}}(\Omega,\mathcal{T}_h)},$$

and when $s = s_i$ for all $1 \leq i \leq N_h$, we write $H^s(\Omega, \mathcal{T}_h)$,$\|v\|_{s,h}$ and $|v|_{s,h}$, respectively. For $s = 0$, we denote the norm by $\|.\|$ which is the standard L^2-norm. We now define the jump and average of a function $v \in H^1(\Omega, \mathcal{T}_h)$ on an edge $e_k \in \Gamma$ as follows. If $e_k \in \Gamma_I$, that is, $e_k = \partial K_i \cap \partial K_j$ $(i > j)$ for some i and j, then we define the jump and average as

$$[v] = v|_{K_i} - v|_{K_j}, \qquad \{v\} = \frac{v|_{K_i} + v|_{K_j}}{2}.$$

In the case $e_k \in \Gamma_\partial$, there exists K_i such that $e_k = \partial K_i \cap \partial\Omega$, and we then define, for notational convenience, the jump and average on e_k as

$$[v] = v|_{K_i\cap\partial\Omega}, \qquad \{v\} = v|_{K_i\cap\partial\Omega}.$$

Let $\hat{P}_{p_i}(\hat{K})$ be the space of polynomials of total degree less than or equal to p_i on the triangle $\hat{K}$, and let $\hat{Q}_{p_i}(\hat{K})$ be the space of polynomials of degree less than or equal to p_i in each variable which are defined on the rectangle $\hat{K}$. Let $\mathcal{Z}_{p_i}(\hat{K})$ denote $\hat{P}_{p_i}(\hat{K})$ or $\hat{Q}_{p_i}(\hat{K})$ whenever $\hat{K}$ is a master triangle or a rectangle, respectively. Now define (see [21], [14])

$$\mathcal{Z}_{p_i}(K_i) = \{v : v = \hat{v} \circ F_i^{-1}, \hat{v} \in \hat{\mathcal{Z}}_{p_i}(\hat{K})\}.$$

The discontinuous finite element space is defined as

$$\mathcal{D}_p(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|K_i \in \mathcal{Z}_{p_i}(K_i)\},$$

where $p = \min\{p_i \geq 1 : 1 \leq i \leq N_h\}$.

We also need the following discontinuous finite element space of piecewise polynomials with uniform degree p:

$$\mathcal{D}_p^*(\mathcal{T}_h) = \{v \in L^2(\Omega) : v|K_i \in \mathcal{Z}_p(K_i) \quad \forall i\}.$$

We also define the following Sobolev space with piecewise polynomial traces:

$$H_p^s(\Omega) = \{v \in H^s(\Omega) : v|_{\partial\Omega} = w|_{\partial\Omega} \quad \text{for some } w \in \mathcal{D}_p^*(\mathcal{T}_h)\}.$$

Assumption P.

1. *The finite element subdivision $\mathcal{T}_h$ satisfies the* bounded local variation *condition in the sense that if $|\partial K_i \cap \partial K_j| > 0$ for any K_i and $K_j \in \mathcal{T}_h$, then there exists a constant κ independent of h_i, h_j such that*

$$\frac{h_i}{h_j} \leq \kappa.$$

In particular, this implies that for any element K_i the number of neighboring elements $K_j \in \mathcal{T}_h$ with $|\partial K_i \cap \partial K_j| > 0$ is bounded by N_κ uniformly for some positive integer N_κ.

2. *The discontinuous finite element space $\mathcal{D}_p(\mathcal{T}_h)$ satisfies the following* bounded local variation: *If $|\partial K_i \cap \partial K_j| > 0$ for any K_i and $K_j \in \mathcal{T}_h$, then there exists a constant ϱ independent of p_i and p_j such that*

$$\frac{p_i}{p_j} \leq \varrho.$$

Here, $|.|$ denotes the one-dimensional Euclidean measure.

We now present some examples which satisfy Assumption P(1).

(i) *Regular subdivision.* A subdivision of Ω into shape regular elements K_i, $1 \leq i \leq N_h$, is such that for any two elements K_i and K_j, the common portion $\partial K_i \cap \partial K_j$ is either empty or a vertex of K_i or an entire edge e of K_i, that is, $e = \partial K_i \cap \partial K_j$ and there is no other element $K_l \in \mathcal{T}_h$ $(l \neq j,\ i)$ such that $|e \cap \partial K_l| > 0$ [10, p. 132].

(ii) 1-*irregular subdivision.* A shape regular subdivision $\{K_i\}_{i=1}^{N_h}$ of Ω is such that for any side of an element K_i, there can be at most one hanging node (cf. Figure 1); see [16] and [17, p. 5].

For $e_k \in \Gamma_I$, there are two elements K_i and K_j such that $e_k = \partial K_i \cap \partial K_j$. Hence, we define the "degree" of polynomial in K_i and K_j restricted to e_k by p_k, by $p_k = (p_i + p_j)/2$. For $e_k \in \Gamma_\partial$, we note that there is one element K_i with $e_k = \partial K_i \cap \partial\Omega$, and hence we denote the degree of polynomial restricted to e_k by $p_k = p_i$.

From Assumption P, it is easy to see that if $e_k \subset \partial K_i$, then there exist constants $c_1(\kappa)$, $c_2(\kappa)$, $c_3(\varrho)$, and $c_4(\varrho)$ which are independent of h and p such that

$$c_1(\kappa)h_i \leq |e_k| \leq c_2(\kappa)h_i, \qquad c_3(\varrho)p_i \leq p_k \leq c_4(\varrho)p_i. \tag{2.1}$$

Let $v \in H^2(\Omega, \mathcal{T}_h)$. We define the following mesh-dependent norms which appear naturally in the analysis of interior penalty discontinuous Galerkin methods:

$$|||v|||^2 = \left(\sum_{i=1}^{N_h} \int_{K_i} |\nabla v|^2 \, \mathrm{dx} \ + \ \mathcal{J}^{1,\beta}(v,\ v) \right) \tag{2.2}$$

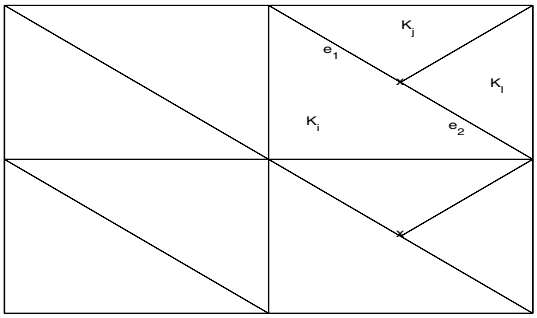


FIG. 1. 1-*irregular mesh.*

and

$$|||v|||_+^2 = \left(\sum_{i=1}^{N_h} \int_{K_i} |\nabla v|^2 \, \mathrm{dx} \; + \; \sum_{e_k \in \Gamma} \frac{|e_k|^\beta}{p_k^2} \int_{e_k} \left\{ \frac{\partial v}{\partial \nu} \right\}^2 \, \mathrm{ds} \; + \; \mathcal{J}^{1,\beta}(v,\; v) \right), \tag{2.3}$$

where

$$\mathcal{J}^{\sigma,\beta}(v,\; w) = \left(\sum_{e_k \in \Gamma} \sigma_k \frac{p_k^2}{|e_k|^\beta} \int_{e_k} [v][w] \, \mathrm{ds} \right),$$

$\sigma|_{e_k} = \sigma_k$, and σ_k, β are positive real numbers.

Approximation properties of finite element spaces. Below, we state without proof a lemma on some *hp*-approximation properties.

LEMMA 2.1. *For $\phi \in H^s(K_i)$, there exists a positive constant C_A (depending on s but independent of ϕ, p_i, and h_i) and a sequence $\phi_{p_i}^{h_i} \in \mathcal{Z}_{p_i}(K_i)$, $p_i = 1, 2, \ldots$, such that*

(i) *for any $0 \le l \le s_i$,*

$$\|\phi - \phi_{p_i}^{h_i}\|_{H^l(K_i)} \le C_A \frac{h_i^{\mu_i - l}}{{p_i}^{s_i - l}} \|\phi\|_{H^{s_i}(K_i)};$$

(ii) *for $s_i > l + \frac{1}{2}$,*

$$\|\phi - \phi_{p_i}^{h_i}\|_{H^l(e_k)} \le C_A \frac{h_i^{\mu_i - l - 1/2}}{p_i^{s_i - l - 1/2}} \|\phi\|_{H^{s_i}(K_i)};$$

(iii) *for $0 \le l \le s_i - 1 + 2/r$,*

$$\|\phi - \phi_{p_i}^{h_i}\|_{W_r^l(K_i)} \le C_A \frac{h_i^{\mu_i - l - 1 + 2/r}}{{p_i}^{s_i - l - 1 + 2/r}} \|\phi\|_{H^{s_i}(K_i)},$$

where $\mu_i = \min(s_i,\; p_i + 1)$.

The proof of properties (i) and (ii) can be found in [6, Lemma 4.5]. Then using properties (1) and (3) in Lemma 1 of [1] and rescaling (see [2, Lemma 2]), it is easy to derive property (iii).

For given $\phi \in H^{\mathbf{s}}(\Omega, \mathcal{T}_h)$, we define $I_h\phi \in \mathcal{D}_p(\mathcal{T}_h)$ by $(I_h\phi)|_{K_i} = \phi_{p_i}^{h_i}(\phi|_{K_i})$ for all $1 \le i \le N_h$. By virtue of Lemma 2.1, $I_h\phi$ satisfies the local approximation properties derived in Lemma 2.1; see [17, p. 737].

Trace inequalities. We state without proof the following trace inequality. For $r = 2$ it is proved in [22, Appendix A.2], and using similar arguments, we can easily obtain the inequality for $r = 4$.

LEMMA 2.2. *Let $\phi \in H^{j+1}(K_i)$, $K_i \in \mathcal{T}_h$. Then there exists a constant $C_{T_1} > 0$ such that*

$$\|\phi\|^r_{W^j_r(e_k)} \le C_{T_1}\left(\frac{1}{h_i}\|\phi\|^r_{W^j_r(K_i)} + \|\phi\|^{r-1}_{W^j_{2r-2}(K_i)}\|\nabla^{(j+1)}\phi\|_{L^2(K_i)}\right), \tag{2.4}$$

where $j = 0, 1$ and $r = 2, 4$.

We recall the following trace inequality on finite element spaces for our future use. For a proof, we refer to [23, Lemma 2.1].

LEMMA 2.3. *Let $v_h \in \mathcal{Z}_{p_i}(K_i)$. Then there exists a constant $C_{T_2} > 0$ such that*

$$\|\nabla^l v_h\|_{L^2(e_k)} \le C_{T_2} p_i h_i^{-1/2}\|\nabla^l v_h\|_{L^2(K_i)}, \quad l = 0, 1. \tag{2.5}$$

Below, we state without proof a lemma on inverse inequalities. For a proof, we refer to [18, p. 6], [7, Theorem 6.1].

LEMMA 2.4 (inverse inequalities). *Let $v_h \in \mathcal{Z}_{p_i}(K_i)$. Then, for $r \ge 2$, there exists a constant $C_I > 0$ such that*

$$\|v_h\|_{L^r(K_i)} \le C_I p_i^{1-2/r} h_i^{(2/r-1)}\|v_h\|_{L^2(K_i)}, \tag{2.6}$$

$$|v_h|_{H^l(K_i)} \le C_I p_i^2 h_i^{-1}|v_h|_{H^{l-1}(K_i)}, \quad l \ge 1, \tag{2.7}$$

and

$$\|v_h\|_{L^r(e_k)} \le C_I p_i^{1-2/r}|e_k|^{(1/r-1/2)}\|v_h\|_{L^2(e_k)}, \tag{2.8}$$

where $e_k \subset \partial K_i$ is an edge.

For our future use, we state the following Poincaré-type inequalities on $H^1(\Omega, \mathcal{T}_h)$. For a proof, we refer to [18, Theorem 3.7]; see also [8] for the case of $r = 2$.

LEMMA 2.5 (Poincaré-type inequalities). *Let $v \in H^1(\Omega, \mathcal{T}_h)$. Then there exists a constant $C_P > 0$ independent of h and v such that, for $1 \le r < \infty$,*

$$\|v\|_{L^r(\Omega)} \le C_P|||v|||.$$

3. Nonselfadjoint linear elliptic problems. For our error analysis of discontinuous Galerkin methods applied to the nonlinear elliptic problem (1.2)–(1.3), we need some results on the corresponding linearized problems. Since the linearized problem is a nonselfadjoint elliptic problem, in this section, we consider the following second order linear nonselfadjoint elliptic partial differential equation:

$$\begin{aligned} -\nabla\cdot(a(x)\nabla u) + \vec{b}(x)\cdot\nabla u + a_0(x)u &= f(x) \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega. \end{aligned} \tag{3.1}$$

We adopt the following assumptions on the problem (3.1).

ASSUMPTION R.

1. *There exists $\alpha > 0$ such that $0 < \alpha \le a(x)$ and $a_0(x) \ge 0$ for all $x \in \bar{\Omega}$.*
2. *$a \in W^1_\infty(\Omega)$ and b, $a_0 \in L^\infty(\Omega)$ with $M = \max\{\|a\|_{L^\infty(\Omega)}, \|b\|_{L^\infty(\Omega)}, \|a_0\|_{L^\infty(\Omega)}\}$.*
3. *$f \in L^2(\Omega)$ and $g \in H^{3/2}(\partial\Omega)$.*

Then, from [15, Lemma 9.17] it is well known that there exists a unique solution $u \in H^2(\Omega)$ to the problem (3.1) satisfying

$$\|u\|_{H^2(\Omega)} \le C\left(\|f\|_{L^2(\Omega)} + \|g\|_{H^{3/2}(\partial\Omega)}\right). \tag{3.2}$$

3.1. Weak formulation. For $w,\ v \in H^2(\Omega, \mathcal{T}_h)$, we consider the bilinear form

$$B(w,\ v) = \sum_{i=1}^{N_h} \int_{K_i} (a\nabla w \cdot \nabla v + a_0 wv + (\vec{b} \cdot \nabla w)v) - \sum_{e_k \in \Gamma} \int_{e_k} \left\{ a \frac{\partial w}{\partial \nu} \right\} [v]$$
$$- \theta \sum_{e_k \in \Gamma} \int_{e_k} \left\{ a \frac{\partial v}{\partial \nu} \right\} [w] + \mathcal{J}^{\sigma,\beta}(w,\ v) + \sum_{e_k \in \Gamma} \int_{e_k} \left\{ \vec{b} \cdot \nu v \right\} [w]$$

and the linear form

$$L(v) = \int_\Omega fv - \theta \sum_{e_k \in \Gamma_\partial} \int_{e_k} \left(a \frac{\partial v}{\partial \nu} \right) g + \sum_{e_k \in \Gamma_\partial} \int_{e_k} \sigma_k \frac{p_k^2}{|e_k|^\beta} vg + \sum_{e_k \in \Gamma_\partial} \int_{e_k} \vec{b} \cdot \nu v g,$$

where $\theta = \pm 1$. When $\vec{b} = 0$, we note that $\theta = +1$ corresponds to a symmetric and $\theta = -1$ to a nonsymmetric interior penalty method.

We define a weak formulation which is suitable for the discontinuous Galerkin methods as follows: Find $u \in H^2(\Omega, \mathcal{T}_h)$ such that

$$B(u,\ v) = L(v) \quad \forall v \in H^2(\Omega, \mathcal{T}_h). \tag{3.3}$$

Now the discontinuous Galerkin approximation of u is to seek $u_h \in \mathcal{D}_p(\mathcal{T}_h)$ such that

$$B(u_h,\ v_h) = L(v_h) \quad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h). \tag{3.4}$$

Below, we examine the consistency of the above scheme (3.4).

THEOREM 3.1. *If the solution u of problem* (3.1) *is in $H^2(\Omega)$, then u satisfies problem* (3.3). *Conversely, if the solution u of problem* (3.3) *is in $H^1(\Omega) \cap H^2(\Omega, \mathcal{T}_h)$, then u satisfies problem* (3.1) *weakly.*

The proof techniques of Rivière, Wheeler, and Girault [23, Lemma 2.2] or [22, Theorem 3.1] can be easily modified to prove Theorem 3.1, and so the proof is omitted. The solvability of (3.4) will be discussed at the end of section 3. From (3.3)–(3.4) and Theorem 3.1, it is easy to check that

$$B(u - u_h,\ v_h) = 0 \quad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h). \tag{3.5}$$

Following [22], [24], we state the following Gårding-type inequality.

LEMMA 3.2. *Let $\beta \geq 1$ and $0 < \sigma_0 \leq \sigma_k \leq \sigma_m$. Further, assume that $\sigma_0 \geq C(\alpha,\ M,\ C_{T_2},\ N_\kappa)$ when $\theta = 1$, and $\sigma_0 > 0$ when $\theta = -1$. Then there exist two constants $C_1 = C(\alpha,\ \sigma_0) > 0$ and $C_2 = C(\alpha,\ \sigma_0,\ M,\ C_{T_2},\ N_\kappa) > 0$ which are independent of h and p such that*

$$B(v_h,\ v_h) \geq C_1 |||v_h|||^2 - C_2 \|v_h\|^2 \quad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h).$$

A straightforward modification of the analysis of Prudhomme, Pascal, and Oden [22, Theorems 3.4 and 3.5] and of Wheeler [24, Lemma 3] yields the proof of Lemma 3.2, and so we omit the proof. Throughout this article, C denotes a generic constant which is independent of h, p, and u_h but may depend on κ, ϱ, σ_0, σ_m, α, M, C_A, C_{T_1}, C_{T_2}, C_I, C_P, C_1, and C_2.

Using the trace inequality (2.5) and (2.1), it is an easy exercise to prove the following Lemma 3.3. For details, see [22, Theorem 3.3].

LEMMA 3.3. *Let $\beta \geq 1$ and $\phi \in H^2(\Omega, \mathcal{T}_h)$. If σ_k is bounded above by a positive number σ_m, then there exists a positive constant C, independent of h and p, such that*

$$|B(\phi,\ v_h)| \leq C|||\phi|||_+\ |||v_h||| \qquad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h).$$

LEMMA 3.4. *Let $\beta = 1$. Then there exists a positive constant C which depends on C_A, but is independent of h and p, such that*

$$|||\phi - \mathcal{I}_h\phi|||_+ \leq C\left(\sum_{i=1}^{N_h} \frac{h_i^{2\mu_i-2}}{p_i^{2s_i-3}}\|\phi\|^2_{H^{s_i}(K_i)}\right)^{1/2},$$

where $\mu_i = \min\{p_i + 1, s_i\}$.

Proof. Let $\eta^* = \phi - \mathcal{I}_h\phi$. Then, using (2.3), Lemma 2.1, and (2.1), we obtain

$$\begin{aligned}|||\eta^*|||^2_+ &= \sum_{i=1}^{N_h}\int_{K_i}|\nabla\eta^*|^2 + \sum_{e_k\in\Gamma}\int_{e_k}\frac{|e_k|^\beta}{p_i^2}\left\{\frac{\partial\eta^*}{\partial\nu}\right\}^2 + \sum_{e_k\in\Gamma}\int_{e_k}\frac{p_i^2}{|e_k|^\beta}[\eta^*]^2\\ &\leq C\sum_{i=1}^{N_h}\left(\frac{h_i^{2\mu_i-2}}{p_i^{2s_i-2}}\|\phi\|^2_{H^{s_i}(K_i)} + \frac{h_i^{2\mu_i-3+\beta}}{p_i^{2s_i-1}}\|\phi\|^2_{H^{s_i}(K_i)} + \frac{h_i^{2\mu_i-1-\beta}}{p_i^{2s_i-3}}\|\phi\|^2_{H^{s_i}(K_i)}\right). \qquad (3.6)\end{aligned}$$

Since $\beta = 1$, the lemma is proved by taking a square root on both sides of (3.6). ∎

We prove the following lemma, which will be used in the proof of a priori error estimates.

LEMMA 3.5. *Let $\beta = 1$ and $q \in L^2(\Omega)$. Then, for sufficiently small h, there exists a unique $\phi_h \in \mathcal{D}_p(\mathcal{T}_h)$ satisfying*

$$B(v_h,\ \phi_h) = \int_\Omega q v_h \qquad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h). \qquad (3.7)$$

Moreover, ϕ_h satisfies

$$|||\phi_h||| \leq C\|q\|. \qquad (3.8)$$

Proof. Note that (3.7) leads to a system of linear algebraic equations. So it is enough to prove uniqueness. Set $v_h = \phi_h$ in (3.7) and use Lemma 3.2 to obtain

$$\begin{aligned}C_1|||\phi_h|||^2 - C_2\|\phi_h\|^2 \leq B(\phi_h,\ \phi_h) &= \int_\Omega q\phi_h\\ &\leq \|q\|\ \|\phi_h\|.\end{aligned}$$

Therefore, we arrive at

$$|||\phi_h||| \leq C_1\|q\| + C_2\|\phi_h\|. \qquad (3.9)$$

To estimate $\|\phi_h\|$ in terms of $|||\phi_h|||$, we apply the standard Aubin–Nitsche duality argument. For $\phi_h \in \mathcal{D}_p(\mathcal{T}_h)$, we consider the following auxiliary problem:

$$\begin{aligned}-\nabla\cdot(a(x)\nabla\psi) + \vec{b}(x)\cdot\nabla\psi + a_0(x)\psi &= \phi_h \quad \text{in } \Omega,\\ \psi &= 0 \quad \text{on } \partial\Omega.\end{aligned} \qquad (3.10)$$

Then Assumption R implies that ψ satisfies the following elliptic regularity:

$$\|\psi\|_{H^2(\Omega)} \leq C\|\phi_h\|_{L^2(\Omega)}. \qquad (3.11)$$

We multiply (3.10) by ϕ_h and integrate over Ω, apply integration by parts, and then use Lemmas 3.3 and 3.4 to obtain

$$\begin{aligned}\|\phi_h\|^2 = B(\psi,\ \phi_h) &= B(\psi - \mathcal{I}_h\psi,\ \phi_h) + B(\mathcal{I}_h\psi,\ \phi_h) = B(\psi - \mathcal{I}_h\psi,\ \phi_h) + \int_\Omega q\mathcal{I}_h\psi \\ &\le C\left(h\ |||\phi_h||| + \|q\|\right)\|\psi\|_{H^2(\Omega)}.\end{aligned}$$

From the elliptic regularity (3.11), we now arrive at

$$\|\phi_h\| \le Ch|||\phi_h||| + \|q\|. \tag{3.12}$$

Substituting (3.12) into (3.9), we obtain the estimate (3.8) for sufficiently small h. Hence, (3.7) has a unique solution and this completes the rest of the proof. ∎

3.2. A priori error estimates. Let $\beta = 1$. Since Lemma 3.2 holds for elements in $\mathcal{D}_p(\mathcal{T}_h)$, we split $e = u - u_h$ into $e = \eta + \chi$, where $\eta = u - \mathcal{I}_h u$ and $\chi = \mathcal{I}_h u - u_h$. Then using Lemmas 3.2 and 3.3 and (3.5), we obtain

$$\begin{aligned}C_1|||\chi|||^2 - C_2\|\chi\|^2 \le B(\chi,\ \chi) &= B((\mathcal{I}_h u - u) + (u - u_h),\ \chi) \\ &= B(\mathcal{I}_h u - u,\ \chi) = B(\eta,\ \chi) \\ &\le C|||\eta|||_+\ |||\chi|||.\end{aligned}$$

Therefore,

$$|||\chi||| \le C|||\eta|||_+ + C_2\|\chi\|. \tag{3.13}$$

In order to estimate $\|\chi\|$, we set $q = \chi$ and $v_h = \chi$ in Lemma 3.5. Using (3.5) and Lemma 3.3, we now obtain

$$\begin{aligned}\|\chi\|^2 = B(\chi,\ \phi_h) &= B(\mathcal{I}_h u - u_h,\ \phi_h) = B(\mathcal{I}_h u - u,\ \phi_h) \\ &\le C|||\eta|||_+\ |||\phi_h|||.\end{aligned}$$

Using (3.8), we arrive at

$$\|\chi\| \le C|||\eta|||_+. \tag{3.14}$$

From the estimates (3.13) and (3.14), we obtain

$$|||\chi||| \le C|||\eta|||_+. \tag{3.15}$$

Now using Lemma 3.4, inequality (3.15), and triangle inequality, we deduce the following theorem.

THEOREM 3.6. *Let $\beta = 1$; then for sufficiently small h, there exists a positive constant C which is independent of h and p such that*

$$|||u - u_h||| \le C\left(\sum_{i=1}^{N_h} \frac{h_i^{2\mu_i - 2}}{p^{2s_i - 3}}\|u\|^2_{H^{s_i}(K_i)}\right)^{1/2},$$

where $\mu_i = \min\{s_i,\ p_i + 1\}$.

Existence and uniqueness. We now prove the existence of a unique solution to problem (3.4) using the discrete dual problem (3.7) stated in Lemma 3.5. Assume that there exist two distinct solutions u_h^1 and u_h^2 for the problem (3.4). Let $\xi = u_h^1 - u_h^2$ and set $q = \xi$, $v_h = \xi$ in (3.7). Since $B(u_h^1 - u_h^2, v_h) = 0$ for all $v_h \in \mathcal{D}_p(\mathcal{T}_h)$, we obtain

$$\|\xi\|^2 = B(\xi,\ \phi_h) = B(u_h^1 - u_h^2,\ \phi_h) = 0.$$

Therefore, $u_h^1 = u_h^2$, which leads to a contradiction. Hence, we conclude that there exists a unique solution u_h for problem (3.4). Now uniqueness implies the existence of a discrete solution u_h to problem (3.4).

4. Quasi-linear elliptic problems. In this section, we consider the following nonlinear elliptic boundary value problem:

$$-\nabla \cdot (a(x,u)\nabla u) = f(x) \quad \text{in } \Omega, \tag{4.1}$$

$$u(x) = g(x) \quad \text{on } \partial\Omega, \tag{4.2}$$

where Ω is a bounded domain in $\mathbb{R}^2$ with smooth boundary $\partial\Omega$. As in [12], we make the following assumptions for problem (4.1)–(4.2). There exist positive constants $\alpha,\ M$ such that $0 < \alpha \le a(x,v) \le M$, $x \in \bar{\Omega},\ v \in \mathbb{R}$, $a(x,v) \in C_b^2(\bar{\Omega} \times \mathbb{R})$, where $C_b^2(\bar{\Omega} \times \mathbb{R})$ is the class of twice continuously differentiable functions on $\bar{\Omega} \times \mathbb{R}$ such that all derivatives of $a(x,v)$ up to and including second order are bounded in $\bar{\Omega} \times \mathbb{R}$. Further, for some $\delta \in (0,1)$, $f \in C^{\delta}(\Omega)$ and g can be extended to Ω to be in $C^{2+\delta}(\Omega)$; then it follows from [11] that there exists a unique weak solution u to (4.1)–(4.2) and $u \in C^{2+\delta}(\bar{\Omega})$, where $C^{m+\delta}(\bar{\Omega})$ consists of all functions whose mth order derivatives are Hölder continuous of order δ on $\bar{\Omega}$.

4.1. Weak formulation. For $\psi,\ w,$ and $v \in H^2(\Omega, \mathcal{T}_h)$, we define the form $B(\psi; w,\ v)$, which is linear in $w,\ v$ for fixed ψ, by

$$\begin{aligned} B(\psi; w,\ v) = &\sum_{i=1}^{N_h} \int_{K_i} a(\psi)\nabla w \cdot \nabla v - \sum_{e_k \in \Gamma_I} \int_{e_k} \left(\left\{ a(\psi)\frac{\partial w}{\partial \nu} \right\} [v] + \theta \left\{ a(\psi)\frac{\partial v}{\partial \nu} \right\} [w] \right) \\ &- \sum_{e_k \in \Gamma_\partial} \int_{e_k} \left(a(g)\frac{\partial w}{\partial \nu} v + \theta a(g)\frac{\partial v}{\partial \nu} w \right) + \mathcal{J}^{\sigma,\beta}(w,\ v), \end{aligned}$$

and the linear functional L on $H^2(\Omega,\ \mathcal{T}_h)$ by

$$L(v) = \int_\Omega fv + \theta \sum_{e_k \in \Gamma_\partial} \int_{e_k} a(g)\frac{\partial v}{\partial \nu} g + \sum_{e_k \in \Gamma_\partial} \int_{e_k} \sigma_k \frac{p_i^2}{|e_k|^\beta} vg,$$

where $\theta = \pm 1$. Since for each fixed ψ, $B(\psi; \cdot,\ \cdot)$ is a bilinear form, we note that $\theta = +1$ corresponds to a symmetric and $\theta = -1$ to a nonsymmetric method. We define the weak formulation of (4.1)–(4.2) which is suitable for applying a discontinuous Galerkin method as follows: Find $u \in H^2(\Omega, \mathcal{T}_h)$ such that

$$B(u; u,\ v) = L(v) \quad \forall v \in H^2(\Omega, \mathcal{T}_h). \tag{4.3}$$

Now the discontinuous Galerkin (SIPG and NIPG) approximation of u is to seek $u_h \in \mathcal{D}_p(\mathcal{T}_h)$ such that

$$B(u_h; u_h,\ v_h) = L(v_h) \quad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h). \tag{4.4}$$

Below, we state without proof the consistency of the above scheme (4.4).

THEOREM 4.1 (equivalence of (4.1)–(4.2) and (4.3)). *If the solution u of* (4.1)–(4.2) *is in $H^2(\Omega)$, then u satisfies* (4.3). *Conversely, if $u \in H^1(\Omega) \cap H^2(\Omega, \mathcal{T}_h)$ is a solution of* (4.3), *then u satisfies* (4.1)–(4.2) *weakly.*

The proof follows along the lines of the proof given in [23, Lemma 2.2] or [22, Theorem 3.1], so we omit it. With $v = v_h \in \mathcal{D}_p(\mathcal{T}_h) \subset H^2(\Omega, \mathcal{T}_h)$ in (4.3), and using (4.4), we obtain

$$B(u; u,\ v_h) = B(u_h; u_h,\ v_h) \quad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h). \tag{4.5}$$

Following Taylor series expansion, we write

$$a(w) = a(u) + \tilde{a}_u(w)(w-u), \tag{4.6}$$

where $\tilde{a}_u(w) = \int_0^1 a_u(w + t[u-w])dt$, and

$$a(w) = a(u) + a_u(u)(w-u) + \tilde{a}_{uu}(w)(w-u)^2, \tag{4.7}$$

where $\tilde{a}_{uu}(w) = \int_0^1 (1-t)a_{uu}(w + t[w-u])dt$.

Note that since $a_u \in C_b^1(\bar{\Omega} \times \mathbb{R})$ and $a_{uu} \in C_b^0(\bar{\Omega} \times \mathbb{R})$, it is easy to see that $\tilde{a}_u \in L^\infty(\Omega \times \mathbb{R})$ and $\tilde{a}_{uu} \in L^\infty(\Omega \times \mathbb{R})$. We use the following notation throughout this section:

$$C_a = \max\left[\ \|\tilde{a}_u\|_{L^\infty(\Omega\times\mathbb{R})},\ \|\tilde{a}_{uu}\|_{L^\infty(\Omega\times\mathbb{R})}\right]. \tag{4.8}$$

For simplicity, we consider the following form $\tilde{B}(\cdot\ ;\ \cdot\ ,\ \cdot)$:

$$\tilde{B}(\psi; w,\ v) = B(\psi; w,\ v) + \sum_{i=1}^{N_h} \int_{K_i} (a_u(\psi)\nabla\psi)w \cdot \nabla v - \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ a_u(\psi)\frac{\partial \psi}{\partial \nu} w \right\} [v].$$

Note that $\tilde{B}$ is linear in w and $v \in H^2(\Omega, \mathcal{T}_h)$ for a fixed ψ. It is clear from the assumptions on $a(u)$ that Lemmas 3.2 and 3.3 hold for $\tilde{B}$. Since $a \in C_b^2(\bar{\Omega} \times \mathbb{R})$ and $u \in C^2(\bar{\Omega})$, there is a unique solution $\psi \in H^2(\Omega)$ to the following elliptic problem:

$$\begin{aligned} -\nabla \cdot (a(u)\nabla\psi + a_u(u)\nabla u\psi) &= \phi_h \quad \text{in } \Omega, \\ \psi &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{4.9}$$

and ψ satisfies the elliptic regularity $\|\psi\|_{H^2(\Omega)} \le C\|\phi_h\|$; see [11, Theorem 2], [12, p. 692]. Hence, Lemma 3.5 holds as well for $\tilde{B}$. Now we linearize problem (4.4) around $\mathcal{I}_h u$ for our subsequent analysis. Set $e = u - u_h$. Subtracting $B(u; u_h,\ v_h)$ from both sides of (4.5), we obtain

$$\begin{aligned} B(u; e,\ v_h) = &\sum_{i=1}^{N_h} \int_{K_i} (a(u_h) - a(u))\nabla u_h \cdot \nabla v_h - \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ (a(u_h) - a(u))\frac{\partial u_h}{\partial \nu} \right\} [v_h] \\ &- \theta \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ (a(u_h) - a(u))\frac{\partial v_h}{\partial \nu} \right\} [u_h]. \end{aligned} \tag{4.10}$$

Since $[u] = 0$ on each $e_k \in \Gamma_I$, we rewrite (4.10) as

$$\begin{aligned} B(u; e,\ v_h) &= \sum_{i=1}^{N_h} \int_{K_i} (a(u_h) - a(u)) \nabla(u_h - u) \cdot \nabla v_h - \int_{\Gamma_I} \left\{ (a(u_h) - a(u)) \frac{\partial u}{\partial \nu} \right\} [v_h] \\ &\quad - \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ (a(u_h) - a(u)) \frac{\partial (u_h - u)}{\partial \nu} \right\} [v_h] + \sum_{i=1}^{N_h} \int_{K_i} (a(u_h) - a(u)) \nabla u \cdot \nabla v_h \\ &\quad - \theta \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ (a(u_h) - a(u)) \frac{\partial v_h}{\partial \nu} \right\} [u_h - u]. \end{aligned} \tag{4.11}$$

Finally, we add the following terms to both sides of (4.11):

$$- \sum_{i=1}^{N_h} \int_{K_i} a_u(u)(u_h - u) \nabla u \cdot \nabla v_h + \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ a_u(u)(u_h - u) \frac{\partial u}{\partial \nu} \right\} [v_h].$$

We split $e = u - u_h = u - \mathcal{I}_h u + \mathcal{I}_h u - u_h$. Now using the Taylor formulae (4.6)–(4.7), equation (4.11) takes the form

$$\tilde{B}(u; \mathcal{I}_h u - u_h,\ v_h) = \tilde{B}(u; \mathcal{I}_h u - u,\ v_h) + \mathcal{F}(u_h; u_h - u,\ v_h), \tag{4.12}$$

where

$$\begin{aligned} \mathcal{F}(u_h; -e,\ v_h) &= \sum_{i=1}^{N_h} \int_{K_i} \tilde{a}_u(u_h) e \nabla e \cdot \nabla v_h - \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ \tilde{a}_u(u_h) e \frac{\partial e}{\partial \nu} \right\} [v_h] \\ &\quad - \theta \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ \tilde{a}_u(u_h) e \frac{\partial v_h}{\partial \nu} \right\} [e] + \sum_{i=1}^{N_h} \int_{K_i} \tilde{a}_{uu}(u_h) e^2 \nabla u \cdot \nabla v_h \\ &\quad - \sum_{e_k \in \Gamma_I} \int_{e_k} \left\{ \tilde{a}_{uu}(u_h) e^2 \frac{\partial u}{\partial \nu} \right\} [v_h]. \end{aligned} \tag{4.13}$$

Note that (4.4) is equivalent to (4.12).

Assumption Q (*hp*-quasi-uniformity [19]). *Along with Assumption* P*, we also assume that the subdivision $\mathcal{T}_h$ and discontinuous space $\mathcal{D}_p(\mathcal{T}_h)$ satisfy the following hp-quasi-uniformity condition:*

$$\left(\max_{1 \le i \le N_h} \frac{h_i}{p_i} \right) \le C_Q \left(\min_{1 \le i \le N_h} \frac{h_i}{p_i} \right), \tag{4.14}$$

where C_Q is a positive constant which is independent of h and p.

Observe that under assumption (4.14), the following holds:

$$\left(\max_{1 \le i \le N_h} \frac{p_i}{h_i} \right) \left(\max_{1 \le i \le N_h} \frac{h_i}{p_i} \right) = \left(\min_{1 \le i \le N_h} \frac{h_i}{p_i} \right)^{-1} \left(\max_{1 \le i \le N_h} \frac{h_i}{p_i} \right) \le C_Q. \tag{4.15}$$

4.2. Existence and uniqueness. For a given $z \in \mathcal{D}_p(\mathcal{T}_h)$, let $S_h : \mathcal{D}_p(\mathcal{T}_h) \to \mathcal{D}_p(\mathcal{T}_h)$ be a map $y = S_h z \in \mathcal{D}_p(\mathcal{T}_h)$ satisfying

$$\tilde{B}(u; \mathcal{I}_h u - y,\ v_h) = \tilde{B}(u; \mathcal{I}_h u - u,\ v_h) + \mathcal{F}(z; z - u,\ v_h) \quad \forall v_h \in \mathcal{D}_p(\mathcal{T}_h). \tag{4.16}$$

For a given z, problem (4.16) leads to a system of linear algebraic equations. So using Lemmas 3.2 and 3.5, it is easy to show that the map S_h is well defined. Now consider the ball

$$\mathcal{O}_\delta(\mathcal{I}_h u) = \{z \in \mathcal{D}_p(\mathcal{T}_h) : \ |||\mathcal{I}_h u - z||| \le \delta\}$$

of radius δ, where δ will be chosen later. We first show that for some $\delta > 0$, S_h maps $\mathcal{O}_\delta(\mathcal{I}_h u)$ into itself. Then appealing to Brouwer's fixed point theorem yields the existence of a solution to problem (4.12), and hence there exists a solution to problem (4.4). The following lemma is a key result for proving the existence of a unique solution to the discrete problem (4.4). Throughout this section we use the following notation to denote the Sobolev norm of u:

$$C_u = \max\left[\ \|u\|_{H^2(\Omega)},\ \|u\|_{H^1(\Omega)}|u|_{W^1_\infty(\Omega)}\right]. \tag{4.17}$$

LEMMA 4.2. *Let $\beta \ge 1$ and $z, v_h \in \mathcal{D}_p(\mathcal{T}_h)$. Set $\chi = z - \mathcal{I}_h u$ and $\eta = u - \mathcal{I}_h u$. Then there exists a constant $C > 0$ which is independent of h and p such that*

$$|\mathcal{F}(z; z-u,\ v_h)| \le CC_a \left[\left(\max_{1\le i\le N_h} \frac{p_i}{h_i}\right)^{1/2} |||\chi|||^2 + C_u h^{1/2}(|||\chi||| + |||\eta|||)\right] |||v_h|||.$$

Proof. Let $z \in \mathcal{D}_p(\mathcal{T}_h)$ and set $\zeta = z - u$. In (4.13), we now replace u_h by z and e by $z - u$ to obtain

$$\begin{aligned}\mathcal{F}(z;\zeta,\ v_h) = &\sum_{i=1}^{N_h}\int_{K_i} \tilde{a}_u(z)\zeta\nabla\zeta\cdot\nabla v_h - \sum_{e_k\in\Gamma_I}\int_{e_k}\left\{\tilde{a}_u(z)\zeta\frac{\partial\zeta}{\partial\nu}\right\}[v_h]\\ &- \theta\sum_{e_k\in\Gamma_I}\int_{e_k}\left\{\tilde{a}_u(z)\zeta\frac{\partial v_h}{\partial\nu}\right\}[\zeta] + \sum_{i=1}^{N_h}\int_{K_i}\tilde{a}_{uu}(z)\zeta^2\nabla u\cdot\nabla v_h\\ &- \sum_{e_k\in\Gamma_I}\int_{e_k}\left\{\tilde{a}_{uu}(z)\zeta^2\frac{\partial u}{\partial\nu}\right\}[v_h].\end{aligned} \tag{4.18}$$

We split $\zeta = \chi - \eta$, where $\chi = z - \mathcal{I}_h u$ and $\eta = u - \mathcal{I}_h u$. Then we estimate from below the bound for each term on the right-hand side of (4.18). For the first term on the right-hand side of (4.18), we split and then bound it as

$$\begin{aligned}\sum_{i=1}^{N_h}\int_{K_i}|\tilde{a}_u(z)\zeta\nabla\zeta\cdot\nabla v_h| \le{}& C_a\sum_{i=1}^{N_h}\int_{K_i}|\chi\nabla\chi\cdot\nabla v_h| + C_a\sum_{i=1}^{N_h}\int_{K_i}|\chi\nabla\eta\cdot\nabla v_h|\\ &+ C_a\sum_{i=1}^{N_h}\int_{K_i}|\eta\nabla\chi\cdot\nabla v_h| + C_a\sum_{i=1}^{N_h}\int_{K_i}|\eta\nabla\eta\cdot\nabla v_h|.\end{aligned} \tag{4.19}$$

Using Hölder's inequality and the inverse inequality (2.6), we estimate the first term on the right-hand side of (4.19) as

$$\begin{aligned}\sum_{i=1}^{N_h}\int_{K_i}|\chi\nabla\chi\cdot\nabla v_h| &\le \sum_{i=1}^{N_h}\|\chi\|_{L^6(K_i)}\|\nabla\chi\|_{L^3(K_i)}\|\nabla v_h\|_{L^2(K_i)}\\ &\le C\sum_{i=1}^{N_h}\|\chi\|_{L^6(K_i)}\frac{p_i^{1/3}}{h_i^{1/3}}\|\nabla\chi\|_{L^2(K_i)}\|\nabla v_h\|_{L^2(K_i)}\end{aligned}$$

$$\leq C\|\chi\|_{L^6(\Omega)}\left(\sum_{i=1}^{N_h}\frac{p_i}{h_i}\|\nabla\chi\|^3_{L^2(K_i)}\right)^{1/3}|v_h|_{1,h}$$

$$\leq C\|\chi\|_{L^6(\Omega)}\left[\max_{K_i}\|\nabla\chi\|^{1/3}_{L^2(K_i)}\left(\sum_{i=1}^{N_h}\frac{p_i}{h_i}\|\nabla\chi\|^2_{L^2(K_i)}\right)^{1/3}\right]|v_h|_{1,h}$$

$$\leq C|||\chi|||\left(\sum_{i=1}^{N_h}\|\nabla\chi\|^2_{L^2(K_i)}\right)^{1/6}\left(\sum_{i=1}^{N_h}\frac{p_i}{h_i}\|\nabla\chi\|^2_{L^2(K_i)}\right)^{1/3}|||v_h|||$$

$$\leq C\left(\max_{1\leq i\leq N_h}\frac{p_i}{h_i}\right)^{1/3}|||\chi|||^2\ |||v_h|||. \tag{4.20}$$

For the second term on the right-hand side of (4.19), we use Hölder's inequality and Lemma 2.1 to obtain

$$\sum_{i=1}^{N_h}\int_{K_i}|\chi\nabla\eta\cdot\nabla v_h| \leq C\sum_{i=1}^{N_h}\|\chi\|_{L^6(K_i)}\|\nabla\eta\|_{L^3(K_i)}\|\nabla v_h\|_{L^2(K_i)}$$

$$\leq C\sum_{i=1}^{N_h}\|\chi\|_{L^6(K_i)}\frac{h_i^{2/3}}{p_i^{2/3}}\|u\|_{H^2(K_i)}\|\nabla v_h\|_{L^2(K_i)}$$

$$\leq C\frac{h^{2/3}}{p^{2/3}}\|\chi\|_{L^6(\Omega)}\left(\sum_{i=1}^{N_h}\|u\|^3_{H^2(K_i)}\right)^{1/3}|v_h|_{1,h}$$

$$\leq C\frac{h^{2/3}}{p^{2/3}}|||\chi|||\left[\max_{K_i}\|u\|^{1/3}_{H^2(K_i)}\left(\sum_{i=1}^{N_h}\|u\|^2_{H^2(K_i)}\right)^{1/3}\right]|||v_h|||$$

$$\leq C\frac{h^{2/3}}{p^{2/3}}|||\chi|||\left(\sum_{i=1}^{N_h}\|u\|^2_{H^2(K_i)}\right)^{1/2}|||v_h|||$$

$$\leq C\frac{h^{2/3}}{p^{2/3}}\|u\|_{H^2(\Omega)}|||\chi|||\ |||v_h|||. \tag{4.21}$$

To estimate the third term on the right-hand side of (4.19), apply Hölder's inequality and the inverse inequality (2.6) to find, following estimate (4.22), that

$$\sum_{i=1}^{N_h}\int_{K_i}|\eta\nabla\chi\cdot\nabla v_h| \leq C\sum_{i=1}^{N_h}\|\eta\|_{L^6(K_i)}\|\nabla\chi\|_{L^3(K_i)}\|\nabla v_h\|_{L^2(K_i)}$$

$$\leq C\sum_{i=1}^{N_h}\frac{h_i^{4/3}}{p_i^{4/3}}\|u\|_{H^2(K_i)}\frac{p_i^{2/3}}{h_i^{2/3}}\|\nabla\chi\|_{L^2(K_i)}\|\nabla v_h\|_{L^2(K_i)}$$

$$\leq C\frac{h^{2/3}}{p^{2/3}}\|u\|_{H^2(\Omega)}|||\chi|||\ |||v_h|||. \tag{4.22}$$

For the last term on the right-hand side of (4.19), we use Lemmas 2.1 and 2.5 to estimate it as

$$\sum_{i=1}^{N_h}\int_{K_i}|\eta\nabla\eta\cdot\nabla v_h| \leq C\sum_{i=1}^{N_h}\|\eta\|_{L^6(K_i)}\|\nabla\eta\|_{L^3(K_i)}\|\nabla v_h\|_{L^2(K_i)}$$

$$\leq C\frac{h^{2/3}}{p^{2/3}}\|u\|_{H^2(\Omega)}|||\eta|||\ |||v_h|||. \tag{4.23}$$

We substitute the estimates (4.20)–(4.23) into (4.19) to obtain

$$\sum_{i=1}^{N_h}\int_{K_i}|\tilde{a}_u(z)\zeta\nabla\zeta\cdot\nabla v_h| \le CC_a\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2\,|||v_h|||$$

$$+\,CC_a\frac{h^{2/3}}{p^{2/3}}\|u\|_{H^2(\Omega)}\left(|||\chi|||+|||\eta|||\right)|||v_h|||. \tag{4.24}$$

As in (4.19) the second term on the right-hand side of (4.18) becomes

$$\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\left\{\tilde{a}_u(z)\zeta\frac{\partial\zeta}{\partial\nu}\right\}[v_h]\right| \le C\left(C_a\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\chi\frac{\partial\chi}{\partial\nu}\right||[v_h]| + C_a\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\chi\frac{\partial\eta}{\partial\nu}\right||[v_h]|\right.$$

$$\left.+\,C_a\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\eta\frac{\partial\chi}{\partial\nu}\right||[v_h]| + C_a\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\eta\frac{\partial\eta}{\partial\nu}\right||[v_h]|\right). \tag{4.25}$$

Using Hölder's inequality, the inverse inequality (2.8), the trace inequalities (2.4)–(2.5), and (2.1), the first term on the right-hand side of (4.25) is estimated as

$$\begin{aligned}\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\chi\frac{\partial\chi}{\partial\nu}\right||[v_h]| &\le C\sum_{e_k\in\Gamma_I}\left(\frac{|e_k|^{\beta/2}}{p_k}\|\chi\|_{L^4(e_k)}\|\nabla\chi\|_{L^4(e_k)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\right)\\ &\le C\sum_{e_k\in\Gamma_I}\frac{|e_k|^{\beta/2-1/4}}{p_k^{1/2}}\|\chi\|_{L^4(e_k)}\|\nabla\chi\|_{L^2(e_k)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\ &\le C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{p_k^{1/2}}{|e_k|^{1-\beta/2}}\|\nabla\chi\|_{L^2(K_i)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\ &\qquad\left(\|\chi\|^4_{L^4(K_i)}+h_i\|\chi\|^3_{L^6(K_i)}\|\nabla\chi\|_{L^2(K_i)}\right)^{1/4}\\ &\le C\left(\max_{1\le i\le N_h}\frac{p_i}{h_i^{2-\beta}}\right)^{1/2}|||\chi|||^2\,|||v_h|||.\end{aligned} \tag{4.26}$$

Similarly, we use Hölder's inequality, (2.1), the trace inequality (2.4), and Lemma 2.1 to estimate the second term on the right-hand side of (4.25) as

$$\begin{aligned}\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\chi\frac{\partial\eta}{\partial\nu}\right||[v_h]| &\le C\sum_{e_k\in\Gamma_I}\left(\frac{|e_k|^{\beta/2}}{p_k}\|\chi\|_{L^4(e_k)}\|\nabla\eta\|_{L^4(e_k)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\right)\\ &\le C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{|e_k|^{\beta/2-1/2}}{p_k}\left(\|\chi\|^4_{L^4(K_i)}+h_i\|\chi\|^3_{L^6(K_i)}\|\nabla\chi\|_{L^2(K_i)}\right)^{1/4}\\ &\qquad\left(\|\nabla\eta\|^4_{L^4(K_i)}+h_i\|\nabla\eta\|^3_{L^6(K_i)}\|\nabla^2\eta\|_{L^2(K_i)}\right)^{1/4}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\ &\le C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{h_i^{\beta/2-1/2}}{p_k}\left(\frac{h_i^2}{p_i^2}\|u\|^4_{H^2(K_i)}+h_i\frac{h_i}{p_i}\|u\|^4_{H^2(K_i)}\right)^{1/4}\\ &\qquad\left(\|\chi\|^4_{L^4(K_i)}+h_i\|\chi\|^3_{L^6(K_i)}\|\nabla\chi\|_{L^2(K_i)}\right)^{1/4}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\ &\le C\|u\|_{H^2(\Omega)}\left(\frac{h^{\beta/2}}{p^{3/2}}+\frac{h^{\beta/2}}{p^{5/4}}\right)|||\chi|||\;|||v_h|||.\end{aligned} \tag{4.27}$$

For the third term on the right-hand side of (4.25), apply Hölder's inequality, the trace inequalities (2.4)–(2.5), and Lemma 2.1 to find that

$$\begin{aligned}\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\eta\frac{\partial\chi}{\partial\nu}\right||[v_h]| &\le C\sum_{e_k\in\Gamma_I}\left(\frac{|e_k|^{\beta/2}}{p_k}\|\eta\|_{L^4(e_k)}\|\nabla\chi\|_{L^4(e_k)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\right)\\ &\le C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{|e_k|^{\beta/2-1/2}}{p_k^{1/2}}\|\nabla\chi\|_{L^2(e_k)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\ &\qquad\left(\|\eta\|^4_{L^4(K_i)}+h_i\|\eta\|^3_{L^6(K_i)}\|\nabla\eta\|_{L^2(K_i)}\right)^{1/4}\\ &\le C\|u\|_{H^2(\Omega)}\left(\frac{h^{1/2+\beta/2}}{p}+\frac{h^{1/2+\beta/2}}{p^{3/4}}\right)|||\chi|||\ |||v_h|||. \end{aligned}\tag{4.28}$$

For the last term on the right-hand side of (4.25), we use an argument similar to that of (4.28) to obtain

$$\begin{aligned}\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\eta\frac{\partial\eta}{\partial\nu}\right||[v_h]| &\le C\sum_{e_k\in\Gamma_I}\left(\frac{|e_k|^{\beta/2}}{p_k}\|\eta\|_{L^4(e_k)}\|\nabla\eta\|_{L^4(e_k)}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\right)\\ &\le C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{|e_k|^{\beta/2-1/2}}{p_k}\left(\|\eta\|^4_{L^4(K_i)}+h_i\|\eta\|^3_{L^6(K_i)}\|\nabla\eta\|_{L^2(K_i)}\right)^{1/4}\\ &\qquad\left(\|\nabla\eta\|^4_{L^4(K_i)}+h_i\|\nabla\eta\|^3_{L^6(K_i)}\|\nabla^2\eta\|_{L^2(K_i)}\right)^{1/4}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\ &\le C\|u\|_{H^2(\Omega)}\left(\frac{h^{\beta/2}}{p^{3/2}}+\frac{h^{\beta/2}}{p^{5/4}}\right)|||\eta|||\ |||v_h|||. \end{aligned}\tag{4.29}$$

We substitute the estimates (4.26)–(4.29) into (4.25). Since $\beta\ge 1$, we obtain

$$\begin{aligned}\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\zeta\frac{\partial\zeta}{\partial\nu}\right||[v_h]| &\le CC_a\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2\ |||v_h|||\\ &\quad+CC_ah^{1/2}\|u\|_{H^2(\Omega)}(|||\chi|||+|||\eta|||)|||v_h|||. \end{aligned}\tag{4.30}$$

In a similar way, we find the following estimates for the third, fourth, and last terms on the right-hand side of (4.18):

$$\begin{aligned}\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\zeta\frac{\partial v_h}{\partial\nu}\right||[\zeta]| &\le CC_a\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2\ |||v_h|||\\ &\quad+CC_ah^{1/2}\|u\|_{H^2(\Omega)}(|||\chi|||+|||\eta|||)|||v_h|||, \end{aligned}\tag{4.31}$$

$$\begin{aligned}\sum_{i=1}^{N_h}\int_{K_i}|(\tilde a_{uu}(z)\nabla u)\zeta^2\cdot\nabla v_h| &\le CC_a|||\chi|||^2\ |||v_h|||\\ &\quad+CC_ah^{1/2}\|u\|_{H^1(\Omega)}|u|_{W^1_\infty(\Omega)}(|||\chi|||+|||\eta|||)|||v_h|||, \end{aligned}\tag{4.32}$$

and

$$\sum_{e_k\in\Gamma}\int_{e_k}\left|\left\{\left(\tilde{a}_{uu}(z)\frac{\partial u}{\partial\nu}\right)\zeta^2\right\}[v_h]\right| \le CC_a|||\chi|||^2\ |||v_h|||$$

$$+\ CC_a h^{1/2}\|u\|_{H^1(\Omega)}|u|_{W^1_\infty(\Omega)}(|||\chi||| + |||\eta|||)|||v_h|||. \tag{4.33}$$

Substituting the estimates (4.24) and (4.30)–(4.33) into (4.18), we complete the rest of the proof. □

LEMMA 4.3. *Let $\beta \ge 1$ and $z \in \mathcal{D}_p(\mathcal{T}_h)$. Set $y = S_h z$. Then there exists a positive constant C which is independent of h and p such that*

$$|||\mathcal{I}_h u - y||| \le CC_a\left[\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\mathcal{I}_h u - z|||^2 + C_u h^{1/2}|||\mathcal{I}_h u - z|||\right]$$
$$+\ CC_a\left[(1 + C_u h^{1/2})|||\mathcal{I}_h u - u|||_+\right].$$

Proof. Let $\chi = \mathcal{I}_h u - z$, $\eta = \mathcal{I}_h u - u$, and $\xi = \mathcal{I}_h u - y$. Set $v_h = \xi$ in (4.16). Then for the first term on the right-hand side of (4.16), use Lemma 3.3 to obtain

$$|\tilde{B}(u;\eta,\xi)| \le C|||\eta|||_+\ |||\xi|||. \tag{4.34}$$

Set $v_h = \xi$ in Lemma 4.2 to arrive at

$$|\mathcal{F}(z; z-u,\ \xi)| \le CC_a\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2\ |||\xi||| \tag{4.35}$$
$$+\ CC_a C_u h^{1/2}\left(|||\chi||| + |||\eta|||\right)\ |||\xi|||.$$

Substituting the estimates (4.34)–(4.35) into (4.16) and using the fact that $|||\eta||| \le |||\eta|||_+$, we obtain

$$|\tilde{B}(u;\xi,\ \xi) \le CC_a\left[\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2 + C_u h^{1/2}\left(|||\chi||| + |||\eta|||\right) + |||\eta|||_+\right]|||\xi|||$$
$$\le CC_a\left[\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2 + C_u h^{1/2}|||\chi|||\right]|||\xi|||$$
$$+\ CC_a(1 + C_u h^{1/2})|||\eta|||_+\ |||\xi|||.$$

Then using the Gårding inequality, that is, Lemma 3.2, we obtain

$$C_1|||\xi||^2 - C_2\|\xi\|^2 \le \tilde{B}(u;\xi,\ \xi)$$
$$\le CC_a\left[\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2 + C_u h^{1/2}|||\chi|||\right]|||\xi|||$$
$$+\ CC_a\left[(1 + C_u h^{1/2})|||\eta|||_+\right]|||\xi|||,$$

and hence

$$|||\xi||| \le CC_a\left[\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}|||\chi|||^2 + C_u h^{1/2}|||\chi|||\right] \tag{4.36}$$
$$+\ CC_a\left[(1 + C_u h^{1/2})|||\eta|||_+\right] + C\|\xi\|.$$

In order to complete the proof of the lemma, it is now sufficient to obtain an estimate for $\|\xi\|$. Setting $q = \xi$ and $v_h = \xi$ in Lemma 3.5, it follows that

$$\begin{aligned}\|\xi\|^2 &= \tilde{B}(u; \mathcal{I}_h u - y,\ \phi_h) = \tilde{B}(u; \mathcal{I}_h u - u,\ \phi_h) + \mathcal{F}(z; z-u, \phi_h)\\ &\le CC_a \left[\left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} |||\chi|||^2 + C_u h^{1/2} |||\chi||| + (1 + C_u h^{1/2}) |||\eta|||_+ \right] |||\phi_h|||.\end{aligned}$$

Therefore, using the fact from Lemma 3.5 that $|||\phi_h||| \le C\|\xi\|$, we obtain

$$\|\xi\| \le CC_a \left[\left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} |||\chi|||^2 + h^{1/2} C_u |||\chi||| + (1 + C_u h^{1/2}) |||\eta|||_+ \right]. \tag{4.37}$$

We combine inequalities (4.36) and (4.37) to complete the rest of the proof. ∎

THEOREM 4.4. *Let $\beta = 1$. Then, for sufficiently small h, there is a $\delta = \delta(h, p)$ such that the map S_h maps $\mathcal{O}_\delta(\mathcal{I}_h u)$ into itself.*

Proof. Let $z \in \mathcal{O}_\delta(\mathcal{I}_h u)$ and set $y = S_h z$. Choose $\delta = \frac{1}{h^\epsilon} |||\mathcal{I}_h u - u|||_+$ for some $0 < \epsilon \le 1/4$. Then using the fact that $z \in \mathcal{O}_\delta(\mathcal{I}_h u)$, and using Lemma 3.4 with $s_i \ge 2,\ p_i \ge 1$ and (4.15), we obtain

$$\begin{aligned}\left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} |||\mathcal{I}_h u - z|||^2 &\le \left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} \delta^2\\ &\le \left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} \frac{1}{h^\epsilon} |||\mathcal{I}_h u - u|||_+ \ \delta\\ &\le C \frac{1}{h^\epsilon} \left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} \left[\sum_{i=1}^{N_h} \frac{h_i^2}{p_i} \|u\|^2_{H^2(K_i)} \right]^{1/2} \delta\\ &\le C \frac{1}{h^\epsilon} \left(\max_{1\le i\le N_h} \frac{p_i}{h_i} \right)^{1/2} \left(\max_{1\le i\le N_h} \frac{h_i^2}{p_i} \right)^{1/2} \|u\|_{H^2(\Omega)}\ \delta\\ &\le CC_Q C_u h^{1/2-\epsilon}\ \delta. \qquad (4.38)\end{aligned}$$

We substitute (4.38) in Lemma 4.3 to obtain

$$\begin{aligned}|||\mathcal{I}_h u - y||| &\le CC_a \left(C_Q C_u h^{1/2-\epsilon}\ \delta + C_u h^{1/2} \delta + (1 + C_u h^{1/2}) h^\epsilon \delta \right)\\ &\le CC_a \left(C_Q C_u h^{1/2-\epsilon} + C_u h^{1/2} + (1 + C_u h^{1/2}) h^\epsilon \right)\ \delta. \qquad (4.39)\end{aligned}$$

Choose h small so that $CC_a \left(C_Q C_u h^{1/2-\epsilon} + C_u h^{1/2} + (1 + C_u h^{1/2}) h^\epsilon \right) \le 1$, and hence S_h maps $\mathcal{O}_\delta(\mathcal{I}_h u)$ into itself. This completes the rest of the proof. ∎

THEOREM 4.5. *Let $\beta = 1$. There is a $\delta = \delta(h,\ p) > 0$ and a positive constant C such that the following holds for any given $z_1, z_2 \in \mathcal{O}_\delta(\mathcal{I}_h u)$ and $0 < \epsilon \le \frac{1}{4}$:*

$$|||S_h z_1 - S_h z_2||| \le CC_a C_Q C_u h^\epsilon\ |||z_1 - z_2|||.$$

Proof. Set $y_1 = S_h z_1$ and $y_2 = S_h z_2$. Using the definition (4.16) of S_h, it is clear that

$$\tilde{B}(u; y_2 - y_1,\ v_h) = \mathcal{F}(z_1; z_1 - u,\ v_h) - \mathcal{F}(z_2; z_2 - u,\ v_h). \tag{4.40}$$

Choose $\delta = \frac{1}{h^\epsilon} |||\eta|||_+$ for some $0 < \epsilon \le 1/4$ with $\eta = u - \mathcal{I}_h u$. Set $\chi = z_1 - z_2$. Using Taylor's formulae (2.4)–(2.5) and (4.18), we rewrite the first terms from each of the

terms on the right-hand side of (4.40) on each K_i as

$$\begin{aligned}\tilde{a}_{uu}(z_1)(z_1-u)^2-\tilde{a}_{uu}(z_2)(z_2-u)^2 &= a(z_1)-a(z_2)+a_u(u)(z_2-z_1)\\ &= a(z_1)-a(z_2)-a_u(z_2)\chi+(a_u(z_2)-a_u(u))\,\chi\\ &= \tilde{R}(z_1,z_2)\chi^2+\tilde{a}_{uu}(z_2)(z_2-u)\chi,\end{aligned}$$

where $\tilde{R}(z_1,z_2)=\int_0^1(1-t)a_{uu}(z_1+t[z_1-z_2])dt$. Similarly, other terms on the right-hand side of (4.40) can be rewritten in a similar fashion. Now, an argument similar to that of Lemma 4.2 implies that

$$\begin{aligned}|\mathcal{F}(z_1;z_1-u,\ v_h)-\mathcal{F}(z_2;z_2-u,\ v_h)| &\le CC_a\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)^{1/2}\Big[\ |||\chi|||^2\ |||v_h|||\\ &\quad+|||z_1-\mathcal{I}_hu|||\ |||\chi|||\ |||v_h|||+|||\mathcal{I}_hu-z_2|||\ |||\chi|||\ |||v_h|||\ \Big]\\ &\le CC_aC_QC_uh^{\epsilon}\ |||\chi|||\ |||v_h|||. \qquad (4.41)\end{aligned}$$

We set $v_h=(y_2-y_1)$ in (4.40) and (4.41). Then, using Lemmas 3.2 and 3.5, we obtain

$$|||y_1-y_2|||\le CC_aC_QC_uh^{\epsilon}\ |||z_1-z_2|||,$$

and this completes the proof. ∎

For sufficiently small h, we deduce from Theorem 4.5 that there is a $\delta>0$ such that the map $S_h:\mathcal{O}_\delta(\mathcal{I}_hu)\to\mathcal{O}_\delta(\mathcal{I}_hu)$ is continuous. Hence, using Theorems 4.4 and 4.5 and Brouwer's fixed point theorem, we conclude for small h that there exists a $u_h\in\mathcal{O}_\delta(\mathcal{I}_hu)$ such that $S_hu_h=u_h$. Then, from Theorem 4.5, it is clear that u_h is the unique fixed point of S_h. Hence, we have proved that there exists a unique solution u_h to problem (4.4).

4.3. A priori error estimates. Note that, from (4.39), u_h satisfies

$$|||\mathcal{I}_hu-u_h|||\le CC_a\left(C_QC_uh^{1/2-\epsilon}+C_uh^{1/2}+(1+C_uh^{1/2})h^{\epsilon}\right)\ \delta.$$

Since $\delta=\frac{1}{h^{\epsilon}}|||\eta|||_+$ for any $0<\epsilon\le 1/4$, we obtain

$$\begin{aligned}|||\mathcal{I}_hu-u_h||| &\le CC_a\left(C_QC_uh^{1/2-\epsilon}+C_uh^{1/2}+(1+C_uh^{1/2})h^{\epsilon}\right)\ \frac{1}{h^{\epsilon}}|||\eta|||_+\\ &\le CC_aC_QC_u|||\eta|||_+. \qquad (4.42)\end{aligned}$$

Using Lemma 3.4, estimate (4.42), and a triangle inequality, we have obtained the following estimate which is optimal in h and suboptimal in p.

THEOREM 4.6. *Let $\beta=1$. Then, for sufficiently small h, there exists a constant $C=C(\alpha,M)$ which is independent of h and p such that the solution u_h of the problem* (4.4) *satisfies*

$$|||u-u_h|||\le CC_aC_QC_u\left(\sum_{i=1}^{N_h}\frac{h_i^{2\mu_i-2}}{p_i^{2s_i-3}}\|u\|^2_{H^{s_i}(K_i)}\right)^{1/2},$$

where $\mu_i=\min\{p_i+1,\ s_i\}$, and C_a, C_Q, and C_u are as in (4.8), (4.14), *and* (4.17), *respectively.*

Remark 4.1. Note that the estimate obtained in Theorem 4.6 is optimal in h and suboptimal in p. However, this results leads to precisely the same h-optimal and p-suboptimal rate of convergence in the broken H^1-norm as in the case of linear elliptic problem, when it is approximated by the NIPG method [23, Theorem 3.1], [16].

4.4. Optimal error estimates in the broken energy norm and the L^2-norm, when $u \in H_p^s(\Omega)$, $s \geq 2$. In the following, with the additional assumptions on the mesh and g, we prove optimal error estimates in the broken energy norm as well as in the L^2-norm. Therefore, along with Assumption Q, we assume that $\mathcal{T}_h$ is a regular subdivision of Ω into triangles or rectangles and $\mathcal{D}_p(\mathcal{T}_h) = \mathcal{D}_p^*(\mathcal{T}_h)$. We also assume that there is a $v \in \mathcal{D}_p^*(\mathcal{T}_h)$ such that $g = v|_{\partial\Omega}$.

We note from [23, pp. 908–913] that by using a continuous interpolant $\mathcal{I}_h^c u \in \mathcal{D}_p(\mathcal{T}_h) \cap \bar{C}^0(\Omega)$ of u instead of $\mathcal{I}_h u$, which may be discontinuous across the edges in Γ_I, the optimal rate of convergence can be recovered. Since the construction of $\mathcal{I}_h^c$ is not discussed in [23], we present below the results related to the construction of $\mathcal{I}_h^c$. The idea of constructing $\mathcal{I}_h^c u \in \mathcal{D}_p^*(\mathcal{T}_h) \cap C^0(\bar{\Omega})$ is to modify the sequence $u_p^{h_i} \in \mathcal{D}_p^*(\mathcal{T}_h)$ in Lemma 2.1 by adding suitable piecewise polynomials on each K_i. For more on the construction of $\mathcal{I}_h^c$, we refer to [5, Theorem 4.1], [6, Theorem 4.6], [1, Theorem 4], and [2, Theorem 3]. Following these constructions, we prove the following lemma.

LEMMA 4.7. *Let $\mathcal{T}_h$ be a regular subdivision. Then, for a given $\phi \in H_p^s(\Omega)$, $s \geq 2$, there exists a positive constant C_{A_c} (depending on s but independent of ϕ, p, and h) and an $\mathcal{I}_h^c\phi \in \mathcal{D}_p^*(\mathcal{T}_h) \cap C^0(\bar{\Omega})$ such that for all K_i and e_k*

(i) $\mathcal{I}_h^c\phi|_{\partial\Omega} = \phi|_{\partial\Omega}$;

(ii) *for any $0 \leq l \leq s$ and $0 \leq l \leq 2$,*

$$\|\phi - \mathcal{I}_h^c\phi\|_{H^l(K_i)} \leq C_{A_c} \frac{h_i^{\mu-l}}{p^{s-l-\delta_1}} \left(\sum_{K_j \in K_i^*} \|\phi\|_{H^s(K_j)}^2 \right)^{1/2},$$

where $\delta_1 = 0$ if $l = 0, 1$ and $\delta_1 = 1$ if $l = 2$;

(iii) *for $s > l + \frac{1}{2}$ and $l = 0, 1$,*

$$|\phi - \mathcal{I}_h^c\phi|_{H^l(e_k)} \leq C_{A_c} \frac{h_i^{\mu-l-1/2}}{p^{s-l-1/2-\delta_2}} \left(\sum_{K_j \in K_i^*} \|\phi\|_{H^s(K_j)}^2 \right)^{1/2},$$

where $\delta_2 = 0$ if $l = 0$ and $\delta_2 = 1/2$ if $l = 1$;

(iv) *for $0 \leq l \leq s - 1 + 2/r$ and $l = 0,\ 1$,*

$$\|\phi - \mathcal{I}_h^c\phi\|_{W_r^l(K_i)} \leq C_{A_c} \frac{h_i^{\mu-l-1+2/r}}{p^{s-l-1+2/r}} \left(\sum_{K_j \in K_i^*} \|\phi\|_{H^s(K_j)}^2 \right)^{1/2},$$

where $\mu = \min(s,\ p+1)$, $K_i^* = \{K_j : |\partial K_i \cap \partial K_j| > 0\}$, and e_k is an edge on ∂K_i.

Remark 4.2. Note that Assumption P(1) implies that the cardinality of K_i^* is bounded by N_κ for all i.

Proof of Lemma 4.7. Statement (i) in the lemma is proved in [6, Theorem 4.6]. For $0 \leq l \leq 1$, the approximation property in (ii) is proved in [2, Theorem 3], [6, Theorem 4.6]. Using the inverse inequality (2.7) and Lemma 2.1, the proof of property (ii) for $l = 2$ is as follows:

$$\begin{aligned}\|\phi - \mathcal{I}_h^c\phi\|_{H^2(K_i)} &\leq \|\phi - \mathcal{I}_h\phi\|_{H^2(K_i)} + \|\mathcal{I}_h\phi - \mathcal{I}_h^c\phi\|_{H^2(K_i)} \\ &\leq \|\phi - \mathcal{I}_h\phi\|_{H^2(K_i)} + C\frac{p^2}{h_i}\|\mathcal{I}_h\phi - \mathcal{I}_h^c\phi\|_{H^1(K_i)}\end{aligned}$$

$$\le \|\phi - \mathcal{I}_h\phi\|_{H^2(K_i)} + C\frac{p^2}{h_i}\|\phi - \mathcal{I}_h\phi\|_{H^1(K_i)} + \frac{p^2}{h_i}\|\phi - \mathcal{I}_h^c\phi\|_{H^1(K_i)}$$
$$\le C\frac{h_i^{\mu-2}}{p^{s-3}}\left(\sum_{K_j\in K^*}\|\phi\|^2_{H^s(K_j)}\right)^{1/2}.$$

Then, using the trace inequality (2.4), we deduce property (iii) of the lemma. Finally, using arguments similar to those of [2, Theorem 3], property (iv) can be easily proved. ∎

Remark 4.3. The approximation property (ii) for $l = 2$ and property (iii) for $l = 1$ are not optimal in terms of p. But as we see in our next analysis, these properties do not affect the accuracy of the approximation u_h.

LEMMA 4.8. *Let $\mathcal{T}_h$ be a regular subdivision and let $\mathcal{D}_p(\mathcal{T}_h) = \mathcal{D}_p^*(\mathcal{T}_h)$. Then, for any $\beta \ge 1$ and given any $\phi \in H_p^s(\Omega)$, $s \ge 2$, there exists a constant C independent of h and p such that*

$$|||\phi - \mathcal{I}_h^c\phi|||_+ \le C\left(\sum_{i=1}^{N_h}\frac{h_i^{2\mu-2}}{p^{2s-2}}\|\phi\|^2_{H^s(K_i)}\right)^{1/2}, \tag{4.43}$$

where $\mu = \min\{p+1, s\}$.

Proof. Let $\eta^* = \phi - \mathcal{I}_h^c\phi$. Since $\mathcal{I}_h^c\phi \in \mathcal{D}_p^*(\mathcal{T}_h) \cap C^0(\bar\Omega)$ and $\mathcal{I}_h^c\phi|_{\partial\Omega} = \phi|_{\partial\Omega}$, the jump $[\phi - \mathcal{I}_h^c\phi] = 0$ on each $e_k \in \Gamma$. Hence, using (2.3) and Lemma 4.7, we obtain

$$|||\phi - \mathcal{I}_h^c\phi|||_+^2 = \sum_{i=1}^{N_h}\int_{K_i}|\nabla\eta^*|^2 + \sum_{e_k\in\Gamma}\int_{e_k}\frac{|e_k|^\beta}{p^2}\left\{\frac{\partial\eta*}{\partial\nu}\right\}^2$$
$$\le C\sum_{i=1}^{N_h}\sum_{K_j\in K_i^*}\left(\frac{h_i^{2\mu-2}}{p^{2s-2}}\|\phi\|^2_{H^s(K_j)} + \frac{h_i^\beta}{p^2}\frac{h_i^{2\mu-3}}{p^{2s-4}}\|\phi\|^2_{H^s(K_j)}\right)$$
$$\le C\sum_{i=1}^{N_h}\left(\frac{h_i^{2\mu-2}}{p^{2s-2}}\|\phi\|^2_{H^s(K_i)} + \frac{h_i^{2\mu-2}}{p^{2s-2}}\|\phi\|^2_{H^s(K_i)}\right).$$

This completes the rest of the proof. ∎

THEOREM 4.9. *Let $\mathcal{T}_h$ be a regular subdivision and let $\mathcal{D}_p(\mathcal{T}_h) = \mathcal{D}_p^*(\mathcal{T}_h)$. Suppose that $u \in H_p^s(\Omega)$, $s \ge 2$. Then, for any $\beta \ge 1$ and for sufficiently small h, there exists a constant $C = C(\alpha, M)$ which is independent of h and p such that the solution u_h of problem* (4.4) *satisfies*

$$|||u - u_h||| \le CC_aC_QC_u\left(\sum_{i=1}^{N_h}\frac{h^{2\mu-2}}{p^{2s-2}}\|u\|^2_{H^s(K_i)}\right)^{1/2},$$

where $\mu = \min\{p+1,\ s\}$, and C_a, C_Q, and C_u are as in (4.8), (4.14), *and* (4.17), *respectively.*

Proof. Under the hypotheses on the mesh, there is an $\mathcal{I}_h^c\phi \in \mathcal{D}_p^*(\mathcal{T}_h)\cap C^0(\bar\Omega)$ such that $\mathcal{I}_h^c\phi|_{\partial\Omega} = \phi|_{\partial\Omega}$. Hence, the jump $[\phi - \mathcal{I}_h^c\phi] = 0$ on each $e_k \in \Gamma$. Then, using Lemmas 3.2 and 4.7, it is easy to prove Lemma 3.5 for any $\beta \ge 1$. Now, note that estimates (4.27) and (4.29) in Lemma 4.2 depend on the approximation property (ii) for $l = 2$. Though there is a suboptimality in this property, we still obtain the results

in Lemma 4.2 by replacing $\mathcal{I}_h u$ by $\mathcal{I}_h^c u$ and for any $\beta \geq 1$. Below, we only indicate the changes to be made in the text of the proof of Lemma 4.2. We now consider the term on the left-hand side of (4.27) with $\eta = u - \mathcal{I}_h^c u$. Then, using Hölder's inequality, the trace inequality (2.4), and Lemma 4.7, we estimate this term as follows:

$$\begin{aligned}
\sum_{e_k\in\Gamma_I}\int_{e_k}\left|\left\{\chi\frac{\partial\eta}{\partial\nu}\right\}[v_h]\right| &\leq C\sum_{e_k\in\Gamma_I}\left(\frac{|e_k|^{\beta/2}}{p}\|\chi\|_{L^4(e_k)}\|\nabla\eta\|_{L^4(e_k)}\left(\int_{e_k}\frac{p^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\right)\\
&\leq C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{|e_k|^{\beta/2-1/2}}{p}\left(\|\chi\|_{L^4(K_i)}^4+h_i\|\chi\|_{L^6(K_i)}^3\|\nabla\chi\|_{L^2(K_i)}\right)^{1/4}\\
&\quad\left(\|\nabla\eta\|_{L^4(K_i)}^4+h_i\|\nabla\eta\|_{L^6(K_i)}^3\|\nabla^2\eta\|_{L^2(K_i)}\right)^{1/4}\left(\int_{e_k}\frac{p^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\
&\leq C\sum_{i=1}^{N_h}\sum_{e_k\in\partial K_i}\frac{h^{\beta/2-1/2}}{p}\left(\frac{h^2}{p^2}\|u\|_{H^2(K_i)}^4+h\frac{h}{p}\frac{1}{p-1}\|u\|_{H^2(K_i)}^4\right)^{1/4}\\
&\quad\left(\|\chi\|_{L^4(K_i)}^4+h_i\|\chi\|_{L^6(K_i)}^3\|\nabla\chi\|_{L^2(K_i)}\right)^{1/4}\left(\int_{e_k}\frac{p^2}{|e_k|^\beta}[v_h]^2\right)^{1/2}\\
&\leq C\|u\|_{H^2(\Omega)}\left(\frac{h^{\beta/2}}{p^{3/2}}+\frac{h^{\beta/2}}{p}\right)|||\chi|||\;|||v_h|||. \qquad (4.44)
\end{aligned}$$

Similarly, a replacement of $\mathcal{I}_h u$ by $\mathcal{I}_h^c u$ in Lemma 4.3 and an application of Lemma 4.8 yield the proofs of Theorems 4.4 and 4.5 for any $\beta \geq 1$. Hence, the estimate (4.42) holds for any $\beta \geq 1$ with $\eta = u - \mathcal{I}_h^c u$. Then an application of Lemma 4.8 completes the rest of the proof. ∎

Now, we proceed to derive the L^2-norm error estimate. Since the SIPG method is adjoint consistent, one can expect optimal L^2-norm error estimates in terms of h. But for the NIPG method, the bilinear form is not adjoint consistent. In general, it may be difficult to prove the optimal L^2 error estimate in terms of h. However, if $u \in H_p^s(\Omega),\ s \geq 2$, it is possible to obtain optimal L^2-norm error estimates in terms of both h and p by increasing the penalty on the uniform regular subdivision. Assume that the hypotheses of Theorem 4.9 hold. Of course, these assumptions are not necessary to derive optimal L^2-norm error estimate in terms of h for the SIPG method. Below, we appeal to the Aubin–Nitsche duality argument to estimate $\|u - u_h\|$.

THEOREM 4.10. *Let $a \in C_b^2(\bar{\Omega}\times\mathbb{R})$ and $u \in W_\infty^1(\Omega)$. Suppose that $\beta \geq 3$ when $\theta = -1$, and $\beta \geq 1$ when $\theta = 1$. Further, assume that the hypotheses of Theorem 4.9 hold. Then there exists a constant $C = C(\alpha,\ M)$ such that for small h*

$$\|u - u_h\| \leq CC_Q C_a C_u^2 \frac{h^\mu}{p^s}\|u\|_{\mathbf{s},h},$$

where $\mu = \min\{p+1,\ s\}$, and C_a, C_Q, and C_u are as in (4.8), (4.14), and (4.17), respectively.

Proof. Our assumptions on a and u imply that there is a unique solution $\phi \in H^2(\Omega)$ to the following linear elliptic problem:

$$\begin{aligned}
-\nabla\cdot(a(u)\nabla\phi) + (a_u(u)\nabla u)\cdot\nabla\phi &= e \quad \text{on } \Omega,\\
\phi &= 0 \quad \text{on } \partial\Omega,
\end{aligned}$$

and ϕ satisfies the following elliptic regularity (see [15, Lemma 9.17]):

$$\|\phi\|_{H^2(\Omega)} \le C\|e\|_{L^2(\Omega)}. \tag{4.45}$$

Note that

$$\|e\|^2 = B(u;e,\ \phi) + \int_\Omega (a_u(u)\nabla u)\cdot e\nabla\phi dx + (\theta-1)\sum_{e_k\in\Gamma}\int_{e_k}\left\{a(u)\frac{\partial\phi}{\partial\nu}\right\}[e]. \tag{4.46}$$

The first term on the right-hand side of (4.46) is rewritten as

$$\begin{aligned} B(u;e,\ \phi) &= B(u;u,\ \phi) - B(u_h;u_h,\ \phi) + B(u_h;u_h,\ \phi) - B(u;u_h,\ \phi)\\ &= (B(u;u,\ \phi-\chi) - B(u_h;u_h,\ \phi-\chi)) + (B(u_h;u_h,\ \phi) - B(u;u_h,\ \phi))\\ &= I + II, \end{aligned} \tag{4.47}$$

where $\chi = \mathcal{I}_h^c\phi$ such that $\chi|_{\partial\Omega} = 0$. For the first term on the right-hand side of (4.47), we note that

$$\begin{aligned} I &= B(u;u,\ \phi-\chi) - B(u_h;u,\ \phi-\chi) + B(u_h;u,\ \phi-\chi) - B(u_h;u_h,\ \phi-\chi)\\ &= \sum_{i=1}^{N_h}\int_{K_i}(a(u)-a(u_h))\nabla u\cdot\nabla(\phi-\chi) + \sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u))\nabla(u-u_h)\cdot\nabla(\phi-\chi)\\ &\quad - \sum_{e_k\in\Gamma}\int_{e_k}\left\{(a(u_h)-a(u))\frac{\partial(\phi-\chi)}{\partial\nu}\right\}[u-u_h] + \sum_{i=1}^{N_h}\int_{K_i}a(u)\nabla(u-u_h)\cdot\nabla(\phi-\chi)\\ &\quad - \sum_{e_k\in\Gamma}\int_{e_k}\left\{a(u)\frac{\partial(\phi-\chi)}{\partial\nu}\right\}[u-u_h]. \end{aligned} \tag{4.48}$$

Since $u\in W^{1,\infty}(\Omega)$, we use the Cauchy–Schwarz inequality, Lemma 4.8, to bound the first and fourth terms on the right-hand side of (4.48) as

$$\begin{aligned} \left|\sum_{i=1}^{N_h}\int_{K_i}(a(u)-a(u_h))\nabla u\cdot\nabla(\phi-\chi)\right| &\le C_u\|e\|\ \|\phi-\chi\|_{H^1(\Omega)}\\ &\le CC_u\frac{h}{p}\|e\|\ \|\phi\|_{H^2(\Omega)} \end{aligned} \tag{4.49}$$

and

$$\begin{aligned} \left|\sum_{i=1}^{N_h}\int_{K_i}a(u)\nabla(u-u_h)\cdot\nabla(\phi-\chi)\right| &\le M|||e|||\ \|\phi-\chi\|_{H^1(\Omega)}\\ &\le C\frac{h}{p}|||e|||\ \|\phi\|_{H^2(\Omega)}. \end{aligned} \tag{4.50}$$

Now, using Hölder's inequality and Lemmas 4.8 and 2.5, we estimate the second term on the right-hand side of (4.48) as

$$\begin{aligned} \left|\sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u))\nabla(u_h-u)\cdot\nabla(\phi-\chi)\right| &\le C\ \|e\|_{L^3(\Omega)}|||e|||\ \|\phi-\chi||_{W_6^1(\Omega)}\\ &\le C\ |||e|||^2\ \|\phi\|_{H^2(\Omega)}. \end{aligned} \tag{4.51}$$

Next, using arguments similar to those of (4.44), we bound the third term on the right-hand side of (4.48) as

$$\begin{aligned}\left|\sum_{e_k\in\Gamma}\int_{e_k}\left\{(a(u_h)-a(u))\frac{\partial(\phi-\chi)}{\partial\nu}\right\}[u-u_h]\right| &\le CC_a\sum_{e_k\in\Gamma}\frac{|e_k|^{\beta/2}}{p_k}\|e\|_{L^4(e_k)}|\phi-\chi|_{W_4^1(e_k)}\\ &\quad\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[e]^2ds\right)^{1/2}\\ &\le CC_a|||e|||^2\ \|\phi\|_{H^2(\Omega)}. \end{aligned}\tag{4.52}$$

Then, using Lemma 4.7, the fifth term on the right-hand side of (4.48) is estimated as

$$\left|\sum_{e_k\in\Gamma}\int_{e_k}\left\{a(u)\frac{\partial(\phi-\chi)}{\partial\nu}\right\}[u-u_h]\right|\le C\frac{h^{\beta/2+1/2}}{p}|||e|||\ \|\phi\|_{H^2(\Omega)}.\tag{4.53}$$

Hence using (4.45), we obtain, for any $\beta\ge 1$,

$$|I|\le CC_uC_a\left(|||e|||^2+\frac{h}{p}|||e||||+\|e\|\right)\|e\|.\tag{4.54}$$

For the second term in (4.47), that is, II, we note that $[u]=0$ on $e_k\in\Gamma_I$. Thus,

$$\begin{aligned}II&=\sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u))\nabla u_h\cdot\nabla\phi-\int_{\Gamma_I}\left\{((a(u_h)-a(u))\frac{\partial\phi}{\partial\nu}\right\}[u_h-u]\\ &=\sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u))\nabla(u_h-u)\cdot\nabla\phi+\sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u))\nabla u\cdot\nabla\phi\\ &\quad-\int_{\Gamma_I}\left\{((a(u_h)-a(u))\frac{\partial\phi}{\partial\nu}\right\}[u_h-u]. \end{aligned}\tag{4.55}$$

Use Hölder's inequality, the Sobolev imbedding theorem, and Lemma 2.5 to estimate the first term on the right-hand side of (4.55) as

$$\begin{aligned}\left|\sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u))\nabla(u_h-u)\cdot\nabla\phi\right| &\le C_a\|e\|_{L^3(\Omega)}|e|_{1,h}\|\phi||_{W_6^1(\Omega)}\\ &\le CC_a|||e|||^2\ \|\phi\|_{H^2(\Omega)}. \end{aligned}\tag{4.56}$$

Now for the third term on the right-hand side of (4.55), using Hölder's inequality, the trace inequality (2.4), and Lemma 2.5, we arrive at

$$\begin{aligned}\left|\sum_{e_k\Gamma_I}\int_{e_k}\left\{((a(u_h)-a(u))\frac{\partial\phi}{\partial\nu}\right\}[u_h-u]\right| &\le CC_a\sum_{e_k\Gamma_I}\int_{e_k}\left|e\frac{\partial\phi}{\partial\nu}\right||[e]|\\ &\le CC_a\sum_{e_k\Gamma_I}\frac{|e_k|^{\beta/2}}{p_k}\|e\|_{L^4(e_k)}\|\phi\|_{W_4^1(e_k)}\mathcal{J}^{1,\beta}(e,e)^{1/2}\\ &\le CC_a\frac{h^{\beta/2-1/2}}{p}|||e|||^2\ \|\phi\|_{H^2(\Omega)}. \end{aligned}\tag{4.57}$$

We rewrite the second term on the right-hand side of (4.55) together with the second term on the right-hand side of (4.46) as

$$\sum_{i=1}^{N_h}\int_{K_i}(a(u_h)-a(u)+a_u(u)(u-u_h))\nabla u\cdot\nabla\phi=\sum_{i=1}^{N_h}\int_{K_i}\tilde{a}_{uu}(u_h)(u-u_h)^2\nabla u\cdot\nabla\phi,$$

and then we use Lemma 2.5 to obtain

$$\begin{aligned}\left|\sum_{i=1}^{N_h}\int_{K_i}\tilde{a}_{uu}(u_h)(u-u_h)^2\nabla u\cdot\nabla\phi\right| &\le C_aC_u\left|\sum_{i=1}^{N_h}\int_{K_i}(u-u_h)^2\cdot\nabla\phi\right|\\ &\le CC_aC_u|||e|||^2\ \|\phi\|_{H^2(\Omega)}. \qquad (4.58)\end{aligned}$$

Hence using (4.45), for any $\beta\ge 1$ we obtain

$$|II|\le CC_uC_a|||e|||^2\ \|e\|. \qquad (4.59)$$

For $\theta=1$, the third term on the right-hand side of (4.46) becomes zero. For $\theta=-1$, using the trace inequality (2.4) and (4.45), the third term on the right-hand side of (3.10) is estimated as

$$\begin{aligned}\sum_{e_k\in\Gamma}\int_{e_k}\left\{a(u)\frac{\partial\phi}{\partial\nu}[e]\right\} &\le C\sum_{e_k\in\Gamma}\left(\int_{e_k}\frac{|e_k|^\beta}{p_k^2}\left|\frac{\partial\phi}{\partial\nu}\right|^2\right)^{1/2}\left(\int_{e_k}\frac{p_k^2}{|e_k|^\beta}[e]^2\right)^{1/2}\\ &\le C\frac{h^{\beta/2-1/2}}{p}|||e|||\ \|e\|. \qquad (4.60)\end{aligned}$$

We combine the estimates (4.54)–(4.60) to obtain

$$\|u-u_h\|\le C\left(|||u-u_h|||+\frac{h}{p}+|\theta-1|\frac{h^{\beta/2-1/2}}{p}\right)|||u-u_h|||.$$

Using Theorem 4.9 completes the rest of the proof. ∎

Remark 4.4. In the proof of Lemma 4.2 and the subsequent results in section 4, we have assumed that the range of $\frac{\partial^l a}{\partial u^l}(x,v),\ x\in\bar{\Omega},\ v\in\mathbb{R},\ l=0,1,2$, is a compact set, say $[m,M]\subset\mathbb{R}$. However, if $u\in H^{5/2}(\Omega)$, we note that asymptotically only the values of $v\in[m_u-\delta^*,M_u+\delta^*]\subset\mathbb{R}$, where $0<\delta^*<1$, $m_u=\inf\{u(x):x\in\bar{\Omega}\}$, and $M_u=\sup\{u(x):x\in\bar{\Omega}\}$, are considered to derive the proof of Lemma 4.2 and the subsequent results. To be more precise, the terms $\tilde{a}_u(z)$ and $\tilde{a}_{uu}(z),\ z\in\mathcal{O}_\delta(\mathcal{I}_hu)$, in (4.18) (see the estimates (4.19)–(4.33)), can be estimated as follows. Since $z\in\mathcal{O}_\delta(\mathcal{I}_hu)$, where $\delta=h^{-\epsilon}|||u-\mathcal{I}_hu|||_+,\ 0<\epsilon\le 1/4$, using the inverse inequality (2.6) and Lemmas 2.5 and 2.1, we find that

$$\begin{aligned}\|z-u\|_{L^\infty(\Omega)} &\le \|z-\mathcal{I}_hu\|_{L^\infty(\Omega)}+\|\mathcal{I}_hu-u\|_{L^\infty(\Omega)}\\ &\le C\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)\|z-\mathcal{I}_hu\|_{L^2(\Omega)}+\|\mathcal{I}_hu-u\|_{L^\infty(\Omega)}\\ &\le C\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)|||z-\mathcal{I}_hu|||+\|\mathcal{I}_hu-u\|_{L^\infty(\Omega)}\\ &\le Ch^{-\epsilon}\left(\max_{1\le i\le N_h}\frac{p_i}{h_i}\right)|||u-\mathcal{I}_hu|||_+ +\|\mathcal{I}_hu-u\|_{L^\infty(\Omega)}\end{aligned}$$

$$\leq Ch^{-\epsilon}\left(\max_{1\leq i\leq N_h}\frac{p_i}{h_i}\right)\left(\sum_{i=1}^{N_h}\frac{h_i^3}{p_i^2}\|u\|^2_{H^{5/2}(K_i)}\right)^{1/2}+C\frac{h}{p}\|u\|_{H^2(\Omega)}$$

$$\leq Ch^{-\epsilon}\left(\max_{1\leq i\leq N_h}\frac{p_i}{h_i}\right)\left(\max_{1\leq i\leq N_h}\frac{h_i^{3/2}}{p_i}\right)\|u\|_{H^{5/2}(\Omega)}+C\frac{h}{p}\|u\|_{H^2(\Omega)}$$

$$\leq Ch^{1/2-\epsilon}\|u\|_{H^{5/2}(\Omega)}. \tag{4.61}$$

Therefore, for sufficiently small h, $\|z\|_{L^\infty(\Omega)} \leq \delta^* + \|u\|_{L^\infty(\Omega)}$, where $0 < \delta^* < 1$. Now, since the nonlinear functions a_u and a_{uu} are continuous, they map the compact set $[m_u - \delta^*, M_u + \delta^*]$ into a compact set in $\mathbb{R}$, and hence the results in Lemma 4.2 and the subsequent results in section 4 remain valid when $a(v)$, $a_u(v)$, and $a_{uu}(v)$ are bounded for bounded u. Finally, we remark that when $u \in H^2(\Omega)$, it may be possible to show the boundedness of $a(v)$ and its derivatives for $v \in [m_u - \delta^*, M_u + \delta^*] \subset \mathbb{R}$ by using better inverse inequalities, say in the first line of (4.61), applying $\|z - \mathcal{I}_h u\|_{L^\infty(K_i)} \leq Cp_i^{1/2}h_i^{-1/4}\|z - \mathcal{I}_h u\|_{L^8(K_i)}$ (see [20, p. 916]), and using the Poincaré inequality in Lemma 2.5 to complete the estimate (4.61).

5. Numerical experiments. In this section, we discuss the performance of the proposed NIPG and SIPG methods for the numerical approximation of the quasi-linear elliptic problem (4.1)–(4.2). For this, we consider the following nonlinear elliptic problem:

$$\begin{aligned} -\nabla\cdot((1+u)\nabla u) &= f \quad \text{in} \quad \Omega,\\ u &= 0 \quad \text{on} \quad \partial\Omega, \end{aligned}$$

where $\Omega = (0,1)\times(0,1)$ and f is taken in such a way that the exact solution is $u = x(1-x)y(1-y)$. We divide Ω into regular uniform triangles. The stabilization parameter σ_k, appearing in the penalty term $\mathcal{J}^{\sigma,\beta}$, is taken as follows: $\sigma_k = 10$ for all e_k. We investigate the convergence of NIPG ($\theta = -1$) and SIPG($\theta = 1$) on a sequence of uniform triangular meshes for each of $p = 1, 2$, and 3, where $p = p_i$ for $1 \leq i \leq N_h$. Similarly, we also investigate the convergence of both methods by enriching the polynomial degree p on a fixed mesh.

Convergence in the broken H^1-norm. We set $\beta = 1$ for both the NIPG and SIPG methods. In Figure 2, we plot the broken H^1-norm of the error against the mesh function h for polynomial degrees $p = 1, 2$, and 3. Here, we observe that for each p, $|||u - u_h|||$ converges to zero at the rate $\mathcal{O}(h^p)$ as the mesh is refined. These experiments illustrate the theoretical results obtained in Theorem 4.6. In Figure 3, we present the convergence of the broken H^1-norm of the error as the degree of the polynomials increases on a fixed mesh.

Convergence in the L^2-norm. According to Theorem 4.10, the NIPG method gives optimal L^2 order of convergence, provided the jump term is superpenalized. We take $\beta = 3$ when $\theta = -1$. Since the SIPG method is optimal in the L^2-norm, we take $\beta = 1$ when $\theta = 1$. We investigate the theoretical results obtained in Theorem 4.10 by performing the experiments with the above values of β. In Figure 4, we plot the L^2-norm of the error against the mesh function h for polynomial degrees $p = 1, 2$, and 3. We note that for each p, $\|u - u_h\|$ converges to zero at the rate $\mathcal{O}(h^{p+1})$ as the mesh is refined. The convergence lines are almost the same for both the NIPG and SIPG methods. These results show that the NIPG method exhibits an optimal order of convergence in the L^2-norm on a regular mesh by imposing the

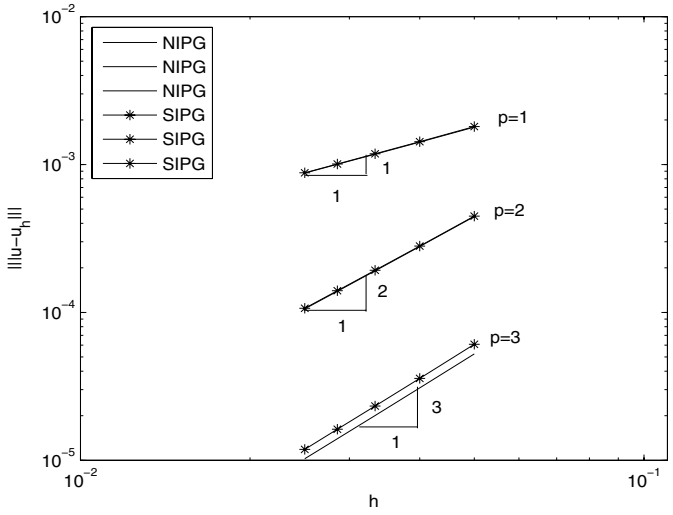


FIG. 2. *Convergence of NIPG and SIPG with h-refinement.*

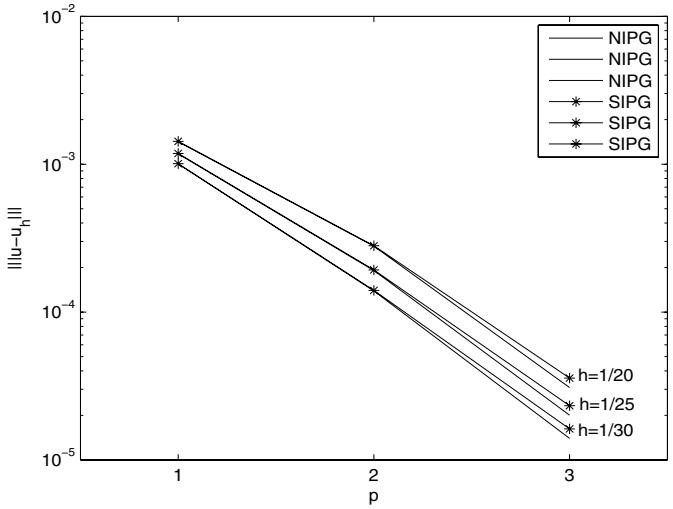


FIG. 3. *Convergence of NIPG and SIPG with p-refinement.*

superpenalty. In Figure 5, we also plot the L^2-norm of the error against the degree of the polynomial p on a fixed mesh. The L^2-norm of the error converges exponentially to zero as p increases. These experiments illustrate the theoretical results obtained in Theorem 4.10.

6. Conclusions. In this paper, we have discussed *hp*-discontinuous Galerkin finite element methods (SIPG and NIPG) for approximating the solutions of nonlinear elliptic boundary value problems of nonmonotone type on a bounded domain in $\mathbb{R}^2$. Using Brouwer's fixed point theorem, we have shown that the discrete problem has a solution under an *hp*-quasi-uniformity condition on the mesh. Further, using Lipschitz

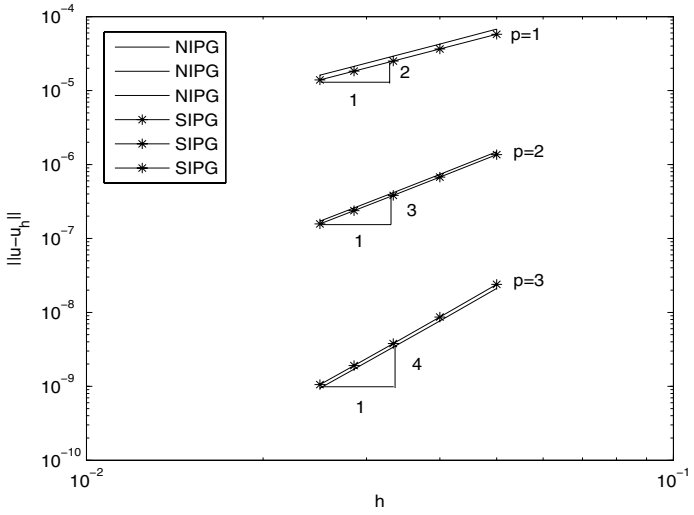


FIG. 4. *Convergence of NIPG and SIPG with h-refinement.*

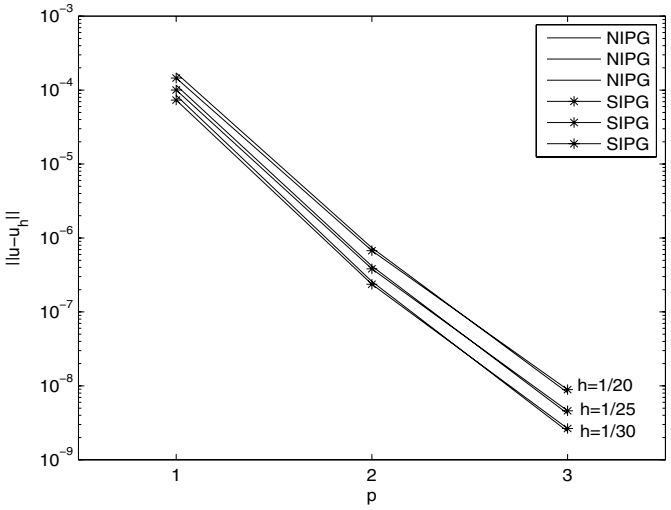


FIG. 5. *Convergence of NIPG and SIPG with p-refinement.*

continuity of the discrete solution map, uniqueness of the discrete solution is also proved. We have proved error estimates in a broken H^1-norm, which are optimal in h and suboptimal in p. These results lead to precisely the same h-optimal and mildly p-suboptimal rates of convergence as in the case of linear elliptic boundary value problems using NIPG methods; see [23], [16]. Further, an optimal error estimate in the L^2-norm on a regular mesh is established by imposing a superpenalty. The results of this article can be easily extended to problems in three space dimensions by making appropriate changes in the analysis. Moreover, it is not difficult to extend our analysis to the problem $-\nabla \cdot (a(u)\nabla u) + f(u) = 0$, where $f(u) \in C_b^2(\bar{\Omega} \times \mathbb{R})$.

With appropriate modifications in the analysis, it is possible to extend the theoretical results of this article to problem (4.1)–(4.2) when $a(u)$ is a bounded uniformly positive-definite matrix.

Acknowledgments. The second author acknowledges the financial support provided by the Isaac Newton Institute for Mathematical Sciences, Cambridge, UK, during May–June 2003 under the program "Computational Challenges in PDEs," where this work was initiated. The second author also thanks Professor E. Süli, one of the organizers of the above program, for fruitful discussions on section 3 of this manuscript. The authors are indebted to both of the referees for their valuable suggestions and constructive comments which helped the authors to improve the manuscript.

REFERENCES

[1] M. Ainsworth and D. Kay, *The approximation theory for the p-version finite element method and application to the nonlinear elliptic PDEs*, Numer. Math., 82 (1999), pp. 351–388.

[2] M. M. Ainsworth and D. Kay, *Approximation theory for the hp-version finite element method and application to the nonlinear Laplacian*, Appl. Numer. Math., 34 (2000), pp. 329–344.

[3] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[4] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[5] I. Babuška and M. Suri, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.

[6] I. Babuška and M. Suri, *The h-p version of the finite element method with quasiuniform meshes*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 199–238.

[7] C. Bernardi, M. Dauge, and Y. Maday, *Polynomials in the Sobolev World*, Preprint R03038, Laboratoire Jacques-Louis Lions, C.N.R.S. et Université Pierre et Marie Curie, Paris, 2003.

[8] S. C. Brenner, *Poincare–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

[9] F. Brezzi, M. Manzini, L. D. Marini, and P. Pietra, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Methods Partial Differential Equations, 16 (2000), pp. 265–278.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978; reprinted as Classics Appl. Math. 40, SIAM, Philadelphia, 2002.

[11] J. Douglas, Jr., T. Dupont, and J. Serrin, *Uniqueness and comparison theorems for nonlinear elliptic equations in divergence form*, Arch. Ration. Mech. Anal., 42 (1971), pp. 157–168.

[12] J. Douglas, Jr., and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (1975), pp. 689–696.

[13] J. Douglas, Jr., and T. Dupont, *Interior penalty procedures for elliptic and parabolic Galerkin methods*, in Computing Methods in Applied Sciences, Lecture Notes in Phys. 58, Springer-Verlag, Berlin, 1976, pp. 207–216.

[14] E. H. Georgoulis, and E. Süli, *Optimal error estimates for the hp-version interior penalty discontinuous Galerkin finite element method*, IMA J. Numer. Anal., 25 (2005), pp. 205–220.

[15] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[16] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[17] P. Houston, J. A. Robson, and E. Süli, *Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems* I: *The scalar case*, IMA J. Numer. Anal., 25 (2005), pp. 726–749.

[18] A. Lasis and E. Süli, *Poincare-Type Inequalities for Broken Sobolev Spaces*, Preprint NI03067-CPD, Isaac Newton Institute for Mathematical Sciences, Cambridge University, Cambridge, UK, 2003.

[19] A. Lasis and E. Süli *One-Parameter Discontinuous Galerkin Finite Element Discretisation of Quasilinear Parabolic Problems*, Research report NA-04/25, Oxford University Computer Laboratory, Oxford, UK, 2004.

[20] F. A. Milner and M. Suri, *Mixed finite element methods for quasilinear elliptic problems: The p-version*, RAIRO Modél. Math. Anal. Numér., 26 (1992), pp. 913–931.
[21] J. T. Oden, I. Babuška, and C. E. Baumann, *A discontinuous hp finite element method for diffusion problems*, J. Comput. Phys., 146 (1998), pp. 491–519.
[22] S. Prudhomme, F. Pascal, and J. T. Oden, *Review of Error Estimation for Discontinuous Galerkin Methods*, Research report 00-27, TICAM, University of Texas, Austin, 2000.
[23] B. Rivière, M. F. Wheeler, and V. Girault, *A priori error estimates for finite element methods based on discontinuous approximation spaces for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 902–931.
[24] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 193–205

A CARDINAL FUNCTION ALGORITHM FOR COMPUTING MULTIVARIATE QUADRATURE POINTS*

MARK A. TAYLOR†, BETH A. WINGATE‡, AND LEN P. BOS§

Abstract. We present a new algorithm for numerically computing quadrature formulas for arbitrary domains which exactly integrate a given polynomial space. An effective method for constructing quadrature formulas has been to numerically solve a nonlinear set of equations for the quadrature points and their associated weights. Symmetry conditions are often used to reduce the number of equations and unknowns. Our algorithm instead relies on the construction of cardinal functions and thus requires that the number of quadrature points N be equal to the dimension of a prescribed lower dimensional polynomial space. The cardinal functions allow us to treat the quadrature weights as dependent variables and remove them, as well as an equivalent number of equations, from the numerical optimization procedure. We give results for the triangle, where for all degrees $d \leq 25$, we find quadrature formulas of this form which have positive weights and contain no points outside the triangle. Seven of these quadrature formulas improve on previously known results.

Key words. multivariate integration, quadrature, cubature, Fekete points, spectral methods, triangle, polynomial approximation

AMS subject classifications. 65D32, 65D30, 65M60, 65M70

DOI. 10.1137/050625801

1. Introduction.

Gauss quadrature points, and related points such as Gauss–Lobatto, are commonly used in numerical methods which rely on both accurate high-order polynomial interpolation and quadrature properties. They are heavily relied on by the diagonal-mass-matrix spectral element method, which has been very successful in geophysical applications dominated by wave propagation [17, 15, 10, 11, 30].

Gauss quadrature points are known only for tensor-product domains such as the line, square and cube. It is unclear how to find Gauss-like points for non–tensor-product domains like triangles or tetrahedrons, which makes it difficult to extend the diagonal-mass-matrix spectral element method to these domains. There are two generalizations that have been studied in some detail. The first involves searching for points in these domains with optimal interpolation properties by minimizing the Lebesgue constant [3, 14, 29]. For high polynomial degree, some of the best results have been obtained for Fekete points, which can be computed in a natural way with a cardinal function algorithm [29]. Cardinal functions, or Lagrange interpolating polynomials, are those which have the value 1 at one of the points and 0 at the remaining points. The second generalization involves searching for points in the domain of interest which give an optimal quadrature formula for the integral of polynomials over the domain. Here we consider a set of N points $\{z_1, z_2, \ldots, z_N\}$ and weights $\{w_1, w_2, \ldots, w_N\}$ to be a quadrature formula of strength d if the quadrature approximation for a

*Received by the editors March 2, 2005; accepted for publication (in revised form) August 1, 2006; published electronically January 12, 2007.
http://www.siam.org/journals/sinum/45-1/62580.html
†Sandia National Laboratory, Exploratory Simulation Technologies, MS 0318, Albuquerque, NM 87185 (mataylo@sandia.gov).
‡Los Alamos National Laboratory, CCS, P.O. Box 1663, Los Alamos, NM 87545 (wingate@lanl.gov).
§Department of Mathematics, University of Calgary, 2500 University Drive NW, Calgary, AB T2N 1N4, Canada (lpbos@math.ucalgary.ca).

domain Ω,

$$\int_\Omega g(\xi)\, d\xi \simeq \sum_{j=1}^{N} w_j g(z_j),$$

is exact for all polynomials g up to degree d. Among all quadrature formulas of strength d, the optimal formulas are those with the fewest possible points N. In this work, we describe an algorithm for computing near-optimal quadrature formulas which is motivated by the Fekete point algorithm and relies heavily on the construction of cardinal functions.

The quadrature problem has been extensively studied independently of spectral element applications and has a long history of both theoretical and numerical development. For a recent review, see [4, 20, 7, 5]. An on-line database containing many of the best-known quadrature formulas is described in [6]. Many of those results were republished in the book [25], where they are available on the included CD-ROM.

One successful approach for numerically finding quadrature formulas dates to [21]. A generalized version was used recently in [31]. Newton's method is used to solve the nonlinear system of algebraic equations for the quadrature weights and locations of the points. Symmetry is used to reduce the complexity of the problem. If the quadrature points are invariant under the action of a group G, then the number of equations can be reduced to the dimension of the subspace of $\mathcal{P}_d$ invariant under G.

Motivated by the cardinal function Fekete point algorithm [29], we propose a new method to reduce the complexity of the quadrature problem: we look for quadrature formulas that have the same number of points as the dimension of a lower dimensional polynomial space. We can then construct a cardinal function basis for this lower dimensional space, make use of the interpolatory quadrature formulas, and derive a remarkable expression analytically relating the variation in the quadrature weights to the variation of the quadrature points. The net result is a significant reduction in the number of equations and unknowns. Symmetry can still be used to further reduce the complexity of the problem if needed. We then apply this algorithm to the triangle, where we are able find optimal quadrature sets of strength 9 through 25, subject only to the cardinal function constraint without imposing any symmetry constraints on the solutions.

2. Orthogonal polynomials. We first define our notation and describe the basis that will be used to represent cardinal functions. Let Ω be a domain in $\Re^n$, with ξ an arbitrary point in Ω. Let $\mathcal{P}_d$ be the finite dimensional vector space of polynomials in the Cartesian components of ξ of at most degree d, and let $N = \dim \mathcal{P}_d$. As an example, if Ω is the right triangle, then

$$\mathcal{P}_d = \operatorname{span}\{\xi_1^n \xi_2^m, m + n \le d\},$$

where $\xi = (\xi_1, \xi_2) \in \Re^2$ and $N = \frac{1}{2}(d+1)(d+2)$.

Our algorithm requires working with a cardinal function basis for $\mathcal{P}_d$. The most practical way to compute cardinal functions numerically is to work instead with their expansions in terms of an orthogonal, easily computed basis for $\mathcal{P}_d$. We denote this basis by $\{g_i(\xi), i = 1, \ldots, N\}$. For simplicity, we require that $g_1(\xi) = 1$. Then since the remaining basis functions are orthogonal to the constant function, we have

$$\int_\Omega g_i\, d\xi = \begin{cases} |\Omega| & \text{for } i = 1, \\ 0 & \text{for } i > 1, \end{cases} \tag{2.1}$$

where $|\Omega|$ is the area of Ω and $d\xi$ represents the uniform area measure.

For this work, all of our numerical calculations were performed in the right triangle. In the triangle there are several suitable choices of orthogonal basis functions. We use the Proriol polynomials $\{g_{m,n}\}$ [22, 18, 8]. The indexes m and n specify the top degree in each coordinate. Here we convert this traditional double index (m, n) into a single index by $i = (m+n+1)(m+n+2)/2-m$, so that $\mathcal{P}_d = \text{span}\{g_i, i = 1, \ldots, N\}$. Recurrence relations to evaluate these polynomials and their derivatives are given in [24].

3. Cardinal functions and quadrature points. We now describe the procedure we use to compute cardinal functions defined by a set of N points

$$\mathbf{z} = \{z_1, z_2, \ldots, z_N\},$$

where each z_i is a point in Ω. If the points are nondegenerate, the cardinal functions can be defined uniquely as the polynomials in ξ which belong to $\mathcal{P}_d$ and satisfy

$$\phi_i(\xi; \mathbf{z}) = \begin{cases} 1 & \text{if } \xi = z_i, \\ 0 & \text{if } \xi = z_j,\ j \neq i. \end{cases} \tag{3.1}$$

The cardinal functions depend implicitly on the defining points, and thus we include a second argument of $\mathbf{z}$. To evaluate cardinal functions numerically, we first express them in terms of the orthogonal basis $\{g_m\}$,

$$\phi_i(\xi; \mathbf{z}) = \sum_m \hat{\phi}_i^m g_m(\xi). \tag{3.2}$$

The expansion coefficients $\hat{\phi}_i^m$ are computed by evaluating (3.2) at the points z_j and solving the $N \times N$ linear system

$$\phi_i(z_j; \mathbf{z}) = \sum_m \hat{\phi}_i^m g_m(z_j).$$

If reasonable care is used, the points $\{z_j\}$ can be chosen so that the system is well conditioned and easily inverted by Gaussian elimination. The resulting cardinal functions form a basis for $\mathcal{P}_d$.

Following conventional spectral method techniques, we evaluate cardinal functions at an arbitrary point ξ by simply evaluating the $g_m(\xi)$ (via recurrence relations) and summing the series in (3.2). Derivatives of cardinal functions with respect to ξ are evaluated in a similar fashion, by first evaluating the derivatives of g_m after differentiating (3.2).

The interpolatory quadrature formula for $\mathcal{P}_d$ can be constructed at these points by solving the system

$$\sum_{j=1}^{N} w_j \phi_i(z_j; \mathbf{z}) = \int_\Omega \phi_i \, d\xi \qquad \forall \phi_i,\ i = 1, \ldots, N.$$

Making use of (2.1) and (3.1), the solution of this system is given by

$$w_i = \int_\Omega \phi_i \, d\xi = |\Omega| \hat{\phi}_i^1. \tag{3.3}$$

By construction, the interpolatory quadrature weights and the points $\{z_i\}$ give a quadrature formula which exactly integrates our N cardinal functions $\{\phi_i\}$. Since $\mathcal{P}_d = \text{span}\{\phi_i\}$, we have

$$\sum_{j=1}^{N} w_j g(z_j) = \int_\Omega g \, d\xi \qquad \forall g \in \mathcal{P}_d.$$

Thus any set of N nondegenerate points $\{z_j\}$ will yield a quadrature formula for $\mathcal{P}_d$, with uniquely determined quadrature weights. The problem now is to find the N points which integrate all of $\mathcal{P}_{d+e}$ for the largest possible e.

4. Derivatives of cardinal functions with respect to z. In the algorithm that follows, we will also need to compute the derivative of a cardinal function ϕ_i with respect to the points z_j used to define ϕ_i. For this we use the following result (which was also derived independently in [23]).

THEOREM 4.1. *For cardinal functions defined by* (3.1)*, we have the vector valued equation*

$$\frac{\partial \phi_i}{\partial z_j}(\xi; \mathbf{z}) = -\phi_j(\xi; \mathbf{z}) \frac{\partial \phi_i}{\partial \xi}(z_j; \mathbf{z}) \tag{4.1}$$

by which we mean

$$\frac{\partial}{\partial (z_i)_k} \phi_i(\xi; \mathbf{z}) = -\phi_i(\xi; \mathbf{z}) \frac{\partial \phi_i}{\partial \xi_k}(z_i; \mathbf{z}), \qquad k = 1 \ldots n,$$

where $(z_i)_k$ is the kth coordinate of the point $z_i \in \Re^n$, and ξ_k is the kth coordinate of the point $\xi \in \Re^n$.

Proof. Consider the derivative with respect to $(z_i)_k$,

$$\frac{\partial}{\partial (z_j)_k} \phi_i(\xi; \mathbf{z}) = \lim_{h \to 0} \frac{\phi_i(\xi; (\mathbf{z} \backslash z_j) \cup \{z_j + h e_k\}) - \phi_i(\xi; \mathbf{z})}{h},$$

where e_k denotes the standard unit direction vector in the kth coordinate. But the difference

$$\phi_i(\xi; (\mathbf{z} \backslash z_j) \cup \{z_j + h e_k\}) - \phi_i(\xi; \mathbf{z})$$

is zero at the points of $\mathbf{z} \backslash z_j$, as is $\phi_j(\xi; \mathbf{z})$, and hence by uniqueness,

$$\phi_i(\xi; (\mathbf{z} \backslash z_j) \cup \{z_j + h e_k\}) - \phi_i(\xi; \mathbf{z}) = C \phi_j(\xi; \mathbf{z})$$

for some constant C.

To evaluate C, first suppose that $j \neq i$. Then evaluate at $\xi = z_j + h e_k$ to obtain (for sufficiently small h)

$$0 - \phi_i(z_j + h e_k; \mathbf{z}) = C \phi_j(z_j + h e_k; \mathbf{z})$$

so that

$$C = -\frac{\phi_i(z_j + h e_k; \mathbf{z})}{\phi_j(z_j + h e_k; \mathbf{z})}.$$

Hence,

$$\begin{aligned}\frac{\partial}{\partial(z_j)_k}\phi_i(\xi;\mathbf{z}) &= \lim_{h\to 0} -\frac{\phi_i(z_j+he_k;\mathbf{z})}{h\phi_j(z_j+he_k;\mathbf{z})}\phi_j(\xi;\mathbf{z}) \\ &= \frac{\phi_j(\xi;\mathbf{z})}{\phi_j(z_j;\mathbf{z})}\lim_{h\to 0}\frac{0-\phi_i(z_j+he_k;\mathbf{z})}{h} \\ &= \frac{\phi_j(\xi;\mathbf{z})}{\phi_j(z_j;\mathbf{z})}\lim_{h\to 0}\frac{\phi_i(z_j;\mathbf{z})-\phi_i(z_j+he_k;\mathbf{z})}{h} \\ &= -\frac{\phi_j(\xi;\mathbf{z})}{1}\frac{\partial\phi_i}{\partial\xi_k}(z_j;\mathbf{z}) \\ &= -\phi_j(\xi;\mathbf{z})\frac{\partial\phi_i}{\partial\xi_k}(z_j;\mathbf{z}).\end{aligned}$$

Similarly, if $j=i$, we evaluate at $\xi = a_i + he_k$ to obtain

$$\begin{aligned}C &= \frac{1-\phi_i(z_i+he_k;\mathbf{z})}{\phi_i(z_i+he_k;\mathbf{z})} \\ &= \frac{\phi_i(z_i;\mathbf{z})-\phi_i(z_i+he_k;\mathbf{z})}{\phi_i(z_i+he_k;\mathbf{z})}\end{aligned}$$

so that

$$\frac{\partial}{\partial(z_i)_k}\phi_i(\xi;\mathbf{z}) = -\phi_i(\xi;\mathbf{z})\frac{\partial\phi_i}{\partial\xi_k}(z_i;\mathbf{z}). \qquad \square$$

Thus the derivative of the ith cardinal function with respect to the jth quadrature point is given by the jth cardinal function times a term independent of ξ and involving only the conventional derivative. The later term can be easily evaluated by differentiating (3.2).

Using (3.3) and (4.1), we can also derive a similar relation showing that the derivative of the ith weight with respect to the jth quadrature point is given by the jth weight times the same term that appears in (4.1):

$$\frac{\partial w_i}{\partial z_j} = -\int_\Omega \phi_j(\xi;\mathbf{z})\frac{\partial\phi_i}{\partial\xi}(z_j;\mathbf{z})\,d\xi = -w_j\frac{\partial\phi_i}{\partial\xi}(z_j;\mathbf{z}). \tag{4.2}$$

5. A cardinal function algorithm for computing quadrature points for $\mathcal{P}_{d+e}$. We now describe an iterative method for improving an initial set of N quadrature points $\mathbf{z}$. In order to integrate a space larger than $\mathcal{P}_d$, we need to find quadrature points which satisfy the nonlinear equation

$$\sum_i w_i g_m(z_i) = \int_\Omega g_m\,d\xi \qquad \forall g_m \in \mathcal{P}_{d+e} \tag{5.1}$$

for some $e > 0$. By using the interpolatory quadrature weights given by (3.3), we automatically integrate all of $\mathcal{P}_d$; thus we need only satisfy the equations for the basis functions in $\mathcal{P}_{d+e}$ which are not in $\mathcal{P}_d$:

$$\sum_i w_i g_m(z_i) = 0 \qquad \forall g_m : d < \text{degree}\, g_m \le d+e$$

and we have replaced the integral on the left-hand side of (5.1) by 0 by virtue of (2.1). Define

$$F_m = \sum_i w_i g_m(z_i)$$

and $F = \{F_m : d < \text{degree}\, g_m \le d + e\}$. Note that since the weights are determined by $\mathbf{z}$, we can treat F as a function solely of $\mathbf{z}$. Then (5.1) is equivalent to $F = 0$, which can be solved using Newton's method:

$$\frac{\partial \mathbf{z}}{\partial t} = -\left(\nabla F\right)^{-1} F.$$

The gradient of F is not necessarily square. In order to invert ∇F we restrict ourselves to the underdetermined case and use the pseudoinverse for $(\nabla F)^{-1}$. If ∇F is of full rank, many solutions exist and this approach gives the minimum norm solution, while in the rank-deficient case it will give the minimum norm least-squares solution [12]. To ensure that the system is underdetermined, we chose e so that there are fewer equations than degrees of freedom in the problem. There is one degree of freedom for each coordinate of each point in $\Re^n$, for a total of $n \dim \mathcal{P}_d$. The number of equations is given by $\dim \mathcal{P}_{d+e} - \dim \mathcal{P}_d$. Thus the degrees-of-freedom constraint is given by $\dim \mathcal{P}_{d+e} \le (n+1) \dim \mathcal{P}_d = (n+1)N$.

Using (4.2), the components of the gradient of F are given by

$$\begin{aligned} \frac{\partial F_m}{\partial z_j} &= \sum_i \left(w_i \frac{\partial g_m(z_i)}{\partial z_j} + \frac{\partial w_i}{\partial z_j} g_m(z_i) \right) \\ &= w_j \frac{\partial g_m}{\partial \xi}(z_j) - w_j \sum_i \frac{\partial \phi_i}{\partial \xi}(\mathbf{z}, z_j) g_m(z_i). \end{aligned} \tag{5.2}$$

Comparing this approach to the traditional Newton method for quadrature, such as in [31], we see that the use of the interpolatory quadrature weights has removed the weights from the iteration and thus reduced the number of unknowns by N. Since these weights exactly integrate $\mathcal{P}_d$, we have also reduced the number of equations by N. The only increase in complexity is the addition of the term involving the derivative of the weights with respect to the quadrature points. But this term is easy to evaluate by virtue of (4.2).

If symmetry is imposed on the quadrature points, then additional reductions in the number of equations and unknowns are possible, as in [21]. However, for the results presented in this paper we typically do not impose any symmetry constraints.

6. Practical considerations. In practice, Newton's method to solve $F = 0$ is used only to accelerate the convergence of a slower, more robust algorithm. We first use the steepest descent algorithm to minimize $\sum_m F_m^2$. This algorithm simply moves the points in the direction of steepest descent given by $\nabla(\sum_m F_m^2)$:

$$\frac{\partial z_j}{\partial t} = -2 \sum_m F_m \frac{\partial F_m}{\partial z_j}.$$

Once this algorithm has found a possible quadrature formula, we switch to Newton's method and iterate until the sequence converges. If the iteration fails to converge, then another initial condition is chosen and the procedure is repeated.

When the steepest descent algorithm is used in such a large (N) dimensional space, it can be expected to find only local minimums, most of which will not represent solutions of $F = 0$. This makes the success of the algorithm highly dependent on the initial condition. Here we will present numerical results for the triangle, where theory offers some guidance as to how to choose initial conditions. Our procedure is the same as that used in [29]. We choose a distribution of points so that their density approximates the extremal measure μ from pluripotential theory [16]. The extremal measure has been connected to the distribution of both quadrature points and Fekete points. For the right triangle $\xi_1 \geq 0$, $\xi_2 \geq 0$, and $\xi_1 + \xi_2 \leq 1$ the extermal measure is given in [1] as

$$\mu(\xi) = \frac{1}{\sqrt{\xi_1 \xi_2 (1 - \xi_1 - \xi_2)}}.$$

It is conjectured that μ is the density of quadrature points with positive weights as the limit N goes to infinity, and it was recently shown that this limit is bounded below by $c\mu$ for some constant c [19]. The same conjecture has also been made for Fekete points [26], where it is known that their density, in the limit as N goes to infinity, is bounded above by $c\mu$ [27].

To distribute a finite set of points to approximate $\mu(\xi)$, we first assume the points lie in a nested family of triangles. We then compute a nested family of triangular shells, each with an area (using the measure $\mu(\xi)d\xi$) proportional to the number of points we have decided to place in that shell. If there are k points to be placed in a given shell, we break that shell into k quadrilateral pieces, all with the same area, and place one point in the center of each piece. For a given number of points, there are a variety of configurations which can be generated by altering the number of points within each shell and the number of shells. The cardinal function algorithm is extremely sensitive to the initial condition, so many of these initial conditions must be tried to find an optimal solution.

7. Results. Our results for the triangle are summarized in Table 1. Except for quadrature formulas associated with $d = 3$ and $d = 4$, we were able to obtain the

TABLE 1

Quadrature points computed with the cardinal function algorithm. In all cases, the quadrature weights are positive and the points are not outside the triangle. Solutions which are not D_3 symmetric are denoted by asym. *Solutions which improve upon previously published results are denoted by* new.

Degree of cardinal functions (d)	Number of points (N)	Degree of exact integration (d+e)	Error	Notes
1	3	2	4.4×10^{-16}	
2	6	4	9.7×10^{-16}	
3	10	5	1.7×10^{-14}	
4	15	7	2.1×10^{-14}	
5	21	9	2.8×10^{-14}	
6	28	11	4.7×10^{-15}	asym
7	36	13	2.2×10^{-14}	asym,new
8	45	14	1.8×10^{-15}	
9	55	16	8.6×10^{-15}	asym,new
10	66	18	3.3×10^{-14}	asym,new
11	78	20	2.8×10^{-14}	asym,new
12	91	21	2.9×10^{-14}	new
13	105	23	3.3×10^{-14}	new
14	120	25	4.3×10^{-14}	asym,new

optimal solution (fewest number of points) subject to the cardinal function constraint on the number of points and the degrees-of-freedom constraint:

$$N = \dim \mathcal{P}_d, \tag{7.1}$$

$$\dim \mathcal{P}_{d+e} \leq 3N. \tag{7.2}$$

All the quadrature points have positive weights and no points lie outside the triangle, although neither of these properties is in any way guaranteed by the cardinal function algorithm. Thus the solutions for $d > 4$ are the best that this algorithm could attain. The truly optimal quadrature points (those with the fewest points for a given strength) would have to be found with a more sophisticated algorithm.

The errors presented in Table 1 are the max norm of the quadrature error over all the orthonormal basis functions:

$$\max_{g_{m,n} \in \mathcal{P}_{d+e}} \left| \sum_i w_i g_{m,n}(z_i) - \int_\Omega g_{m,n} \, d\xi \right|$$

with normalization $\int_\Omega g_{m,n}^2 \, d\xi = |\Omega|$. Many of the quadrature sets are invariant under the symmetry group of rotations and reflections of the triangle, D_3. The solutions which do not have this symmetry are denoted with *asym* in the table.

Quadrature formulas denoted by *new* in the table represent formulas which improve upon the best previously published results, as taken from the extensive database described in [6] and the quadrature points presented in [31] (which were included in the database as of this writing). The new solutions for integration degree $d + e$ from 18 to 25 have fewer points than the previously published results. For $d + e = 13$ and 16, the results presented here have the fewest points among formulas with positive weights and no points outside the triangle. For $d + e = 13$, the previous result with the fewest quadrature points has $N = 36$, but some of those points are outside the triangle and not all the weights are positive [2]. For $d + e = 16$, previous results include formulas with 52 points, some of which are outside the triangle [9], and 55 points, some of which have negative weights [13].

In Table 2, we summarize the results for the triangle from [6], [31], and the cardinal function algorithm.

The coordinates of the points for the first four *new* quadrature formulas are given in Appendix A. The coordinates for all the formulas in the table are available electronically from [28]. Plots for the first four of these quadrature points are shown in Figure 1. In the figure, the right triangle has been mapped linearly to the equilateral triangle in order to make the asymmetry in the points more visible.

8. Summary. We have presented a cardinal function algorithm for computing multivariate quadrature points. The key ideas involve the use of the interpolatory quadrature weights expressed as integrals of cardinal functions and a formula relating the derivatives of cardinal functions with respect to z_i (their defining points) to conventional derivatives in ξ. These two ideas allow us to reduce the number of equations and number of unknowns by N, while still retaining analytic expressions for the gradients necessary to apply steepest decent or Newton iterations. The algorithm was applied to the triangle, where optimal (in the sense of (7.1) and (7.2)) formulas were constructed for integrating polynomials up to degree 25. Seven of these quadrature formulas improve on previously known results. Of these new formulas, 5 out of 7 are asymmetric. It remains an open question whether symmetric formulas can be found

TABLE 2

Comparisons of known quadrature formula for the triangle. The column labeled Cools *gives the results collected in* [6]. *The column labeled* Wandzura and Xiao *gives the results from* [31]. *Formulas with negative weights or points that lie outside the triangle are denoted with a* †.

Degree of exact integration	Number of points (N)		
	Cools	Wandzura and Xiao	Cardinal function algorithm
2	3	3	3
3	4	6	
4	6	6	6
5	7	7	10
6	10†	12	
7	12	15	15
8	15†	16	
9	19	19	21
10	22†	25	
11	27†	28	28
12	33	36	
13	36†	40	36
14	42	46	45
15	48†	54	
16	52†	58	55
17	61	66	
18	67†	73	66
19	73	82	
20	79†	85	78
21		93	91
22		100	
23		106	105
24		118	
25		126	120
26		138	
27		145	
28		154	
29		166	
30		175	

of equal strength and, if not, what is the minimum amount of asymmetry required to obtain a given strength.

By construction, the number of points required for each quadrature formula is given by $N = \dim \mathcal{P}_d$ for some d, meaning they can also be used for interpolation in $\mathcal{P}_d$. This is a useful property for many finite element methods. These discretizations commonly result in equations for functions which are assumed (within each triangular element) to be in the space $\mathcal{P}_d$.

To apply the algorithm to other domains and more than two dimensions requires only the knowledge of an orthogonal basis of polynomials and the ability to evaluate the basis functions at arbitrary points. The use of cardinal functions requires that $N = \dim \mathcal{P}_d$. This constraint can be relaxed by replacing $\mathcal{P}_d$ with any subspace $\mathcal{P}' \subset \mathcal{P}_{d+e}$. The algorithm is unmodified other than that one needs to compute cardinal functions and interpolatory weights for the space $\mathcal{P}'$ instead of $\mathcal{P}_d$.

Appendix A. Tables of quadrature points. We now list the coordinates of the first four quadrature formulas marked with *new* in Table 1. Coordinates for all formulas are available electronically in [28]. For each line, we give the first two barycentric coordinates of each point (equivalent to the x and y coordinates after an equilateral triangle is linearly mapped to the unit right triangle $x \geq 0$, $y \geq 0$,

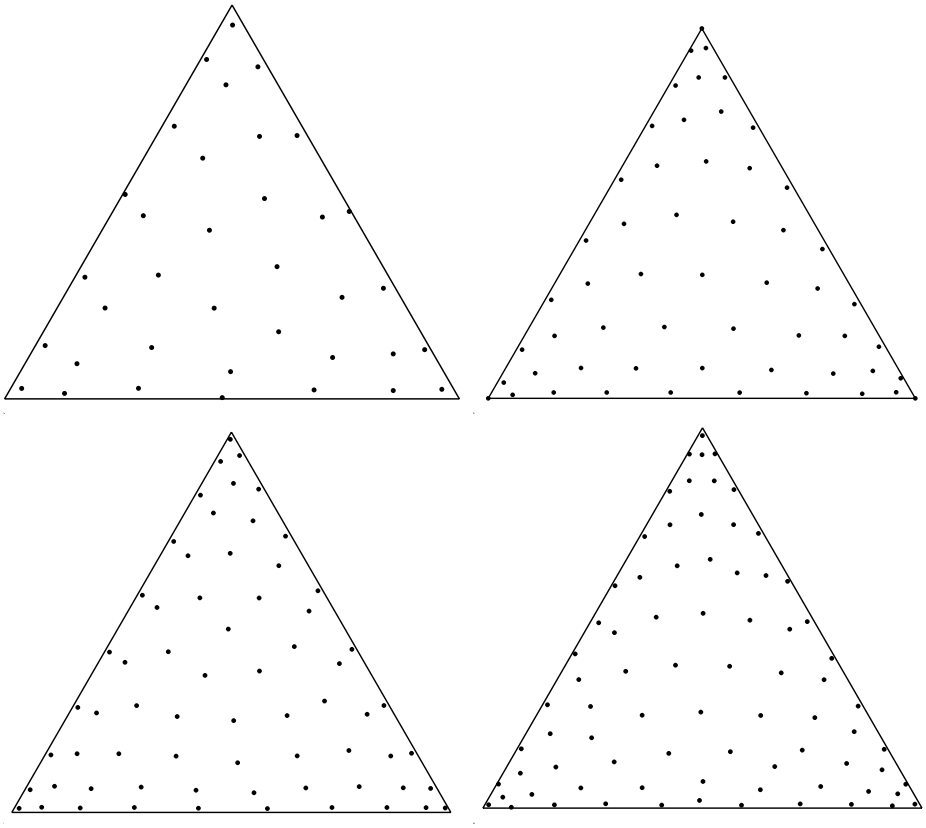


FIG. 1. *Quadrature points for the triangle which, from left to right and top to bottom, exactly integrate polynomials of degree* 13, 16, 18, *and* 20. *No points are outside the triangle, and all quadrature weights are positive.*

and $x + y \leq 1$) followed by the associated quadrature weight. The third barycentric coordinate is defined such that the sum of all three coordinates is one.

Tables of points

```
integration degree=13 N=36:
0.0242935351590 0.9493059293846 0.0166240998757
0.0265193427722 0.0242695130640 0.0166811699778
0.9492126023551 0.0265067966437 0.0166830569067
0.0033775763749 0.4767316412363 0.0175680870083
0.4757672298101 0.5198921829102 0.0184474661845
0.5190783193471 0.0055912706202 0.0197942410188
0.8616839745321 0.0133996048618 0.0203540395855
0.1249209759926 0.8613054321334 0.0206852863940
0.0138565453861 0.1247733717358 0.0208271366086
0.0211887064222 0.8438438351223 0.0317819778279
0.8432296787219 0.1354563645830 0.0320472035241
0.1354231797865 0.0213482820656 0.0320607681146
0.3088853510679 0.0221919663014 0.0430765959183
0.6685057595169 0.3089012879389 0.0438473415339
0.0226545012557 0.6691709943321 0.0439209672733
0.2808515408772 0.6924718155106 0.0479951923691
0.6922446749051 0.0268723345026 0.0483806260733
0.0268617447119 0.2810093973222 0.0484867423375
0.1141778485470 0.7973581413586 0.0556964488024
0.7974807922061 0.0879806508791 0.0561026364356
0.0892807293894 0.1145020561128 0.0565190123693
0.1052487892455 0.6686904119922 0.0689289890670
0.6663022280740 0.2275051631832 0.0717213336089
0.2307803737547 0.1054572561221 0.0727453920976
0.1705059157540 0.5174064398658 0.0788807336737
0.5086593973043 0.3170523855209 0.0810114345512
0.3141823862281 0.1810706361659 0.0825725299055
0.4617460817864 0.4678594539804 0.0842044567330
0.0693087496081 0.4622856042085 0.0843585533305
0.4651955259268 0.0724357805669 0.0851969868488
0.2578625857893 0.6131395039177 0.0902845328052
0.6112627766779 0.1300360834609 0.0914283143485
0.1305182135934 0.2581713828884 0.0916279065409
0.4281437991828 0.2362005969817 0.1025573374896
0.3356995783730 0.4311026308588 0.1033159661413
0.2305424298836 0.3456013949376 0.1035854367193
integration degree=16 N=55:
1.0000000000000 0.0000000000000 0.0006202599851
0.0000000000000 1.0000000000000 0.0006315174712
0.0000000000000 0.0000000000000 0.0007086601559
0.9398863583577 0.0049848744634 0.0055163716168
0.0543806683058 0.9386405618617 0.0062692407656
0.0093940049164 0.0526424462697 0.0078531408826
0.0164345086362 0.9469035517351 0.0094551483864
```

```
0.9469487269862 0.0363373677167 0.0097824511271
0.0426604005768 0.0151224541799 0.0099861643489
0.0122269495439 0.8693773510664 0.0137553818816
0.8673696521047 0.1204917285774 0.0140979178040
0.8456744021389 0.0157763967870 0.0149646864337
0.1395759632103 0.8448120870375 0.0156097503612
0.1317821743231 0.0135009605584 0.0157683693348
0.0157955126300 0.1455274938536 0.0175794546383
0.7365462884436 0.0155697540908 0.0204113840270
0.0139688430330 0.7379836894450 0.0209562878616
0.2547895186039 0.7297615689771 0.0210713412998
0.7316386522555 0.2543076683315 0.0217646760202
0.0157253728951 0.2696239795791 0.0222288408699
0.2662302843647 0.0144783956308 0.0224186693682
0.8673504065214 0.0591679410400 0.0230122616993
0.0741493666957 0.8634782575061 0.0236813902500
0.0159285948360 0.4191238955238 0.0257464643368
0.0156061028068 0.5809222921146 0.0257956801608
0.5910094817484 0.0159251452651 0.0258072327610
0.4034771496889 0.5806700368104 0.0260343232059
0.5694745628526 0.4149495146302 0.0265768141609
0.0678493700650 0.0761218678591 0.0265784761831
0.4265968590272 0.0157509692312 0.0267532329238
0.0670982507890 0.7741898312421 0.0375787806641
0.7528310231480 0.0819119495639 0.0383065894195
0.7753727783557 0.1577128457292 0.0384849695025
0.1689073157787 0.7503943099742 0.0389619825852
0.1687335832919 0.0708311507268 0.0394604111547
0.0821244708436 0.1762996626771 0.0412364778098
0.6288705363345 0.0807744953317 0.0512872438483
0.0811413015266 0.3054373589776 0.0516405641935
0.2969112065080 0.6227485988871 0.0518230042269
0.0767542314171 0.6247247149546 0.0528527988181
0.6223022333845 0.3011485821166 0.0538505573027
0.3103786288051 0.0779098365079 0.0541895329319
0.0819218215187 0.4603633038351 0.0584737146444
0.4717022665013 0.0821554006797 0.0592863168363
0.4546603415250 0.4637565033890 0.0594358276749
0.1701091339237 0.6422277808188 0.0631800255863
0.6406004329487 0.1898293537256 0.0632926845153
0.1912267583717 0.1739955685343 0.0640707361772
0.1885315767070 0.4798914070406 0.0812040595918
0.4772929957691 0.3348356598119 0.0814437513530
0.3126974621760 0.4957972197259 0.0814679201241
0.4961225945946 0.1927553668904 0.0815050548084
0.1928805312867 0.3161015807261 0.0815164664939
0.3360041453816 0.1894892801290 0.0816931059623
0.3337280550848 0.3343571021811 0.0923218334531
integration degree=18 N=66:
0.0116731059668 0.9812565951289 0.0025165756986
0.9810030858388 0.0071462504863 0.0025273452007
0.0106966317092 0.0115153933376 0.0033269295333
0.9382476983551 0.0495570591341 0.0081503492125
0.0126627518417 0.9370123620615 0.0086135525742
0.0598109409984 0.0121364578922 0.0087786746179
0.0137363297927 0.0612783625597 0.0097099585562
0.9229527959405 0.0141128270602 0.0102466211915
0.0633107354993 0.9220197291727 0.0108397688341
0.0117265100335 0.1500520475229 0.0129385390176
0.1554720587323 0.8325147121589 0.0136339823583
0.8343293888982 0.0125228158759 0.0138477328147
0.8501638031957 0.1371997508736 0.0139421540105
0.0128816350522 0.8477627063479 0.0144121399968
0.1510801608959 0.0136526924039 0.0153703455534
0.0101917879217 0.5770438618345 0.0162489802253
0.2813372399303 0.7066853759623 0.0169718304280
0.7124374628501 0.0124569780990 0.0170088532421
0.2763025250863 0.0121741311386 0.0170953520675
0.0109658368561 0.4194306712466 0.0173888854559
0.4289110517884 0.5599616067469 0.0174543962439
0.4215420555115 0.0116475994785 0.0178406757287
0.5711258590444 0.0118218313989 0.0178446863879
0.5826868270511 0.4057889581177 0.0179046337552
0.0130567806713 0.2725023750868 0.0181259756201
0.0130760400964 0.7224712523233 0.0184784838882
0.7263437062407 0.2602984019251 0.0185793564371
0.0687230068637 0.0631417277210 0.0203217151777
0.8652302101529 0.0720611837338 0.0213771661809
0.0648599071037 0.8590433543910 0.0231916854098
0.1483494943362 0.7888788352240 0.0274426710859
0.0624359898396 0.1493935499354 0.0290301922340
0.7871369011735 0.0656382042757 0.0294522738505
0.0519104921610 0.5255635695605 0.0299436251629
0.1543129927444 0.0716383926917 0.0307026948119
0.2617842745603 0.0621479485288 0.0325263365863
0.7667257872813 0.1658211554831 0.0327884208506
0.2582103676627 0.6800119766139 0.0331234675192
0.0679065925147 0.7571515437782 0.0346167526875
0.5293578274804 0.4121503841107 0.0347081373976
0.0666036150484 0.2612513087886 0.0347372049404
0.0585675461899 0.3902236114535 0.0348528762454
0.0644535360411 0.6373626559761 0.0348601561186
0.6748138429151 0.0637583342061 0.0355471569975
0.3914602310369 0.5503238090563 0.0360182996383
0.6487701492307 0.2836728360263 0.0362926285843
0.3946498220408 0.0605175522554 0.0381897702083
0.5390137151933 0.0611990176936 0.0392252800118
0.1627895082785 0.6861322141035 0.0482710125888
0.6812436322641 0.1567968345899 0.0489912121566
0.1542832878020 0.1667512624020 0.0497220833872
0.2522727750445 0.2504803933395 0.0507065736986
0.2547981532407 0.4994090649043 0.0509771994043
0.1485580549194 0.5756023096087 0.0521360063667
0.2930239606436 0.5656897354162 0.0523460874925
0.2808991272310 0.1437921574248 0.0524440683552
0.4820989592971 0.2518557535865 0.0527459644823
0.5641878245444 0.1462966743153 0.0529449063728
0.1307699644344 0.4489577586117 0.0542395594501
0.1479692221948 0.3001174386829 0.0543470203419
0.5638684222946 0.2813772089298 0.0547100548639
0.4361157428790 0.4252053446420 0.0557288345913
0.3603263935285 0.2599190004889 0.0577734264233
0.4224188334674 0.1453238443303 0.0585393781623
0.3719001833052 0.3780122703567 0.0609039250680
0.2413645006928 0.3847563284940 0.0637273964449
integration degree=20 N=78:
0.0089411337112 0.0086983293702 0.0021744545399
0.9792622629807 0.0102644133744 0.0028987135265
0.0105475382112 0.9785514202515 0.0030846029337
0.0023777061947 0.0636551098604 0.0034401633104
0.0630425115795 0.0041506347509 0.0041898472012
0.9308422496730 0.0048053482263 0.0044738051498
0.0629076555490 0.9316790069481 0.0047054420814
0.9315962246381 0.0626264881801 0.0048867935750
0.0061951689415 0.9293587058564 0.0051927643369
0.0287125819237 0.0310202122997 0.0074073058981
0.9293844478305 0.0342152968219 0.0079755410301
0.0375457566621 0.9257868884669 0.0083550522910
0.0086895739064 0.1584971251510 0.0096166660864
0.1547597053965 0.8363606657688 0.0096318257850
0.8331025294185 0.0089257244824 0.0098577460758
0.8374231073526 0.1529167304078 0.0102657880301
0.1559362505234 0.0094966240058 0.0103188103111
0.0098599642095 0.8342211493596 0.0106291001630
0.4055873733289 0.0074389302008 0.0106881306895
0.5964727898618 0.3956330809311 0.0106969021010
0.0080747800416 0.4031319425903 0.0109026461714
0.0075073977721 0.5851609594681 0.0109899783575
0.3936764519237 0.5974896592899 0.0113423055229
0.5846530726212 0.0087250464968 0.0120535642930
0.4870804112120 0.0202129229912 0.0139619193821
0.2683512811785 0.7202340088668 0.0141147991536
0.7223956288748 0.2662399366456 0.0141930347046
0.2716826742357 0.0112882698808 0.0144212676268
0.0112580842046 0.7169695963325 0.0144704346855
0.0115034734370 0.2740067110166 0.0144949769872
0.7140525900564 0.0113511560497 0.0145386775694
0.4902871053112 0.4936491841468 0.0145964190926
```

```
0.0201423425209 0.4832573459601 0.0147314578466    0.0538297481158 0.3358616826849 0.0350393454927
0.0361107464859 0.0935679501582 0.0167463963304    0.1848840324117 0.1551831523851 0.0350717420310
0.8607998819851 0.0397379067075 0.0168955500458    0.3376267104744 0.6081402596294 0.0352129215334
0.1005891526001 0.8586343419352 0.0169422662884    0.6067102034499 0.0542632795598 0.0352615504981
0.0918740717058 0.0395513001973 0.0173070172095    0.4612614085496 0.0688176670722 0.0366403220343
0.8604888296191 0.0966224057079 0.0174524546493    0.1525465365671 0.6510240845749 0.0367733107670
0.0439842178673 0.8561886349107 0.0177217222159    0.0700582543543 0.4661904392742 0.0371675662937
0.2011017606735 0.7449115835626 0.0282824024023    0.4704201379032 0.4634826455353 0.0373371571606
0.7449993726263 0.0536865638166 0.0284996712488    0.1216461693746 0.2381494875516 0.0403973346588
0.0532186641310 0.1963754275935 0.0285005646539    0.6371404052702 0.1238399384513 0.0413580040638
0.7453984647401 0.1982065805550 0.0300647223478    0.2379904515119 0.6370216452326 0.0421957791870
0.1957289932876 0.0555713833156 0.0302031277082    0.1483929857177 0.4894188577780 0.0495451004037
0.1092532057988 0.6100036182413 0.0303987136077    0.3598069571550 0.1452880866253 0.0500419261141
0.0567625702001 0.7409121894959 0.0305668796074    0.4941441055095 0.3610216383818 0.0505794587115
0.0483837933475 0.6075135660978 0.0306067413002    0.1440630687981 0.3513508341887 0.0520037210188
0.1080612809760 0.1122081510437 0.0309330068201    0.5019764440004 0.1435491663293 0.0521533567886
0.6185605900991 0.2698753703035 0.0309773820835    0.3555423834298 0.5016491599502 0.0524899152358
0.7721296013497 0.1114117395333 0.0313146250545    0.2443439540771 0.2406052129104 0.0599159762516
0.6115734801133 0.3389367677931 0.0313573493392    0.2437064989342 0.5109017277055 0.0599609997426
0.3381326103376 0.0494693938787 0.0314320469287    0.5122200807321 0.2452737973543 0.0599915272129
0.1173084128254 0.7696451309795 0.0315182143894    0.2526038315178 0.3700319555094 0.0634133183449
0.2674551260596 0.1115718808154 0.0324248137985    0.3759895652851 0.2505406611631 0.0635311861108
0.6542100160026 0.1906548314700 0.0347512152386    0.3729077987144 0.3753750277549 0.0637206605672
```

Acknowledgments. We thank R. Cools for many useful comments on drafts of this manuscript and D. M. Gay for helpful discussions on optimization algorithms. We would also like to thank the anonymous reviewers for improvements due to their careful reading and helpful suggestions.

REFERENCES

[1] M. Baran, *Complex equilibrium measure and Bernstein type theorems for compact sets in* $\mathbf{r}^n$, Proc. Amer. Math. Soc., 123 (1994), pp. 485–494.

[2] J. Berntsen and T. Espelid, *Degree* 13 *Symmetric Quadrature Rules for the Triangle*, Technical report, Reports in Informatics 44, Department of Informatics, University of Bergen, Bergen, Norway, 1990.

[3] Q. Chen and I. Babuška, *Approximate optimal points for polynomial interpolation of real functions in an interval and in a triangle*, Comput. Methods Appl. Mech. Engrg., 128 (1995), pp. 405–417.

[4] R. Cools, *Constructing Cubature Formulae: The science behind the art*, in Acta Numer. 6, Cambridge University Press, Cambridge, UK, 1997, pp. 1–54.

[5] R. Cools, *Monomial cubature rules since "Stroud": A compilaton.* II, J. Comput. and Appl. Math., 112 (1999), pp. 21–27.

[6] R. Cools, *An encyclopaedia of cubature formulas*, J. Complexity, 19 (2003), pp. 445–453; also available online from http://www.cs.kuleuven.be/~nines/ecf.

[7] R. Cools and P. Rabinowitz, *Monomial cubature rules since "Stroud": A compilaton*, J. Comput. Appl. Math., 48 (1993), pp. 309–326.

[8] M. Dubiner, *Spectral methods on triangles and other domains*, J. Sci. Comput., 6 (1991), pp. 345–390.

[9] D. Dunavant, *High degree efficient symmetrical Gaussian quadrature rules for the triangle*, Internat. J. Numer. Methods Engrg., 21 (1985), pp. 1129–1148.

[10] A. Fournier, M. Taylor, and J. Tribbia, *The spectral element atmosphere model (SEAM): High-resolution parallel computation and localized resolution of regional dynamics*, Mon. Wea. Rev., 132 (2004), pp. 726–748.

[11] F. X. Giraldo and T. E. Rosmond, *A scalable spectral element Eulerian atmospheric model (SEE-AM) for NWP: Dynamical core tests*, Mon. Wea. Rev., 132 (2004), pp. 133–153.

[12] G. H. Golub and C. F. V. Loan, *Matrix Computations*, 2nd ed., The Johns Hopkins University Press, Baltimore, MD, 1989.

[13] S. Heo and Y. Xu, *Constructing symmetric cubature formulae on a triangle*, in Advances in Computational Mathematics, Z. Chen et al., ed., Dekker, New York, 1999, pp. 203–221.

[14] J. S. Hesthaven, *From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex*, SIAM J. Numer. Anal., 35 (1998), pp. 655–676.

[15] M. Iskandarani, D. Haidvogel, and J. Levin, *A three-dimensional spectral element model for the solution of the hydrostatic primitive equations*, J. Comput. Phys., 186 (2003).
[16] M. Klimek, *Pluripotential Theory*, Lond. Math. Soc. Monogr. (NS) 6, The Clarendon Press, Oxford University Press, New York, 1991.
[17] D. Komatitsch and J. Tromp, *Spectral-element simulations of global seismic wave propagation—*I *Validation*, Geophys. J. Int., 149 (2002), pp. 390–412.
[18] T. Koornwinder, *Two-variable analogues of the classical orthogonal polynomials*, in Theory and Applications of Special Functions, R. A. Askey, ed., Academic Press, New York, 1975, pp. 435–495.
[19] G. Kuperberg, *Numerical cubature from Archimedes' hat-box theorem*, 2004; available online from http://www.arxiv.org/abs/math.NA/0405366.
[20] J. Lyness and R. Cools, *A survey of numerical cubature over triangles*, Appl. Math., 48 (1994), pp. 127–150.
[21] J. Lyness and D. Jespersen, *Moderate degree symmetric quadrature rules for the triangle*, J. Inst. Math. Appl., 15 (1975), pp. 19–32.
[22] J. Proriol, *Sur une famille de polynomes à deux variables orthogonaux dans un triangle*, C. R. Acad. Sci. Paris, 245 (1957), pp. 2459–2461.
[23] M. J. Roth, *Nodal Configurations and Voronoi Tessellations for Triangular Spectral Elements*, Ph.D. thesis, University of Victoria, Victoria, BC, Canada, 2005.
[24] S. Sherwin and G. Karniadakis, *A triangular spectral element method; applications to the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 123 (1995), pp. 189–229.
[25] P. Šolín, K. Segeth, and I. Doležel, *Higher-Order Finite Element Methods*, Chapman & Hall/CRC Press, Boca Raton, FL, 2004.
[26] C. C. T. Bloom, L. Bos, and N. Levenberg, *Polynomial interpolation of holomorphic functions in* $\mathbf{c}$ *and* $\mathbf{c}^n$, Rocky Mountain J. Math., 22 (1992), pp. 441–470.
[27] B. A. Taylor, *private communication*, 1998.
[28] M. A. Taylor, B. A. Wingate, and L. P. Bos, *Several new quadrature formulas for polynomial integration in the triangle*, 2005; available online from http://www.arxiv.org/abs/math.NA/0501496.
[29] M. A. Taylor, B. A. Wingate, and R. Vincent, *An algorithm for computing Fekete points in the triangle*, SIAM J. Numer. Anal., 38 (2000), pp. 1707–1720.
[30] S. Thomas and R. Loft, *The NCAR spectral element climate dynamical core: Semi-implicit Eulerian formulation*, J. Sci. Comput., 25 (2005), pp. 307–322.
[31] S. Wandzura and H. Xiao, *Symmetric quadrature rules on a triangle*, Comput. Math. Appl., 45 (2003), pp. 1829–1840.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 206–222

BPX-TYPE PRECONDITIONERS FOR SECOND AND FOURTH ORDER ELLIPTIC PROBLEMS ON THE SPHERE*

JAN MAES†, ANGELA KUNOTH‡, AND ADHEMAR BULTHEEL†

Abstract. We develop two Bramble–Pasciak–Xu-type preconditioners for second (resp., fourth) order elliptic problems on the surface of the two-sphere. To discretize the second order problem we construct C^0 linear elements on the sphere, and for the fourth order problem we construct C^1 finite elements of Powell–Sabin type on the sphere. The main reason these BPX preconditioners work depends on this particular choice of basis. We prove optimality and provide numerical examples. Furthermore we numerically compare the BPX preconditioners with the suboptimal hierarchical basis preconditioners.

Key words. BPX preconditioner, C^0 and C^1 finite elements, elliptic equations on surfaces

AMS subject classifications. 65F10, 65F35, 65N30, 35J20, 35J35

DOI. 10.1137/050647414

1. Introduction.

The aim of the present paper is the development of two Bramble–Pasciak–Xu (BPX) [7] preconditioners for second (resp., fourth) order elliptic problems on the two-dimensional sphere. Such problems arise from several applications in physical geodesy, oceanography, and meteorology [8], and they are even of interest for the graphics community, since surface meshes are often parameterized by using so-called harmonic weights, which correspond to a finite element discretization of the Laplace–Beltrami operator; see, e.g., [1] and references therein.

The geometry of the sphere is a major obstacle in constructing suitable approximation spaces for solving partial differential equations. Often a transformation into spherical coordinates is used which gives rise to singularities at the "poles" of the sphere. This complication is induced by the spherical coordinate system itself. Therefore, an important point in our method is the use of homogeneous polynomials in $\mathbb{R}^3$ that allows us to stick with Cartesian coordinates; hence the "pole problem" is avoided. In order to develop the theory we shall restrict ourselves to the following two most simple equations:

$$-\Delta_S u = f \quad \text{on} \quad S, \tag{1.1}$$

and

$$\Delta_S^2 u = f \quad \text{on} \quad S, \tag{1.2}$$

where Δ_S is the Laplace–Beltrami operator on the two-sphere S. In order to work with Cartesian coordinates we write the Laplace–Beltrami operator in terms of the

*Received by the editors December 13, 2005; accepted for publication (in revised form) September 6, 2006; published electronically January 12, 2007. This work was partially supported by the Flemish Fund for Scientific Research (FWO Vlaanderen) projects MISS (G.0211.02) and SMID (G.0431.05), by the European Community's Human Potential Programme under contract HPRN–CT–2002–00286 (Breaking Complexity), and by the Belgian Programme on Interuniversity Attraction Poles, initiated by the Belgian Federal Science Policy Office. The scientific responsibility rests with the authors.
http://www.siam.org/journals/sinum/45-1/64741.html

†Department of Computer Science, Katholieke Universiteit Leuven, Celestijnenlaan 200A, B-3001 Heverlee, Belgium (jan.maes@cs.kuleuven.be, adhemar.bultheel@cs.kuleuven.be).

‡Institut für Angewandte Mathematik and Institut für Numerische Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany (kunoth@iam.uni-bonn.de).

tangential gradient

$$\nabla_S u := \nabla u - (n \cdot \nabla u)n,$$

with n the outward normal to S. The Laplace–Beltrami operator on S can now be defined as

$$\Delta_S := \nabla_S \cdot \nabla_S.$$

We use C^0 continuous piecewise linear spherical polynomials to discretize the variational problem

$$\int_S \nabla_S u \nabla_S v \; d\omega = \int_S f v \; d\omega \quad \text{for all } v \in H^1(S) \tag{1.3}$$

corresponding to (1.1), and C^1 continuous piecewise quadratic spherical polynomials to discretize the variational problem

$$\int_S \Delta_S u \Delta_S v \; d\omega = \int_S f v \; d\omega \quad \text{for all } v \in H^2(S) \tag{1.4}$$

corresponding to (1.2). For every $f \in L_2(S)$ with $\int_S f \; d\omega = 0$ there exist a weak solution $u \in H^1(S)$ of (1.3) and a weak solution $u \in H^2(S)$ of (1.4). In both cases u is unique up to a constant; see, e.g., [5, 16].

So let $m \in \{1, 2\}$, and suppose $V \subset H^m(S)$ is a space of conforming C^{m-1} finite elements defined on a spherical triangulation of S with mesh size h. Define $a(u, v)$ as the bilinear form induced by (1.3) (resp., (1.4)) given $m = 1$ (resp., $m = 2$), and let A denote the positive definite self-adjoint operator on V defined by

$$a(u, v) = (Au, v), \quad v \in V, \tag{1.5}$$

where $(\cdot, \cdot)$ denotes the inner product of $L_2(S)$. Then we have to solve the linear operator equation

$$Au = b \tag{1.6}$$

for some $u \in V$, where $b \in V$ is defined by $(b, v) = (f, v)$, $v \in V$. The conjugate gradient method is a very efficient solver for large linear systems arising from problems such as (1.6). However, because of stability reasons, it is necessary that these systems have been suitably preconditioned. It is a known fact (see, e.g., [12]) that if for some constants $0 < \gamma, \Gamma < \infty$ and some invertible operator C

$$\gamma(C^{-1}u, u) \le a(u, u) \le \Gamma(C^{-1}u, u), \quad u \in V, \tag{1.7}$$

then the spectral condition number $\kappa(C^{1/2}AC^{1/2})$ is bounded by Γ/γ.

Let us represent the operator A by the stiffness matrix $A_\Phi := (a(\phi_i, \phi_j))_{i,j \in I}$ with respect to some typical nodal basis $\Phi := \{\phi_i \; : \; i \in I\}$ of V. Then it is known that $\kappa(A_\Phi) = \mathcal{O}(h^{-2})$ for problem (1.3) and $\kappa(A_\Phi) = \mathcal{O}(h^{-4})$ for problem (1.4). In order to precondition the system

$$A_\Phi y = b_\Phi, \quad (b_\Phi)_i := (f, \phi_i), \; i \in I, \tag{1.8}$$

one can perform a change of basis. So let $\Psi = \{\psi_i \; : \; i \in I\}$ be another basis of V, and let L be the transfer matrix between the two bases. Then

$$A_\Psi = L^T A_\Phi L,$$

which suggests the use of $C = LL^T$ as preconditioner for the nodal basis discretization.

Several approaches exist to construct a suitable preconditioner, such as the hierarchical basis (HB) preconditioner [31] and the closely related BPX preconditioner [7]. The growth rate of the condition numbers was shown to be logarithmic in the size of the problem for the HB preconditioner [31] and uniformly bounded for the BPX preconditioner in [12, 27]. Originally, these results were formulated for second order problems on two-dimensional planar domains, but they could also be established for fourth order problems on the plane [13, 20, 26]. Recently, we constructed an HB preconditioner for fourth order elliptic problems on the surface of the sphere in [23]. The growth rate of the condition number was shown to be logarithmic which is, as expected, similar to the planar case. It is the aim of the present paper to prove optimality of a BPX preconditioner for problems (1.3) and (1.4), independent of the discretization, and to give numerical evidence of this optimality. We emphasize that the crucial steps in the optimality proof depend on the particular choice of basis and, thus, are not valid for arbitrary C^0 or C^1 finite element constructions on the sphere. For both problems we explicitly construct a suitable basis that is easy to implement.

The outline of the remaining sections is as follows. In section 2, we introduce the C^0 continuous piecewise linear and C^1 continuous piecewise quadratic spherical polynomials that will be used to discretize problem (1.3) (resp., (1.4)). The corresponding BPX preconditioners are constructed in section 3, and we prove their optimality. Finally, in section 4 we conclude with some numerical experiments that confirm the theory with small absolute condition and iteration numbers.

We finish this introduction with a note about notation. We always mean by $a \sim b$ that $a \lesssim b$ and $a \gtrsim b$ hold, where $a \lesssim b$ means that a can be bounded by a constant multiple of b uniformly in any parameters on which a,b may depend, and $a \gtrsim b$ means $b \lesssim a$.

2. Suitable elements on the sphere. In a series of papers [2, 3, 4], Alfeld, Neamtu, and Schumaker develop spline spaces on triangulations on the sphere analogous to the classical spline spaces on planar triangulations. The idea is to work with homogeneous Bernstein–Bézier polynomials in $\mathbb{R}^3$ which are then restricted to the sphere. A function f defined on $\mathbb{R}^3$ is *homogeneous of degree* d provided that $f(\alpha v) = \alpha^d f(v)$ for all real α and all $v \in \mathbb{R}^3$. The space $\mathbb{H}_d$ of *trivariate polynomials of degree* d *that are homogeneous of degree* d is a $\binom{d+2}{2}$ dimensional subspace of the space of trivariate polymials of degree d. Let $\{v_1, v_2, v_3\}$ be a set of linearly independent unit vectors in $\mathbb{R}^3$. We call

$$\mathcal{T} := \{v \in \mathbb{R}^3 \mid v = b_1(v)v_1 + b_2(v)v_2 + b_3(v)v_3 \quad \text{with} \quad b_i(v) \geq 0\}$$

the *trihedron* generated by $\{v_1, v_2, v_3\}$. Each $v \in \mathbb{R}^3$ can be written in the form

$$v = b_1(v)v_1 + b_2(v)v_2 + b_3(v)v_3, \tag{2.1}$$

and we call $b_1(v), b_2(v), b_3(v)$ the *trihedral coordinates of* v *with respect to* $\mathcal{T}$. Given an integer $d \geq 0$, the *homogeneous Bernstein basis polynomials of degree* d *on* $\mathcal{T}$ are the polynomials

$$B^d_{ijk}(v) := \frac{d!}{i!j!k!} b_1(v)^i b_2(v)^k b_3(v)^k, \qquad i + j + k = d,$$

and they form a basis for $\mathbb{H}_d$. We define a *spherical triangle* as the restriction of a trihedron $\mathcal{T}$ to the unit sphere S. The restrictions of the trihedral coordinates (2.1) to a spherical triangle with vertices v_1, v_2, and v_3 are called *spherical barycentric*

coordinates. Any homogeneous polynomial p of degree d and its restriction to a spherical triangle τ has a *Bernstein–Bézier representation* with respect to τ:

$$p(v) := \sum_{i+j+k=d} c_{ijk} B^d_{ijk}(v), \tag{2.2}$$

where the coefficients c_{ijk} are the *Bézier ordinates*.

Homogeneous polynomials in their Bernstein–Bézier representation can be evaluated efficiently using the classical de Casteljau algorithm:

$$p(v) = c^d_{000}(v),$$

where for $1 \le l \le d$

$$\begin{aligned} &c^0_{ijk}(v) := c_{ijk}, \\ &c^l_{ijk}(v) := b_1(v)c^{l-1}_{i+1,j,k} + b_2(v)c^{l-1}_{i,j+1,k} + b_3(v)c^{l-1}_{i,j,k+1}, \quad i+j+k = d-l. \end{aligned}$$

Also continuity conditions can be expressed analogously to the classical bivariate case. Let $\mathcal{T}$ and $\widetilde{\mathcal{T}}$ be trihedra with vertices $\{v_1, v_2, v_3\}$ and $\{v_4, v_2, v_3\}$. A necessary and sufficient condition for p and $\widetilde{p}$ to be C^r continuous across the common boundary is

$$\widetilde{c}_{ijk} = c^i_{0jk}(v_4), \quad i = 0, 1, \dots, r, \quad i+j+k = d. \tag{2.3}$$

We write $\mathbb{H}_d(\Omega)$ for the restriction of $\mathbb{H}_d$ to any subset Ω of the unit sphere S, and refer to $\mathbb{H}_d(\Omega)$ as the *space of spherical polynomials of degree d*. Similarly, we write $\mathbb{H}_d(H)$ for the restriction of $\mathbb{H}_d$ to any hyperplane H in $\mathbb{R}^3 \setminus \{0\}$. This is just the well-known space of bivariate polynomials. All these spaces have the same dimension $\binom{d+2}{2}$. Let Δ be a conforming spherical triangulation of $\Omega \subset S$. Then we define the *space of spherical splines of degree d and smoothness r associated with* Δ to be

$$S^r_d(\Delta) := \{s \in C^r(S) \ : \ s|_\tau \in \mathbb{H}_d(\tau), \ \tau \in \Delta\},$$

where $s|_\tau$ denotes the restriction of s to the spherical triangle τ.

2.1. C^0 linear elements on the sphere. The C^0 continuous piecewise linear spherical polynomials that we describe here are a natural extension of the well-known linear elements introduced by Courant [10]. However, our approach differs significantly from previous constructions (e.g., [6, 16]); see Remark 2.4. Suppose that we are given an initial triangulation Δ_0 of S and that

$$\Delta_0 \subset \Delta_1 \subset \cdots \subset \Delta_j \subset \cdots, \quad j = 0, 1, \dots,$$

is a sequence of dyadically refined triangulations obtained by subdividing the triangles at level j (i.e., the triangles of Δ_j) into 4 congruent subtriangles of level $j+1$. This refinement is regular; i.e. the minimum angle condition is satisfied and

$$\operatorname{diam} \tau \sim 2^{-j}, \quad \tau \in \Delta_j, \quad j = 0, 1, \dots.$$

For each $j = 0, 1, \dots$ we define $v_{i,j}$, $i = 1, \dots, N_j$, as the vertices of the triangulation Δ_j. We create suitable basis functions for the nested spherical spline spaces

$$S^0_1(\Delta_0) \subset S^0_1(\Delta_1) \subset \cdots \subset S^0_1(\Delta_j) \subset \cdots, \quad j = 0, 1, \dots,$$

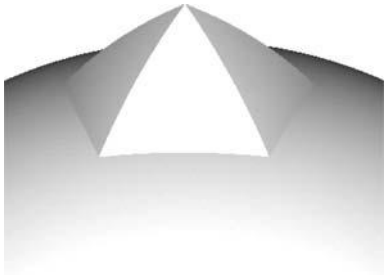


FIG. 2.1. *Graph of* $(\phi_{i,j}(v)+1)v$ *with* $v \in S$ *and* $\phi_{i,j}$ *the spherical Courant element.*

and this approach allows us to point out a strong connection with the classical Courant elements on the plane.

So let us define a nodal basis for $S_1^0(\Delta_j)$ by solving the following interpolation problem: find functions $\phi_{i,j} \in S_1^0(\Delta_j)$, $i = 1, \ldots, N_j$, such that $\phi_{i,j}(v_{k,j}) = \delta_{i,k}$. Obviously this interpolation problem has a unique solution. If we restrict the spline $\phi_{i,j}$ to any spherical triangle τ in Δ_j, we get a spherical Bernstein–Bézier polynomial (2.2) of degree $d = 1$. The interpolation problem determines the three Bézier ordinates c_{ijk} in (2.2) (with $d = 1$) in a unique way. If the spherical triangle τ does not contain vertex $v_{i,j}$, then the three Bézier ordinates equal zero, and hence $\phi_{i,j}$ has local support. If τ contains vertex $v_{i,j}$, then the Bézier ordinate that is associated with this vertex takes the value 1, and the other two Bézier ordinates take the value 0. It is easily checked that the continuity conditions (2.3) for $r = 0$ are satisfied. Figure 2.1 shows the spherical Courant element.

We can look at each spherical basis function $\phi_{i,j}$ as the restriction of a trivariate homogeneous function to the sphere S. In particular, let f be any spherical function and let $d \in \mathbb{N}$; then we define $(f)_d$ as its *homogeneous extension of degree* d, i.e.,

$$(f)_d(v) := |v|^d f\left(\frac{v}{|v|}\right), \quad v \in \mathbb{R}^3 \setminus \{0\}. \tag{2.4}$$

If we restrict the homogeneous extension of degree 1 of $\phi_{i,j}$ to the sphere S, we recover $\phi_{i,j}$, i.e., $\phi_{i,j} \equiv (\phi_{i,j})_1|_S$. Moreover, we even have the following theorem.

THEOREM 2.1. *The restriction of* $(\phi_{i,j})_1$ *to the tangent plane touching* S *at* $v_{i,j}$ *is a classical bivariate Courant element defined on this tangent plane centered around the vertex* $v_{i,j}$.

Proof. First we define the *radial projection* R_T from any plane T that is tangent to S onto S by

$$R_T\overline{v} := v := \frac{\overline{v}}{|\overline{v}|} \in S, \quad \overline{v} \in T, \tag{2.5}$$

where $|v|$ denotes the Euclidean norm of v. Let $T_{i,j}$ be the tangent plane touching S at vertex $v_{i,j} \in \Delta_j$. Because the mapping $R_{T_{i,j}}$ is one-to-one, the inverse $R_{T_{i,j}}^{-1}$ is well defined. Define $\Delta_{i,j}$ as the 1-ring of vertex $v_{i,j}$ in Δ_j. Let $\overline{\Delta}_{i,j}$ be the image of $\Delta_{i,j}$ under $R_{T_{i,j}}^{-1}$. Since great circles are mapped onto straight lines under $R_{T_{i,j}}^{-1}$, $\overline{\Delta}_{i,j}$ consists of planar neighboring triangles with one common vertex $v_{i,j}$. The spline space $S_1^0(\overline{\Delta}_{i,j})$ is just the well-known bivariate linear spline space on the triangulation $\overline{\Delta}_{i,j}$. Let τ be a spherical triangle in $\Delta_{i,j}$ and denote its vertices by v_1, v_2, v_3. Let

$\mathcal{T}$ be the trihedron generated by $\{v_1, v_2, v_3\}$. Then for some $a \in \{1, 2, 3\}$ we have $v_a = v_{i,j}$ (since $\tau \in \Delta_{i,j}$) and

$$(\phi_{i,j})_1(w) = b_a(w)^i, \qquad w \in \mathcal{T},$$

with (b_1, b_2, b_3) the trihedral coordinates of w with respect to $\mathcal{T}$. Consequently, $(\phi_{i,j})_1$ equals zero at the vertices of $\overline{\Delta}_{i,j}$, except that for vertex $v_{i,j} \in \overline{\Delta}_{i,j}$ we get $(\phi_{i,j})_1(v_{i,j}) = 1$. Since $\phi_{i,j} = (\phi_{i,j})_1|_S$, we have $\phi_{i,j}|_\tau \in \mathbb{H}_1(S)$. Let $\overline{\tau}$ be the image of τ under $R^{-1}_{T_{i,j}}$. We find that $(\phi_{i,j})_1|_{\overline{\tau}} \in \mathbb{H}_1(\overline{\tau})$ and thus $(\phi_{i,j})_1|_{T_{i,j}} \in S^0_1(\overline{\Delta}_{i,j})$. This proves that $(\phi_{i,j})_1|_{T_{i,j}}$ is just the well-known classical bivariate Courant element. ∎

This idea can be exploited to extend several properties of the classical Courant elements to the spherical elements $\phi_{i,j}$, such as

$$0 \le \phi_{i,j}(v) \le 1, \quad v \in S.$$

The following lemma is obvious.

LEMMA 2.2 (Riesz L_∞-stability). *The nodal basis functions* $\{\phi_{i,j} \mid i = 1, \dots, N_j\}$ *satisfy*

$$\left\| \sum_{i=1}^{N_j} c_{i,j}\phi_{i,j} \right\|_{L_\infty} \sim \max_i |c_{i,j}|.$$

Proof. There exist a triangle $\tau \in \Delta_j$ and a point $v \in \tau$ such that

$$\left\| \sum_{i=1}^{N_j} c_{i,j}\phi_{i,j} \right\|_{L_\infty} = \left| \sum_{i=1}^{N_j} c_{i,j}\phi_{i,j}(v) \right| \le \max_i |c_{i,j}| \sum_{i|v_{i,j}\in\tau} \|\phi_{i,j}\|_{L_\infty} \lesssim \max_i |c_{i,j}|.$$

The other inequality follows from $|c_{k,j}| = |\sum_{i=1}^{N_j} c_{i,j}\phi_{i,j}(v_{k,j})| \le \|\sum_{i=1}^{N_j} c_{i,j}\phi_{i,j}\|_{L_\infty}$. This completes the proof. ∎

To derive the optimality of the BPX preconditioner we will need the following theorem.

THEOREM 2.3 (Riesz L_p-stability). *For any* $1 < p < \infty$ *we have*

$$\left\| \sum_{i=1}^{N_j} c_{i,j}\phi_{i,j} \right\|^p_{L_p} \sim 2^{-2j} \sum_{i=1}^{N_j} |c_{i,j}|^p.$$

Proof. Since we have already established Riesz L_∞-stability of the basis (Lemma 2.2), the proof is identical to the corresponding proof for the classical Courant elements on the plane from [10]. ∎

Remark 2.4. There exist other constructions of C^0 spherical finite elements in the literature. In [16] problem (1.1) is discretized by approximating the sphere S by a polyhedron S_h. Then linear elements on the surface S_h are used. In [6] spherical linear elements are created, but another definition for spherical barycentric coordinates is used. In [6] the spherical barycentric coordinates are required to form a partition of unity, and therefore they inevitably fail to have many of the important properties that the spherical barycentric coordinates of [2] have. The optimality proof of the BPX preconditioner that we give in section 3 works only for our construction.

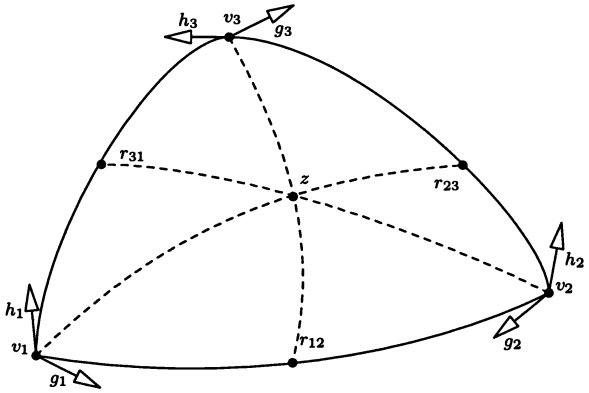


FIG. 2.2. *The spherical Powell–Sabin macroelement.*

2.2. C^1 Powell–Sabin elements on the sphere. In general, maintaining C^1 continuity conditions (2.3) between neighboring triangles results in nontrivial relations and is not always possible for arbitrary given triangulations; see, e.g., [17]. Therefore, to overcome this problem, we will focus on the *Powell–Sabin 6-split* of a triangulation. Starting from an arbitrary spherical triangulation Δ, we introduce further substructures by subdividing each triangle of Δ into 6 subtriangles in a prescribed way. Because of the special structure of this refined triangulation one introduces sufficient degrees of freedom to maintain overall C^1 continuity. The Powell–Sabin 6-split is obtained as follows:

1. Define for each triangle τ_k in Δ an interior point z_k such that if two triangles τ_k and τ_l have a common edge (circle segment), then the arc that joins z_k and z_l intersects this common edge (circle segment) at a point r_{kl} between its vertices. The arc between two points on S is defined as the circle segment connecting these two points obtained as the intersection of S with a plane passing through the two points and the origin.
2. Join the points z_k to the vertices of τ_k.
3. For each edge (circle segment) of τ_k
   - that belongs to the boundary $\partial\Omega$, join z_k to some point of the edge,
   - that is common to a triangle τ_l, join z_k to r_{kl}.

Figure 2.2 shows the split of one triangle. We will refer to this new triangulation as Δ^{PS}. The spline space $S_2^1(\Delta^{PS})$ of piecewise quadratic C^1 spherical polynomials over Δ^{PS} will be called the space of *spherical Powell–Sabin (PS) splines*. Let g_i and h_i be independent unit vectors lying in the tangent plane of S at the vertices v_i, $i = 1, \ldots, N$, of the triangulation Δ. The following interpolation problem can be considered for spherical PS splines. Given any set of values $(\alpha_i, \beta_i, \gamma_i)$, $i = 1, \ldots, N$, find $s(v) \in S_2^1(\Delta^{PS})$ such that

$$s(v_i) = \alpha_i, \quad \frac{\partial s(v_i)}{\partial g_i} = \beta_i, \quad \frac{\partial s(v_i)}{\partial h_i} = \gamma_i, \tag{2.6}$$

for all $i = 1, \ldots, N$. Maes and Bultheel [23] have shown that this interpolation problem has a unique solution, and hence the classical result of [28] can be extended to spherical domains; i.e., the dimension of the spherical spline space $S_2^1(\Delta^{PS})$ equals $3N$.

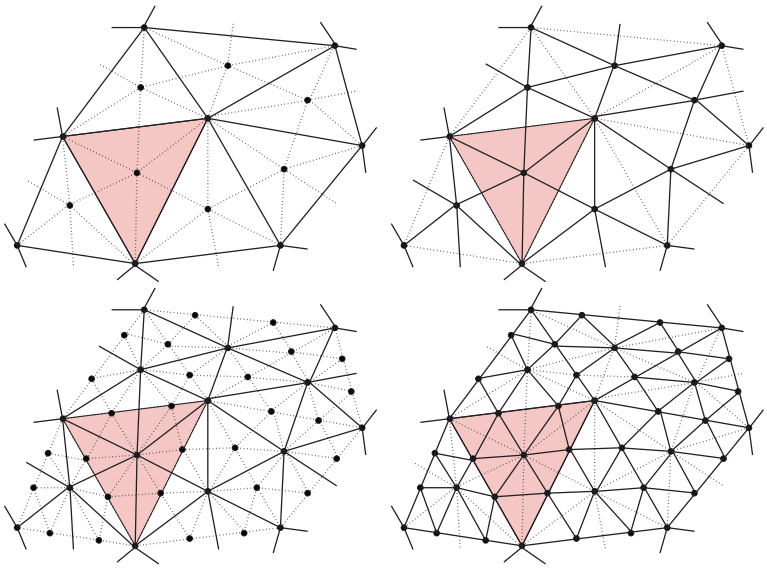


FIG. 2.3. *Principle of $\sqrt{3}$ subdivision. Applying the $\sqrt{3}$ subdivision twice results in triadic subdivision.*

In order to create nested spherical PS spline spaces

$$S_2^1(\Delta_0^{PS}) \subset S_2^1(\Delta_1^{PS}) \subset S_2^1(\Delta_2^{PS}) \subset \cdots$$

it is sufficient that we find a refinement procedure that yields nested sequences

$$\begin{aligned} &\Delta_0^{PS} \subset \Delta_1^{PS} \subset \Delta_2^{PS} \subset \cdots, \\ &\{v_i \in \Delta_0\} \subset \{v_i \in \Delta_1\} \subset \{v_i \in \Delta_2\} \subset \cdots. \end{aligned} \tag{2.7}$$

It was pointed out by Vanraes et al. [30] that applying a $\sqrt{3}$ *refinement scheme* yields nested PS spline spaces. Applying the $\sqrt{3}$ scheme twice yields a triadic scheme. The $\sqrt{3}$ scheme was first introduced by Kobbelt [19] and Labsik and Greiner [21]. Instead of splitting each edge in Δ_0 and performing a 1-to-4 split for each triangle (dyadic refinement), we compute a new vertex for each triangle and retriangulate the old and new vertices. Figure 2.3 shows the principle. Note that the new edges in Δ_1 coincide with the lines of the PS 6-split Δ_0^{PS}. In the new triangles new interior points must be chosen on the one line of the new PS 6-split Δ_1^{PS} that is already fixed, that is, the original edge that crosses the triangle.

Remark 2.5. Although the $\sqrt{3}$ refinement is applicable to arbitrary (spherical) triangulations, it is not rigorously proven whether the corresponding sequence (2.7) satisfies the minimum angle condition and whether

$$\operatorname{diam} \tau \sim \sqrt{3}^{-j}, \quad \tau \in \Delta_j^{PS}, \quad j = 0, 1, \ldots.$$

In our mathematical analysis we will always assume that we are given a nested sequence (2.7) that is regular. By using a PS 12-split as in [26] instead of the PS 6-split one can obtain a provably regularly refined sequence of PS spline spaces by applying dyadic refinement. However, we opt for the PS 6-split because the construction of the corresponding basis functions is less complicated, certainly on the sphere, and

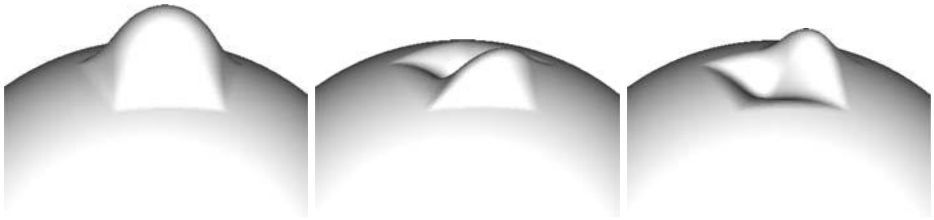


FIG. 2.4. *Graphs of* $(B^k_{i,j}(v)+1)v$ *with* $v \in S$ *for* $k = 1, 2, 3$ *with* $B^k_{i,j}$ *the spherical Hermite PS basis function.*

because the $\sqrt{3}$ refinement is a topologically slower refinement than the dyadic refinement; thus we have more levels of resolution if a prescribed target complexity of the PS spline space must not be exceeded.

With each vertex $v_{i,j} \in \Delta_j$ we associate two directions $g_{i,j}$ and $h_{i,j}$ such that the set $(v_{i,j}, g_{i,j}, h_{i,j})$ forms an orthonormal basis for $\mathbb{R}^3$. For instance, suppose that $v_{i,j}$ has spherical coordinates $(\cos\theta \sin\phi, \sin\theta \sin\phi, \cos\phi)$, $\theta \in [0, 2\pi]$, $\phi \in [0, \pi]$; then take $g_{i,j} = (\cos\theta\cos\phi, \sin\theta\cos\phi, -\sin\phi)$ and $h_{i,j} = (-\sin\theta, \cos\theta, 0)$. Let us introduce the functionals

$$\lambda^1_{i,j}(f) := f(v_{i,j}), \quad \lambda^2_{i,j}(f) := \frac{\partial f(v_{i,j})}{\partial g_{i,j}}, \quad \lambda^3_{i,j}(f) := \frac{\partial f(v_{i,j})}{\partial h_{i,j}}, \quad f \in C^1(S).$$

Then we construct a nodal basis for $S^1_2(\Delta^{PS}_j)$ by solving the following interpolation problems of the form (2.6): find functions $B^k_{i,j} \in S^1_2(\Delta^{PS}_j)$, $k = 1, 2, 3$, $i = 1, \ldots, N_j$, such that

$$\begin{aligned} \lambda^1_{m,j}(B^k_{i,j}) &= \sqrt{3}^{-j}\delta_{k,1}\delta_{i,m}, \\ \lambda^2_{m,j}(B^k_{i,j}) &= \delta_{k,2}\delta_{i,m}, \\ \lambda^3_{m,j}(B^k_{i,j}) &= \delta_{k,3}\delta_{i,m} \end{aligned} \tag{2.8}$$

for all $m = 1, \ldots, N_j$. Note that these basis functions satisfy $B^k_{i,j} \equiv (B^k_{i,j})_2|_S$; i.e. the spherical basis function $B^k_{i,j}$ is equal to the restriction of its homogeneous extension (2.4) of degree 2 to the sphere S. If we restrict $(B^k_{i,j})_2$ to the tangent plane touching S at $v_{i,j}$, we get the corresponding planar Hermite basis function of [29] defined on this tangent plane. The proof is similar to the proof of Theorem 2.1. A detailed proof in a more general setting can be found in [23, Theorem 4.1]. Figure 2.4 shows the three spherical Hermite PS basis functions that are associated with one vertex.

We now show some stability properties of the nodal basis (2.8) that will be useful in the optimality proof of the BPX preconditioner.

LEMMA 2.6 (Riesz L_∞-stability). *The nodal basis defined by* (2.8) *satisfies*

$$\left\| \sum_{i=1}^{N_j} \sum_{k=1}^{3} c^k_{i,j} B^k_{i,j} \right\|_{L_\infty} \sim \sqrt{3}^{-j} \max_{i,k} |c^k_{i,j}|.$$

Proof. First we note that this result is well known for the classical bivariate Hermite basis of PS type on planar triangulations. Indeed, the inequality $\gtrsim$ can be shown using the Markov inequality for polynomials [9], and the inequality $\lesssim$ can be deduced, for instance, from the work in [29, section 6.2]. This result for the bivariate

planar setting can be extended easily to the spherical setting by exploiting the fact that the restriction of $(B^k_{i,j})_2$ to the tangent plane touching S at $v_{i,j}$ is a classical bivariate Hermite basis function. For a detailed proof, see [23, Corollary 4.2]. ∎

THEOREM 2.7 (Riesz L_p-stability). *If s is in $S^1_2(\Delta^{PS}_j)$, then for any $1 < p < \infty$ we have*

$$\|s\|^p_{L_p} \sim \sqrt{3}^{-2j} \left(\sum_{i=1}^{N_j} |\lambda^1_{i,j}(s)|^p + \sqrt{3}^{-jp} \sum_{i=1}^{N_j} \sum_{k=2}^{3} |\lambda^k_{i,j}(s)|^p \right).$$

Proof. Using the Markov inequality for spherical polynomials [25, Prop. 4.3], we infer that $|\lambda^k_{i,j}(s)| \lesssim \sqrt{3}^j \|s\|_{L_\infty(\tau_i)}$ for $k = 2, 3$ with $\tau_i \in \Delta^{PS}_j$ such that $v_{i,j} \in \tau_i$. By mapping τ_i to a standard reference triangle and using the fact that all norms on the finite-dimensional space of polynomials are equivalent, we find that $\|s\|_{L_\infty(\tau_i)} \lesssim \sqrt{3}^{2j/p} \|s\|_{L_p(\tau_i)}$, which implies

$$\sqrt{3}^{-2j} \left(\sum_{i=1}^{N_j} |\lambda^1_{i,j}(s)|^p + \sqrt{3}^{-jp} \sum_{i=1}^{N_j} \sum_{k=2}^{3} |\lambda^k_{i,j}(s)|^p \right) \lesssim \sum_{i=1}^{N_j} \sum_{k=1}^{3} \|s\|^p_{L_p(\tau_i)} \lesssim \|s\|^p_{L_p}.$$

The other inequality follows from the observation that

$$\begin{aligned} |s(v)|^p &= \left| \sum_{i=1}^{N_j} \left(\sqrt{3}^j \lambda^1_{i,j}(s) B^1_{i,j}(v) + \sum_{k=2}^{3} \lambda^k_{i,j}(s) B^k_{i,j}(v) \right) \right|^p \\ &\lesssim \sum_{i=1}^{N_j} \left(\sqrt{3}^{jp} |\lambda^1_{i,j}(s)|^p |B^1_{i,j}(v)|^p + \sum_{k=2}^{3} |\lambda^k_{i,j}(s)|^p |B^k_{i,j}(v)|^p \right), \end{aligned}$$

which holds because at any $v \in S$ there are at most nine nonzero basis functions. We find that

$$\begin{aligned} \|s\|^p_{L_p} &\lesssim \sum_{i=1}^{N_j} \left(\sqrt{3}^{jp} |\lambda^1_{i,j}(s)|^p \int_S |B^1_{i,j}(v)|^p dv + \sum_{k=2}^{3} |\lambda^k_{i,j}(s)|^p \int_S |B^k_{i,j}(v)|^p dv \right) \\ &\lesssim \sqrt{3}^{-2j} \left(\sum_{i=1}^{N_j} |\lambda^1_{i,j}(s)|^p + \sqrt{3}^{-jp} \sum_{i=1}^{N_j} \sum_{k=2}^{3} |\lambda^k_{i,j}(s)|^p \right), \end{aligned}$$

where we have used that $\|B^k_{i,j}\|_{L_\infty} \lesssim \sqrt{3}^{-j}$, which follows from the Riesz L_∞-stability of the basis (Lemma 2.6). ∎

3. Construction of the BPX preconditioners. In this section we construct BPX preconditioners for problems (1.3) and (1.4) and prove that these preconditioners are optimal. Let $m \in \{1, 2\}$, and let $\rho \in \mathbb{R}$ be a scaling factor. If m equals 1, we define S_j as the spline space $S^0_1(\Delta_j)$, we set the scaling factor ρ equal to 2, and we solve problem (1.3). If m equals 2, we define S_j as the spline space $S^1_2(\Delta^{PS}_j)$ and we set the scaling factor ρ equal to $\sqrt{3}$, leading to problem (1.4).

Let Q_j, $j = 0, 1, \ldots$, be a sequence of projectors on S_j which are orthogonal with respect to the inner product $(\cdot, \cdot)$, and let $Q_{-1} \equiv 0$. Let Ω be a subset of the sphere

S, and let $H^m(\Omega)$, $H^m(S)$ be the spherical Sobolev spaces as defined in [22, 25]. We prove the following theorem.

THEOREM 3.1. *Suppose $s \in S_J$. Then*

$$\|s\|^2_{H^m(S)} \sim \sum_{j=0}^{J} \rho^{2mj} \|(Q_j - Q_{j-1})s\|^2_{L_2(S)} . \tag{3.1}$$

Proof. Let Ω be a subset of S such that $\operatorname{diam}(\Omega) \leq 1$. Let T_Ω be the tangent plane touching S at r_Ω, with r_Ω the center of a spherical cap of smallest possible radius containing Ω. Here a spherical cap is defined as the region of a sphere which lies on one side of a given plane that intersects with the sphere. Recall the definition of the radial projection R_{T_Ω} from (2.5). Let $\overline{\Omega}$ be the image of Ω under $R^{-1}_{T_\Omega}$ and define $H^m(\overline{\Omega})$ as the usual Sobolev space on domains in $\mathbb{R}^2$. Let $(s)_m$ be the homogeneous extension (2.4) of degree m of s, and define $\overline{s}$ as the restriction of $(s)_m$ to $\overline{\Omega}$. The norm equivalence $\|s\|_{H^m(\Omega)} \sim \|\overline{s}\|_{H^m(\overline{\Omega})}$ holds; see Lemma 3.2 in [25]. Furthermore, we also have $\|s\|_{L_2(\Omega)} \sim \|\overline{s}\|_{L_2(\overline{\Omega})}$; see Lemma 3.1 in [25]. Now let $s = \sum_{j=0}^{J} s_j$ with each $s_j \in S_j$. Then it follows from the theory of homogeneous polynomials [2, 3] that $(s)_m = \sum_{j=0}^{J} (s_j)_m$; hence $\overline{s} = \sum_{j=0}^{J} \overline{s}_j$ with $\overline{s}_j$ the restriction of $(s_j)_m$ to $\overline{\Omega}$. Furthermore, each $\overline{s}_j$ is a member of the planar spline space $\overline{S}_j$ which is defined as $S^0_1(R^{-1}_{T_\Omega}(\Delta_j|_\Omega))$ for $m = 1$ and as $S^1_2(R^{-1}_{T_\Omega}(\Delta^{PS}_j|_\Omega))$ for $m = 2$. Proposition 2 in [26] claims that

$$\|\overline{s}\|^2_{H^m(\overline{\Omega})} \sim \inf \sum_{j=0}^{J} \rho^{2mj} \|\overline{s}_j\|^2_{L_2(\overline{\Omega})} ,$$

where the infimum must be taken with respect to all admissible representations $\sum_{j=0}^{J} \overline{s}_j$ of $\overline{s}$. From the norm equivalences above, we get

$$\|s\|^2_{H^m(\Omega)} \sim \inf \sum_{j=0}^{J} \rho^{2mj} \|s_j\|^2_{L_2(\Omega)} . \tag{3.2}$$

Now consider a finite collection of domains Ω_k with $\operatorname{diam}(\Omega_k) \leq 1$, covering S. Equation (3.2) is valid for each subdomain Ω_k. Furthermore, we have the equivalences $\|s_j\|^2_{L_2(S)} \sim \sum_k \|s_j\|^2_{L_2(\Omega_k)}$ and $\|s\|^2_{H^m(S)} \sim \sum_k \|s\|^2_{H^m(\Omega_k)}$. Hence,

$$\|s\|^2_{H^m(S)} \sim \inf \sum_{j=0}^{J} \rho^{2mj} \|s_j\|^2_{L_2(S)} ,$$

which immediately implies (3.1); see [15, 26]. ∎

Remark 3.2. Proposition 2 in [26] is formulated in terms of C^1 finite elements, but it is clear that a similar result holds for C^0 finite elements (with obvious modifications).

In view of (1.7) let us define the self-adjoint positive definite operator C_J^{-1} on S_J by

$$(C_J^{-1}u, v) = \sum_{j=0}^{J} \rho^{2mj}((Q_j - Q_{j-1})u, (Q_j - Q_{j-1})v), \tag{3.3}$$

and let A_J be the operator defined by (1.5) for $V = S_J$. By Poincaré's inequality on S, we have

$$a(u,u) \sim \|u\|^2_{H^m(S)} \tag{3.4}$$

under the constraint $\int_S u \, d\omega = 0$. Then Theorem 3.1 and (1.7) imply that

$$\kappa(C_J^{1/2} A_J C_J^{1/2}) = \mathcal{O}(1) \tag{3.5}$$

under the constraint that we fix the solution u of (1.3), (1.4) such that $\int_S u \, d\omega = 0$.

We now replace C_J by a spectrally equivalent and computationally simpler preconditioner $\widehat{C}_J$ given by

$$\widehat{C}_J := \sum_{j=0}^{J} \sum_{i=1}^{N_j} (\cdot, \phi_{i,j}) \phi_{i,j} \tag{3.6}$$

for problem (1.3) and by

$$\widehat{C}_J := \sum_{j=0}^{J} \sum_{i=1}^{N_j} \sum_{k=1}^{3} (\cdot, B_{i,j}^k) B_{i,j}^k \tag{3.7}$$

for problem (1.4). We say that two operators A and B are *spectrally equivalent* if

$$\frac{(Av,v)}{(v,v)} \sim \frac{(Bv,v)}{(v,v)}.$$

Note that, by (3.3), we have

$$C_J^{-1} := \sum_{j=0}^{J} \rho^{2mj} (Q_j - Q_{j-1}). \tag{3.8}$$

Indeed, by the L_2-orthogonality of the operators Q_j we find

$$\begin{aligned}
(C_J^{-1} u, v) &= \left(\sum_{j=0}^{J} \rho^{2mj} (Q_j - Q_{j-1}) u, v \right) \\
&= \left(\sum_{j=0}^{J} \rho^{2mj} (Q_j - Q_{j-1}) u, \sum_{j=0}^{J} (Q_j - Q_{j-1}) v \right) \\
&= \sum_{j=0}^{J} \rho^{2mj} ((Q_j - Q_{j-1}) u, (Q_j - Q_{j-1}) v).
\end{aligned}$$

Let us focus on the biharmonic problem (1.4); i.e., take $m = 2$. By the orthogonality of the projectors Q_j, one finds from (3.8) that

$$C_J = \sum_{j=0}^{J} \sqrt{3}^{-4j} (Q_j - Q_{j-1}).$$

Because of the decaying scaling factors we are allowed to replace C_J by the spectrally equivalent operator

$$\widetilde{C}_J := \sum_{j=0}^{J} \sqrt{3}^{-4j} Q_j.$$

From Theorem 2.7, we have the Riesz L_2-stability

$$\left\| \sum_{i=1}^{N_j} \sum_{k=1}^{3} c_{i,j}^k B_{i,j}^k \right\|_{L_2}^2 \sim \sqrt{3}^{-4j} \sum_{i=1}^{N_j} \sum_{k=1}^{3} |c_{i,j}^k|^2, \tag{3.9}$$

and by the Riesz representation theorem this implies the existence of a dual or biorthogonal basis $\{\widetilde{B}_{i,j}^k\}$ such that

$$\left\| \sum_{i=1}^{N_j} \sum_{k=1}^{3} c_{i,j}^k \widetilde{B}_{i,j}^k \right\|_{L_2}^2 \sim \sqrt{3}^{4j} \sum_{i=1}^{N_j} \sum_{k=1}^{3} |c_{i,j}^k|^2. \tag{3.10}$$

The orthogonal projector Q_j has the representation

$$Q_j f = \sum_{i=1}^{N_j} \sum_{k=1}^{3} (f, \widetilde{B}_{i,j}^k) B_{i,j}^k.$$

Hence,

$$(Q_j f, f) = (Q_j f, Q_j f) = \|Q_j f\|_{L_2}^2 \sim \sqrt{3}^{4j} \sum_{i=1}^{N_j} \sum_{k=1}^{3} |(f, B_{i,j}^k)|^2 = (\widehat{Q}_j f, f),$$

with

$$\widehat{Q}_j := \sqrt{3}^{4j} \sum_{i=1}^{N_j} \sum_{k=1}^{3} (\cdot, B_{i,j}^k) B_{i,j}^k,$$

which shows that $\widehat{C}_J$ defined in (3.7) is spectrally equivalent to C_J such that, by (3.5),

$$\kappa(\widehat{C}_J^{1/2} A_J \widehat{C}_J^{1/2}) = \mathcal{O}(1).$$

The optimality of the preconditioner defined in (3.6) for problem (1.3) can be derived analogously using Theorem 2.3. We have thus proved the main result of this paper.

THEOREM 3.3. *The BPX preconditioners given by*

$$\sum_{j=0}^{J} \sum_{i=1}^{N_j} (\cdot, \phi_{i,j}) \phi_{i,j} \quad \text{and} \quad \sum_{j=0}^{J} \sum_{i=1}^{N_j} \sum_{k=1}^{3} (\cdot, B_{i,j}^k) B_{i,j}^k$$

yield uniformly bounded condition numbers for problem (1.3) *(resp.,* (1.4)).

COROLLARY 3.4. *Any basis of the general form in* [23] *which is stable in the sense of Theorem* 2.7 *gives rise to an optimal BPX preconditioner for* (1.4).

Remark 3.5. In the paper [18] Griebel shows that the conjugate gradient method for the semidefinite system that arises from the Galerkin scheme using the nodal basis

functions of the finest level and of all coarser levels of discretization is equivalent to the BPX-preconditioned conjugate gradient method for the linear system that arises from the Galerkin scheme using only the nodal basis functions of the finest level. In our numerical experiments we take the approach of the semidefinite system using the nodal basis functions at all resolution levels. For details on efficient implementation we refer to [18].

Remark 3.6. The above derivation depends heavily on the use of a biorthogonal basis. Its existence is guaranteed by the Riesz representation theorem. Note the weight change from $\sqrt{3}^{-4j}$ in (3.9) for the finite element basis to $\sqrt{3}^{4j}$ in (3.10) for the biorthogonal basis. For properties of Riesz bases in connection with biorthogonality and multiresolution we refer the reader to [14].

4. Numerical results. In this section we provide the results of numerical experiments illustrating the optimality of the BPX preconditioners developed in the earlier sections. We also compare the results of the BPX preconditioners with those obtained using the corresponding hierarchical preconditioners which are suboptimal.

The first problem that we solve is given by

$$-\Delta_S u = 2x \quad \text{on} \quad S, \tag{4.1}$$

and the exact solution u equals x, which can easily be checked since spherical harmonics are eigenfunctions of the Laplace–Beltrami operator on S [24]. To discretize problem (4.1) we use the basis functions $\phi_{i,j}$. We start from an almost uniform triangulation Δ_0 by projecting the twelve vertices of the regular icosahedron onto the sphere. These twelve points define a mesh consisting of twenty equal spherical triangles; cf. [6]. The finer triangulations Δ_j are constructed by subdividing the triangles of the previous coarser triangulation into four equal subtriangles. Hence the dimension of the spline space increases like $2 + 10 \cdot 4^j$ with the refinement level j. Inner products of the form $(\nabla_S \phi_{i_1,j}, \nabla_S \phi_{i_2,j})$ will have to be computed. Hereto, we use a third order Gaussian quadrature formula on a triangle; see also [4, Prop. 4.1].

The second problem that we solve is given by

$$\Delta_S^2 u = 36xy \quad \text{on} \quad S, \tag{4.2}$$

and the exact solution u equals xy. In order to discretize (4.2) we have to compute inner products of the form $(\Delta_S B_{i_1,j}^{k_1}, \Delta_S B_{i_2,j}^{k_2})$. Since the basis functions $B_{i,j}^k$ are piecewise quadratic polynomials, we can use the formula

$$\Delta_S B_{i,j}^k(v) = \Delta B_{i,j}^k(v) - 6B_{i,j}^k(v), \quad v \in S,$$

with Δ the usual Laplace operator on $\mathbb{R}^3$; see [24]. Then, to evaluate the inner products, we use again a third order Gaussian quadrature formula on a triangle. We show results for the $\sqrt{3}$ refinement procedure where we start from the same quasi-uniform triangulation Δ_0 as in the first problem (4.1). The dimension of the spline space increases like $6 + 30 \cdot 3^j$ with the refinement level j.

Note that the solution u in (4.1) and (4.2) is unique only up to a constant. From [2, Prop. 7.2] we find that constant functions on the sphere are contained in the spherical PS spline space $S_2^1(\Delta_j^{PS})$ but not in the spherical piecewise linear spline space $S_1^0(\Delta_j)$. Hence, the stiffness matrix corresponding to the nodal basis $\{B_{i,j}^k\}$ will have one zero eigenvalue with an eigenvector corresponding to the constant function. The stiffness matrix corresponding to the nodal basis $\{\phi_{i,j}\}$ will have an eigenvalue of $\mathcal{O}(h^2)$ with an

TABLE 4.1
Iteration history for problem (4.1).

		BPX			HB		
dim	J	κ	residual	#iter	κ	residual	#iter
42	1	3.1	2.4897e-05	12	7.6	2.4974e-05	17
162	2	3.7	1.6766e-05	9	10.7	1.9546e-05	16
642	3	4.6	4.7350e-06	11	15.2	8.6198e-06	20
2562	4	5.5	4.5474e-06	11	22.2	5.2361e-06	22
10242	5	6.2	1.6705e-06	12	31.9	3.0622e-06	23
40962	6	6.7	1.0193e-06	12	44.9	1.4750e-06	25
163842	7	7.0	6.2720e-07	12	60.9	6.5043e-07	26
655362	8	7.4	1.6451e-07	13	84.2	3.4960e-07	24

TABLE 4.2
Iteration history for problem (4.2).

		BPX			HB		
dim	J	κ	residual	#iter	κ	residual	#iter
96	1	51.0	3.0555e-11	10	60.7	1.5555e-03	5
276	2	65.7	8.7509e-04	7	82.0	4.9572e-04	9
816	3	79.5	3.9398e-04	8	103.7	5.4025e-04	15
2436	4	88.4	2.6345e-04	11	123.6	2.5576e-04	20
7296	5	96.8	1.7065e-04	11	152.0	1.8184e-04	26
21876	6	103.4	8.9656e-05	13	192.1	1.0873e-04	31
65616	7	107.7	5.0634e-05	13	237.0	6.3035e-05	36
196836	8	110.2	3.2635e-05	14	310.7	3.5946e-05	44

eigenvector that approximates the constant function up to discretization error $\mathcal{O}(h^2)$ with respect to the L_2-norm. Note that the condition numbers that we compute are given by $\kappa(C^{1/2}AC^{1/2}) = \lambda_{\max}/\lambda_{\min}$, where $\lambda_{\max}$ denotes the largest eigenvalue of $C^{1/2}AC^{1/2}$ and $\lambda_{\min}$ its smallest nonzero eigenvalue. For obvious reasons we also omit the smallest eigenvalue of $\mathcal{O}(h^2)$ for the Poisson equation. Note that, from Theorem 3.3 and Remark 3.5, the BPX preconditioner uses all nodal basis functions on all levels. For each redundant basis function we will get a zero eigenvalue.

Tables 4.1 and 4.2 show the results. We have used a *nested iteration conjugate gradient method* to solve the problem; i.e., by means of an outer iteration loop going from a coarse resolution level to the finest resolution level we compute the solution to (4.1) or (4.2) at each level with the BPX-preconditioned conjugate gradient method and we use the solution obtained at the previous coarser level as an initial guess. At each level we stop the conjugate gradient iteration if the H^m-norm of the residual is proportional to the discretization error which is of $\mathcal{O}(h)$. In [11] arguments are given for the fact that nested iteration is an asymptotically optimal method in the sense that it provides the solution u at the finest resolution level J up to discretization error in an overall number of $\mathcal{O}(N_J)$ operations, provided that an optimal preconditioner is used.

Each table has the same setup. The first column shows the dimension of the spline space, and the second column contains the resolution level J. Then we distinguish between the results for the BPX preconditioner and the results for the HB preconditioner. For each preconditioner we display the spectral condition number κ of the system matrix for the linear system of equations that is solved. Moreover, we show the H^m-norm of the residuals corresponding to the approximate solution, and the number of iterations that are needed on this level to reach discretization error accuracy.

Remark 4.1. Computing the H^m-norm of the residual is easy. Let us concentrate on problem (1.2). We have that

$$\begin{aligned}\|u_J - u\|_{H^2}^2 &\sim \|Au_J - b\|_{(H^2)'}^2 \\ &= \left\| \sum_{j=0}^{J} \sum_{i=1}^{N_j} \sum_{k=1}^{3} \langle B_{i,j}^k, A(u_j - u)\rangle \widetilde{B}_{i,j}^k \right\|_{(H^2)'}^2 \\ &\sim \sum_{j=0}^{J} \sum_{i=1}^{N_j} \sum_{k=1}^{3} |\langle B_{i,j}^k, A(u_j - u)\rangle|^2 \\ &= \sum_{j=0}^{J} \sum_{i=1}^{N_j} \sum_{k=1}^{3} \left|\langle B_{i,j}^k, Au_j\rangle - \langle B_{i,j}^k, b\rangle\right|^2 .\end{aligned}$$

Here $\{\widetilde{B}_{i,j}^k\}$ is the dual frame to $\{B_{i,j}^k\}$. The first equivalence is due to the ellipticity of the operator A. The second equivalence is because the dual frame is a Riesz frame for the dual function space $\left(H^2\right)'$. The last expression is just the l_2-norm of the residual of the system (1.8) with respect to the frame $\{B_{i,j}^k\}$ (see also Remark 3.5). This trick works only for elliptic partial differential equations, because the first equivalence above makes use of the ellipticity condition (3.4).

REFERENCES

[1] B. Aksoylu, A. Khodakovsky, and P. Schröder, *Multilevel solvers for unstructured surface meshes*, SIAM J. Sci. Comput., 26 (2005), pp. 1146–1165.

[2] P. Alfeld, M. Neamtu, and L. L. Schumaker, *Bernstein–Bézier polynomials on spheres and sphere-like surfaces*, Comput. Aided Geom. Design, 13 (1996), pp. 333–349.

[3] P. Alfeld, M. Neamtu, and L. L. Schumaker, *Dimension and local bases of homogeneous spline spaces*, SIAM J. Math. Anal., 27 (1996), pp. 1482–1501.

[4] P. Alfeld, M. Neamtu, and L. L. Schumaker, *Fitting scattered data on sphere-like surfaces using spherical splines*, J. Comput. Appl. Math., 73 (1996), pp. 5–43.

[5] T. Aubin, *Nonlinear Analysis on Manifolds: Monge-Ampère Equations*, Springer-Verlag, New York, 1982.

[6] J. R. Baumgardner and P. O. Frederickson, *Icosahedral discretization of the two-sphere*, SIAM J. Numer. Anal., 22 (1985), pp. 1107–1115.

[7] J. H. Bramble, J. E. Pasciak, and J. Xu, *Parallel multilevel preconditioners*, Math. Comp., 55 (1990), pp. 1–22.

[8] D. Castaño, G. Constantinescu, A. Kunoth, and W. D. Schuh, *Approximate continuation of harmonic functions*, manuscript, 2006.

[9] Ch. Coatmélec, *Approximation et interpolation des fonctions différentiables de plusieurs variables*, Ann. Sci. École Norm. Sup. (4), 83 (1966), pp. 271–341.

[10] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[11] W. Dahmen, A. Kunoth, and R. Schneider, *Wavelet least squares methods for boundary value problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1985–2013.

[12] W. Dahmen and A. Kunoth, *Multilevel preconditioning*, Numer. Math., 63 (1992), pp. 315–344.

[13] W. Dahmen, P. Oswald, and X.-Q. Shi, *C^1-hierarchical bases*, J. Comput. Appl. Math., 51 (1994), pp. 37–56.

[14] W. Dahmen, *Some remarks on multiscale transformations, stability, and biorthogonality*, in Wavelets, Images, and Surface Fittings, P. J. Laurent, A. Le Méhauté, and L. L. Schumaker, eds., A. K. Peters, Wellesley, MA, 1994, pp. 157–188.

[15] R. A. DeVore and V. A. Popov, *Interpolation of Besov spaces*, Trans. Amer. Math. Soc., 305 (1988), pp. 397–414.

[16] G. Dziuk, *Finite elements for the Beltrami operator on arbitrary surfaces*, in Partial Differential Equations and Calculus of Variations, Lecture Notes in Math. 1357, S. Hildebrandt and R. Leis, eds., Springer-Verlag, Berlin, 1988, pp. 142–155.
[17] G. Farin, *Curves And Surfaces For Computer Aided Geometric Design: A Practical Guide*, 4th ed., Academic Press, Boston, 1997.
[18] M. Griebel, *Multilevel algorithms considered as iterative methods on semidefinite systems*, SIAM J. Sci. Comput., 15 (1994), pp. 547–565.
[19] L. Kobbelt, $\sqrt{3}$*-subdivision*, in Proceedings of the 27th Annual Conference on Graphics and Interactive Techniques, ACM Press, Addison–Wesley, New York, 2000, pp. 103–112.
[20] A. Kunoth, *Multilevel Preconditioning*, Ph.D. thesis, Verlag Shaker, Aachen, Germany, 1994.
[21] U. Labsik and G. Greiner, *Interpolatory* $\sqrt{3}$*-subdivision*, Comput. Graph. Forum, 19 (2000), pp. 131–138.
[22] J.-L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications, Vol.* I, Springer-Verlag, New York, 1972.
[23] J. Maes and A. Bultheel, *A hierarchical basis preconditioner for the biharmonic equation on the sphere*, IMA J. Numer. Anal., 26 (2006), pp. 563–583.
[24] C. Müller, *Spherical Harmonics*, Lecture Notes in Math. 17, Springer-Verlag, Berlin, New York, 1966.
[25] M. Neamtu and L. L. Schumaker, *On the approximation order of splines on spherical triangulations*, Adv. Comput. Math., 21 (2004), pp. 3–20.
[26] P. Oswald, *Hierarchical conforming finite element methods for the biharmonic equation*, SIAM J. Numer. Anal., 29 (1992), pp. 1610–1625.
[27] P. Oswald, *On discrete norm estimates related to multilevel preconditioners in the finite element method*, in Proceedings of the International Conference on Constructive Theory of Functions, Bulgarian Academy of Science, Bulgaria, K. G. Ivanov, P. Petrushev, and B. Sendov, eds., Sofia, (Varna, 1991) 1992, pp. 203–214.
[28] M. J. D. Powell and M. A. Sabin, *Piecewise quadratic approximations on triangles*, ACM Trans. Math. Software, 3 (1977), pp. 316–325.
[29] P. Sablonnière, *Error bounds for Hermite interpolation by quadratic splines on an* α*-triangulation*, IMA J. Numer. Anal., 7 (1987), pp. 495–508.
[30] E. Vanraes, J. Windmolders, A. Bultheel, and P. Dierckx, *Automatic construction of control triangles for subdivided Powell–Sabin splines*, Comput. Aided Geom. Design, 21 (2004), pp. 671–682.
[31] H. Yserentant, *On the multi-level splitting of finite element spaces*, Numer. Math., 49 (1986), pp. 379–412.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 223–235

SUPERCONVERGENCE FOR CONTROL-VOLUME MIXED FINITE ELEMENT METHODS ON RECTANGULAR GRIDS*

THOMAS F. RUSSELL†, MARY F. WHEELER‡, AND IVAN YOTOV§

Abstract. We consider control-volume mixed finite element methods for the approximation of second-order elliptic problems on rectangular grids. These methods associate control volumes (covolumes) with the vector variable as well as the scalar, obtaining local algebraic representation of the vector equation (e.g., Darcy's law) as well as the scalar equation (e.g., conservation of mass). We establish $O(h^2)$ superconvergence for both the scalar variable in a discrete L^2-norm and the vector variable in a discrete $H(\mathrm{div})$-norm. The analysis exploits a relationship between control-volume mixed finite element methods and the lowest order Raviart–Thomas mixed finite element methods.

AMS subject classifications. 65N06, 65N12, 65N15, 65N30, 76S05

Key words. control volume, mixed finite element, error estimates, superconvergence

DOI. 10.1137/050646330

1. Introduction.

We consider the second-order elliptic problem in a domain $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, written as a first-order system

$$\mathbf{u} = -K\nabla p \text{ in } \Omega, \tag{1.1}$$

$$\nabla \cdot \mathbf{u} = f \text{ in } \Omega, \tag{1.2}$$

$$\mathbf{u} \cdot \mathbf{n} = 0 \text{ on } \partial\Omega. \tag{1.3}$$

The above equations model single-phase flow in porous media, where p is the fluid pressure, the vector $\mathbf{u}$ is the Darcy velocity, K is a symmetric uniformly positive definite and bounded diagonal tensor, representing the rock permeability divided by the fluid viscosity, $\mathbf{n}$ is the outward unit normal to $\partial\Omega$, and f is the source term satisfying the compatibility condition

$$\int_\Omega f\,dx = 0.$$

The choice of homogeneous Neumann boundary condition corresponds to an impermeable boundary, which is the typical physical situation.

In this paper we consider discretizations for (1.1)–(1.3) based on control-volume mixed finite element methods (CVMFEM) and establish $O(h^2)$ superconvergence for the pressure and velocity in a discrete L^2-norm and $H(\mathrm{div})$-norm, respectively. Most of the arguments can be extended to Dirichlet boundary conditions. However, some

*Received by the editors November 29, 2005; accepted for publication (in revised form) August 4, 2006; published electronically January 22, 2007.
http://www.siam.org/journals/sinum/45-1/64633.html

†Department of Mathematics, University of Colorado at Denver, Denver, CO 80217, and Division of Mathematical Sciences, National Science Foundation, Arlington, VA 22230 (trussell@nsf.gov). This author was partially supported by NSF grants DMS 0084438 and DMS 0222300.

‡Institute for Computational Engineering and Sciences, Department of Aerospace Engineering and Engineering Mechanics, and Department of Petroleum and Geosystems Engineering, University of Texas, Austin, TX 78712 (mfw@ices.utexas.edu). This author was partially supported by DOE grant DE-FGO2-04ER25617 and NSF grants EIA 0121523 and DMS 0411413.

§Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (yotov@math.pitt.edu). This author was partially supported by DOE grant DE-FG02-04ER25618 and NSF grants DMS 0107389 and DMS 0411694.

loss of superconvergence occurs on the boundary in that case. Global $O(h)$ convergence has been shown by Chou et al. [9, 10]; here we obtain the $O(h^2)$ rate suggested by various numerical results (e.g., [8, 19, 24, 22]). Superconvergence is proved by $O(h^2)$ estimates of the differences between the scalar and vector discrete solutions and appropriate projections of the exact solutions.

CVMFEM, first introduced in [8], can be viewed as a type of mixed covolume method [9, 10, 11]. CVMFEM are closely related to the Raviart–Thomas mixed finite element methods (MFEM) [27, 7, 28], cell-centered finite difference (CCFD) methods [29, 30, 4], mimetic finite difference (MFD) methods [5, 21, 6], and multipoint flux approximation (MPFA) methods [1, 17]. Some of these relationships are explored in detail in [22].

Like MFEM, CVMFEM are designed to provide simultaneous (accurate) approximations of pressure and velocity, and local mass conservation, $\int_Q \nabla\cdot\mathbf{u}_h = \int_Q f$ on each finite element Q, where $\mathbf{u}_h$ is the computed velocity. These properties can be difficult to obtain when K is heterogeneous (in particular, discontinuous) and/or anisotropic, especially when it incorporates irregular geological features. The methods listed above seek to accomplish this for flow in porous media, among other applications.

Unlike MFEM, CVMFEM have vector control volumes (covolumes) that give rise to a local discrete Darcy law analogous to (1.1). An engineer measuring the permeability of a core sample will typically impose a pressure at each end and observe the flux through the core. The discrete CVMFEM control volume that corresponds to the discrete flux unknown through a face, consisting of the two adjacent halves of the elements on either side of the face (see Figure 1), plays the role of this core, with the element pressures representing the imposed pressures at the ends. The vector test function associated with the control volume is essentially a piecewise-constant vector field, similar to a unit vector in the control volume and a zero vector outside it. The algebraic equation produced by this test function is the local discrete Darcy law. Thus, CVMFEM represent both physical principles in (1.1)–(1.3) locally.

In MFEM, the test vector belongs to the vector trial space and therefore has a continuous normal component. Because the test and trial spaces are the same, the mass matrix is symmetric and positive definite (SPD). In CVMFEM, the normal component of the test vector is discontinuous at the ends of the control volume, and can also be discontinuous at the element face for general distorted grids. If K is elementwise constant and the elements are affine (parallelograms in two dimensions), the mass matrix is SPD, despite the distinct test and trial spaces; in general, it is not symmetric, but symmetry can be restored by appropriate numerical integration [19].

On a uniform grid with constant K, the lowest-order Raviart–Thomas MFEM, denoted RT_0, yields a tridiagonal mass matrix with weights 1/6, 2/3, 1/6, and the basic CCFD results in a diagonal mass matrix. As will be seen below, CVMFEM leads to weights 1/8, 3/4, 1/8. These are all of the form c, $1-2c$, c, where $c = 0$ (CCFD), 1/6 (MFEM), or 1/8 (CVMFEM). In [19], some heuristic reasons to favor $c = 1/8$ are presented: on a uniform grid, the second-order truncation error term is half that of $c = 0$ and $c = 1/6$; on a nonuniform grid, only $c = 1/8$ matches one-sided compact finite differences, avoiding any first-order local truncation error; in terms of Fourier modes, the ratio of the discrete eigenvalue to the continuous eigenvalue is generally closer to 1 for $c = 1/8$. Numerical results in [22] for homogeneous K show second-order convergence for both MFEM and CVMFEM; on orthogonal grids, the flux error for CVMFEM improves on that of MFEM by a factor of approximately 2.6; on the distorted grids used, CVMFEM is worse by a factor of about 1.3.

The rest of the paper is organized as follows. In the next section we recall the Raviart–Thomas MFEM for (1.1)–(1.3). Section 3 describes the CVMFEM and its relation to the Raviart–Thomas MFEM. Superconvergence for the velocity is established in section 4. Section 5 is devoted to superconvergence for the pressure.

2. Mixed finite element methods. We will make use of the following standard notation. For a subdomain $G \subset \mathbb{R}^d$, the $L^2(G)$ inner product (or duality pairing) for scalar and vector valued functions is denoted by $(\cdot,\cdot)_G$. We denote the norm in the Sobolev space $W_p^k(G)$, $k \in \mathbb{R}$, $1 \le p \le \infty$ [2], by $\|\cdot\|_{k,p,G}$. Let $\|\cdot\|_{k,G}$ be the norm of the Hilbert space $H^k(G) = W_2^k(G)$. We omit G in the subscript if $G = \Omega$. For a section of a subdomain boundary $S \subset \mathbb{R}^{d-1}$ we write $\langle\cdot,\cdot\rangle_S$ and $\|\cdot\|_{0,S}$ for the $L^2(S)$ inner product (or duality pairing) and norm, respectively.

The mixed variational formulation, which is the basis for the MFEM is as follows. Find $\mathbf{u} \in \mathbf{V}$ and $p \in W$ such that

$$(K^{-1}\mathbf{u}, \mathbf{v}) = (p, \nabla\cdot\mathbf{v}), \quad \mathbf{v} \in \mathbf{V}, \tag{2.1}$$

$$(\nabla\cdot\mathbf{u}, w) = (f, w), \quad w \in W, \tag{2.2}$$

where

$$\mathbf{V} = \{\mathbf{v} \in H(\mathrm{div};\Omega) : \mathbf{v}\cdot\mathbf{n} = 0 \text{ on } \partial\Omega\}, \quad W = L_0^2(\Omega) = \left\{w \in L^2(\Omega) : \int_\Omega w\,dx = 0\right\},$$

and

$$H(\mathrm{div};\Omega) = \{\mathbf{v} : \mathbf{v} \in (L^2(\Omega))^2,\ \nabla\cdot\mathbf{v} \in L^2(\Omega)\}$$

with a norm

$$\|\mathbf{v}\|_{\mathbf{V}} = (\|\mathbf{v}\|^2 + \|\nabla\cdot\mathbf{v}\|^2)^{1/2}.$$

We assume that Ω can be exactly covered by a rectangular-type finite element partition $\mathcal{T}_h$. Let $\mathbf{V}_h \times W_h \subset \mathbf{V} \times W$ be the lowest-order Raviart–Thomas (RT_0) mixed finite element spaces on $\mathcal{T}_h$ [27]. More precisely, for all $Q \in \mathcal{T}_h$,

$$\mathbf{V}_h(Q) = \{\mathbf{v} = (a_1{+}b_1x, a_2{+}b_2y, a_3{+}b_3z)^T \text{ on } Q\}, \quad W_h(Q) = \{w = \text{constant on } Q\},$$

$$\mathbf{V}_h = \{\mathbf{v} \in \mathbf{V} : \mathbf{v}|_Q \in \mathbf{V}_h(Q)\ \forall\, Q \in \mathcal{T}_h\}, \quad W_h = \{w \in W : w|_Q \in W_h(Q)\ \forall\, Q \in \mathcal{T}_h\},$$

where the third component of $\mathbf{v}$ should be removed if $d = 2$. The degrees of freedom of $\mathbf{V}_h$ are the constant normal components on the sides. If these are continuous, then $\mathbf{v} \in H(\mathrm{div};\Omega)$. Key properties of the RT_0 spaces are

$$\nabla\cdot\mathbf{V}_h = W_h \tag{2.3}$$

and the existence of an interpolation operator $\Pi : (H^1(\Omega))^d \to \mathbf{V}_h$ (see [27, 7]) such that for $\mathbf{q} \in (H^1(\Omega))^2$

$$(\nabla\cdot(\Pi\mathbf{q} - \mathbf{q}), w) = 0 \quad \forall\, w \in W_h \tag{2.4}$$

and which satisfies the continuity and approximation properties

$$\|\Pi\mathbf{q}\|_{\mathbf{V}} \le C\|\mathbf{q}\|_1, \tag{2.5}$$

$$\|\mathbf{q} - \Pi\mathbf{q}\|_0 \le Ch|\mathbf{q}|_1. \tag{2.6}$$

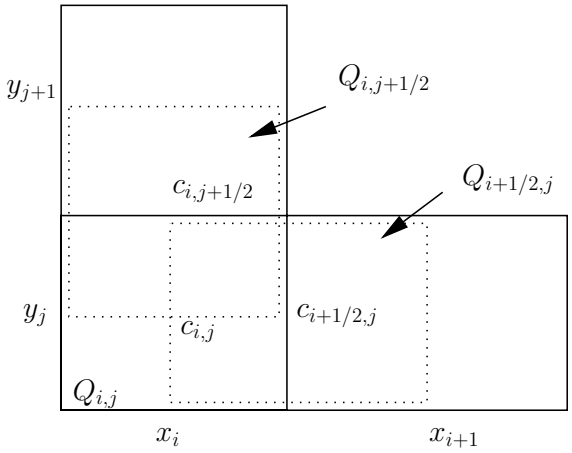


FIG. 1. *Computational grid and control volumes.*

The MFEM for approximating (2.1)–(2.2) is as follows. Find $\tilde{\mathbf{u}}_h \in \mathbf{V}_h$, $\tilde{p}_h \in W_h$ such that

$$(K^{-1}\tilde{\mathbf{u}}_h, \mathbf{v}) = (\tilde{p}_h, \nabla \cdot \mathbf{v}), \quad \mathbf{v} \in \mathbf{V}_h, \tag{2.7}$$

$$(\nabla \cdot \tilde{\mathbf{u}}_h, w) = (f, w), \quad w \in W_h. \tag{2.8}$$

It has been shown in [27] that (2.7)–(2.8) has a unique solution and

$$\|p - \tilde{p}_h\|_W + \|\mathbf{u} - \tilde{\mathbf{u}}_h\|_{\mathbf{V}} = O(h).$$

A number of authors have studied superconvergence for the above method or the closely related CCFD method [25, 14, 30, 15, 16, 18, 4] and have shown results of the form

$$|||p - \tilde{p}_h|||_W + |||\mathbf{u} - \tilde{\mathbf{u}}_h|||_{\mathbf{V}} = O(h^2),$$

where $|||\cdot|||_W$ and $|||\cdot|||_{\mathbf{V}}$ are discrete norms defined in (4.8) and (4.9) below (or some variants of them). The goal of this paper is to obtain similar superconvergence results for the CVMFEM.

3. Control volume mixed finite element methods. Denote the elements of $\mathcal{T}_h$ by $Q_{i,j}$ for $d = 2$ or by $Q_{i,j,k}$ for $d = 3$; see Figure 1 for $d = 2$. For simplicity, in most of the paper we will use the notation and present the arguments for $d = 2$. The case $d = 3$ is a trivial extension.

The center of $Q_{i,j}$ is denoted by $c_{i,j}$. The midpoints of the left and right edges are denoted by $c_{i-1/2,j}$ and $c_{i+1/2,j}$, respectively, with similar notation for the bottom and top edges. With each edge we associate a control volume, where Darcy's law (1.1) is approximated. In particular, letting $c_{i+1/2,j} = (x_{i+1/2}, y_j)$, $c_{i,j} = (x_i, y_j)$, etc., define

$$Q_{i+1/2,j} := (x_i, x_{i+1}) \times (y_{j-1/2}, y_{j+1/2}) \cap \Omega, \tag{3.1}$$

$$Q_{i,j+1/2} := (x_{i-1/2}, x_{i+1/2}) \times (y_i, y_{i+1}) \cap \Omega. \tag{3.2}$$

The control volumes $Q_{i+1/2,j}$ and $Q_{i,j+1/2}$ are referred to as v_1-volumes and v_2-volumes, respectively. The control volumes that have at least one edge on $\partial\Omega$ are called border volumes.

Define the velocity test space

$$\begin{aligned}\mathbf{Y}_h = \{(v_h^1, v_h^2) : v_h^1|_{Q_{i+1/2,j}} &= \text{ constant } \forall\, Q_{i+1/2,j},\ v_h^1 = 0 \text{ on border } v_1\text{-volumes}\\ v_h^2|_{Q_{i,j+1/2}} &= \text{ constant } \forall\, Q_{i,j+1/2},\ v_h^2 = 0 \text{ on border } v_2\text{-volumes}\}.\end{aligned}$$

Thus, for example, the basis function $\mathbf{y}_{i+1/2,j} \in \mathbf{Y}_h$ associated with $c_{i+1/2,j}$ is the vector $(\chi_{i+1/2,j}, 0)$, i.e., $(1,0)$ on $Q_{i+1/2,j}$, $(0,0)$ elsewhere. To see the form of the associated algebraic equation, write (1.1) as $K^{-1}\mathbf{u} + \nabla p = 0$, form the inner product with $\mathbf{y}_{i+1/2,j}$, and integrate

$$\int_{x_i}^{x_{i+1}} \int_{y_{j-1/2}}^{y_{j+1/2}} (K^1)^{-1} u^1 \, dy\, dx + \int_{y_{j-1/2}}^{y_{j+1/2}} (p(x_{i+1}, y) - p(x_i, y))\, dy = 0,$$

where $\mathbf{u} = (u^1, u^2)$ and $K = \operatorname{diag}(K^1, K^2)$. Suppose that K is elementwise constant on $Q_{i,j}$ and $Q_{i+1,j}$. Taking $\mathbf{u} = \mathbf{v}_{i-1/2,j}, \mathbf{v}_{i+1/2,j}, \mathbf{v}_{i+3/2,j} \in \mathbf{V}_h$, the usual RT_0 vector basis functions, we obtain the tridiagonal mass-matrix coefficients

$$1/8\,(K^1_{i,j})^{-1} h_i^x h_j^y,\ 3/8\,(K^1_{i,j})^{-1} h_i^x h_j^y + 3/8\,(K^1_{i+1,j})^{-1} h_{i+1}^x h_j^y,\ 1/8\,(K^1_{i+1,j})^{-1} h_{i+1}^x h_j^y,$$

where h^x and h^y are the element dimensions. For homogeneous K and a uniform grid, this reduces to 1/8, 3/4, 1/8, as noted above.

3.1. Variational formulation for CVMFEM. Following [9], define the bilinear forms $a(\cdot,\cdot) : (L^2(\Omega))^d \times (L^2(\Omega))^d \to \mathbb{R}$, $b(\cdot,\cdot) : \mathbf{Y}_h \times H^1(\Omega) \to \mathbb{R}$, and $c(\cdot,\cdot) : H(\operatorname{div};\Omega) \times L^2(\Omega) \to \mathbb{R}$ as follows:

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) &:= (K^{-1}\mathbf{u}, \mathbf{v}),\\ b(\mathbf{v}, p) &:= \sum_{i,j} \langle p, (v^1, 0)^T \cdot \mathbf{n}\rangle_{\partial Q_{i+1/2,j}} + \sum_{i,j} \langle p, (0, v^2)^T \cdot \mathbf{n}\rangle_{\partial Q_{i,j+1/2}},\\ c(\mathbf{u}, w) &:= (\nabla \cdot \mathbf{u}, w).\end{aligned}$$

LEMMA 3.1. *If* $(\mathbf{u}, p) \in H(\operatorname{div};\Omega) \times H^1(\Omega)$ *solves* (1.1)–(1.3)*, then* $(\mathbf{u}, p)$ *satisfies the variational formulation*

$$a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) = 0, \quad \mathbf{v} \in \mathbf{Y}_h, \tag{3.3}$$

$$c(\mathbf{u}, w) = (f, w), \quad w \in W_h. \tag{3.4}$$

Proof. Equation (1.1) implies, for $\mathbf{v} \in \mathbf{Y}_h$,

$$\begin{aligned}(K^{-1}\mathbf{u}, \mathbf{v}) = (-\nabla p, \mathbf{v}) &= \sum_{i,j} (-\nabla p, (v^1, 0)^T)_{Q_{i+1/2,j}} + \sum_{i,j} (-\nabla p, (0, v^2)^T)_{Q_{i,j+1/2}}\\ &= -\sum_{i,j} \langle p, (v^1, 0)^T \cdot \mathbf{n}\rangle_{\partial Q_{i+1/2,j}} - \sum_{i,j} \langle p, (0, v^2)^T \cdot \mathbf{n}\rangle_{\partial Q_{i,j+1/2}},\end{aligned}$$

giving (3.3). Equation (3.4) follows trivially from (1.2). ∎

The CVMFEM may be formulated as follows. Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times W_h$ such that

$$a(\mathbf{u}_h, \mathbf{v}) + b(\mathbf{v}, p_h) = 0, \quad \mathbf{v} \in \mathbf{Y}_h, \tag{3.5}$$

$$c(\mathbf{u}_h, w) = (f, w), \quad w \in W_h. \tag{3.6}$$

Note that (3.5) is a Petrov–Galerkin FEM, since the test functions differ from the trial functions. We next recall the transfer operator $\gamma_h : \mathbf{V}_h \to \mathbf{Y}_h$, introduced in [9]. Define, for all $\mathbf{v} \in \mathbf{V}_h$,

$$\gamma_h \mathbf{v} = \left(\sum_{i,j} v^1(c_{i+1/2,j}) \chi_{i+1/2,j}, \sum_{i,j} v^2(c_{i,j+1/2}) \chi_{i,j+1/2} \right).$$

It has been shown in [9] that for constants $\alpha > 0$ and C independent of h,

$$b(\gamma_h \mathbf{v}, w) = -c(\mathbf{v}, w) \quad \forall\ \mathbf{v} \in \mathbf{V}_h, w \in W_h, \tag{3.7}$$

$$a(\mathbf{v}, \gamma_h \mathbf{v}) \geq \alpha \|\mathbf{v}\|_0^2 \quad \forall\ \mathbf{v} \in \mathbf{V}_h, \tag{3.8}$$

$$\|\gamma_h \mathbf{v}\|_0 \leq C \|\mathbf{v}\|_0. \tag{3.9}$$

4. Velocity superconvergence analysis. In this section we establish superconvergence for the velocity in the CVMFEM. In the treatment of the permeability K we will make use of the following piecewise smooth space. Let $W^{\alpha}_{\mathcal{T}_h}$ consist of functions φ such that $\varphi|_Q \in W^{\alpha}(Q)$ for all $Q \in \mathcal{T}_h$ and $\|\varphi\|_{\alpha,Q}$ is uniformly bounded, independently of h. Let

$$|||\varphi|||_{\alpha} = \max_{Q \in \mathcal{T}_h} \|\varphi\|_{\alpha,Q}.$$

Subtracting (3.5)–(3.6) from (3.3)–(3.4) gives the error equations

$$a(\mathbf{u} - \mathbf{u}_h, \mathbf{v}) + b(\mathbf{v}, p - p_h) = 0, \quad \mathbf{v} \in \mathbf{Y}_h, \tag{4.1}$$

$$c(\mathbf{u} - \mathbf{u}_h, w) = 0, \quad w \in W_h. \tag{4.2}$$

We first note that (4.2) implies

$$0 = c(\mathbf{u} - \mathbf{u}_h, w) = (\nabla \cdot (\mathbf{u} - \mathbf{u}_h), w) = (\nabla \cdot (\Pi \mathbf{u} - \mathbf{u}_h), w) \quad \forall\, w \in W_h$$

using (2.4). Therefore, using (2.3),

$$\nabla \cdot (\Pi \mathbf{u} - \mathbf{u}_h) = 0. \tag{4.3}$$

Let P_h be the L^2-orthogonal projection onto W_h, satisfying for any $\varphi \in L^2(\Omega)$

$$(\varphi - P_h \varphi, w) = 0 \quad \forall w \in W_h. \tag{4.4}$$

Taking $\mathbf{v} = \gamma_h(\Pi \mathbf{u} - \mathbf{u}_h)$ and $w = P_h p - p_h$ in (4.1)–(4.2) implies

$$\begin{aligned} &a(\Pi \mathbf{u} - \mathbf{u}_h, \gamma_h(\Pi \mathbf{u} - \mathbf{u}_h)) \\ &\qquad = -a(\mathbf{u} - \Pi \mathbf{u}, \gamma_h(\Pi \mathbf{u} - \mathbf{u}_h)) - b(\gamma_h(\Pi \mathbf{u} - \mathbf{u}_h), p - p_h), \end{aligned} \tag{4.5}$$

$$c(\Pi \mathbf{u} - \mathbf{u}_h, P_h p - p_h) = 0. \tag{4.6}$$

The second term on the right in (4.5) can be manipulated as follows:

$$\begin{aligned} b(\gamma_h(\Pi \mathbf{u} - \mathbf{u}_h), p - p_h) &= b(\gamma_h(\Pi \mathbf{u} - \mathbf{u}_h), p - P_h p) + b(\gamma_h(\Pi \mathbf{u} - \mathbf{u}_h), P_h p - p_h) \\ &= b(\gamma_h(\Pi \mathbf{u} - \mathbf{u}_h), p - P_h p) - c(\Pi \mathbf{u} - \mathbf{u}_h, P_h p - p_h) \\ &= b(\gamma_h(\Pi \mathbf{u} - \mathbf{u}_h), p - P_h p), \end{aligned}$$

using (3.7) and (4.6) in the last equality. Therefore (4.5) gives

(4.7)
$$a(\Pi\mathbf{u}-\mathbf{u}_h,\gamma_h(\Pi\mathbf{u}-\mathbf{u}_h)) = -a(\mathbf{u}-\Pi\mathbf{u},\gamma_h(\Pi\mathbf{u}-\mathbf{u}_h)) - b(\gamma_h(\Pi\mathbf{u}-\mathbf{u}_h),p-P_hp).$$

Lemma 4.4 implies that

$$|a(\mathbf{u}-\Pi\mathbf{u},\gamma_h(\Pi\mathbf{u}-\mathbf{u}_h)| \le Ch^2|||K^{-1}|||_{1,\infty}\|\mathbf{u}\|_2\|\Pi\mathbf{u}-\mathbf{u}_h\|_0.$$

Using (4.3), Lemma 4.5 gives

$$|b(\gamma_h(\Pi\mathbf{u}-\mathbf{u}_h),p-P_hp)| \le Ch^2\|p\|_3\|\Pi\mathbf{u}-\mathbf{u}_h\|_0.$$

With the above two bounds and (3.8), (4.7) implies the following superconvergence result.

THEOREM 4.1. *For the CVMFEM approximation* $(\mathbf{u}_h,p_h)$*, there exists a constant* C *independent of* h *such that*

$$\|\Pi\mathbf{u}-\mathbf{u}_h\|_0 \le Ch^2|||K^{-1}|||_{1,\infty}(\|\mathbf{u}\|_2+\|p\|_3).$$

Remark 4.1. The velocity superconvergence result of Theorem 4.1 and the pressure superconvergence bound of Theorem 5.1 require global smoothness of $\mathbf{u}$ and p. There are practical cases when the solution is locally smooth on a given region but possesses reduced regularity globally, such as aquifers with faults or multiple rock layers. Such cases could be treated by establishing interior and negative norm bounds, using techniques developed in [26, 14].

The above result immediately implies superconvergence for the velocity in an L^2 sense along the Gaussian lines. Consider an element $Q=[a_1,b_1]\times[a_2,b_2]$. Following [18, 16], for a vector $\mathbf{q}=(q_1,q_2)$ define

$$|||q_1|||_{1,Q}^2 = (b_2-a_2)\int_{a_1}^{b_1}\left|q_1\left(x_1,\frac{a_2+b_2}{2}\right)\right|^2 dx_1,$$

$$|||q_2|||_{2,Q}^2 = (b_1-a_1)\int_{a_2}^{b_2}\left|q_2\left(\frac{a_1+b_1}{2},x_2\right)\right|^2 dx_2,$$

$$|||\mathbf{q}|||^2 = \sum_{i=1}^{2}\sum_{Q\in\mathcal{T}_h}|||q_i|||_{i,Q}^2.$$

Note that for $\mathbf{q}\in\mathbf{V}_h$, $|||\mathbf{q}||| = \|\mathbf{q}\|_0$.

COROLLARY 4.2. *There exists a constant* C *independent of* h *such that*

$$|||\mathbf{u}-\mathbf{u}_h||| \le Ch^2|||K^{-1}|||_{1,\infty}(\|\mathbf{u}\|_2+\|p\|_3).$$

Proof: It was shown in [16] that

$$|||\mathbf{u}-\Pi\mathbf{u}||| \le Ch^2|\mathbf{u}|_2,$$

where $|\cdot|_2$ denotes the H^2-seminorm. Also, using Theorem 4.1,

$$|||\Pi\mathbf{u}-\mathbf{u}_h||| = \|\Pi\mathbf{u}-\mathbf{u}_h\|_0 \le Ch^2|||K^{-1}|||_{1,\infty}(\|\mathbf{u}\|_2+\|p\|_3).$$

The assertion of the corollary follows from the above two bounds and the triangle inequality. ∎

It is also easy to see that $\nabla \cdot (\mathbf{u} - \mathbf{u}_h)$ is superconvergent at the midpoints of the elements. Define, for a scalar function g,

$$|||g||| = \left(\sum_{i,j} |Q_{i,j}| g(c_{i,j})^2 \right)^{1/2}. \tag{4.8}$$

Using (4.3) and (2.4),

$$|||\nabla \cdot (\mathbf{u} - \mathbf{u}_h)||| = |||\nabla \cdot (\mathbf{u} - \Pi\mathbf{u})||| = |||\nabla \cdot \mathbf{u} - \widehat{\nabla \cdot \mathbf{u}}||| \le Ch^2 \|\nabla \cdot \mathbf{u}\|_{2,\infty},$$

where the last inequality follows from Lemma 4.6. Defining

$$|||\mathbf{q}|||_{\mathbf{V}}^2 = |||\mathbf{q}|||^2 + |||\nabla \cdot \mathbf{q}|||^2, \tag{4.9}$$

the above results can be summarized as follows.

COROLLARY 4.3. *There exists a constant C independent of h such that*

$$|||\mathbf{u} - \mathbf{u}_h|||_{\mathbf{V}} \le Ch^2 (\|\mathbf{u}\|_2 + \|\nabla \cdot \mathbf{u}\|_{2,\infty} + \|p\|_3). \tag{4.10}$$

We next proceed with the three lemmas needed in the proof of Theorem 4.1.

LEMMA 4.4. *There exists a constant C independent of h such that, for all $\mathbf{v} \in \mathbf{V}_h$,*

$$|a(\mathbf{u} - \Pi\mathbf{u}, \gamma_h \mathbf{v})| \le Ch^2 |||K^{-1}|||_{1,\infty} \|\mathbf{u}\|_2 \|\mathbf{v}\|_0.$$

Proof. We first show that if $\mathbf{q} \in (P_1(Q))^2$, where P_k is the space of polynomials of degree $\le k$, then

$$\int_Q (\mathbf{q} - \Pi\mathbf{q}) \gamma_h \mathbf{v} \, dx \, dy = 0 \quad \forall \mathbf{v} \in \mathbf{V}_h, \ Q \in \mathcal{T}_h. \tag{4.11}$$

The argument follows the proof of Lemma 3.1 in [16]. Let $Q = [a,b] \times [c,d]$ and let $L_1(x)$ and $\tilde{L}_1(y)$ be the linear Legendre polynomials on $[a,b]$ and $[c,d]$, respectively. It is easy to see that any $\mathbf{q} \in (P^1(Q))^2$ can be decomposed into

$$\mathbf{q}(x,y) = \bar{\mathbf{q}}(x,y) + (\alpha \tilde{L}_1(y), \beta L_1(x))^T,$$

where $\bar{\mathbf{q}} \in \mathbf{V}_h(Q)$. Since $\bar{\mathbf{q}} - \Pi\bar{\mathbf{q}} = 0$, it is enough to establish (4.11) for $\mathbf{q}(x,y) = (\alpha \tilde{L}_1(y), \beta L_1(x))^T$. It is shown in [16] that in this case $\Pi\mathbf{q} = 0$. Therefore

$$\begin{aligned} \int_Q (\mathbf{q} - \Pi\mathbf{q}) \gamma_h \mathbf{v} \, dx \, dy &= \int_Q \mathbf{q} \gamma_h \mathbf{v} \, dx \, dy \\ &= \int_Q (\alpha \tilde{L}_1(y) (\gamma_h \mathbf{v})^1(x,y) + \beta L_1(x) (\gamma_h \mathbf{v})^2(x,y)) \, dx \, dy = 0, \end{aligned}$$

using that for any fixed $x_0 \in [a,b]$, $(\gamma_h \mathbf{v})^1(x_0, y) \in P_0[c,d]$, that for any fixed $y_0 \in [c,d]$, $(\gamma_h \mathbf{v})^2(x, y_0) \in P_0[a,b]$, and the orthogonality properties of $L_1(x)$ and $\tilde{L}_1(y)$.

We now have

$$\begin{aligned} a(\mathbf{u} - \Pi\mathbf{u}, \gamma_h \mathbf{v}) &= (K^{-1}(\mathbf{u} - \Pi\mathbf{u}), \gamma_h \mathbf{v}) \\ &= \sum_{Q \in \mathcal{T}_h} [K_Q^{-1}(\mathbf{u} - \Pi\mathbf{u}, \gamma_h \mathbf{v})_Q + ((K^{-1} - K_Q^{-1})(\mathbf{u} - \Pi\mathbf{u}), \gamma_h \mathbf{v})_Q], \end{aligned}$$

where K_Q^{-1} is the value of K^{-1} at the center of Q. Therefore

$$|a(\mathbf{u}-\Pi\mathbf{u},\gamma_h\mathbf{v})| \le C\|K^{-1}\|_{0,\infty}\sum_{Q\in\mathcal{T}_h}|(\mathbf{u}-\Pi\mathbf{u},\gamma_h\mathbf{v})_Q| + Ch|||K^{-1}|||_{1,\infty}\|\mathbf{u}-\Pi\mathbf{u}\|_0\|\gamma_h\mathbf{v}\|_0. \tag{4.12}$$

Using (4.11), an application of the Bramble–Hilbert lemma [12] implies

$$|(\mathbf{u}-\Pi\mathbf{u},\gamma_h\mathbf{v})_Q| \le Ch^2|\mathbf{u}|_{2,Q}\|\gamma_h\mathbf{v}\|_{0,Q},$$

which combined with (4.12), (2.6), and (3.9) completes the proof. ∎

LEMMA 4.5. *There exists a constant C independent of h such that for all $\mathbf{v}\in\mathbf{V}_h$,*

$$|b(\gamma_h\mathbf{v},p-P_hp)| \le Ch^2\|p\|_3\|\mathbf{v}\|_{\mathbf{V}}.$$

Proof. Let $e_{i+1/2,j}=\partial Q_{i+1/2,j}\cap Q_{i,j}$ and $e_{i,j+1/2}=\partial Q_{i,j+1/2}\cap Q_{i,j}$. Note that in the sums in

$$\begin{aligned} &b(\gamma_h\mathbf{v},p-P_hp)\\ &=\sum_{i,j}\langle p-P_hp,((\gamma_h\mathbf{v})^1,0)^T\cdot\mathbf{n}\rangle_{\partial Q_{i+1/2,j}}+\sum_{i,j}\langle p-P_hp,(0,(\gamma_h\mathbf{v})^2)^T\cdot\mathbf{n}\rangle_{\partial Q_{i,j+1/2}}, \end{aligned}$$

every edge $e_{i+1/2,j}$ and $e_{i,j+1/2}$ appears twice (from the two neighboring covolumes). Using that $\frac{\partial v_1}{\partial x}$ and $\frac{\partial v_2}{\partial y}$ are constants on each element $Q_{i,j}$, we have

$$\begin{aligned} &b(\gamma_h\mathbf{v},p-P_hp)\\ &=\sum_{i,j}\left(h_i^x\frac{\partial v_1}{\partial x}\int_{e_{i+1/2,j}}(p-P_hp)\,dy+h_j^y\frac{\partial v_2}{\partial y}\int_{e_{i,j+1/2}}(p-P_hp)\,dx\right)\\ &=\sum_{i,j}\left(\frac{\partial v_1}{\partial x}\left(h_i^x\int_{e_{i+1/2,j}}p\,dy-\int_{Q_{i,j}}p\,dxdy\right)\right.\\ &\qquad\left.+\frac{\partial v_2}{\partial y}\left(h_j^y\int_{e_{i,j+1/2}}p\,dx-\int_{Q_{i,j}}p\,dxdy\right)\right)\\ &=\sum_{i,j}\left(\left(p,\frac{\partial v_1}{\partial x}\right)_{Q_{i,j},M_x}-\left(p,\frac{\partial v_1}{\partial x}\right)_{Q_{i,j}}+\left(p,\frac{\partial v_2}{\partial y}\right)_{Q_{i,j},M_y}-\left(p,\frac{\partial v_2}{\partial y}\right)_{Q_{i,j}}\right), \end{aligned} \tag{4.13}$$

where $(\cdot,\cdot)_{Q,M_x}$ is the quadrature rule on Q which uses the midpoint rule in x and exact integration in y, and $(\cdot,\cdot)_{Q,M_y}$ uses exact integration in x and the midpoint rule in y. Since the midpoint rule is exact for linear polynomials, the Peano kernel theorem [13, Theorem 3.7.1] implies

$$\begin{aligned} &\left(p,\frac{\partial v_1}{\partial x}\right)_{Q_{i,j},M_x}-\left(p,\frac{\partial v_1}{\partial x}\right)_{Q_{i,j}}=\int_{Q_{i,j}}\varphi(x)\frac{\partial^2 p}{\partial x^2}(x,y)\frac{\partial v_1}{\partial x}\,dxdy\\ &=\int_{Q_{i,j}}\varphi(x)\frac{\partial^2 p}{\partial x^2}(x,y)\nabla\cdot\mathbf{v}\,dxdy-\int_{Q_{i,j}}\varphi(x)\frac{\partial^2 p}{\partial x^2}(x,y)\frac{\partial v_2}{\partial y}\,dxdy\equiv T_1+T_2, \end{aligned} \tag{4.14}$$

where

$$\varphi(x)=\begin{cases}(x-x_{i-1/2})^2/2, & x_{i-1/2}\le x\le x_i,\\ (x-x_{i+1/2})^2/2, & x_i\le x\le x_{i+1/2}.\end{cases}$$

For the first term we have

$$|T_1| \le Ch^2 \|p\|_{2,Q_{i,j}} \|\nabla \cdot \mathbf{v}\|_{0,Q_{i,j}}. \tag{4.15}$$

Integrating by parts in T_2 gives

$$\begin{aligned} T_2 = &\int_{Q_{i,j}} \varphi(x) \frac{\partial^3 p}{\partial x^2 \partial y}(x,y) v_2(x,y)\, dxdy \\ &- \left(\int_{e_{i,j,t}} - \int_{e_{i,j,b}} \right) \varphi(x) \frac{\partial^2 p}{\partial x^2}(x,y) v_2(x,y)\, dx \equiv T_{2,1} + T_{2,2}, \end{aligned} \tag{4.16}$$

where $e_{i,j,t}$ and $e_{i,j,b}$ are the top and the bottom edge of $Q_{i,j}$, respectively. For $T_{2,1}$ we have

$$|T_{2,1}| \le Ch^2 \|p\|_{3,Q_{i,j}} \|\mathbf{v}\|_{0,Q_{i,j}}. \tag{4.17}$$

For $T_{2,2}$ we notice that v_2 is continuous across horizontal edges and the assumed regularity of $p(x,y)$ implies that the trace of $\frac{\partial^2 p}{\partial x^2}$ is well defined. When summing over all elements, each edge integral will appear twice from the expressions for the two neighboring elements, with opposite signs. Therefore

$$\sum_{i,j} T_{2,2} = 0. \tag{4.18}$$

Combining (4.14)–(4.18) implies

$$\sum_{i,j} \left(\left(p, \frac{\partial v_1}{\partial x} \right)_{Q_{i,j}, M_x} - \left(p, \frac{\partial v_1}{\partial x} \right)_{Q_{i,j}} \right) \le Ch^2 \|p\|_3 \|\mathbf{v}\|_{\mathbf{V}}.$$

The second error term in (4.13) can be bounded in a similar way. Note that for $d = 3$, a similar argument goes through with two terms analogous to T_2. ∎

LEMMA 4.6. *For all $g \in W^2_\infty$ there exists a constant C independent of h such that*

$$|||g - P_h g||| \le Ch^2 \|g\|_{2,\infty}.$$

Proof. Let $Q \in \mathcal{T}_h$. The Taylor expansion with integral remainder about the midpoint (x_0, y_0) of Q gives for any $(x,y) \in Q$

$$g(x,y) = g(x_0,y_0) + (x - x_0) \frac{\partial g}{\partial x}(x_0,y_0) + (y - y_0) \frac{\partial g}{\partial y}(x_0,y_0) + R(x,y),$$

where $|R(x,y)| \le Ch^2 \|g\|_{2,\infty,Q}$. Integrating the above equation over Q and using that $\int_Q g = \int_Q P_h g$ gives

$$|Q|(P_h g(x_0,y_0) - g(x_0,y_0)) = \int_Q R(x,y)\, dxdy,$$

which implies

$$|P_h g(x_0,y_0) - g(x_0,y_0)| \le Ch^2 \|g\|_{2,\infty,Q}.$$

The statement of the lemma now follows from the definition (4.8) of $|||\cdot|||$. ∎

5. Pressure superconvergence analysis. In this section we employ a duality argument to derive superconvergence for the pressure at the cell centers. We will make use of the following continuity property of Π [23, 3]. For any $\varepsilon > 0$,

$$\|\Pi \mathbf{q}\|_0 \le C(\|\mathbf{q}\|_\varepsilon + \|\nabla \cdot \mathbf{q}\|_0). \tag{5.1}$$

Consider the auxiliary problem

$$\begin{aligned} -\nabla \cdot K\nabla\varphi &= P_h p - p_h \quad \text{in } \Omega, \\ -K\nabla\varphi \cdot \mathbf{n} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{5.2}$$

which is well posed since $\int_\Omega P_h p = \int_\Omega p_h = 0$. Elliptic regularity [20] implies that there exists $\varepsilon > 0$ such that

$$\|\varphi\|_{1+\varepsilon} \le C\|P_h p - p_h\|_0. \tag{5.3}$$

Note that (5.3) holds for L-shaped domains. Let $\phi = -K\nabla\varphi$. We have

$$\begin{aligned} \|P_h p - p_h\|_0^2 &= (P_h p - p_h, \nabla \cdot \phi) = (P_h p - p_h, \nabla \cdot \Pi\phi) = c(\Pi\phi, P_h p - p_h) \\ &= -b(\gamma_h \Pi\phi, P_h p - p_h) = -b(\gamma_h \Pi\phi, P_h p - p) - b(\gamma_h \Pi\phi, p - p_h) \\ &= -b(\gamma_h \Pi\phi, P_h p - p) + a(\mathbf{u} - \mathbf{u}_h, \gamma_h \Pi\phi), \end{aligned} \tag{5.4}$$

using (4.1) with $\mathbf{v} = \gamma_h \Pi\phi$. By Lemma 4.5,

$$\begin{aligned} |b(\gamma_h \Pi\phi, P_h p - p)| &\le Ch^2 \|p\|_3 \|\Pi\phi\|_{\mathbf{V}} \\ &\le Ch^2 \|p\|_3 (\|\phi\|_\varepsilon + \|\nabla \cdot \phi\|_0) \le Ch^2 |||K|||_{\epsilon,\infty} \|p\|_3 \|P_h p - p_h\|_0 \end{aligned}$$

using (5.1), (5.3), and that $\|\nabla \cdot \Pi\phi\|_0 \le \|\nabla \cdot \phi\|_0$, which follows from $\nabla \cdot \Pi\phi = P_h \nabla \cdot \phi$. For the last term in (5.4) we write

$$\begin{aligned} |a(\mathbf{u} - \mathbf{u}_h, \gamma_h \Pi\phi)| &= |a(\mathbf{u} - \Pi\mathbf{u}, \gamma_h \Pi\phi) + a(\Pi\mathbf{u} - \mathbf{u}_h, \gamma_h \Pi\phi)| \\ &\le C(h^2 |||K^{-1}|||_{1,\infty} \|\mathbf{u}\|_2 \|\Pi\phi\|_0 + \|K^{-1}\|_{0,\infty} \|\Pi\mathbf{u} - \mathbf{u}_h\|_0 \|\gamma_h \Pi\phi\|_0) \\ &\le Ch^2 |||K^{-1}|||_{1,\infty} (\|\mathbf{u}\|_2 + \|p\|_3) \|\Pi\phi\|_0 \\ &\le Ch^2 |||K|||_{\epsilon,\infty} |||K^{-1}|||_{1,\infty} (\|\mathbf{u}\|_2 + \|p\|_3) \|P_h p - p_h\|_0 \end{aligned}$$

using Lemma 4.4, Theorem 4.1, (3.9), (5.1), (5.3), and (5.2). A combination of (5.4) and the above two bounds gives the following pressure superconvergence result.

THEOREM 5.1. *For the CVMFEM approximation* $(\mathbf{u}_h, p_h)$*, there exists a constant* C *independent of* h *such that*

$$\|P_h p - p_h\|_0 \le Ch^2 |||K|||_{\epsilon,\infty} |||K^{-1}|||_{1,\infty} (\|\mathbf{u}\|_2 + \|p\|_3).$$

It is now easy to obtain superconvergence for the pressure at the midpoints of the elements. Let $|||w|||_W = |||w|||$, where $|||w|||$ is defined in (4.8), and note that $|||w|||_W = \|w\|_0$ for all $w \in W_h$.

COROLLARY 5.2. *There exists a constant* C *independent of* h *such that*

$$|||p - p_h|||_W \le Ch^2 |||K|||_{\epsilon,\infty} |||K^{-1}|||_{1,\infty} (\|\mathbf{u}\|_2 + \|p\|_{2,\infty} + \|p\|_3).$$

Proof. The result follows immediately from the triangle inequality

$$|||p - p_h|||_W \le |||p - P_h p|||_W + |||P_h p - p_h|||_W,$$

Lemma 4.6, and Theorem 5.1. ∎

REFERENCES

[1] I. Aavatsmark, T. Barkve, O. Bøe, and T. Mannseth, *Discretization on unstructured grids for inhomogeneous, anisotropic media.* I. *Derivation of the methods*, SIAM J. Sci. Comput., 19 (1998), pp. 1700–1716.

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

[3] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov, *Mixed finite element methods on nonmatching multiblock grids*, SIAM J. Numer. Anal., 37 (2000), pp. 1295–1315.

[4] T. Arbogast, M. F. Wheeler, and I. Yotov, *Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences*, SIAM J. Numer. Anal., 34 (1997), pp. 828–852.

[5] M. Berndt, K. Lipnikov, D. Moulton, and M. Shashkov, *Convergence of mimetic finite difference discretizations of the diffusion equation*, East-West J. Numer. Math., 9 (2001), pp. 265–284.

[6] M. Berndt, K. Lipnikov, M. Shashkov, M. F. Wheeler, and I. Yotov, *Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals*, SIAM J. Numer. Anal., 43 (2005), pp. 1728–1749.

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

[8] Z. Cai, J. E. Jones, S. F. McCormick, and T. F. Russell, *Control-volume mixed finite element methods*, Comput. Geosci., 1 (1997), pp. 289–315.

[9] S.-H. Chou and D. Y. Kwak, *Mixed covolume methods on rectangular grids for elliptic problems*, SIAM J. Numer. Anal., 37 (2000), pp. 758–771.

[10] S.-H. Chou, D. Y. Kwak, and K. Y. Kim, *A general framework for constructing and analyzing mixed finite volume methods on quadrilateral grids: The overlapping covolume case*, SIAM J. Numer. Anal., 39 (2001), pp. 1170–1196.

[11] S.-H. Chou and P. S. Vassilevski, *A general mixed covolume framework for constructing conservative schemes for elliptic problems*, Math. Comp., 68 (1999), pp. 991–1011.

[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.

[13] P. J. Davis, *Interpolation and Approximation*, Dover Publications, New York, 1975.

[14] J. Douglas, Jr., and F. A. Milner, *Interior and superconvergence estimates for mixed methods for second order elliptic problems*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 397–428.

[15] J. Douglas, Jr., and J. Wang, *Superconvergence of mixed finite element methods on rectangular domains*, Calcolo, 26 (1989), pp. 121–133.

[16] R. Durán, *Superconvergence for rectangular mixed finite elements*, Numer. Math., 58 (1990), pp. 287–298.

[17] M. G. Edwards and C. F. Rogers, *Finite volume discretization with imposed flux continuity for the general tensor pressure equation*, Comput. Geosci., 2 (1998), pp. 259–290.

[18] R. E. Ewing, R. D. Lazarov, and J. Wang, *Superconvergence of the velocity along the Gauss lines in mixed finite element methods*, SIAM J. Numer. Anal., 28 (1991), pp. 1015–1029.

[19] V. A. Garanzha and V. N. Konshin, *Approximation Schemes and Discrete Well Models for the Numerical Simulation of the* 2*-D Non-Darcy Fluid Flows in Porous Media*, Tech. rep., Comm. Appl. Math., Computer Centre, Russian Academy of Sciences, Moscow, 1999.

[20] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[21] J. Hyman, J. Morel, M. Shashkov, and S. Steinberg, *Mimetic finite difference methods for diffusion equations*, Comput. Geosci., 6 (2002), pp. 333–352.

[22] R. A. Klausen and T. F. Russell, *Relationships among some locally conservative discretization methods which handle discontinuous coefficients*, Comput. Geosci., 8 (2004), pp. 341–377.

[23] T. P. Mathew, *Domain Decomposition and Iterative Refinement Methods for Mixed Finite Element Discretizations of Elliptic Problems*, Ph.D. thesis, Courant Institute of Mathematical Sciences, New York University, 1989.

[24] R. L. Naff, T. F. Russell, and J. D. Wilson, *Shape functions for velocity interpolation in general hexahedral cells*, Comput. Geosci., 6 (2002), pp. 285–314.

[25] M. Nakata, A. Weiser, and M. F. Wheeler, *Some superconvergence results for mixed finite element methods for elliptic problems on rectangular domains*, in The Mathematics of Finite Elements and Applications V, J. R. Whiteman, ed., Academic Press, London, 1985, pp. 367–389.

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

[27] R. A. Raviart and J. M. Thomas, *A Mixed Finite Element Method for 2nd Order Elliptic Problems*, in Mathematical Aspects of the Finite Element Method, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[28] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. Lions, eds., Elsevier Science Publishers B.V., Amsterdam, 1991, pp. 523–639.

[29] T. F. Russell and M. F. Wheeler, *Finite element and finite difference methods for continuous flows in porous media*, in The Mathematics of Reservoir Simulation, R. E. Ewing, ed., SIAM, Philadelphia, 1983, pp. 35–106.

[30] A. Weiser and M. F. Wheeler, *On convergence of block-centered finite-differences for elliptic problems*, SIAM J. Numer. Anal., 25 (1988), pp. 351–375.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 236–262

ON TAYLOR MODEL BASED INTEGRATION OF ODES*

M. NEHER†, K. R. JACKSON‡, AND N. S. NEDIALKOV§

Abstract. Interval methods for verified integration of initial value problems (IVPs) for ODEs have been used for more than 40 years. For many classes of IVPs, these methods are able to compute guaranteed error bounds for the flow of an ODE, where traditional methods provide only approximations to a solution. Overestimation, however, is a potential drawback of verified methods. For some problems, the computed error bounds become overly pessimistic, or the integration even breaks down. The dependency problem and the wrapping effect are particular sources of overestimations in interval computations. Berz and his coworkers have developed Taylor model methods, which extend interval arithmetic with symbolic computations. The latter is an effective tool for reducing both the dependency problem and the wrapping effect. By construction, Taylor model methods appear particularly suitable for integrating nonlinear ODEs. We analyze Taylor model based integration of ODEs and compare Taylor model methods with traditional enclosure methods for IVPs for ODEs.

Key words. Taylor model methods, verified integration, ODEs, IVPs

AMS subject classifications. 65G40, 65L05, 65L70

DOI. 10.1137/050638448

1. Introduction.

The numerical solution of initial value problems (IVPs) for ODEs is one of the fundamental problems in scientific computation. Today, there are many well-established algorithms for approximate solution of IVPs. However, traditional integration methods usually provide only approximate values for the solution. Precise error bounds are rarely available. The error estimates, which are sometimes delivered, are not guaranteed to be accurate and are sometimes unreliable.

In contrast, reliable integration computes guaranteed bounds for the flow of an ODE, including all discretization and roundoff errors in the computation. Originated by Moore in the 1960s [33], interval computations are a particularly useful tool for this purpose. There is a vast literature on interval methods for verified integration [6, 8, 9, 10, 12, 19, 21, 22, 24, 29, 31, 32, 33, 35, 36, 37, 38, 39, 40, 44, 45, 46, 47], but there are still many open questions. The results of interval arithmetic computations are often impaired by overestimation caused by the dependency problem and by the wrapping effect. In verified integration, overestimation may degrade the computed enclosure of the flow, enforce miniscule step sizes, or even bring about premature abortion of an integration.

Berz and his coworkers have developed Taylor model methods, which combine interval arithmetic with symbolic computations [2, 5, 25, 27, 28]. In Taylor model methods, the basic data type is not a single interval, but a *Taylor model* $\mathcal{U} := p_n(x)+\boldsymbol{i}$ consisting of a multivariate polynomial $p_n(x)$ of order n in m variables, and a remainder interval $\boldsymbol{i}$. In computations that involve $\mathcal{U}$, the polynomial part is propagated by symbolic calculations wherever possible and thus not significantly affected by the

*Received by the editors August 19, 2005; accepted for publication (in revised form) August 21, 2006; published electronically January 22, 2007.
http://www.siam.org/journals/sinum/45-1/63844.html
†Institut für Angewandte und Numerische Mathematik, Universität Karlsruhe (TH), 76128 Karlsruhe, Germany (markus.neher@math.uni-karlsruhe.de).
‡Computer Science Department, University of Toronto, 10 King's College Rd., Toronto, ON, M5S 3G4, Canada (krj@cs.toronto.edu).
§Department of Computing and Software, McMaster University, Hamilton, ON, L8S 4L7, Canada (nedialk@mcmaster.ca).

dependency problem or the wrapping effect. Only the interval remainder term and polynomial terms of order higher than n, which are usually small, are bounded using interval arithmetic.

Taylor model arithmetic is an extension of interval arithmetic with a comprehensive variety of applicable enclosure sets. Nevertheless, there has been some debate about the usefulness and the limitations of Taylor model methods [42]. To some extent, this may be due to the sometimes cursory description of technical details of Taylor model arithmetic, which may be obvious to the experts of Taylor models, but which are less trivial to others.

The motivation of this paper is to analyze Taylor model methods for the verified integration of ODEs and to compare these methods with existing interval methods. Taylor models are better suited for integrating ODEs than interval methods whenever richness in available enclosure sets and reduction of the dependency problem is an advantage. This is usually the case for IVPs for nonlinear ODEs, especially in combination with large initial sets or with large integration domains. Although parameter intervals or initial sets can be handled by subdivision, this approach is only practical in low dimensions.

The advantage of Taylor model methods is less obvious for linear ODEs, where interval methods should perform equally well. Nevertheless, we include a discussion of Taylor model methods for linear ODEs in this paper for two reasons. First, the discussion is simpler for linear ODEs than for nonlinear ones. Second, if Taylor model methods failed on linear ODEs, they would likely fail on nonlinear ODEs as well. However, some of the most advantageous properties of Taylor models take effect only on nonlinear problems. We use a simple nonlinear model problem to illustrate these advantages.

The paper is structured as follows. In the next section, basic concepts of interval arithmetic and Taylor model methods are reviewed. Interval methods for ODEs are presented in section 3. The naive Taylor model method is described in section 4, which is followed by a discussion of Taylor model methods for linear ODEs. A nonlinear model problem is used to explain preconditioned Taylor model methods for ODEs in section 6. In the last section, numerical examples for linear ODEs are given.

2. Preliminaries.

2.1. Interval arithmetic. Interval arithmetic [1, 14, 33, 41] is a powerful tool for verified computations. In interval arithmetic, operations between intervals are employed to calculate guaranteed bounds for continuous problems with a finite number of basic arithmetic operations. We assume that the reader is familiar with real interval arithmetic and floating-point interval arithmetic. The latter is based on a screen of floating-point numbers. Rigor of a computation is achieved by enclosing real numbers by floating-point intervals (that is, intervals with floating-point upper and lower bounds) and by performing all calculations with directed rounding according to the rules of interval arithmetic [20]. Successful software implementations of floating-point interval arithmetic have for example been given in [3, 17, 18].

The set of compact real intervals is denoted by

$$\mathbb{IR} = \{\, \boldsymbol{x} = [\underline{x}, \overline{x}] \mid \underline{x}, \overline{x} \in \mathbb{R},\ \underline{x} \le \overline{x} \,\}.$$

A real number x is identified with a point interval $\boldsymbol{x} = [x, x]$. The *midpoint* and the *width* of an interval $\boldsymbol{x}$ are denoted by $\mathrm{m}(\boldsymbol{x}) := (\overline{x} + \underline{x})/2$ and $\mathrm{w}(\boldsymbol{x}) := \overline{x} - \underline{x}$, respectively. The set of all m-dimensional interval vectors is denoted by $\mathbb{IR}^m$. In this

paper, intervals are denoted by boldface. Lower-case letters are used for denoting scalars and vectors. Matrices are denoted by upper-case letters.

2.2. Dependency problem and wrapping effect. Interval methods are sometimes affected by overestimation, whence the computed error bounds may be overly pessimistic. Overestimation is often caused by the *dependency problem*, that is, the failure of interval arithmetic to identify different occurrences of the same variable. For example, the range of $f(x) := x/(1+x)$ on $\boldsymbol{x} = [1,2]$ is $[1/2, 2/3]$, but interval arithmetic evaluation yields

$$\frac{\boldsymbol{x}}{1+\boldsymbol{x}} = \frac{[1,2]}{[2,3]} = \left[\frac{1}{3}, 1\right].$$

In general, the dependency problem is not easily removed. To diminish overestimation, alternative evaluation schemes, such as centered forms [33], have been developed. A discussion of computer methods for the range of functions is given in [43].

A second source of overestimation is the *wrapping effect*, which appears when intermediate results of a computation are enclosed by intervals. The wrapping effect was first observed by Moore in 1965 [32]; a recent analysis has been given by Lohner [23].

2.3. Taylor model arithmetic. For reducing both the dependency problem and the wrapping effect, interval arithmetic has been extended with symbolic computations. Symbolic-numeric computations have been proposed under various names since the 1980s [11, 16, 25]. Early implementations in software were also given [11, 15], but to the authors' knowledge, these packages have not been widely distributed and are not available today.

Starting in the 1990s, Berz and his group developed a rigorous multivariate Taylor arithmetic [2, 25, 28]. In these references, a *Taylor model* is defined in the following way. Let $f : D \subset \mathbb{R}^m \to \mathbb{R}$ be a function that is $(n+1)$ times continuously differentiable in an open set containing the box $\boldsymbol{x}$. Let x_0 be a point in $\boldsymbol{x}$, let p_n denote the nth order Taylor polynomial of f around x_0, and let $\boldsymbol{i}$ be an interval such that

$$f(x) \in p_n(x - x_0) + \boldsymbol{i} \quad \text{for all} \quad x \in \boldsymbol{x}. \tag{2.1}$$

Then the pair $(p_n, \boldsymbol{i})$ is called an nth order Taylor model of f around x_0 on $\boldsymbol{x}$.

This original definition of a Taylor model is useful for computations in exact arithmetic, but it must be extended for floating-point computations. For example, there is no Taylor model of $e^x \approx 1 + x + (1/2)x^2 + (1/6)x^3 + \ldots$ of order $n \geq 3$ in IEEE 754 floating-point arithmetic, since the coefficient of x^3 is not exactly representable as a floating-point number. In [29], instead of the Taylor polynomial of f, an arbitrary polynomial p_n with floating-point coefficients is used in (2.1), but the definition of a Taylor model in [29] assumes that the width of $\boldsymbol{i}$ is of order $O\big(\|\mathrm{w}(\boldsymbol{x})\|^n\big)$. In this paper, such an assumption on the width of $\boldsymbol{i}$ is not required.

We use calligraphy letters for denoting Taylor models:

$$\mathcal{U} := p_n(x) + \boldsymbol{i}, \quad x \in \boldsymbol{x},$$

where $\boldsymbol{x} \in \mathbb{IR}^m$, $\boldsymbol{i} \in \mathbb{IR}$ are intervals, and p_n is an m-variate polynomial of order n. $\boldsymbol{x}$ is called the *domain interval* of $\mathcal{U}$, and $\boldsymbol{i}$ is its *remainder interval*. A Taylor model is the set of all m-variate continuous functions f such that

$$f(x) \in p_n(x) + \boldsymbol{i}$$

holds for all $x \in \boldsymbol{x}$. Evaluating $\mathcal{U}$ for all $x \in \boldsymbol{x}$, we obtain the *range* of $\mathcal{U}$:

$$\mathrm{Rg}\,(\mathcal{U}) := \{z = p(x) + \iota \mid x \in \boldsymbol{x}, \iota \in \boldsymbol{i}\}.$$

Example 2.1. *Taylor models of* e^x *and* $\cos x$. Let $\boldsymbol{x} := [-\frac{1}{2}, \frac{1}{2}]$ and $x_0 := 0$. Then Taylor's theorem is a natural starting point for constructing Taylor models. We have

$$e^x = 1 + x + \frac{1}{2}x^2 + \frac{1}{6}x^3 e^{\xi}, \quad \cos x = 1 - \frac{1}{2}x^2 + \frac{1}{6}x^3 \sin\xi, \quad x,\, \xi \in \boldsymbol{x},$$

from which we derive Taylor models for $f_1(x) := e^x$ and $f_2(x) := \cos x$:

$$\mathcal{U}_1(x) := 1 + x + \tfrac{1}{2}x^2 + [-0.035, 0.035], \quad \mathcal{U}_2(x) := 1 - \tfrac{1}{2}x^2 + [-0.010, 0.010], \quad x \in \boldsymbol{x},$$

respectively.

Taylor model arithmetic has been defined in [2, 25, 28]. We use the same arithmetic rules, even though our Taylor models differ slightly from the Taylor models defined in these references. The difference affects only the function set that is defined by a Taylor model.

In computations that involve a Taylor model $\mathcal{U}$, the polynomial part is propagated by symbolic calculations wherever possible. In floating-point computations, the roundoff errors of the symbolic operations are rigorously estimated and the estimate is added to the remainder interval of the final result. This part of the computation is hardly affected by the dependency problem or the wrapping effect. Only the interval remainder term and polynomial terms of order higher than n (which in applications are usually small) are processed according to the rules of interval arithmetic.

Example 2.2. *Multiplication of two univariate Taylor models of order* 2. Let $\boldsymbol{x} := [-\frac{1}{2}, \frac{1}{2}]$ and

$$\mathcal{U}_1(x) := 1 + x + \tfrac{1}{2}x^2 + [-0.035, 0.035], \;\; \mathcal{U}_2(x) := 1 - \tfrac{1}{2}x^2 + [-0.010, 0.010],$$

where $x \in \boldsymbol{x}$. For all $x \in \boldsymbol{x}$, it holds that

$$\begin{aligned}
\mathcal{U}_1(x) \cdot \mathcal{U}_2(x) &\subseteq (1 + x + \tfrac{1}{2}x^2)(1 - \tfrac{1}{2}x^2) + \big(\tfrac{1}{2} + \tfrac{1}{2}(1+x)^2\big)[-0.010, 0.010] \\
&\quad + (1 - \tfrac{1}{2}x^2)[-0.035, 0.035] + [-0.035, 0.035] \cdot [-0.010, 0.010] \\
&\subseteq (1 + x) - \tfrac{1}{2}x^3 - \tfrac{1}{4}x^4 + [0.625, 1.625] \cdot [-0.010, 0.010] + [0.875, 1] \cdot [-0.035, 0.035] \\
&\quad + [-0.004, 0.004] \\
&\subseteq 1 + x - [-0.063, 0.063] - [-0.016, 0.016] + [-0.202, 0.202] = 1 + x \\
&\quad + [-0.281, 0.281],
\end{aligned}$$

so we may define

$$\mathcal{U}_1(x) \cdot \mathcal{U}_2(x) := 1 + x + [-0.281, 0.281].$$

This product is a Taylor model for the function $e^x \cos x$, $x \in \boldsymbol{x}$:

$$e^x \cos x \in 1 + x + [-0.281, 0.281], \quad x \in \boldsymbol{x}.$$

In Example 2.2, direct interval evaluation for computing the remainder interval of the product has been used for simplicity. Due to the dependency problem, this does not always yield optimal bounds. More accurate estimation schemes have been proposed in [30].

Compositions $\mathcal{U}_1 \circ \mathcal{U}_2$ of Taylor models are evaluated in a similar way as products; $\circ$ denotes the composition operator for functions, namely,

$$(f \circ g)(x) = f\big(g(x)\big).$$

Example 2.3. *Composition of two univariate Taylor models of order* 2. Let $\boldsymbol{x} := [-\frac{1}{2}, \frac{1}{2}]$ and

$$\mathcal{U}_1(x) := 1 + x + \tfrac{1}{2}x^2 + [-0.035, 0.035], \quad \mathcal{U}_2(x) := 1 - \tfrac{1}{2}x^2 + [-0.010, 0.010],$$

where $x \in \boldsymbol{x}$. It is tempting to compute the composition $\mathcal{U}_1 \circ \mathcal{U}_2$ in the following manner:

$$\begin{aligned}
\mathcal{U}_1(x) \circ \mathcal{U}_2(x) &\subseteq 1 + (1 - \tfrac{1}{2}x^2 + [-0.010, 0.010]) + \tfrac{1}{2}(1 - \tfrac{1}{2}x^2 + [-0.010, 0.010])^2 \\
&\quad + [-0.035, 0.035] \\
&\subseteq 2 - \tfrac{1}{2}x^2 + [-0.045, 0.045] + \tfrac{1}{2}(1 - x^2 + \tfrac{1}{4}x^4 + [-0.020, 0.020] - x^2[-0.010, 0.010] \\
&\quad + [-0.001, 0.001]) \\
&\subseteq \tfrac{5}{2} - x^2 + \tfrac{1}{8}x^4 - x^2[-0.005, 0.005] + [-0.056, 0.056] \\
&\subseteq \tfrac{5}{2} - x^2 + [0, 0.008] - [-0.002, 0.002] + [-0.056, 0.056] = \tfrac{5}{2} - x^2 + [-0.058, 0.066].
\end{aligned}$$

Hence, we may define

$$\mathcal{U}_1(x) \circ \mathcal{U}_2(x) := \frac{5}{2} - x^2 + [-0.058, 0.066]. \tag{2.2}$$

However, the above computation does not yield a Taylor model for $e^{\cos x}$ for all $x \in \boldsymbol{x}$. Evaluating (2.2) at $x = 0$, we obtain

$$\mathcal{U}_1(0) \circ \mathcal{U}_2(0) = [2.442, 2.566] \not\ni e = e^{\cos 0}.$$

The reason for this failure lies in the range of $\mathcal{U}_2$, which is not contained in $\boldsymbol{x}$. Compositions of Taylor models are indeed computed as above, but it is required that the domain of $\mathcal{U}_1$ contains the range of $\mathcal{U}_2$.

In our example, it suffices to compute the remainder term for the exponential function on the interval $[-1, 1]$. Using Lagrange's representation of the remainder term, we have

$$\frac{e^\xi}{3!}x^3 \in \Big[-\frac{e}{6}, \frac{e}{6}\Big] \subseteq [-0.454, 0.454] \quad \text{for all } \xi \in [-1, 1] \text{ and all } x \in [-1, 1].$$

Using $[-0.454, 0.454]$ instead of $[-0.035, 0.035]$ in the derivation of (2.2) yields

$$\mathcal{U}_1(x) \circ \mathcal{U}_2(x) := \frac{5}{2} - x^2 + [-0.477, 0.485],$$

which is a verified enclosure of $\mathcal{U}_1(x) \circ \mathcal{U}_2(x)$ for all $x \in \boldsymbol{x}$. Note that it is still not a verified enclosure for all $x \in [-1, 1]$. The latter requires that the interval term of $\mathcal{U}_2$ is also computed for $x \in [-1, 1]$.

A *Taylor model vector* is a vector with Taylor model components. When no ambiguity arises, we call a Taylor model vector simply a Taylor model. Arithmetic operations for Taylor model vectors are defined componentwise.

2.3.1. Floating-point Taylor model arithmetic. On a computer with floating-point arithmetic, a Taylor model is defined by a polynomial with machine representable coefficients and a suitable remainder interval that takes account for the roundoff errors. These roundoff errors can occur

- when a function is represented by a Taylor model, or
- when operations between Taylor models are executed.

Example 2.4. *Addition of two univariate floating-point Taylor models.* For simplicity, we use Taylor models of order 1 and a floating-point number system with a mantissa of four decimal digits. Let

$$\boldsymbol{x} := [-1,1], \quad f_1(x) := 1 + x + \frac{1}{8}x^2,\ x \in \boldsymbol{x}, \quad f_2(x) := 1 + \frac{1}{3}x,\ x \in \boldsymbol{x}.$$

Then linear Taylor models for f_1 and f_2 are given by

$$\mathcal{U}_1(x) := 1 + x + [0, 0.125],\ \ \mathcal{U}_2(x) := 1 + 0.3333x + [-0.0001, 0.0001],\ \ x \in \boldsymbol{x}.$$

For $j = 1, 2$, the inclusion condition

$$f_j(x) \in \mathcal{U}_j(x) \ \text{ for all } \ x \in \boldsymbol{x}$$

does not define $\mathcal{U}_1$ and $\mathcal{U}_2$ uniquely. For example,

$$\widetilde{\mathcal{U}}_1(x) := 1 + x + [-0.125, 0.125],\ \ x \in \boldsymbol{x},$$

is also a valid, but less accurate, Taylor model for f_1.

A Taylor model for $f_1 + f_2$ is obtained by performing $\mathcal{U}_1 + \mathcal{U}_2$ with suitable outward rounding. The interval bound for the roundoff error in $x + 0.3333x$ depends on the domain $\boldsymbol{x}$.

$$\begin{aligned}\mathcal{U}_1(x) + \mathcal{U}_2(x) &\subseteq 2 + (x + 0.3333x) + [-0.0001, 0.1251] \\ &\subseteq 2 + (1.333x + [-0.0003, 0.0003]) + [-0.0001, 0.1251] \\ &= 2 + 1.333x + [-0.0004, 0.1254].\end{aligned}$$

A software implementation of Taylor model arithmetic has been developed by Berz and Makino [3, 26] in the COSY INFINITY package [4]. Using COSY INFINITY, Taylor models have been applied with success to a variety of problems, including global optimization [34], verified multidimensional integration [7], and the verified solution of ODEs and DAEs [6, 13].

2.4. Representation of intervals by Taylor models. For a given vector $c \in \mathbb{R}^m$ and a given diagonal matrix $C \in \mathbb{R}^{m \times m}$ with nonnegative diagonal elements, the range of the Taylor model vector

$$\mathcal{U} := c + Cx, \quad x \in \boldsymbol{x}, \tag{2.3}$$

is an m-dimensional interval vector. Vice versa, each interval vector $\boldsymbol{z} \in \mathbb{IR}^m$ can be represented by a Taylor model vector of the form (2.3). There is freedom of choice in selecting c, C, and $\boldsymbol{x}$. A convenient choice is

$$c = \mathrm{m}(\boldsymbol{z}), \quad C = \operatorname{diag}\left(\frac{1}{2}\mathrm{w}(\boldsymbol{z})\right), \quad \boldsymbol{x} = [-1,1]^m,$$

where $[-1,1]^m$ denotes an interval vector with $[-1,1]$ in each component.

Example 2.5. Let $\boldsymbol{z} = ([1,2],[-2,2])^T$. Then we have

$$\boldsymbol{z} = \mathrm{Rg}\left(\begin{pmatrix}\frac{3}{2}\\ 0\end{pmatrix} + \begin{pmatrix}\frac{1}{2} & 0\\ 0 & 2\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}\right), \quad \begin{pmatrix}x\\ y\end{pmatrix} \in [-1,1]^2.$$

3. Interval methods for ODEs.

3.1. Interval IVPs. We consider the smooth interval IVP

$$u' = f(t,u), \quad u(t_0) \in \boldsymbol{u}_0, \quad t \in \boldsymbol{t} = [t_0, t_{\text{end}}], \tag{3.1}$$

where $f : \mathbb{R} \times \mathbb{R}^m \to \mathbb{R}^m$ is a sufficiently smooth function, $\boldsymbol{u}_0 \in \mathbb{IR}^m$ is a given interval vector in the space variables, and $t_{\text{end}} > t_0$ is a given endpoint of the time interval. (The case $t_{\text{end}} < t_0$ is handled similarly.)

While the ODE is defined in the traditional way, the initial value is allowed to vary in the interval $\boldsymbol{u}_0$. In applications, this variability is used for modeling uncertainties in initial conditions. For each $u_0 \in \boldsymbol{u}_0$, the point IVP

$$u' = f(t,u), \quad u(t_0) = u_0$$

has a classical solution, which is denoted by $u(t; t_0, u_0)$. In the following, we assume that $u(t; t_0, u_0)$ exists and is bounded for all $t \in \boldsymbol{t}$ and for all $u_0 \in \boldsymbol{u}_0$.

Our goal when solving (3.1) is to calculate bounds on the flow of the interval IVP. For each $t \in \boldsymbol{t}$, we wish to calculate an interval $\boldsymbol{u}(t)$ such that

$$u(t; t_0, u_0) \in \boldsymbol{u}(t)$$

holds for all $u_0 \in \boldsymbol{u}_0$. The tube $\boldsymbol{u}(t)$, $t \in \boldsymbol{t}$, then contains all solutions of $u' = f(t,u)$ that emerge from $\boldsymbol{u}_0$.

3.2. Interval methods for IVPs. All enclosure methods for ODEs that we are aware of subdivide the domain of integration into subintervals. At each grid point, the flow of the given ODE is enclosed by a set with a certain geometric structure, for example, an m-dimensional rectangle. In the general case, the shape of the flow has a different geometry, so that the flow is wrapped by some larger set, which serves as the initial set for the next time step. To maintain the validity of the method, all solutions of the ODE emerging from the increased initial set must be enclosed in subsequent time steps. The method thus picks up additional solutions of the ODE (that is, solutions not emerging from the original initial set) during the integration process. If the accumulated flow becomes too large, the method may break down because it can no longer compute a sufficiently tight enclosure. It is essential for any verified integration method to minimize the excess introduced by the wrapping of intermediate enclosures of the flow.

In Moore's *direct interval method* [31, 32, 33], the widths of the enclosures at subsequent time steps are always increasing, even for shrinking flows. For linear autonomous ODEs, the direct interval method is only suited for pure contractions. If the flow is rotated, the rotation of the initial set usually provokes exponential growth of the widths of the computed interval enclosures.

In the *parallelepiped method* [32, 33, 12, 21], the flow of the ODE at intermediate time steps is enclosed by parallelepipeds instead of rectangular boxes. This choice is motivated by the shape of the flow of a linear ODE with interval initial values, which is a parallelepiped at any time. For this problem, the only source of overestimation is the remainder interval accounting for the discretization error and the accumulated roundoff errors, if the computation is performed in floating-point arithmetic. These quantities must be enclosed by the final parallelepiped enclosure, but the wrapping affects only small quantities. The algebraic crux of the parallelepiped method is the verified inversion of certain matrices A_j [21, 36], which often tend to become singular

after some time steps, so that the method breaks down either due to excessive wrapping or because the verified matrix inversion is no longer feasible. Hence, breakdown of the parallelepiped method is a rule rather than an exception.

To preserve good condition numbers in the matrices A_j, Lohner [21] developed the *QR method*. His idea was to stabilize the iteration by orthogonalization of the matrices, so that the algebraic problem of inverting the matrices is reduced to taking the transpose.

Various other interval methods have been proposed to fight the wrapping effect, and there are several techniques which are effective in reducing overestimation of the flow for some problem classes [12, 19, 21, 32, 33]. Nevertheless, the ability of interval methods to minimize wrapping is limited by the fact that interval-based enclosure sets are convex. If the flow is a nonconvex set, as may arise for nonlinear ODEs, any interval wrap must be at least as large as the convex hull of the flow.

4. Taylor model methods for ODEs. Taylor model methods use multivariate polynomials in the initial values plus a small interval remainder term to represent the flow of an IVP. Thus, it is possible to work with nonlinear boundary curves, including nonconvex enclosure sets for crescent-shaped or twisted flows. For nonlinear ODEs, this increased flexibility in admissible boundary curves is an intrinsic advantage of Taylor model methods over traditional interval methods, making Taylor model methods very effective in some cases in reducing the wrapping effect.

We refer to the recent paper of Makino and Berz [29] for the general description of Taylor model methods for ODEs. Our intention here is to explain the fundamental difference between interval methods and Taylor model methods with a simple nonlinear example.

4.1. Quadratic model problem. We consider the quadratic model problem

$$\begin{aligned} u' &= v, \qquad u(0) \in [0.95, 1.05], \\ v' &= u^2, \qquad v(0) \in [-1.05, -0.95], \end{aligned} \tag{4.1}$$

where the differentiation is with respect to t. In an interval method, one would use interval initial values $\boldsymbol{u}_0 = [0.95, 1.05]$ and $\boldsymbol{v}_0 = [-1.05, -0.95]$. In the Taylor model method, the initial set is described by parameters, which we call a and b, and which we choose in the interval $[-0.05, 0.05]$. The initial conditions of the IVP (4.1) at $t = t_0$ are thus given by

$$\begin{aligned} u_0(a, b) &:= 1 + a, \qquad a \in \boldsymbol{a} := [-0.05, 0.05], \\ v_0(a, b) &:= -1 + b, \qquad b \in \boldsymbol{b} := [-0.05, 0.05]. \end{aligned}$$

For illustration, we use order $n = 3$ and step size $h = 0.1$ in the Taylor model integration of (4.1). All numbers are displayed here rounded to six decimal digits. In each integration step, the multivariate Taylor series (with respect to t, a, and b) of the solution of (4.1) is employed. The third-order Taylor polynomial serves as an approximate solution. The truncation error of the series is enclosed by a suitable remainder interval.

The first integration step consists of integrating the IVP

$$\begin{aligned} u' &= v, \qquad u(0) = 1 + a, \\ v' &= u^2, \qquad v(0) = -1 + b \end{aligned} \tag{4.2}$$

for $0 \le t \le h$. We use the Picard iteration to calculate a multivariate Taylor polynomial approximation of the solution to (4.2). Using the initial approximations

$$u^{(0)}(\tau, a, b) = 1 + a,$$
$$v^{(0)}(\tau, a, b) = -1 + b$$

(τ is time), the first step of the Picard iteration yields

$$u^{(1)}(\tau, a, b) = u_0(a, b) + \int_0^\tau v^{(0)}(s, a, b)\, ds = 1 + a - \tau + b\tau,$$
$$v^{(1)}(\tau, a, b) = v_0(a, b) + \int_0^\tau \left(u^{(0)}(s, a, b)\right)^2 ds = -1 + b + \tau + 2a\tau + a^2\tau.$$

After two more Picard iterations (and omitting the higher order terms), we obtain the third order Taylor polynomials

$$u^{(3)}(\tau, a, b) = 1 + a - \tau + b\tau + \frac{1}{2}\tau^2 + a\tau^2 - \frac{1}{3}\tau^3,$$
$$v^{(3)}(\tau, a, b) = -1 + b + \tau + 2a\tau - \tau^2 + a^2\tau - a\tau^2 + b\tau^2 + \frac{2}{3}\tau^3,$$

as multivariate approximations to the solution of (4.2). For a verified enclosure of the flow, the Taylor polynomials have to be furnished with suitable remainder bounds. Their derivation is based on a fixed point iteration [24]. Intervals $\boldsymbol{i}_0$ and $\boldsymbol{j}_0$ are sought such that the inclusions

$$u_0 + \int_0^\tau \left(v^{(3)}(s, a, b) + \boldsymbol{j}_0\right) ds \subseteq u^{(3)}(\tau, a, b) + \boldsymbol{i}_0,$$
$$v_0 + \int_0^\tau \left(u^{(3)}(s, a, b) + \boldsymbol{i}_0\right)^2 ds \subseteq v^{(3)}(\tau, a, b) + \boldsymbol{j}_0$$

simultaneously hold for all $a \in \boldsymbol{a}$, for all $b \in \boldsymbol{b}$, and for all $\tau \in [0, 0.1]$. For the details of the computation of the remainder interval, we refer to [24]. In our example, these inclusions are fulfilled, for example, for

$$\boldsymbol{i}_0 = [-5.09307\text{E-}5, 7.86167\text{E-}5] \quad \text{and} \quad \boldsymbol{j}_0 = [-1.75707\text{E-}4, 1.60933\text{E-}4].$$

An enclosure of the flow of the IVP (4.2) for $t \in [0, 0.1]$ is given by the Taylor models

$$\widetilde{\mathcal{U}}_1(\tau, a, b) := 1 + a - \tau + b\tau + \frac{1}{2}\tau^2 + a\tau^2 - \frac{1}{3}\tau^3 + \boldsymbol{i}_0,$$
$$\widetilde{\mathcal{V}}_1(\tau, a, b) := -1 + b + \tau + 2a\tau - \tau^2 + a^2\tau - a\tau^2 + b\tau^2 + \frac{2}{3}\tau^3 + \boldsymbol{j}_0,$$

where a, $b \in [-0.05, 0.05]$, $\tau \in [0, 0.1]$, and $t = \tau$.

Evaluating $\widetilde{\mathcal{U}}_1$ and $\widetilde{\mathcal{V}}_1$ at $\tau = h = 0.1$, we obtain the enclosure of the flow at $t_1 = 0.1$ (Taylor models are of order at most 2 in the space variables):

$$\begin{aligned} \mathcal{U}_1(a, b) &:= \widetilde{\mathcal{U}}_1(0.1, a, b) = 0.904667 + 1.01a + 0.1b + \boldsymbol{i}_0, \\ \mathcal{V}_1(a, b) &:= \widetilde{\mathcal{V}}_1(0.1, a, b) = -0.909333 + 0.19a + 1.01b + 0.1a^2 + \boldsymbol{j}_0, \end{aligned} \tag{4.3}$$

which is the initial set for the second integration step. The latter is performed with a slight modification. We do not use the interval remainder terms in $\mathcal{U}_1$ and $\mathcal{V}_1$ when computing the polynomial part of the Taylor model in the space and time variables. The Picard iteration is again performed for $\tau \in [0, 0.1]$, with initial approximations

$$u^{(0)}(\tau, a, b) = 0.904667 + 1.01a + 0.1b,$$
$$v^{(0)}(\tau, a, b) = -0.909333 + 0.19a + 1.01b + 0.1a^2.$$

After three iterations (and again omitting higher order terms), we obtain

$$\begin{aligned} u^{(3)}(\tau, a, b) = {} & 0.904667 + 1.01a + 0.1b - 0.909333\tau + 0.19a\tau + 1.01b\tau + 0.409211\tau^2 \\ & + 0.1a^2\tau + 0.913713a\tau^2 + 0.0904667b\tau^2 - 0.274215\tau^3, \\ v^{(3)}(\tau, a, b) = {} & -0.909333 + 0.19a + 1.01b + 0.818422\tau + 0.1a^2 + 1.82743a\tau \\ & + 0.180933b\tau - 0.822644\tau^2 \\ & + 1.0201a^2\tau + 0.202ab\tau + 0.01b^2\tau - 0.74654a\tau^2 + 0.82278b\tau^2 \\ & + 0.522429\tau^3. \end{aligned}$$

To compute the interval remainder term, we must find intervals $\boldsymbol{i}_1$ and $\boldsymbol{j}_1$ fulfilling the inclusions

$$\begin{aligned} \mathcal{U}_1(a, b) + \textstyle\int_0^\tau \left(v^{(3)}(s, a, b) + \boldsymbol{j}_1\right) ds \subseteq u^{(3)}(\tau, a, b) + \boldsymbol{i}_1, \\ \mathcal{V}_1(a, b) + \textstyle\int_0^\tau \left(u^{(3)}(s, a, b) + \boldsymbol{i}_1\right)^2 ds \subseteq v^{(3)}(\tau, a, b) + \boldsymbol{j}_1 \end{aligned} \tag{4.4}$$

for all a, $b \in [-0.05, 0.05]$ and for all $\tau \in [0, 0.1]$. (Note that $\boldsymbol{i}_0$ and $\boldsymbol{j}_0$ are contained in $\mathcal{U}_1$ and $\mathcal{V}_1$, respectively, from (4.3).) Suitable remainder intervals are, for example,

$$\boldsymbol{i}_1 = [-1.12850\text{E-}4, 1.65751\text{E-}4], \quad \boldsymbol{j}_1 = [-3.31917\text{E-}4, 3.24724\text{E-}4].$$

Thus, the flow of the IVP (4.2) for $t \in [0.1, 0.2]$ is contained in the Taylor models

$$\widetilde{\mathcal{U}}_2(\tau, a, b) = u^{(3)}(\tau, a, b) + \boldsymbol{i}_1,$$
$$\widetilde{\mathcal{V}}_2(\tau, a, b) = v^{(3)}(\tau, a, b) + \boldsymbol{j}_1,$$

where a, $b \in [-0.05, 0.05]$, $\tau \in [0, 0.1]$, $t = \tau + 0.1$.

Evaluating at $\tau = 0.1$, we obtain the enclosure of the flow at $t_2 = 0.2$ (Taylor models are of order at most 2 in the space variables):

$$\begin{aligned} \mathcal{U}_2(a, b) := \widetilde{\mathcal{U}}_2(0.1, a, b) = {} & 0.817551 + 1.03814a + 0.201905b + 0.01a^2 + \boldsymbol{i}_1, \\ \mathcal{V}_2(a, b) := \widetilde{\mathcal{V}}_2(0.1, a, b) = {} & -0.835195 + 0.365277a + 1.03632b \\ & + 0.20201a^2 + 0.0202ab + 0.001b^2 + \boldsymbol{j}_1. \end{aligned}$$

For larger values of t, the integration can be continued as in the second integration step described above.

Remark 4.1.

1. The sets $(\mathcal{U}_j, \mathcal{V}_j)$ containing the flow of the IVP (4.2) generally become more and more irregular for increasing j. Integration over a larger domain is shown in Figure 6.1.

2. In the above calculations, the polynomial parts of the Taylor models are independent of the initial domain intervals for a and b and independent of the step size h, but the interval remainder bounds are not.
3. The order of the method refers to the order of the multivariate Taylor polynomials with respect to space and time variables that are calculated in the integration step. When the initial sets are defined by linear functions in a and b, then it follows by induction that the maximum order of the polynomials representing the flow at the grid points (obtained after evaluating t) is always at least one less than the order of the method.

In the above example, we have used the so-called *naive* Taylor model integration method to illustrate the qualitative difference of interval methods and Taylor model methods for solving IVPs. For practical computations, the naive Taylor model method is not very useful. The interval remainder terms are propagated as in the direct interval method. The inclusion (4.4) implies that the diameters of the interval remainder terms are nondecreasing. Often, these diameters grow exponentially, and the method breaks down early. More advanced Taylor model integration methods are discussed in the next section. For clarity, we summarize the major steps of the naive Taylor model method as Algorithm 4.1.

Algorithm 4.1 (naive Taylor model method)

Let the initial set be given as a Taylor model vector in m space variables.

For $j := 0, 1 \ldots, j_{\max} - 1$:

1. Compute the Taylor polynomial p_n (of dimension m in $m+1$ variables) of the solution of the $j+1$st time step, using Picard iteration.
2. Compute a remainder interval vector $\boldsymbol{i}$, using Schauder's fixed point theorem (via interval iteration based on Picard iteration).
3. Evaluate $\widetilde{\mathcal{U}} = p_n + \boldsymbol{i}$ at t_{j+1}. The resulting m-dimensional Taylor model $\mathcal{U}$ contains the flow of the IVP and serves as initial set for the next time step.

4.2. Shrink wrapping and preconditioning. For successful integration over long time spans, sophisticated treatment of the interval terms is required. For this purpose, Berz and Makino invented two schemes which they call *shrink wrapping* and *preconditioning*. Shrink wrapping is a method to absorb the interval remainder term into the symbolic part of the Taylor model. From a geometric viewpoint, it resembles the parallelepiped method. Shrink wrapping uses the same linear map as the parallelepiped method so that it has the same limitations when this map becomes ill-conditioned. Preconditioning aims at maintaining a small condition number for the shrink wrapping map. Thus it stabilizes the integration process, like the QR interval method does.

For clarity of the presentation, we describe shrink wrapping and preconditioning for the special case of linear autonomous ODEs. The generalization to nonlinear ODEs is straightforward. We refer to [29] for the details.

5. Taylor model methods for linear ODEs. For a linear ODE, the flow of an interval IVP is a parallelepiped for all time, so Taylor models seem to have no obvious advantage over interval methods. On the other hand, if Taylor model methods failed on linear ODEs, they would probably not be effective for nonlinear ODEs. The purpose of this section is to show that they can be as good as interval methods for linear ODEs.

We consider the linear autonomous ODE

$$\begin{aligned} u' &= B\,u, \\ u(0) &= \mathcal{U}_0, \end{aligned} \tag{5.1}$$

where B is a given real matrix, $\boldsymbol{x}$ is a given interval vector, and $\mathcal{U}_0 = p_n(x)$, $x \in \boldsymbol{x}$, is a Taylor model vector with zero remainder interval describing the initial set. x is used to denote the vector of the space variables. We assume that the enclosure step in the Taylor model method is feasible with some constant step size $h > 0$ and some order $n \in \mathbb{N}$.

5.1. Naive Taylor model method. In the first integration step, Picard iteration of order n is used to compute the multivariate Taylor polynomial

$$u_{1,n} := P_n(tB)\,p_n(x), \quad \text{where} \quad P_n(tB) := \sum_{k=0}^{n} \frac{(tB)^k}{k!}.$$

Introducing $T := P_n(hB)$, the verification step consists of finding an interval vector $\boldsymbol{i}_1$ such that

$$p_n(x) + \int_0^h B\big(P_n(\tau B)\,p_n(x) + \boldsymbol{i}_1\big)\,d\tau \;\subseteq\; P_n(hB)\,p_n(x) + \boldsymbol{i}_1 = T p_n(x) + \boldsymbol{i}_1$$

holds for all $x \in \boldsymbol{x}$ (see, for example, [24, Ch. 6]). At $t_1 = h$, the flow of the IVP (5.1) is then enclosed by the Taylor model

$$\mathcal{U}_1 := T\,p_n(x) + \boldsymbol{i}_1.$$

Subsequent integration steps are performed in the same manner, but with a slight modification in the verification step. In the jth integration step, $j \geq 2$, $\boldsymbol{i}_j$ is sought such that the inclusion

$$T^{j-1}p_n(x) + \boldsymbol{i}_{j-1} + \int_0^h B\left(P_n(\tau B)\,T^{j-1}p_n(x) + \boldsymbol{i}_j\right)\,d\tau \subseteq T^j p_n(x) + \boldsymbol{i}_j$$

is fulfilled for all $x \in \boldsymbol{x}$. Letting

$$\mathcal{U}_j := T\,\mathcal{U}_{j-1} + \boldsymbol{i}_j, \quad j = 1, 2, \ldots,$$

the *naive Taylor model method* for (5.1) consists of the iteration

$$\mathcal{U}_j = T^j\,\mathcal{U}_0 + \sum_{k=1}^{j} (T\circ)^{j-k}\boldsymbol{i}_k, \quad j = 1, 2, \ldots, \tag{5.2}$$

where

$$(T\circ)^0\boldsymbol{x} := \boldsymbol{x}, \quad (T\circ)^k\boldsymbol{x} := T \cdot \big((T\circ)^{k-1}\boldsymbol{x}\big), \quad k \in \mathbb{N}.$$

Apart from the different computation of the remainder interval, for the initial value problem (5.1), the naive Taylor model method (5.2) coincides with the direct interval method that occurs in [36]. Hence, the naive Taylor model method (5.2) has the same divergence property as the direct interval method, for which it was shown in [36] that after j steps we have

$$\mathrm{w}\big((T\circ)^{j-1}\boldsymbol{i}_1\big) \;=\; |T|^{\,j-1}\,\mathrm{w}(\boldsymbol{i}_1)$$

(for $A = (a_{ij})$, we denote by $|A|$ the matrix with components $|a_{ij}|$). The key point here is that the spectral radius of $|T|^{j-1}$ may be much larger than the spectral radius of T^{j-1}, which describes the natural error growth of a point method. If this is the case, the error bounds for the naive Taylor model method may be much larger than the true error.

5.2. Naive Taylor model method with shrink wrapping. Berz and Makino [29] defined shrink wrapping as a method for absorbing the interval part of the Taylor model into the polynomial part by modifying the polynomial coefficients. The set defined by the sum of the given polynomial and interval is wrapped by a set defined by a pure polynomial. The new set may be larger than the initial set, but it is less prone to the dependency problem and to the wrapping effect in succeeding calculations.

In the verified integration of ODEs, shrink wrapping is usually applied to the Taylor model enclosures of the flow at the grid points, before continuing the integration. In practical computations, shrink wrapping is performed when the size of the interval remainder term exceeds some heuristically chosen bound. After shrink wrapping, the initial set of the subsequent integration step is purely symbolic, which removes the dependency problem and simplifies the verification step. The success of the Taylor model based integration method depends on the successful reduction of the excess introduced in the shrink wrapping process.

The process of applying shrink wrapping to a Taylor model vector

$$\mathcal{U} := p(x) + \boldsymbol{i}, \quad x \in \boldsymbol{x},$$

is described in [29]. Here, we outline only its four basic steps. First, let $\widetilde{\mathcal{U}}$ denote the Taylor model that is obtained when the constant part of p is removed. Second, multiply $\widetilde{\mathcal{U}}$ by the inverse of the matrix associated with its linear part and obtain the Taylor model $\widehat{\mathcal{U}}$. Third, estimate the nonlinear part of $\widehat{\mathcal{U}}$, its Jacobian, and the interval term of $\widehat{\mathcal{U}}$ to obtain the shrink wrap factor $q \geq 1$. Fourth, multiply the polynomial part of $\widetilde{\mathcal{U}}$ with q and add the constant part of $\mathcal{U}$.

We illustrate shrink wrapping with the following nonlinear example. For clarity, we use two scalar Taylor models $\mathcal{U}$ and $\mathcal{V}$ instead of a Taylor model vector. The symbolic variables are denoted by a and b (instead of the vector x).

Example 5.1. Absorption of the interval part into the symbolic part of a Taylor model. We consider the Taylor model vector $(\mathcal{U}, \mathcal{V})^T$, where

$$\left.\begin{aligned} \mathcal{U}(a,b) &:= 2 + 4a + \tfrac{1}{2}a^2 + [-0.2, 0.2], \\ \mathcal{V}(a,b) &:= 1 + 3b + ab + [-0.1, 0.1], \end{aligned}\right\} \quad a,\, b \in [-1, 1]. \tag{5.3}$$

The set defined by (5.3) is shown in Figure 5.1. Following the above outline, we obtain

$$\begin{aligned} \widetilde{\mathcal{U}}(a,b) &= 4a + \tfrac{1}{2}a^2 + [-0.2, 0.2], \\ \widetilde{\mathcal{V}}(a,b) &= 3b + ab + [-0.1, 0.1]. \end{aligned} \tag{5.4}$$

The matrix associated with the linear part of the Taylor model (5.4) is

$$C := \begin{pmatrix} 4 & 0 \\ 0 & 3 \end{pmatrix}.$$

Multiplying (5.4) with C^{-1}, we have

$$\begin{aligned} \widehat{\mathcal{U}}(a,b) &= a + \tfrac{1}{8}a^2 + [-0.05, 0.05], \\ \widehat{\mathcal{V}}(a,b) &= b + \tfrac{1}{3}ab + [-0.034, 0.034]. \end{aligned}$$

Estimating the nonlinear part and the interval terms as described in [29], we compute numbers s, t, and d satisfying

$$s \geq \left|\frac{1}{8}a^2\right|,\ s \geq \left|\frac{1}{3}ab\right| \quad \text{for all } a, b \in [-1, 1],$$
$$t \geq \left|\frac{1}{4}a\right|,\ t \geq \left|\frac{1}{3}b\right|,\ t \geq \left|\frac{1}{3}a\right| \quad \text{for all } a, b \in [-1, 1],$$
$$d \geq 0.05,\ d \geq 0.034.$$

These conditions are fulfilled for $s = t = \frac{1}{3}$ and $d = 0.05$, from which we deduce the shrink wrap factor [29]

$$q = 1 + d \cdot \frac{1}{(1-t)(1-s)} = \frac{89}{80}.$$

The final Taylor model after shrink wrapping is

$$\begin{aligned} \mathcal{U}_{\mathrm{sw}}(a, b) &:= 2 + \frac{89}{20}a + \frac{89}{160}a^2, \\ \mathcal{V}_{\mathrm{sw}}(a, b) &:= 1 + \frac{287}{80}b + \frac{89}{80}ab. \end{aligned} \tag{5.5}$$

As Figure 5.1 shows, the set defined by (5.3) is contained in the set defined by (5.5).

Applying shrink wrapping in the linear model problem (5.1) is rather simple. For simplicity, let us assume that shrink wrapping is performed in every integration step. Then we must compute [29] $q_j := 1 + d_j/2$, where

$$d_j := \left\| \mathrm{w}\big((T^j)^{-1}\boldsymbol{i}_j\big) \right\|_\infty .$$

If T is sufficiently well-conditioned, and if the interval terms are sufficiently small, then the factors d_j are almost zero, and shrink wrapping is feasible for many integration steps.

The naive Taylor model method with shrink wrapping resembles the parallelepiped method. By multiplying the nonconstant coefficients of the Taylor polynomial, for linear autonomous ODEs the interval term is absorbed as in the parallelepiped method.

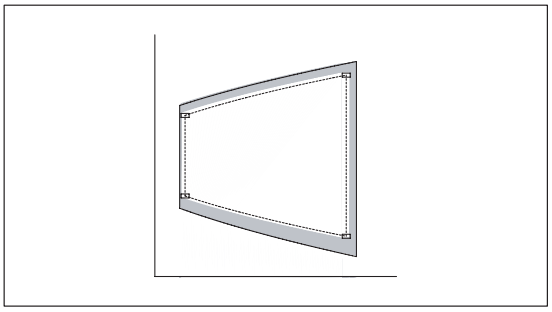


FIG. 5.1. *Sets of the Taylor models before (see* (5.3)*) and after shrink wrapping (see* (5.5)*). The dotted line is the boundary of the set that is described by the polynomial of the original Taylor model. The white area is the set described by the original Taylor model, including the interval term. The excess area introduced by shrink wrapping is shaded in grey.*

While T^j is well-conditioned, d_j is small, and so is the excess area. On the other hand, q_j (and the excess area) becomes large if T^j becomes ill conditioned, which is eventually the case if T has eigenvalues of different magnitude. In this case the integration breaks down due to the growth of the Taylor polynomial coefficients.

The naive Taylor model method with shrink wrapping is outlined as Algorithm 5.1.

Algorithm 5.1 (naive Taylor model method with shrink wrapping)

Let the initial set be given as a Taylor model vector in m space variables.

For $j := 0, 1 \ldots, j_{\max} - 1$:

1. Compute the m-dimensional Taylor model $\mathcal{U} = p_n + \boldsymbol{i}$ (containing the flow of the IVP at t_{j+1}) as in the naive Taylor model method.
2. Absorb $\boldsymbol{i}$ into p_n by shrink wrapping.
3. Continue the integration with the modified polynomial as the initial set for the next time step.

5.3. Preconditioned Taylor models. We showed in the previous section that shrink wrapping has the same limitations as the parallelepiped method in traditional interval arithmetic. To make Taylor model based integration successful for a larger class of IVPs, some stabilization process similar to the QR interval method is required. For restoring good condition numbers of the maps defined by the linear parts of the Taylor models in the integration process, Berz and Makino developed preconditioned Taylor models [29].

In the naive Taylor model method with or without shrink wrapping, the flow of the ODE $u' = f(t, u)$ is represented by a single Taylor model at each grid point. In the preconditioned Taylor model method, the flow of the ODE at $t = t_j$ is represented by a composition of a left and a right Taylor model

$$\mathcal{U}_l \circ \mathcal{U}_r = (p_{l,j} + \boldsymbol{i}_{l,j}) \circ (p_{r,j} + \boldsymbol{i}_{r,j}).$$

DEFINITION 5.2. *The composition*

$$\mathcal{U}(x) := \big(p_l(x) + \boldsymbol{i}_l\big) \circ \big(p_r(x) + \boldsymbol{i}_r\big) \tag{5.6}$$

of two Taylor models

$$\begin{aligned} \mathcal{U}_l(x) &:= p_l(x) + \boldsymbol{i}_l, \ \ x \in \boldsymbol{x}_l, \\ \mathcal{U}_r(x) &:= p_r(x) + \boldsymbol{i}_r, \ x \in \boldsymbol{x}_r, \end{aligned}$$

is called a preconditioned Taylor model *if*

$$\mathrm{Rg}\,(\mathcal{U}_r) \subseteq \boldsymbol{x}_l. \tag{5.7}$$

The range enclosure condition (5.7) is essential in verified integration with preconditioned Taylor models (see discussion below). The factorization into a left and a right Taylor model is not unique. Two preconditioned Taylor models of the form (5.6) can have the same domain $\boldsymbol{z}$ and the same range, but different polynomials and remainder intervals. In verified integration, preconditioning is used to replace some

representation of the flow at an intermediate grid point by a different set of initial values that is more suitable for continuing the integration. Here preconditioning is essentially a substitution in space variables. In the continuation of the integration, the right Taylor model is not involved at all. The following theorem is a reformulation of a proposition given without a proof by Makino and Berz [29].

THEOREM 5.3. *If the initial set of an IVP is given by a preconditioned Taylor model, then integrating the flow of the ODE acts only on the left Taylor model.*

For better understanding of this theorem, which is the key point of the preconditioned integration method, we present first a formal proof, then an example with symbolic integration, and finally a numerical example.

Proof. The space variables are parameters in the integration with respect to time. If $F(x,t)$ is a primitive of $f(x,t)$, that is, if

$$\int f(x,t)\,dt = F(x,t),$$

then substituting $x = g(u)$ does not affect F:

$$\int f(g(u),t)\,dt = F(g(u),t).$$

Preconditioned integration uses $x = (p_{l,j} + \boldsymbol{i}_{l,j})$ and $g(u) = (p_{r,j} + \boldsymbol{i}_{r,j})$. □

Example 5.4. Preconditioned symbolic integration over two time steps. We consider the IVP

$$\begin{aligned} x' &= x(x+y), \quad x(0) = 1+a,\\ y' &= -x(x+y), \quad y(0) = -1+b. \end{aligned}$$

Its unique solution is

$$\begin{aligned} x(t) &= (1+a)e^{(a+b)t},\\ y(t) &= a+b-(1+a)e^{(a+b)t}, \end{aligned}$$

so that at $t = 1$,

$$x(1) = (1+a)e^{a+b}, \quad y(1) = a+b-(1+a)e^{a+b}.$$

To continue the integration, we use the IVP

$$\begin{aligned} u' &= u(u+v), \qquad u(0) = \alpha,\\ v' &= -u(u+v), \quad v(0) = \beta, \end{aligned}$$

and obtain

$$u(1) = \alpha e^{\alpha+\beta}, \quad v(1) = \alpha+\beta-\alpha e^{\alpha+\beta}.$$

Due to the substitution rule, $u(1) = x(2)$ and $v(1) = y(2)$. Indeed, letting

$$\begin{aligned} \alpha &= (1+a)e^{a+b},\\ \beta &= a+b-(1+a)e^{a+b}, \end{aligned}$$

we obtain

$$\begin{aligned} u(1) &= (1+a)e^{2(a+b)} = x(2),\\ v(1) &= (a+b)-(1+a)e^{2(a+b)} = y(2). \end{aligned}$$

The same variable substitution as in Example 5.4 is applied when the initial set for an ODE is given by some preconditioned Taylor model $\mathcal{U}_l \circ \mathcal{U}_r$. To compute an enclosure of the flow, it suffices to integrate the given ODE for the initial values defined by $\mathrm{Rg}\,(\mathcal{U}_l)$, and to compose the integrated Taylor model with $\mathcal{U}_r$. If higher order terms appear in the composition process, they are included in the remainder interval of the result, as in Example 2.2.

In practice, preconditioning is used to replace the integrated preconditioned flow at the end of the jth integration step,

$$\left(\oint \mathcal{U}_{l,j}\right) \circ \mathcal{U}_{r,j}$$

(where $\oint \mathcal{U}$ denotes integrated flow with respect to the given ODE), by a different preconditioned Taylor model

$$\mathcal{U}_{l,j+1} \circ \mathcal{U}_{r,j+1}.$$

The initial set for the $(j+1)$st integration step is defined by $\mathrm{Rg}\,(\mathcal{U}_{l,j+1})$. The method is successful if

- the amount of overestimation in the wrapping of $\left(\oint \mathcal{U}_{l,j}\right) \circ \mathcal{U}_{r,j}$ by $\mathcal{U}_{l,j+1} \circ \mathcal{U}_{r,j+1}$ is sufficiently small, and if
- $\mathrm{Rg}\,(\mathcal{U}_{l,j+1})$ is better suited for continuing the integration than $\oint \mathcal{U}_{l,j}$. For example, preconditioning can be used to reduce the condition number of certain matrices that control the propagation of the global error (see example below) or to reduce the number of nonzero elements in the polynomial part of the left Taylor model.

In Lohner's QR method, an ill-conditioned parallelepiped is wrapped by some well-conditioned m-dimensional rectangle. For preconditioning Taylor models, a large variety of well-conditioned wraps is conceivable. The optimal choice is still an open question for future research.

One important aspect of preconditioned integration is the computation of the remainder bounds in the Picard iteration. If the initial set is given by (5.6), the validity of the enclosure is already guaranteed if the remainder intervals hold for $x \in \mathrm{Rg}\,(\mathcal{U}_r)$. In practice, the remainder bounds are calculated for $x \in \boldsymbol{x}$, a larger set and a potential source of overestimation. In practical computations, overestimation (loss of accuracy) is usually converted to costs (increase of computation time). A common strategy is to limit the admissible size of the remainder intervals by some prescribed bound. Using a larger initial set then has the effect of reducing step sizes and increasing overall computation time.

A simple choice for the left Taylor model (the initial set) in each integration step is a well-conditioned linear map (a parallelepiped). The following description of preconditioned integration is a simplified version of the presentation in [29]. We consider the linear autonomous IVP

$$\begin{aligned} u' &= B\,u, \\ u(0) &= u_0 = c_0 + C_0 x, \end{aligned} \tag{5.8}$$

where B is a real matrix, c_0 is a real vector, C_0 is a diagonal matrix, and x is contained in $[-1,1]^m$. The initial set is given by a Taylor model vector of the form (2.3). A suitable preconditioned Taylor model for this initial set is

$$p_{l,0}(x) = c_0 + C_0 x, \quad \boldsymbol{i}_{l,0} = 0, \quad p_{r,0}(x) = x, \quad \boldsymbol{i}_{r,0} = 0.$$

We assume that the flow at t_j is given by the preconditioned Taylor model

$$\mathcal{U}_j := (p_{l,j} + \boldsymbol{i}_{l,j}) \circ (p_{r,j} + \boldsymbol{i}_{r,j}) = (c_{l,j} + C_{l,j}\, x + \boldsymbol{i}_{l,j}) \circ (c_{r,j} + C_{r,j}\, x + \boldsymbol{i}_{r,j}),$$

where $c_{l,j}$ and $c_{r,j}$ are real vectors, and $C_{l,j}$ and $C_{r,j}$ are real matrices. Using the matrix T from section 5.1, the flow after integration is given by

$$\mathcal{U}_{j+1} := (Tc_{l,j} + TC_{l,j}\, x + \boldsymbol{i}_{l,j+1}) \circ (p_{r,j} + \boldsymbol{i}_{r,j}).$$

For $c_{l,j+1} := Tc_{l,j}$ and any nonsingular matrix $C_{l,j+1}$, the preconditioned Taylor model $\mathcal{U}_{j+1}$ can be rewritten as

$$\begin{aligned}
\mathcal{U}_{j+1} &= (Tc_{l,j} + C_{l,j+1}\, x + [0,0]) \circ \left\{ \left[C_{l,j+1}^{-1} TC_{l,j}\, x + C_{l,j+1}^{-1} \boldsymbol{i}_{l,j+1} \right] \circ (p_{r,j} + \boldsymbol{i}_{r,j}) \right\} \\
&= (c_{l,j+1} + C_{l,j+1}\, x + [0,0]) \circ \Big\{ \left[C_{l,j+1}^{-1} TC_{l,j}\, x + C_{l,j+1}^{-1} \boldsymbol{i}_{l,j+1} \right] \\
&\qquad\qquad \circ (c_{r,j} + C_{r,j}\, x + \boldsymbol{i}_{r,j}) \Big\} \\
&= (c_{l,j+1} + C_{l,j+1}\, x + [0,0]) \circ \left\{ C_{l,j+1}^{-1} TC_{l,j}\ (c_{r,j} + C_{r,j}\, x + \boldsymbol{i}_{r,j}) + C_{l,j+1}^{-1} \boldsymbol{i}_{l,j+1} \right\} \\
&= (c_{l,j+1} + C_{l,j+1}\, x + [0,0]) \\
&\quad \circ \left\{ C_{l,j+1}^{-1} TC_{l,j}\, c_{r,j} + C_{l,j+1}^{-1} TC_{l,j} C_{r,j}\, x + C_{l,j+1}^{-1} TC_{l,j}\, \boldsymbol{i}_{r,j} + C_{l,j+1}^{-1} \boldsymbol{i}_{l,j+1} \right\} \\
&=: (c_{l,j+1} + C_{l,j+1}\, x + [0,0]) \circ (c_{r,j+1} + C_{r,j+1}\, x + \boldsymbol{i}_{r,j+1}).
\end{aligned}$$

The interval term $\boldsymbol{i}_{r,j}$ in the preconditioned Taylor model integration of (5.8) is propagated as the interval term in the parallelepiped and QR interval iteration, if $C_{l,j+1}$ is chosen as in those methods. For $C_{l,j+1} = TC_{l,j}$, the parallelepiped method is obtained, for $TC_{l,j}P_j = Q_j R_j$ (where P_j is a permutation matrix for sorting the columns of $TC_{l,j}$) and $C_{l,j+1} = Q_j$, the QR method. Numerical examples confirming these relations are presented in section 7.

For nonlinear ODEs, the nonlinear terms in the left Taylor model can be shifted to the right Taylor model in the same manner [29]. However, the resulting Taylor model methods then differ from the corresponding interval methods. First, the symbolic parts of the composed Taylor models describe nonlinear enclosures sets of the flow, which need not be convex, in contrast to interval methods. Second, the nonlinear terms in the left Taylor models then also act on the interval terms in the right Taylor models. An analysis of the resulting interval propagation will be the subject of future research.

6. Preconditioned quadratic example. We now demonstrate QR preconditioned Taylor model integration for the quadratic model problem of section 4.1, namely,

$$\begin{aligned}
u' &= v, \qquad u(0) \in [0.95, 1.05], \\
v' &= u^2, \qquad v(0) \in [-1.05, -0.95].
\end{aligned}$$

In each integration step, the left Taylor models are constructed via a QR factorization of the linear parts of the integrated Taylor models of the previous integration step. As in the naive integration of this IVP in section 4.1, order $n = 3$ and step size $h = 0.1$ are used, and all numbers are displayed rounded to six decimal digits.

In the first integration step, the initial set is described by the left Taylor model in space variables at t_0. The right Taylor model at t_0 is the identity map in space variables. Hence, the first integration step is performed as in the naive Taylor model method (cf. section 4.1), and we obtain the integrated left Taylor models (4.3), namely,

$$\left.\begin{aligned} \widetilde{\mathcal{U}}_{l,1}(a,b) &:= 0.904667 + 1.01a + 0.1b + \widetilde{\boldsymbol{i}}_0, \\ \widetilde{\mathcal{V}}_{l,1}(a,b) &:= -0.909333 + 0.19a + 1.01b + 0.1a^2 + \widetilde{\boldsymbol{j}}_0, \end{aligned}\right\} \quad a,\ b \in [-0.05, 0.05],$$

where

$$\widetilde{\boldsymbol{i}}_0 = [-5.09307\text{E-}5, 7.86167\text{E-}5], \quad \widetilde{\boldsymbol{j}}_0 = [-1.75707\text{E-}4, 1.60933\text{E-}4].$$

For reasons that will soon become clear, we normalize the domain such that a and b are contained in $[-1, 1]$. Doing so (without changing the names of the variables), we have

$$\left.\begin{aligned} \widetilde{\mathcal{U}}_{l,1}(a,b) &:= 0.904667 + 0.0505a + 0.005b + \widetilde{\boldsymbol{i}}_0, \\ \widetilde{\mathcal{V}}_{l,1}(a,b) &:= -0.909333 + 0.0095a + 0.0505b + 0.00025a^2 + \widetilde{\boldsymbol{j}}_0, \end{aligned}\right\} \quad a,\ b \in [-1, 1].$$

So far, the right Taylor models have been unaffected by the integration process. Before continuing the integration, however, we precondition the left Taylor models. We extract the linear parts of $\widetilde{\mathcal{U}}_{l,1}$ and $\widetilde{\mathcal{V}}_{l,1}$, and obtain the matrix $C_{l,1}$, from which we compute a QR factorization.

$$\begin{aligned} C_{l,1} := \begin{pmatrix} 0.0505 & 0.005 \\ 0.0095 & 0.0505 \end{pmatrix} &= \begin{pmatrix} 0.982762 & -0.184876 \\ 0.184876 & 0.982762 \end{pmatrix} \cdot \begin{pmatrix} 0.0513858 & 0.0142500 \\ 0 & 0.0487051 \end{pmatrix} \\ &=: QR. \end{aligned}$$

The left Taylor models in the second integration step are built from the constant terms of $\widetilde{\mathcal{U}}_{l,1}$ and $\widetilde{\mathcal{V}}_{l,1}$ and from Q. Thus we get

$$\begin{aligned} \overline{\mathcal{U}}_{l,1}(a,b) &:= 0.904667 + 0.982762a - 0.184876b, \\ \overline{\mathcal{V}}_{l,1}(a,b) &:= -0.909333 + 0.184876a + 0.982762b. \end{aligned}$$

The nonlinear term $0.00025a^2$ in $\widetilde{\mathcal{V}}_{l,1}$ and the interval terms $\widetilde{\boldsymbol{i}}_0$, $\widetilde{\boldsymbol{j}}_0$ are collected in the right Taylor models, which are multiplied by Q^T. We obtain

$$Q^T \cdot \begin{pmatrix} 0 \\ 0.00025a^2 \end{pmatrix} = \begin{pmatrix} 0.0000462190a^2 \\ 0.000245691a^2 \end{pmatrix}$$

and

$$\begin{pmatrix} \overline{\boldsymbol{i}}_0 \\ \overline{\boldsymbol{j}}_0 \end{pmatrix} := Q^T \cdot \begin{pmatrix} \widetilde{\boldsymbol{i}}_0 \\ \widetilde{\boldsymbol{j}}_0 \end{pmatrix} = \begin{pmatrix} [-8.25368\text{E-}5, 1.07014\text{E-}4] \\ [-1.87213\text{E-}4, 1.67575\text{E-}4] \end{pmatrix},$$

which yields

$$\left.\begin{aligned} \overline{\mathcal{U}}_{r,1}(a,b) &:= 0.0513858a + 0.0142500b + 0.0000462190a^2 + \overline{\boldsymbol{i}}_0, \\ \overline{\mathcal{V}}_{r,1}(a,b) &:= 0.0487051b + 0.000245691a^2 + \overline{\boldsymbol{j}}_0, \end{aligned}\right\} \quad a,\ b \in [-1, 1].$$

Before we can continue the integration, we must further modify the preconditioned Taylor models. This is probably the most surprising part of the algorithm. It is also crucial for the validity of the method. After the first time step, the flow of the IVP is contained in the composition of the left and right Taylor models. For continuing the integration, we want to drop the right Taylor model. On one hand, this is only feasible if the left Taylor model contains the flow of the IVP. On the other hand, the set defined by the left Taylor model should not be much larger than the current flow, because that would mean large overestimation. There are two potential solutions for ensuring the desired inclusion property. We can modify the domain of the independent variables, or we may modify the left Taylor model by an additional transformation. We describe both alternatives in the following.

The starting point of the transformation is the range of the right Taylor model. We have

$$\begin{aligned}\mathrm{Rg}\left(\overline{\mathcal{U}}_{r,1}\right) &\subseteq 0.0513858\cdot[-1,1]+0.0142500\cdot[-1,1]+0.0000462190\cdot[0,1]\\ &\qquad +[-8.25368\text{E-}5, 1.07014\text{E-}4]\\ &= [-0.0657183368, 0.065789033]\subseteq[-0.0657183, 0.0657890],\\ \mathrm{Rg}\left(\overline{\mathcal{V}}_{r,1}\right) &\subseteq 0.0487051\cdot[-1,1]+0.000245691\cdot[0,1]+[-1.87213\text{E-}4, 1.67575\text{E-}4]\\ &= [-0.048892151, 0.049118366]\subseteq[-0.0488922, 0.0491184].\end{aligned}$$

Thus we may continue the integration with the initial set for the second time step given by

$$\left.\begin{aligned}\widehat{\mathcal{U}}_{l,1}(a,b) &:= 0.904667+0.982762a-0.184876b,\\ \widehat{\mathcal{V}}_{l,1}(a,b) &:= -0.909333+0.184876a+0.982762b,\end{aligned}\right\}\quad \begin{array}{l}a\in[-0.0657183, 0.0657890],\\ b\in[-0.0488922, 0.0491184]\end{array}$$

(unchanged polynomials, but modified domain).

Alternatively, we can apply a linear transformation on the left and the right Taylor models by a scaling matrix [29]. It is convenient here to denote the linear map (that is, a linear Taylor model $\mathcal{S}$ with zero constant part and zero interval remainder term) associated with a matrix S by the matrix itself. First note that for any nonsingular matrix S,

$$\begin{aligned}(\overline{\mathcal{U}}_{l,1},\overline{\mathcal{V}}_{l,1})\circ(\overline{\mathcal{U}}_{r,1},\overline{\mathcal{V}}_{r,1}) &= (\overline{\mathcal{U}}_{l,1},\overline{\mathcal{V}}_{l,1})\circ(S\circ S^{-1})\circ(\overline{\mathcal{U}}_{r,1},\overline{\mathcal{V}}_{r,1})\\ &\subseteq((\overline{\mathcal{U}}_{l,1},\overline{\mathcal{V}}_{l,1})\circ S)\circ(S^{-1}\circ(\overline{\mathcal{U}}_{r,1},\overline{\mathcal{V}}_{r,1})),\end{aligned}$$

where the subset property is induced by the subdistributivity law of interval arithmetic [1, p. 3]. Letting

$$S := \begin{pmatrix}0.0657890 & 0\\ 0 & 0.0491184\end{pmatrix},$$

we obtain

$$\begin{aligned}(\overline{\mathcal{U}}_{l,1},\overline{\mathcal{V}}_{l,1})\circ S &= \begin{pmatrix}0.904667\\ -0.909333\end{pmatrix}\\ &\quad+\begin{pmatrix}0.982762 & -0.184876\\ 0.184876 & 0.982762\end{pmatrix}\begin{pmatrix}0.0657890 & 0\\ 0 & 0.0491184\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}\\ &= \begin{pmatrix}0.904667\\ -0.909333\end{pmatrix}+\begin{pmatrix}0.0646550 & -0.00908081\\ 0.0121628 & 0.0482716\end{pmatrix}\begin{pmatrix}a\\ b\end{pmatrix}.\end{aligned}$$

Since S has been determined such that the range of each component of $S^{-1}\circ(\overline{\mathcal{U}}_{r,1}, \overline{\mathcal{V}}_{r,1})$ is contained in $[-1,1]$, it is feasible to continue the integration with the left Taylor models

$$\left.\begin{aligned} \mathcal{U}_{l,1}(a,b) &:= 0.904667 + 0.0646550a - 0.00908081b, \\ \mathcal{V}_{l,1}(a,b) &:= -0.909333 + 0.0121628a + 0.0482716b, \end{aligned}\right\} \quad a,\ b \in [-1,1]$$

as initial set for the second time step (modified polynomials, but original domain). The corresponding right Taylor models are

$$\begin{aligned} \begin{pmatrix} \mathcal{U}_{r,1} \\ \mathcal{V}_{r,1} \end{pmatrix} &:= S^{-1} \circ (\overline{\mathcal{U}}_{r,1}, \overline{\mathcal{V}}_{r,1}) \\ &= \begin{pmatrix} 15.2001 & 0 \\ 0 & 20.3590 \end{pmatrix} \begin{pmatrix} 0.0513858a + 0.01425b + 0.000046219a^2 + \overline{\boldsymbol{i}}_0 \\ 0.0487051b + 0.000245691a^2 + \overline{\boldsymbol{j}}_0 \end{pmatrix} \\ &= \begin{pmatrix} 0.781070a + 0.216602b + 0.000702534a^2 + [-0.00125457, 0.00162662] \\ 0.991586b + 0.00500202a^2 + [-0.00381146, 0.00341165] \end{pmatrix}. \end{aligned}$$

Remark 6.1. From a mathematical viewpoint, modification of the domain or of the polynomials are equivalent approaches for factorizing preconditioned Taylor models, but maintaining the integration domain via the scaling matrices is advantageous for the software implementation of the method, because it simplifies the estimation of the higher order terms in the integration step.

In the second integration step, we use the initial set defined by $\mathcal{U}_{l,1}$ and $\mathcal{V}_{l,1}$. Proceeding as before, we obtain the integrated left Taylor models (for $a,\ b \in [-1,1]$)

$$\begin{aligned} \widetilde{\mathcal{U}}_{l,2}(a,b) &:= 0.817551 + 0.0664561a - 0.00433580b + \widetilde{\boldsymbol{i}}_1, \\ \widetilde{\mathcal{V}}_{l,2}(a,b) &:= -0.835195 + 0.0233831a + 0.0471479b \\ &\quad + 0.000418026a^2 - 0.000117424ab + 0.00000824612b^2 + \widetilde{\boldsymbol{j}}_1, \end{aligned}$$

where

$$\widetilde{\boldsymbol{i}}_1 = [-5.72276\text{E-}5, 9.15947\text{E-}5], \quad \widetilde{\boldsymbol{j}}_1 = [-1.80914\text{E-}4, 1.80850\text{E-}4].$$

Finally, the flow at t_2 is made up by the composition of the integrated left Taylor models and the previous right Taylor models. We have

$$\begin{aligned} \mathcal{U}_2(a,b) &:= \widetilde{\mathcal{U}}_{l,2}(\mathcal{U}_{r,1}(a,b), \mathcal{V}_{r,1}(a,b)) = 0.817551 + 0.0519069a + 0.0100952b \\ &\quad + 0.000025a^2 + \ [-3.48708\text{E-}4, 4.09534\text{E-}4], \end{aligned}$$

$$\begin{aligned} \mathcal{V}_2(a,b) &:= \widetilde{\mathcal{V}}_{l,2}(\mathcal{U}_{r,1}(a,b), \mathcal{V}_{r,1}(a,b)) = -0.835195 + 0.0182638a + 0.0518160b \\ &\quad + 0.000507287a^2 - 0.0000505ab - 0.0000025b^2 + [-4.38606\text{E-}4, 4.28392\text{E-}4], \end{aligned}$$

where $a,\ b \in [-1,1]$.

Algorithm 6.1 (QR preconditioned Taylor model method)

Let the initial set be given as a preconditioned Taylor model vector $\mathcal{U}_{l,0} \circ \mathcal{U}_{r,0}$ in m space variables, with $\mathcal{U}_{r,0}$ the identity map and symbolic variables in $[-1, 1]$.

For $j := 0, 1 \ldots, j_{\max} - 1$:

1. Integrate $\mathcal{U}_{l,j}$ (containing the flow of the IVP at t_j) as in the naive Taylor model method. Denote the integrated left Taylor model (containing the flow of the IVP at t_{j+1}) by $\widetilde{\mathcal{U}}_{l,j+1}$. The flow is also contained in $\widetilde{\mathcal{U}}_{l,l+1} \circ \mathcal{U}_{r,j}$.
2. Replace $\widetilde{\mathcal{U}}_{l,j+1} \circ \mathcal{U}_{r,j}$ by $\mathcal{U}_{l,j+1} \circ \mathcal{U}_{r,j+1}$:
   - (i) Compute the QR factorization of the linear part of $\widetilde{\mathcal{U}}_{l,j+1}$.
   - (ii) Shift all but the constant part of $\widetilde{\mathcal{U}}_{l,j+1}$ to $\mathcal{U}_{r,j}$. Make Q the linear part of $\widetilde{\mathcal{U}}_{l,j+1}$. Apply Q^{-1} on $\mathcal{U}_{r,j}$.
   - (iii) Bound the range of the new $\mathcal{U}_{r,j}$.
   - (iv) Apply a scaling matrix S_{j+1} on $\mathcal{U}_{r,j}$ such that each component of the range of $\mathcal{U}_{r,j+1} := S_{j+1}^{-1} \circ \mathcal{U}_{r,j}$ is contained in $[-1, 1]$ and spans $[-1, 1]$ approximately.
   - (v) Set $\mathcal{U}_{l,j+1} := \widetilde{\mathcal{U}}_{l,j+1} \circ S_{j+1}$.

Compared with the naive Taylor model integration performed in section 4.1, the polynomial coefficients are identical except for roundoff errors. This does not invalidate the computations, since all roundoff errors are rigorously bounded by the interval terms. Even though preconditioned integration is the superior method with respect to accuracy in the long run, the interval terms after two integration steps are larger here. The advantage of preconditioning becomes apparent only after several integration steps (see section 6.1). Algorithm 6.1 summarizes the preconditioned Taylor model method with domain normalization.

6.1. Numerical comparison with the QR interval method. Finally, we compare the performance of Lohner's software AWA [21] with the COSY INFINITY integrator written by Makino. We use the quadratic model IVP (4.1) for the comparison. For the computation, Taylor expansions of order 18 were used in both programs. In both programs, the QR method (QR preconditioning) is used. The computed enclosure sets are shown in Figure 6.1.

In the left picture, integration is performed in the time interval $[0, 2.8]$. In the beginning, the enclosures from AWA (rectangular boxes) and COSY INFINITY (nonlinear sets) are of similar quality. Near the end of the integration domain, the enclosures from AWA start exploding. While AWA aborts integration at $t = 3.75$, COSY INFINITY is able to continue the integration much longer (right picture; enclosures of AWA are not shown). We attribute this to the ability of Taylor model methods to use nonconvex enclosure sets of the flow.

This example shows that Taylor model methods may perform much better than interval methods on some problems, but this is not always the case. For some problems, interval methods can be as effective. Moreover, if they succeed, interval methods are often faster than Taylor model methods, because symbolic computations with multivariate polynomials are expensive.

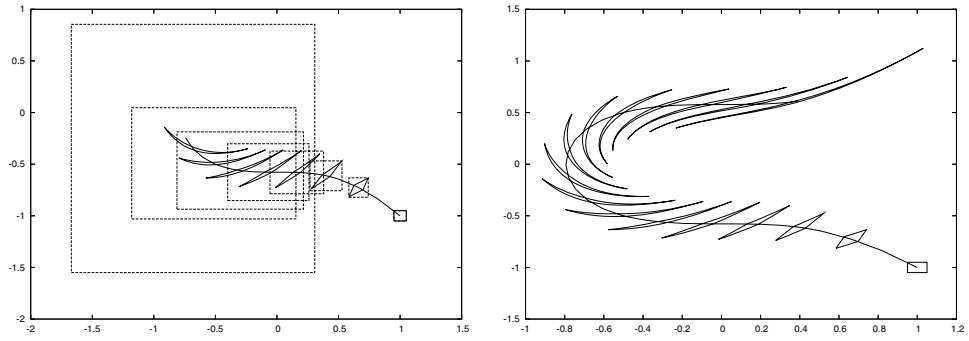


FIG. 6.1. *Integration of quadratic model IVP with AWA and COSY INFINITY for* $t \in [0, 2.8]$ *(left), and with COSY INFINITY for* $t \in [0, 6]$ *(right). Enclosures of the flow are shown for* $t_k = 0.4k$, $k = 0, 1, \ldots$. *The solid line in each picture belongs to the approximate solution that was computed with a Runge–Kutta method (for the model ODE with point initial values).*

7. Linear numerical examples. We compare interval methods and Taylor model methods for the linear autonomous ODE

$$u' = B\,u,$$

where B is a real 3×3 matrix. Numerical results are displayed for three different choices of B. In all examples, the initial values

$$\boldsymbol{u}_0 = \begin{pmatrix} [0.999, 1.001] \\ [0.999, 1.001] \\ [0.999, 1.001] \end{pmatrix}$$

were used. The computations were performed with AWA and with the COSY INFINITY integrator. In all examples, order 12 was chosen for the Taylor polynomial. Using lower orders (6 and 9 were tested) gave less accurate results, and using higher orders (15 was tested) increased the computation times but not the accuracy of the results. For integration with COSY INFINITY, the minimal step size was set to 0.25.

In the tables, the following notation is used.

- AWA iv, AWA pe, and AWA QR denote the direct interval method, the parallelepiped method, and the QR method, respectively.
- TM na, TM sw, and TM QR denote the naive Taylor model method without shrink wrapping, the naive Taylor model method with shrink wrapping, and the Taylor model method with QR preconditioning, respectively.

The observed performance of the methods is in agreement with the theoretical considerations in this paper. Naive Taylor model integration without shrink wrapping performs as the direct interval method (except for Example 1), naive Taylor model integration with shrink wrapping like the parallelepiped method, and QR preconditioned Taylor model integration similar to the QR method.

We call two matrices A and B *floating-point similar* if A is obtained from B by a similarity transform executed in floating-point arithmetic. Floating-point similar matrices are denoted by $A \approx B$. Intervals are sometimes displayed using a short notation with upper and lower indexes. For example, 1.4^{7301}_{5593}E-001 denotes the interval [0.145593,0.147301].

TABLE 7.1
Numerical results for Example 7.1.

Method	t_{end}	Steps	$y_1(t_{\text{end}})$
AWA iv	100	216	1.4^{7301}_{5593}E-001
AWA pe	52.6	131	aborted
AWA QR	100	216	1.4^{7301}_{5593}E-001
TM na	100	391	[−2.378E+26, 2.378E+26]
TM sw	100	272	[−2.282E+112, 2.282E+112]
TM QR	100	122	1.4^{7301}_{5593}E-001

Example 7.1. *Pure contraction (*see Table 7.1*).*

$$B = \begin{pmatrix} -0.4375 & 0.0625 & -0.2651650429 \\ 0.0625 & -0.4375 & -0.2651650429 \\ -0.2651650429 & -0.2651650429 & -0.375 \end{pmatrix} \approx \begin{pmatrix} -\frac{1}{2} & 0 & 0 \\ 0 & -\frac{3}{4} & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

B has three distinct real eigenvalues so that B describes a contraction without rotation. For such problems, the parallelepiped method is not well suited, because the matrices A_j, which have to be inverted, become nearly singular. The interval method breaks down, and the corresponding naive Taylor model method with shrink wrapping computes a practically useless enclosure of the solution.

The direct interval method succeeds here. We also would have expected the naive Taylor model method without shrink wrapping to succeed. While the reason for its failure is not clear, it provides further evidence for our judgement that this method is not very effective. Both the QR interval method and the QR preconditioned Taylor model method succeed here.

TABLE 7.2
Numerical results for Example 7.2.

Method	t	Steps	$y_1(t_{\text{end}})$
AWA iv	76.5	393	aborted
AWA pe	100	449	1.49^{522}_{222}E+000
AWA QR	100	449	1.49^{522}_{222}E+000
TM na	100	396	[−1.517E+45, 1.517E+45]
TM sw	100	343	1.49^{522}_{222}E+000
TM QR	100	343	1.49^{522}_{222}E+000

Example 7.2. *Pure rotation (*see Table 7.2*).*

$$B = \begin{pmatrix} 0 & -0.7071067810 & -0.5 \\ 0.7071067810 & 0 & 0.5 \\ 0.5 & -0.5 & 0 \end{pmatrix} \approx \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

B has eigenvalues $\pm i$ and 0. The flow of this IVP is a rotating interval box. As expected, the direct interval method and the naive Taylor model method fail, whereas the parallelepiped method and the naive Taylor model method with shrink wrapping (and also the QR based methods) succeed.

TABLE 7.3
Numerical results for Example 7.3.

Method	t	Steps	$y_1(t_{\text{end}})$
AWA iv	85.5	507	aborted
AWA pe	75.2	404	aborted
AWA QR	100	516	1.34^{862}_{592}E+000
TM na	100	397	[−1.605E+55, 1.605E+55]
TM sw	100	357	[−3.566E+106, 3.566E+106]
TM QR	100	362	1.34^{862}_{592}E+000

Example 7.3. *Contraction and rotation (*see Table 7.3*).*

$$B = \begin{pmatrix} -0.125 & -0.8321067810 & -0.3232233048 \\ 0.5821067810 & -0.125 & 0.6767766952 \\ 0.6767766952 & -0.3232233048 & -0.25 \end{pmatrix} \approx \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & -\frac{1}{2} \end{pmatrix}$$

In our last example, B has eigenvalues $\pm i$ and $-1/2$, so contraction and rotation are combined. Here, the direct interval method and the naive Taylor model method are bound to fail because of the rotation, whereas the contraction causes the parallelepiped method and the Taylor model method with shrink wrapping to fail.

Only the QR based methods can successfully deal with both contraction and rotation of the initial set. For these methods, the overestimation of the final flow is hardly noticeable. This agrees with the general observation that the QR decomposition is a very effective tool in fighting the wrapping effect, both for the interval method and for the preconditioned Taylor model method.

Conclusion. We have compared traditional enclosure methods with Taylor model based integration. For the verified solution of initial value problems for ODEs, we have shown how Taylor model methods benefit from symbolic computations. Increased flexibility in admissible boundary curves of enclosures is an intrinsic advantage over traditional interval methods, and not only for the solution of ODEs. In future research, we hope to contribute to the further development and increased use of Taylor model methods.

Acknowledgments. The authors thank Martin Berz and Kyoko Makino for several very helpful clarifying discussions on Taylor models, and for making the COSY INFINITY integrator available. Our thanks also go to the referees for helpful comments.

REFERENCES

[1] G. ALEFELD AND J. HERZBERGER, *Introduction to Interval Computations*, Academic Press, New York, 1983.
[2] M. BERZ, *From Taylor series to Taylor models*, in Beam Stability and Nonlinear Dynamics, AIP Conf. Proc. 405, American Institute of Physics, Melville, NY, East Lansing, MI, 1997, pp. 1–23.
[3] M. BERZ, *Cosy Infinity Version* 8 *Reference Manual*, NSCL Technical report MSUCL-1088, Michigan State University, East Lansing, MI, 1998.
[4] M. BERZ, *COSY INFINITY*, available online at http://bt.pa.msu.edu/index_files/cosy.html, 2006.
[5] M. BERZ AND G. HOFFSTÄTTER, *Computation and application of Taylor polynomials with interval remainder bounds*, Reliab. Comput., 4 (1998), pp. 83–97.

[6] M. Berz and K. Makino, *Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor models*, Reliab. Comput., 4 (1998), pp. 361–369.
[7] M. Berz and K. Makino, *New methods for high-dimensional verified quadrature*, Reliab. Comput., 5 (1999), pp. 13–22.
[8] M. Berz and K. Makino, *Performance of Taylor model methods for validated integration of ODEs*, in On the Approximation of Interval Functions, Lecture Notes in Comput. Sci. 3732, Springer-Verlag, New York, 2006, pp. 65–73.
[9] G. F. Corliss, *Survey of interval algorithms for ordinary differential equations*, Appl. Math. Comput., 31 (1989), pp. 112–120.
[10] G. F. Corliss, *Guaranteed error bounds for ordinary differential equations*, in Theory and Numerics of Ordinary and Partial Differential Equations, M. Ainsworth, J. Levesley, W. A. Light, and M. Marletta, eds., Clarendon Press, Oxford, UK, 1995.
[11] J.-P. Eckmann, H. Koch, and P. Wittwer, *A computer-assisted proof of universality in area-preserving maps*, Mem. Amer. Math. Soc., 47 (1984).
[12] P. Eijgenraam, *The Solution of Initial Value Problems Using Interval Arithmetic*, Mathematical Centre Tracts 144, Stichting Mathematisch Centrum, Amsterdam, The Netherlands, 1981.
[13] J. Hoefkens, M. Berz, and K. Makino, *Verified high-order integration of DAEs and higher-order ODEs*, in Scientific Computing, Validated Numerics, and Interval Methods, W. Krämer and J. Wolff von Gudenberg, eds., Kluwer, Dordrecht, The Netherlands, 2001, pp. 281–292.
[14] L. Jaulin, M. Kieffer, O. Didrit, and E. Walter, *Applied Interval Analysis*, Springer-Verlag, London, 2001.
[15] E. Kaucher, *Self-validating computation of ordinary and partial differential equations*, in Computerarithmetic: Scientific Computation and Programming Languages, E. Kaucher, U. Kulisch, and Ch. Ullrich, eds., Teubner, Stuttgart, 1987, pp. 221–254.
[16] E. W. Kaucher and W. L. Miranker, *Self-Validating Numerics for Function Space Problems*, Academic Press, New York, 1984.
[17] R. Klatte, U. Kulisch, Ch. Lawo, M. Rauch, and A. Wiethoff, *C-XSC: A C++ Class Library for Extended Scientific Computing*, Springer-Verlag, Berlin, 1993.
[18] R. Klatte, U. Kulisch, M. Neaga, D. Ratz, and Ch. Ullrich, *Pascal-XSC—Language Reference with Examples*, Springer-Verlag, Berlin, 1992.
[19] W. Kühn, *Rigorously computed orbits of dynamical systems without the wrapping effect*, Computing, 61 (1998), pp. 47–67.
[20] U. Kulisch and W. L. Miranker, *Computer Arithmetic in Theory and Practice*, Academic Press, New York, 1981.
[21] R. Lohner, *Einschließung der Lösung gewöhnlicher Anfangs- und Randwertaufgaben und Anwendungen*, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, Germany, 1988.
[22] R. Lohner, *Computation of guaranteed solutions of ordinary initial and boundary value problems*, in Computational Ordinary Differential Equations, J. R. Cash and I. Gladwell, eds., Clarendon Press, Oxford, UK, 1992, pp. 425–435.
[23] R. Lohner, *On the ubiquity of the wrapping effect in the computation of error bounds*, in Perspectives of Enclosure Methods, U. Kulisch, R. Lohner, and A. Facius, eds., Springer-Verlag, Vienna, 2001, pp. 201–217.
[24] K. Makino, *Rigorous Analysis of Nonlinear Motion in Particle Accelerators*, Ph.D. thesis, Michigan State University, East Lansing, MI, 1998.
[25] K. Makino and M. Berz, *Remainder differential algebras and their applications*, in Computational Differentiation: Techniques, Applications, and Tools, M. Berz, C. Bischof, G. Corliss, and A. Griewank, eds., SIAM, Philadelphia, 1996, pp. 63–74.
[26] K. Makino and M. Berz, *COSY INFINITY version* 8, Nuclear Instruments and Methods in Physics Research A, 427 (1999), pp. 338–343.
[27] K. Makino and M. Berz, *Efficient control of the dependency problem based on Taylor model methods*, Reliab. Comput., 5 (1999), pp. 3–12.
[28] K. Makino and M. Berz, *Taylor models and other validated functional inclusion methods*, Int. J. Pure Appl. Math., 4 (2003), pp. 379–456.
[29] K. Makino and M. Berz, *Suppression of the Wrapping Effect by Taylor Model-Based Validated Integrators*, MSU Report MSUHEP 40910, Michigan State University, East Lansing, MI, 2004.
[30] K. Makino, M. Berz, and Y. Kim, *Range bounding with Taylor models—some case studies*, WSEAS Transactions on Mathematics, 3 (2004), pp. 137–145.

[31] R. E. Moore, *The automatic analysis and control of error in digital computation based on the use of interval numbers*, in Error in Digital Computation, Vol. I, L. B. Rall, ed., John Wiley and Sons, New York, 1965, pp. 61–130.
[32] R. E. Moore, *Automatic local coordinate transformations to reduce the growth of error bounds in interval computation of solutions of ordinary differential equations*, in Error in Digital Computation, Vol. II, L. B. Rall, ed., John Wiley and Sons, New York, 1965, pp. 103–140.
[33] R. E. Moore, *Interval Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1966.
[34] P. S. V. Nataraj and K. Kotecha, *Global optimization with higher order inclusion function forms. Part* 1: *A combined Taylor-Bernstein form*, Reliab. Comput., 10 (2004), pp. 27–44.
[35] N. S. Nedialkov, *Computing Rigorous Bounds on the Solution of an IVP for an ODE*, Ph.D. thesis, University of Toronto, Toronto, Canada, 1999.
[36] N. S. Nedialkov and K. R. Jackson, *A new perspective on the wrapping effect in interval methods for initial value problems for ordinary differential equations*, in Perspectives of Enclosure Methods, U. Kulisch, R. Lohner, and A. Facius, eds., Springer-Verlag, Vienna, 2001, pp. 219–264.
[37] N. S. Nedialkov and K. R. Jackson, *Some recent advances in validated methods for IVPs for ODEs*, Appl. Numer. Math., 42 (2003), pp. 269–284.
[38] N. S. Nedialkov, K. R. Jackson, and G. F. Corliss, *Validated solutions of initial value problems for ordinary differential equations*, Appl. Math. Comput., 105 (1999), pp. 21–68.
[39] N. S. Nedialkov, K. R. Jackson, and J. Pryce, *An effective high-order interval method for validating existence and uniqueness of the solution of an IVP for an ODE*, Reliab. Comput., 7 (2001), pp. 449–465.
[40] M. Neher, *Geometric series bounds for the local errors of Taylor methods for linear nth order ODEs*, in Symbolic Algebraic Methods and Verification Methods, G. Alefeld, J. Rohn, S. Rump, and T. Yamamoto, eds., Springer-Verlag, Vienna, 2001, pp. 183–193.
[41] A. Neumaier, *Interval Methods for Systems of Equations*, Cambridge University Press, Cambridge, UK, 1990.
[42] A. Neumaier, *Taylor forms—use and limits*, Reliab. Comput., 9 (2003), pp. 43–79.
[43] H. Ratschek and J. Rokne, *Computer Methods for the Range of Functions*, Ellis Horwood Limited, Chichester, UK, 1984.
[44] R. Rihm, *Über Einschließungsverfahren für gewöhnliche Anfangswertprobleme und ihre Anwendung auf Differentialgleichungen mit unstetiger rechter Seite*, Ph.D. thesis, Universität Karlsruhe, Karlsruhe, Germany, 1993.
[45] R. Rihm, *Interval methods for initial value problems in ODEs*, in Topics in Validated Computations, J. Herzberger, ed., Elsevier, Amsterdam, The Netherlands, 1994, pp. 173–207.
[46] H. J. Stetter, *Validated solution of initial value problems for ODE*, Notes Rep. Math. Sci. Engrg., 7 (1990), pp. 171–193.
[47] N. F. Stewart, *A heuristic to reduce the wrapping effect in the numerical solution of* $x' = f(t, x)$, BIT, 11 (1971), pp. 328–337.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 263–276

CONDITION ESTIMATES FOR PSEUDO-ARCLENGTH CONTINUATION*

K. I. DICKSON†, C. T. KELLEY†, I. C. F. IPSEN†, AND I. G. KEVREKIDIS‡

Abstract. We bound the condition number of the Jacobian in pseudo-arclength continuation problems, and we quantify the effect of this condition number on the linear system solution in a Newton-GMRES solve. Pseudo-arclength continuation solves parameter dependent nonlinear equations $G(u, \lambda) = 0$ by introducing a new parameter s, which approximates arclength, and viewing the vector $x = (u, \lambda)$ as a function of s. In this way simple fold singularities can be computed directly by solving a larger system $F(x, s) = 0$ by simple continuation in the new parameter s. It is known that the Jacobian F_x of F with respect to $x = (u, \lambda)$ is nonsingular if the path contains only regular points and simple fold singularities. We introduce a new characterization of simple folds in terms of the singular value decomposition, and we use it to derive a new bound for the norm of F_x^{-1}. We also show that the convergence rate of GMRES in a Newton step for $F(x, s) = 0$ is essentially the same as that of the original problem $G(u, \lambda) = 0$. In particular, we prove that the bounds on the degrees of the minimal polynomials of the Jacobians F_x and G_u differ by at most 2. We illustrate the effectiveness of our bounds with an example from radiative transfer theory.

Key words. pseudo-arclength continuation, singularity, GMRES, singular vectors, eigenvalues, rank-one update, turning point, simple fold, fold point, limit point

AMS subject classifications. 65H10, 65H17, 65H20, 65F10, 65F15

DOI. 10.1137/060654384

1. Introduction.

Numerical continuation is the process of solving systems of nonlinear equations $G(u, \lambda) = 0$ for various values of a real parameter λ. Here $u \in R^N$, λ is a real scalar, and $G : R^{N+1} \to R^N$. An obvious approach for implementing numerical continuation, called *parameter continuation* [11, 13, 19], traces out a solution path by repeatedly incrementing λ until the desired value of λ is reached. In each such iteration, the current solution u is used as an initial iterate for the next value of λ. Although parameter continuation is simple and intuitive, it fails at points (u, λ) where the Jacobian G_u is singular. In this paper we consider singularities which are simple folds.

The standard way to remedy the failure of parameter continuation at simple folds is to reparameterize the problem by introducing an approximate arclength parameter, s, so that both u and λ depend on s. This idea, known as *pseudo-arclength continuation* [11, 13, 19], introduces a new parameter s and treats the vector $x = (u, \lambda)$ as a function of s. We then solve a new system $F(x, s) = 0$ by parameter continuation in s. In order for this approach to succeed, the Jacobian F_x of F must be nonsingular. It is known that F_x is nonsingular at simple folds and points where G_u is nonsingular [13].

*Received by the editors March 15, 2006; accepted for publication (in revised form) August 29, 2006; published electronically January 22, 2007.
http://www.siam.org/journals/sinum/45-1/65438.html

†Center for Research in Scientific Computation and Department of Mathematics, North Carolina State University, Box 8205, Raleigh, NC 27695-8205 (kidickso@unity.ncsu.edu, Tim_Kelley@ncsu.edu, ipsen@math.ncsu.edu). The work of these authors has been partially supported by National Science Foundation grants DMS-0404537 and DMS-0209695, and Army Research Office grants DAAD19-01-1-0592, W911NF-04-1-0276, W911NF-05-1-0171, and W911NF-06-1-0096.

‡Department of Chemical Engineering and PACM, Princeton University, Princeton, NJ 08544 (yannis@princeton.edu). The work of this author was supported in part by AFOSR and an NSF/ITR grant.

Our first goal (section 3) is to quantify this nonsingularity. To this end we provide a new characterization of simple folds in terms of the singular value decomposition (SVD) of G_u. From the SVD, we derive a new bound for $\|F_x^{-1}\|_2$. This bound can be used to limit the arclength step in Newton's method. As a byproduct we obtain a refinement of Weyl's monotonicity theorem [23] for the smallest eigenvalue of a symmetric positive semidefinite matrix (section 3.1).

We also examine in section 4 how the conditioning of F_x affects the convergence of the inner GMRES [26] iteration in a Newton-GMRES solver [2, 3, 14, 15]. We show that the eigenvalue clustering of the Jacobian F_x in the reformulated problem is not much different from that of the Jacobian G_u in the original problem. This implies [4, 17] that the convergence speed of GMRES, when used as a linear solver for the Newton step, is not degraded when parameter continuation is replaced by pseudo-arclength continuation.

Finally, in section 5, we illustrate our findings with a numerical example from radiative transfer theory.

2. Background. We briefly review theory and algorithms for solving numerical continuation problems $G(u, \lambda) = 0$, where $\lambda \in R$, $u \in R^N$, and $G : R^{N+1} \to R^N$. We discuss parameter continuation in section 2.1 and pseudo-arclength continuation in section 2.2. We use the abbreviations

$$G_u \equiv \frac{\partial G}{\partial u}, \qquad G_\lambda \equiv \frac{\partial G}{\partial \lambda}.$$

2.1. Simple parameter continuation. Parameter continuation [11, 13, 19] is the simplest method for solving $G(u, \lambda) = 0$. The idea is to start at a point $\lambda = \lambda_{init}$ and solve $G(u, \lambda) = 0$ for $u(\lambda)$, say, by Newton's method. Use the solution $u(\lambda)$ as the initial iterate to solve the next problem $G(u, \lambda + d\lambda) = 0$. Algorithm **paramc** below is a simple implementation of parameter continuation from λ_{init} to $\lambda_{end} = \lambda_{init} + n\, d\lambda$ where n denotes the maximum number of continuation iterations.

---

paramc$(u, G, \lambda_{init}, \lambda_{end}, d\lambda)$
  Set $\lambda = \lambda_{init}$, $u_0 = u$
  while $\lambda \leq \lambda_{end}$ **do**
    Solve $G(u, \lambda) = 0$ with u_0 as the initial iterate to obtain $u(\lambda)$
    $u_0 = u(\lambda)$
    $\lambda = \lambda + d\lambda$
  end while

Corollary 2.1 is a consequence of the implicit function theorem [13, 22] and states that parameter continuation, as realized in Algorithm **paramc**, will succeed near a solution at which G_u is nonsingular. Parameter continuation may fail if the arc of solutions contains singular points, i.e., solutions at which G_u is singular.

COROLLARY 2.1. *Let G be Lipschitz continuously differentiable, $G(u_0, \lambda_0) = 0$, and $G_u(u_0, \lambda_0)$ be nonsingular. Then there is $\delta > 0$, which depends only on $\|G_u^{-1}(u_0, \lambda_0)\|$ and the Lipschitz constants of G_u and G_λ, such that if $|\lambda - \lambda_0| < \delta$ then Newton's method with initial iterate u_0 converges q-quadratically to the solution $u(\lambda)$ of $G(u, \lambda) = 0$, i.e.,*

$$\|u_{n+1} - u(\lambda)\| = O(\|u_n - u(\lambda)\|^2), \tag{2.1}$$

where, for $n \geq 0$,

$$u_{n+1} = u_n - G_u(u_n, \lambda)^{-1} G(u_n, \lambda).$$

Proof. Define the Lipschitz constant

$$\|G_u(u, \lambda) - G_u(v, \mu)\| \leq \gamma_G(\|u - v\| + |\lambda - \mu|).$$

Differentiating $G(u, \lambda) = 0$ with respect to λ gives

$$du/d\lambda = -G_u^{-1} G_\lambda.$$

The implicit function theorem implies that there is δ_1 such that if

$$|\lambda - \lambda_0| \leq \delta_1$$

then there is a solution arc $u(\lambda)$ defined for $|\lambda - \lambda_0| \leq \delta_1$. Since $G_u^{-1} G_\lambda$ is Lipschitz continuous, there is γ_u, which depends only on $\|G_u^{-1}(u_0, \lambda_0)\|$ and the Lipschitz constants of G_u and G_λ, such that

$$\|du/d\lambda\| = \|G_u^{-1} G_\lambda\| \leq \gamma_u.$$

A lower bound for the radius of the ball of attraction for the Newton iteration is [14]

$$\frac{1}{2\gamma_G \|G_u^{-1}(u_0, \lambda_0)\|},$$

so choosing

$$\delta = \min\left(\delta_1, \frac{1}{2\gamma_u \gamma_G \|G_u^{-1}(u_0, \lambda_0)\|}\right)$$

completes the proof. ∎

The implicit function theorem and Corollary 2.1 fail near most singular points. Our objective in this paper is to investigate the simplest class of singular points at which the implicit function theorem fails.

2.2. Pseudo-arclength continuation. Pseudo-arclength continuation [11, 13, 19] avoids the problems of Algorithm **paramc** at singular points by using an approximation of arclength parameterization. The curve in Figure 5.1, for instance, has a singularity with respect to the parameter λ. If we choose arclength s as the parameter λ, and $x = (u^T, \lambda)^T$ in place of u, we can compute the curve with simple parameter continuation. The curve in Figure 5.1 has a simple fold, which is the singularity of interest for this paper. Formally, a simple fold (or fold point, turning point, or limit point) is defined as follows [5, 13, 20, 24].

DEFINITION 2.2. *A solution* (u_0, λ_0) *of* $G(u, \lambda) = 0$ *is a simple fold if*

- $\dim(\mathrm{Ker}(G_u(u_0, \lambda_0))) = 1$ *and*
- $G_\lambda(u_0, \lambda_0) \notin \mathrm{Range}(G_u(u_0, \lambda_0))$.

To develop a pseudo-arclength continuation method, we assume that x depends smoothly on s. Then one can differentiate $G(u, \lambda) = 0$ with respect to s and obtain

$$\frac{dG(u(s), \lambda(s))}{ds} = G_u \dot{u} + G_\lambda \dot{\lambda} = 0. \tag{2.2}$$

Equivalently, one can differentiate $G(x) = 0$ and obtain $G_x\dot{x} = 0$. Here, $\dot{x}$ denotes the derivative with respect to s. Since s is arclength in the Euclidean norm,

$$\|\dot{x}\|^2 = \|\dot{u}\|^2 + |\dot{\lambda}|^2 = 1. \tag{2.3}$$

Having introduced a new parameter s, one adds an equation to $G(u, \lambda) = 0$ so that the number of equations equals the number of unknowns. To do this one introduces the extended system

$$F(x, s) = \begin{pmatrix} G(x) \\ \mathcal{N}(x, s) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{2.4}$$

The normalization equation $\mathcal{N} = 0$ is an approximation of (2.3) where

$$\mathcal{N}(x, s) = \dot{x}_0^T(x - x_0) - (s - s_0) = 0. \tag{2.5}$$

Equation (2.5) says that the new point on the path lies on a hyperplane orthogonal to the tangent vector through the current point x_0, and the intersection of that hyperplane with the tangent vector is a distance $ds = s - s_0$ from x_0.

While we prove our results using the normalization (2.5), the bounds also apply to other normalizations [7, 11, 13, 25, 27], which are asymptotically equivalent to (2.5).

Given a known point (x_0, s_0), the pseudo-arclength continuation method increments arclength by ds, and solves (2.4) with the normalization (2.5) by Newton's method with initial iterate x_0. Algorithm **psarc** is a simple implementation of pseudo-arclength continuation.

---

psarc(u, F, s_{end}, ds)
  Set $s = 0$, $x_0 = (u_0^T, \lambda_0)^T$
  while $s \leq s_{end}$ **do**
    Approximate $\dot{x}$
    Solve $F(x, s) = 0$ with fixed s and x_0 as the initial iterate to obtain $x(s)$
    $x_0 = x(s)$
    $s = s + ds$
  end while

Since pseudo-arclength continuation is just simple parameter continuation applied to F with s as the parameter, Corollary 2.1 gives conditions for the convergence of Newton's method in pseudo-arclength continuation, and we restate the corollary in terms of F for completeness.

COROLLARY 2.3. *Let the assumptions of Corollary* 2.1 *hold for* F. *Then there is* $\delta > 0$, *which depends only on* $\|F_x^{-1}(x_0, s_0)\|$ *and the Lipschitz constants of* F_x *and* x, *such that if* $|s - s_0| < \delta$ *then Newton's method with initial iterate* x_0 *converges q-quadratically to the solution.*

One consequence of Corollary 2.3 is that a bound on $\|F_x^{-1}\|$ is an important factor in bounding the arclength step. In the next section we present the main result of this paper, a new bound on $\|F_x^{-1}\|$.

3. Nonsingularity of F_x. For a solution $x_0 = (u_0, \lambda_0)$ to $G(u, \lambda) = 0$, we present an upper bound on $\|F_x^{-1}(x_0, s_0)\|$ in the case that

- $G_u(u_0, \lambda_0)$ is nonsingular or
- (u_0, λ_0) is a simple fold of $G(u, \lambda) = 0$.

In order to derive the bound, we introduce a new characterization of simple folds, which is based on the SVD of G_u. We prove the bound in section 3.2. In section 3.1 we refine Weyl's monotonicity theorem for the smallest eigenvalue of a symmetric positive semidefinite matrix, which we need for the proof.

Let

$$G_u(u,\lambda) = U\Sigma V^T$$

be an SVD of $G_u(u,\lambda)$, where

$$\Sigma = \mathrm{diag}(\sigma_1,\sigma_2,\ldots,\sigma_N), \qquad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_N, \qquad u_N \equiv Ue_N,$$

and e_N is the last column of the $N \times N$ identity matrix. The trailing column u_N of U is a left singular vector associated with the smallest singular value σ_N. When necessary, we will make the dependence on λ or u explicit, by writing, for example, $\sigma_N(u,\lambda)$ or $u_N(u,\lambda)$.

Since the singular values are continuous functions of the elements in $G_u(u,\lambda)$, they are also continuous in λ. If

$$\sigma_{N-1} \geq \bar{\sigma} > 0$$

for all (u,λ) then the nullity of $G_u(u,\lambda)$ is at most one. If in addition $\sigma_N = 0$ then u_N spans the left nullspace of $G_u(u,\lambda)$. From the direct sum

$$\mathrm{Ker}(G_u^T(u_0,\lambda_0)) \oplus \mathrm{Range}(G_u(u_0,\lambda_0)) = R^N$$

we see that $G_\lambda(u_0,\lambda_0)$ is not in the Range $(G_u(u_0,\lambda_0))$ if and only if $G_\lambda(u_0,\lambda_0)^T u_N \neq 0$. Hence we have a new, equivalent definition of a simple fold.

DEFINITION 3.1 (simple fold via SVD). *Let (u_0,λ_0) be a solution of $G(u,\lambda)=0$, and let $u_N(u_0,\lambda_0)$ be a left singular vector of $G_u(u_0,\lambda_0)$ associated with σ_N.*

Then (u_0,λ_0) is a simple fold if

- $\sigma_{N-1}(u_0,\lambda_0) > 0$,
- $\sigma_N = 0$, *and*
- $u_N(u_0,\lambda_0)^T G_\lambda(u_0,\lambda_0) \neq 0$.

We will use Definition 3.1 to motivate the assumptions in Theorem 3.2. Suppose (u_0,λ_0) is a regular point (G_u nonsingular) or a simple fold. Since G is Lipschitz continuously differentiable, we can, by requiring $u_N^T(\lambda_0)u_N(\lambda) > 0$, for example, define u_N as a continuous function of u and λ. Hence $G_\lambda(u,\lambda)^T u_N(u,\lambda)$ is a continuous function of (u,λ). So there is $\alpha > 0$ such that for all (u,λ) sufficiently near (u_0,λ_0),

$$\max\left(\sigma_N(u,\lambda)^2, |u_N(u,\lambda)^T G_\lambda(u,\lambda)|^2 \frac{\mathrm{gap}}{\mathrm{gap}+\xi^2}\right) \geq \alpha > 0, \tag{3.1}$$

where

$$\mathrm{gap} \equiv \sigma_{N-1}(u,\lambda)^2 - \sigma_N(u,\lambda)^2 \tag{3.2}$$

and

$$\xi \equiv |u_N(u,\lambda)^T G_\lambda(u,\lambda)| + \|(I - u_N(u,\lambda)u_N(u,\lambda)^T)G_\lambda(u,\lambda)\|. \tag{3.3}$$

Inequality (3.1) is a way to quantify the statement that all points on a solution arc are either regular points or simple folds by saying that either $\sigma_N > 0$ (regular point) or the conditions in Definition 3.1 hold.

The main result of this paper is the following theorem.

THEOREM 3.2. *Let $\bar{\Omega}$ be the closure of an open subset $\Omega \in R^{N+1}$, and let G be continuously differentiable in $\bar{\Omega}$. Let $x_0 = (u_0, \lambda_0)$ in $\bar{\Omega}$ be a solution to $G(u_0, \lambda_0) = 0$, and $\mathcal{N}(x_0, s_0) = 0$ with $\|\dot{x}_0\| = 1$. Let $\tau \geq 0$ be such that $\|G_u(u_0, \lambda_0)\dot{u}_0 + G_\lambda(u_0, \lambda_0)\dot{\lambda}_0\| \leq \tau$.*

Assume that for all (u, λ) in $\bar{\Omega}$ there exists $\alpha > 0$ such that

$$\sigma_{N-1}(u, \lambda) > 0 \text{ and } \max\left(\sigma_N(u, \lambda)^2, |u_N(u, \lambda)^T G_\lambda(u, \lambda)|^2 \frac{\text{gap}}{\text{gap} + \xi^2}\right) \geq \alpha,$$

where gap *and ξ are defined by* (3.2) *and* (3.3).

If $\tau < \alpha$, then for all $x = (u, \lambda)$ in $\bar{\Omega}$, the smallest singular value $\sigma_{\min}(F_x)$ of the Jacobian F_x of $F(x, s)$ is bounded from below with

$$\sigma_{\min}(F_x) \geq \sqrt{1 - \tau \max\left\{\frac{1}{\alpha}, 1\right\}}.$$

We postpone the proof of Theorem 3.2 until section 3.1 in order to derive an auxiliary result first.

3.1. Lower bound for the smallest eigenvalue. We derive a lower bound for the smallest eigenvalue of the rank-one update $A + yy^T$, where A is a real symmetric positive semidefinite matrix of order N, and y is a real $N \times 1$ vector.

Let $\beta_1 \geq \cdots \geq \beta_N \geq 0$ be the eigenvalues of A. Weyl's monotonicity theorem [23, Theorem (10.3.1)] implies bounds for the smallest eigenvalue of $A + yy^T$:

$$\beta_N \leq \lambda_{\min}(A + yy^T) \leq \beta_{N-1}.$$

Intuitively one would expect that $\lambda_{\min}(A+yy^T)$ is larger if y is close to an eigenvector of β_N. We confirm this by deriving lower bounds for $\lambda_{\min}(A + yy^T)$ that incorporate the contribution of y in the eigenspace of β_N.

THEOREM 3.3. *Let A be an $N \times N$ real symmetric positive semidefinite matrix, u_N an eigenvector of A associated with β_N, $\|u_N\| = 1$, and $y \neq 0$ a real $N \times 1$ vector. Set $y_N \equiv u_N^T y$. Then*

$$\lambda_{\min}(A + yy^T) \geq \max\left\{\beta_N,\ y_N^2\ \frac{\text{gap}}{\text{gap} + \xi^2}\right\}, \tag{3.4}$$

where gap $\equiv \beta_{N-1} - \beta_N$ *and* $\xi \equiv |y_N| + \sqrt{\|y\|^2 - y_N^2}$.

Proof. We first show that

$$\lambda_{\min}(A + yy^T) \geq \min\left\{\beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},\ \beta_{N-1} \frac{y_N^2}{\xi^2}\right\} \tag{3.5}$$

is a lower bound for $\lambda_{\min}(A + yy^T) = \min_{\|x\|=1} x^T(A + yy^T)x$.

Let

$$A = U \begin{pmatrix} \beta_1 & & \\ & \ddots & \\ & & \beta_N \end{pmatrix} U^T$$

be an eigendecomposition of A, and x be any real vector with $\|x\| = 1$. Partition

$$U^T x = \begin{pmatrix} \bar{x} \\ x_N \end{pmatrix}, \qquad U^T y = \begin{pmatrix} \bar{y} \\ y_N \end{pmatrix}$$

so that $\xi = |y_N| + \|\bar{y}\|$. Then

$$x^T(A + yy^T)x \geq \beta_{N-1}\|\bar{x}\|^2 + \beta_N x_N^2 + (y^T x)^2.$$

If $\|\bar{x}\| \geq |y_N|/\xi$ then

$$x^T(A + yy^T)x \geq (\beta_{N-1} y_N^2)/\xi^2,$$

which proves the second part of the bound in (3.5).

If $\|\bar{x}\| < |y_N|/\xi$ then $|y_N| - \|\bar{x}\|\xi > 0$, and it makes sense to use $|x_N| \geq 1 - \|\bar{x}\|$ in

$$|y^T x| = |y_N x_N + \bar{y}^T \bar{x}| \geq |y_N x_N| - \|\bar{x}\|\|\bar{y}\| \geq |y_N| - \|\bar{x}\|\xi.$$

Hence

$$x^T(A + yy^T)x \geq \beta_{N-1}\|\bar{x}\|^2 + \beta_N x_N^2 + (y^T x)^2 \geq \beta_N + y_N^2 + (\text{gap} + \xi^2)\|\bar{x}\|^2 - 2\xi\|\bar{x}\||y_N|.$$

This is a function of $\|\bar{x}\|$ which has a minimum at $\|\bar{x}\| = |y_N|\xi/(\text{gap} + \xi^2)$. Hence

$$x^T(A + yy^T)x \geq \beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},$$

which proves the first part of the bound in (3.5).

With the help of (3.5) we now show the desired bound (3.4). Weyl's theorem [23, Theorem (10.3.1)] implies $\lambda_{\min}(A + yy^T) \geq \beta_N$, which proves the first part of the bound in (3.4). For the second part of the bound in (3.4), we use the fact that the eigenvalues of A are nonnegative; hence $\beta_{N-1} \geq \text{gap}$ and

$$\frac{\beta_{N-1}}{\xi^2} \geq \frac{\text{gap}}{\text{gap} + \xi^2}.$$

Substituting this into (3.5) gives the second part of the bound in (3.4):

$$\begin{aligned} \min(A + yy^T) &\geq \min\left\{ \beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},\, y_N^2 \frac{\beta_{N-1}}{\xi^2} \right\} \\ &\geq \min\left\{ \beta_N + y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2},\, y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2} \right\} = y_N^2 \frac{\text{gap}}{\text{gap} + \xi^2}. \qquad \square \end{aligned}$$

The quantity gap in Theorem 3.3 is the absolute gap between the smallest and next smallest eigenvalues, and ξ is an approximation for $\|y\|$ since $\|y\| \leq \xi \leq \sqrt{2}\|y\|$. The theorem shows that $\lambda_{\min}(A + yy^T)$ is likely to be larger if y has a substantial contribution in the eigenspace of β_N.

Now we are in a position to complete the proof of Theorem 3.2.

3.2. Proof of Theorem 3.2. Define the residual

$$r \equiv G_u(u_0, \lambda_0)\dot{u}_0 + G_\lambda(u_0, \lambda_0)\dot{\lambda}_0.$$

Letting $G_u = G_u(u, \lambda)$, $G_\lambda = G_\lambda(u, \lambda)$, and $F_x = F_x(x, s)$, we have

$$F_x F_x^T = \begin{pmatrix} G_u & G_\lambda \\ \dot{u}_0^T & \dot{\lambda}_0 \end{pmatrix} \begin{pmatrix} G_u^T & \dot{u}_0 \\ G_\lambda^T & \dot{\lambda}_0 \end{pmatrix} = \begin{pmatrix} G_u G_u^T + G_\lambda G_\lambda^T & r \\ r^T & 1 \end{pmatrix}.$$

The eigenvalues of $F_x F_x^T$ are the squares of the singular values of F_x. Applying Theorem 3.3 to $G_u G_u^T + G_\lambda G_\lambda^T$ with $A = G_u G_u^T$, $y = G_\lambda$, $\beta_N = \sigma_N(u,\lambda)^2$, $\beta_{N-1} = \sigma_{N-1}^2(u,\lambda)$, and gap $= \sigma_{N-1}(u,\lambda)^2 - \sigma_N(u,\lambda)^2$ shows $\lambda_{\min}(G_u G_u^T + G_\lambda G_\lambda^T) \geq \alpha$. Hence we can write

$$\begin{pmatrix} G_u G_u^T + G_\lambda G_\lambda^T & 0 \\ 0 & 1 \end{pmatrix}^{-1} F_x F_x^T = I + E,$$

where $\|E\| \leq \tau \max\left\{\frac{1}{\alpha}, 1\right\}$. If $\tau < \min\{\alpha, 1\}$ then $\|E\| < 1$, $I + E$ is nonsingular, and

$$\frac{1}{\|(F_x F_x^T)^{-1}\|} \geq 1 - \tau \max\left\{\frac{1}{\alpha}, 1\right\}.$$

4. Newton-GMRES and eigenvalue clustering. This section discusses the performance of the inner GMRES iteration in the context of continuation with a Newton-GMRES nonlinear solver. Theorem 3.2 gives bounds on the smallest singular value of F_x in terms of the singular values of G_u. These lower bounds lead to bounds on the condition number of F_x. While the results in the previous section address conditioning, they do not directly translate into the performance of iterative methods [12, 14, 30], especially in the nonnormal case. However, we can go further to see that the eigenvalue clustering properties of the matrix F_x do not stray far from those of G_u.

Suppose the eigenvalues of G_u are nicely clustered (in the sense of [4, 17]). Even in the singular case, this would mean that the zero eigenvalue of G_u is an "outlier." We seek to show that adding the row and column does not significantly increase the number of outliers, and that we can then use the estimates in [4, 17].

The idea is that [16]

$$G_u = I + K(u) + E, \tag{4.1}$$

where K_u is a low-rank operator, say, of rank p, and E is small. We then want to write F_x in the same way, and then compare the number of outliers by comparing the ranks of the K-terms. The assumption that (4.1) holds is clearly valid if G_u is a compact perturbation of the identity; examples of this are nonlinear integral equations as well as the compact maps which are implicitly defined by the time-steppers as described in [1, 9, 10, 18, 28, 29].

Assume that E is small enough so that the eigenvalues of $I - K$ are "outliers" in the sense of [4]. Since the degree of the minimal polynomial of $I - K$ is at most $p + 1$, we have a bound for the sequence of residuals $\{r_l\}$ of the GMRES iteration of the form

$$\|r_{\hat{p}+k}\| \leq C\|E\|^k \|r_0\|, \tag{4.2}$$

where $\hat{p} \leq p + 1$ GMRES iterations are needed to remove the contribution of the outlying eigenvalues.

Theorem 4.1 states that the spectral properties of F_x are similar to those of G_u.

THEOREM 4.1. *Let the assumptions of Theorem* 3.2 *hold. Assume that* (4.1) *holds with* $\operatorname{rank}(K(u)) = p$. *Then there is* $\mathcal{K}(u)$ *having rank at most* $p + 2$ *such that*

$$\|F_x - I - \mathcal{K}(u)\| \leq \|E\|.$$

Proof. We write [16]

$$F_x = I_{(N+1)\times(N+1)} + \begin{pmatrix} K & G_\lambda \\ \dot{u}^T & \dot{\lambda} \end{pmatrix} + \begin{pmatrix} E & 0 \\ 0 & 0 \end{pmatrix}.$$

The range of

$$\mathcal{K} = \begin{pmatrix} K & G_\lambda \\ \dot{u}^T & \dot{\lambda} \end{pmatrix}$$

is

$$\begin{pmatrix} \text{Range}(K) \\ 0 \end{pmatrix} + \text{span}\left\{ \begin{pmatrix} G_\lambda \\ 0 \end{pmatrix} \right\} + \text{span}\left\{ \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right\},$$

and hence the rank of $\mathcal{K}$ is at most $p+2$. ∎

So, while the eigenvalues may change, we have not increased the degree of the minimal polynomial of the main term (K versus $\mathcal{K}$) beyond $p+3$. Hence, the methods of [4] can be applied to obtain a bound like (4.2) with $\hat{p} \le p+3$.

5. Example: Chandrasekhar *H*-equation. We now present an example of a solution path containing a simple fold. The equation of interest is called the Chandrasekhar H-equation [6, 14, 21] from radiative transfer theory:

$$H(\mu) = 1 - \left(\frac{c}{2} \int_0^1 H(\nu) \frac{d\nu\mu}{\mu+\nu} \right)^{-1}. \tag{5.1}$$

The goal is to compute $H(\mu)$ for $\mu \in [0,1]$ as a function of c. There is a simple fold at $c = 1$ [21], and the same analysis shows that this is also the case for any discretization of the equation which uses a quadrature rule that integrates constant functions exactly.

In this section we use a Newton-GMRES version of pseudo-arclength continuation [8], fixing the step in arclength to $ds = .02$, using a secant predictor [13], and beginning the continuation at $c = 0$, where $H = 1$ is the solution. The vector with components all equal to one is the solution of the discrete problem as well. We discretize the integral with the composite midpoint rule using 200 nodes. A consequence of this discretization is that all scalar products of discretized functions in the GMRES solves were scaled by $1/200$. Because we do this, all the singular value results will converge as the quadrature rule is refined.

Figure 5.1 is a plot of $\|H\|_1$ against c. A fixed value of ds, as we use here, causes problems as the L^1 norm of H increases. The reason for this is that the solution develops very large derivatives, and the predictor becomes very poor. We stopped the continuation at $c = .9$ on the upper branch for that reason.

For this example we can also compute the L^1 norm as a function of c analytically, and verify the results in Figure 5.1. We can rewrite (5.1) as

$$H(\mu) = 1 + \frac{c}{2} \int_0^1 H(\mu) H(\nu) \frac{d\nu\mu}{\mu+\nu}. \tag{5.2}$$

Integrating (5.2) with respect to μ yields

$$\|H\|_1 = 1 + \frac{c}{2} \int_0^1 \int_0^1 \frac{H(\mu)H(\mu)\mu \, d\mu \, d\nu}{\mu+\nu} = 1 + \frac{c}{4} \|H\|_1^2,$$

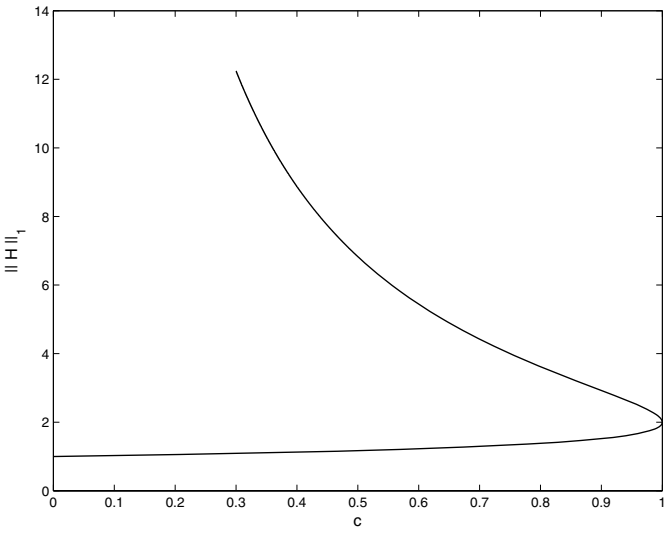


FIG. 5.1. $\|H\|_1$ *as a function of* c.

and so

$$\|H\|_1 = \frac{1 \pm \sqrt{1-c}}{c/2}. \tag{5.3}$$

As a demonstration of the result in section 3, we calculate the smallest singular value of the Jacobian matrix associated with the augmented system for the H-equation with each continuation iteration. We used the MATLAB `svds` command for this. In the language of section 3, we find $\sigma_{\min}(F_{(H,c)})$ for various c where $F_{(H,c)}$ denotes the Jacobian of

$$\begin{pmatrix} G(H,c) \\ \mathcal{N}(H,c,s) \end{pmatrix}.$$

Figure 5.2 shows that the smallest singular value of $F_{(H,c)}$ for each c stays away from zero keeping $F_{(H,c)}$ nonsingular, even at the simple fold ($c = 1$).

It is interesting to compare Figure 5.2 with a plot of the smallest singular value of G_H, which we can also compute on the path. In Figure 5.3, one can see the singularity at $c = 1$ and also see that G_H is becoming more and more poorly conditioned as the L^1 norm of H increases.

The consequences of the remarks in section 4 are that for a problem like the H-equation, which is a nonlinear compact fixed point problem, the number of GMRES iterations per Newton step should be bounded. One must take this expectation with a grain of salt because as one moves along the path, the norm of the solution increases, and so the number of outliers may increase slowly. The observations we present illustrate this.

In Figure 5.4 we plot the average number of GMRES iterations per Newton iteration as a function of c. The lower curve corresponds to the continuation from $c = 0$ to $c = 1$, and the upper from $c = 1$ to $c = .3$. The computation in this figure was done for pseudo-arclength continuation, which we compare with parameter continuation in Figure 5.5. As one moves further on the path, the predictor becomes less effective,

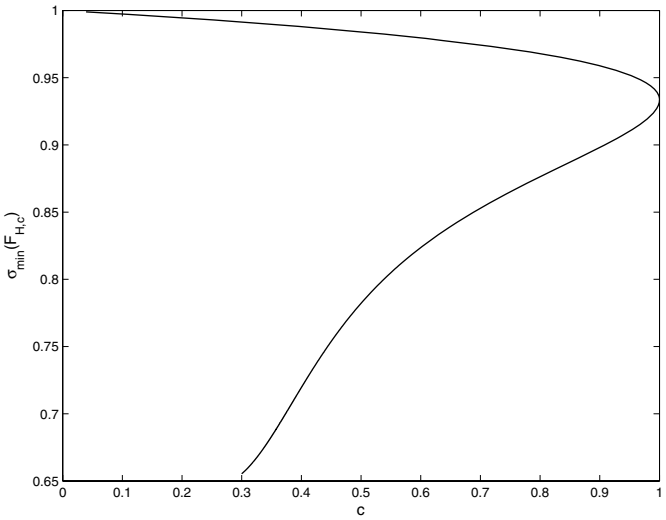


FIG. 5.2. $\sigma_{\min}(F_{(H,c)})$ *as a function of* c.

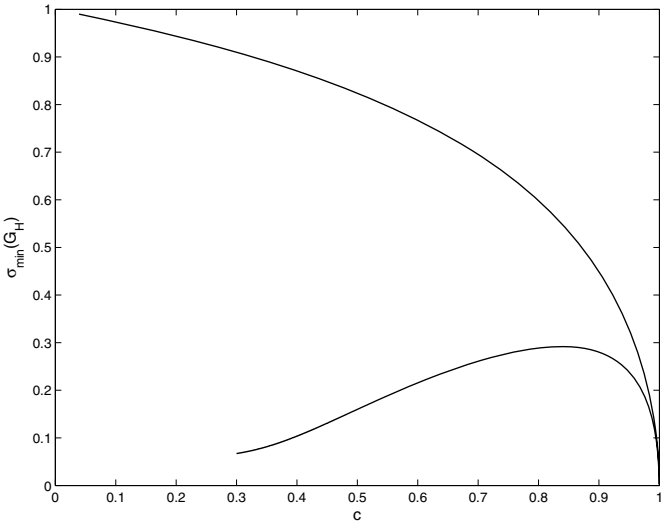


FIG. 5.3. $\sigma_{\min}(G_H)$ *as a function of* c.

and the number of Newton iterations increases. The predictor is also different for the first two points on the path, because we do not have the data we need to build the secant predictor before we have computed two points. The initial point for $c = 0$ is the vector with 1 in each component, which is the solution, so the plots begin with the first nonzero value of c.

Figure 5.5 is the result of a simple parameter continuation for each of the upper and lower branches. The lower curve is for values of $c \in [.3, .9]$, where the problem is quite easy. The linear solver takes fewer GMRES iterations per Newton iteration on this branch, and we observe that the difference in linear iterations from the lower branch in Figure 5.4 is at most 1, consistent with the theory. On the upper branch,

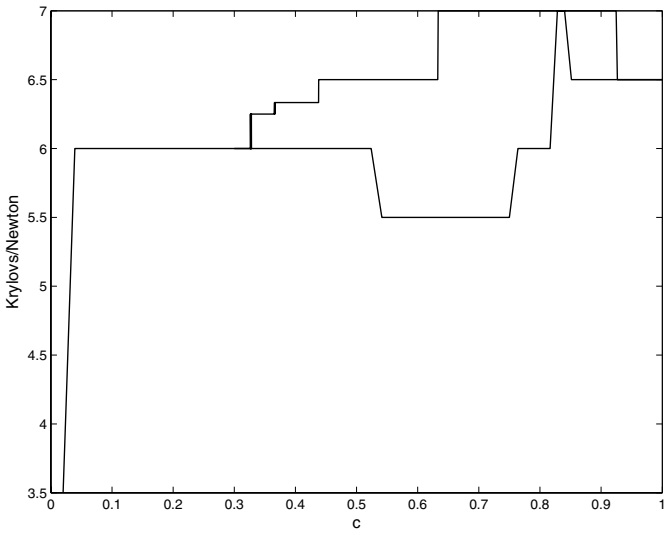


FIG. 5.4. *Krylov's per Newton: Pseudo-arclength continuation.*

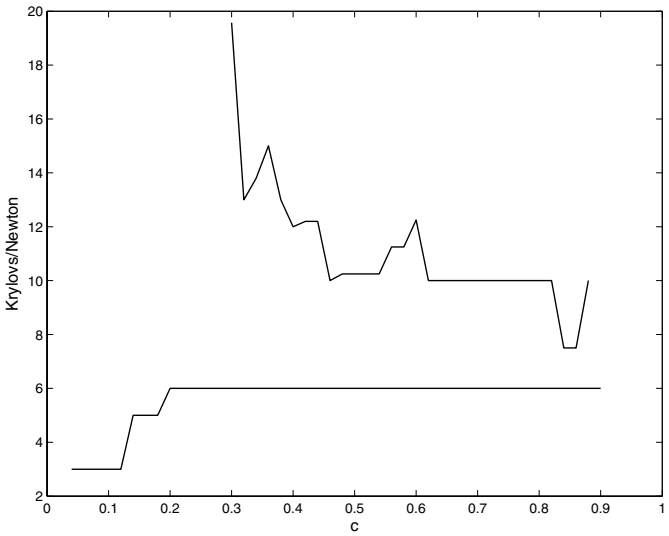


FIG. 5.5. *Krylov's per Newton: Parameter continuation in c.*

where $c \in [.3, .9]$, the performance of parameter continuation is significantly worse than that of pseudo-arclength continuation, and the linear solver performs significantly less well in the parameter continuation solver. This is consistent with the singular value results in Figure 5.3.

6. Conclusion. For simple fold singularities, we have given new bounds on the conditioning of the extended system of nonlinear equations that arise in pseudo-arclength continuation. The two bounds are a lower estimate on the smallest singular value of the Jacobian of the extended system, and an upper bound on the number of eigenvalues that lie outside a cluster of eigenvalues for the Jacobian of the origi-

nal system. The latter of these two bounds implies an upper bound on the number of GMRES iterations needed to achieve a certain termination criterion (4.2). We illustrate the bounds with a numerical experiment.

Acknowledgment. The authors thank Alastair Spence for many helpful discussions.

REFERENCES

[1] A. Armaou and I. G. Kevrekidis, *Equation-free optimal switching policies for bistable reacting systems using coarse time-steppers*, Internat. J. Robust Nonlinear Control, 15 (2005), pp. 713–726; also available online from http://arxiv.org/abs/math.OC/0410467, 2004.
[2] P. N. Brown and Y. Saad, *Hybrid Krylov methods for nonlinear systems of equations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 450–481.
[3] P. N. Brown and Y. Saad, *Convergence theory of nonlinear Newton–Krylov algorithms*, SIAM J. Optim., 4 (1994), pp. 297–330.
[4] S. L. Campbell, I. C. F. Ipsen, C. T. Kelley, and C. D. Meyer, *GMRES and the minimal polynomial*, BIT, 36 (1996), pp. 664–675.
[5] T. F. C. Chan and H. B. Keller, *Arc-length continuation and multi-grid techniques for nonlinear elliptic eigenvalue problems*, SIAM J. Sci. Statist. Comput., 3 (1982), pp. 173–194.
[6] S. Chandrasekhar, *Radiative Transfer*, Dover, New York, 1960.
[7] E. J. Doedel and J. P. Kernévez, *AUTO: Software for Continuation and Bifurcation Problems in Ordinary Differential Equations*, Tech. rep., California Institute of Technology, Pasadena, CA, 1986.
[8] W. R. Ferng and C. T. Kelley, *Mesh independence of matrix-free methods for path following*, SIAM J. Sci. Comput., 21 (2000), pp. 1835–1850.
[9] C. W. Gear and I. G. Kevrekidis, *Telescopic projective methods for parabolic differential equations*, J. Comput. Phys., 187 (2003), pp. 95–109.
[10] C. W. Gear, I. G. Kevrekidis, and K. Theodoropoulos, *Coarse integration/bifurcation analysis via microscopic simulators: Micro-Galerkin methods*, Comp. Chem. Engrg., 26 (2002), pp. 941–963.
[11] W. J. F. Govaerts, *Numerical Methods for Bifurcations of Dynamic Equilibria*, SIAM, Philadelphia, 2000.
[12] I. Ipsen and C. Meyer, *The idea behind Krylov methods*, Amer. Math. Monthly, 105 (1998), pp. 889–99.
[13] H. B. Keller, *Lectures on Numerical Methods in Bifurcation Theory*, Tata Institute of Fundamental Research Lectures on Mathematics and Physics 79, Springer-Verlag, Berlin, 1987.
[14] C. T. Kelley, *Iterative Methods for Linear and Nonlinear Equations*, Frontiers Appl. Math. 16, SIAM, Philadelphia, 1995.
[15] C. T. Kelley, *Solving Nonlinear Equations with Newton's Method*, Fundamentals of Algorithms 1, SIAM, Philadelphia, 2003.
[16] C. T. Kelley, I. G. Kevrekidis, and L. Qiao, *Newton-Krylov Solvers for Time-Steppers*, Tech. rep. CRSC-TR04-10, Center for Research in Scientific Computation, North Carolina State University, Raleigh, NC, 2004.
[17] C. T. Kelley and Z. Q. Xue, *GMRES and integral operators*, SIAM J. Sci. Comput., 17 (1996), pp. 217–226.
[18] I. G. Kevrekidis, C. W. Gear, J. M. Hyman, P. G. Kevrekidis, O. Runborg, and C. Theodoropoulos, *Equation-free coarse-grained multiscale computation: Enabling microscopic simulators to perform system-level analysis*, Commun. Math. Sci., 1 (2003), pp. 715–762.
[19] Y. A. Kuznetsov, *Elements of Applied Bifurcation Theory*, Springer-Verlag, New York, 1998.
[20] R. Menzel and H. Schwetlick, *Zur Lösung parameterabhängiger nichtlinearer Gleichungen mit singulären Jacobi-Matrizen*, Numer. Math., 30 (1978), pp. 65–79.
[21] T. W. Mullikin, *Some probability distributions for neutron transport in a half space*, J. Appl. Probab., 5 (1968), pp. 357–374.
[22] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York, 1970.
[23] B. N. Parlett, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

[24] W. C. Rheinboldt, *Solution fields of nonlinear equations and continuation methods*, SIAM J. Numer. Anal., 17 (1980), pp. 221–237.
[25] W. C. Rheinboldt, *Numerical Analysis of Parametrized Nonlinear Equations*, John Wiley and Sons, New York, 1986.
[26] Y. Saad and M. H. Schultz, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Statist. Comput., 7 (1986), pp. 856–869.
[27] A. G. Salinger, N. M. Bou-Rabee, R. P. Pawlowski, E. D. Wilkes, E. A. Burroughs, R. B. Lehoucq, and L. A. Romero, *LOCA* 1.0 *Library of Continuation Algorithms: Theory and Implementation Manual*, Tech. rep. SAND2002-0396, Sandia National Laboratory, Albuquergue, NM, 2002.
[28] C. I. Siettos, C. C. Pantelides, and I. G. Kevrekidis, *Enabling dynamic process simulators to perform alternative tasks: A time-stepper based toolkit for computer-aided analysis*, Ind. Eng. Chem. Res., 42 (2003), pp. 6795–6801.
[29] K. Theodoropoulos, Y.-H. Qian, and I. G. Kevrekidis, *Coarse stability and bifurcation analysis using timesteppers: A reaction diffusion example*, Proc. Natl. Acad. Sci., 97 (2000), pp. 9840–9843.
[30] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 277–295

FINITE ELEMENT APPROXIMATIONS IN A NONLIPSCHITZ DOMAIN*

GABRIEL ACOSTA†, MARÍA G. ARMENTANO‡, RICARDO G. DURÁN‡, AND ARIEL L. LOMBARDI‡

Abstract. In this paper we analyze the approximation by standard piecewise linear finite elements of a nonhomogeneous Neumann problem in a cuspidal domain. Since the domain is not Lipschitz, many of the results on Sobolev spaces, which are fundamental in the usual error analysis, do not apply. Therefore, we need to work with weighted Sobolev spaces and to develop some new theorems on traces and extensions. We show that, in the domain considered here, suboptimal order can be obtained with quasi-uniform meshes even when the exact solution is in H^2, and we prove that the optimal order with respect to the number of nodes can be recovered by using appropriate graded meshes.

Key words. cuspidal domains, Neumann problem, finite elements, graded meshes

AMS subject classifications. 65N30, 46E35

DOI. 10.1137/050647797

1. Introduction.

The finite element method has been widely analyzed in its different forms for all kind of partial differential equations. However, as far as we know, all analyses are restricted to the case of polygonal or smooth domains, and no results have been obtained for the case in which the domain is nonLipschitz, with the exception of the well-known fracture problems.

The goal of this paper is to initiate the analysis of finite element approximations in nonLipschitz domains. As a first step in this direction we consider a model problem in a plane domain with an external cusp.

Several difficulties arise in this problem because many of the results on Sobolev spaces, which are fundamental in the analysis of partial differential equations in variational form, do not apply. For example, the standard trace theorems do not hold in this case, and this fact makes the analysis of nonhomogeneous Neumann problems more difficult.

Given $\alpha > 1$, let $\Omega \subset \mathbb{R}^2$ be the domain defined by

$$\Omega = \{(x, y) : 0 < x < 1,\ 0 < y < x^\alpha\},$$

and let $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$ be its boundary, with

$$\Gamma_1 = \{0 \le x \le 1, y = 0\},\quad \Gamma_2 = \{x = 1, 0 \le y \le 1\},\quad \text{and}\quad \Gamma_3 = \{0 \le x \le 1, y = x^\alpha\}$$

(see Figure 1).

*Received by the editors December 16, 2005; accepted for publication (in revised form) August 4, 2006; published electronically February 2, 2007. Supported by ANPCyT under grant PICT 03-13719, Fundación Antorchas, by Universidad de Buenos Aires under grant X052, and by CONICET under grant PIP 5478.

http://www.siam.org/journals/sinum/45-1/64779.html

†Instituto de Ciencias, Universidad Nacional de General Sarmiento, J.M. Gutierrez 1150, Los Polvorines, B1613GSX Provincia de Buenos Aires, Argentina (gacosta@ungs.edu.ar). This author is partially supported by grant PICT 03-10724.

‡Departamento de Matemática, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, 1428 Buenos Aires, Pab. I, Ciudad Universitaria, Argentenia (garmenta@dm.uba.ar,rduran@dm.uba.ar, aldoc7@dm.uba.ar). These authors are members of CONICET, Argentina.

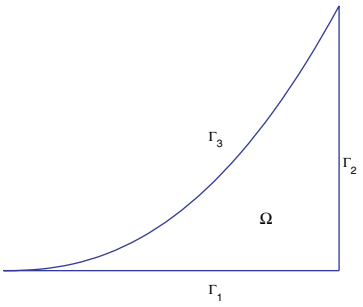


FIG. 1. *Cuspidal domain.*

Some of our arguments require that $\alpha < 3$, and so our main result will be valid under this restriction.

Our model problem is

$$\begin{cases} -\Delta u = f & \text{in } \Omega, \\ \dfrac{\partial u}{\partial \nu} = g & \text{on } \Gamma_3, \\ \dfrac{\partial u}{\partial \nu} = 0 & \text{on } \Gamma_1, \\ u = 0 & \text{on } \Gamma_2, \end{cases} \tag{1.1}$$

where ν denotes the outside normal.

A natural way to approximate the solution of problem (1.1) is to replace Ω with a polygonal domain and to use the standard linear finite element method. It is known that, under appropriate conditions on the data, the solution of this problem is in $H^2(\Omega)$ (see [1]). Therefore, based on the experience of and theory for smooth domains, one would expect that the optimal order of convergence could be obtained by using quasi-uniform meshes. However, numerical examples show that this is not the case (see section 2). The reason for this behavior seems to be the fact that the solution cannot be extended to an H^2 function on the polygonal domain approximating the original domain. Indeed, it is known that the standard extension theorems in Sobolev spaces do not apply for our domain (see, for example, [12]).

We will show that the optimal order with respect to the number of nodes in the H^1 norm can be recovered by using appropriate graded meshes. To obtain this result, we will first prove an extension theorem for the domain Ω which shows that the solution of problem (1.1) can be extended to a function in a weighted H^2 space, with the weight being a power of the distance to the cuspidal point.

The rest of the paper is organized as follows. In section 2 we introduce the finite element approximation of our problem and show that the use of quasi-uniform meshes can give bad results. Section 3 deals with some extension and trace theorems in weighted Sobolev spaces that we need for our error analysis. Finally, in section 4 we prove that optimal order approximations are obtained by using appropriate graded meshes.

2. Finite element approximations. In this section we introduce the finite element approximation of our model problem and show that, if the meshes are quasi uniform, the approximation may be of suboptimal order even when the exact solution is in $H^2(\Omega)$.

Introducing the space

$$V = \{v \in H^1(\Omega) : v|_{\Gamma_2} = 0\},$$

we see that the weak form of problem (1.1) is to find $u \in V$ such that

$$\int_\Omega \nabla u \cdot \nabla v = \int_\Omega fv + \int_{\Gamma_3} gv \qquad \forall v \in V. \tag{2.1}$$

The following existence and regularity results have been proved in [1]. Define $z(t) := g(t, t^\alpha)$. If $f \in L^2(\Omega)$ and $z\, t^{-\frac{\alpha}{2}} \in L^2(0,1)$, this problem has a unique solution. If in addition we assume that $z'\, t^{1-\frac{\alpha}{2}} \in L^2(0,1)$, the solution is in $H^2(\Omega)$ and there exists a constant C such that

$$\|u\|_{H^2(\Omega)} \le C \left\{ \|f\|_{L^2(\Omega)} + \|z\, t^{-\frac{\alpha}{2}}\|_{L^2(0,1)} + \|z'\, t^{1-\frac{\alpha}{2}}\|_{L^2(0,1)} \right\}. \tag{2.2}$$

To approximate the solution of (1.1) we replace Ω with a polygonal domain Ω_h and use the standard linear finite element method. We will construct Ω_h in such a way that $\Omega \subset \Omega_h$ and the nodes on Γ_h, the boundary of Ω_h, are also on Γ.

Let $\{\mathcal{T}_h\}$ be a family of triangulations of Ω_h satisfying the maximum angle condition. Associated with $\{\mathcal{T}_h\}$ we have the finite element space

$$V_h = \{v \in H^1(\Omega_h) : v|_{\Gamma_2} = 0 \,\text{and}\, v|_T \in \mathcal{P}_1 \quad \forall\, T \in \mathcal{T}_h\},$$

where $\mathcal{P}_1$ denotes the space of linear polynomials.

Denote by $\Gamma_{3,h}$ the part of Γ_h approximating Γ_3 and by I_h the piecewise linear interpolation at the endpoints of the segments which lie on $\Gamma_{3,h}$.

Then, our discrete problem is to find $u_h \in V_h$ such that

$$\int_{\Omega_h} \nabla u_h \cdot \nabla v = \int_\Omega fv + \int_{\Gamma_{3,h}} I_h(gv) \quad \forall v \in V_h. \tag{2.3}$$

Observe that the discrete problem corresponds to a boundary problem on Ω_h if we consider f as being extended by zero outside Ω.

One could think that, when the solution is in $H^2(\Omega)$, the numerical approximation obtained with quasi-uniform meshes would be of optimal order. However, the following example shows that this is not the case.

Example 2.1. Consider

$$f(x,y) = s(s-1)(1+y^2/2)x^{s-2} + x^s - 1$$

and

$$z(t) = g(t, t^\alpha) = \frac{-s\alpha t^{\alpha+s-2}(1+t^{2\alpha}/2) + (1-t^s)t^\alpha}{\sqrt{1+\alpha^2 t^{2(\alpha-1)}}}.$$

Then, the solution of (1.1) is

$$u(x,y) = (1-x^s)(1+y^2/2),$$

and an easy calculation shows that $u \in H^2(\Omega)$ whenever $s > \frac{3-\alpha}{2}$.

We take $\alpha = 2$, and different values of s, with $\frac{1}{2} < s < 1$, and solve problem (2.3) by using quasi-uniform meshes. Table 1 shows the order of the error in the H^1 norm in terms of number of nodes and in terms of mesh size.

The reason for this behavior seems to be the fact that the solution cannot be extended to an H^2 function on Ω_h. Indeed, it is well known that the standard extension theorems in Sobolev spaces do not apply for our domain (see, for example, [12]).

TABLE 1
H^1 *order using quasi-uniform meshes for* $\alpha = 2$.

Value of s	Order in number of nodes	Order in h
0.55	0.324	0.626
0.6	0.335	0.647
0.65	0.347	0.671
0.7	0.362	0.698
0.75	0.380	0.733
0.8	0.404	0.781
0.85	0.440	0.849
0.9	0.491	0.948
0.95	0.545	1.053

3. Extension and trace theorems. The standard results on extensions and restrictions in Sobolev spaces do not apply for domains with external cusps. In this section we prove some weaker results using weighted norms.

First, we develop an extension theorem in a weighted Sobolev space for $H^2(\Omega)$ functions with a vanishing normal derivative on Γ_1. In particular, our theorem applies to solutions of (1.1) which, in view of (2.2), are in $H^2(\Omega)$ under appropriate assumptions on the data.

Second, we prove a trace theorem for functions in $H^1(\Omega)$ which will be useful for estimating the error due to the approximation of the nonhomogeneous Neumann-type boundary condition.

Given a domain $\mathcal{D} \subset \mathbb{R}^2$ we introduce the weighted Sobolev space

$$H^2_\alpha(\mathcal{D}) = \left\{ v : r^{\frac{\alpha-1}{2}} D^\gamma v \in L^2(\mathcal{D}) \ \ \forall\, \gamma\,, |\gamma| \le 2 \right\},$$

where $r = \sqrt{x^2+y^2}$, and its natural norm

$$\|v\|^2_{H^2_\alpha(\mathcal{D})} = \sum_{|\gamma|\le 2} \|r^{\frac{\alpha-1}{2}} D^\gamma v\|^2_{L^2(\mathcal{D})}.$$

Our argument proceeds in two steps. First, we extend the given function to the Lipschitz domain

$$\mathcal{D} := \{(x,y) \in \mathbb{R}^2 : \ \ -x < y < x^\alpha\,, \ \ 0 < x < 1\}$$

(see Figure 2) in such a way that the extension belongs to $H^2_\alpha(\mathcal{D})$. Then, we apply known theorems for weighted Sobolev spaces on Lipschitz domains to obtain an extension which belongs to $H^2_\alpha(\mathbb{R}^2)$.

We call W the subspace of $H^2(\Omega)$ defined by

$$W = \left\{ u \in H^2(\Omega) : \ \ \frac{\partial u}{\partial \nu} = 0 \ \text{ on } \ \Gamma_1 \right\}.$$

LEMMA 3.1. *Given* $u \in W$ *there exists a function* $\tilde{u} \in H^2_\alpha(\mathcal{D})$ *such that* $\tilde{u}|_\Omega = u$ *and*

$$\|\tilde{u}\|_{H^2_\alpha(\mathcal{D})} \le C \|u\|_{H^2(\Omega)}.$$

Proof. We extend u in the following way. Given $(x,y) \in \mathcal{D}$ with $y \le 0$, let $\eta = -x^{\alpha-1} y$. Observe that $(x,\eta) \in \Omega$, and therefore we can define

$$\begin{cases} \tilde{u}(x,y) = u(x,y) & \text{for} \quad (x,y) \in \Omega, \\ \tilde{u}(x,y) = u(x,\eta) & \text{for} \quad (x,y) \in \mathcal{D} \setminus \Omega. \end{cases}$$

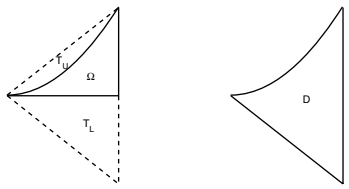


FIG. 2.

To simplify notation define $T_L := \mathcal{D} \setminus \overline{\Omega}$.

We claim that $\tilde{u} \in H^2_\alpha(T_L)$. Observe first that for $(x, y) \in T_L$ we have $x \sim r$, and therefore we can replace the weight $r^{\alpha-1}$ with $x^{\alpha-1}$ in our estimates.

By a change of variables we obtain

$$\int_{T_L} \tilde{u}^2(x,y) x^{\alpha-1} dxdy = \int_\Omega u^2(x,\eta) dxd\eta = \|u\|^2_{L^2(\Omega)}.$$

Now, for $(x, y) \in T_L$ we have

$$\frac{\partial \tilde{u}}{\partial x}(x,y) = \frac{\partial u}{\partial x}(x,\eta) - \frac{\partial u}{\partial \eta}(x,\eta)(\alpha-1)x^{\alpha-2}y$$

and

$$\frac{\partial \tilde{u}}{\partial y}(x,y) = -\frac{\partial u}{\partial \eta}(x,\eta)x^{\alpha-1}.$$

Then, recalling that $\eta = -x^{\alpha-1}y$, we obtain

$$\int_{T_L} \left(\frac{\partial \tilde{u}}{\partial x}\right)^2 x^{\alpha-1} dxdy \le C \left\{ \int_\Omega \left(\frac{\partial u}{\partial x}\right)^2 dxd\eta + \int_\Omega \left(\frac{\partial u}{\partial \eta}\right)^2 \left(\frac{\eta}{x}\right)^2 dxd\eta \right\}$$

but, since $(x, \eta) \in \Omega$, we have $\frac{\eta}{x} \le x^{\alpha-1} \le 1$, and then

$$\int_{T_L} \left(\frac{\partial \tilde{u}}{\partial x}\right)^2 x^{\alpha-1} dxdy \le C\|\nabla u\|^2_{L^2(\Omega)}.$$

Analogously we get

$$\int_{T_L} \left(\frac{\partial \tilde{u}}{\partial y}\right)^2 x^{\alpha-1} dxdy \le C\|\nabla u\|^2_{L^2(\Omega)}.$$

Bounds for the second derivatives of $\tilde{u}$ follow similarly. For instance, we have

$$\begin{aligned}\frac{\partial^2 \tilde{u}}{\partial x^2}(x,y) = {} & \frac{\partial^2 u}{\partial x^2}(x,\eta) - 2(\alpha-1)\frac{\partial^2 u}{\partial \eta \partial x}(x,\eta)x^{\alpha-2}y \\ & - \frac{\partial^2 u}{\partial \eta^2}(x,\eta)(\alpha-1)^2 x^{2(\alpha-2)}y^2 - \frac{\partial u}{\partial \eta}(x,\eta)(\alpha-2)(\alpha-1)x^{\alpha-3}y,\end{aligned}$$

and hence,

$$\begin{aligned}\int_{T_L} \left(\frac{\partial^2 \tilde{u}}{\partial x^2}\right)^2 x^{\alpha-1} dxdy \le C \Bigg\{ & \int_\Omega \left(\frac{\partial^2 u}{\partial x^2}\right)^2 dxd\eta + \int_\Omega \left(\frac{\partial^2 u}{\partial \eta \partial x}\right)^2 \left(\frac{\eta}{x}\right)^2 dxd\eta \\ & + \int_\Omega \left(\frac{\partial^2 u}{\partial \eta^2}\right)^2 \left(\frac{\eta}{x}\right)^4 dxd\eta + \int_\Omega \left(\frac{\partial u}{\partial \eta}\right)^2 \left(\frac{\eta}{x^2}\right)^2 dxd\eta \Bigg\}.\end{aligned}$$

Now, the first three terms on the right-hand side can be bounded by using again that $\frac{\eta}{x} \leq 1$. For the last term we have

$$\int_\Omega \left(\frac{\partial u}{\partial \eta}\right)^2 \left(\frac{\eta}{x^2}\right)^2 dxdy \leq \int_0^1 \int_0^{x^\alpha} \left(\frac{\partial u}{\partial \eta}\right)^2 \frac{1}{\eta^2} d\eta dx \leq C \int_0^1 \int_0^{x^\alpha} \left(\frac{\partial^2 u}{\partial \eta^2}\right)^2 d\eta dx,$$

where the last inequality follows from the Hardy inequality [10] and the fact that $\frac{\partial u}{\partial \eta}(x,0) = 0$. Hence,

$$\int_{T_L} \left(\frac{\partial^2 \tilde{u}}{\partial x^2}\right)^2 x^{\alpha-1} dxdy \leq C|u|^2_{H^2(\Omega)}.$$

In a similar way we can show that

$$\int_{T_L} \left(\frac{\partial^2 \tilde{u}}{\partial y \partial x}\right)^2 x^{\alpha-1} dxdy \leq C|u|^2_{H^2(\Omega)}$$

and

$$\int_{T_L} \left(\frac{\partial^2 \tilde{u}}{\partial y^2}\right)^2 x^{\alpha-1} dxdy \leq C|u|^2_{H^2(\Omega)},$$

where $|\cdot|_{H^2(\Omega)}$ denotes the H^2-seminorm in Ω.

Therefore, we have proved that $\tilde{u} \in H^2_\alpha(T_L)$ and that

$$\|\tilde{u}\|_{H^2_\alpha(T_L)} \leq C\|u\|_{H^2(\Omega)}.$$

On the other hand, using that $\frac{\partial u}{\partial \nu} = 0$ on Γ_1, it is easy to see that $\tilde{u} \in H^2_\alpha(\mathcal{D})$, thus concluding the proof. ∎

Now, using known extension theorems for weighted Sobolev spaces on Lipschitz domains due to Chua [6], we can extend functions in W to $H^2_\alpha(\mathbb{R}^2)$.

THEOREM 3.1. *If $\alpha < 3$ and $u \in W$, there exists a function $\tilde{u} \in H^2_\alpha(\mathbb{R}^2)$ such that $\tilde{u}|_\Omega = u$ and*

$$\|\tilde{u}\|_{H^2_\alpha(\mathbb{R}^2)} \leq C\|u\|_{H^2(\Omega)}.$$

Proof. In view of Lemma 3.1 we have only to show that for $v \in H^2_\alpha(\mathcal{D})$ there exists an extension $\tilde{v} \in H^2_\alpha(\mathbb{R}^2)$ such that

$$\|\tilde{v}\|_{H^2_\alpha(\mathbb{R}^2)} \leq C\|v\|_{H^2_\alpha(\mathcal{D})}.$$

But this follows immediately from the results in [6] because, for $1 < \alpha < 3$, our weight belongs to the class considered in that paper (the Muckenhoupt class A_2) [7, page 145]. ∎

In the rest of this section we prove a trace theorem for functions in $H^1(\Omega)$. In [1] it was proved that

$$\|u\|_{L^2(\Gamma)} \leq C(\|ux^{-\frac{\alpha}{2}}\|_{L^2(\Omega)} + \|\nabla u x^{\frac{\alpha}{2}}\|_{L^2(\Omega)}). \tag{3.1}$$

Our trace theorem is a consequence of this result and the known imbedding theorem

$$H^1(\Omega) \subset L^r(\Omega) \quad \text{for} \quad 2 \leq r \leq \frac{2(\alpha+1)}{\alpha-1}, \tag{3.2}$$

which is a particular case of the results given in [2].

THEOREM 3.2. *Let $u \in H^1(\Omega)$.*
(1) *If $\alpha < 2$, then $u \in L^2(\Gamma)$ and $\|u\|_{L^2(\Gamma)} \le C\|u\|_{H^1(\Omega)}$.*
(2) *If $\alpha \ge 2$, then $x^\beta u \in L^2(\Gamma)$ and $\|x^\beta u\|_{L^2(\Gamma)} \le C\|u\|_{H^1(\Omega)}$ $\forall \beta > \alpha/2 - 1$.*
Proof. Part (1) was proved in [1]. Therefore, we will prove only (2) here.
Using (3.1) for the function $x^\beta u$ we have

$$\|x^\beta u\|_{L^2(\Gamma)} \le C(\|x^\beta u x^{-\frac{\alpha}{2}}\|_{L^2(\Omega)} + \|\nabla(x^\beta u)x^{\frac{\alpha}{2}}\|_{L^2(\Omega)}).$$

It is easy to see that the second term on the right-hand side is bounded by $\|u\|_{H^1(\Omega)}$ because $\alpha \ge 2$ and $\beta > \alpha/2 - 1$. Then, it is enough to show that

$$\|x^\beta u x^{-\frac{\alpha}{2}}\|_{L^2(\Omega)} \le \|u\|_{H^1(\Omega)}. \tag{3.3}$$

Using the Hölder inequality we have

$$\int_\Omega u^2 x^{2\beta-\alpha} \le \left(\int_\Omega u^{2q}\right)^{\frac{1}{q}} \left(\int_\Omega x^{(2\beta-\alpha)\frac{q}{q-1}}\right)^{\frac{q-1}{q}}.$$

Choosing $q = r/2$ with $r = 2(\alpha+1)/(\alpha-1)$ and using the imbedding theorem (3.2) we obtain

$$\|x^\beta u x^{-\frac{\alpha}{2}}\|_{L^2(\Omega)} \le \left(\int_\Omega x^{(2\beta-\alpha)\frac{q}{q-1}}\right)^{\frac{q-1}{2q}} \|u\|_{H^1(\Omega)}.$$

But

$$\int_\Omega x^{(2\beta-\alpha)\frac{q}{q-1}} = \int_\Omega x^{(2\beta-\alpha)\frac{\alpha+1}{2}} < \infty$$

because $\beta > \alpha/2 - 1$, and therefore (3.3) holds. ∎

4. Optimal approximations using graded meshes. In this section we obtain error estimates in H^1 of quasi-optimal order (i.e., optimal up to a logarithmic factor) with respect to the number of nodes by using appropriate graded meshes.

Finite element methods using graded meshes of the type considered here have been analyzed for problems with corner-type singularities in [3, 4, 9]. In [4, 9] the error estimates were obtained under the classic regularity condition on the meshes (the minimum angle condition). This hypothesis has been relaxed in [3], where the author obtained error estimates under the maximum angle condition. This generalization is very important for our problem because we cannot avoid small angles in those elements which are near the cusp.

Consider $1 < \alpha < 3$ and define $\gamma = (\alpha - 1)/2$. Let Ω_h be an approximating polygon and $\mathcal{T}_h$ a triangulation of it, where $h > 0$ is a parameter that goes to 0. For each $T \in \mathcal{T}_h$ we denote by h_T its diameter and by β_T its maximum angle. We assume that there exist positive constants σ and $\beta < \pi$, independent of h, such that the following hypotheses hold:
(1) $\beta_T < \beta$ for all $T \in \mathcal{T}_h$ (the maximal angle condition).
(2) $h_T \sim \sigma\, h^{\frac{1}{1-\gamma}}$ if $(0,0) \in T$.
(3) $h_T \le \sigma\, h \inf_T x^\gamma$ if $(0,0) \notin T$.

Since we know that the solution of our problem has an extension $\tilde{u} \in H^2_\alpha(\Omega_h)$, we are interested in interpolation error estimates for functions in this space. We call $\Pi v \in V_h$ the piecewise linear Lagrange interpolation of v.

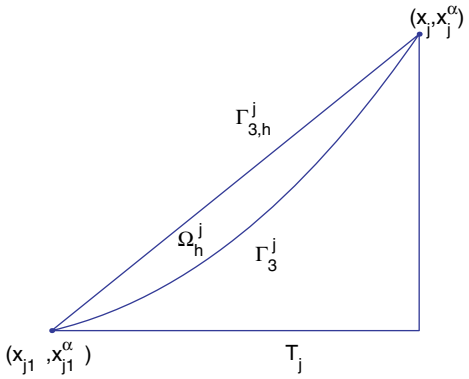


FIG. 3.

THEOREM 4.1. *If $v \in H^2_\alpha(\Omega_h)$ and the family of triangulations satisfies conditions* (1), (2), *and* (3), *there exists a constant C depending only on β, σ, and α such that*

$$\|v - \Pi v\|_{H^1(\Omega_h)} \leq Ch\|v\|_{H^2_\alpha(\Omega_h)}.$$

Proof. The proof follows as in [9, page 392] but using the error estimates obtained by Apel under the maximum angle condition (see Theorem 2.4 in [3, page 63]). ∎

Now we introduce some notation which will be used in the rest of this section. We denote by $\Gamma^j_{3,h}$, $1 \leq j \leq n$, the edges on the boundary of Ω_h, by $(x_{j-1}, x^\alpha_{j-1})$ and (x_j, x^α_j) their endpoints with $x_0 = 0$ and $x_n = 1$, and by Γ^j_3 the part on Γ_3 with the same endpoints. Let Ω^j_h be the region bounded by Γ^j_3 and $\Gamma^j_{3,h}$.

In addition to assumptions (1), (2), and (3) we will need for our error analysis the following hypothesis on the meshes:

(H) For $1 \leq j \leq n$ the region Ω^j_h is contained in only one triangle.

We denote by T_j the triangle containing Ω^j_h and by h_j its diameter (see Figure 3).

It can be seen from our hypotheses that there exists a constant C, independent of h, such that, for $2 \leq j \leq n$,

$$x_j \leq C x_{j-1}. \tag{4.1}$$

In fact, from (H) we have $x_j - x_{j-1} \leq C|\Gamma^j_{3,h}|$ for some constant C depending only on α. Then, $x_j - x_{j-1} \leq Ch_j$, and therefore from assumption (3) we have

$$x_j \leq x_{j-1}\left(1 + Chx^{\gamma-1}_{j-1}\right),$$

and since $j \geq 2$, $x_{j-1} \geq x_1 \sim h^{1/(1-\gamma)}$ by assumption (2), we obtain (4.1).

We will show below that meshes satisfying all our assumptions can indeed be constructed.

The next lemma deals with the error arising from the approximation of the domain by polygonal domains. We will work with an extension $\tilde{u}$ of the solution u of (1.1). Since $u \in W$ we know from Theorem 3.1 that there exists $\tilde{u} \in H^2_\alpha(\mathbb{R}^2)$ such that $\tilde{u}|_\Omega = u$ and

$$\|\tilde{u}\|_{H^2_\alpha(\mathbb{R}^2)} \leq C\|u\|_{H^2(\Omega)}. \tag{4.2}$$

We will use the well-known imbedding $H^1(\mathcal{D}) \subset L^p(\mathcal{D})$ for planar Lipchitz domains and $1 \le p < \infty$ and use the explicit dependence on p of the constant in the continuity of this inclusion (see, for example, [8]), namely,

$$\|v\|_{L^p(\mathcal{D})} \le C\sqrt{p}\,\|v\|_{H^1(\mathcal{D})}. \tag{4.3}$$

LEMMA 4.1. *If $1 < \alpha < 3$, then there exists a constant C, which depends only on α, β, and σ, such that*

$$\|\nabla \tilde{u}\|_{L^2(\Omega_h \setminus \Omega)} \le Ch\sqrt{\log(1/h)}\,\|u\|_{H^2(\Omega)}.$$

Proof. Clearly, for every h, the polygonal domain Ω_h is contained in the triangle

$$T_U = \{0 \le x \le 1,\, 0 \le y \le x\}$$

(see Figure 2). Writing

$$\int_{T_U} |v|^p = \int_{T_U} |v|^p x^{p(\frac{\alpha-1}{2})} x^{-p(\frac{\alpha-1}{2})}$$

and applying the Hölder inequality with $2/p$ and its dual exponent, we obtain

$$\|v\|_{L^p(T_U)} \le C\|v x^{\frac{\alpha-1}{2}}\|_{L^2(T_U)}$$

for any function v and $1 \le p < \frac{4}{\alpha+1}$. Therefore, using (4.2) we conclude that $\tilde{u} \in W^{2,p}(T_U)$ and that

$$\|\tilde{u}\|_{W^{2,p}(T_U)} \le C\|u\|_{H^2(\Omega)}. \tag{4.4}$$

As a consequence, we obtain that, for $\beta > \frac{\alpha-1}{2}$, $\nabla \tilde{u}\, x^\beta \in H^1(T_U)$ and

$$\|\nabla \tilde{u}\, x^\beta\|_{H^1(T_U)} \le C\|u\|_{H^2(\Omega)}. \tag{4.5}$$

Indeed, since $\tilde{u} \in H^2_\alpha(\mathbb{R}^2)$, we already know that $\nabla \tilde{u}\, x^\beta \in L^2(T_U)$, and so we have only to see that the first derivatives of $\nabla \tilde{u}\, x^\beta$ belong to $L^2(T_U)$. But, taking the derivative of $\nabla \tilde{u}\, x^\beta$ and using again that $\tilde{u} \in H^2_\alpha(\mathbb{R}^2)$, we see that it remains only to prove that $\nabla \tilde{u}\, x^{\beta-1} \in L^2(T_U)$.

Now, from (4.4) and a well-known Sobolev imbedding theorem we obtain that $\nabla \tilde{u} \in L^{p^*}(T_U)$ for $1 \le p < \frac{4}{\alpha+1}$ and $p^* = \frac{2p}{2-p}$. Moreover,

$$\|\nabla \tilde{u}\|_{L^{p^*}(T_U)} \le C\|u\|_{H^2(\Omega)}.$$

Therefore, applying the Hölder inequality with $p^*/2$ and its dual exponent q, we have

$$\int_{T_U} |\nabla \tilde{u}|^2 x^{2(\beta-1)} \le \|\nabla \tilde{u}\|^2_{L^{p^*}(T_U)} \|x^{2(\beta-1)}\|_{L^q(T_U)},$$

but since $\beta > \frac{\alpha-1}{2}$, it is possible to choose $p < \frac{4}{\alpha+1}$ such that $\|x^{2(\beta-1)}\|_{L^q(T_U)}$ is finite, thus concluding the proof of (4.5).

Now, let $\beta > \frac{\alpha-1}{2}$ and $2 \le p < \infty$, to be chosen below. Applying the Hölder inequality for $p/2$ and its dual exponent q, we have

$$\int_{\Omega_h \setminus \Omega} |\nabla \tilde{u}|^2 \le \Big(\int_{\Omega_h \setminus \Omega} |\nabla \tilde{u}|^p\, x^{\beta p} \Big)^{\frac{2}{p}} \Big(\int_{\Omega_h \setminus \Omega} x^{-2\beta q} \Big)^{\frac{1}{q}}, \tag{4.6}$$

and therefore, from the Sobolev imbedding (4.3) and (4.5), we obtain

$$\int_{\Omega_h\setminus\Omega} |\nabla \tilde{u}|^2 \le \frac{C}{q-1}\,\|u\|^2_{H^2(\Omega)}\Big(\int_{\Omega_h\setminus\Omega} x^{-2\beta q}\Big)^{\frac{1}{q}} \tag{4.7}$$

for $q \to 1$. Then, we have to estimate

$$\int_{\Omega_h\setminus\Omega} x^{-2\beta q} = \sum_{j=1}^{N}\int_{\Omega_h^j} x^{-2\beta q}. \tag{4.8}$$

Since $\gamma = \frac{\alpha-1}{2}$ and $1<\alpha<3$ we can choose β and $q>1$ such that

$$\gamma < \beta < \min\{2\gamma, 1\} \qquad \text{and} \qquad \beta q < \min\{2\gamma, 1\}.$$

Let us estimate each term in the right-hand side of (4.8). Since $\Omega_h^1 \subset T_1$ we have

$$\int_{\Omega_h^1} x^{-2\beta q} \le \int_{T_1} x^{-2\beta q}.$$

Hence, using now that $h_1 \le \sigma\, h^{\frac{1}{1-\gamma}}$, we obtain

$$\int_{T_1} x^{-2\beta q} \le C h_1^{2(\gamma+1-\beta q)} \le C h^{2\,\frac{\gamma+1-\beta q}{1-\gamma}},$$

and therefore

$$\int_{T_1} x^{-2\beta q} \le C h^2$$

because $q\beta < 2\gamma$.

On the other hand, we have

$$\sum_{j>1}\int_{\Omega_h^j} x^{-2\beta q} \le \sum_{j>1} x_{j-1}^{-2\beta q}|\Omega_h^j|,$$

but by using the well-known error formula for the trapezoidal rule, we obtain

$$|\Omega_h^j| \le C h_j^3 x_{j-1}^{\alpha-2} = C h_j^3 x_{j-1}^{2\gamma-1},$$

where in the case $\alpha > 2$ we have used (4.1). Therefore, since $h_j \le \sigma h\, x_{j-1}^{\gamma}$, we have

$$\begin{aligned}\sum_{j>1}\int_{\Omega_h^j} x^{-2\beta q} &\le C\sum_{j>1} x_{j-1}^{-2\beta q+2\gamma-1} h_j^3 \le C h^2 \sum_{j>1} x_{j-1}^{-2\beta q+4\gamma-1} h_j \\ &\le C h^2 \int_0^1 x^{-2\beta q+4\gamma-1},\end{aligned}$$

where we have used again (4.1). But the last integral is finite because $\beta q < 2\gamma$. Moreover, it is bounded by a constant which remains bounded when $q \to 1$.

Therefore, summing up the estimates obtained, it follows from (4.7) that

$$\|\nabla \tilde{u}\|_{L^2(\Omega_h\setminus\Omega)} \le \frac{C}{\sqrt{q-1}}\,\|u\|_{H^2(\Omega)} h^{\frac{1}{q}}$$

with a constant C which does not blow up when $q \to 1$.

The proof concludes with a standard extrapolation argument taking

$$q = \frac{2\log(1/h)}{2\log(1/h) - 1}. \qquad \square$$

Now, we want to estimate the error arising in the numerical integration of the boundary term. With this goal we introduce an extension $\tilde{g}$ of the function g to $\Gamma_{3,h}$. The procedure is similar to that used in [5]. Calling $\phi(t) = (t, t^\alpha)$ we define $\tilde{g}$ on each $\Gamma_{3,h}^j$ as

$$\tilde{g}(\psi_j(t)) := g(\phi(t)) = z(t), \qquad x_{j-1} \le t \le x_j,$$

where

$$\psi_j(t) = (t, t^\alpha + \delta_j(t))$$

with

$$\delta_j(t) = \frac{x_j^\alpha - x_{j-1}^\alpha}{x_j - x_{j-1}}(t - x_{j-1}) + x_{j-1}^\alpha - t^\alpha.$$

The following lemma gives some estimates for the functions δ_j and their derivatives that will be useful in our error analysis.

LEMMA 4.2. *There exists a constant C, which depends only on α, such that*

(i) $|\delta_1(t)| \le 2h_1^\alpha$ *and* $|\delta_1'(t)| \le Ch_1^{\alpha-1}$.

(ii) $|\delta_j(t)| \le Ch_j^2 x_j^{\alpha-2}$ *and* $|\delta_j'(t)| \le Ch_j x_j^{\alpha-2}$, $2 \le j \le n$.

Proof. The estimates in (i) follow immediately from $\delta_1(t) = x_1^{\alpha-1}t - t^\alpha$, $0 \le t \le x_1$, and $x_1 \le h_1$. Consider now $2 \le j \le n$. Since $\delta_j(x_{j-1}) = \delta_j(x_j) = 0$, δ_j' vanishes at some point in the interval (x_{j-1}, x_j), and therefore

$$|\delta_j'(t)| \le C(x_j - x_{j-1})x_j^{\alpha-2},$$

where we have used (4.1) to bound δ'' in the case $\alpha < 2$. So, the second part of (ii) follows from $x_j - x_{j-1} \le h_j$. Finally, the bound for δ_j follows immediately from the bound for its derivative using again that $\delta_j(x_{j-1}) = 0$ and $x_j - x_{j-1} \le h_j$. $\square$

Observe that if we apply a standard trace result in the polygonal domain Ω_h, the constant depends on h. However, since $\Gamma_{3,h}$ approximates Γ_3, a trace theorem with a constant independent of h can be derived from Theorem 3.2. This is the object of the next lemma.

LEMMA 4.3. *There exists a constant C independent of h such that, for all $v \in V_h$,*

$$\|x^r v\|_{L^2(\Gamma_{3,h})} \le C \|v\|_{H^1(\Omega_h)}$$

for $r > \alpha/2 - 1$ if $\alpha \ge 2$ and $r = 0$ if $\alpha < 2$.

Proof. Since $h_j \le Cx_j$, it follows from (ii) of Lemma 4.2 that

$$|\delta_j(t)| \le Cx_j^{\alpha-1}h_j.$$

Then, since v is linear in each triangle, we have

$$\begin{aligned}\int_{\Gamma_{3,h}^j} v^2 x^{2r} &= \int_{x_{j-1}}^{x_j} \left| v(\phi(t)) + \delta_j(t)\frac{\partial v}{\partial y}(\phi(t)) \right|^2 t^{2r}|\psi_j'(t)| \\ &\le C\int_{x_{j-1}}^{x_j} |v(\phi(t))|^2 t^{2r}|\psi_j'(t)| + C\int_{x_{j-1}}^{x_j} \left| \frac{\partial v}{\partial y}|_{\Gamma_3^j} \right|^2 |\delta_j(t)|^2 t^{2r}|\psi_j'(t)|.\end{aligned}$$

Since

$$\psi_j(t) = \left(t, \frac{x_j^\alpha - x_{j-1}^\alpha}{x_j - x_{j-1}}(t - x_{j-1}) + x_{j-1}^\alpha\right),$$

it follows that $|\psi_j'(t)| \sim |\phi'(t)| \sim C$, and thus

$$\int_{\Gamma_{3,h}^j} v^2 x^{2r} \le C\|x^r v\|_{0,\Gamma_3^j}^2 + Ch_j^3 x_j^{2\alpha-2+2r} \left|\frac{\partial v}{\partial y}|_{\Gamma_3^j}\right|^2. \tag{4.9}$$

If $j = 1$, we have

$$\left|\frac{\partial v}{\partial y}|_{\Gamma_3^1}\right|^2 \sim \left\|\frac{\partial v}{\partial y}\right\|_{L^2(T_1)}^2 h_1^{-1-\alpha},$$

and using that $h_1 \sim x_1$ we obtain

$$\|x^r v\|_{L^2(\Gamma_{3,h}^1)}^2 \le C\|x^r v\|_{L^2(\Gamma_3^1)}^2 + Ch_1^{\alpha+2r} \left\|\frac{\partial v}{\partial y}\right\|_{L^2(T_1)}^2,$$

while if $j > 1$ we have

$$\left|\frac{\partial v}{\partial y}|_{\Gamma_3^j}\right|^2 \sim \left\|\frac{\partial v}{\partial y}\right\|_{0,T_j}^2 h_j^{-2} x_j^{1-\alpha},$$

and then

$$\|x^r v\|_{L^2(\Gamma_{3,h}^j)}^2 \le C\|x^r v\|_{L^2(\Gamma_3^j)}^2 + Ch_j x_j^{\alpha-1+2r} \left\|\frac{\partial v}{\partial y}\right\|_{L^2(T_j)}^2.$$

Therefore, for every j we have

$$\|x^r v\|_{L^2(\Gamma_{3,h}^j)}^2 \le C\left(\|x^r v\|_{L^2(\Gamma_3^j)}^2 + \left\|\frac{\partial v}{\partial y}\right\|_{L^2(T_j)}^2\right), \qquad j = 1, \dots, n,$$

and the lemma follows by summing up the previous inequalities for $j = 1, \dots, n$ and using Theorem 3.2. □

LEMMA 4.4. *There exists a constant C independent of h such that, for all $v \in V_h$,*

(i) *if $\alpha < 2$ and $z' \in L^2(0,1)$,*

$$\left|\int_{\Gamma_3} gv - \int_{\Gamma_{3,h}} I_h(gv)\right| \le Ch\|z'\|_{L^2(0,1)}\|v\|_{H^1(\Omega_h)}.$$

(ii) *if $2 \le \alpha < 3$, $\beta > \alpha/2 - 1$, and $z' t^{-\beta} \in L^2(0,1)$,*

$$\left|\int_{\Gamma_3} gv - \int_{\Gamma_{3,h}} I_h(gv)\right| \le Ch\|z'\, t^{-\beta}\|_{L^2(0,1)}\|v\|_{H^1(\Omega_h)}.$$

Proof. First, we observe that since g and $\tilde{g}$ agree at the nodes on $\Gamma_3 \cap \Gamma_{3,h}$, we have

$$\begin{aligned}\left|\int_{\Gamma_3} gv - \int_{\Gamma_{3,h}} I_h(gv)\right| &= \left|\int_{\Gamma_3} gv - \int_{\Gamma_{3,h}} \tilde{g}v + \int_{\Gamma_{3,h}} (\tilde{g}v - I_h(\tilde{g}v))\right| \\ &\le \sum_{j=1}^{n} \left|\int_{\Gamma_3^j} gv - \int_{\Gamma_{3,h}^j} \tilde{g}v\right| + \sum_{j=1}^{n} \int_{\Gamma_{3,h}^j} |\tilde{g}v - I_h(\tilde{g}v)| \\ &=: I + II. \end{aligned} \tag{4.10}$$

For any $v \in V_h$, we have

$$\begin{aligned} I = \sum_{j=1}^{n} \left|\int_{\Gamma_3^j} gv - \int_{\Gamma_{3,h}^j} \tilde{g}v\right| &\le \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} |z(t)| \left|v(\phi(t))|\phi'(t)| - v(\psi_j(t))|\psi_j'(t)|\right| \\ &\le \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} |z(t)| \, |v(\phi(t)) - v(\psi_j(t))| \, |\phi'(t)| \\ &\qquad + \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} |z(t)||v(\psi_j(t))| \left||\phi'(t)| - |\psi_j'(t)|\right| \\ &\le \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} |z(t)| \left|\frac{\partial v}{\partial y}|_{\Gamma_3^j}\right| |\delta_j(t)| \, |\phi'(t)| \\ &\qquad + C \int_{x_{j-1}}^{x_j} |z(t)||v(\psi_j(t))| \left|\delta_j'(t)\right| \\ &=: \sum_{j=1}^{n} A_j + B_j \end{aligned}$$

and

$$II = \sum_{j=1}^{n} \int_{\Gamma_{3,h}^j} |\tilde{g}v - I_h(\tilde{g}v)| \le \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} |z(t)v(\psi_j(t)) - I_h(z(v \circ \psi_j))(t)| \, |\psi_j'(t)|.$$

Let

$$w_j(t) = (z(t) - \bar{z_j})v(\psi_j(t)), \qquad t \in I_j, j = 1, \ldots, n,$$

where $\bar{z_j}, j = 1, \ldots, n$, are constants to be chosen below. It follows that

$$II \le C \sum_{j=1}^{n} \int_{x_{j-1}}^{x_j} |w_j(t) - I_h w_j(t)| \le C \sum_{j=1}^{n} h_j \int_{x_{j-1}}^{x_j} |w_j'(t)|,$$

where we have used a standard L^1 interpolation error estimate. Since, for $t \in I_j$,

$$\begin{aligned} |w_j'(t)| &\le |z'(t)||v(\psi_j(t))| + |z(t) - \bar{z_j}| \left|\frac{\partial v}{\partial x}(\psi_j(t)) + \frac{\partial v}{\partial y}(\psi_j(t)) \frac{x_j^\alpha - x_{j-1}^\alpha}{x_j - x_{j-1}}\right| \\ &\le |z'(t)||v(\psi_j(t))| + C|z(t) - \bar{z_j}||\nabla v(\psi_j(t))|, \end{aligned}$$

thus,

$$II \le C \sum_{j=1}^{n} h_j \left(\int_{x_{j-1}}^{x_j} |z'(t)||v(\psi_j(t))| + \int_{x_{j-1}}^{x_j} |z(t) - \bar{z_j}||\nabla v(\psi_j(t))| \right).$$

Clearly, since $h_j \le h$ for any $j = 1, \dots, n$ it follows that

$$\begin{aligned} II &\le Ch \sum_{j=1}^{n} \left(\int_{x_{j-1}}^{x_j} |z'(t)||v(\psi_j(t))| + \int_{x_{j-1}}^{x_j} |z(t) - \bar{z_j}||\nabla v(\psi_j(t))| \right) \\ &=: Ch \sum_{j=1}^{n} C_j + D_j. \end{aligned}$$

Now, we consider the case $\alpha < 2$ and we prove the result given in (i).

For $j = 1$, using (i) of Lemma 4.2 and that

$$|\nabla v(\phi(t))| \sim \|\nabla v\|_{L^2(T_1)} h_1^{-\frac{\alpha+1}{2}}, \quad t \in I_1, \quad v \in V_h, \tag{4.11}$$

we obtain, for $\beta = \max\{0, \frac{3}{2} - \alpha\}$,

$$\begin{aligned} A_1 &\le Ch_1^{\alpha+\frac{1}{2}+\beta} \|zt^{-\beta}\|_{L^2(I_1)} |\nabla v|_{T_1}| \le Ch_1^{\frac{\alpha}{2}+\beta} \|zt^{-\beta}\|_{L^2(I_1)} \|\nabla v\|_{L^2(T_1)} \\ &\le Ch^{\left(\frac{\alpha}{2}+\beta\right)\frac{2}{3-\alpha}} \|zt^{-\beta}\|_{L^2(I_1)} \|\nabla v\|_{L^2(T_1)}. \end{aligned}$$

Hence, since $\alpha < 2$, then $\left(\frac{\alpha}{2} + \beta\right) \frac{2}{3-\alpha} \ge 1$, and therefore,

$$A_1 \le Ch \|zt^{-\beta}\|_{L^2(I_1)} \|\nabla v\|_{L^2(T_1)}. \tag{4.12}$$

Similary, for $\beta = \max\{0, \frac{5}{2} - \frac{3}{2}\alpha\}$ we have

$$\begin{aligned} B_1 &\le Ch_1^{\alpha+\beta-1} \|zt^{-\beta}\|_{L^2(I_1)} \|v\|_{L^2(\Gamma^1_{3,h})} \le h^{(\alpha+\beta-1)\frac{2}{3-\alpha}} \|zt^{-\beta}\|_{L^2(I_1)} \|v\|_{L^2(\Gamma^1_{3,h})} \\ &\le Ch \|zt^{-\beta}\|_{L^2(I_1)} \|v\|_{L^2(\Gamma^1_{3,h})}. \end{aligned}$$

For $j > 1$, using (ii) of Lemma 4.2 and that

$$|\nabla v(\phi(t))| \sim h_j^{-1} x_j^{\frac{1-\alpha}{2}} \|\nabla v\|_{L^2(T_j)}, \qquad t \in I_j, \tag{4.13}$$

since $h_j \le Chx_j^{\frac{\alpha-1}{2}}$, we have, for $\beta = \max\{0, \frac{3}{2} - \alpha\}$,

$$\begin{aligned} A_j &\le Ch_j^2 x_j^{\alpha+\beta-\frac{3}{2}} \|zt^{-\beta}\|_{L^2(I_j)} |\nabla v|_{T_j}| = Ch_j x_j^{\beta+\frac{\alpha}{2}-1} \|zt^{-\beta}\|_{L^2(I_j)} \|\nabla v\|_{L^2(T_j)} \\ &\le Chx_j^{\beta+\alpha-\frac{3}{2}} \|zt^{-\beta}\|_{L^2(I_j)} \|\nabla v\|_{L^2(T_j)} \\ &\le Ch \|zt^{-\beta}\|_{L^2(I_j)} \|\nabla v\|_{L^2(T_j)}. \end{aligned} \tag{4.14}$$

Similarly, for $j > 1$ and $\beta = \max\{0, \frac{5}{2} - \frac{3}{2}\alpha\}$, applying the Cauchy–Schwarz inequality we have that

$$\begin{aligned} B_j &\le Ch_j x_j^{\alpha-2+\beta} \|zt^{-\beta}\|_{L^2(I_j)} \|v\|_{L^2(\Gamma^j_{3,h})} \le Chx_j^{\frac{3}{2}\alpha-\frac{5}{2}+\beta} \|zt^{-\beta}\|_{L^2(I_j)} \|v\|_{L^2(\Gamma^j_{3,h})} \\ &\le Ch \|zt^{-\beta}\|_{L^2(I_j)} \|v\|_{L^2(\Gamma^j_{3,h})}. \end{aligned} \tag{4.15}$$

So, since $\frac{3}{2} - \alpha < \frac{5}{2} - \frac{3}{2}\alpha$, if we take $\beta = \max\{0, \frac{5}{2} - \frac{3}{2}\alpha\}$, we obtain for any j

$$\left| \int_{\Gamma_3^j} gv - \int_{\Gamma^j_{3,h}} \tilde{g}v \right| \le Ch \|zt^{-\beta}\|_{L^2(I_j)} \left(\|\nabla v\|_{L^2(T_j)} + \|v\|_{L^2(\Gamma^j_{3,h})} \right), \tag{4.16}$$

and adding for $j = 1, \dots, n$, we have

$$I \le Ch\|zt^{-\beta}\|_{L^2(0,1)} \left(\|\nabla v\|_{L^2(\Omega_h)} + \|v\|_{L^2(\Gamma_{3,h})}\right).$$

Now, since $z(0) = 0$ and $\beta < 1$, by using the Hardy inequality

$$\|zt^{-\beta}\|_{L^2(0,1)} \le C\|z'\|_{L^2(0,1)} \tag{4.17}$$

and Lemma 4.3 for the case $\alpha < 2$, we conclude that

$$I \le Ch\|z'\|_{L^2(0,1)}\|v\|_{H^1(\Omega_h)}. \tag{4.18}$$

On the other hand, we have that for all $j \ge 1$,

$$C_j = \int_{x_{j-1}}^{x_j} |z'(t)||v(\psi_j(t))| \le \|z'\|_{L^2(I_j)} \|v\|_{L^2(\Gamma_{3,h}^j)}. \tag{4.19}$$

Taking $\bar{z_j} = \frac{1}{x_j - x_{j-1}} \int_{x_{j-1}}^{x_j} z$, it follows from the Poincaré inequality that

$$\|z - \bar{z_j}\|_{L^2(I_j)} \le Ch_j\|z'\|_{L^2(I_j)}.$$

Then, using (4.11) for $j = 1$ and (4.13) for $j > 1$, we obtain

$$\begin{aligned} D_1 &\le Ch_1^{1-\frac{\alpha}{2}}\|z'\|_{L^2(I_1)}\|\nabla v\|_{L^2(T_1)} \le Ch^{\frac{2-\alpha}{3-\alpha}}\|z'\|_{L^2(I_1)}\|\nabla v\|_{L^2(T_1)}, \\ D_j &\le Ch_j^{\frac{1}{2}} x_j^{\frac{1-\alpha}{2}}\|z'\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)} \le Ch^{\frac{1}{2}}\|z'\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)}, \qquad j > 1, \end{aligned}$$

and therefore

$$D_j \le C\|z'\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)} \qquad \forall j \ge 1. \tag{4.20}$$

So, adding inequalities (4.19) and (4.20) for $j = 1, \dots, n$, we have that

$$II \le Ch\left(\|z'\|_{L^2(0,1)}\|v\|_{L^2(\Gamma_{3,h})} + \|z'\|_{L^2(0,1)}\|\nabla v\|_{L^2(\Omega_h)}\right),$$

and using Lemma 4.3 again we conclude that

$$II \le Ch\|z'\|_{L^2(0,1)}\|v\|_{H^1(\Omega_h)}. \tag{4.21}$$

From this inequality, (4.18), and (4.10) the proof of (i) concludes.

Now, consider the case $\alpha \ge 2$. By the same arguments used in the previous case we have

$$\begin{aligned} A_1 &\le Ch_1^{\alpha+\frac{1}{2}}\|z\|_{L^2(I_1)}|\nabla v|_{T_1}| \le Ch_1^{\frac{\alpha}{2}}\|z\|_{L^2(I_1)}\|\nabla v\|_{L^2(T_1)}, \\ &\le Ch^{\frac{\alpha}{3-\alpha}}\|z\|_{L^2(I_1)}\|\nabla v\|_{L^2(T_1)}, \end{aligned}$$

but $\frac{\alpha}{3-\alpha} \ge 1$, so

$$A_1 \le Ch\|z\|_{L^2(I_1)}\|\nabla v\|_{L^2(T_1)}. \tag{4.22}$$

For $j > 1$,

$$\begin{aligned} A_j &\le Ch_j^2 x_j^{\alpha-\frac{3}{2}}\|z\|_{L^2(I_j)}|\nabla v|_{T_j}| \le Ch_j x_j^{\frac{\alpha}{2}-1}\|z\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)} \\ &\le Chx_j^{\alpha-\frac{3}{2}}\|z\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)} \le Ch\|z\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)}. \end{aligned} \tag{4.23}$$

Similarly, for any $\beta \geq 0$ we have

$$B_1 \leq Ch_1^{\alpha-1}\|zt^{-\beta}\|_{L^2(I_1)}\|vx^{\beta}\|_{L^2(\Gamma^1_{3,h})} \leq h^{(\alpha-1)\frac{2}{3-\alpha}}\|zt^{-\beta}\|_{L^2(I_1)}\|vx^{\beta}\|_{L^2(\Gamma^1_{3,h})}$$

$$\leq Ch\|zt^{-\beta}\|_{L^2(I_1)}\|vx^{\beta}\|_{L^2(\Gamma^1_{3,h})} \tag{4.24}$$

and

$$B_j \leq Ch_j x_j^{\alpha-2}\|zt^{-\beta}\|_{L^2(I_j)}\|vx^{\beta}\|_{L^2(\Gamma^j_{3,h})} \leq Chx_j^{\frac{3}{2}\alpha-\frac{5}{2}}\|zt^{-\beta}\|_{L^2(I_j)}\|vx^{\beta}\|_{L^2(\Gamma^j_{3,h})}$$

$$\leq Ch\|zt^{-\beta}\|_{L^2(I_j)}\|vx^{\beta}\|_{L^2(\Gamma^j_{3,h})}. \tag{4.25}$$

Thus, adding inequalities (4.22), (4.23), (4.24), and (4.25) for $j = 1, \ldots, n$ we conclude that for any $\beta \geq 0$,

$$I \leq Ch\left(\|z\|_{L^2(0,1)}\|\nabla v\|_{L^2(\Omega_h)} + \|zt^{-\beta}\|_{L^2(0,1)}\|vx^{\beta}\|_{L^2(\Gamma_{3,h})}\right).$$

Taking $\frac{\alpha}{2} - 1 < \beta < 1$ and using the Hardy inequality (4.17) and our trace result for the case $2 \leq \alpha < 3$, we obtain

$$I \leq Ch\|z'\|_{L^2(0,1)}\|v\|_{H^1(\Omega_h)}. \tag{4.26}$$

On the other hand, for any j and $\frac{\alpha}{2} - 1 < \beta < 1$ it follows that

$$C_j \leq \|z't^{-\beta}\|_{L^2(I_j)}\|vx^{\beta}\|_{L^2(\Gamma^j_{3,h})},$$

and by using (4.11) for $j = 1$ and (4.13) for $j > 1$ we get

$$D_j \leq C\|z'\|_{L^2(I_j)}\|\nabla v\|_{L^2(T_j)}, \qquad j \geq 1.$$

Therefore, we conclude that for $\frac{\alpha}{2} - 1 < \beta < 1$,

$$II \leq Ch\left(\|z't^{-\beta}\|_{L^2(0,1)}\|vx^{\beta}\|_{L^2(\Gamma_{3,h})} + \|z'\|_{L^2(0,1)}\|\nabla v\|_{L^2(\Omega_h)}\right).$$

Hence, using Lemma 4.3 again we obtain

$$II \leq Ch\|z't^{-\beta}\|_{L^2(0,1)}\|v\|_{H^1(\Omega_h)}, \tag{4.27}$$

and thus, using (4.26) and (4.27) in (4.10), we conclude the proof of (ii). ∎

We can now prove our main theorem, which gives quasi-optimal error estimates in H^1 for the piecewise linear approximation on appropriate graded meshes.

THEOREM 4.2. *Let u be the solution of (2.1), and let $u_h \in V_h$ be its finite element approximation using the mesh $\mathcal{T}_h$. Assume $\alpha < 3$, $f \in L^2(\Omega)$, $z\,t^{-\frac{\alpha}{2}} \in L^2(0,1)$, and $z't^{-r} \in L^2(0,1)$, with $r = 0$ when $\alpha < 2$ and $r > \alpha/2 - 1$ when $\alpha \geq 2$.*

If the family of meshes satisfies (1), (2), (3), and (H), then there exists a constant C depending only on α, β, and σ such that

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch\sqrt{\log(1/h)}\,\left\{\|f\|_{L^2(\Omega)} + \|z\,t^{-(\alpha/2)}\|_{L^2(0,1)} + \|z'\,t^{-r}\|_{L^2(0,1)}\right\}.$$

Proof. In view of (2.2) and since $r > \alpha/2 - 1$, it is enough to prove

$$\|u - u_h\|_{H^1(\Omega)} \leq Ch\sqrt{\log(1/h)}\,\left\{\|u\|_{H^2(\Omega)} + \|z'\,t^{-r}\|_{L^2(0,1)}\right\}.$$

Since $\Omega \subset \Omega_h$, we have

$$\|u - u_h\|_{H^1(\Omega)} \le \|\tilde{u} - u_h\|_{H^1(\Omega_h)},$$

and therefore it is enough to prove that

$$\|\tilde{u} - u_h\|_{H^1(\Omega_h)} \le Ch\sqrt{\log(1/h)}\, \left\{\|u\|_{H^2(\Omega)} + \|z'\, t^{-r}\|_{L^2(0,1)}\right\}. \tag{4.28}$$

Using the Poincaré inequality, we have

$$\begin{aligned} \|\tilde{u} - u_h\|^2_{H^1(\Omega_h)} &\le C|\tilde{u} - u_h|^2_{H^1(\Omega_h)} \\ &= C\left[\int_{\Omega_h} \nabla(\tilde{u} - u_h)\cdot\nabla(\tilde{u} - \Pi\tilde{u}) + \int_{\Omega_h} \nabla(\tilde{u} - u_h)\cdot\nabla(\Pi\tilde{u} - u_h)\right], \end{aligned} \tag{4.29}$$

but we know from (4.2) and Theorem 4.1 that

$$|\tilde{u} - \Pi\tilde{u}|_{H^1(\Omega_h)} \le Ch\|\tilde{u}\|_{H^2_\alpha(\Omega_h)} \le Ch\|u\|_{H^2(\Omega)}. \tag{4.30}$$

Thus, for the first term in (4.29), using the Young inequality, we have

$$\int_{\Omega_h} \nabla(\tilde{u} - u_h)\cdot\nabla(\tilde{u} - \Pi\tilde{u}) \le \varepsilon|\tilde{u} - u_h|^2_{H^1(\Omega_h)} + C_\varepsilon h^2\|u\|^2_{H^2(\Omega)} \tag{4.31}$$

with ε to be chosen below.

Then, we have only to estimate the second term of (4.29). To simplify notation we introduce $w_h := \Pi\tilde{u} - u_h$. From (2.1) and (2.3) we have

$$\begin{aligned} \int_{\Omega_h} \nabla(\tilde{u} - u_h)\cdot\nabla w_h &= \int_{\Omega} \nabla(\tilde{u} - u_h)\cdot\nabla w_h + \int_{\Omega_h\setminus\Omega} \nabla(\tilde{u} - u_h)\cdot\nabla w_h \\ &= \int_{\Omega} \nabla u\cdot\nabla w_h + \int_{\Omega_h\setminus\Omega} \nabla\tilde{u}\cdot\nabla w_h - \int_{\Omega_h} \nabla u_h\cdot\nabla w_h \\ &= \int_{\Omega_h\setminus\Omega} \nabla\tilde{u}\cdot\nabla w_h + \int_{\Gamma_3} g w_h - \int_{\Gamma_{3,h}} I_h(g w_h). \end{aligned}$$

Then, from Lemmas 4.1 and 4.4, using (4.2) and again the Young inequality we obtain

$$\int_{\Omega_h} \nabla(\tilde{u} - u_h)\cdot\nabla w_h \le C_\varepsilon h^2 \log 1/h \left\{\|u\|^2_{H^2(\Omega)} + \|z'\, t^{-r}\|^2_{L^2(0,1)}\right\} + \varepsilon|w_h|^2_{H^1(\Omega_h)}. \tag{4.32}$$

Hence, from (4.30)

$$\begin{aligned} |w_h|^2_{H^1(\Omega_h)} &\le 2(|\Pi\tilde{u} - \tilde{u}|^2_{H^1(\Omega_h)} + |\tilde{u} - u_h|^2_{H^1(\Omega_h)}) \\ &\le Ch^2\|u\|^2_{H^2(\Omega)} + 2|\tilde{u} - u_h|^2_{H^1(\Omega_h)}. \end{aligned} \tag{4.33}$$

Therefore, from (4.31), (4.32), and (4.33) we get

$$\begin{aligned} &|\tilde{u} - u_h|^2_{H^1(\Omega_h)} \\ &\quad\le C_\varepsilon h^2\|u\|^2_{H^2(\Omega)} + C_\varepsilon h^2 \log 1/h \left\{\|u\|^2_{H^2(\Omega)} + \|z'\, t^{-r}\|^2_{L^2(0,1)}\right\} + C\varepsilon|\tilde{u} - u_h|^2_{H^1(\Omega_h)}, \end{aligned}$$

and so the result follows by choosing ε small enough and using the estimates given in (2.2). ∎

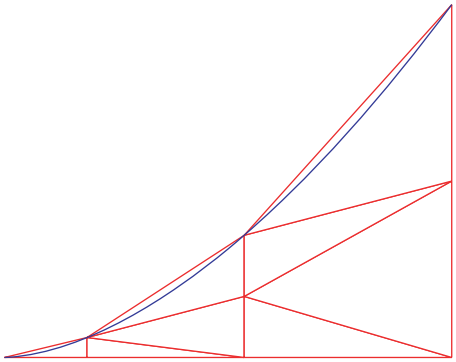


FIG. 4. *Graded mesh with* $\alpha = 2$ *and* $n = 3$.

TABLE 2
H^1 *order using graded meshes for* $\alpha = 2$.

Value of s	Order in number of nodes	Order in h
0.55	0.588	1.054
0.6	0.585	1.049
0.65	0.584	1.047
0.7	0.584	1.046
0.75	0.584	1.047
0.8	0.585	1.048
0.85	0.586	1.049
0.9	0.586	1.051
0.95	0.587	1.052

Now we show that meshes satisfying the hypotheses (1)–(3) and (H) can be constructed. To define the mesh $\mathcal{T}_h$, with $h = 1/n$ we use the following method given in [9, page 393] and [11].

1. Introduce the partition of the interval $(0, 1)$ given by

$$x_j = \left(\frac{j}{n}\right)^{\frac{2}{3-\alpha}} \qquad 0 \le j \le n.$$

2. Take the points $(x_j, 0)$ in Γ_1, (x_j, x_j^α) in Γ_3, and for $j > 1$, divide each of the vertical lines $\{(x_j, y) : 0 \le y \le x_j^\alpha\}$ uniformly into subintervals such that each has length $\sim x_j - x_{j-1}$.

Figure 4 shows an example of one of these meshes.

If N is the number of nodes in the partition $\mathcal{T}_h$, it can be proved that $h^2 \sim 1/N$ [9, page 393], [11]. Therefore, using these meshes we have the following error estimate in terms of the number of nodes:

$$\|u - u_h\|_{H^1(\Omega)} \le C\sqrt{\frac{\log N}{N}}\left\{\|f\|_{L^2(\Omega)} + \|z\, t^{-\frac{\alpha}{2}}\|_{L^2(0,1)} + \|z'\, t^{-r}\|_{L^2(0,1)}\right\}.$$

Observe that this estimate is quasi optimal. Indeed, up to the logarithmic factor, the order with respect to the number of nodes is the same as that obtained for a smooth problem using quasi-uniform meshes.

Table 2 shows the numerical results obtained with these graded meshes for example (2.1).

REFERENCES

[1] G. Acosta, M. G. Armentano, R. G. Durán, and A. L. Lombardi, *Nonhomogeneous Neumann problem for the Poisson equation in domains with an external cusp*, J. Math. Anal. Appl., 310 (2005), pp. 397–411.

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

[3] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., B. G. Teubner, Stuttgart, 1999.

[4] I. Babuška, R. B. Kellog, and J. Pitkaranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.

[5] J. H. Bramble and J. T. King, *A robust finite element method for nonhomogeneous Dirichlet problems in domains with curved boundaries*, Math. Comp., 63 (1994), pp. 1–17.

[6] S. K. Chua, *Extension theorems on weighted Sobolev spaces*, Indiana Math. J., 41 (1992), pp. 1027–1076.

[7] J. Duoandikoetxea Zuazo, *Análisis de Fourier*, Addison-Wesley Iberoamericana, Madrid, 1995.

[8] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.

[9] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.

[10] G. Hardy, J. E. Littlewood, and G. Polya, *Inequalities*, Cambridge University Press, Cambridge, UK, 1952.

[11] G. Raugel, *Résolution numérique par une méthode d'éléments finis du problème Dirichlet pour le laplacien dans un polygone*, C. R. Acad. Sci. Paris Ser. A, 286 (1978), pp. A791–A794.

[12] E. M. Stein, *Singular Integrals and Differentiability Properties of Functions*, Princeton University Press, Princeton, NJ, 1970.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 296–312

TWO-SCALE BOOLEAN GALERKIN DISCRETIZATIONS FOR FREDHOLM INTEGRAL EQUATIONS OF THE SECOND KIND*

FANG LIU† AND AIHUI ZHOU‡

Abstract. In this paper, some two-scale Boolean Galerkin discretizations are proposed and analyzed for a class of Fredholm integral equations of the second kind in multidimensions. It is shown by both theory and numerics that this type of multiscale discretization algorithm not only significantly reduces the number of degrees of freedom but also produces very accurate approximations.

Key words. Boolean Galerkin discretization, Fredholm integral equation, multidimension, two-scale

AMS subject classification. 65R20

DOI. 10.1137/050633007

1. Introduction.

It is known that it is a very challenging task to solve an integral equation in multidimensions by using the standard Galerkin discretizations, due to the fact that the resulting linear algebraic systems of integral equations involve dense matrices. However, integral equations of the second kind with smooth kernels in multidimensions have important applications in many areas such as physics, engineering, and finance; see, e.g., [3, 9, 14, 18, 19, 25, 26] and the references cited therein. Hence, an accurate and economic numerical scheme for solving an integral equation in multidimensions is highly desired. In this paper, we propose and analyze some two-scale finite element discretizations for solving multidimensional integral equations of the second kind with smooth kernels. These discretizations are nothing but several coupled standard Galerkin discretizations of two scales. The approximations obtained from the coupled two-scale discretizations have almost the same approximation accuracy, but the computational cost is reduced significantly, as compared with the standard Galerkin approximations. Moreover, it may be significant that this kind of discretization can be carried out in parallel.

We now use the three-dimensional case as an example to demonstrate the key idea of our discretizations. Let $R_{h_1,h_2,h_3}u$ be the standard Galerkin approximation to the exact solution u of an equation on a uniform cuboid grid with mesh size h_1 in the x-direction, h_2 in the y-direction, and h_3 in the z-direction, respectively. Then, a two-scale Boolean Galerkin approximation to u, which is a linear combination of different standard Galerkin approximations over two scales meshes, is constructed as follows:

$$B^h_{H,H,H}u = R_{h,H,H}u + R_{H,h,H}u + R_{H,H,h}u - 2R_{H,H,H}u.$$

*Received by the editors June 3, 2005; accepted for publication (in revised form) August 16, 2006; published electronically February 2, 2007. This work was supported in part by the Chinese National Natural Science Foundation under grant 10425105 and the National Basic Research Program under grant 2005CB321704.

http://www.siam.org/journals/sinum/45-1/63300.html

†LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, China, and Graduate School, Chinese Academy of Sciences, Beijing 100080, China (fliu@lsec.cc.ac.cn).

‡LSEC, Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, P.O. Box 2719, Beijing 100080, China (azhou@lsec.cc.ac.cn).

For this new approximation $B^h_{H,H,H}u$, we may establish the estimation (see Theorem 3.1)

$$\|B^h_{H,H,H}u - u_{h,h,h}\|_{0,2} = O(h + H^2)$$

if the piecewise constant functions are used as the approximate spaces and the exact solution u is smooth. Consequently, for example, we obtain an asymptotically optimal approximation $B^h_{H,H,H}u$ in parallel with only $O(h^{-2})$ degrees of freedom when $H = O(\sqrt{h})$ is taken, while the degrees of freedom for $u_{h,h,h}$ are of $O(h^{-3})$.

The two-scale Boolean Galerkin approximation studied in this paper is closely related to the multiscale Boolean interpolation which is first constructed in [10]. This multiscale Boolean technique is originally applied in [12] to reduce computational complexity in the numerical solution of partial differential equations (see also [4, 5, 6, 15, 20, 21, 22]). In the context of solving integral equations, the one-scale Boolean sum technique is used in [13] in conjunction with the degenerate kernel scheme to achieve a higher order of convergence and in [26] in designing a fast multiscale Boolean approximation scheme to get fast approximations. This multiscale Boolean technique is also analogous to the sparse grid method and the multiparameter extrapolation method discussed in [2, 11, 23, 28, 29]. Indeed, the sparse grid method may be viewed as an implicit version of the multiscale Boolean method in [4, 5, 6, 12, 20, 21, 22]. Instead of the multiscale Boolean technique, in the current work, we adopt a two-scale Boolean approach. Very recently, the two-scale Boolean discretization idea has been introduced to numerical partial differential equations in [16, 17]. It is shown that the two-scale Boolean approach is more flexible than the multiscale Boolean technique, which is a key for us to introduce the multiscale techniques to nonlinear problems [16]. Moreover, since the two-scale finite element approximations are computed on regular meshes, existing solvers can be used without any need for an explicit discretization on a sparse grid.

The rest of the paper is organized as follows. In section 2, some preliminary materials are provided. In section 3, three two-scale Boolean Galerkin discretizations are proposed and analyzed for solving multidimensional integral equations of the second kind. In section 4, several numerical experiments, which support our theory, are reported. Finally, in section 5, some concluding remarks are presented.

2. Preliminaries. We begin with the definition of notation. Let $\square = (0,1)^d (d \geq 2)$ be the unit cube in $\mathbb{R}^d$. We use $W^{s,p}(\square)$ to denote the standard Sobolev spaces of functions whose derivatives of order less than or equal to s are in $L^p(\square)$. We denote by $\mathbb{N}_0$ the set of all nonnegative integers. For a function $v \in W^{s,p}(\square)$, a point $x = (x_1, x_2, \ldots, x_d) \in \square$, and an index $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}_0^d$, we let

$$(D^\alpha w)(x) = \left(\frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_d} w}{\partial x_d^{\alpha_d}} \right)(x).$$

The norms and seminorms for the space $W^{s,p}(\square)$ are defined by

$$\|w\|_{s,p} = \begin{cases} \left(\sum_{|\alpha| \leq s} \|D^\alpha w\|_p^p \right)^{\frac{1}{p}} & \text{if } 1 \leq p < \infty, \\ \max_{|\alpha| \leq s} \|D^\alpha w\|_\infty \text{ if } p = \infty \end{cases}$$

and

$$|w|_{s,p} = \begin{cases} \left(\sum_{|\alpha|=s} \|D^\alpha w\|_p^p \right)^{\frac{1}{p}} & \text{if } 1 \le p < \infty, \\ \max_{|\alpha|=s} \|D^\alpha w\|_\infty \text{ if } p = \infty, \end{cases}$$

respectively (see, e.g., [1, 8]). When $p = 2$, we let $H^s(\Box) = W^{s,p}(\Box)$, $\|\cdot\|_s = \|\cdot\|_{s,p}$, and $\|\cdot\| = \|\cdot\|_0$, and we use $(\cdot,\cdot)$ for the standard $L^2(\Box)$ inner product. We will also use the negative norm $\|\cdot\|_{-1}$, which is defined for $w \in H^{-1}(\Box) = (H^1(\Box))^*$ by

$$\|w\|_{-1} = \sup_{\phi \in H^1(\Box)} \frac{(w,\phi)}{\|\phi\|_1}.$$

Throughout this paper, we shall use the letter C to denote a generic positive constant which may stand for different values at its different occurrences.

2.1. Interpolation and L^2-projection. The purposes of this subsection are to describe the multidimensional tensor product interpolation and L^2-projection operators. To this end, we first define the interpolation and L^2-projection operators in one dimension. Due to the higher-dimensional nature of these integral equations, we will use the spaces of the piecewise constant functions as our approximate ones. For a positive integer n we let $\mathbb{Z}_n = \{1, 2, \ldots, n\}$. Let $T^h((0,1))$ be a mesh of the interval $(0,1)$ with the mesh size $h \in [0,1)$, i.e.,

$$T^h((0,1)) = \{(x_{i-1}, x_i) : i \in \mathbb{Z}_n, x_0 = 0, x_n = 1\},$$

$$h = \max\{x_i - x_{i-1} : i \in \mathbb{Z}_n\}.$$

We use $\partial^2 T^h((0,1))$ to denote the set of the midpoints of the subintervals in the mesh $T^h((0,1))$, namely,

$$\partial^2 T^h((0,1)) = \left\{ \frac{x_{i-1} + x_i}{2} : i \in \mathbb{Z}_n \right\}.$$

Define the space of piecewise constant functions in $L^\infty((0,1))$ by setting

$$S^h((0,1)) = \{v \in L^\infty((0,1)) : v\,|_\tau \text{ is constant, } \tau \in T^h((0,1))\}.$$

Let $I_h : C((0,1)) \to S^h((0,1))$ be the Lagrange interpolation and $P_h : L^2((0,1)) \to S^h((0,1))$ be the L^2-projection operator defined by

$$(I_h w)(t) = w(t) \quad \forall t \in \partial^2 T^h((0,1)), \quad \forall w \in C((0,1))$$

and

$$\int_0^1 (P_h w - w)(t) v(t) dt = 0 \quad \forall v \in S^h((0,1)), \quad \forall w \in L^2((0,1)),$$

respectively.

We next describe the multidimensional notation. For $\mathbf{h} = (h_1, \ldots, h_d)$, where $h_j \in [0,1)$, construct a mesh of the unit cube $\Box$ in $\mathbb{R}^d$ by

$$T^{\mathbf{h}}(\Box) = T^{h_1}((0,1)) \times \cdots \times T^{h_d}((0,1))$$

with the associated space of piecewise constant functions on $\square$ by

$$S^{\mathbf{h}}(\square) = S^{h_1}((0,1)) \otimes \cdots \otimes S^{h_d}((0,1)).$$

We remark that $S^{\mathbf{h}}(\square)$ is the tensor product space of the spaces of piecewise constant functions on the interval $(0,1)$. The interpolation operator $I_{\mathbf{h}}$ from $C(\square)$ onto $S^{\mathbf{h}}(\square)$ is constructed by

$$I_{\mathbf{h}} = I_{h_1} \circ \cdots \circ I_{h_d},$$

while the L^2-projection operator $P_{\mathbf{h}}$ from $L^2(\square)$ onto $S^{\mathbf{h}}(\square)$ is set to be

$$P_{\mathbf{h}} = P_{h_1} \circ \cdots \circ P_{h_d}.$$

It is easy to prove by definition that for every $w \in L^2(\square)$, there holds

$$(w - P_{\mathbf{h}} w, v) = 0 \quad \forall\ v \in S^{\mathbf{h}}(\square). \tag{2.1}$$

For $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}_0^d$, we set

$$|\alpha| = \alpha_1 + \cdots + \alpha_d,$$

$$\mathbf{h}^\alpha = h_1^{\alpha_1} \cdots h_d^{\alpha_d}$$

and

$$\mathbf{h}\alpha = (h_1\alpha_1, \ldots, h_d\alpha_d).$$

We define the order $\alpha \le \beta$ for the elements $\alpha, \beta \in \{0,1\}^d$ by $\alpha_i \le \beta_i$ for all $i \in \mathbb{Z}_d$. Furthermore, we denote $\mathbf{0} = (0, \ldots, 0) \in \mathbb{R}^d$ and $\mathbf{e} = (1, \ldots, 1) \in \mathbb{R}^d$ and for $i \in \mathbb{Z}_d, \hat{\mathbf{e}}_i = \mathbf{e} - \mathbf{e}_i$ and $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^d$ whose ith component is one and zero otherwise.

We also need the notion of the mixed Sobolev space (see [16, 21, 22, 26]) defined for $\alpha \in \{0,1\}^d$ and $1 \le p \le \infty$ by

$$W^{\alpha,p}_{\mathrm{mix}}(\square) = \{w \in L^p(\square) : D^\beta w \in L^p(\square) \ \ \forall \ \ \beta \ \text{ with } \ \mathbf{0} \le \beta \le \alpha\}$$

with the associated norm given for $w \in W^{\alpha,p}_{\mathrm{mix}}(\square)$ by

$$\|w\|_{W^{\alpha,p}_{\mathrm{mix}}} = \left(\sum_{\mathbf{0} \le \beta \le \alpha} \|D^\beta w\|_{0,p}^2 \right)^{\frac{1}{2}}.$$

In particular, we denote

$$\mathcal{H}^\alpha(\square) = W^{\alpha,2}_{\mathrm{mix}}(\square).$$

In our discussion, we will also need the space

$$H^{1,2}(\square) = \left\{ w \in H^1(\square) : \frac{\partial^2 w}{\partial x_i \partial x_j} \in L^2(\square),\ i,j = 1, \ldots, d,\ i \ne j \right\} \tag{2.2}$$

with norm

$$\|w\|_{H^{1,2}} = \|w\|_1 + \sum_{\substack{i,j=1 \\ i \ne j}}^{d} \left\| \frac{\partial^2 w}{\partial x_i \partial x_j} \right\|.$$

It is seen that for all $i \in \mathbb{Z}_d$, there holds

$$\|w - P_{\mathbf{he}_i} w\|_{-1} + h_i \|w - P_{\mathbf{he}_i} w\| + h_i \|w - I_{\mathbf{he}_i} w\| \le C h_i^2 \|D^{\mathbf{e}_i} w\| \qquad \forall\ w \in \mathcal{H}^{\mathbf{e}_i}(\square). \tag{2.3}$$

Consequently, if $w \in H^1(\square)$, then

$$\begin{aligned} &\|w - P_{\mathbf{h}} w\|_{-1} + \max\{h_1, \dots, h_d\}(\|w - P_{\mathbf{h}} w\| + \|w - I_{\mathbf{h}} w\|) \\ &\qquad \le C \max\{h_1^2, \dots, h_d^2\} \|w\|_1. \end{aligned} \tag{2.4}$$

It can be verified that for any $\mathbf{0} \le \alpha, \beta \le \mathbf{e}$ with $\alpha + \beta \le \mathbf{e}$, there holds the identity that for every $w \in \mathcal{H}^{\alpha}(\square)$

$$D^{\alpha} I_{\mathbf{h}\beta} w = I_{\mathbf{h}\beta} D^{\alpha} w, \tag{2.5}$$

$$D^{\alpha} P_{\mathbf{h}\beta} w = P_{\mathbf{h}\beta} D^{\alpha} w. \tag{2.6}$$

Thus from the above basic properties of $P_{\mathbf{h}}$, we have the following proposition.

PROPOSITION 2.1. *Assume that $\mathbf{0} \le \alpha \le \mathbf{e}$ with $|\alpha| \ge 2$. If $w \in H^{1,2}(\square)$, then*

$$\left\| \prod_{0 \le \beta \le \alpha, |\beta| = 1} (I - P_{\mathbf{h}\beta}) w \right\|_{-1} + \max_{i \in \mathbb{Z}_d} h_i \left\| \prod_{0 \le \beta \le \alpha, |\beta| = 1} (I - P_{\mathbf{h}\beta}) w \right\| \le C \max_{i \in \mathbb{Z}_d} h_i^3 \|w\|_{H^{1,2}},$$

where I is the identity operator.

Given $\tau \in (0, 1)$, let $w_{\mathbf{h}\alpha + \tau\beta} \in S^{\mathbf{h}\alpha + \tau\beta}(\square)(\mathbf{0} \le \alpha, \beta \le \mathbf{e}$ and $\alpha + \beta = \mathbf{e})$, and set

$$\delta_{\tau}^{\alpha} w_{\mathbf{h}} = \prod_{\alpha_i \ne 0} \delta_{\tau}^{\mathbf{e}_i} w_{\mathbf{h}},$$

where

$$\delta_{\tau}^{\mathbf{e}_i} w_{\mathbf{h}} = w_{\mathbf{h}} - w_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i}, \quad i \in \mathbb{Z}_d.$$

If $d = 2$ and $\mathbf{h} = (h_1, h_2)$, for instance, then

$$\begin{aligned} \delta_{\tau}^{(1,0)} w_{h_1, h_2} &= w_{h_1, h_2} - w_{\tau, h_2}, \\ \delta_{\tau}^{(1,1)} w_{h_1, h_2} &= w_{h_1, h_2} - w_{h_1, \tau} - w_{\tau, h_2} + w_{\tau, \tau}. \end{aligned}$$

Given $h, H \in (0, 1)$, let $w_{H\mathbf{e}} \in S^{H\mathbf{e}}(\square)$, $w_{h\mathbf{e}} \in S^{h\mathbf{e}}(\square)$, and $w_{h\alpha + H\beta} \in S^{h\alpha + H\beta}(\square)$ $(\mathbf{0} \le \alpha, \beta \le \mathbf{e}, \alpha + \beta = \mathbf{e})$, and define

$$B_H^h w_{h\mathbf{e}} = w_{H\mathbf{e}} - \sum_{i=1}^{d} \delta_h^{\mathbf{e}_i} w_{H\mathbf{e}}.$$

PROPOSITION 2.2. *Let $i \in \mathbb{Z}_d$. If $w \in \mathcal{H}^{\mathbf{e}_i}(\square)$, then*

$$\|\delta_H^{\mathbf{e}_i} I_{\mathbf{h}} w\| + \|\delta_H^{\mathbf{e}_i} P_{\mathbf{h}} w\| \le C \max\{h_i, H\} \|D^{\mathbf{e}_i} w\| \tag{2.7}$$

and

$$\|\delta_H^{\mathbf{e}_i} P_{\mathbf{h}} w\|_{-1} \le C \max\{h_i^2, H^2\} \|D^{\mathbf{e}_i} w\|. \tag{2.8}$$

Proof. The estimate (2.7) follows directly from the definition of $\delta_H^{\mathbf{e}_i}$ and (2.3). We now prove (2.8). For $\phi \in H^1(\square)$, since

$$((P_{\mathbf{h}} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w, P_{H\mathbf{e}_i} P_{\mathbf{h}\hat{\mathbf{e}}_i}\phi) = 0$$

and since there holds

$$\|(I - P_{H\mathbf{e}_i})P_{\mathbf{h}\hat{\mathbf{e}}_i}\phi\| \le CH\|D^{\mathbf{e}_i}P_{\mathbf{h}\hat{\mathbf{e}}_i}\phi\| \le CH\|P_{\mathbf{h}\hat{\mathbf{e}}_i}D^{\mathbf{e}_i}\phi\| \le CH\|\phi\|_1,$$

we conclude that

$$\begin{aligned}|((P_{\mathbf{h}} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w, \phi)| &= |((P_{\mathbf{h}} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w, P_{\mathbf{h}\hat{\mathbf{e}}_i}\phi)| \\ &= |((P_{\mathbf{h}} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w, (I - P_{H\mathbf{e}_i})P_{\mathbf{h}\hat{\mathbf{e}}_i}\phi)| \le CH\|(P_{\mathbf{h}} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w\|\|\phi\|_1,\end{aligned}$$

which ensures that

$$\|(P_{\mathbf{h}} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w\|_{-1} \le CH(\|(P_{\mathbf{h}\hat{\mathbf{e}}_i} - P_{\mathbf{h}\hat{\mathbf{e}}_i + h_i\mathbf{e}_i})w\| + \|(P_{\mathbf{h}\hat{\mathbf{e}}_i} - P_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i})w\|).$$

Using the estimation

$$\|(P_{\mathbf{h}\hat{\mathbf{e}}_i} - P_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i})w\| \le C\tau\|D^{\mathbf{e}_i}w\|$$

for $\tau \in \{h_i, H\}$, we then have (2.8). This completes the proof. ∎

From Proposition 2.2, we immediately obtain the following proposition.

PROPOSITION 2.3. *If $w \in H^{1,2}(\square)$, then*

$$\|B_H^h I_{h\mathbf{e}}w - I_{h\mathbf{e}}w\| + \|B_H^h P_{h\mathbf{e}}w - P_{h\mathbf{e}}w\| \le C\max\{h_i^2, H^2\}\|w\|_{H^{1,2}} \tag{2.9}$$

and

$$\|B_H^h P_{h\mathbf{e}}w - P_{h\mathbf{e}}w\|_{-1} \le C\max\{h_i^3, H^3\}\|w\|_{H^{1,2}}. \tag{2.10}$$

Proof. For $w_\tau = I_\tau w$ or $P_\tau w$ when $\tau = (\tau_1, \tau_2, \ldots, \tau_d)$, we have

$$w_{h\mathbf{e}} = w_{H\mathbf{e}} + \sum_{|\alpha|=1}^{d} (-1)^{|\alpha|}\delta_h^\alpha w_{H\mathbf{e}}.$$

Thus we obtain

$$B_H^h w_{h\mathbf{e}} - w_{h\mathbf{e}} = -\sum_{|\alpha|=2}^{d} (-1)^{|\alpha|}\delta_h^\alpha w_{H\mathbf{e}}, \tag{2.11}$$

which together with Proposition 2.2 completes the proof. ∎

2.2. The standard Galerkin approximation. Suppose that the kernel $k \in L^2(\square \times \square)$. Then the operator $K : L^2(\square) \to L^2(\square)$ defined by

$$(Ku)(x) = \int_\square k(x,y)u(y)dy, \quad x \in \square, \quad u \in L^2(\square),$$

is a compact integral operator (see, e.g., page 277 of [27]). Consider the Fredholm integral equation of the second kind

$$u + Ku = f, \tag{2.12}$$

where $u \in L^2(\square)$ is the unknown and $f \in L^2(\square)$ is a given function. We assume that -1 is not an eigenvalue of K; namely, the inverse operator $(I+K)^{-1}$ exists as a bounded operator on $L^2(\square)$. In other words, (2.12) has a unique solution u in $L^2(\square)$ satisfying

$$\|u\| \le C\|f\|.$$

The standard Galerkin projection $R_{\mathbf{h}} : L^2(\square) \to S^{\mathbf{h}}(\square)$ is defined by

$$((I+K)(R_{\mathbf{h}}u - u), v) = 0 \quad \forall \ \ v \in S^{\mathbf{h}}(\square), \tag{2.13}$$

namely,

$$(I + P_{\mathbf{h}}K)R_{\mathbf{h}}u = P_{\mathbf{h}}(u + Ku). \tag{2.14}$$

Associated with K and $R_{\mathbf{h}}$, we may define $K^* : L^2(\square) \to L^2(\square)$ by

$$(K^*u)(x) = \int_{\square} k(y,x)u(y)dy, \quad x \in \square, \quad u \in L^2(\square),$$

and $R^*_{\mathbf{h}} : L^2(\square) \to S^{\mathbf{h}}(\square)$ by

$$((I+K^*)(R^*_{\mathbf{h}}u - u), v) = 0 \ \ \forall \ \ v \in S^{\mathbf{h}}(\square).$$

It is seen that both $(I+K)^{-1}$ and $(I+K^*)^{-1}$ exist as bounded operators on $H^1(\square)$ if $k \in H^1(\square \times \square)$ (see, e.g., [24]).

In the next proposition, we provide some standard estimates for the standard Galerkin projections $R_{\mathbf{h}}u$ and $R^*_{\mathbf{h}}u$ for any $u \in \mathcal{H}^{\mathbf{e}}(\square)$ (see, e.g., [7, 14, 24, 26]).

PROPOSITION 2.4. *If $k \in L^2(\square \times \square)$ and $u \in H^1(\square)$, then*

$$\|u - R_{\mathbf{h}}u\| + \|u - R^*_{\mathbf{h}}u\| \le C \inf_{v \in S^{\mathbf{h}}(\square)} \|u - v\| \le C \max\{h_1, \ldots, h_d\}\|u\|_1. \tag{2.15}$$

For the Galerkin approximation $u_{\mathbf{h}} \equiv R_{\mathbf{h}}u$ to the exact solution u of (2.12), we define the *iterated Galerkin approximation* by

$$\tilde{u}_{\mathbf{h}} = f - Ku_{\mathbf{h}}. \tag{2.16}$$

The following estimation is also classic and can be found in the literature (see, e.g., [24, 26]).

PROPOSITION 2.5. *If $k \in H^1(\square \times \square)$, then*

$$\|u - u_{\mathbf{h}}\|_{-1} + \|u - \tilde{u}_{\mathbf{h}}\| \le C \max\{h_1, \ldots, h_d\}\|u - u_{\mathbf{h}}\|.$$

In our analysis, we need the following two results.

PROPOSITION 2.6. *Assume that $u \in \mathcal{H}^{\mathbf{e}_i}(\square)(i \in \mathbb{Z}_d)$. If $k \in H^1(\square \times \square)$, then*

$$\|D^{\mathbf{e}_i} R_{\mathbf{h}\hat{\mathbf{e}}_i} u\| \le C(\|D^{\mathbf{e}_i}u\| + \|u\|). \tag{2.17}$$

Proof. It is obtained from (2.14) that

$$D^{\mathbf{e}_i}(I + P_{\mathbf{h}\hat{\mathbf{e}}_i}K)R_{\mathbf{h}\hat{\mathbf{e}}_i}u = D^{\mathbf{e}_i}P_{\mathbf{h}\hat{\mathbf{e}}_i}(I+K)u,$$

which together with (2.6) leads to

$$D^{\mathbf{e}_i} R_{\mathbf{h}\hat{\mathbf{e}}_i} u = P_{\mathbf{h}\hat{\mathbf{e}}_i}\left(D^{\mathbf{e}_i}(I+K)u - D^{\mathbf{e}_i}KR_{\mathbf{h}\hat{\mathbf{e}}_i}u\right).$$

Hence from

$$\|P_{\mathbf{h}\hat{\mathbf{e}}_i}u\| + \|R_{\mathbf{h}\hat{\mathbf{e}}_i}u\| \le C\|w\|,$$

we arrive at (2.17). This completes the proof. ∎

PROPOSITION 2.7. *Assume that* $u \in \mathcal{H}^{\mathbf{e}_i}(\square)(i \in \mathbb{Z}_d)$.

(1) *If* $k \in L^2(\square \times \square)$, *then*

$$\|\delta_H^{\mathbf{e}_i} R_{\mathbf{h}} u\| \le C \max\{h_i, H\}\|D^{\mathbf{e}_i}u\|. \tag{2.18}$$

(2) *If* $k \in H^1(\square \times \square)$, *then*

$$\|\delta_H^{\mathbf{e}_i} R_{\mathbf{h}} u\|_{-1} \le C \max\{h_i^2, H^2\}\|D^{\mathbf{e}_i}u\|. \tag{2.19}$$

Proof. The estimate (2.18) can be obtained directly from (2.15). It remains to prove (2.19).

Because of the triangle inequality

$$\|\delta_H^{\mathbf{e}_i} R_{\mathbf{h}} u\|_{-1} \le \|R_{\mathbf{h}}u - R_{\mathbf{h}\hat{\mathbf{e}}_i}u\|_{-1} + \|R_{\mathbf{h}\hat{\mathbf{e}}_i}u - R_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i}u\|_{-1}, \tag{2.20}$$

we shall estimate $\|R_{\mathbf{h}}u - R_{\mathbf{h}\hat{\mathbf{e}}_i}u\|_{-1}$ and $\|R_{\mathbf{h}\hat{\mathbf{e}}_i}u - R_{\mathbf{h}\hat{\mathbf{e}}_i + H\mathbf{e}_i}u\|_{-1}$.

To this end, for any $\phi \in H^1(\square)$, we let

$$\psi = R^*_{\mathbf{h}\hat{\mathbf{e}}_i}(I + K^*)^{-1}\phi.$$

Since Proposition 2.6 is also true when $R_{\mathbf{h}\hat{\mathbf{e}}_i}$ is replaced by $R^*_{\mathbf{h}\hat{\mathbf{e}}_i}$, we have

$$\|D^{\mathbf{e}_i} R^*_{\mathbf{h}\hat{\mathbf{e}}_i}(I + K^*)^{-1}\phi\| \le C(\|D^{\mathbf{e}_i}(I + K^*)^{-1}\phi\| + \|(I + K^*)^{-1}\phi\|),$$

which together with (see, e.g., [24])

$$\|(I + K^*)^{-1}\phi\|_1 \le C\|\phi\|_1$$

yields

$$\|D^{\mathbf{e}_i}\psi\| \le C\|\phi\|_1. \tag{2.21}$$

Note that $P_{\tau\mathbf{e}_i}\psi \in S^{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i}(\square)$ for $\tau = h_i$ or H, we obtain

$$\begin{aligned}((R_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i} - R_{\mathbf{h}\hat{\mathbf{e}}_i})u, \phi) &= ((R_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i} - R_{\mathbf{h}\hat{\mathbf{e}}_i})u, (I + K^*)\psi) \\ &= ((I + K)(R_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i} - R_{\mathbf{h}\hat{\mathbf{e}}_i})u, \psi) \\ &= ((I + K)(R_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i} - R_{\mathbf{h}\hat{\mathbf{e}}_i})u, (I - P_{\tau\mathbf{e}_i})\psi)\,.\end{aligned}$$

Since (2.18) is also true when H is replaced by h_i, it follows from using (2.18) and the approximation property of $S^{\tau\mathbf{e}_i}(\square)$ that

$$|((R_{\mathbf{h}\hat{\mathbf{e}}_i + \tau\mathbf{e}_i} - R_{\mathbf{h}\hat{\mathbf{e}}_i})u, \phi)| \le C\tau^2\|D^{\mathbf{e}_i}u\|\|D^{\mathbf{e}_i}\psi\| \le C\tau^2\|D^{\mathbf{e}_i}u\|\|\phi\|_1,$$

which leads to (2.19). ∎

3. The two-scale Boolean Galerkin approximation. This section is devoted to designing and analyzing the two-scale Boolean Galerkin approximation method for solving (2.12). Suppose that $k \in H^1(\square \times \square)$.

Given $H \in (0,1)$, we assume that $H \gg h$ and $T^H((0,1)) \subset T^h((0,1))$. Then the *two-scale Boolean Galerkin approximation* and the *two-scale Boolean iterated Galerkin approximation* are constructed as follows:

$$B^h_{H\mathbf{e}}u = \sum_{i=1}^{d} R_{H\hat{\mathbf{e}}_i + h\mathbf{e}_i}u - (d-1)R_{H\mathbf{e}}u,$$

$$\tilde{B}^h_{H\mathbf{e}}u = \sum_{i=1}^{d} \tilde{R}_{H\hat{\mathbf{e}}_i + h\mathbf{e}_i}u - (d-1)\tilde{R}_{H\mathbf{e}}u.$$

For instance, when $d = 3$, we have

$$B^h_{H,H,H}u = B^h_{H\mathbf{e}}u = R_{h,H,H}u + R_{H,h,H}u + R_{H,H,h}u - 2R_{H,H,H}u,$$

$$\tilde{B}^h_{H,H,H}u = \tilde{B}^h_{H\mathbf{e}}u = \tilde{R}_{h,H,H}u + \tilde{R}_{H,h,H}u + \tilde{R}_{H,H,h}u - 2\tilde{R}_{H,H,H}u,$$

where $\tilde{R}_{h,k,l}u = f - Ku_{h,k,l}$. We want to discuss the superclose property of the two-scale Boolean Galerkin approximation $B^h_{H\mathbf{e}}u$ and investigate the superconvergence property of the Boolean iterated Galerkin approximation $\tilde{B}^h_{H\mathbf{e}}u$.

To analyze the error of the two-scale Boolean Galerkin approximations, we need to establish some estimates for the standard Galerkin projection. First, we need the following conclusions for the L^2-projection.

LEMMA 3.1. *For $i \in \mathbb{Z}_d$, set*

$$g = P_{\mathbf{he}_i}(I + K)(I - P_{\mathbf{he}_i})u. \tag{3.1}$$

If $u \in \mathcal{H}^{\mathbf{e}_i}(\square)$, then

$$\|g\| + h_i\|D^{\mathbf{e}_j}g\| \le Ch_i^2\|D^{\mathbf{e}_i}u\| \quad \forall j \in \mathbb{Z}_d \setminus \{i\}. \tag{3.2}$$

Proof. For $v \in S^{\mathbf{he}_i}(\square)$, we set

$$\phi = (I + K)^* v.$$

Therefore, using the definition of g, we conclude that

$$(g, v) = ((I + K)(I - P_{\mathbf{he}_i})u, v) = ((I - P_{\mathbf{he}_i})u, \phi)\,.$$

Noting that there holds the identity $(I - P_{\mathbf{he}_i})^2 = I - P_{\mathbf{he}_i}$ and noting that the operator $I - P_{\mathbf{he}_i}$ is self-adjoint, it follows that

$$(g, v) = ((I - P_{\mathbf{he}_i})u, (I - P_{\mathbf{he}_i})\phi)\,.$$

The fact that

$$v = P_{\mathbf{he}_i}v \quad \forall v \in S^{\mathbf{he}_i}(\square)$$

implies

$$(g, v) = ((I - P_{\mathbf{he}_i})u, (I - P_{\mathbf{he}_i})K^* v) \quad \forall v \ \in S^{\mathbf{he}_i}(\square).$$

Consequently,

$$|(g,v)| \le Ch_i^2 \|D^{\mathbf{e}_i} u\| \|v\| \quad \forall v \ \in S^{\mathbf{he}_i}(\square).$$

Choosing $v = g$ in the last inequality produces

$$\|g\| \le Ch_i^2 \|D^{\mathbf{e}_i} u\|.$$

Finally, from $P_{\mathbf{he}_i}(I - P_{\mathbf{he}_i}) = 0$, we obtain

$$D^{\mathbf{e}_j} g = D^{\mathbf{e}_j} P_{\mathbf{he}_i} K(I - P_{\mathbf{he}_i})u = P_{\mathbf{he}_i} D^{\mathbf{e}_j} K(I - P_{\mathbf{he}_i})u$$

if $j \in \mathbb{Z}_d \setminus \{i\}$. Hence we have

$$\|D^{\mathbf{e}_j} g\| \le C\|(I - P_{\mathbf{he}_i})u\|, \quad j \in \mathbb{Z}_d \setminus \{i\},$$

which together with (2.3) leads to

$$\|D^{\mathbf{e}_j} g\| \le Ch_i \|D^{\mathbf{e}_i} u\|.$$

This completes the proof. ∎

Using Lemma 3.1, we then obtain the following proposition.

PROPOSITION 3.1. *Suppose that $u \in \mathcal{H}^{\mathbf{e}_i}(\square)(i \in \mathbb{Z}_d)$. Then there exists a function $\psi \in \mathcal{H}^{\mathbf{e}_i}(\square)$ satisfying*

$$R_{\mathbf{h}}((I - P_{\mathbf{he}_i})u) = R_{\mathbf{h}}\psi \tag{3.3}$$

and

$$\|\psi\| + h_i \|D^{\mathbf{e}_j}\psi\| \le Ch_i^2 \|D^{\mathbf{e}_i} u\| \quad \forall j \in \mathbb{Z}_d \setminus \{i\}. \tag{3.4}$$

Proof. Set $g = P_{\mathbf{he}_i}(I + K)(I - P_{\mathbf{he}_i})u$. Then from Lemma 3.1, we conclude that

$$\|g\| + h_i \|D^{\mathbf{e}_j} g\| \le Ch_i^2 \|D^{\mathbf{e}_i} u\| \quad \forall j \in \mathbb{Z}_d \setminus \{i\}. \tag{3.5}$$

If $\psi = (I + K)^{-1} g$, then $\psi \in \mathcal{H}^{\mathbf{e}_i}(\square)$ satisfies (3.3) (see [24]) and has the estimates

$$\|\psi\| \le C\|g\|, \tag{3.6}$$

$$\|D^{\mathbf{e}_j}\psi\| \le C(\|D^{\mathbf{e}_j} g\| + \|g\|) \quad \forall j \in \mathbb{Z}_d \setminus \{i\}. \tag{3.7}$$

Combining (3.5), (3.6), and (3.7), we get (3.4). ∎

Now we turn to studying the hierarchical surplus, the difference between the two-scale Boolean Galerkin and the standard Galerkin approximations.

PROPOSITION 3.2. *Suppose that $\tau \in \{h, H\}^d, i, j \in \mathbb{Z}_d$, and $i \ne j$. If $u \in H^1(\square)$, then*

$$\|\delta_h^{\mathbf{e}_j} R_\tau (I - P_{\tau\mathbf{e}_i})u\| \le CH^2 \|u\|_1 \tag{3.8}$$

and

$$\|\delta_h^{\mathbf{e}_j} R_\tau (I - P_{\tau\mathbf{e}_i})u\|_{-1} \le CH^3 \|w\|_1. \tag{3.9}$$

Proof. Let $\psi \in H^1(\square)$ satisfy (3.3) when $\mathbf{h}$ is replaced by τ. By Propositions 3.1 and 2.7, we have

$$\|\delta_h^{\mathbf{e}_j} R_\tau (I - P_{\tau\mathbf{e}_i})u\| = \|\delta_h^{\mathbf{e}_j} R_\tau \psi\| \le CH \|D^{\mathbf{e}_j}\psi\| \le CH^2 \|u\|_1$$

and

$$\|\delta_h^{\mathbf{e}_j} R_\tau (I - P_{\tau \mathbf{e}_i}) u\|_{-1} = \|\delta_h^{\mathbf{e}_j} R_\tau \psi\|_{-1} \le C H^2 \|D^{\mathbf{e}_j} \psi\| \le C H^3 \|u\|_1. \qquad \square$$

Next, we estimate the error of the two-scale Boolean Galerkin approximation.

THEOREM 3.1. *If $u \in H^{1,2}(\square)$, then*

$$\|B_{H\mathbf{e}}^h u - R_{h\mathbf{e}} u\| \le C H^2 \|u\|_{H^{1,2}}$$

and

$$\|B_{H\mathbf{e}}^h u - R_{h\mathbf{e}} u\|_{-1} \le C H^3 \|u\|_{H^{1,2}}.$$

Consequently,

$$\|u - B_{H\mathbf{e}}^h u\| \le C(h + H^2) \|u\|_{H^{1,2}}$$

and

$$\|u - B_{H\mathbf{e}}^h u\|_{-1} \le C(h^2 + H^3) \|u\|_{H^{1,2}}.$$

Proof. For $w_{h\mathbf{e}} \in S^{h\mathbf{e}}(\square)$, define

$$\delta_H w_{h\mathbf{e}} = B_H^h w_{h\mathbf{e}} - w_{h\mathbf{e}}.$$

Then (2.11) implies

$$\delta_H w_{h\mathbf{e}} = - \sum_{|\beta|=2}^{d} (-1)^{|\beta|} \delta_h^\beta w_{H\mathbf{e}}. \tag{3.10}$$

From the identity

$$I - P_{\mathbf{h}} = - \sum_{0 \le \alpha \le \mathbf{e}, |\alpha| \ge 1} (-1)^{|\alpha|} \prod_{0 \le \beta \le \alpha, |\beta| = 1} (I - P_{\mathbf{h}\beta}),$$

we obtain

$$\|\delta_H R_{h\mathbf{e}} u\| \le \|\delta_H R_{h\mathbf{e}} P_{h\mathbf{e}} u\| + \sum_{0 \le \alpha \le \mathbf{e}, |\alpha| \ge 1} \left\| \delta_H R_{h\mathbf{e}} \prod_{0 \le \beta \le \alpha, |\beta| = 1} (I - P_{h\beta}) u \right\|$$

or

$$\|\delta_H R_{h\mathbf{e}} u\| \le \|\delta_H P_{h\mathbf{e}} u\| + \sum_{0 \le \alpha, |\alpha| = 1} \|\delta_H R_{h\mathbf{e}} (I - P_{h\alpha}) u\|$$

$$+ \sum_{0 \le \alpha \le \mathbf{e}, |\alpha| \ge 2} \left\| \delta_H R_{h\mathbf{e}} \prod_{0 \le \beta \le \alpha, |\beta| = 1} (I - P_{h\beta}) u \right\|.$$

Using the stability

$$\|R_{\mathbf{h}} w\| + \|P_{\mathbf{h}} w\| \le C \|w\| \quad \forall w \in L^2(\square),$$

we can then estimate the term

$$\sum_{0\le\alpha\le\mathbf{e},|\alpha|\ge 2}\left\|\delta_H R_{h\mathbf{e}}\prod_{0\le\beta\le\alpha,|\beta|=1}(I-P_{h\beta})u\right\|$$

as follows:

$$\sum_{0\le\alpha\le\mathbf{e},|\alpha|\ge 2}\left\|\delta_H R_{h\mathbf{e}}\prod_{0\le\beta\le\alpha,|\beta|=1}(I-P_{h\beta})u\right\|$$
$$\le C\sum_{0\le\alpha\le\mathbf{e},|\alpha|\ge 2}\max_{\tau\in\{h,H\}^d}\left\|\prod_{0\le\beta\le\alpha,|\beta|=1}(I-P_{\tau\beta})u\right\|,$$

while the identity

$$\sum_{0\le\alpha,|\alpha|=1}\delta_H R_{h\mathbf{e}}(I-P_{h\alpha})=-\sum_{0\le\alpha,|\alpha|=1}\sum_{|\beta|=2}^{d}(-1)^{|\beta|}\delta_h^{\beta}R_{H\mathbf{e}}(I-P_{H\alpha}),$$

which can be derived directly from (3.10), leads to

$$\left\|\sum_{0\le\alpha,|\alpha|=1}\delta_H R_{h\mathbf{e}}(I-P_{h\alpha})u\right\|\le C\max_{\tau\in\{h,H\}^d}\sum_{i,j\in\mathbb{Z}_d,i\ne j}\|\delta_h^{\mathbf{e}_j}R_\tau(I-P_{\tau\mathbf{e}_i})u\|.$$

Hence we have

$$\|\delta_H R_{h\mathbf{e}}u\|\le\|\delta_H P_{h\mathbf{e}}u\|+C\sum_{i,j\in\mathbb{Z}_d,i\ne j}\|\delta_h^{\mathbf{e}_j}R_{H\mathbf{e}}(I-P_{H\mathbf{e}_i})u\|$$
$$+C\sum_{0\le\alpha\le\mathbf{e},|\alpha|\ge 2}\max_{\tau\in\{h,H\}^d}\left\|\prod_{0\le\beta\le\alpha,|\beta|=1}(I-P_{\tau\beta})u\right\|.$$

Finally, we prove the first estimate of this theorem from the above estimate and Propositions 2.1, 2.3, and 3.2. The second estimate can be proved similarly. ∎

As for the two-scale Boolean iterated Galerkin approximation $\tilde{B}^h_{H\mathbf{e}}u$, we obtain the following theorem.

THEOREM 3.2. *If $u\in H^{1,2}(\square)$, then*

$$\|\tilde{B}^h_{H\mathbf{e}}u-\tilde{R}_{h\mathbf{e}}u\|\le CH^3\|u\|_{H^{1,2}}.$$

Consequently,

$$\|u-\tilde{B}^h_{H\mathbf{e}}u\|\le C(h^2+H^3)\|u\|_{H^{1,2}}.$$

Proof. By the definition of $\tilde{R}_{h\mathbf{e}}u$ and $B^h_{H\mathbf{e}}u$, we have the identity

$$\tilde{B}^h_{H\mathbf{e}}u-\tilde{R}_{h\mathbf{e}}u=-K(B^h_{H\mathbf{e}}u-R_{h\mathbf{e}}u).$$

Hence we obtain

$$\|\tilde{B}^h_{H\mathbf{e}}u-\tilde{R}_{h\mathbf{e}}u\|=\|-K(B^h_{H\mathbf{e}}u-R_{h\mathbf{e}}u)\|$$
$$\le C\|B^h_{H\mathbf{e}}u-R_{h\mathbf{e}}u\|_{-1}\le CH^3\|u\|_{H^{1,2}}.$$

This completes the proof. □

Finally, we construct a defect correction approximation $R_{H\mathbf{e}}u+\tilde{B}^h_{H\mathbf{e}}u-R_{H\mathbf{e}}\tilde{B}^h_{H\mathbf{e}}u$, for which we have the following theorem.

THEOREM 3.3. *If* $u \in H^{1,2}(\square)$, *then*

$$\|u - R_{H\mathbf{e}}u - \tilde{B}^h_{H\mathbf{e}}u + R_{H\mathbf{e}}\tilde{B}^h_{H\mathbf{e}}u\| \le CH(h^2 + H^3)\|u\|_{H^{1,2}}. \tag{3.11}$$

Proof. From the definition of $\tilde{B}^h_{H,H,H}u$, we get

$$(I - R_{H\mathbf{e}})(u - \tilde{B}^h_{H\mathbf{e}}u) = -(I - R_{H\mathbf{e}})K(u - B^h_{H\mathbf{e}}u).$$

Thus, Proposition 2.4 and Theorem 3.2 yield

$$\begin{aligned}\|(I - R_{H\mathbf{e}})(u - \tilde{B}^h_{H\mathbf{e}}u)\| &\le CH\| - K(u - B^h_{H\mathbf{e}}u)\|_1 \\ &\le CH\|u - B^h_{H\mathbf{e}}u\|_{-1} \le CH(h^2 + H^3)\|u\|_{H^{1,2}}.\end{aligned}$$

Noting that $u - R_{H\mathbf{e}}u - \tilde{B}^h_{H\mathbf{e}}u + R_{H\mathbf{e}}\tilde{B}^h_{H\mathbf{e}}u$ is nothing but $(I - R_{H\mathbf{e}})(u - \tilde{B}^h_{H\mathbf{e}}u)$, we conclude that (3.11) is valid. □

4. Numerical experiments. In this paper, we have presented and analyzed three two-scale finite element discretizations for a class of Fredholm integral equations. It is perhaps a little too much of an undertaking to carry out and report numerical experiments for less smooth solutions in the current work, since there are some practical issues that need to be taken into account carefully. For illustration, we choose to report some two- and three-dimensional numerical experiments only for smooth solutions.

In two-dimensional examples, we use four piecewise constant finite elements that are of mesh sizes $h \times H$, $H \times h$, $H \times H$, and $h \times h$, respectively. Our two-scale finite element approximation is denoted by $B^h_{H,H}u = R_{h,H}u + R_{H,h}u - R_{H,H}u$. In three-dimensional cases, we use five piecewise constant finite elements that are of mesh sizes $h \times H \times H$, $H \times h \times H$, $H \times H \times h$, $H \times H \times H$, and $h \times h \times h$, respectively. The two-scale finite element approximation is constructed by $B^h_{H,H,H}u = R_{h,H,H}u+R_{H,h,H}u+R_{H,H,h}u - 2R_{H,H,H}u$. In all of our numerical experiments, we choose $h = H^2$. Our numerical experiments are carried out on SGI Origin 3800 at the State Key Laboratory of Scientific and Engineering Computing, Chinese Academy of Sciences.

Example 1. Consider an integral equation of the second kind in $\mathbb{R}^2$:

$$u(x_1, x_2) + \int_0^1 \int_0^1 x_1 x_2 \exp(y_1 + y_2)u(y_1, y_2)dy_1dy_2 = \exp(-x_1 - x_2) \text{ in } \square \tag{4.1}$$

with an exact solution $u = \exp(-x_1 - x_2) - \frac{1}{2}x_1x_2$, where $\square = (0, 1)^2$.

It is observed from Table 1 that not only does the two-scale finite element approximation $B^h_{H,H}u$ have a high accuracy, but also the number of the degrees of freedom for obtaining $B^h_{H,H}u$ is only of $O(1/h \times 1/H) = O(h^{-3/2})$, while that for the standard finite element solution $B_{h,h}u$ is of $O(h^{-2})$ when $h = H^2$. For instance, the accuracy of the two-scale finite element approximation $B^h_{H,H}u$ with 1,000 degrees of freedom is asymptotically the same as that of the standard finite element solution $R_{h,h}u$ with 10,000 degrees of freedom. The numerical results, stated in Table 1, support our theory (see Theorem 3.1). Hence $B^h_{H,H}u$ is a much better solution in terms of computational cost. Moreover, we can carry out the major computation in parallel. As a result, the computational scale is reduced and the computational time is saved.

TABLE 1
Example 1: L^2-*estimates for* $B^h_{H,H}u$.

$\frac{1}{h} \times \frac{1}{H}$	$\|B^h_{H,H}u - R_{h,h}u\|$	$\|u - R_{h,h}u\|$
4×2	0.00318054	0.06741261
16×4	0.00096040	0.01687140
64×8	0.00024886	0.00421813
81×9	0.00019709	0.00333285
100×10	0.00015991	0.00269961

TABLE 2
Example 1: L^2-*estimates for* $\tilde{B}^h_{H,H}u$.

$\frac{1}{h} \times \frac{1}{H}$	$\|\tilde{B}^h_{H,H}u - \tilde{R}_{h,h}u\|$	$\|u - \tilde{R}_{h,h}u\|$
4×2	0.00011373	0.00325378
16×4	0.00001086	0.00020183
64×8	0.00000074	0.00001260
81×9	0.00000046	0.00000787
100×10	0.00000030	0.00000516

It is shown from Theorem 3.1 that the iterated Galerkin approximation $\tilde{R}_{h,h}u$ and the corresponding two-scale combination solution $\tilde{B}^h_{H,H}u$ have a higher accuracy than the standard Galerkin approximation $R_{h,h}u$ and $B^h_{H,H}u$, respectively, which are supported by our numerical results, too (see Table 2).

For a three-dimensional example, we consider the integral equation

$$u(x_1,x_2,x_3) + \int_0^1\int_0^1\int_0^1 x_1x_2x_3\exp(y_1+y_2+y_3)u(y_1,y_2,y_3)dy_1dy_2dy_3 = \exp(-x_1-x_2-x_3) \text{ in } \square \tag{4.2}$$

with an exact solution $u = \exp(-x_1 - x_2 - x_3) - \frac{1}{2}x_1x_2x_3$, where $\square = (0,1)^3$.

It is not easy to solve (4.2), since the linear system for (4.2) is a dense matrix and the degree of freedom increases rapidly when the mesh size h decreases. More precisely, it is getting more and more difficult to compute $R_{h,h,h}u$ when h is smaller and smaller because of the memory limit and the speed of solving the huge linear system. However, it is relatively easy to get an approximation $B^h_{H,H,H}u$ that is asymptotically the same as that of $R_{h,h,h}u$ when $h = H^2$ (see Table 3). For instance, when $h = 1/100$, it is not possible for us to obtain $\|u - R_{h,h,h}u\|$ through computing $R_{h,h,h}u$ (in our SGI Origin 3800), but it is very easy to get $\|u - B^h_{H,H,H}u\|$.

It is observed from Table 3 that the approximate accuracy of the two-scale finite element approximation $B^h_{H,H,H}u$ with 10,000 degrees of freedom is asymptotically the same as that of the standard finite element solution $R_{h,h,h}u$ with 1,000,000 degrees of freedom, which coincides with our theory (see Theorem 3.1). It is shown by Table 3 that not only is the degree of freedom for obtaining $B^h_{H,H,H}u$ only of $O(1/h \times 1/H \times 1/H) = O(h^{-2})$ while that for the standard finite element solution $R_{h,h,h}u$ is of $O(h^{-3})$ when $h = H^2$, but also the two-scale combination solution $B^h_{H,H,H}u$ has high accuracy. Hence $B^h_{H,H,H}u$ is a much better solution in terms of computational cost. Moreover, it is shown from Theorem 3.2 that the iterated Galerkin solutions $\tilde{R}_{h,h,h}u$ and $\tilde{B}^h_{H,H,H}u$ are more accurate than $R_{h,h,h}u$ and $B^h_{H,H,H}u$, respectively. The results shown in Table 3 are in accordance with our theory, too.

In the computation, however, it takes much more time to get $\tilde{R}_{h,h,h}u$ and $\tilde{B}^h_{H,H,H}u$

TABLE 3
Example 1: L^2*-estimates for* $B^h_{H,H,H}u$ *and* $\tilde{B}^h_{H,H,H}u$.

$\frac{1}{h}\times\frac{1}{H}\times\frac{1}{H}$	$\|u-B^h_{H,H,H}u\|$	$\|u-\tilde{B}^h_{H,H,H}u\|$	$\|u-R_{h,h,h}u\|$	$\|u-\tilde{R}_{h,h,h}u\|$
4×2×2	0.04918028	0.00262928	0.04875677	0.00282859
16×4×4	0.01237953	0.00015595	0.01221095	0.00017483
64×8×8	0.00309985			
81×9×9	0.00244954			
100×10×10	0.00198428			

than to get $R_{h,h,h}u$ and $B^h_{H,H,H}u$, respectively. The reason is that it is time-consuming to compute $KR_{h,h,h}u$ and $KB^h_{H,H,H}u$. As a result, we may conclude that $B^h_{H,H,H}u$ is better than $\tilde{B}^h_{H,H,H}u$ in terms of the efficiency of computation. Similarly, it also takes much more time to get $R_{H,H,H}u-\tilde{B}^h_{H,H,H}u+R_{H,H,H}\tilde{B}^h_{H,H,H}u$ than to get $B^h_{H,H,H}u$, though $R_{H,H,H}u-\tilde{B}^h_{H,H,H}u+R_{H,H,H}\tilde{B}^h_{H,H,H}u$ has a higher accuracy than $B^h_{H,H,H}u$ (see Theorem 3.3). Taking the efficiency into account, it may be concluded that $B^h_{H,H,H}u$ would be a very good approximation and would be recommended.

Example 2. Consider an integral equation of the second kind in $\mathbb{R}^2$:

$$u(x_1,x_2)-\int_0^1\int_0^1\exp(x_1y_1+x_2y_2)u(y_1,y_2)dy_1dy_2=f(x_1,x_2)\ \text{ in }\square \tag{4.3}$$

with an exact solution $u=\exp(x_1+x_2)$ and

$$f(x_1,x_2)=\exp(x_1+x_2)-\prod_{i=1}^{2}((\exp(x_i+1)-1)/(x_i+1)),$$

where $\square=(0,1)^2$. It is noted that the global stiff matrix of (4.3) is symmetric but not positive definite. As a result, we adopt the GMRES method to solve the discrete system of (4.3).

In the three-dimensional case, we consider the following integral equation:

$$u(x_1,x_2,x_3)-\int_0^1\int_0^1\int_0^1\exp\left(\sum_{i=1}^{3}(x_iy_i)\right)u(y_1,y_2,y_3)dy_1dy_2dy_3=f(x_1,x_2,x_3)\text{ in }\square \tag{4.4}$$

with an exact solution $u=\exp(x_1+x_2+x_3)$ and

$$f(x_1,x_2,x_3)=\exp(x_1+x_2+x_3)-\prod_{i=1}^{3}((\exp(x_i+1)-1)/(x_i+1)),$$

where $\square=(0,1)^3$.

It is shown by Tables 4, 5, and 6 that the numerical results of Example 2 support our theory again.

5. Concluding remarks. In this paper, we have proposed and analyzed several two-scale Boolean Galerkin discretizations for Fredholm integral equations of the second kind. It is shown by both theory and numerics that these new discretizations are very efficient for solving integral equations in multidimensions. Since the computational cost and storage requirement of the two-scale discretizations still grow

TABLE 4
Example 2: L^2*-estimates for* $B^h_{H,H}u$.

$\frac{1}{h} \times \frac{1}{H}$	$\|B^h_{H,H}u - R_{h,h}u\|$	$\|u - R_{h,h}u\|$
4×2	0.04842409	0.32735166
16×4	0.01553697	0.08153128
64×8	0.00409148	0.02037784
81×9	0.00324410	0.01610091
100×10	0.00263427	0.01304169

TABLE 5
Example 2: L^2*-estimates for* $\tilde{B}^h_{H,H}u$.

$\frac{1}{h} \times \frac{1}{H}$	$\|\tilde{B}^h_{H,H}u - \tilde{R}_{h,h}u\|$	$\|u - \tilde{R}_{h,h}u\|$
4×2	0.00088787	0.04532941
16×4	0.00008909	0.00285472
64×8	0.00000617	0.00017851
81×9	0.00000388	0.00011144
100×10	0.00000256	0.00007312

TABLE 6
Example 2: L^2*-estimates for* $B^h_{H,H,H}u$ *and* $\tilde{B}^h_{H,H,H}u$.

$\frac{1}{h} \times \frac{1}{H} \times \frac{1}{H}$	$\|u - B^h_{H,H,H}u\|$	$\|u - \tilde{B}^h_{H,H,H}u\|$	$\|u - R_{h,h,h}u\|$	$\|u - \tilde{R}_{h,h,h}u\|$
4×2×2	0.73017738	0.08240124	0.71543805	0.09778049
16×4×4	0.18486246	0.00465276	0.17845676	0.00619129
64×8×8	0.04637518			
81×9×9	0.03665092			
100×10×10	0.02969234			

exponentially with the dimensionality, however, our methods may not be applicable for very high-dimensional problems. What we discussed here is only for piecewise constant elements. Indeed, similar results can be expected for higher-order elements. It should also be mentioned that the same discretizations can be applied to solve integral eigenvalue problems as well as nonlinear integral equations, which is our ongoing research project.

Acknowledgment. The authors would like to thank the referee for his/her valuable comments and suggestions that have improved the presentation of this paper.

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] K. BITTNER, *Fast algorithms for periodic spline wavelets on sparse grids*, SIAM J. Sci. Comput., 20 (1999), pp. 1192–1213.
[3] H. BRUNNER, *On the numerical solution of nonlinear Volterra–Fredholm integral equations by collocation methods*, SIAM J. Numer. Anal., 27 (1990), pp. 987–1000.
[4] H. J. BUNGARTZ AND M. GRIEBEL, *Sparse grids*, Acta Numer., 13 (2004), pp. 1–123.
[5] H. BUNGARTZ, M. GRIEBEL, D. RÖSCHKE, AND C. ZENGER, *A proof of convergence for the combination technique for the Laplace equation using tools of symbolic computation*, Math. Comput. Simulation, 42 (1996), pp. 595–605.
[6] H. BUNGARTZ, M. GRIEBEL, AND U. RÜDE, *Extrapolation, combination, and sparse grid techniques for elliptic boundary value problems*, Comput. Methods Appl. Mech. Engrg., 116 (1994), pp. 243–252.
[7] Z. CHEN AND Y. XU, *The Petrov–Galerkin and iterated Petrov–Galerkin methods for second-kind integral equations*, SIAM J. Numer. Anal., 35 (1998), pp. 406–434.

[8] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, SIAM, Philadelphia, 2002.
[9] F. Cuker and S. Smale, *On the mathematical foundations of learning*, Bull. Amer. Math. Soc. (N.S.), 39 (2002), pp. 1–49.
[10] F.-J. Delvos, *d-variate Boolean interpolation*, J. Approx. Theory, 34 (1982), pp. 99–114.
[11] M. Griebel, P. Oswald, and T. Schiekofer, *Sparse grids for boundary integral equations*, Numer. Math., 83 (1999), pp. 279–312.
[12] M. Griebel, M. Schneider, and C. Zenger, *A combination technique for the solution of sparse grid problems*, in Proceedings of the IMACS International Symposium on Iterative Methods in Linear Algebra, P. de Groen and R. Beauwens, eds., Elsevier, Amsterdam, 1992, pp. 263–281.
[13] H. Kaneko and Y. Xu, *Degenerate kernel method for Hemmerstein equations*, Math. Comp., 56 (1991), pp. 141–148.
[14] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.
[15] Q. Lin, N. Yan, and A. Zhou, *A sparse finite element method with high accuracy: Part* I, Numer. Math., 88 (2001), pp. 731–742.
[16] F. Liu and A. Zhou, *Two-scale finite element discretizations for partial differential equations*, J. Comput. Math., 24 (2006), pp. 373–392.
[17] F. Liu and A. Zhou, *Localization and parallelizations for two-scale finite element discretizations*, Comm. Pure Appl. Anal., to appear.
[18] K. T. Mynbaev, *Approximate solution of a Fredholm integral equation of the second kind in a multidimensional domain*, Soviet Math. Dokl., 43 (1991), pp. 543–546.
[19] C.-H. Pao and N. E. Bickers, *Renormalization-group acceleration of self-consistent field solutions: Two-dimensional Hubbard model*, Phys. Rev. B, 49 (1994), pp. 1586–1599.
[20] C. Pflaum, *Diskretisierung Elliptischer Differentialgleichungen mit Dünnen Gittern*, Dissertation, Technische Universität München, München, Germany, 1995.
[21] C. Pflaum, *Convergence of the combination technique for second-order elliptic differential equations*, SIAM J. Numer. Anal., 34 (1997), pp. 2431–2455.
[22] C. Pflaum and A. Zhou, *Error analysis of the combination technique*, Numer. Math., 84 (1999), pp. 327–350.
[23] U. Rüde and A. Zhou, *Multiparameter extrapolation methods for boundary integral equations*, Adv. Comput. Math., 9 (1998), pp. 173–190.
[24] I. H. Solan and V. Thomée, *Superconvergence of the Galerkin iterates for integral equations of the second kind*, J. Integral Equations, 9 (1985), pp. 1–23.
[25] G. Vainikko, *Multidimensional Weakly Singular Integral Equations*, Lectures Notes Math. 1549, Springer-Verlag, Berlin, 1993.
[26] Y. Xu and A. Zhou, *Fast Boolean approximations methods for solving integral equations in high dimensions*, J. Integral Equations Appl., 16 (2004), pp. 83–110.
[27] K. Yosida, *Functional Analysis*, 6th ed., Springer-Verlag, Berlin, 1980.
[28] C. Zenger, *Sparse grids*, in Parallel Algorithms for Partial Differential Equations: Proceedings of the Sixth GAMM-Seminar (Kiel, 1990), Notes Numer. Fluid Mech. 31, W. Hackbusch, ed., Vieweg, Braunschweig, 1991, pp. 241–251.
[29] A. Zhou, *Multi-parameter error resolution for the collocation method of Volterra integral equations*, BIT, 37 (1997), pp. 978–987.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 313–332

CONVERGENCE ANALYSIS OF A QUASI-CONTINUUM APPROXIMATION FOR A TWO-DIMENSIONAL MATERIAL WITHOUT DEFECTS*

PING LIN†

Abstract. In many applications, materials are modeled by a large number of particles (or atoms), where any particle can interact with any other. The computational cost is very high since the number of atoms is huge. Recently much attention has been paid to a so-called quasi-continuum (QC) method, which is a mixed atomistic/continuum model. The QC method uses an adaptive finite element framework to effectively integrate the majority of the atomistic degrees of freedom in regions where there is no serious defect. However, numerical analysis of this method is still in its infancy. In this paper we will conduct a convergence analysis of the QC method in the case when there is no defect. We will also remark on the case when the defect region is small. The difference between our analysis and conventional analysis is that our exact atomistic solution is not a solution of a continuous partial differential equation, but a discrete lattice scale solution which is not approximately related to any conventional partial differential equation.

Key words. lattice statics, Lennard–Jones potential, global minimization, finite element method, quasi-continuum approximation, material defects

AMS subject classifications. 65K10, 65N15, 65N30, 70C20, 74G15, 74G65, 74N15, 74Q05

DOI. 10.1137/050636772

1. Introduction.

The analysis of the structure of material, and defects of material such as dislocations or fractures, often involves the effect of the lattice on the scale. Directly solving the whole system (lattice statics) provides an accurate solution for analysis on this scale. However, because the number of atomistic particles in a material is huge, it is often impossible to directly solve the whole system to obtain the material properties. The fact that in many practical problems defects occur only in some local and small regions may help with the design of approximation or reduction methods for the original huge problem. The quasi-continuum (QC) approximation recently gained attention in the engineering literature (cf. [13, 3, 21, 17]). The idea is that we can consider the region (called the local approximation region), where no defects occur, at the macroscopic scale, and the theory of continuum material elasticity may apply. The model enables a treatment of lattice defects—should these defects arise—and exhibits a continuous transition from the lattice to the continuum realms at intermediate length scales. It is incorporated with the (nonconforming) finite element method and is expected to be an approximation of the full lattice-scale model. There are other models that couple atomistic/finite element methods and that employ some sort of handshaking region at the atomistic/FEM interface [12, 1, 18]. A major strength of the QC method is its ability to adaptively mesh as the deformation gradient changes, such that regions can switch between microscopic and continuum measures of the energy as needed. Further improvement of the method can also be founded in the literature; see, e.g., [20].

While a significant body of knowledge about QC related models and their experimental and numerical tests has been accumulated, not much has been reported

*Received by the editors July 24, 2005; accepted for publication (in revised form) June 29, 2006; published electronically February 9, 2007.
http://www.siam.org/journals/sinum/45-1/63677.html
†Department of Mathematics, The National University of Singapore, Singapore 117543 (matlinp@nus.edu.sg).

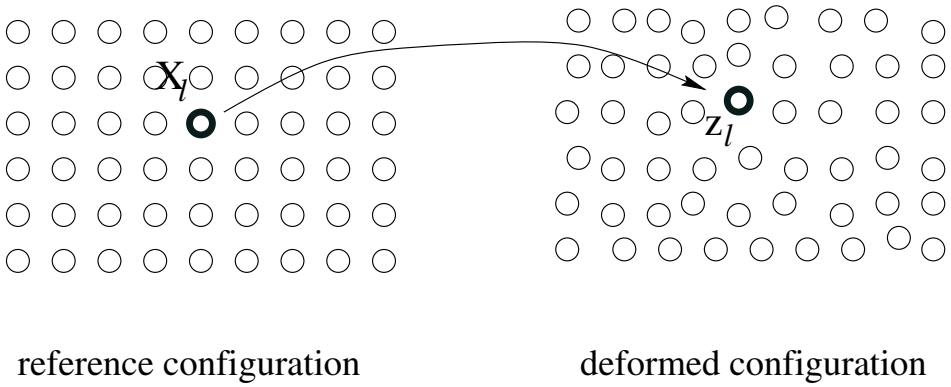


FIG. 1. *Atomic positions at the reference and deformed configurations.*

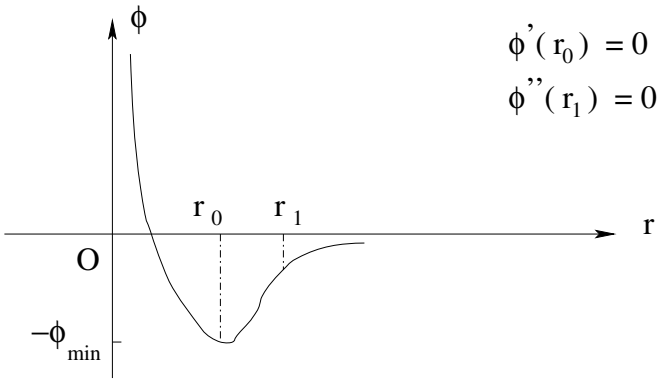


FIG. 2. *The energy function $\phi(r)$.*

on the analysis of the approximation error of these models. The convergence rate of the QC method is analyzed in [11] via computational results of a specific nanoindentation problem. Other numerical tests can also be found in [21, 19]. In [7] the QC method has been related to a heterogeneous multiscale framework based on a linear homogenization technique. In [14] the QC method is analyzed for a one-dimensional atomic chain without external forces. An analysis with an external force is conducted in [4] for the nearest neighbor interaction. Other relevant work may be found in [8, 2, 9, 6, 15]. The aim of this paper is to present an error analysis between the solution of the original lattice-scale atomic system and the solution of its QC approximation for multidimensional materials with conservative external forces. We will consider only a two-dimensional material in this paper.

Let $X_\ell = (x_\ell, y_\ell)$ represent the coordinates of an atom or particle ℓ (i.e., the location of the atom ℓ in a reference configuration). We collect all these reference positions of atoms in a vector X. The position of the atom ℓ in the deformed configuration is denoted as z_ℓ. We collect all deformed atomic positions in a vector z. See Figure 1.

We shall assume that the atoms interact with each other via a pair-interacting potential $\phi(r)$, where r is the distance between the pairs of atoms. The shape of the function, shown in Figure 2, is usually nonconvex. The popular 6-12 Lennard–Jones

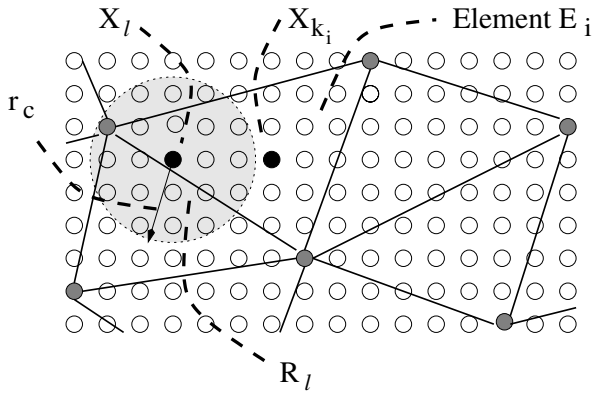


FIG. 3. *Triangulation and notation.*

potential (see Frank and van der Merwe [10])

$$\phi(r) = \phi_{min}\left[-2\left(\frac{r_0}{r}\right)^6 + \left(\frac{r_0}{r}\right)^{12}\right] \tag{1}$$

is one example of such a function. We shall do our analysis for a potential function of this type. Obviously r_0 is the minimal point of $\phi(r)$. We may assume that r_0 is the grid size of atomic position X in the reference configuration (called a lattice constant). Later we shall also assume that the nearest neighbor atomic interaction dominates farther interactions based on the solid material property.

The meaning of the *defect* in this paper is given as follows. If there is no defect, any pair of nearest neighbor atoms, say located at z_k and z_ℓ, should have such a distance $r_{k\ell} = |z_k - z_\ell|$, where $\phi(r_{k\ell})$ is convex (i.e., $r_{k\ell} < r_1$). Basically, if $r_{k\ell} < r_1$ but is close to r_1, the material has or starts to have a defect; if $r_{k\ell} \geq r_1$ (i.e., $r_{k\ell}$ is in the nonconvex region of $\phi(r)$), then we think that the material has a serious defect. If the external forces on the material are so strong that some pairs of nearest neighbor atoms are in the nonconvex region of $\phi(r)$, the convergence analysis may be problematic (if not impossible) since the solution at the lattice scale may not be unique. So in our analysis we shall assume that pairs of nearest neighbor atoms are in the convex region of $\phi(r)$, as described in section 3. In this paper we mainly consider the material with no serious defects. We shall remark on the defect case as well.

Later, when considering the QC approximation we need to triangulate the reference domain, as the finite element method usually does. In order to describe the total potential energy and the approximate energy in a consistent way, we consider a triangulation of the domain at this early stage, which is shown in Figure 3. We shall assume that the triangular mesh satisfies usual regular conditions. R_ℓ is a disc region with radius r_c centered at the atom X_ℓ. Only atoms in the disc R_ℓ will be counted for interaction with the atom X_ℓ. In practice, r_c is taken as twice the potential cut-off radius.

Assume that the external force is conservative and its corresponding external potential is denoted as $F(z)$. Define $f(z) = F_z(z)$. The total potential energy reads

$$E(z) = \frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{m_i}\sum_{\ell\in R_k}\phi(|z_k - z_\ell|) + \sum_{i=1}^{N}\sum_{k=1}^{m_i}F(z_k), \tag{2}$$

where, and in what follows, the letter i is specially assigned to the index for the ith triangular element E_i, and the index of the three vertices of the element E_i will be denoted as i_1, i_2, and i_3. k is the index[1] for atoms in the element E_i, m_i is the number of atoms in E_i, and N is the number of triangular elements. $|\cdot|$ represents the Euclidean length of a vector in $\mathbf{R}^2$. Obviously, we should also have $\ell \neq k$ in the summation $\Sigma_{\ell \in R_k}$ since an atom cannot interact with itself. For simplicity of notation we just drop the self-interaction condition $\ell \neq k$ throughout the paper. We shall also assume Dirichlet boundary conditions in a bounded material domain Ω.

According to physical principle (cf. [21]) stable configuration of the material is identified with the minimizer of the potential energy $E(z)$:

$$E(\hat{z}) = \min_z E(z). \tag{3}$$

Let $z = \hat{z} + tv$ (or $z_\ell = \hat{z}_\ell + tv_\ell$). From

$$\frac{d}{dt} E(\hat{z} + tv)|_{t=0} = 0,$$

we can obtain the variational formulation of (3):

$$a(\hat{z}, v) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{m_i} \sum_{\ell \in R_k} \frac{\phi'(|\hat{z}_\ell - \hat{z}_k|)}{|\hat{z}_\ell - \hat{z}_k|} (\hat{z}_\ell - \hat{z}_k) \cdot (v_\ell - v_k) = -\sum_{i=1}^{N} \sum_{k=1}^{m_i} f(\hat{z}_k) \cdot v_k \quad \forall v, \tag{4}$$

where the solution $\hat{z}$ satisfies Dirichlet boundary conditions.

The rest of the paper is organized as follows. In section 2 we introduce the QC approximation similarly to the original lattice-scale problem given above. Roughly speaking, it corresponds to a nonconforming finite element approximation of (4). In section 3 we present a few assumptions and justifications for existence and uniqueness of the lattice-scale solution. Finally, in section 4 we estimate the error of the QC approximation in cases with no serious defect. Roughly speaking, we mainly assume (i) all nearest neighbor pairs of atoms in a convex region of ϕ, and (ii) nearest neighbor interaction dominates further interaction. Under these assumptions we show that the error of QC approximation is of $O(h)$ plus a nonconforming error term which is related to the number of atoms in each element and each element's boundaries. Here h is the largest side of the triangulation. We also remark on a possible error estimate in the case when the above assumptions do not hold or serious defects occur in a relatively small region.

2. QC approximation. Now we describe the QC approximation combined with a nonconforming finite element idea to problem (3). Our description follows [21], but our unknown vector is the deformed atomic position vector in order to be consistent with the original lattice-scale problem described above. Essentially our formulation should be the same as that in [21]. A triangulation is already given in the previous section. Assume in each triangle E_i, shown in Figure 4, that the atoms are deformed linearly, which corresponds to using a piecewise linear function to approximate the solution in the finite element context. As mentioned in the previous section, we denote the three vertices of the element E_i by X_{i_1}, X_{i_2}, and X_{i_3}. Let $X_{i_j} = (x_{i_j}, y_{i_j})^T$,

[1] We should use a double index such as (i, k) to represent an atom k in the element E_i. For simplicity of notation without significant ambiguity we use k instead of (i, k).

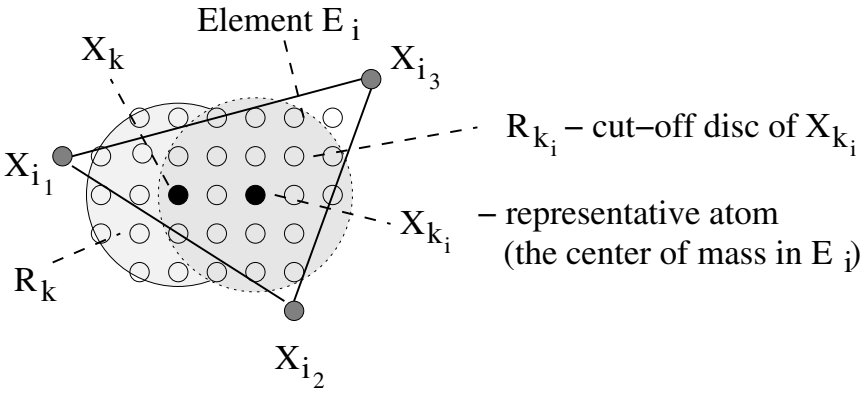


FIG. 4. *One triangular element.*

$j = 1, 2, 3$, be vertices of the triangular element E_i and Δ_i be its area. Define piecewise linear basis functions (denoting $X_\ell = (x_\ell, y_\ell)^T$):

$$\psi_{i_1}(X_\ell) = \frac{x_{i_2}y_{i_3} - x_{i_3}y_{i_2}}{2\Delta_i} + \frac{y_{i_2} - y_{i_3}}{2\Delta_i}x_\ell + \frac{x_{i_3} - x_{i_2}}{2\Delta_i}y_\ell,$$
$$\psi_{i_2}(X_\ell) = \frac{x_{i_3}y_{i_1} - x_{i_1}y_{i_3}}{2\Delta_i} + \frac{y_{i_3} - y_{i_1}}{2\Delta_i}x_\ell + \frac{x_{i_1} - x_{i_3}}{2\Delta_i}y_\ell,$$
$$\psi_{i_3}(X_\ell) = \frac{x_{i_1}y_{i_2} - x_{i_2}y_{i_1}}{2\Delta_i} + \frac{y_{i_1} - y_{i_2}}{2\Delta_i}x_\ell + \frac{x_{i_2} - x_{i_1}}{2\Delta_i}y_\ell.$$

In each element E_i the derivatives $\nabla_{x,i}\psi_{i_j}$ and $\nabla_{y,i}\psi_{i_j}$ of ψ_{i_j} are constant, and we denote the derivative within the element i as $\nabla_i\psi_{i_j} = (\nabla_{x,i}\psi_{i_j}, \nabla_{y,i}\psi_{i_j})^T$. We can express the position of the representative atom k_i (the atom closest to the center of mass of the element E_i) and any atom ℓ in the element E_i using these basis functions, i.e.,

$$Z_{k_i} = \psi_{i_1}(X_{k_i})Z^h_{i_1} + \psi_{i_2}(X_{k_i})Z^h_{i_2} + \psi_{i_3}(X_{k_i})Z^h_{i_3}, \tag{5}$$
$$Z_\ell = \psi_{i_1}(X_\ell)Z^h_{i_1} + \psi_{i_2}(X_\ell)Z^h_{i_2} + \psi_{i_3}(X_\ell)Z^h_{i_3}, \tag{6}$$

where Z^h is a vector collecting all positions of atoms at the vertices of the triangulation, and $Z = (Z_\ell)$ is a vector collecting all positions of atoms (defined by (6)) at every lattice node X_ℓ. We can define their derivatives (constant) as well:

$$\nabla^h_{s,i}Z = \nabla^h_{s,i}Z_{k_i} = \nabla^h_{s,i}Z_\ell = \sum_{j=1}^{3} \nabla_{s,i}\psi_{i_j}Z^h_{i_j}, \tag{7}$$

where s may be x or y and $\ell \in E_i$. We can also write

$$Z_\ell - Z_{k_i} = \sum_{j=1}^{3} \left(\psi_{i_j}(X_\ell) - \psi_{i_j}(X_{k_i})\right) Z^h_{i_j} = \sum_{j=1}^{3} (X_\ell - X_{k_i}) \cdot \nabla_i\psi_{i_j}Z^h_{i_j}$$
$$= (x_\ell - x_{k_i})\nabla^h_{x,i}Z + (y_\ell - y_{k_i})\nabla^h_{y,i}Z. \tag{8}$$

The cut-off disc R_{k_i} could include atoms of a few triangular elements in the so-called nonlocal approximation (cf. [21]). If an atom $\ell \in R_{k_i}$ is not in the element E_i but in

another element $E_{i'}$, then we can express its approximate position in a similar way according to the vertices and basis function associated with $E_{i'}$:

$$Z_\ell = \psi_{i'_1}(X_\ell)Z^h_{i'_1} + \psi_{i'_2}(X_\ell)Z^h_{i'_2} + \psi_{i'_3}(X_\ell)Z^h_{i'_3}.$$

In the case when there is no serious defect, local approximation is enough and no nonlocal approximation is necessary.

The idea of the QC method is the following: The potential energy associated with any atom k in the triangular element E_i is approximately equal to the potential energy associated with the representative atom k_i. That is,

$$\sum_{\ell\in R_k} \phi(|z_\ell - z_k|) \approx \sum_{\ell\in R_{k_i}} \phi(|z_\ell - z_{k_i}|), \tag{9}$$

where R_k and R_{k_i} are cut-off discs of atomic interaction associated with the atom k and the representative atom k_i, respectively. With the QC approximation we can write an approximate total potential energy of (2) as follows:

$$E_{qc}(Z^h) = \frac{1}{2}\sum_{i=1}^{N} m_i \sum_{\ell\in R_{k_i}} \phi(|Z_\ell - Z_{k_i}|) + \sum_{i=1}^{N}\sum_{k=1}^{m_i} F(Z_k), \tag{10}$$

where Z_ℓ, Z_k, and Z_{k_i} are defined as in (6) and (5), and Z^h is a vector collecting all positions of atoms at the vertices of the triangulation, e.g., $Z^h_{i_1}$, $Z^h_{i_2}$, $Z^h_{i_3}$, etc. The approximate stable configuration $\hat{Z}^h$ is identified with the minimizer of the approximate potential energy $E_{qc}(Z^h)$:

$$E_{qc}(\hat{Z}^h) = \min_{Z^h} E_{qc}(Z^h). \tag{11}$$

Note that in [21] deformation gradients defined on triangles are used as minimization variables. We do the minimization with respect to positions of atoms at the vertices of the triangulation in order to make it consistent with the original problem (3), and consequently convergence analysis may be conducted with less difficulty. Let $Z^h = \hat{Z}^h + tV^h$ (or $Z^h_{i_j} = \hat{Z}^h_{i_j} + tV^h_{i_j}$). From $\frac{d}{dt}E_{qc}(\hat{Z}^h + tV^h)|_{t=0} = 0$ we then obtain the approximate variational formulation

$$\begin{aligned} a_h(\hat{Z}, V) &= \frac{1}{2}\sum_{i=1}^{N} m_i \sum_{\ell\in R_{k_i}} \frac{\phi'(|\hat{Z}_\ell - \hat{Z}_{k_i}|)}{|\hat{Z}_\ell - \hat{Z}_{k_i}|}(\hat{Z}_\ell - \hat{Z}_{k_i})\cdot(V_\ell - V_{k_i}) \\ &= -\sum_{i=1}^{N}\sum_{k=1}^{m_i} f(\hat{Z}_k)\cdot V_k \quad \forall V, \end{aligned} \tag{12}$$

where the solution $\hat{Z}^h$ satisfies Dirichlet boundary conditions, and $\hat{Z}_\ell$, $\hat{Z}_{k_i}$, $\hat{Z}_k$, V_ℓ, V_{k_i}, and V_k are defined similarly to Z_ℓ and Z_{k_i} as in (6) and (5) (the only difference is that $Z^h_{i_j}$ is replaced by $\hat{Z}^h_{i_j}$ and $V^h_{i_j}$, $j = 1, 2, 3$). It corresponds to a nonconforming method in the finite element context. The difference between our analysis and conventional finite element analysis is that our exact solution is not a solution of a continuous partial differential equation but a discrete lattice-scale solution which may not be related to any conventional partial differential equation. Let $g(\alpha) = \phi_\alpha(|\alpha|) = \frac{\phi'(|\alpha|)}{|\alpha|}\alpha$. We can write

$$a_h(\hat{Z}, V) = \frac{1}{2}\sum_{i=1}^{N} m_i \sum_{\ell\in R_{k_i}} g(\hat{Z}_\ell - \hat{Z}_{k_i})\cdot(V_\ell - V_{k_i}) \quad \forall V. \tag{13}$$

3. Assumptions and uniqueness of the lattice-scale solution. The objective of this section is mainly to make a few assumptions for conducting the error analysis to be presented in the next section. The derivative of $g(\alpha)$ is important in the analysis. We first state a simple lemma.

LEMMA 1. *If two vectors $\alpha = (\alpha_1, \alpha_2)^T$ and $\beta = (\beta_1, \beta_2)^T$ are not in the same direction, i.e., $\alpha_1\beta_2 \neq \alpha_2\beta_1$, then the matrix*

$$\begin{pmatrix} \alpha_1^2 & \alpha_1\alpha_2 \\ \alpha_1\alpha_2 & \alpha_2^2 \end{pmatrix} + \begin{pmatrix} \beta_1^2 & \beta_1\beta_2 \\ \beta_1\beta_2 & \beta_2^2 \end{pmatrix}$$

is symmetric, positive definite, and has eigenvalues $\geq \frac{|\alpha|^2|\beta|^2}{|\alpha|^2+|\beta|^2}\sin^2\gamma$ and $\leq |\alpha|^2+|\beta|^2$, where γ is the angle between the two vectors α and β.

We can explicitly calculate the derivative of $g(\alpha)$ and obtain

$$\begin{aligned} \phi_{\alpha\alpha}(|\alpha|) = g_\alpha(\alpha) &= \tfrac{\phi''(|\alpha|)}{|\alpha|^2}\begin{pmatrix} \alpha_1^2 & \alpha_1\alpha_2 \\ \alpha_2\alpha_1 & \alpha_2^2 \end{pmatrix} + \tfrac{\phi'(|\alpha|)}{|\alpha|^3}\begin{pmatrix} \alpha_2^2 & -\alpha_1\alpha_2 \\ -\alpha_2\alpha_1 & \alpha_1^2 \end{pmatrix} \\ &= g_{\alpha 1}(\alpha) + g_{\alpha 2}(\alpha), \end{aligned} \tag{14}$$

where $g_{\alpha 1}$ and $g_{\alpha 2}$ represent the first and second terms in the expression of g_α. If $|\alpha|$ is in the convex region of $\phi(|\alpha|)$, then $\phi''(|\alpha|) > 0$ and $\phi'(|\alpha|) \geq 0$ if $|\alpha| \geq r_0$. Therefore, the second term $g_{\alpha 2}(\alpha)$ of (14) is positive semidefinite when $|\alpha| \geq r_0$. In the case $|\alpha| < r_0$,

$$\phi''(|\alpha|) - \frac{|\phi'(|\alpha|)|}{|\alpha|} = \phi''(|\alpha|) + \frac{\phi'(|\alpha|)}{|\alpha|} > 72\phi_{min}\frac{r_0^6}{|\alpha|^8}. \tag{15}$$

Hence, $\phi'(|\alpha|)/|\alpha|$ would be much smaller than $\phi''(|\alpha|)$ when $|\alpha| < r_0$. On the other hand, the potential energy function $\phi(|\alpha|)$ vanishes quickly after the lattice distance r_0 from atomic property (see the figure in [5, p. 143], which suggests nearest neighbor dominance in atomic interactions of solid materials. These facts together with Lemma 1 motivate the assumption below that the nearest neighbor sum of $g_{\alpha 1}$ dominates the sum of g_α.

We now write down our assumptions, and some further explanation follows afterward. We shall analyze the error of the QC approximation in the next section, based on these assumptions.

Assumptions. *Let $\hat{z}$ and Z^h be the lattice-scale solution and the QC approximate solution satisfying Dirichlet boundary conditions, and let $\hat{Z}$ be the piecewise linear interpolation based on Z^h as defined in* (6). *Also, a square lattice grid is used in the reference configuration* (*see Figure* 1).

1. *The distance of any pair of nearest neighbor atoms $\hat{z}_\ell - \hat{z}_{nb_\ell}$ and $\hat{Z}_\ell - \hat{Z}_{nb_\ell}$ is located in the convex region of $\phi(|\alpha|)$. More precisely, the nearest neighbor distance is in the region*

   $$c_1 r_0 < |\alpha| < C_1 r_0, \tag{16}$$

   where $\alpha = \hat{z}_\ell - \hat{z}_{nb_\ell}$ or $\hat{Z}_\ell - \hat{Z}_{nb_\ell}$ and $0 < c_1 < C_1 < r_1/r_0$ (r_1 is the inflection point of $\phi(r)$, e.g., for the 6-12 *Lennard–Jones potential $r_1 = \sqrt[6]{13/7}\ r_0$). Under our problem setting (square reference grid) nearest neighbors of $\ell = (\ell_x, \ell_y)$ should at least include $\{(\ell_x-1, \ell_y), (\ell_x+1, \ell_y), (\ell_x, \ell_y-1), (\ell_x, \ell_y+1)\}$.*
2. *The nearest neighbor interaction dominates farther interactions. In our analysis we use the following for any atom $k = (k_x, k_y)$): $\sum_{\ell\in R_k}(x_\ell - x_k)^2 g_\alpha(\xi_\ell)$ and $\sum_{\ell\in R_k}(y_\ell - y_k)^2 g_\alpha(\xi_\ell)$ are dominated by $\sum_{nb_k}(x_{nb_k} - x_k)^2 g_{\alpha 1}(\xi_{nb_k})$ and*

$\sum_{nb_k}(y_{nb_k}-y_k)^2 g_{\alpha 1}(\xi_{nb_k})$ *(summing over all nearest neighbor atoms* nb_k*), respectively, where* $\xi_\ell = \mu_\ell(\hat{Z}_\ell - \hat{Z}_k) + (1-\mu_\ell)(\hat{z}_\ell - \hat{z}_k)$*. Also, among nearest neighbor atoms* $\{(k_x-1,k_y),(k_x+1,k_y),(k_x,k_y-1),(k_x,k_y+1)\}$ *there are at least two* ξ_ℓ*'s in distinct directions.*

3. *As we mentioned earlier, the external potential* $F(z)$ *is smooth and strictly convex or* $F_{zz}(z) = f_z(z)$ *is positive definite.*

The upper and lower bounds for $|\alpha|$ in assumption 1 can be showed for one-dimensional problems (see [14]). $|\alpha|$ located in the convex region of the pair potential energy may not be true in general but is necessary for the uniqueness of the solution. A lower bound of $|\alpha|$ away from zero is expected since if pairs of atoms are too close, the energy will be too large to reach its minimum. Assumption 2 tries to provide a concrete mathematical description (or understanding) of nearest neighbor dominance for a solid material. In Assumption 3 the strictly convex assumption may be relaxed to being just convex (see the remark after Theorem 1). Thus common external forces, e.g., gravity, Coulomb forces, and the nanoindentation example given in [11], are included in this study.

The existence of the solution of minimization problems (satisfying a finite boundary condition) is obvious from the property of a continuous function since the pair potential energy function ϕ is continuous and has a lower bound and the external energy is continuous in the bounded material domain. Under the assumptions we can also show the uniqueness of the solution of variational problems (4) and (12). For example, assume that if there are two solutions z^1 and z^2 of problem (4), then both z^1 and z^2 satisfy (4). Subtracting the two equations for z^1 and z^2, denoting $d = z^1 - z^2$, and taking $v = d$ we have

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{m_i}\sum_{\ell\in R_k} g_\alpha(\xi_k)(d_k-d_\ell)\cdot(d_k-d_\ell) + \sum_{i=1}^{N}\sum_{k=1}^{m_i} f_z(\eta_k)d_k\cdot d_k = 0,$$

where $\xi_k = \mu(z_k^1 - z_\ell^1) + (1-\mu)(z_k^2 - z_\ell^2)$ $(0<\mu<1)$ and η_k is between z_k^1 and z_k^2. According to assumption 1 (noting that we use only the second inequality $|\alpha| \le C_1 r_0$), the difference of nearest neighbors of both $z_k^1 - z_\ell^1$ and $z_k^2 - z_\ell^2$ is in the convex region of ϕ (i.e., $\le C_1 r_0$) and F is convex so both $\sum_{\ell\in R_k} g_\alpha(\xi_k)(d_k-d_\ell)\cdot(d_k-d_\ell)$ (according to the assumption of the nearest neighbor dominance) and $f_z(\eta_k)d_k\cdot d_k$ are nonnegative. We can then conclude $d=0$ or uniqueness of the solution.

We would like to mention here that using the square lattice grid in the reference configuration does not imply that a square lattice structure is assumed. The real location of atoms is in the deformed configuration, where triangular or hexagonal lattice solutions and other types of lattice solutions are possible, depending on the Dirichlet boundary condition, the external force, and the cut-off of the atomic interaction. In the next section we shall analyze error in the QC approximation.

4. Error analysis of the QC method. Our goal in this section is to estimate the error $\hat{Z} - \hat{z}$, where $\hat{z}$ is the solution of the original problem (3) or (4) and $\hat{Z}$ is the piecewise linear interpolation (defined as (5) or (6)) of the solution $\hat{Z}^h$ of the approximate problem (11) or (12). Define $\tilde{z}$ to be a piecewise linear interpolation of the solution $\hat{z}$, i.e.,

$$\tilde{z}_\ell = \psi_{i_1}(X_\ell)\hat{z}_{i_1} + \psi_{i_2}(X_\ell)\hat{z}_{i_2} + \psi_{i_3}(X_\ell)\hat{z}_{i_3}, \tag{17}$$

if $X_\ell = (x_\ell, y_\ell)^T$ is a point in a triangular element E_i. We first estimate the interpolation error $\hat{z} - \tilde{z}$. In order to write down a discrete Taylor's expansion we introduce

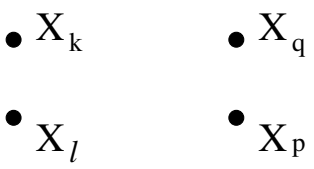


FIG. 5. *Location of four points.*

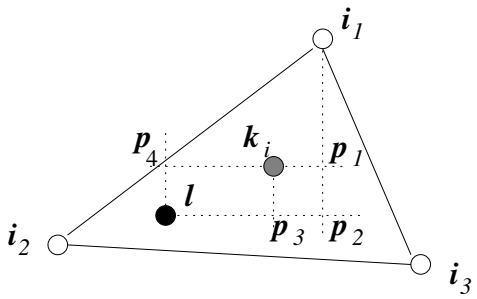


FIG. 6. *Derivation of discrete Taylor's formula.*

notation of the first and second order divided differences as follows. If three points X_ℓ, X_k, and X_p are located in the horizontal direction (say, X_k in the middle) we define

$$D_{x,k}u_\ell = \frac{u_k - u_\ell}{x_k - x_\ell}, \quad D_{x,p}D_{x,k}u_\ell = \frac{D_{x,p}u_k - D_{x,k}u_\ell}{\frac{1}{2}(x_p - x_\ell)}. \tag{18}$$

$D_{y,k}u_\ell$ and $D_{y,p}D_{y,k}u_\ell$ can be defined similarly if three atomic points X_ℓ, X_k, and X_p are located in the vertical direction. If four atomic points X_ℓ, X_k, X_p, and X_q are located, as in Figure 5, we define

$$D_{x,p}D_{y,k}u_\ell = \frac{D_{y,q}u_p - D_{y,k}u_\ell}{x_p - x_\ell}, \quad D_{y,k}D_{x,p}u_\ell = \frac{D_{x,q}u_k - D_{x,p}u_\ell}{y_k - y_\ell}. \tag{19}$$

Denote $\hat{z} = (\hat{z}^1, \hat{z}^2)^T$, and let indices i_1, i_2, and i_3 be the vertices of the element E_i. Let ℓ be the index of any atomic point in the element and representative atom k_i, and let atoms p_1, p_2, p_3, and p_4 be positioned as in Figure 6. Obviously, $x_{i_1} = x_{p_1} = x_{p_2}$, $x_{k_i} = x_{p_3}$, $x_{p_4} = x_\ell$, $y_{p_2} = y_{p_3} = y_\ell$, and $y_{p_1} = y_{k_i} = y_{p_4}$. For simplicity we will consider only ℓ located in a position as shown in Figure 6. For any other locations of ℓ and any other different relative positions of atoms i_1, k_i, and ℓ, the discussion is similar. Mimicking the continuous Taylor's theorem, we can have

$$\hat{z}^1_{i_1} = \hat{z}^1_\ell + \hat{z}^1_{p_2} - \hat{z}^1_\ell + \hat{z}^1_{i_1} - \hat{z}^1_{p_2} = \hat{z}^1_\ell + D_{x,p_2}\hat{z}^1_\ell\,(x_{p_2} - x_\ell) + D_{y,i_1}\hat{z}^1_{p_2}\,(y_{i_1} - y_{p_2}) \tag{20}$$

We can also obtain

$$\begin{aligned}
D_{x,p_2}\hat{z}^1_\ell &= D_{x,p_3}\hat{z}^1_\ell + D_{x,p_2}\hat{z}^1_\ell - D_{x,p_3}\hat{z}^1_\ell \\
&= D_{x,p_3}\hat{z}^1_\ell + \frac{x_{p_2} - x_{p_3}}{x_{p_2} - x_\ell}D_{x,p_2}\hat{z}^1_{p_3} + \frac{x_{p_3} - x_\ell}{x_{p_2} - x_\ell}D_{x,p_3}\hat{z}^1_\ell - D_{x,p_3}\hat{z}^1_\ell \\
&= D_{x,p_3}\hat{z}^1_\ell + \frac{x_{p_2} - x_{p_3}}{x_{p_2} - x_\ell}(D_{x,p_2}\hat{z}^1_{p_3} - D_{x,p_3}\hat{z}^1_\ell) \\
&= D_{x,p_3}\hat{z}^1_\ell + \frac{x_{p_2} - x_{p_3}}{x_{p_2} - x_\ell}D_{x,p_2}D_{x,p_3}\hat{z}^1_\ell\,\frac{1}{2}(x_{p_2} - x_\ell)
\end{aligned}$$

and

$$D_{y,i_1}\hat{z}^1_{p_2} = D_{y,p_4}\hat{z}^1_\ell + D_{y,i_1}\hat{z}^1_{p_2} - D_{y,p_1}\hat{z}^1_{p_2} + D_{y,p_1}\hat{z}^1_{p_2} - D_{y,p_4}\hat{z}^1_\ell = D_{y,p_4}\hat{z}^1_\ell + \frac{y_{i_1}-y_{p_1}}{y_{i_1}-y_{p_2}} D_{y,i_1}D_{y,p_1}\hat{z}^1_{p_2}\ \frac{1}{2}(y_{i_1}-y_{p_2}) + D_{x,p_2}D_{y,p_4}\hat{z}^1_\ell\ (x_{p_2}-x_\ell).$$

Therefore, using $x_{i_1} - x_\ell = x_{p_2} - x_\ell$ and $y_{i_1} - y_{p_2} = y_{i_1} - y_\ell$, we have

$$\begin{aligned}\hat{z}^1_{i_1} &= \hat{z}^1_\ell + D_{x,p_3}\hat{z}^1_\ell\ (x_{i_1}-x_\ell) + D_{y,p_4}\hat{z}^1_\ell\ (y_{i_1}-y_\ell) \\ &\quad + \frac{x_{p_2}-x_{p_3}}{x_{p_2}-x_\ell} D_{x,p_2}D_{x,p_3}\hat{z}^1_\ell\ \frac{1}{2}(x_{p_2}-x_\ell)^2 \\ &\quad + \frac{y_{i_1}-y_{p_1}}{y_{i_1}-y_{p_2}} D_{y,i_1}D_{y,p_1}\hat{z}^1_{p_2}\ \frac{1}{2}(y_{i_1}-y_{p_2})^2 + D_{x,p_2}D_{y,p_4}\hat{z}^1_\ell\ (x_{p_2}-x_\ell)(y_{i_1}-y_{p_2}),\end{aligned} \tag{21}$$

where it is easy to see

$$\frac{x_{p_2}-x_{p_3}}{x_{p_2}-x_\ell} \le 1, \quad \frac{y_{i_1}-y_{p_1}}{y_{i_1}-y_{p_2}} \le 1.$$

Similar expansions can be done for $\hat{z}^1_{i_2}$, $\hat{z}^1_{i_3}$, $\hat{z}^2_{i_1}$, $\hat{z}^2_{i_2}$, and $\hat{z}^2_{i_3}$. Also, denote all the second order divided difference terms in these discrete Taylor's expansion as $D^2\hat{z}^1_{i_j,k_i,\ell}$, $j = 1, 2, 3$. Hence,

$$\begin{aligned}\tilde{z}^1_\ell &= \psi_{i_1}(X_\ell)\hat{z}^1_{i_1} + \psi_{i_2}(X_\ell)\hat{z}^1_{i_2} + \psi_{i_3}(X_\ell)\hat{z}^1_{i_3} \\ &= \psi_{i_1}(X_\ell)\left(\hat{z}^1_\ell + D_{x,p_3}\hat{z}^1_\ell(x_{i_1}-x_\ell) + D_{y,p_4}\hat{z}^1_\ell(y_{i_1}-y_\ell) + O(h^2)D^2\hat{z}^1_{i_1,k_i,\ell}\right) \\ &\quad + \psi_{i_2}(X_\ell)\left(\hat{z}^1_\ell + D_{x,p_3}\hat{z}^1_\ell(x_{i_2}-x_\ell) + D_{y,p_4}\hat{z}^1_\ell(y_{i_2}-y_\ell) + O(h^2)D^2\hat{z}^1_{i_2,k_i,\ell}\right) \\ &\quad + \psi_{i_3}(X_\ell)\left(\hat{z}^1_\ell + D_{x,p_3}\hat{z}^1_\ell(x_{i_3}-x_\ell) + D_{y,p_4}\hat{z}^1_\ell(y_{i_3}-y_\ell) + O(h^2)D^2\hat{z}^1_{i_3,k_i,\ell}\right).\end{aligned} \tag{22}$$

We can obtain a similar expression for $\tilde{z}^2$. Noting that $\sum_{j=1}^3 \psi_{i_j}(X_\ell) = 1$, $\sum_{j=1}^3 (x_{i_j} - x_\ell)\psi_{i_j}(X_\ell) = 0$, and $\sum_{j=1}^3 (y_{i_j} - y_\ell)\psi_{i_j}(X_\ell) = 0$ we then have

$$|\hat{z}^1_\ell - \tilde{z}^1_\ell| \le Ch^2 \sum_{j=1}^3 |D^2\hat{z}^1_{i_j,k_i,\ell}|, \quad |\hat{z}^2_\ell - \tilde{z}^2_\ell| \le Ch^2 \sum_{j=1}^3 |D^2\hat{z}^2_{i_j,k_i,\ell}|, \tag{23}$$

where $D^2\hat{z}_{i_j,k_i,\ell}$ are defined above as a sum of a number of second order divided differences involved with three points in the element E_i, C is a generic positive constant, and h is the maximum size of all triangular elements. Similarly, we can estimate discrete derivatives of the interpolation error,

$$\begin{aligned}\nabla^h_{s,i}\tilde{z} = \nabla^h_{s,i}\tilde{z}_\ell &= \sum_{j=1}^3 \nabla_{s,i}\psi_{i_j}\hat{z}_{i_j} \\ &= \sum_{j=1}^3 \nabla_{s,i}\psi_{i_j}\left(\hat{z}_\ell + D_{x,p_3}\hat{z}_\ell(x_{i_j}-x_\ell) + D_{y,p_4}\hat{z}_\ell(y_{i_j}-y_\ell) + O(h^2)D^2\hat{z}_{i_j,k_i,\ell}\right),\end{aligned}$$

where s represents x or y. Noting that $\sum_{j=1}^3 \nabla_{s,i}\psi_{i_j} = 0$, $\sum_{j=1}^3 (x_{i_j} - x_\ell)\nabla_{x,i}\psi_{i_j} = 1$, $\sum_{j=1}^3 (y_{i_j} - y_\ell)\nabla_{x,i}\psi_{i_j} = 0$, $\sum_{j=1}^3 (x_{i_j} - x_\ell)\nabla_{y,i}\psi_{i_j} = 0$, and $\sum_{j=1}^3 (y_{i_j} - y_\ell)\nabla_{y,i}\psi_{i_j} = 1$ we have

$$|\nabla_{s,i}\hat{z}_\ell - \nabla^h_{s,i}\tilde{z}_\ell| \le Ch \sum_{j=1}^3 |D^2\hat{z}_{i_j,k_i,\ell}|, \text{ for } X_\ell \in E_i, \tag{24}$$

where s is x or y and

$$\nabla_{x,i}\hat{z}_\ell = (D_{x,p_3}\hat{z}^1_\ell, D_{x,p_3}\hat{z}^2_\ell)^T, \quad \nabla_{y,i}\hat{z}_\ell = (D_{y,p_4}\hat{z}^1_\ell, D_{y,p_4}\hat{z}^2_\ell)^T. \tag{25}$$

Note that if ℓ is located at a position such that $x_{k_i} = x_{p_3} = x_\ell$ or $y_{k_i} = y_{p_4} = y_\ell$, then $D_{x,p_3}\hat{z}_\ell$ or $D_{y,p_4}\hat{z}_\ell$ may not be defined, respectively. For these ℓ's we can change $k_i = (k_x, k_y)$ into its neighbor, e.g., $k_i = (k_x + 1, k_y)$ or $k_i = (k_x, k_y + 1)$, in definition (25), respectively. All previous arguments and results about the interpolation error will still be valid. We thus have

$$\|\hat{z} - \tilde{z}\|_2 \le C_2 h^2, \quad \|\nabla_s \hat{z} - \nabla^h_s \tilde{z}\|_2 \le C_2 h, \tag{26}$$

where s represents x or y, $\|\cdot\|_2 = \sqrt{\frac{1}{M}\sum_{i=1}^N \sum_{\ell=1}^{m_i} |(\cdot)_\ell|^2}$ (where $M = \sum_{i=1}^N m_i$), C_2 is a positive constant proportional to the ℓ_2 norm of second order divided differences of $\hat{z}$, and $\nabla_s \hat{z}$ and $\nabla^h_s \tilde{z}$ are vectors with components $\nabla_{s,i}\hat{z}_\ell$ and $\nabla^h_{s,i}\tilde{z}_\ell$, respectively.

Therefore, to estimate $\hat{Z} - \hat{z}$ we need only estimate $\hat{Z} - \tilde{z}$. We first have

$$\begin{aligned} &a_h(\hat{Z}, \hat{Z} - \tilde{z}) - a_h(\tilde{z}, \hat{Z} - \tilde{z}) \\ &= \frac{1}{2}\sum_{i=1}^N m_i \sum_{\ell \in R_{k_i}} \left[g(\hat{Z}_\ell - \hat{Z}_{k_i}) - g(\tilde{z}_\ell - \tilde{z}_{k_i})\right] \cdot \left[(\hat{Z}_\ell - \hat{Z}_{k_i}) - (\tilde{z}_\ell - \tilde{z}_{k_i})\right] \\ &= \frac{1}{2}\sum_{i=1}^N m_i \sum_{\ell \in R_{k_i}} \left[(\hat{Z}_\ell - \hat{Z}_{k_i}) - (\tilde{z}_\ell - \tilde{z}_{k_i})\right]^T g_\alpha(\xi_\ell) \left[(\hat{Z}_\ell - \hat{Z}_{k_i}) - (\tilde{z}_\ell - \tilde{z}_{k_i})\right], \end{aligned} \tag{27}$$

where $\xi_\ell = \mu_\ell(\hat{Z}_\ell - \hat{Z}_{k_i}) + (1 - \mu_\ell)(\tilde{z}_\ell - \tilde{z}_{k_i})$, $0 < \mu_\ell < 1$. From (5), (6), and (17) and similarly to (8), if ℓ is in the element E_i, then

$$\begin{aligned} \hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i}) &= (x_\ell - x_{k_i}) \sum_{j=1}^3 \nabla_{x,i}\psi_{i_j}(\hat{Z}^h_{i_j} - \hat{z}_{i_j}) \\ &\quad + (y_\ell - y_{k_i}) \sum_{j=1}^3 \nabla_{y,i}\psi_{i_j}(\hat{Z}^h_{i_j} - \hat{z}_{i_j}) \\ &= (x_\ell - x_{k_i})\nabla^h_{x,i}(\hat{Z} - \tilde{z}) + (y_\ell - y_{k_i})\nabla^h_{y,i}(\hat{Z} - \tilde{z}). \end{aligned} \tag{28}$$

Due to assumptions given in the previous section, at the nearest neighbors $\ell \in nb_{k_i}$ of $k_i = (k_{ix}, k_{iy})$, $\hat{Z}_\ell - \hat{Z}_{k_i}$ and $\tilde{z}_\ell - \tilde{z}_{k_i}$ are located in the convex region of the energy function $\phi(|\alpha|)$. Then corresponding to the nearest neighbors $\ell \in nb_{k_i}$, $\xi_\ell = \mu_\ell(\hat{Z}_\ell - \hat{Z}_{k_i}) + (1 - \mu_\ell)(\tilde{z}_\ell - \tilde{z}_{k_i})$ would be in the convex region (see (16)). Define $nb'_{k_i} = \{(k_{ix} - 1, k_{iy}), (k_{ix} + 1, k_{iy}), (k_{ix}, k_{iy} - 1), (k_{ix}, k_{iy} + 1)\} \subseteq nb_{k_i}$. From (27) and assumption 2, $a_h(\hat{Z}, \hat{Z} - \tilde{z}) - a_h(\tilde{z}, \hat{Z} - \tilde{z})$ is dominated by

$$\begin{aligned} &\sum_{\ell \in nb_{k_i}} \left[(\hat{Z}_\ell - \hat{Z}_{k_i}) - (\tilde{z}_\ell - \tilde{z}_{k_i})\right]^T g_{\alpha 1}(\xi_\ell) \left[(\hat{Z}_\ell - \hat{Z}_{k_i}) - (\tilde{z}_\ell - \tilde{z}_{k_i})\right] \\ &\quad \ge \nabla^h_{x,i}(\hat{Z} - \tilde{z})^T \sum_{\ell \in nb'_{k_i}} (x_\ell - x_{k_i})^2 g_{\alpha 1}(\xi_\ell) \nabla^h_{x,i}(\hat{Z} - \tilde{z}) + \nabla^h_{y,i}(\hat{Z} - \tilde{z})^T \\ &\quad \times \sum_{\ell \in nb'_{k_i}} (y_\ell - y_{k_i})^2 g_{\alpha 1}(\xi_\ell) \nabla^h_{y,i}(\hat{Z} - \tilde{z}). \end{aligned}$$

By easy calculation we have eigenvalues

$$\lambda\left(\sum_{\ell\in nb'_{k_i}}(x_\ell - x_{k_i})^2 g_{\alpha 1}(\xi_\ell)\right) \text{ and } \lambda\left(\sum_{\ell\in nb'_{k_i}}(y_\ell - y_{k_i})^2 g_{\alpha 1}(\xi_\ell)\right) > p > 0. \tag{29}$$

From Lemma 1, (14), and (1), p is independent of lattice constant r_0 but depends only on $\sin^2\gamma$ and $\min_{t\in(0,C_1]}\phi_1''(t)$, where $\phi_1(t) = \phi(r_0 t)$, and the meaning of C_1 and γ are given in Lemma 1 and assumptions 1 and 2. Therefore

$$a_h(\hat{Z}, \hat{Z} - \tilde{z}) - a_h(\tilde{z}, \hat{Z} - \tilde{z}) > \frac{p}{2}\sum_{i=1}^{N} m_i \left(|\nabla^h_{x,i}(\hat{Z} - \tilde{z})|^2 + |\nabla^h_{y,i}(\hat{Z} - \tilde{z})|^2\right), \tag{30}$$

where $|\cdot|$ is the Euclidean length of $\mathbf{R}^2$.

On the other hand, from (12) (taking $V = \hat{Z} - \tilde{z}$) and using (4) (taking $v = \hat{Z} - \tilde{z}$), and noting

$$a_h(\hat{Z}, \hat{Z} - \tilde{z}) = -\sum_{i=1}^{N}\sum_{k=1}^{m_i} f(\hat{Z}_k)\cdot(\hat{Z}_k - \tilde{z}_k), \quad a(\hat{z}, \hat{Z} - \tilde{z}) = -\sum_{i=1}^{N}\sum_{k=1}^{m_i} f(\hat{z}_k)\cdot(\hat{Z}_k - \tilde{z}_k),$$

we have

$$\begin{aligned} &a_h(\hat{Z}, \hat{Z} - \tilde{z}) - a_h(\tilde{z}, \hat{Z} - \tilde{z}) \\ &\quad = \left(a_h(\hat{z}, \hat{Z} - \tilde{z}) - a_h(\tilde{z}, \hat{Z} - \tilde{z})\right) - \left(a_h(\hat{z}, \hat{Z} - \tilde{z}) - a(\hat{z}, \hat{Z} - \tilde{z})\right) \\ &\qquad + \sum_{i=1}^{N}\sum_{k=1}^{m_i}(f(\hat{z}_k) - f(\hat{Z}_k))\cdot(\hat{Z}_k - \tilde{z}_k) = T_1 - T_2 + T_3, \end{aligned}$$

where T_1 and T_2 correspond to the first two bracketed terms in the expression, respectively, and T_3 is the double sum term. We now treat them one by one. We can easily have

$$\begin{aligned} T_1 &= \frac{1}{2}\sum_{i=1}^{N} m_i \sum_{\ell\in R_{k_i}} (g(\hat{z}_\ell - \hat{z}_{k_i}) - g(\tilde{z}_\ell - \tilde{z}_{k_i}))\cdot\left(\hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i})\right) \\ &= \frac{1}{2}\sum_{i=1}^{N} m_i \sum_{\ell\in R_{k_i}} (\hat{z}_\ell - \hat{z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i}))^T\, g_\alpha(\xi'_\ell)\left(\hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i})\right), \end{aligned}$$

where $\xi'_\ell = \mu'_\ell(\hat{z}_\ell - \hat{z}_{k_i}) + (1 - \mu'_\ell)(\tilde{z}_\ell - \tilde{z}_{k_i})$, $0 < \mu'_\ell < 1$. We can write

$$\hat{z}_\ell - \hat{z}_{k_i} = D_{x,\ell}\hat{z}_{(k_{ix},\ell_y)}(x_\ell - x_{k_i}) + D_{y,(k_{ix},\ell_y)}\hat{z}_{k_i}(y_\ell - y_{k_i}),$$

where k_{ix} is the first component of the index k_i and ℓ_y is the second component of the index ℓ. Hence,

$$\begin{aligned} \hat{z}_\ell - \hat{z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i}) = &\left(D_{x,\ell}\hat{z}_{(k_{ix},\ell_y)} - \nabla^h_{x,i}\tilde{z}\right)(x_\ell - x_{k_i}) \\ &+ \left(D_{y,(k_{ix},\ell_y)}\hat{z}_{k_i} - \nabla^h_{y,i}\tilde{z}\right)(y_\ell - y_{k_i}). \end{aligned} \tag{31}$$

So similarly to the argument of (24), in estimating the interpolation error we have $\tilde{z}_\ell - \tilde{z}_{k_i} = \hat{z}_\ell - \hat{z}_{k_i} + O(hr_0)$. Taking the inner product of it with $\hat{z}_\ell - \hat{z}_{k_i}$ and using assumption 1 (the first inequality $c_1 r_0 \le |\alpha|$), we can obtain

$$(\hat{z}_\ell - \hat{z}_{k_i}) \cdot (\tilde{z}_\ell - \tilde{z}_k) \ge |\hat{z}_\ell - \hat{z}_{k_i}|^2 - O(hr_0)|\hat{z}_\ell - \hat{z}_{k_i}| \ge c_1^2 r_0^2 - O(hr_0) C_1 r_0 > 0 \tag{32}$$

if h is sufficiently small. Using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} |T_1| \le \frac{1}{2} &\left(\sum_{i=1}^{N} m_i \sum_{\ell \in R_{k_i}} (\hat{z}_\ell - \hat{z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i}))^T \, g_\alpha(\xi'_\ell) \, (\hat{z}_\ell - \hat{z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i})) \right)^{\frac{1}{2}} \\ &\times \left(\sum_{i=1}^{N} m_i \sum_{\ell \in R_{k_i}} \left(\hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i}) \right)^T g_\alpha(\xi'_\ell) \left(\hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i}) \right) \right)^{\frac{1}{2}}. \end{aligned}$$

To get an upper bound of T_1 we need to get an upper bound of eigenvalues in the left-hand side of (29). We then need a lower bound of the size of ξ'_ℓ. From assumption 1 we have $|\hat{z}_\ell - \hat{z}_{k_i}| \ge \bar{c}_1 r_0$ and $|\tilde{z}_\ell - \tilde{z}_{k_i}| \ge \bar{c}_1 r_0$ (using (31)), where $\bar{c}_1 = \min\{c_1, c_1 + O(h)\} > 0$. Then using (32) yields

$$\begin{aligned} |\xi'_\ell|^2 &= \mu_\ell^{'2} |\hat{z}_\ell - \hat{z}_{k_i}|^2 + (1-\mu'_\ell)^2 |\tilde{z}_\ell - \tilde{z}_{k_i}|^2 + 2\mu'_\ell (1-\mu'_\ell)(\hat{z}_\ell - \hat{z}_{k_i}) \cdot (\tilde{z}_\ell - \tilde{z}_{k_i}) \\ &\ge (\mu_\ell^{'2} + (1-\mu'_\ell)^2) \bar{c}_1^2 r_0^2 \ge \frac{1}{2} \bar{c}_1^2 r_0^2. \end{aligned}$$

Combining it with the second inequality of (16) in assumption 1, we thus have

$$\frac{1}{\sqrt{2}} \bar{c}_1 r_0 \le |\xi'_\ell| \le C_1 r_0. \tag{33}$$

Based on this and assumption 2, we can easily verify the eigenvalues

$$\lambda \left(\sum_{\ell \in R_{k_i}} (x_\ell - x_{k_i})^2 g_\alpha(\xi'_\ell) \right) \text{ and } \lambda \left(\sum_{\ell \in R_{k_i}} (y_\ell - y_{k_i})^2 g_\alpha(\xi'_\ell) \right) \le P, \tag{34}$$

where P is a positive constant depending on $\max_{t \in [\frac{1}{\sqrt{2}} \bar{c}_1, C_1]} \phi_1''(t)$ but is independent of lattice constant r_0. From (34), (31), (24), and (28) we then have

$$|T_1| \le C_2 h M^{\frac{1}{2}} \left(\sum_{i=1}^{N} m_i \left(|\nabla_{x,i}^h (\hat{Z} - \tilde{z})|^2 + |\nabla_{y,i}^h (\hat{Z} - \tilde{z})|^2 \right) \right)^{\frac{1}{2}}, \tag{35}$$

where $M = \sum_{i=1}^{N} m_i$ is the total number of atoms and the positive constant C_2 is proportional to the ℓ_2 norm of second order divided difference of $\hat{z}$ as defined in (26) (noting that the constant C in (24) and the constant P in (34) are absorbed into the constant C_2). The third term can be written as

$$T_3 = \sum_{i=1}^{N} \sum_{k=1}^{m_i} (f(\tilde{z}_k) - f(\hat{Z}_k)) \cdot (\hat{Z}_k - \tilde{z}_k) + \sum_{i=1}^{N} \sum_{k=1}^{m_i} (f(\hat{z}_k) - f(\tilde{z}_k)) \cdot (\hat{Z}_k - \tilde{z}_k), \tag{36}$$

where the first sum is less than $-F''_{min}\sum_{i=1}^{N}\sum_{k=1}^{m_i}|\hat{Z}_k-\tilde{z}_k|^2$ ($F''_{min}>0$) due to the strictly convex assumption of the external energy F. The second sum is less than

$$\frac{1}{2F''_{min}}\sum_{i=1}^{N}\sum_{k=1}^{m_i}|f(\hat{z}_k)-f(\tilde{z}_k)|^2+\frac{F''_{min}}{2}\sum_{i=1}^{N}\sum_{k=1}^{m_i}|\hat{Z}_k-\tilde{z}_k|^2$$

due to the Cauchy inequality. From (23) and (26) we then have

$$T_3\le C_2^2h^4M-\frac{F''_{min}}{2}\sum_{i=1}^{N}\sum_{k=1}^{m_i}|\hat{Z}_k-\tilde{z}_k|^2=C_2^2h^4M-\frac{F''_{min}}{2}M\|\hat{Z}-\tilde{z}\|_2^2. \tag{37}$$

Next we consider the second term T_2 (nonconforming part of the error).

From (4) we have

$$\begin{aligned}T_2&=a_h(\hat{z},\hat{Z}-\tilde{z})-a(\hat{z},\hat{Z}-\tilde{z})\\&=\frac{1}{2}\sum_{i=1}^{N}m_i\sum_{\ell\in R_{k_i}}g(\hat{z}_\ell-\hat{z}_{k_i})\cdot(\hat{Z}_\ell-\hat{Z}_{k_i}-(\tilde{z}_\ell-\tilde{z}_{k_i}))\\&\quad-\frac{1}{2}\sum_{i=1}^{N}\sum_{k=1}^{m_i}\sum_{j\in R_k}g(\hat{z}_j-\hat{z}_k)\cdot(\hat{Z}_j-\hat{Z}_k-(\tilde{z}_j-\tilde{z}_k)).\end{aligned}$$

Define

$$Y_j=\psi_{i_1}(X_j)\hat{Z}^h_{i_1}+\psi_{i_2}(X_j)\hat{Z}^h_{i_2}+\psi_{i_3}(X_j)\hat{Z}^h_{i_3}\quad\forall j\in R_k,X_k\in E_i, \tag{38}$$

$$\tilde{y}_j=\psi_{i_1}(X_j)\hat{z}_{i_1}+\psi_{i_2}(X_j)\hat{z}_{i_2}+\psi_{i_3}(X_j)\hat{z}_{i_3}\quad\forall j\in R_k,X_k\in E_i, \tag{39}$$

where Y_j (or $\tilde{y}_j$) is a linear extension of the linear lattice function $\hat{Z}$ (or $\tilde{z}$, respectively) from the triangular element E_i to the neighbor elements; i.e., we have

$$Y_j=\hat{Z}_j\quad\text{and}\quad\tilde{y}_j=\tilde{z}_j\quad\text{if}\quad X_j\in E_i. \tag{40}$$

Then we can write

$$T_2=\frac{1}{2}(A_1+A_2),$$

where

$$\begin{aligned}A_1&=\sum_{i=1}^{N}m_i\sum_{\ell\in R_{k_i}}g(\hat{z}_\ell-\hat{z}_{k_i})\cdot(\hat{Z}_\ell-\hat{Z}_{k_i}-(\tilde{z}_\ell-\tilde{z}_{k_i}))\\&\quad-\sum_{i=1}^{N}\sum_{k=1}^{m_i}\sum_{j\in R_k}g(\hat{z}_j-\hat{z}_k)\cdot(Y_j-Y_k-(\tilde{y}_j-\tilde{y}_k)),\\A_2&=\sum_{i=1}^{N}\sum_{k=1}^{m_i}\sum_{j\in R_k}g(\hat{z}_j-\hat{z}_k)\left[(Y_j-Y_k-(\tilde{y}_j-\tilde{y}_k))-(\hat{Z}_j-\hat{Z}_k-(\tilde{z}_j-\tilde{z}_k))\right].\end{aligned}$$

The term A_1 can be estimated by using Taylor's theorem to the function g and shifting $\hat{z}_k$ to $\hat{z}_{k_i}$ and $\hat{z}_j$ to $\hat{z}_\ell$ with $X_j-X_k=X_\ell-X_{k_i}$ (see Figure 7 for relative positions of atoms j, k, ℓ, and k_i). From the definitions of Y_j and $\tilde{y}_j$ in (38) and (39), and noting

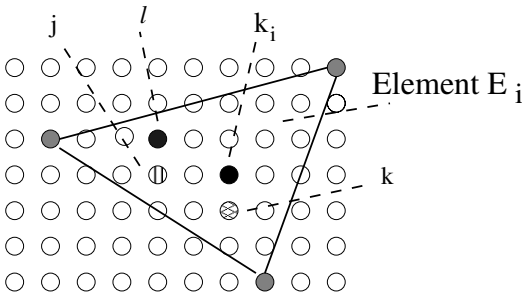


FIG. 7. *Shifting of atom k to k_i and of atom j to ℓ.*

that $X_j - X_k = X_\ell - X_{k_i}$ and ψ_{i_j} $(j = 1, 2, 3)$ are linear, we should have

$$Y_j - Y_k = \hat{Z}_\ell - \hat{Z}_{k_i}, \quad \tilde{y}_j - \tilde{y}_k = \tilde{z}_\ell - \tilde{z}_{k_i}.$$

Therefore,

$$\begin{aligned} A_1 &= \sum_{i=1}^{N} \sum_{k=1}^{m_i} \sum_{j \in R_k} \left(g(\hat{z}_j - \hat{z}_k) - g(\hat{z}_\ell - \hat{z}_{k_i})\right) \left(\hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i})\right) \\ &= \sum_{i=1}^{N} \sum_{k=1}^{m_i} \sum_{j \in R_k} \left(\hat{z}_j - \hat{z}_k - (\hat{z}_\ell - \hat{z}_{k_i})\right)^T g_\alpha(\xi_A)(\hat{Z}_\ell - \hat{Z}_{k_i} - (\tilde{z}_\ell - \tilde{z}_{k_i})), \end{aligned}$$

where $\xi_A = \mu_A(\hat{z}_j - \hat{z}_k) + (1 - \mu_A)(\hat{z}_\ell - \hat{z}_{k_i})$ and $\mu_A \in [0, 1]$. We can write

$$\begin{aligned} \hat{z}_j - \hat{z}_k - (\hat{z}_\ell - \hat{z}_{k_i}) = &(D_{x,j}\hat{z}_{(k_x, j_y)} - D_{x,\ell}\hat{z}_{(k_{ix}, \ell_y)})(x_\ell - x_{k_i}) \\ &+ (D_{y,(k_x,j_y)}\hat{z}_k - D_{y,(k_{ix},\ell_y)}\hat{z}_{k_i})(y_\ell - y_{k_i}), \end{aligned}$$

where k_x and k_{ix} are the first components of k and k_i, respectively; j_y and l_y are the second components of j and l, respectively, and the differences of first order divided differences are of order $O(h)$. Similarly to the arguments of (32) and (33), we can have $(\hat{z}_j - \hat{z}_k) \cdot (\hat{z}_\ell - \hat{z}_{k_i}) > 0$ and then $\frac{1}{\sqrt{2}}\bar{c}_1 r_0 \le |\xi_A| \le C_1 r_0$. Applying the Cauchy–Schwarz inequality and following the steps in estimating T_1, we can thus have

$$|A_1| \le C_2 h M^{\frac{1}{2}} \left(\sum_{i=1}^{N} m_i \left(|\nabla^h_{x,i}(\hat{Z} - \tilde{z})|^2 + |\nabla^h_{y,i}(\hat{Z} - \tilde{z})|^2 \right) \right)^{\frac{1}{2}}. \tag{41}$$

Next we estimate A_2. Noting (40) we have

$$\begin{aligned} A_2 &= \sum_{i=1}^{N} \sum_{k=1}^{m_i} \sum_{j \in R_k} g(\hat{z}_j - \hat{z}_k) \cdot (Y_j - \tilde{y}_j - (\hat{Z}_j - \tilde{z}_j)) \\ &= \sum_{i=1}^{N} \sum_{j \in S_i} g(\hat{z}_j - \hat{z}_k) \cdot (Y_j - \tilde{y}_j - (\hat{Z}_j - \tilde{z}_j)), \end{aligned}$$

where $S_i = \{j : j \in R_k,\ X_k \in E_i,\ X_j \bar{\in} E_i\}$ and R_k is the cut-off disc of atom X_k. The number of atoms in the set S_i is the number of atoms near the boundary of triangle

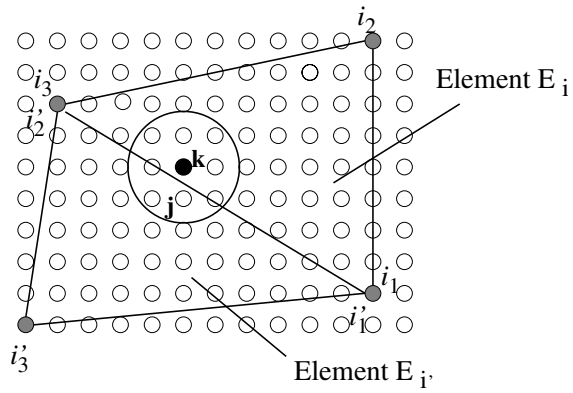


FIG. 8. *Cut-off discs near the boundary of two neighboring elements.*

E_i (see Figure 8). Denote the number as b_i. Also denote the line connecting atoms i_1 and i_3 (or i_1' and i_2') as i_1i_3 and the line connecting j and k as jk and choose an atom $X_{j'}$ on the line i_1i_3 to be the closest to the intersection point of lines i_1i_3 and jk. We then have

$$\psi_{i_1}(X_{j'}) = \psi_{i_1'}(X_{j'}),\ \psi_{i_3}(X_{j'}) = \psi_{i_2'}(X_{j'}),\ \psi_{i_2}(X_{j'}) = 0,\ \psi_{i_3'}(X_{j'}) = 0,$$

and $Y_{j'} - \tilde{y}_{j'} = \hat{Z}_{j'} - \tilde{z}_{j'}$. Therefore,

$$\begin{aligned} Y_j - \tilde{y}_j - (Y_{j'} - \tilde{y}_{j'}) &= \sum_{q=1}^{3} (\psi_{i_q}(X_j) - \psi_{i_q}(X_{j'}))(\hat{Z}^h_{i_q} - \hat{z}_{i_q}) \\ &= (x_j - x_{j'})\nabla^h_{x,i}(\hat{Z} - \tilde{z}) + (y_j - y_{j'})\nabla^h_{y,i}(\hat{Z} - \tilde{z}) \\ \hat{Z}_j - \tilde{z}_j - (Y_{j'} - \tilde{y}_{j'}) &= \sum_{q=1}^{3} (\psi_{i_q'}(X_j) - \psi_{i_q'}(X_{j'}))(\hat{Z}^h_{i_q'} - \hat{z}_{i_q'}) \\ &= (x_j - x_{j'})\nabla^h_{x,i'}(\hat{Z} - \tilde{z}) + (y_j - y_{j'})\nabla^h_{y,i'}(\hat{Z} - \tilde{z}). \end{aligned}$$

From the nearest neighbor dominance and nearest neighbor pairs located in the convex region of ϕ (assumptions 1–3) we can obtain that $|\sum_{j\in R_k} g(\hat{z}_j - \hat{z}_k)(x_j - x_{j'})|$ and $|\sum_{j\in R_k} g(\hat{z}_j - \hat{z}_k)(y_j - y_{j'})|$ are bounded by a constant independent of lattice scale and mesh size (like P in (34)). Then, noting

$$Y_j - \tilde{y}_j - (\hat{Z}_j - \tilde{z}_j) = Y_j - \tilde{y}_j - (Y_{j'} - \tilde{y}_{j'}) + (Y_{j'} - \tilde{y}_{j'}) - (\hat{Z}_j - \tilde{z}_j),$$

we have

$$\begin{aligned} |A_2| &\le C \sum_{i=1}^{N} b_i \left(|\nabla^h_{x,i}(\hat{Z} - \tilde{z})| + |\nabla^h_{y,i}(\hat{Z} - \tilde{z})| + |\nabla^h_{x,i'}(\hat{Z} - \tilde{z})| + |\nabla^h_{y,i'}(\hat{Z} - \tilde{z})| \right) \\ &\le C \left(\sum_{i=1}^{N} \frac{b_i^2}{m_i} \right)^{\frac{1}{2}} \left(\sum_{i=1}^{N} m_i (|\nabla^h_{x,i}(\hat{Z} - \tilde{z})|^2 + |\nabla^h_{y,i}(\hat{Z} - \tilde{z})|^2) \right)^{\frac{1}{2}}. \end{aligned} \tag{42}$$

Hence,

$$|T_2| \le \frac{1}{2}(|A_1| + |A_2|)$$

$$\le C\left(C_2 h M^{\frac{1}{2}} + \sqrt{\sum_{i=1}^{N} \frac{b_i^2}{m_i}}\right)\left(\sum_{i=1}^{N} m_i(|\nabla_{x,i}^h(\hat{Z}-\tilde{z})|^2 + |\nabla_{y,i}^h(\hat{Z}-\tilde{z})|^2)\right)^{\frac{1}{2}}. \tag{43}$$

We can then have the following error estimate.

THEOREM 1. *Let $\hat{Z}$ and $\hat{z}$ be the solutions of* (4) *and* (12)*, respectively, satisfying Dirichlet boundary conditions, and let assumptions* 1–3 *hold. Then we have*

$$\|\hat{Z}-\hat{z}\|_2 + \|\nabla_x^h \hat{Z} - \nabla_x \hat{z}\|_2 + \|\nabla_y^h \hat{Z} - \nabla_y \hat{z}\|_2 \le C\left(C_2 h + \sqrt{\frac{\sum_{i=1}^{N} \frac{b_i^2}{m_i}}{M}}\right), \tag{44}$$

where C is a generic positive constant, ∇_x^h, ∇_y^h, ∇_x, and ∇_y are discrete derivatives, norm $\|\cdot\|_2$ and constant C_2 are defined as in (26) *and* (25)*, $M = \sum_{i=1}^{N} m_i$, is the total number of atoms, m_i is the number of atoms in the triangular element E_i, and b_i is the number of atoms near at least one boundary of E_i and is less than a constant times the number of atoms located at the longest side of E_i.*

Proof. Using (30) and estimates (35), (43), and (37) for T_1, T_2, and T_3, respectively, and letting

$$H_1 = \left(\frac{1}{M}\sum_{i=1}^{N} m_i(|\nabla_{x,i}^h(\hat{Z}-\tilde{z})|^2 + |\nabla_{y,i}^h(\hat{Z}-\tilde{z})|^2)\right)^{\frac{1}{2}} \text{ and } H_2 = \|\hat{Z}-\tilde{z}\|_2,$$

we have

$$\frac{p}{2}H_1^2 M < a_h(\hat{Z}, \hat{Z}-\tilde{z}) - a_h(\tilde{z}, \hat{Z}-\tilde{z}) \le C\left(C_2 h + \sqrt{\frac{\sum_{i=1}^{N} \frac{b_i^2}{m_i}}{M}}\right) H_1 M$$

$$- \frac{F''_{min}}{2} H_2^2 M + C_2^2 h^4 M$$

or (noting $(H_1^2 + H_2^2) \ge (H_1 + H_2)^2/2$ and $H_1 \le H_1 + H_2$)

$$p_{min}(H_1+H_2)^2 < C\left(C_2 h + \sqrt{\frac{\sum_{i=1}^{N} \frac{b_i^2}{m_i}}{M}}\right)(H_1+H_2) + C_2^2 h^4,$$

where $p_{min} = \min\{p, F''_{min}\}/4$. Solving this inequality for $H_1 + H_2$, we thus have

$$H_1 + H_2 < C\left(C_2 h + \sqrt{\frac{\sum_{i=1}^{N} \frac{b_i^2}{m_i}}{M}}\right).$$

Then $\|\nabla_x^h \hat{Z} - \nabla_x^h \tilde{z}\|_2 + \|\nabla_y^h \hat{Z} - \nabla_y^h \tilde{z}\|_2$ can have the same estimate since

$$\|\nabla_x^h \hat{Z} - \nabla_x^h \tilde{z}\|_2 + \|\nabla_y^h \hat{Z} - \nabla_y^h \tilde{z}\|_2 \le \sqrt{2}\left(\frac{1}{M}\sum_{i=1}^{N}\sum_{k=1}^{m_i}(|\nabla_{x,i}^h(\hat{Z}-\tilde{z})|^2 + |\nabla_{y,i}^h(\hat{Z}-\tilde{z})|^2)\right)^{\frac{1}{2}}$$

$$= \sqrt{2}\, H_1.$$

Combining this with the interpolation error estimates (26), we obtain (44). □

The constant C_2 in the error estimate should be of reasonable size if all the components of the solution $\hat{z}$ defined on a reference configuration, as shown in the introduction, formulate a not-too-rough surface.

Remark 1. If the external potential $F(z)$ is convex but not strictly convex, the analysis may still be done with some small modification. For example, the body force energy $F(z) = -\sum_{i=1}^{N}\sum_{k=1}^{m_i} f_k z_k$ is convex but not strictly convex, where the constant f_k is the force applied on atom k. In this case, $T_3 = 0$ in (36) and there is no need to estimate it further. The theorem without the first term $\|\hat{Z} - \hat{z}\|_2$ in (44) remains true. For a general convex (but not strictly convex) external energy $F(z)$, we need to use a Poincaré type of inequality to control $\|\hat{Z} - \tilde{z}\|_2$ resulting from estimating T_3 in (36). Note that $\hat{Z} - \tilde{z}$ is a piecewise linear interpolation based on Z^h and values of $\hat{z}$ at vertices of the triangulation. So $\hat{Z} - \tilde{z} \in C^0$ or H^1 (and satisfying the homogeneous Dirichlet boundary condition) from the finite element theory. Then from the Poincaré inequality, we have

$$\int_\Omega |\hat{Z} - \tilde{z}|^2 \le C_d \left(\int_\Omega |\nabla_x^h(\hat{Z} - \tilde{z})|^2 + |\nabla_y^h(\hat{Z} - \tilde{z})|^2 \right),$$

where Ω is the material domain in the reference configuration and C_d is a generic constant depending on the size of the domain. Since first order derivatives of $\hat{Z} - \tilde{z}$ are constant in each element, we may derive from above inequality the Poincaré inequality in the discrete ℓ_2 norm $\|\cdot\|_2$:

$$\|\hat{Z} - \tilde{z}\|_2^2 \le \bar{C}_d \left(\|\nabla_x^h(\hat{Z} - \tilde{z})\|_2^2 + \|\nabla_y^h(\hat{Z} - \tilde{z})\|_2^2 \right).$$

Then we can estimate T_3 in (36) as the following:

$$|T_3| \le (\delta - F''_{min}) \sum_{i=1}^{N}\sum_{k=1}^{m_i} |\hat{Z}_k - \tilde{z}_k|^2 + \frac{1}{4\delta} \sum_{i=1}^{N}\sum_{k=1}^{m_i} |f(\hat{z}_k) - f(\tilde{z}_k)|^2,$$

where $F''_{min} \ge 0$. Taking $\delta < p/4\bar{C}_d$ (where p is the constant defined in (30)) we may obtain the result of Theorem 1 in the case of the convex external energy.

Remark 2. When a material has defects, there will be a certain amount of pairs of nearest neighbor atoms whose distance is located in the nonconvex region of the pair potential energy ϕ. In this case we may divide the domain of ϕ into two parts I (convex region, $\phi'' > 0$) and II (nonconvex region) (cf. [16]). We call the material parts corresponding to I and II the nondefect and defect parts, respectively, and collect all indices of triangular elements in the nondefect and defect parts into two sets R^I and R^{II}, respectively. Denote $a_h^I(\cdot,\cdot)$ and $a_h^{II}(\cdot,\cdot)$ as the parts of $a_h(\cdot,\cdot)$ corresponding to R^I and R^{II}, respectively. Then our previous argument may apply to $a_h^I(\cdot,\cdot) = a_h(\cdot,\cdot) - a^{II}(\cdot,\cdot)$ under assumptions 1 and 2 (in the nondefect part), assumption 3, and an additional assumption that in the defect part, $\hat{z}$, $\hat{Z}$, $\nabla_x^h\hat{Z}$, $\nabla_x\hat{z}$, $\nabla_y^h\hat{Z}$, and $\nabla_y\hat{z}$ are all bounded. Then we may have the estimate

(45)

$$\|\hat{Z} - \hat{z}\|_2 + \|\nabla_x^h\hat{Z} - \nabla_x\hat{z}\|_2 + \|\nabla_y^h\hat{Z} - \nabla_y\hat{z}\|_2 \le C \left(C_2 h + \sqrt{\frac{\sum_{i\in R^I} \frac{b_i^2}{m_i}}{M}} + \sqrt{\frac{n^d}{M}} \right),$$

where n^d is the number of atoms in the defect part (indices in R^{II}) of the material. So to have a small error of the QC approximation, the number of atoms n^d in the defect part of the material should be relatively small in comparison with the total number of atoms. That is, serious defects should occur only in a small region in order for the QC method to work. The estimate (45) is a macroscopic scale error estimate. It does not provide an error estimate inside the defect region, where a much smaller scale needs to be used. In fact, due to the nonconvex property in the defect region, the lattice solution may not even be unique. Nevertheless, it seems to us that a macroscopic error estimate is all we may expect for the model.

5. Conclusion. The QC approximation (or method) is a kind of representative of a number of recent atomistic/continuum models for steady-state material problems. It may be a useful technique for model reduction of large scale problems in other fields. Numerical analysis is in its infancy despite its great success in simulation of material problems in engineering literature. In this paper we introduce a mathematical description of the method based on the energy minimization and a nonconforming finite element framework. We prove the convergence of the solution of the QC approximation to the solution of the original atomic scale energy minimization problem in two dimensions (using the Lennard–Jones pair potential). Some mathematical understanding of usual atomistic assumptions for solid materials such as nearest neighbor interaction dominance and convex nearest neighbor pair potential energy is made into assumptions for convergence analysis. In the case that these assumptions do not hold, an expected convergence result is also given as a remark. The framework of analysis may apply to problems with different pair potential energies (or with material impurity) and to three-dimensional problems.

Acknowledgment. The author would like to thank the anonymous referees for their valuable suggestions in improving the paper.

REFERENCES

[1] F. F. Abraham, J. Q. Broughton, N. Bernstein, and E. Kaxiras, *Spanning the length scales in dynamic simulation*, Comput. Phys., 12 (1998), pp. 538–546.

[2] M. Arndt and M. Griebel, *Derivation of high order gradient continuum models from atomistic models for crystalline solids*, Multiscale Model. Simul., 4 (2005), pp. 531–562.

[3] J. L. Bassani, V. Vitek, and E. S. Alber, *Atomic-level elastic properties of interfaces and their relation to continua*, Acta Metall. Mater., 40 (1992), pp. S307–S320.

[4] X. Blanc, C. Le Bris, and F. Legoll, *Analysis of a prototypical multiscale method coupling atomistic and continuum mechanics*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 797–826.

[5] M. Born and K. Huang, *Dynamical Theory of Crystal Lattices*, Oxford University Press, Oxford, UK, 1954.

[6] A. Braides, G. Dal Maso, and A. Garroni, *Variational formulation of softening phenomena in fracture mechanics: The one-dimensional case*, Arch. Ration. Mech. Anal., 146 (1999), pp. 23–58.

[7] W. E and B. Engquist, *The heterogeneous multi-scale methods*, Commun. Math. Sci., 1 (2003), pp. 87–133.

[8] W. E and P. B. Ming, *Analysis of the multiscale method*, J. Comput Math., 19 (2004), pp. 209–220.

[9] W. E and P. B. Ming, *Cauchy-Born Rule and the Stability of Crystals: Static Problem*, Research report 05-23, Institute of Computational Mathematics and Scientific/Engineering Computing, China, 2005.

[10] F. C. Frank and J. H. van der Merwe, Proc. Roy. Soc. London Ser. A, 198 (1949), p. 205, p. 216.

[11] J. Knap and M. Ortiz, *An analysis of the quasicontinuum method*, J Mech. Phys. Solids, 49 (2001), pp. 1899-1923.

[12] S. Kohlhoff, P. Gumbsch, and H. F. Fischmeister, *Crack propagation in bcc crystals studied with a combined finite element and atomistic model*, Philos. Mag. A, 64 (1991), pp. 851–878.
[13] I. A. Kunin, *Elastic Media with Microstructure*, Springer-Verlag, New York, 1982.
[14] P. Lin, *Theoretical and numerical analysis for the quasi-continuum approximation of a material particle model*, Math. Comp., 72 (2003), pp. 657–675.
[15] P. Lin, *A Nonlinear Wave Equation of Mixed Type for Fracture Dynamics*, Research Report 777, Department of Mathematics, National University of Singapore, 2000.
[16] P. Lin and C.-W. Shu, *Numerical solution of a virtual internal bond model for material fracture*, Phys. D, 2912 (2002), pp. 1–21.
[17] R Miller and E. B. Tadmor, *The quasicontinuum method: Overview, applications and current directions*, J. Comput. Aided Material Design, 9 (2002), pp. 203–239.
[18] R. E. Rudd and J. Q. Broughton, *Concurrent coupling of length scales in solid state systems*, Phys. Statist. Sol. B, 217 (2000), pp. 251–291.
[19] V. B. Shenoy, R. Miller, E. B. Tadmor, D. Rodney, R. Phillips, and M. Ortiz, *An adaptive finite element approach to atomic-scale mechanics—the quasicontinuum method*, J. Mech. Phys. Solids, 47 (1999), pp. 611–642.
[20] L. E. Shilkrot, W. A. Curtin, and R. E. Miller, *A coupled atomisic/continuum model of defects in solids*, J. Mech. Phys. Solids, 50 (2002), pp. 2085–2106.
[21] E. B. Tadmor, M. Ortiz, and R. Phillips, *Quasicontinuum analysis of defects in solids*, Philos. Mag. A, 73 (1996), pp. 1529–1563.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 333–351

CONVERGENCE OF UNSYMMETRIC KERNEL-BASED MESHLESS COLLOCATION METHODS*

ROBERT SCHABACK†

Abstract. This paper proves convergence of variations of the unsymmetric kernel-based collocation method introduced by Kansa in 1986. Since then, this method has been very successfully used in many applications, though it may theoretically fail in special situations, and though it had no error bound or convergence proof up to now. Thus it is necessary to add assumptions or to make modifications. Our modifications prevent numerical failure by dropping strict collocation and allow a rigorous mathematical analysis proving error bounds and convergence rates. These rates improve with the smoothness of the solution, the domain, and the kernel providing the trial spaces, but they are currently not yet optimal and deserve refinement. They are based on rates of approximation to the residuals by nonstationary meshless kernel-based trial spaces, and they are independent of the type of differential operator. The results are applicable to large classes of linear problems in strong form, provided that there is a smooth solution and the test and trial discretizations are chosen with some care. Our analysis does not require assumptions like ellipticity, and it can be extended to ill-posed problems.

Key words. Kansa method, error bounds, stability, ill-posed problems, greedy adaptive solver

AMS subject classifications. 65N12, 65N35, 65N22, 65F22

DOI. 10.1137/050633366

1. Introduction.

The final goal of this paper is to prove error bounds and convergence of certain numerical techniques that approximately solve a PDE problem via an unsymmetric or even nonsquare system of linear collocation equations involving meshless kernel-based trial functions. The most popular method of this kind was first proposed by Kansa [8] in 1986, and there are many follow-up papers in engineering journals (see, e.g., [5] for a selection) that can easily be retrieved via the Internet. This is why this paper does not supply additional numerical examples. So far, the method is quite successful in applications with smooth solutions, but it can fail [7] in specially constructed situations. Consequently, it has neither error bounds nor convergence proofs for its original form, and a rigorous mathematical analysis will either require some additional assumptions or make changes to the method itself. We shall do both, but we shall stay general enough not to spoil the applicability to elliptic, parabolic, and hyperbolic problems. Therefore we need a somewhat nonstandard framework, which we sketch here first, to make sure that the reader does not get lost in the technical details we have to provide later.

Consider a linear operator equation

$$L(u) = f,\ L\ :\ U \to F \tag{1.1}$$

between normed linear spaces U and F which is to be solved for any given $f \in F$. The map L takes a *solution* $u \in U$ to its *data* $L(u)$ in F. Thus F will usually be a Cartesian product of *trace* spaces of functions prescribed on the domain or on parts of the boundary. We shall consider a large class of unsymmetric discretization methods to solve such equations approximately, and we need five essential ingredients.

*Received by the editors June 10, 2005; accepted for publication (in revised form) August 18, 2006; published electronically February 9, 2007.
http://www.siam.org/journals/sinum/45-1/63336.html
†Institut für Numerische und Angewandte Mathematik, Lotzestraße 16–18, D–37083 Göttingen, Germany (schaback@math.uni-goettingen.de, http://www.num.math.uni-goettingen.de/schaback).

The first ingredient requires the problem to be *continuously dependent on the data*. In quantitative form, this means that the inverse L^{-1} is a bounded linear map from F to U. In particular, we assume an inequality of the form

$$\|u\|_U \le C_a \|L(u)\|_F \text{ for all } u \in U \tag{1.2}$$

with a positive constant C_a describing the stability of the problem. In practical applications this will imply that the *solution space* U and the *data space* F have to be chosen with some care. In particular, U and F must often be chosen on a theoretical basis, e.g., as quite large spaces in which certain general existence results hold and which carry only rather weak norms. Usually, U will be a Sobolev space $W_2^\mu(\Omega)$ while F is a Cartesian product of Sobolev spaces which provide the right-hand sides for the differential equation and the boundary data via trace operators. Continuous dependence serves here as a replacement for more specific analytic assumptions like coercivity of a bilinear form. However, in section 9 we shall abandon the assumption of continuous dependence to be able to treat a certain class of ill-posed problems.

The second ingredient is some *additional regularity*. The actual solution u of a specific problem will often have more regularity than needed for the spaces U and F defining continuous dependence, and therefore we shall focus on a subspace $U_R \subseteq W_2^m(\Omega) \subset W_2^\mu(\Omega) =: U$ of U which we call the *regularity* subspace. The additional regularity of order $m - \mu > 0$ will be the driving force behind convergence rates, as we shall prove later.

The third ingredient is a scale of finite-dimensional *trial* subspaces U_r of U for a *trial discretization parameter* $r > 0$ which uses the additional regularity to provide a *convergent scheme for data approximation*. This is formalized by not necessarily linear maps $I_r : U_R \to U_r$ with error bounds

$$\|L(u - I_r(u))\|_F \le \epsilon_r(u) \text{ for all } u \in U_R. \tag{1.3}$$

It will be this approximation property that yields our convergence rates driven by additional regularity. Note that we do not use a single discretization parameter like the usual h here, because we need two different discretization parameters r and s for trial and test discretization. The trial spaces U_r can be chosen independent of the operator L, and we do not approximate the solution directly but rather the data via the linear operator L.

The fourth ingredient is a scale of *stable test discretizations* F_s of the data space F with respect to the scale of trial spaces U_r. This is formalized by a *test discretization parameter* $s > 0$ and linear maps $\Pi_s : F \to F_s$ into a scale of finite-dimensional *test data* spaces F_s such that the inequalities

$$\begin{aligned} \|L(u_r)\|_F &\le C(r,s)\|\Pi_s L(u_r)\|_{F_s} \text{ for all } u_r \in U_r, \\ c(s)\|\Pi_s L(u)\|_{F_s} &\le \|L(u)\|_F \text{ for all } u \in U \end{aligned} \tag{1.4}$$

hold. We call a specific choice of trial and test discretization schemes *uniformly stable* if both constants can be chosen independent of r and s for a certain range of these parameters. When restricted to the finite-dimensional trial spaces U_r, the inequalities express equivalence of discrete and nondiscrete norms on the data $L(u_r)$. The second inequality will be easily satisfied by discretization, but the first one will be hard, because it bounds a nondiscrete norm by a discrete norm, and this can work only for finite-dimensional spaces. It also implies uniqueness of solutions of the discretized finite systems $\Pi_s L(u_r) = \Pi_s L(u)$, which is a serious problem [7].

To simplify some of the later arguments, we outline here how we proceed to prove the first inequality of (1.4).

THEOREM 1.1. *Assume a* Poincaré–Friedrichs *inequality*

$$\|f\|_F \le c_1(s)\|f\|_{F_R} + c_2(s)\|\Pi_s f\|_{F_s} \text{ for all } f \in F_R \subset F \tag{1.5}$$

on a regularity subspace F_R *of the data space* F. *Second, assume a* Markov–Bernstein *inequality*

$$\|Lu_r\|_{F_R} \le c_3(r)\|Lu_r\|_F \text{ for all } u_r \in U_r \subset U_R \tag{1.6}$$

on a scale of trial spaces $U_r \subset U_R$ *with* $L(U_R) \subseteq F_R$. *Third, let the trial and test discretization parameters* r *and* s *satisfy the* stability criterion

$$c_1(s)c_3(r) < \frac{1}{2}. \tag{1.7}$$

Then the first inequality of (1.4) *holds with* $C(r,s) \le 2c_2(s)$.

Proof. Just consider

$$\begin{aligned} \|L(u_r)\|_F &\le c_1(s)\|L(u_r)\|_{F_R} + c_2(s)\|\Pi_s L(u_r)\|_{F_s} \\ &\le c_1(s)c_3(r)\|Lu_r\|_F + c_2(s)\|\Pi_s L(u_r)\|_{F_s} \\ &\le \tfrac{1}{2}\|Lu_r\|_F + c_2(s)\|\Pi_s L(u_r)\|_{F_s} \end{aligned}$$

for all $u_r \in U_R$. ∎

Note that in (1.5) we shall have $c_1(s) \to 0$ for $s \to 0$, because the inequality means that a function is small in a weak norm if it is bounded in a strong norm and is small on a large discrete set. In (1.6) we have to expect $c_3(r) \to \infty$ for $r \to 0$ because c_3 bounds a strong norm by a weak one on a finite-dimensional space. Thus the stability criterion (1.7) will usually be satisfied if the test discretization is "fine enough" with respect to the trial discretization.

The final ingredient is the class of *numerical methods*. We do not specify details in this overview, but we can always find nonunique trial functions $u^*_{r,s} \in U_r$ with

$$\|\Pi_s L(u - u^*_{r,s})\|_{F_s} \le \delta_{r,s} \tag{1.8}$$

solving $\Pi_s L(u_r) = \Pi_s L(u)$ approximately with a small tolerance. In fact, the approximation $I_r(u)$ is a solution if we have (1.4) and

$$c(s)\epsilon_r(u) \le \delta_{r,s}. \tag{1.9}$$

Note that we do not attempt to solve the discrete system $\Pi_s L(u_r) = \Pi_s L(u)$ exactly, because it will be overdetermined and unsolvable in general. However, under the assumption (1.9) we know that the relaxed problem (1.8) is solvable. In section 8 we show how to tackle such problems. Altogether, we replace strict collocation by a generalized form of "almost interpolation." Now we can formulate the core result of this paper.

THEOREM 1.2. *If the analytic problem is solvable by* $u \in U_R$ *and if we solve* (1.8) *by some* $u^*_{r,s} \in U_r$, *then the following error bound holds:*

$$\|u - u^*_{r,s}\|_U \le C_a \left(\epsilon_r(u) \left(1 + \frac{C(r,s)}{c(s)} \right) + c(s)\delta_{r,s} \right)$$

provided that all of the above ingredients are available. If the discretization is uniformly stable with constants

$$\begin{array}{rcll} & & C(r,s) & \leq & C, \\ 0 < c & \leq & c(s) & \leq & \tilde{c}, \end{array}$$

and if we choose $\tilde{c}\epsilon_r(u) \leq \delta_{r,s} \leq 2\tilde{c}\epsilon_r(u)$ *to satisfy* (1.9)*, we get the bound*

$$\|u - u^*_{r,s}\|_U \leq \epsilon_r(u) C_a \left(1 + \frac{C}{c} + 2\tilde{c}^2\right),$$

which behaves asymptotically like the trial approximation error $\epsilon_r(u)$*.*

Proof. The assertion follows from a simple chain of inequalities:

$$\begin{array}{rcl} \frac{1}{C_a}\|u - u^*_{r,s}\|_U & \leq & \|L(u - u^*_{r,s})\|_F \\ & \leq & \|L(u - I_r(u))\|_F + \|L(I_r(u) - u^*_{r,s})\|_F \\ & \leq & \epsilon_r(u) + c(s)\|\Pi_s L(I_r(u) - u^*_{r,s})\|_{F_s} \\ & \leq & \epsilon_r(u) + c(s)\|\Pi_s L(u - u^*_{r,s})\|_{F_s} \\ & & + c(s)\|\Pi_s L(I_r(u) - u)\|_{F_s} \\ & \leq & \epsilon_r(u) + c(s)\delta_{r,s} + \frac{C(r,s)}{c(s)}\|L(I_r(u) - u)\|_F \\ & \leq & \epsilon_r(u)\left(1 + \frac{C(r,s)}{c(s)}\right) + c(s)\delta_{r,s}. \qquad \square \end{array}$$

But now we shall have to show how this abstract machinery can be set to work. We shall finally derive specific convergence rates for unsymmetric collocation techniques solving strongly posed problems with continuous dependence in Sobolev spaces, including the Poisson problem with Dirichlet data as an illustration. But note that the above formalism is much more general, and there may be various other future ways to apply the framework, e.g., to unsymmetric methods solving problems in weak form.

The following sections will treat the above ingredients one by one, and then we shall patch the results together. Our key tools will be *nonstationary meshless kernel-based trial spaces* which allow approximation schemes with high-order convergence rates while maintaining stability if paired with sufficiently rich test discretizations. It turns out that the use of smooth kernels makes the final convergence order dependent mainly on the regularity of the solution and the problem. The numerical methods for solving (1.8) will consist of certain variations of the original unsymmetric collocation method, and we already have solvability via (1.9). This saves us from the degeneration problems of the standard unsymmetric collocation technique [7].

2. Well-posed problems and regularity. For example, consider a standard Poisson boundary value problem

$$\begin{array}{rcll} -\Delta u & = & f_\Omega & \text{in } \Omega, \\ u & = & f_D & \text{on } \partial\Omega \end{array} \tag{2.1}$$

on a bounded domain $\Omega \subset \mathbb{R}^d$ with Dirichlet data f_D on the piecewise smooth boundary $\partial\Omega$. In such problems, we consider the equations as being given in strong form; i.e., we assume the solution u to be regular enough to pose the equations pointwise as

$$\begin{array}{rcccll} (-\Delta u)(x) & = & (\delta_x \circ (-\Delta))(u) & = & f_\Omega(x) & \text{for all } x \in \Omega, \\ u(x) & = & (\delta_x \circ Id)(u) & = & f_D(x) & \text{for all } x \in \partial\Omega. \end{array} \tag{2.2}$$

This leads to (1.1) if we define $L(u) := (-\Delta u_{|\Omega}, u_{|\partial\Omega})$ on $U := W_2^\mu(\Omega)$ with values in $F := W_2^{\mu-2}(\Omega) \times W_2^{\mu-1/2}(\partial\Omega)$, with given data $f = (f_\Omega, f_D) \in F$.

But we allow much more general linear equations and boundary value operators. Formally, we follow the notation of [3] and others to combine differential and boundary operators into just one equation and write the latter as (1.1) where u is a function from some normed space U of functions. The mapping $L : U \to F$ maps solutions $u \in U$ to their data $L(u) \in F$, and the given problem consists in the inversion of L.

When aiming at methods with error bounds and convergence, we have to take a closer look at the given analytic problem (1.1). In particular, we shall assume that the problem (1.1) is well-posed in the sense that the solution u depends continuously on the data f of the right-hand side of (1.1). But we have to make this more precise. This can be done in various ways, e.g., by *total sets* of data functionals, but this is not quantitative. For later use we impose a norm $\|.\|_F$ on $F := L(U)$ in a suitable way and assume (1.2) with a positive "analytic" constant C_a which describes the norm of the linear map L^{-1} that takes the data $f \in F$ and maps them back to the solution u in the function space U. Clearly, for such a priori inequalities we must be careful with the choice of norms, because they depend on regularity theory, and they always imply that the homogeneous equation has only the trivial solution. The numerical methods following below will work on discretized versions F_s of F, and thus the proper choice of F will also have practical consequences.

So far we have not mentioned any specific numerical algorithm. But if any numerical method has produced an approximate solution $\tilde{u} \in U$ to the problem (1.1), one can calculate the data $\tilde{f} = L(\tilde{u}) \in F$ and the norm $\|\tilde{f} - f\|_F$ to get the a posteriori error bound

$$\|u - \tilde{u}\|_U \le C_a \cdot \|L(u - \tilde{u})\|_F = C_a \cdot \|f - \tilde{f}\|_F \tag{2.3}$$

for free, since the *residuals* $L(u - \tilde{u}) = f - \tilde{f}$ are explicitly known. It means that errors in the solution are bounded by the norm of the residuals, multiplied with the analytic constant. *Thus any numerical technique that produces approximate solutions of well-posed problems with small residuals will automatically guarantee small errors in the solution.* This trivial observation is well known in numerical analysis and serves as a basis for *defect correction* and *residual minimization* techniques, and is important for providing a safe a posteriori foundation for many unsafe and ad hoc numerical calculations published in science and engineering journals. If the underlying problem is continuously dependent on the data and if the naive user at least checks the residuals carefully, the calculations are on the safe side. But, unfortunately, there is no handbook listing all known inequalities of the form (1.2) for typical applications in science and engineering. In particular, it would be very useful to have proven upper bounds for the analytic constants.

Guided by regularity theory for elliptic problems, we focus on operator equations (1.1) where the linear map L splits into maps $L^1, \ldots, L^n$ with $L^i : U \to F^i$, $1 \le i \le n$, such that $F = F^1 \times \cdots \times F^n$ is the data space. We assume $U = W_2^\mu(\Omega)$ for some bounded domain $\Omega \subset \mathbb{R}^d$ and $F^i := W_2^{\mu-\mu_i}(\Omega^i)$, where μ_i is defined via a trace theorem by the order of the operator L^i and the dimension $d_i \le d$ of the partial domain $\Omega^i \subset \overline{\Omega} \subset \mathbb{R}^d$. The space F is then equipped with the sup of the norms of the spaces F^i. The regularity subspace U_R occurring later will then be a space $U_R \subseteq W_2^m(\Omega) \subset U := W_2^\mu(\Omega)$ for some $m \ge \mu$.

In the standard Poisson problem with Dirichlet data we take $L(u) = (-\Delta u, u_{\partial\Omega})$ mapping $U \subseteq W_2^\mu(\Omega)$ into $F = W_2^{\mu-2}(\Omega) \times W_2^{\mu-1/2}(\partial\Omega)$. This is a well-established

continuous dependence setting if the domain is smooth enough to make trace theorems and the regularity order μ valid. See, e.g., [3, 12] for early references which also allow distributional data and negative μ.

3. Approximation from trial spaces. The second ingredient of our framework is some additional regularity defined via a subspace U_R of the space U occurring in the continuous dependence bound (1.2). At this point, the regularity space can be quite general, but we also want an approximation property like (1.3) to hold. Thus we now have to consider our third ingredient, i.e., techniques that construct approximate solutions $\tilde{u}_r$ from a scale of *trial spaces* $U_r \subseteq U$ with a trial discretization parameter r. Note that this includes plenty of methods, with or without meshes, like finite elements, Petrov–Galerkin schemes, spectral methods, and all variations of collocation. It is trivial that the choice of the trial space should be such that the true solution u can be approximated easily by functions from the trial space. In case of solutions with singularities, like for Poisson problems on domains with incoming vertices, one should make sure that the trial space contains the expected singular functions.

One way to make this more precise is to assume that there is a mapping $A_r : U_R \to U_r$ with

$$\|u - A_r(u)\|_U =: \delta_r(u) \text{ for all } u \in U_R \tag{3.1}$$

with a certain error $\delta_r(u)$ which will depend on the regularity subspace U_R.

But the previous section teaches us that we do not need to approximate the exact solution u in the space U by functions $u_r \in U_r \subset U$ directly. It suffices to make sure that the residuals $L(u) - L(u_r)$ are small. Thus the crucial quantity is the residual error $\|L(u - u_r)\|_F$ for any $u \in U$ and an approximation $u_r \in U_r$. In contrast to the theory of finite elements, we do not consider optimal approximations of u by u_r here, nor do we attempt to minimize the above error with respect to u_r. Instead, we are satisfied if the trial space U_r contains for each function $u \in U$ an approximation $u_r := I_r(u)$ with small *residual error* $\epsilon_r(u)$ as in (1.3).

Of course, if an L-independent approximation operator A_r with (3.1) is available, one can take $I_r = A_r$ and assume $\epsilon_r(u) \le \|L\|\delta_r(u)$ because of

$$\epsilon_r(u) = \|L(u - I_r(u))\|_W \le \|L\|\|(u - A_r(u))\|_U \le \|L\|\delta_r(u).$$

But there may be better choices of I_r if L and the special structure of the residual space W are taken into account.

Inspection of (1.3) for F being a Cartesian product of Sobolev spaces reveals that the special approximation $I_r(u)$ should approximate u well *including its derivatives*, as far as they occur in the collection of data spaces F^i forming the space F. In fact, if we go back to our special case $U = W_2^\mu(\Omega)$ and $F = F^1 \times \cdots \times F^n$ with $F^i := L^i(U) := W_2^{\mu-\mu_i}(\Omega^i)$ with a regularity subspace $U_R \subseteq W_2^m(\Omega)$ and $m > \mu$, we should expect approximation bounds like

$$\begin{aligned} \|L^i(u) - L^i(I_r(u))\|_{W_2^{\mu-\mu_i}(\Omega^i)} &\le Cr^{m-\mu}\|L^i(u)\|_{W_2^{m-\mu_i}(\Omega^i)} \\ &\le Cr^{m-\mu}\|u\|_{W_2^m(\Omega)} \text{ for all } u \in W_2^m(\Omega), \end{aligned} \tag{3.2}$$

and this should work for a reasonable choice of $0 \le \mu \le m$ and with rates that just depend on the *regularity gap* $m - \mu$ and not on the order of the operators involved.

Note that the standard trial spaces of h-type finite element techniques consisting of piecewise linear functions fail to provide approximations of more than first-order

derivatives. In contrast to this, trial spaces generated by sufficiently smooth kernel functions can contain approximations to derivatives of any order, without any additional work needed. We shall explain this in the following sections.

In contrast to many engineering applications where a rather simple solution function is calculated via a huge finite element method system of millions of unknowns, we tend to argue in favor of small trial spaces designed to capture the essential features of the solution without taking the detour via a fine-grained space discretization. The consequence will be that the linear systems get unsymmetric, because any solution from a small *trial* space must be *tested* on a fine-grained space discretization, asking for many more degrees of freedom on the "test side" than on the "trial side." Unsymmetry of a method can be a feature instead of a bug. In what follows we shall investigate the relation of test and trial spaces more closely.

4. Kernel-based trial spaces. Now it is time to study maps I_r or A_r with good approximation properties for certain trial spaces U_r in the sense of (1.3) and (3.1). This is independent of PDE solving, and we shall see that nonstationary scales of meshless kernel-based trial spaces work perfectly.

DEFINITION 4.1. *A* kernel *is a function of the form* $K : \Omega \times \Omega \to \mathbb{R}$ *with* $\Omega \subseteq \mathbb{R}^d$. *It is* translation-invariant *if* $K(x,y) = \Phi(x-y)$ *with* $\Omega = \mathbb{R}^d$ *and* $\Phi : \mathbb{R}^d \to \mathbb{R}$. *It is* radial *if it is translation-invariant and of the form*

$$K(x,y) = \Phi(x-y) = \phi(\|x-y\|_2) \text{ with } \phi : [0,\infty) \to \mathbb{R} \text{ and } x,y \in \mathbb{R}^d.$$

Radial kernels are also called radial basis functions.

Note that radial basis functions ϕ can in principle be used in any space dimension, but certain properties of the associated translation-invariant kernel Φ on $\mathbb{R}^d$ may depend [6, 19] on the dimension d.

Kernels provide excellent tools in various disciplines, including approximation theory, partial differential equations, and machine learning [17]. The most important kernels are *reproducing kernels* of some Hilbert space which can be called the "native" Hilbert space for the kernel. Any Hilbert space H of functions on a domain Ω with continuous and linearly independent point evaluations has a kernel K with the reproduction property

$$f(x) = (f, K(x,\cdot)) \text{ for all } f \in H,\ x \in \Omega.$$

Conversely, any (strictly) positive definite [6, 19] and continuous kernel K on Ω is the reproducing kernel of a *native* Hilbert space N_K of continuous functions on Ω. We denote the norm on the native space N_K by $\|.\|_K$.

We focus here on *trial spaces* provided by kernels. Like in wavelet theory, the notions of *translation* and *dilation* play an important role. First, a general kernel $K : \Omega \times \Omega \to \mathbb{R}$ can be translated to points $y_1, \ldots, y_M \in \Omega$ called *centers* to provide trial functions $u_j(x) := K(x, y_j)$, $1 \le j \le M$, on Ω. In many cases, the set $Y := \{y_j : 1 \le j \le M\}$ of centers should fill a bounded domain Ω in such a way that the centers get dense when $M \to \infty$. This is expressed by the *fill distance*

$$h := h(Y,\Omega) := \sup_{x\in\Omega} \inf_{y\in Y} \|x-y\|_2$$

depending on Ω and the M centers in Y, which should converge to zero if M tends to infinity. The fill distance is the radius of the largest open ball with center in Ω that contains none of the centers y_j from Y. We use the notation h here, but later

we shall have two different fill distances for trial and test centers, and then we shall use r and s for clarity.

A *nonstationary* scale of kernel-based trial spaces can now be defined as

$$U_r := \text{span}\,\{K(\cdot, y_j) \,:\, 1 \le j \le M\} \text{ with } r := h(Y, \Omega), \tag{4.1}$$

where the dependence on the location and number of the centers is suppressed in the notation.

In the above *nonstationary* situation, only translations, but no dilations are used. The translated kernel is fixed and independent of the fill distance. There is no rescaling, if the fill distance gets small. This is in contrast to the *stationary* technique in standard and general finite elements [4]. There, the basis functions are rescaled when the fill distance changes, and in the translation-invariant kernel-based case this can be described by a scale of trial spaces

$$U_r := \text{span}\,\left\{\Phi\left(\frac{x - y_j}{r}\right) \,:\, 1 \le j \le M\right\},$$

where now the wavelet style interaction of translation and dilation is apparent.

The mathematics of the stationary and nonstationary cases are quite different. This often leads to misunderstandings. The stationary situation, as included in the meshless generalized finite element method [4], uses polynomial reproduction and the Bramble–Hilbert lemma. If centers are on a grid, it applies the Strang–Fix theory. Convergence orders are closely tied to polynomial reproduction properties, and the choice of kernels is quite restricted, because integrable kernels like the Gaussian do not yield convergent stationary approximations for $h \to 0$ [6]. Stability is usually much better than in the nonstationary case, but convergence rates (if they exist at all, e.g., for thin-plate splines or multiquadrics) are much smaller. We focus on nonstationary kernel-based trial spaces here, because condition problems can be overcome [5, 9], and we are heading for methods with high approximation orders.

We define a map $I_r \,:\, u \to u_r := I_r(u) \in U_r$ of (1.3) via *interpolation in the trial centers* by solving the system

$$u_r(y_k) := \sum_{j=1}^{M} \alpha_j K(y_j, y_k) = u(y_k),\ 1 \le k \le M, \tag{4.2}$$

for the coefficients $\alpha_1, \ldots, \alpha_M$ defining the function $u_r := I_r(u)$ in terms of the basis functions of the nonstationary trial space U_r of (4.1). This interpolation problem is solvable by definition if the kernel K is symmetric and positive definite [6, 19], because then the $M \times M$ matrix with entries $K(y_j, y_k)$ is symmetric and positive definite. Table 4.1 gives some examples. We ignore *conditionally* positive definite kernels here and refer to the literature [6, 16, 19] for details.

The interpolation system (4.2) makes sense for all functions u which have well-defined function values at the trial centers y_k. Thus the mapping I_r is at least defined on $C(\Omega)$, but for solutions u of PDE problems in strong form we use it on a regularity subspace U_R of $C(\Omega) \subset U$.

The book [19] contains a fairly complete account of interpolation error bounds in the nonstationary setting, while bounds for stationary and regular cases are in [6]. But in view of (1.3) and (3.2), we need very general error bounds in Sobolev spaces which are not covered in these books. Here (on the trial side) and later (on the test

TABLE 4.1
Radial basis functions $\phi(r)$, positive definite on $\mathbb{R}^d$.

Function	$\phi(r)$	Range	Smoothness β
Gaussian	$\exp(-r^2)$	$d \geq 1$	all β
inverse multiquadric	$(r^2+c^2)^\gamma,\ \gamma < -d/2,\ c > 0$	$d \geq 1$	all β
Sobolev for $W_2^k(\mathbb{R}^d)$	$r^{k-d/2}K_{k-d/2}(r),\ k > d/2$	$d \geq 1$	$\beta = 2k-d$
Wendland C^2 [18]	$(1-r)_+^4(1+4r)$	$d \leq 3$	3
	$(1-r)_+^5(1+5r)$	$d \leq 5$	3
Wendland C^4 [18]	$(1-r)_+^6(3+18r+35r^2)$	$d \leq 3$	5
	$(1-r)_+^7(1+7r+16r^2)$	$d \leq 5$	5

side for proving (1.5)) we use a general result from [13] which was extended in [20], while the range of admissible parameters was enlarged in [14].

THEOREM 4.2. *Suppose $\Omega \subset \mathbb{R}^d$ is a bounded domain with an interior cone condition. Choose $q \in [1,\infty]$ and constants*

$$0 \leq \mu < \mu + d/2 < \lfloor m \rfloor \tag{4.3}$$

with μ being an integer. Then there are positive constants C, h_0 such that

$$|u|_{W_q^\mu(\Omega)} \leq C\left(h^{m-\mu-d(1/2-1/q)_+}|u|_{W_2^m(\Omega)} + h^{-\mu}\|u\|_{\infty,Y_h}\right) \tag{4.4}$$

holds for every discrete set Y_h in Ω with fill distance at most $h \leq h_0$ and every $u \in W_2^m(\Omega)$.

This can be seen as a quantitative Poincaré–Friedrichs inequality for functions which are small on a finite subset, and it is independent of any trial space. If we replace seminorms by norms and extract $h^{-\mu}$ out of the right-hand side, the rest is a μ-independent norm on $W_2^m(\Omega)$ and we can apply interpolation theory to replace (4.4) by

$$\|u\|_{W_q^\mu(\Omega)} \leq C\left(h^{m-\mu-d(1/2-1/q)_+}\|u\|_{W_2^m(\Omega)} + h^{-\mu}\|u\|_{\infty,Y_h}\right) \tag{4.5}$$

under the assumptions (4.3) *without* the restriction of μ being an integer.

Now we take $h = r$ because we discretize the *trial* side, and we interpolate a function u on Y_r by $I_r(u)$ using points from Y_r as translations in (4.1) to get

$$\|u - I_r(u)\|_{W_2^\mu(\Omega)} \leq Cr^{m-\mu}\|u - I_r(u)\|_{W_2^m(\Omega)} \text{ for all } u \in W_2^m(\Omega). \tag{4.6}$$

Then we use the standard fact that the interpolant $I_r(u)$ solves the minimization problem

$$\|v\|_K \to \min,\ v \in K,\ v(y_j) = u(y_j) \text{ for all } y_j \in Y_r,$$

implying that $\|I_r(u)\|_K \leq \|u\|_K$ holds if we assume u to be in the native space N_K for the kernel K.

Therefore we strengthen the requirement on the regularization subspace U_R and on the regularity of our solution u to

$$u \in N_K = U_R \subseteq W_2^m(\Omega) \subseteq U \tag{4.7}$$

with bounded embeddings, where we always assume (4.3). This is easy if the kernel is smooth enough, and for the kernels in Table 4.1 the inequality

$$2m \leq \beta + d \tag{4.8}$$

is a sufficient condition.

We can now replace (4.6) by

$$\|u - I_r(u)\|_{W_2^\mu(\Omega)} \le C r^{m-\mu} \|u\|_K \text{ for all } u \in N_K$$

with a different constant. This inequality can be coupled with trace theorems for the operators $L^i : W_2^m(\Omega) \to W_2^{m-\mu_i}(\Omega^i)$ to get

$$\|L^i(u - I_r(u))\|_{W_2^{\mu-\mu_i}(\Omega^i)} \le C\|u - I_r(u)\|_{W_2^\mu(\Omega)} \le C r^{m-\mu} \|u\|_K \text{ for all } u \in N_K$$

which yields (3.2) in a slightly restricted form and our third ingredient (1.3) as

$$\|L(u - I_r(u))\|_F \le C r^{m-\mu} \|u\|_K =: \epsilon_r(u) \text{ for all } u \in U_R = N_K \tag{4.9}$$

under the assumptions (4.3), (4.7), and (4.8).

5. Stability of kernel-based test discretizations. We now consider the stability conditions (1.4), our fourth ingredient. We do this for meshless kernel-based trial spaces and for our running example generalizing the Poisson equation. The trial discretization via U_r and a set Y_r of centers is chosen as in section 4. We assume (4.7) and have the approximation result (4.9). On the test side, we use a set X_s of test centers which has a fill distance s on all of $\overline{\Omega}$. For all the operators L^i that arise in L, we will have a selection $X_s^i := X_s \cap \Omega^i$ of points with the same fill distance with respect to Ω^i, because we can assume that all Ω^i are subsets of $\overline{\Omega}$. The projectors Π_s^i on F^i just map functions from $F^i = W_2^{\mu-\mu_i}(\Omega^i)$ to their values on X_s^i. We thus have to assume Sobolev embedding conditions

$$2(\mu - \mu_i) > d_i := \dim(\Omega^i) \le d := \dim(\Omega),\ 1 \le i \le n. \tag{5.1}$$

The discretized spaces F_s^i will be $\mathbb{R}^{|X_s^i|}$ with the L_∞ norm, and we have

$$\Pi_s^i L^i(u) = (L^i(u))(X_s^i),\ \|\Pi_s^i L^i(u)\|_{F_s^i} = \|L^i(u)\|_{\infty, X_s^i}.$$

This implies by Sobolev embedding

$$\|\Pi_s^i L^i(u)\|_{F_s^i} = \|L^i(u)\|_{\infty, X_s^i} \le \|L^i(u)\|_{\infty,\Omega^i} \le C_i \|L^i(u)\|_{W_2^{\mu-\mu_i}(\Omega^i)},$$

where the constant is independent of u and s. We now assemble this into a discretization $F_s := F_s^1 \times \cdots \times F_s^n$ with $\Pi_s := \Pi_s^1 \times \cdots \times \Pi_s^n$ of $F = F^1 \times \cdots \times F^n$ and take the sup norm of the components. Then we have

$$\begin{aligned} \|\Pi_s L(u)\|_{F_s} &= \sup_{1\le i\le n} \|\Pi_s^i L^i(u)\|_{F_s^i} \\ &\le C \sup_{1\le i\le n} \|L^i(u)\|_{W_2^{\mu-\mu_i}(\Omega^i)} \\ &= C\|L(u)\|_F \end{aligned}$$

and get the second inequality of (1.4) with a constant that is independent of s and dependent only on Sobolev embedding. This leaves us to prove the first inequality of (1.4) via Theorem 1.1.

Fortunately, the inequality (4.5) holds for general Sobolev spaces, and we can apply it on the test side for different operators. We get

$$\|L^i(u)\|_{W_2^{\mu-\mu_i}(\Omega^i)} \le C\left(s^{m-\mu}\|L^i(u)\|_{W_2^{m-\mu_i}(\Omega^i)} + s^{-(\mu-\mu_i)}\|L^i(u)\|_{\infty, X_s^i}\right)$$

under the assumptions (5.1) and

$$d_i/2 < \mu - \mu_i < \mu - \mu_i + d_i/2 < \lfloor m - \mu_i \rfloor \text{ for all } i,\ 1 \le i \le n, \tag{5.2}$$

which pose no serious problems if m is large enough. Now we define F_R to be the range of $L(U_R)$ in the Cartesian product of all spaces $W_2^{m-\mu_i}(\Omega^i)$, taking the sup of the component norms. With part of the notation

$$\min_i \mu_i =: \underline{\mu} \le \mu_i \le \overline{\mu} := \max_i \mu_i, \tag{5.3}$$

this yields (1.5) in the form

$$\|L(u)\|_F \quad \le \quad C\left(s^{m-\mu}\|L(u)\|_{F_R} + s^{\underline{\mu}-\mu}\|\Pi_s L(u)\|_{F_s}\right).$$

We now want to consider (1.6). Assume K to be a translation-invariant positive definite kernel of finite smoothness which is Fourier-transformable in $\mathbb{R}^d$ with an exact decay

$$c(1+\|\omega\|_2)^{-\beta-d} \le \hat{K}(\omega) \le C(1+\|\omega\|_2)^{-\beta-d} \text{ for all } \omega \in \mathbb{R}^d, \tag{5.4}$$

where the constants β can be read from Table 4.1. If we again assume (4.8), we can cite the Bernstein-type inequality

$$\|u_r\|_{W_2^m(\Omega)} \le C \cdot \|u_r\|_{W_2^{(d+\beta)/2}(\mathbb{R}^d)} \le Cr^{-(d+\beta)/2}\|u_r\|_{L_\infty(\Omega)} \text{ for all } u_r \in U_r$$

from [15] provided that the trial centers in $Y_r \subset \Omega$ are not too wildly scattered in the sense that the minimal separation distance $q(Y_r)$ is uniformly bounded below by the fill distance $h(Y_r, \Omega)$ via

$$q(Y_r) := \min_{y_j \ne y_i \in Y_r} \|y_j - y_i\|_2 \ge C \sup_{y \in \Omega} \min_{y_i \in Y_r} \|y - y_i\|_2 =: h(Y_r, \Omega) \tag{5.5}$$

such that both quantities behave asymptotically like the trial discretization parameter r. This yields

$$\begin{aligned}
\|L(u_r)\|_{F_R} &= \max_i \|L_i(u_r)\|_{W_2^{m-\mu_i}(\Omega^i)} \\
&\le C\|u_r\|_{W_2^m(\Omega)} \\
&\le Cr^{-(d+\beta)/2}\|u_r\|_{L_\infty(\Omega)} \\
&\le Cr^{-(d+\beta)/2}\|u_r\|_{W_2^\mu(\Omega)} \\
&= Cr^{-(d+\beta)/2}\|u_r\|_U \\
&\le C_a C r^{-(d+\beta)/2}\|L(u_r)\|_F \text{ for all } i,\ 1 \le i \le n,
\end{aligned}$$

under the additional assumption $\mu > d/2$. This result is far from optimal, but it establishes (1.6) and allows us to apply Theorem 1.1 under a stability condition of the form

$$Cs^{m-\mu}r^{-(d+\beta)/2} < \frac{1}{2}. \tag{5.6}$$

This finally implies the following theorem.

THEOREM 5.1. *Assume that the trial kernel K with* (5.4) *is smooth enough to satisfy* (4.8). *Let the trial space consist of quasi-uniform translates on Ω with discretization parameter r, and consider a test discretization on $\overline{\Omega}$ with fill distance s. Then, under the notation* (5.3), *and the additional conditions* (4.3), (5.2), (5.6), *and $\mu > d/2$, the first inequality of* (1.4) *is satisfied with*

$$\|L(u_r)\|_F \quad \le \quad Cs^{\underline{\mu}-\mu}\|\Pi_s L(u_r)\|_{F_s}$$

for all $u_r \in U_r$.

6. Strong convergence in Sobolev spaces. We now assemble what we have in case of our running example with continuous dependence in Sobolev norms. In contrast to the introduction, we proceed here from the user's point of view.

We start with the analytic problem. Consider an operator equation $L(u) = f$ as in (1.1) whose solution u is continuously dependent on the data f. We assume continuous dependence in the sense of (1.2) to hold if we pick spaces $U = W_2^\mu(\Omega)$ and F defined as a Cartesian product of Sobolev trace spaces F^i as in section 2. But note that users have to make sure in each application problem that the a priori inequalities composing (1.2) are actually satisfied. If several choices of μ are possible, the user should know that the final convergence will take place in $U = W_2^\mu(\Omega)$, but large μ have to be paid for by regularity. If convergence of higher-order derivatives is of importance, a sufficiently large μ must be chosen. Since we solve problems in strong form via evaluation of residuals, we have to pick μ large enough to let all data have continuous point evaluations. This is expressed by the requirement (5.1). At this point, the lower bounds for μ will rule out problems with low regularity. Such problems should be tackled with methods using weak data functionals and involving integration. We plan to deal with such methods in the future, in particular with the unsymmetric meshless Petrov–Galerkin method of Atluri and his collaborators [1].

The next step concerns regularity. We assume that the solution should have at least a $U_R \subseteq W_2^m(\Omega)$ regularity with some $m > \mu$. By standard arguments from approximation theory, the difference between m and μ is the driving force for the possible convergence rates. The user has to decide which m is adequate. Larger m will improve the convergence rates, but they may not be justified by the smoothness of the problem.

Then we pick a kernel K which is smooth enough to have its native space N_K contained in $W_2^m(\Omega)$. In view of Table 4.1, this requires (5.4) and (4.8). The solution u must have at least the regularity of $W_2^m(\Omega)$, because it should be in $U_R = N_K \subseteq W_2^m(\Omega)$. The excess regularity of N_K over $W_2^m(\Omega)$ does not pay off later, and thus it is a good idea to stay with a kernel satisfying $\beta + d = 2m$ to have norm equivalence between $W_2^m(\Omega)$ and N_K. Note that the compactly supported radial polynomial kernels of Wendland [18] satisfy this for certain choices of m, β, and d.

Now it is time to pick a meshless trial discretization U_r via a set Y_r of trial centers with fill distance r using the kernel K. Then we can expect an error behavior $\epsilon_r(u) \leq Cr^{m-\mu}$ for $u \in N_K$ for a direct interpolant to the regular solution in the points of Y_r. This rate is the ideal goal we want to achieve for our numerical solution of the given operator equation.

The next step is to pick a test discretization via a set X_s of test centers in $\overline{\Omega}$ with fill distance s. The second inequality of (1.4) holds with $c(s)$ independent of s because we assume continuous residuals and corresponding Sobolev embedding theorems. By Theorem 5.1, the first inequality of (1.4) will then hold with $C(r,s) \leq Cs^{\mu-\mu}$ using (5.3). But we have to make the test discretization fine enough to satisfy (5.6). As expected, this means that the test discretizations must be somewhat finer than the trial discretizations, and the required relation between s and r is

$$s < c \cdot r^{1+\frac{\mu}{m-\mu}}. \tag{6.1}$$

There is plenty of leeway for small trial and large test spaces.

We are now ready to put everything into Theorem 1.2, while we assume that we solve the discretized problem (1.8) with accuracy $\delta_{r,s}$. With new generic constants

we get

$$\|u - u^*_{r,s}\|_{W_2^\mu(\Omega)} \leq C\left(r^{m-\mu}\left(1 + Cs^{\underline{\mu}-\mu}\right) + \delta_{r,s}\right)\|u\|_K. \tag{6.2}$$

Note that this bound has the proper approximation error of order $r^{m-\mu}$ holding between Sobolev spaces $U = W_2^\mu(\Omega)$ and $U_R = N_K \subseteq W_2^m(\Omega)$, but there also is a counteracting term $s^{\underline{\mu}-\mu}$ which is the price we have to pay for working on discrete residuals in the L_∞ norm while bounding the residual error in the norm on Sobolev trace spaces $W_2^{\mu-\mu_i}(\Omega^i)$ in order to use continuous dependence on Sobolev space data. If we choose $\delta_{r,s}$ properly via (1.9) and s via (5.6), we have solvability of the system and an error bound

$$\|u - u^*_{r,s}\|_{W_2^\mu(\Omega)} \leq Cr^{m-2\mu+\underline{\mu}-\frac{\mu(\mu-\underline{\mu})}{m-\mu}}\|u\|_{W_2^m(\Omega)}.$$

Of course, this is not an optimal bound because μ must be positive and even larger than $d/2$. Consequently, there is quite some future work necessary on this bound, though it is improving when m is much larger than μ. In this context, it is not surprising that most of the practical applications of unsymmetric collocation methods have very regular solutions.

To show the minimum regularity requirements for the results of this section, we should track the possible range of m for the Poisson problem in d dimensions. For operators $L^1 := -\Delta$ and L^2 providing Dirichlet boundary data, we have $\mu_1 = 2$ and $\mu_2 = 1/2$ with $d_1 = d$ and $d_2 = d - 1$. Then (5.2) requires $d/2 < \mu - 2$ while (4.3) leads to $\lfloor m \rfloor > 2 + d$ as the minimum regularity requirement. This clearly needs improvement by future work, but it should be mentioned that the resulting error bound is strong enough to include derivatives up to order μ with $2 + d/2 < \mu < \lfloor m \rfloor - d/2 \geq m - 1 - d/2$.

7. Weak convergence in Sobolev spaces. Analysis of the previous section shows that the term $s^{\underline{\mu}-\mu}$ in (6.2) with some positive μ satisfying (5.1) and $\mu > d/2$ makes the final bound worse than expected. Tracing this back to (4.4) shows that one should better look at another variation which allows $\mu = 0$ at that point without spoiling the assumption that the data are still continuous. In fact, (4.4) also allows

$$\|u\|_{L_\infty(\Omega)} \leq C\left(h^{m-d/2}\|u\|_{W_2^m(\Omega)} + \|u\|_{\infty,Y_h}\right) \text{ for all } u \in W_2^m(\Omega) \tag{7.1}$$

if $d/2 < \lfloor m \rfloor$ holds.

But this does not easily fit into the framework required for continuous dependence. Thus we start anew, defining the data spaces F^i as spaces $C(\Omega^i)$ of continuous functions under the L_∞ norm. To make continuous dependence valid, we use embeddings $C(\Omega^i) \subset L_2(\Omega^i) = W_2^0(\Omega^i) \subseteq W_2^{\mu-\mu_i}(\Omega^i)$ for $\mu := \underline{\mu} = \min \mu_i$. Then we apply the standard continuous dependence relating the standard solution space $W_2^\mu(\Omega)$ to the trace spaces $L^i(W_2^\mu(\Omega)) \subseteq W_2^{\mu-\mu_i}(\Omega^i)$, which fortunately hold for small and even negative μ, if the domain is smooth [12, 3]. This yields a new continuous dependence relation via

$$\begin{aligned}
\|u\|_{W_2^\mu(\Omega)} &\leq C \max_i \|L^i(u)\|_{W_2^{\mu-\mu_i}(\Omega^i)} \\
&\leq C \max_i \|L^i(u)\|_{W_2^0(\Omega^i)} \\
&\leq C \max_i \|L^i(u)\|_{C(\Omega^i)} \\
&= C \max_i \|L^i(u)\|_{F^i} \\
&= C\|L(u)\|_F,
\end{aligned}$$

holding only on the subspace U of functions u in $W_2^\mu(\Omega)$ with continuous data $L(u)$. Note that this will lead to a weak convergence result in $U \subset W_2^\mu(\Omega)$, though the problem formulation is still strong. For instance, a problem with Dirichlet data will lead to $\mu = 1/2$ due to the trace map $W_2^\mu(\Omega) \to W_2^{\mu-1/2}(\partial\Omega)$ if all other trace or differential operators have a larger loss in the order of the respective Sobolev trace spaces.

Thus we now repeat our basic argument for $U \subset W_2^\mu(\Omega)$ with $\mu = \underline{\mu} = \min_i \mu_i$. Our choice of regularity space U_R and the kernel K will be as above. This fixes β and m. To derive the approximation order in (1.3) we apply (4.5) to get

$$\begin{aligned}\|L^i(u - I_r(u))\|_{W_\infty^0(\Omega^i)} &\le Cr^{m-\mu_i-d/2}\|L^i(u - I_r(u))\|_{W_2^{m-\mu_i}(\Omega^i)} \\ &\le Cr^{m-\mu_i-d/2}\|u - I_r(u)\|_{W_2^m(\Omega)} \\ &\le Cr^{m-\overline{\mu}-d/2}\|u\|_{W_2^m(\Omega)},\end{aligned}$$

which requires only

$$d/2 < \lfloor m - \mu_i \rfloor,\ 1 \le i \le n, \tag{7.2}$$

and where we now also need $\overline{\mu}$ from (5.3). Thus we get

$$\epsilon_r(u) \le Cr^{m-\overline{\mu}-d/2}$$

for (1.3).

The discretization of the $F^i = C(\Omega^i)$ spaces is again by pointwise evaluation on a set X_s^i of test centers, taking the discrete L_∞ norm, but we now can skip Sobolev embedding which led to the inequalities (5.1) we want to avoid now. Since every data space is equipped with the L_∞ norm, we have $c(s) = 1$ in the second inequality of (1.4). The proof of the first inequality of (1.4) again proceeds via Theorem 1.1. To prove (1.5) we start with (7.1) on the various data:

$$\|L^i(u)\|_{W_\infty^0(\Omega^i)} \le C\left(s^{m-\mu_i-d_i/2}\|L^i(u)\|_{W_2^{m-\mu_i}(\Omega^i)} + \|L^i(u)\|_{\infty,X_s^i}\right)$$

for all $u \in U_R = N_K$ where we need (7.2) again. We define F_R as in the previous section, and then we have (1.5) in the form

$$\|L(u)\|_F \le C\left(s^{m-\overline{\mu}-d/2}\|L(u)\|_{F_R} + \|\Pi_s L(u)\|_{F_s}\right).$$

Unfortunately, the proof for (1.6) given in section 5 proceeds via $\|.\|_\infty$ and thus needs $\mu > d/2$. Of course there is always some a priori inequality of the form (1.6), but we currently have no explicit upper bounds for $c_3(r)$ in terms of r. Anyway, if we take s small enough to satisfy (1.7) with $c_1(s) = Cs^{m-\overline{\mu}-d/2}$, Theorem 1.1 still is valid and yields the first inequality of (1.4) with $C(r, s)$ independent of r and s provided that s is small enough.

Then we continue as in the proof of Theorem 5.1. Since the discretization scheme is uniformly stable for sufficiently small test discretizations s, Theorem 1.2 now gives the error bound

$$\|u - u_{r,s}^*\|_{W_2^\mu(\Omega)} \le Cr^{m-\overline{\mu}-d/2}\|u\|_{W_2^m(\Omega)} \text{ for all } u \in W_2^m(\Omega)$$

provided that the numerical solution of (1.8) observes (1.9) and the test discretization is fine enough in a way we currently cannot specify explicitly. The left-hand norm

is rather weak here, and the approximation order can probably be improved. For instance, a standard two-dimensional Poisson problem with Dirichlet data would lead to

$$\|u - u^*_{r,s}\|_{W_2^{1/2}(\Omega)} \le C r^{m-3} \|u\|_{W_2^m(\Omega)} \text{ for all } u \in W_2^m(\Omega),$$

but an optimal rate for the Sobolev spaces involved would be $m-1/2$ instead of $m-3$. The minimum regularity in this case is $m = 4$ because of (7.2).

Future work should improve the results of this and the previous sections. This may be done by better choices of spaces and norms, plus better versions of the Markov–Bernstein inequality (1.6) which are currently investigated.

8. Numerical methods. We now look at techniques to solve the discrete problem (1.8). It amounts to solving the n linear problems

$$\Pi_s^i L^i(u - u^*_{r,s}) = 0,\ 1 \le i \le n,$$

approximately, where we discretized the operators L^i on the domains Ω^i by taking only point evaluations. This takes the form of collocation

$$L^i(u)(x_{ji}) = L^i(u^*_{r,s})(x_{ji}) = 0,\ 1 \le i \le n,\ 1 \le j \le N_i,$$

where the points of the test discretization X_s are the union of the sets

$$X_s^i := \{x_{j1}, \dots, x_{jN_i}\},\ 1 \le j \le n,$$

and where we dropped the dependence on s in the notation for the x_{ji} and for N_i. For a shorthand notation, we introduce the functionals

$$\lambda_{ji} \ :\ v \mapsto L^i(v)(x_{ji})$$

and rearrange them into a single-indexed list $\lambda_1, \dots, \lambda_N$ with $N = N_1 + \cdots + N_n$.

Since $u^*_{r,s}$ should be in the trial space U_r generated by translations of the kernel K at trial centers forming $Y_r := \{y_1, \dots, y_M\}$, we arrive at a system

$$\sum_{m=1}^{M} \alpha_m \lambda_i^z K(z, y_m) = \lambda_i^z u(z),\ 1 \le i \le N,$$

with M unknowns and N equations. In case $M = N$ this is exactly the unsymmetric collocation technique dating back to Kansa in 1986 [8]. It has no rigid foundation yet, and it can fail in specially constructed situations [7], though it works fine in many applications. In the first years it was applied to small problems with smooth solutions due to serious condition problems, but recently there have been results on preconditioning [5, 9] that allow a wider range of applicability.

In view of Theorem 5.1 and the two previous sections we know that $N \ge M$ holds and the system has full rank M, provided that our stability conditions are satisfied, calling for a somewhat finer discretization on the test than on the trial side. Thus the system will be unsymmetric and overdetermined, but at least there is no rank loss. Furthermore, we know by (1.3) and (1.9) that there is a good approximate solution to the full system. This means that we can allow any numerical method that produces a solution with similar or less deviation.

Since our convergence analysis worked with the L_∞ norm on the discretized F_s spaces, a first choice would be to go for a best L_∞ approximation of the right-hand side. This means solving a linear optimization problem which minimizes η under the constraints

$$-\eta \le \sum_{m=1}^{M} \alpha_m \lambda_i^z K(z, y_m) - \lambda_i^z u(z) \le \eta, \ 1 \le i \le N, \tag{8.1}$$

where $\alpha_1, \ldots, \alpha_M$ are the other variables. If the revised simplex method is applied to the dual problem, each step has an $\mathcal{O}(M^2)$ complexity. The Kuhn–Tucker conditions ensure that one can work with at most $M+1$ active test conditions at each time. This makes the number $N >> M$ of test centers much less relevant than M, and for nicely chosen low-dimensional trial spaces one can get away with rather small computational complexity, as demonstrated in [10, 11].

But one can also try all other techniques that somehow provide a function $u_{r,s}^* \in U_r$ which by a posteriori inspection leads to a small residual norm $\delta_{r,s}$ in (1.8). This can happen to the original Kansa method when executed on a subset of M test points, or by adaptive bootstrapping techniques like the one in [10, 11] which picks suitable test centers and trial centers one by one. Other alternatives are to use pivoting with row exchange or to go for a least-squares solution first. Anyway, if the resulting residual norm $\delta_{r,s}$ is small, the result of Theorem 1.2 is still valid, proving that one actually has a good approximation to the real solution.

As an aside, we note that a simpler theory is possible if we optimize over a nondiscrete residual norm on F. Section 5 will then be obsolete, but one has to solve semi-infinite optimization problems (if F carries a sup-norm) or apply least-squares methods with integrations (if F carries an inner product). Another strategy to avoid stability problems is to add the numerically accessible quadratic constraint $\|u_r\|_K^2 \le C$ to any method trying to make residuals small. This regularization trick has connections to machine learning [17] and should be investigated in future work.

9. Ill-posed problems. For ill-posed problems, continuous dependence fails, but our method and its analysis will still be useful. We assume that the problem still has the form (1.1), but we now assume that the "true solution" $u \in U$ satisfies only

$$L(u) = f + \rho \in F, \tag{9.1}$$

where F contains the available data f and a small residual ρ. The problem $L(u) = f$ may be unsolvable, and (1.2) is not available. We consider a function $\tilde{u} \in U$ to be acceptable as a "solution" if

$$\|L(\tilde{u}) - L(u)\|_F = \|L(\tilde{u}) - f - \rho\|_F$$

is not much larger than $\|\rho\|_F$. We still assume (1.3) and (1.4), but we have to replace (1.8) by

$$\|\Pi_s(f - L(u_{r,s}^*))\|_{F_s} \le \delta_{r,s}, \tag{9.2}$$

because $L(u)$ now is unknown and does not coincide with f. Furthermore, solvability of the above system now requires

$$c(s)(\|\rho\|_F + \epsilon_r(u)) \le \delta_{r,s} \tag{9.3}$$

instead of (1.9) as a sufficient condition. The proof technique of Theorem 1.2 then still implies the following theorem.

THEOREM 9.1. *If the analytic problem is ill-posed, but solvable by $u \in U_R$ in the sense of* (9.1)*, and if we solve* (9.2) *by some $u^*_{r,s} \in U_r$, then there is a bound*

$$\|L(u - u^*_{r,s})\|_U \le c(s)\|\rho\|_F + \left(\epsilon_r(u)\left(1 + \frac{C(r,s)}{c(s)}\right) + c(s)\delta_{r,s}\right).$$

If the discretization is uniformly stable, then there is a choice of $\delta_{r,s}$ via (9.3) *such that the above residual error behaves asymptotically like the trial approximation error $\epsilon_r(u)$ plus $\|\rho\|_F$.*

Proof. We modify the proof of Theorem 1.2 to get

$$\begin{aligned}
\|L(u - u^*_{r,s})\|_F &\le \|L(u - I_r(u))\|_F + \|L(I_r(u) - u^*_{r,s})\|_F \\
&\le \epsilon_r(u) + c(s)\|\Pi_s L(I_r(u) - u^*_{r,s})\|_{F_s} \\
&\le \epsilon_r(u) + c(s)\|\Pi_s L(I_r(u) - u)\|_{F_s} \\
&\quad + c(s)\|\Pi_s(L(u) - f)\|_{F_s} \\
&\quad + c(s)\|\Pi_s(f - L(u^*_{r,s}))\|_{F_s} \\
&\le \epsilon_r(u) + c(s)\delta_{r,s} + \tfrac{C(r,s)}{c(s)}\|L(I_r(u) - u)\|_F + c(s)\|\Pi_s\|\|\rho\|_F \\
&\le c(s)\|\rho\|_F + \epsilon_r(u)\left(1 + \tfrac{C(r,s)}{c(s)}\right) + c(s)\delta_{r,s}. \qquad \square
\end{aligned}$$

For simplicity of the above presentation, we have replaced the second inequality of (1.4) by

$$c(s)\|\Pi_s g\|_{F_s} \le \|g\|_F \text{ for all } g \in F$$

which is no serious complication. However, we should comment on what happens with Theorems 1.1 and 5.1 if we have no analytic constant C_a for carrying out the proof. We replace C_a by the constant $C_a(r)$ arising in a finite-dimensional version

$$\|u_r\|_U \le C_a(r)\|L(u_r)\|_F \text{ for all } u_r \in U_r$$

of (1.2). This is feasible due to norm equivalence, but we leave it to future research to derive upper bounds for $C_a(r)$.

10. Conclusions. We provided convergence proofs for a generalized nonsquare version of Kansa's collocation method, showing that the convergence rates are determined by approximation results for nonstationary meshless kernel-based trial spaces. The rates improve with the smoothness of the solution, the domain, the differential operator, and the kernel. They hold for large classes of analytic problems, provided that there is continuous dependence on the data, and they result from a fairly general framework that possibly has applications to other unsymmetric methods. On the downside, the results still need improvement by proving better a priori inequalities to plug into the framework.

There are many possibilities for enhancement and extension of these results:

1. Find sufficient conditions for nonsingularity of square Kansa-type collocation matrices.
2. Introduce discretization-dependent weights for different parts of residuals into the theory of this paper in order to align dimension- and order-dependent convergence rates.

3. For important problems of applied analysis, state the continuous dependence of the solution on the data in precise form and derive upper bounds for the analytic constants.
4. Implement algorithms of this paper as local components of a global algorithm using localization features like domain decomposition or partitions of unity and efficiency-enhancing features like preconditioning and iterative solvers.
5. For such a global algorithm, perform large-scale numerical experiments and compare observed convergence rates with the theoretical ones of this paper.
6. Generalize all of this to unsymmetric methods for weak problems like the meshless local Petrov–Galerkin (MLPG) method of Atluri and collaborators [2].

Acknowledgments. The author thanks I. Babuška, C. S. Chen, Y. C. Hon, E. Kansa, L. Ling, G. Lube, and H. Wendland for several very stimulating discussions.

REFERENCES

[1] S. N. Atluri and T. L. Zhu, *A new meshless local Petrov–Galerkin (MLPG) approach in computational mechanics*, Comput. Mech., 22 (1998), pp. 117–127.
[2] S. N. Atluri and T. L. Zhu, *A new meshless local Petrov–Galerkin (MLPG) approach to nonlinear problems in computer modeling and simulation*, Computer Modeling and Simulation in Engineering, 3 (1998), pp. 187–196.
[3] J. P. Aubin, *Approximation of Elliptic Boundary-Value Problems*, Pure Appl. Math., 26, Wiley-Interscience, New York, 1972.
[4] I. Babuška, U. Banerjee, and J. Osborn, *Survey of meshless and generalized finite element methods: A unified approach*, in Acta Numerica 2003, Acta Numer. 12, Cambridge University Press, Cambridge, UK, 2003, pp. 1–215.
[5] D. Brown, L. Ling, E. Kansa, and J. Levesley, *On approximate cardinal preconditioning methods for solving PDEs with radial basis functions*, Engineering Analysis with Boundary Elements, 29 (2005), pp. 343–353.
[6] M. D. Buhmann, *Radial Basis Functions*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2003.
[7] Y. C. Hon and R. Schaback, *On unsymmetric collocation by radial basis functions*, Appl. Math. Comp., 119 (2001), pp. 177–186.
[8] E. J. Kansa, *Application of Hardy's multiquadric interpolation to hydrodynamics*, in Proceedings of the 1986 Annual Simulations Conference, Vol. 4, San Diego, CA, 1986, pp. 111–117.
[9] L. Ling and E. Kansa, *Preconditioning for radial basis functions with domain decomposition methods*, Math. Comput. Modelling, 40 (2005), pp. 1413–1427.
[10] L. L. Ling, R. Opfer, and R. Schaback, *Results on meshless collocation techniques*, Engineering Analysis with Boundary Elements, 30 (2006), pp. 247–253.
[11] L. Ling and R. Schaback, *On adaptive unsymmetric meshless collocation*, in Proceedings of the 2004 Annual International Conference on Computational and Experimental Engineering and Sciences, Advances in Computational and Experimental Engineering and Sciences, S. N. Atluri and A. J. B. Tadeu, eds., Tech Science Press, Forsyth, GA, 2004, CD-Rom edition, paper # 270.
[12] J. L. Lions and E. Magenes, *Problèmes aux limites non homogènes at applications,* Travaux et Recherches Mathématiques 1, Dunod, Paris, 1968.
[13] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Math. Comp., 74 (2005), pp. 743–763.
[14] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions*, Constr. Approx., 24 (2006), pp. 175–186.
[15] R. Schaback and H. Wendland, *Inverse and saturation theorems for radial basis function interpolation*, Math. Comp., 71 (1998), pp. 669–681.
[16] R. Schaback and H. Wendland, *Characterization and construction of radial basis functions*, in Multivariate Approximation and Applications, N. Dyn, D. Leviatan, D. Levin, and A. Pinkus, eds., Cambridge University Press, Cambridge, UK, 2001, pp. 1–24.

[17] R. Schaback and H. Wendland, *Kernel techniques: From machine learning to meshless methods*, in Acta Numerica 2006, Acta Numer. 15, Cambridge University Press, Cambridge, UK, 2006, pp. 543–639.

[18] H. Wendland, *Piecewise polynomial, positive definite, and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396.

[19] H. Wendland, *Scattered Data Approximation*, Cambridge Monographs on Applied and Computational Mathematics, Cambridge University Press, Cambridge, UK, 2005.

[20] H. Wendland and C. Rieger, *Approximate interpolation with applications to selecting smoothing parameters*, Numer. Math., 101 (2005), pp. 729–748.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 352–370

HIGH RESOLUTION SCHEMES FOR A HIERARCHICAL SIZE-STRUCTURED MODEL*

JUN SHEN†, CHI-WANG SHU‡, AND MENGPING ZHANG†

Abstract. In this paper we discuss two explicit finite difference schemes, namely a first order upwind scheme and a second order high resolution scheme, for solving a hierarchical size-structured population model with nonlinear growth, mortality, and reproduction rates. We prove stability and convergence for both schemes and provide numerical examples to demonstrate their capability in solving smooth and discontinuous solutions.

Key words. hierarchical size-structured population model, upwind scheme, high resolution scheme, stability, convergence

AMS subject classifications. 65M06, 65M12, 92-08

DOI. 10.1137/050638126

1. Introduction.

In this paper we develop stable and convergent finite difference schemes for a hierarchical size-structured population model given by the equations

$$
\begin{aligned}
&u_t + \left(g(x, Q(x,t))\, u\right)_x + m(x, Q(x,t))\, u = 0, \qquad (x,t) \in (0,L] \times (0,T], \\
&g(0, Q(0,t))u(0,t) = C(t) + \int_0^L \beta(x, Q(x,t))\, u(x,t)dx, \qquad t \in (0,T], \\
&u(x,0) = u^0(x), \qquad x \in [0,L],
\end{aligned} \tag{1.1}
$$

where $u(x,t)$ is the density of individuals having size x at time t, and the nonlocal term $Q(x,t)$ is defined by

$$
Q(x,t) = \alpha \int_0^x w(\xi)u(\xi,t)d\xi + \int_x^L w(\xi)u(\xi,t)d\xi, \qquad 0 \le \alpha < 1, \tag{1.2}
$$

for some given function w. $Q(x,t)$ depends on the density u in a global way and is usually referred to as the *environment*.

A special feature of (1.1) is the boundary condition at size $x = 0$, which involves the function g representing the growth rate of an individual, and a global dependency on the density $u(x,t)$ for all $x \in (0,L]$. The function m in (1.1) represents the mortality rate of an individual. The function β in the boundary condition of (1.1) represents the reproduction rate of an individual, and the function C represents the inflow rate of zero-size individual from an external source. We assume that the functions g, m,

*Received by the editors August 13, 2005; accepted for publication (in revised form) August 23, 2006; published electronically February 9, 2007.
http://www.siam.org/journals/sinum/45-1/63812.html

†Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, P. R. China (jshen3@mail.ustc.edu.cn, mpzhang@ustc.edu.cn). The research of the third author was partially supported by the Chinese Academy of Sciences grant 2004-1-8.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (shu@dam.brown.edu). The research of this author was partially supported by the Chinese Academy of Sciences while the author was in residence at the University of Science and Technology of China (grant 2004-1-8) and at the Institute of Computational Mathematics and Scientific / Engineering Computing. Additional support was provided by ARO grant W911NF-04-1-0291, NSF grant DMS-0510345, and AFOSR grant FA9550-05-1-0123.

and β are functions of both the size x and the environment Q, which in turn depends globally on the density u; hence the problem is highly nonlinear.

The hierarchical structured population model (1.1) describes population dynamics in which the size of an individual determines its access to resources and hence to its growth or decay. This dependency is based on the environment which is a global function of the density for all sizes. The hierarchy is determined by the size; for example, in a population of animals, very often the size of an individual determines what species its prey can be and by what species it can be eaten, or in a population of forest trees, the taller the tree, the higher the availability of light it has. In (1.2), since $\alpha < 1$, we observe that the environment for size x has a larger weight for the density of those larger than x than for the density of those smaller than x, representing a particular instance of the size hierarchy. We refer to, e.g., [3] for a more detailed discussion of the background and application of the hierarchical size-structured population models.

Hierarchical structured population models have been studied in the literature in, e.g., [2, 3, 5, 6, 11, 13, 19], usually with more restrictive assumptions on the functions g, β, and m. For example, in [3], the model (1.1) was considered for the special situation $g = g(Q)$, $\beta = \beta(Q)$, $m = m(Q)$, and $C = 0$. In [2], the model (1.1) was studied with the functions g and β depending linearly on the size x, m independent of x, and $C = 0$. In [13], (1.1) was investigated with $\alpha = 0$. The model (1.1) with the complete generality as stated above was studied in [1], in which an implicit first order finite difference scheme was analyzed and its stability and convergence, as well as the existence, uniqueness, and well-posedness (in the L^1 norm) of bounded variation weak solutions for (1.1) were proved. However, the scheme in [1] is not very practical for actual numerical simulation, because it is implicit and only first order accurate.

In this paper we develop and analyze two explicit finite difference schemes, namely, a first order upwind scheme and a second order high resolution scheme, for solving (1.1). We prove stability and convergence for both schemes. Many aspects of our proof are based on the techniques in [1, 4, 8, 15], but it is not a routine generalization because of the complication due to the explicit time marching, second order accuracy, and global constraints in the equation. We also provide numerical examples to demonstrate the capability of these schemes in solving smooth and discontinuous solutions. We remark that discontinuous solutions for (1.1) are generic (see, for example, the numerical example in section 4), unless the boundary condition (e.g., the inflow rate C of a zero-size individual from an external source) happens to be compatible with the initial condition.

As in [1], we make the following assumptions on the model functions:

- (H1) $g(x, Q)$ is twice continuously differentiable with respect to x and Q, $g(x, Q) > 0$ for $x \in [0, L)$, $g(L, Q) = 0$, and $g_Q(x, Q) \le 0$.
- (H2) $m(x, Q)$ is nonnegative continuously differentiable with respect to x and Q.
- (H3) $\beta(x, Q)$ is nonnegative continuously differentiable with respect to x and Q. Furthermore, there is a constant $\omega_1 > 0$ such that $\sup_{(x,Q)\in[0,L]\times[0,\infty)} \beta(x, Q) \le \omega_1$.
- (H4) $w(x)$ is nonnegative continuously differentiable.
- (H5) $C(t)$ is nonnegative continuously differentiable.
- (H6) $u^0 \in BV[0, L]$ and $u^0(x) \ge 0$.

In section 2, we present an explicit, first order upwind scheme for solving (1.1) and state its stability and convergence. To save space we omit most details of the proof in this section, since we will provide a similar but technically more complicated proof in the following section for the second order scheme. In section 3, we present an

explicit, second order high resolution scheme for solving (1.1) and prove its stability and convergence. Section 4 contains numerical examples demonstrating the capability of these two numerical schemes. Concluding remarks are given in section 5.

2. A first order upwind finite difference scheme. First, we briefly describe the standard notation to be used in this paper. We assume the spatial domain $[0, L]$ is divided into N cells with cell boundary points denoted by x_j for $0 \le j \le N$, $x_0 = 0$, and $x_N = L$. For simplicity of presentation we will assume that the mesh is uniform of size Δx, namely, $x_j = j\Delta x$. This assumption is not essential for the analysis or the numerical computation; more general meshes can be easily considered. We also denote the time step by Δt. In fact, this time step $\Delta t = \Delta t^n = t^{n+1} - t^n$ could change from one step to the next step, based on stability conditions, but we use the same notation Δt without the superscript n since we will consider only one-step time discretizations (forward Euler or Runge–Kutta). We shall denote by u_j^n and Q_j^n the finite difference approximations of $u(x_j, t^n)$ and $Q(x_j, t^n)$, respectively. We also denote

$$g_j^n = g(x_j, Q_j^n), \;\; \beta_j^n = \beta(x_j, Q_j^n), \;\; m_j^n = m(x_j, Q_j^n), \;\; w_j = w(x_j), \;\; C^n = C(t^n).$$

We define the standard finite difference operators

$$D^-(u_j^n) = \frac{u_j^n - u_{j-1}^n}{\Delta x}, \qquad \Delta_+(u_j^n) = u_{j+1}^n - u_j^n, \qquad \Delta_-(u_j^n) = u_j^n - u_{j-1}^n,$$

and we define the standard discrete L^1 and L^∞ norms and TV seminorm of the grid function u^n by

$$\|u^n\|_1 = \sum_{j=1}^{N} |u_j^n| \Delta x, \qquad \|u^n\|_\infty = \max_{0 \le j \le N} |u_j^n|, \qquad TV(u^n) = \sum_{j=0}^{N-1} |u_{j+1}^n - u_j^n|.$$

The explicit, first order upwind finite difference scheme for (1.1) that we consider in this section is defined by

$$\frac{u_j^{n+1} - u_j^n}{\Delta t} + \frac{g_j^n u_j^n - g_{j-1}^n u_{j-1}^n}{\Delta x} + m_j^n u_j^n = 0, \qquad 1 \le j \le N, \tag{2.1}$$

with the left boundary condition implemented by

$$g_0^n u_0^n = C^n + \sum_{j=1}^{N} \beta_j^n u_j^n \Delta x, \tag{2.2}$$

the environment is computed by

$$Q_j^n = \alpha \sum_{i=1}^{j} w_i u_i^n \Delta x + \sum_{i=j+1}^{N} w_i u_i^n \Delta x, \tag{2.3}$$

and the initial condition is taken as

$$u_j^0 = u^0(x_j), \qquad j = 1, 2, \ldots, N.$$

We denote $\lambda = \frac{\Delta t}{\Delta x}$, and rewrite the scheme (2.1) as

$$u_j^{n+1} = (1 - \lambda g_j^n - \Delta t m_j^n) u_j^n + \lambda g_{j-1}^n u_{j-1}^n, \quad j \ge 1. \tag{2.4}$$

Since we consider only one-step explicit schemes, the right side of (2.4) contains only terms at time level t^n. Hence sometimes we will omit the superscript n when it does not cause confusion.

We first state the L^1 boundedness of the numerical solution u^n for $t^n \le T$, under the assumption that $u_j^n \ge 0$. We will prove the validity of this assumption later.

PROPOSITION 2.1. *If $u_j^n \ge 0$, then $\|u^n\|_1$ is bounded when $t^n \le T$.*

We omit the proof of this proposition since it is similar to and simpler than the proof of Proposition 3.1 in the next section. The L^1 bound is estimated as

$$\|u^n\|_1 \le e^{\omega_1 T}\|u^0\|_1 + \frac{Ce^{\omega_1 T}}{\omega_1} \equiv \omega_2,$$

where ω_1 is the upper bound of $\beta(x,Q)$ given in assumption H3, and the constant ω_2, as well as a sequence of such constants ω_k to be defined later, depend only on the given functions g, m, C, β, and w, the final time T, and the initial condition u^0.

With this L^1 bound on the density u_j^n, we can easily obtain the following upper bound for the environment Q:

$$\begin{aligned} |Q_j^n| &= \left| \alpha \sum_{i=1}^{j} w_i u_i^n \Delta x + \sum_{i=j+1}^{N} w_i u_i^n \Delta x \right| \\ &\le \|w\|_\infty \max_n \|u^n\|_1 \le \omega_2 \|w\|_\infty \equiv Q_{\max}. \end{aligned}$$

We now have a bounded closed domain $\mathcal{D} = \{(x,Q) \in [0,L] \times [0,Q_{max}]\}$ that x and Q reside in; hence by the smoothness assumptions of g, m, β, and w, we have a fixed constant ω_3 such that

$$\sup_{\mathcal{D}} |f(x,Q)| \le \omega_3, \qquad \sup_{0\le x\le L} |h(x)| \le \omega_3$$

for

$$\begin{aligned} f(x,Q) = g(x,Q),\, &g_x(x,Q),\, g_Q(x,Q),\, g_{xx}(x,Q),\, g_{xQ}(x,Q),\, g_{QQ}(x,Q), \\ &m(x,Q),\, m_x(x,Q),\, m_Q(x,Q),\, \beta(x,Q), \\ h(x) = w(x),\, &w'(x). \end{aligned}$$

Thus when $\Delta t \le \Delta t_0 \equiv \frac{1}{2\omega_3}$ and $\lambda \le \lambda_0 \equiv \frac{1}{2\omega_3}$, we have

$$1 - \lambda g_j^n - \Delta t m_j^n \ge 0, \qquad 1 \le j \le N. \tag{2.5}$$

This clearly implies $u_j^n \ge 0$ by (2.4). Notice that we can choose $\lambda = \lambda_0$ as either a constant or a variable depending on the time level t^n. We have thus verified the assumption made in Proposition 2.1 about the nonnegativity of u_j^n.

Next we will state the L^∞ boundedness of the numerical solution.

PROPOSITION 2.2. *$\|u^n\|_\infty$ is bounded for $t^n \le T$.*

We again omit the details of the proof of this proposition since it is similar to and simpler than the proof of Proposition 3.2 in the next section. We point out only that, since g is continuous and $g(0,Q) > 0$ by assumption (H1), we can take

$$\mu = \min_{Q\in[0,Q_{max}]} g(0,Q) > 0. \tag{2.6}$$

We would then have

$$|u_0^n| \le \frac{C+\omega_1\omega_2}{\mu} \tag{2.7}$$

and, for $j \ge 1$,

$$\begin{aligned} |u_j^n| &\le \|u^{n-1}\|_\infty + \sup_{\mathcal{D}} |g_x(x,Q)|\, \|u^{n-1}\|_\infty \Delta t \\ &\le (1+\omega_3\Delta t)\|u^{n-1}\|_\infty. \end{aligned}$$

This, together with (2.7), provides the L^∞ bound as

$$\|u^n\|_\infty \le \max\left\{ e^{\omega_3 T}\|u^0\|_\infty, \frac{1}{\mu}(C+\omega_1\omega_2)\right\} \equiv \omega_4.$$

In order to prove the total variation stability of the scheme, we would need the following results.

LEMMA 2.3. *There exist positive constants ω_5, ω_6, and ω_7 such that*

$$\begin{aligned} &\max_{1\le j\le N} |Q_j^n - Q_{j-1}^n| \le \omega_5\Delta x, \qquad \max_{1\le j\le N} |g_j^n - g_{j-1}^n| \le \omega_5\Delta x, \\ &\max_{1\le j\le N} |m_j^n - m_{j-1}^n| \le \omega_5\Delta x \end{aligned} \tag{2.8}$$

for $1 \le j \le N$,

$$|g_{j+1}^n - 2g_j^n + g_{j-1}^n| \le \omega_6\left(\Delta x^2 + \Delta x|u_{j+1}^n - u_j^n|\right) \tag{2.9}$$

for $1 \le j \le N-1$, and

$$\max_{1\le j\le N} |Q_j^{n+1} - Q_j^n| \le \omega_7\Delta t\left(1+TV(u^n)\right), \quad \max_{1\le j\le N} |g_j^{n+1} - g_j^n| \le \omega_7\Delta t\left(1+TV(u^n)\right),$$

$$\max_{1\le j\le N} |\beta_j^{n+1} - \beta_j^n| \le \omega_7\Delta t\left(1+TV(u^n)\right) \tag{2.10}$$

for $0 \le j \le N$.

We omit the details of the proof of this lemma since it is similar to and simpler than the proof of Lemma 3.3 in the next section. We point out only that ω_5 in (2.8) is given by

$$\omega_5 = \max\left(\omega_3\omega_4,\ \omega_3(1+\omega_3\omega_4)\right),$$

ω_6 in (2.9) is given by

$$\omega_6 = \max\left(\omega_3(2+\omega_5+2\omega_3\omega_4+2\omega_3\omega_4\omega_5),\ \omega_3^2\right),$$

and ω_7 in (2.10) is given by

$$\omega_7 = \max(\omega_3^2,\ \omega_3^3,\ \omega_3\omega_4(\omega_3+\omega_5)L,\ \omega_3^2\omega_4(\omega_3+\omega_5)L).$$

We are now ready to state the total variation stability of the scheme.

PROPOSITION 2.4. *$TV(u^n)$ is bounded for $t^n \le T$.*

We once again omit the details of the proof of this proposition since it is similar to and simpler than the proof of Proposition 3.4 in the next section. We point out only that with

$$\omega_8 = \max\left((\omega_3+\omega_6)\omega_4 L,\ \omega_3+\omega_5+\omega_4\omega_6\right),$$

$$\omega_9 = \omega_8 + \omega_4(\omega_3 + \omega_5),$$

$$\omega_{10} = \max\left(\frac{\omega_4\omega_7}{\mu}(1+L) + \frac{\omega_3^2}{\mu},\ \frac{\omega_4\omega_7}{\mu}(1+L) + \frac{\omega_3}{\mu} + \frac{\omega_3\omega_4}{\mu}(\omega_3+\omega_5)L\right),$$

and

$$\omega_{11} = \omega_9 + \omega_{10},$$

we have

$$TV(u^{n+1}) \le (1 + \omega_{11}\Delta t)TV(u^n) + \omega_{11}\Delta t,$$

which implies the boundedness of $TV(u^n)$ for $t^n \le T$.

Next, we show the Lipschitz stability in t.

PROPOSITION 2.5. *There exists a positive constant M such that for any $q > p$, we have*

$$\sum_{j=1}^{N}\left|\frac{u_j^q - u_j^p}{\Delta t}\right|\Delta x \le M(q-p).$$

Proof. Using (2.1) and Lemma 2.3, we have

$$\begin{aligned}\sum_{j=1}^{N}\left|\frac{u_j^{n+1} - u_j^n}{\Delta t}\right|\Delta x &= \sum_{j=1}^{N}\left|D^-(g_j^n u_j^n) + m_j^n u_j^n\right|\Delta x \\ &= \sum_{j=1}^{N}\left|\left(\frac{g_j^n - g_{j-1}^n}{\Delta x} + m_j^n\right)u_j^n + g_{j-1}^n D^-(u_j^n)\right|\Delta x \\ &\le \omega_4\omega_5 L + \omega_3\omega_4 L + \omega_3 TV(u^n) \le M.\end{aligned}$$

Thus,

$$\sum_{j=1}^{N}\left|\frac{u_j^q - u_j^p}{\Delta t}\right|\Delta x \le \sum_{n=p}^{q-1}\sum_{j=1}^{N}\left|\frac{u_j^{n+1} - u_j^n}{\Delta t}\right|\Delta x \le M(q-p). \qquad \square$$

Following [18] we can define a family of functions $\{U_{\Delta x,\Delta t}\}$ by

$$U_{\Delta x,\Delta t}(x,t) = u_j^n$$

for $x \in [x_{j-1}, x_j)$, $t \in [t^{n-1}, t^n)$, $j = 1,\dots,N$, and $n = 1,\dots,l$. Then, the set of functions $\{U_{\Delta x,\Delta t}\}$ is compact in the topology of $\mathcal{L}^1((0,L)\times(0,T))$, and we have the following result of convergence.

PROPOSITION 2.6. *Under the time step restriction for the validity of previous propositions in this section, there exists a subsequence $\{U_{\Delta x_i,\Delta t_i}\} \subset \{U_{\Delta x,\Delta t}\}$ which converges to a $BV([0,L]\times[0,T])$ function $u(x,t)$ in the sense that*

$$\int_0^L |U_{\Delta x_i,\Delta t_i}(x,0) - u^0(x)|dx \to 0$$

and

$$\int_0^T\int_0^L |U_{\Delta x_i,\Delta t_i}(x,t) - u(x,t)|dxdt \to 0$$

as $i \to \infty$. Furthermore, the function u, satisfying

$$\|u\|_{BV([0,L]\times[0,T])} \le E(\|u^0\|_{BV[0,L]}, \|C\|_{\mathcal{C}^1[0,T]}),$$

is the unique $BV([0,L]\times[0,T])$ solution $u(x,t)$ for (1.1), and the numerical solution $\{U_{\Delta x,\Delta t}\}$ converges to it when $\Delta x \to 0$.

Proof. The convergence of a subsequence to a BV function $u(x,t)$ and the fact that $u(x,t)$ is a BV weak solution of (1.1) follow from Propositions 2.1, 2.2, 2.4, and 2.5 and [18]. The uniqueness of bounded variation weak solutions of (1.1) is proved in [1]. Using this uniqueness we easily deduce the convergence of the numerical solution $\{U_{\Delta x,\Delta t}\}$ toward $u(x,t)$ when $\Delta x \to 0$. ∎

3. A second order high resolution finite difference scheme. The first order scheme defined in the previous section is very diffusive and would need many grid points to achieve acceptable resolution. In this section we develop and analyze a second order high resolution finite difference scheme for (1.1), following the minmod based MUSCL schemes [8, 14]. We remark, however, that the analysis is significantly more complicated because of the global constraints in (1.1). We note that our scheme can be easily generalized to the more accurate generalized MUSCL-type scheme similar to the one in [15] and the total variation bounded modified minmod based scheme in [16] without affecting the analysis. The second order high resolution finite difference scheme that we consider in this section is defined by

$$\frac{u_j^{n+1}-u_j^n}{\Delta t} + \frac{\hat{f}_{j+1/2}^n - \hat{f}_{j-1/2}^n}{\Delta x} + m_j^n u_j^n = 0, \quad 1 \le j \le N, \tag{3.1}$$

where the numerical flux $\hat{f}_{j+1/2}^n$ is defined by

$$\hat{f}_{j+1/2}^n = \begin{cases} g_j^n u_j^n + \frac{1}{2}(g_{j+1}^n - g_j^n)u_j^n + \frac{1}{2} g_j^n mm(\Delta_+ u_j^n, \Delta_- u_j^n) & : \quad j = 2, \ldots, N-2, \\ g_j^n u_j^n & : \quad j = 0, 1, N-1, N, \end{cases}$$

where the minmod function mm is defined by [8]

$$mm(a,b) = \frac{\text{sign}(a) + \text{sign}(b)}{2} \min(|a|, |b|). \tag{3.2}$$

Clearly, this scheme is second order accurate except at the boundary, where it is first order accurate. This guarantees second order accuracy in the global L^1 norm. The global boundary condition at the left is implemented by a second order composite trapezoid rule

$$g_0^n u_0^n = C^n + {\sum_{j=0}^{N}}' \beta_j^n u_j^n \Delta x, \tag{3.3}$$

where the special summation notation is defined by

$${\sum_{j=j_1}^{j_2}}' a_j = \frac{1}{2} a_{j_1} + \frac{1}{2} a_{j_2} + \sum_{j=j_1+1}^{j_2-1} a_j$$

if $j_2 - j_1 \ge 1$, and of course

$${\sum_{j=j_1}^{j_2}}' a_j = 0 \qquad \text{if } \ j_2 \le j_1.$$

The environment is computed also by a second order composite trapezoid rule, except for the integral over the first interval which is computed by the right-ended rectangular rule to avoid using u_0^n. That is,

$$Q_0^n = \omega_1 u_1^n \Delta x + \sum_{i=1}^{N}{}' \, w_i u_i^n \Delta x, \qquad Q_1^n = \alpha\omega_1 u_1^n \Delta x + \sum_{i=1}^{N}{}' \, w_i u_i^n \Delta x,$$

$$Q_j^n = \alpha\omega_1 u_1^n \Delta x + \alpha \sum_{i=1}^{j}{}' \, w_i u_i^n \Delta x + \sum_{i=j}^{N}{}' \, w_i u_i^n \Delta x, \qquad 2 \le j \le N. \tag{3.4}$$

Notice that this approximation to Q_j^n is second order accurate. The initial condition is still taken as

$$u_j^0 = u^0(x_j), \qquad j = 1, 2, \ldots, N.$$

Still using the notation $\lambda = \frac{\Delta t}{\Delta x}$, we can write the scheme (3.1) as

$$u_j^{n+1} = u_j^n - \lambda(\hat{f}_{j+1/2}^n - \hat{f}_{j-1/2}^n) - \Delta t m_j^n u_j^n, \quad j \ge 1. \tag{3.5}$$

We denote

$$A_j^n = \begin{cases} \frac{1}{2}\left(g_{j+1}^n + g_j^n + g_j^n \frac{mm(\Delta_+ u_j^n, \Delta_- u_j^n)}{\Delta_- u_j^n} - g_{j-1}^n \frac{mm(\Delta_- u_j^n, \Delta_- u_{j-1}^n)}{\Delta_- u_j^n}\right) : j = 3, \ldots, N-2, \\ \frac{1}{2}\left(g_{j+1}^n + g_j^n + g_j^n \frac{mm(\Delta_+ u_j^n, \Delta_- u_j^n)}{\Delta_- u_j^n}\right) : j = 2, \\ \frac{1}{2}\left(2g_j^n - g_{j-1}^n \frac{mm(\Delta_- u_j^n, \Delta_- u_{j-1}^n)}{\Delta_- u_j^n}\right) : j = N-1, \\ g_j^n : j = 1, N, \end{cases}$$

$$B_j^n = \begin{cases} \frac{1}{2}(\Delta_+ g_j^n + \Delta_- g_j^n) & : \quad j = 3, \ldots, N-2, \\ \frac{1}{2}\Delta_+ g_j^n & : \quad j = 2, \\ \frac{1}{2}\Delta_- g_j^n & : \quad j = N-1, \\ \Delta_- g_j^n & : \quad j = 1, N, \end{cases}$$

and rewrite the scheme (3.5) as

$$u_j^{n+1} = (1 - \lambda A_j^n - m_j^n \Delta t) u_j^n + \lambda(A_j^n - B_j^n) u_{j-1}^n, \quad j \ge 1. \tag{3.6}$$

We first prove the L^1 boundedness of the numerical solution u^n for $t^n \le T$, under the assumption that $u_j^n \ge 0$. We will prove the validity of this assumption later.

PROPOSITION 3.1. *If $u_j^n \ge 0$, then $\|u^n\|_1$ is bounded when $t^n \le T$.*

Proof. As before, since $u_j^n \ge 0$, $m_j^n \ge 0$, and $g_N^n = 0$, we have

$$\begin{aligned} \frac{\|u^{n+1}\|_1 - \|u^n\|_1}{\Delta t} &= \sum_{j=1}^{N} \frac{u_j^{n+1} - u_j^n}{\Delta t} \Delta x \\ &= -\sum_{j=1}^{N} (\hat{f}_{j+1/2}^n - \hat{f}_{j-1/2}^n) - \sum_{j=1}^{N} m_j^n u_j^n \Delta x \\ &\le -\sum_{j=1}^{N} (\hat{f}_{j+1/2}^n - \hat{f}_{j-1/2}^n) \end{aligned}$$

$$
\begin{aligned}
&= g_0^n u_0^n \\
&= C^n + \frac{1}{2}\beta_0^n u_0^n \Delta x + \frac{1}{2}\beta_N^n u_N^n \Delta x + \sum_{j=1}^{N-1} \beta_j^n u_j^n \Delta x \\
&\le C^n + \frac{1}{2}\beta_0^n u_0^n \Delta x + \omega_1 \|u^n\|_1.
\end{aligned}
$$

We now assume, for the time being, that $u_0^k \le \theta$, where θ is a constant. This assumption will be justified later. If $\Delta x \le 2C/\omega_1\theta$, where again C denotes the upper bound of $C(t)$ for $t \in [0,T]$ and ω_1 is the upper bound of $\beta(x,Q)$, then we have $\frac{1}{2}\beta_0^k u_0^k \Delta x \le C$. For constant Δt, we then immediately have

$$
\begin{aligned}
\|u^n\|_1 &\le (1+\omega_1\Delta t)\|u^{n-1}\|_1 + 2C\Delta t \\
&\le (1+\omega_1\Delta t)^n \|u^0\|_1 + \sum_{j=0}^{n-1} (1+\omega_1\Delta t)^j 2C\,\Delta t \\
&\le e^{\omega_1 T}\|u^0\|_1 + \frac{2Ce^{\omega_1 T}}{\omega_1} \equiv M_2,
\end{aligned}
$$

where the constant M_2, as well as a sequence of such constants M_k to be defined later, depend only on the given functions g, m, C, β, and w, the final time T, and the initial condition u^0. This proof is clearly also valid, with a minor modification, for the situation when Δt is not a constant.

We now look at the bound of Q_j^n. For $0 \le j \le N$, we have, by the definition of Q_j^n in (3.4), that

$$
Q_j^n \le \omega_1 u_1^n \Delta x + \sum_{i=1}^{N}{}' \, w_i u_i^n \Delta x \le \frac{3}{2}\omega_3 M_2 \equiv Q_{\max},
$$

and therefore

$$
g_0^n \ge \min_{0 \le Q \le Q_{\max}} g(0,Q) \equiv \mu > 0. \tag{3.7}
$$

Thus if $\Delta x \le \mu/\omega_1$, we have $g_0^n - \frac{1}{2}\beta_0 \Delta x \ge \mu/2$; hence from (3.3) we deduce

$$
u_0^n \le \frac{2}{\mu}(\omega_1 M_2 + C). \tag{3.8}
$$

The constants on the right-hand side of the inequality above do not depend on θ; hence the assumption on the boundedness of u_0^k is justified. □

As before, we now have a bounded closed domain $\mathcal{D} = \{(x,Q) \in [0,L] \times [0, Q_{\max}]\}$ that x and Q reside in; hence by the smoothness assumptions of g, m, β, and w, we have a fixed constant M_3 such that

$$
\sup_{\mathcal{D}} |f(x,Q)| \le M_3, \qquad \sup_{0 \le x \le L} |h(x)| \le M_3, \qquad \sup_{0 \le t \le T} |\eta(t)| \le M_3
$$

for

$$
\begin{aligned}
f(x,Q) = {} & g(x,Q),\, g_x(x,Q),\, g_Q(x,Q),\, g_{xx}(x,Q),\, g_{xQ}(x,Q),\, g_{QQ}(x,Q), \\
& m(x,Q),\, m_x(x,Q),\, m_Q(x,Q),\, \beta(x,Q), \\
h(x) = {} & w(x),\, w'(x), \qquad \eta(t) = C(t),\, C'(t).
\end{aligned}
$$

It can be easily shown that

$$|A_j^n| \le \frac{3}{2}\max_{\mathcal{D}}|g(x,Q)| \le \frac{3}{2}M_3;$$

thus when $\Delta t \le \Delta t_0 \equiv \frac{1}{2M_3}$ and $\lambda \le \lambda_0 \equiv \frac{1}{3M_3}$, we have

$$1 - \lambda A_j^n - \Delta t m_j^n \ge 0, \qquad 1 \le j \le N. \tag{3.9}$$

Notice that

$$2(A_j^n - B_j^n)$$

$$= \begin{cases} g_j^n\left(1 + \frac{mm(\Delta_+ u_j^n, \Delta_- u_j^n)}{\Delta_- u_j^n}\right) + g_{j-1}^n\left(1 - \frac{mm(\Delta_- u_j^n, \Delta_- u_{j-1}^n)}{\Delta_- u_j^n}\right) & : \quad j = 3, \ldots, N-2, \\ g_j^n\left(2 + \frac{mm(\Delta_+ u_j^n, \Delta_- u_j^n)}{\Delta_- u_j^n}\right) & : \quad j = 2, \\ g_j^n + g_{j-1}^n\left(1 - \frac{mm(\Delta_- u_j^n, \Delta_- u_{j-1}^n)}{\Delta_- u_j^n}\right) & : \quad j = N-1, \\ 2g_{j-1}^n & : \quad j = 1, N, \end{cases}$$

which implies, by the definition of the minmod function (3.2), that

$$A_j^n - B_j^n \ge 0, \qquad 1 \le j \le N. \tag{3.10}$$

This, together with (3.9), clearly implies $u_j^n \ge 0$ by (3.6). Notice that we can choose $\lambda = \lambda_0$ as either a constant or a variable depending on the time level t^n. We have thus verified the assumption made in Proposition 3.1 about the nonnegativity of u_j^n.

Next we will prove the L^∞ boundedness of the numerical solution.

PROPOSITION 3.2. *$\|u^n\|_\infty$ is bounded for $t^n \le T$.*

Proof. First, we have already shown the boundedness of u_0^n in (3.8). As for $j \ge 1$, we use (3.6), (3.9), (3.10), and the nonnegativity of m to obtain

$$\begin{aligned} |u_j^n| &\le (1 - \lambda A_j^{n-1} - \Delta t m_j^{n-1})\|u^{n-1}\|_\infty + \lambda(A_j^{n-1} - B_j^{n-1})\|u^{n-1}\|_\infty \\ &\le \|u^{n-1}\|_\infty - \lambda B_j^{n-1}\|u^{n-1}\|_\infty. \end{aligned}$$

We can easily verify that, for $2 \le j \le N$,

$$Q_j - Q_{j-1} = \frac{1}{2}(\alpha - 1)(w_j u_j + w_{j-1}u_{j-1})\Delta x. \tag{3.11}$$

For $j = 1$, we have a similar formula:

$$Q_1 - Q_0 = (\alpha - 1)\omega_1 u_1. \tag{3.12}$$

Therefore, we have

$$\begin{aligned} g_j^{n-1} - g_{j-1}^{n-1} &= g(x_j, Q_j^{n-1}) - g(x_{j-1}, Q_j^{n-1}) + g(x_{j-1}, Q_j^{n-1}) - g(x_{j-1}, Q_{j-1}^{n-1}) \\ &= g_x(\hat{x}_j, Q_j^{n-1})\Delta x + g_Q(x_{j-1}, \hat{Q}_j^{n-1})(Q_j^{n-1} - Q_{j-1}^{n-1}). \end{aligned}$$

Here and below $\hat{z}_j$ denotes a value between z_{j-1} and z_j for $z = x$ or $z = Q$. By assumption, $\alpha < 1$, $g_Q(x,Q) \le 0$. We clearly have, for $2 \le j \le N$,

$$\begin{aligned} -g_Q(x_{j-1}, \hat{Q}_j^{n-1})(Q_j^{n-1} - Q_{j-1}^{n-1}) &= -\frac{1}{2}g_Q(x_{j-1}, \hat{Q}_j^{n-1})(\alpha - 1)(w_j u_j^{n-1} + w_{j-1}u_{j-1}^{n-1})\Delta x \\ &\le 0, \end{aligned}$$

and also

$$-g_Q(x_0,\hat{Q}_1^{n-1})(Q_1^{n-1}-Q_0^{n-1}) = -g_Q(x_0,\hat{Q}_1^{n-1})(\alpha-1)\omega_1 u_1^{n-1}\Delta x \le 0.$$

Now, noticing that $-B_j \le \max_i(g_{i-1}-g_i)$ for $j \ge 1$, we obtain immediately, for $j \ge 1$,

$$|u_j^n| \le \|u^{n-1}\|_\infty + \sup_{\mathcal{D}} |g_x(x,Q)|\,\|u^{n-1}\|_\infty \Delta t \le (1+M_3\Delta t)\|u^{n-1}\|_\infty.$$

This, together with (3.8), clearly implies

$$\|u^n\|_\infty \le \max\left\{e^{M_3 T}\|u^0\|_\infty,\ \frac{2}{\mu}(\omega_1 M_2 + C)\right\} \equiv M_4. \qquad \square$$

Before proving the total variation stability of the scheme, we would need to prove the following results.

LEMMA 3.3. *There exist positive constants M_5, M_6, and M_7 such that*

$$\begin{aligned} &\max_{1\le j\le N} |Q_j^n - Q_{j-1}^n| \le M_5\Delta x, \qquad \max_{1\le j\le N} |g_j^n - g_{j-1}^n| \le M_5\Delta x, \\ &\max_{1\le j\le N} |m_j^n - m_{j-1}^n| \le M_5\Delta x, \end{aligned} \tag{3.13}$$

for $1 \le j \le N$;

$$\begin{aligned} &|g_{j+1}^n - 2g_j^n + g_{j-1}^n| \le M_6\Delta x\left(\Delta x + |u_j^n - u_{j-1}^n| + |u_{j+1}^n - u_j^n|\right), \\ &\qquad 1 \le j \le N-1, \\ &\quad |B_j^n - B_{j-1}^n| \le M_6\Delta x\left(\Delta x + |u_j^n - u_{j-1}^n| + |u_{j+1}^n - u_j^n|\right), \\ &\quad 4 \le j \le N-2; \end{aligned} \tag{3.14}$$

and

$$|Q_j^{n+1} - Q_j^n| \le M_7 TV(u^n)\Delta t + M_7\Delta t, \qquad |g_j^{n+1} - g_j^n| \le M_7 TV(u^n)\Delta t + M_7\Delta t,$$

$$|\beta_j^{n+1} - \beta_j^n| \le M_7 TV(u^n)\Delta t + M_7\Delta t \tag{3.15}$$

for $0 \le j \le N$.

Proof. By (3.11), we have, for $2 \le j \le N$,

$$|Q_j^n - Q_{j-1}^n| = \left|\frac{1}{2}(\alpha-1)(w_j u_j^n + w_{j-1}u_{j-1}^n)\right|\Delta x \le \|w\|_\infty\|u^n\|_\infty\Delta x \le M_3 M_4\Delta x,$$

which is clearly also valid for $j=0$ by (3.12). Therefore,

$$\begin{aligned} |g_j^n - g_{j-1}^n| &= \left|g(x_j,Q_j^n) - g(x_{j-1},Q_j^n) + g(x_{j-1},Q_j^n) - g(x_{j-1},Q_{j-1}^n)\right| \\ &\le |g_x(\hat{x}_j,Q_j^n)|\Delta x + |g_Q(x_{j-1},\hat{Q}_j)|\,|Q_j^n - Q_{j-1}^n| \\ &\le M_3\Delta x + M_3(M_3M_4\Delta x) = M_3(1+M_3M_4)\Delta x, \end{aligned}$$

$$\begin{aligned} |m_j^n - m_{j-1}^n| &= \left|m(x_j,Q_j^n) - m(x_{j-1},Q_j^n) + m(x_{j-1},Q_j^n) - m(x_{j-1},Q_{j-1}^n)\right| \\ &\le |m_x(\hat{x}_j,Q_j^n)|\Delta x + |m_Q(x_{j-1},\hat{Q}_j)|\,|Q_j^n - Q_{j-1}^n| \\ &\le M_3\Delta x + M_3(M_3M_4\Delta x) = M_3(1+M_3M_4)\Delta x. \end{aligned}$$

We have thus proved (3.13) with

$$M_5 = \max\left(M_3M_4,\ M_3(1+M_3M_4)\right).$$

As to (3.14), for $4 \le j \le N-2$, we can easily verify

$$\begin{aligned}|B_j^n - B_{j-1}^n| &= \left|\frac{1}{2}(g_{j+1}^n - 2g_j^n + g_{j-1}^n) + \frac{1}{2}(g_j^n - 2g_{j-1}^n + g_{j-2}^n)\right| \\ &\le \max_i |g_{i+1}^n - 2g_i^n + g_{i-1}^n|;\end{aligned}$$

hence we need only to prove the first inequality in (3.14). Using (3.11), we have, for $1 \le j \le N-1$,

$$\begin{aligned}|g_{j+1}^n - 2g_j^n + g_{j-1}^n| &= |(g_{j+1}^n - g_j^n) - (g_j^n - g_{j-1}^n)| \\ &= \left|\Delta_+\left(g_x(\hat{x}_j, Q_j^n)\Delta x + g_Q(x_{j-1}, \hat{Q}_j^n)(Q_j^n - Q_{j-1}^n)\right)\right| \\ &\le |g_x(\hat{x}_{j+1}, Q_{j+1}^n) - g_x(\hat{x}_j, Q_j^n)|\Delta x \\ &\quad + |g_Q(x_j, \hat{Q}_{j+1}^n)w_{j+1}u_{j+1}^n - g_Q(x_{j-1}, \hat{Q}_j^n)w_ju_j^n|\frac{(1-\alpha)}{2}\Delta x \\ &\quad + |g_Q(x_j, \hat{Q}_{j+1}^n)w_ju_j^n - g_Q(x_{j-1}, \hat{Q}_j^n)w_{j-1}u_{j-1}^n|\frac{(1-\alpha)}{2}\Delta x \\ &= I + II + III,\end{aligned}$$

where

$$\begin{aligned}I &= |g_x(\hat{x}_{j+1}, Q_{j+1}^n) - g_x(\hat{x}_j, Q_j^n)|\Delta x \\ &= |g_{xx}(\overline{x}_{j+1}, Q_{j+1}^n)(\hat{x}_{j+1} - \hat{x}_j) + g_{xQ}(\hat{x}_j, \overline{Q}_{j+1}^n)(Q_{j+1}^n - Q_j^n)|\Delta x \\ &\le 2M_3\Delta x^2 + M_3M_5\Delta x^2;\end{aligned}$$

and

$$\begin{aligned}II &= \left|g_Q(x_j, \hat{Q}_{j+1}^n)w_{j+1}u_{j+1}^n - g_Q(x_{j-1}, \hat{Q}_j^n)w_ju_j^n\right|\frac{(1-\alpha)}{2}\Delta x \\ &= \Big|(g_Q(x_j, \hat{Q}_{j+1}^n) - g_Q(x_{j-1}, \hat{Q}_j^n))w_{j+1}u_{j+1}^n + g_Q(x_{j-1}, \hat{Q}_j^n)u_{j+1}^n(w_{j+1} - w_j) \\ &\qquad + g_Q(x_{j-1}, \hat{Q}_j^n)w_j(u_{j+1}^n - u_j^n)\Big|\frac{(1-\alpha)}{2}\Delta x \\ &\le \frac{1}{2}\left|g_{Qx}(\hat{x}_j, \hat{Q}_{j+1}^n)\Delta x + g_{QQ}(x_{j-1}, \overline{Q}_{j+1}^n)(\hat{Q}_{j+1}^n - \hat{Q}_j^n)\right| \|w\|_\infty\|u^n\|_\infty\Delta x \\ &\quad + \frac{M_3}{2}\|u^n\|_\infty\|w_x\|_\infty\Delta x^2 + \frac{M_3}{2}\|w\|_\infty|u_{j+1}^n - u_j^n|\Delta x \\ &\le \frac{M_3^2}{2}M_4\Delta x^2 + \frac{M_3^2M_4}{2}\Delta x(2M_5\Delta x) + \frac{M_3^2M_4}{2}\Delta x^2 + \frac{M_3^2}{2}|u_{j+1}^n - u_j^n|\Delta x \\ &\le M_3^2M_4(1+M_5)\Delta x^2 + \frac{M_3^2}{2}\Delta x|u_{j+1}^n - u_j^n|.\end{aligned}$$

Similarly,

$$\begin{aligned}III &= \left|g_Q(x_j, \hat{Q}_{j+1}^n)w_ju_j^n - g_Q(x_{j-1}, \hat{Q}_j^n)w_{j-1}u_{j-1}^n\right|\frac{(1-\alpha)}{2}\Delta x \\ &\le M_3^2M_4(1+M_5)\Delta x^2 + \frac{M_3^2}{2}\Delta x|u_{j+1}^n - u_{j-1}^n|.\end{aligned}$$

Hence we have proved (3.14) with

$$M_6 = \max\left(M_3(2 + M_5 + 2M_3M_4 + 2M_3M_4M_5),\ \frac{M_3^2}{2}\right).$$

For (3.15), $0 \le j \le N$, we have, by the definition of Q_j in (3.4), that

$$|Q_j^{n+1} - Q_j^n| \le \frac{3}{2}\sum_{i=1}^{N} |u_i^{n+1} - u_i^n| w_i \Delta x \le \frac{3}{2} M_3 \sum_{i=1}^{N} |u_i^{n+1} - u_i^n| \Delta x.$$

From (3.6) and the definition of A_i^n and B_i^n, we have, for $1 \le i \le N$,

$$\begin{aligned}
|u_i^{n+1} - u_i^n| &= \left|-\lambda A_i^n u_i^n + \lambda(A_i^n - B_i^n)u_{i-1}^n - m_i^n u_i^n \Delta t\right| \\
&\le \lambda |A_i^n|\,|u_i^n - u_{i-1}^n| + \lambda |B_i^n| |u_{i-1}^n + \|m^n\|_\infty \|u^n\|_\infty \Delta t \\
&\le 2\lambda \sup_{\mathcal{D}} |g(x,Q)|\,|u_i^n - u_{i-1}^n| + \lambda M_4 \max_k |g_k^n - g_{k-1}^n| + M_3 M_4 \Delta t \\
&\le 2\lambda M_3 |u_i^n - u_{i-1}^n| + M_4 M_5 \Delta t + M_3 M_4 \Delta t. \qquad (3.16)
\end{aligned}$$

Thus we have

$$\begin{aligned}
|Q_j^{n+1} - Q_j^n| &\le \frac{3}{2} M_3 \sum_{i=1}^{N} (2\lambda M_3 |u_i^n - u_{i-1}^n| + M_4 M_5 \Delta t + M_3 M_4 \Delta t)\Delta x \\
&= 3M_3^2 TV(u^n)\Delta t + \frac{3}{2} M_3 M_4 (M_3 + M_5) L \Delta t,
\end{aligned}$$

which implies

$$\begin{aligned}
|g_j^{n+1} - g_j^n| &= |g(x_j, Q_j^{n+1}) - g(x_j, Q_j^n)| \\
&= |g_Q(x_j, \tilde{Q}_j)|\,|Q_j^{n+1} - Q_j^n| \le M_3 |Q_j^{n+1} - Q_j^n| \\
&\le 3M_3^3 TV(u^n)\Delta t + \frac{3}{2} M_3^2 M_4 (M_3 + M_5) L \Delta t, \\
|\beta_j^{n+1} - \beta_j^n| &= |\beta(x_j, Q_j^{n+1}) - \beta(x_j, Q_j^n)| \\
&= |\beta_Q(x_j, \tilde{Q}_j)|\,|Q_j^{n+1} - Q_j^n| \le M_3 |Q_j^{n+1} - Q_j^n| \\
&\le 3M_3^3 TV(u^n)\Delta t + \frac{3}{2} M_3^2 M_4 (M_3 + M_5) L \Delta t.
\end{aligned}$$

We have thus proved (3.15) with

$$M_7 = \max\left(3M_3^2,\, 3M_3^3,\, \frac{3}{2} M_3 M_4 (M_3 + M_5) L,\, \frac{3}{2} M_3^2 M_4 (M_3 + M_5) L\right). \qquad \square$$

We are now ready to prove the total variation stability of the scheme.

PROPOSITION 3.4. *$TV(u^n)$ is bounded for $t^n \le T$.*

Proof. First, we rewrite the scheme (3.6) as

$$u_j^{n+1} = u_j^n - \lambda A_j^n (u_j^n - u_{j-1}^n) - \lambda B_j^n u_{j-1}^n - \Delta t m_j^n u_j^n, \qquad j \ge 1.$$

We then have

$$\begin{aligned}
u_{j+1}^{n+1} - u_j^{n+1} &= \left[\left(1 - \lambda A_{j+1}^n\right)\left(u_{j+1}^n - u_j^n\right) + \lambda\left(A_j^n - B_j^n\right)\left(u_j^n - u_{j-1}^n\right)\right] \\
&\quad + \left[-\lambda u_j^n \left(B_{j+1}^n - B_j^n\right)\right] + \left[-\Delta t (m_{j+1}^n u_{j+1}^n - m_j^n u_j^n)\right] \\
&= D_j^n + E_j^n + F_j^n, \qquad j = 1, 2, \ldots, N-1.
\end{aligned}$$

Hence

$$TV(u^{n+1}) = \sum_{j=0}^{N-1} |u_{j+1}^{n+1} - u_j^{n+1}| \le \sum_{j=1}^{N-1} |D_j^n| + \sum_{j=1}^{N-1} |E_j^n| + \sum_{j=1}^{N-1} |F_j^n| + |u_1^{n+1} - u_0^{n+1}|.$$

We now estimate each term separately. First we have

$$\begin{aligned}\sum_{j=1}^{N-1} |D_j^n| &\le \sum_{j=1}^{N-1} (1 - \lambda A_{j+1}^n)|u_{j+1}^n - u_j^n| + \lambda(A_j^n - B_j^n)|u_j^n - u_{j-1}^n| \\ &= \sum_{j=1}^{N-1} |u_{j+1}^n - u_j^n| - \lambda \sum_{j=1}^{N-1} B_j^n |u_j^n - u_{j-1}^n| + \lambda g_1^n |u_1^n - u_0^n| - \lambda g_N^n |u_N^n - u_{N-1}^n| \\ &\le \sum_{j=1}^{N-1} |u_{j+1}^n - u_j^n| + M_5 \Delta t TV(u^n) + \lambda g_1^n |u_1^n - u_0^n|,\end{aligned}$$

where in the first inequality we have used (3.9) and (3.10), and in the last inequality we have used Lemma 3.3, the fact that

$$|B_j| \le \max_i |g_i - g_{i-1}|, \qquad 1 \le j \le N, \tag{3.17}$$

and the fact that $g(x_N, Q) = 0$. Using again Lemma 3.3 and (3.17), we have

$$\begin{aligned}\sum_{j=1}^{N-1} |E_j^n| &\le \sum_{j=4}^{N-2} \lambda |B_{j+1}^n - B_j^n|\,|u_j^n| + \sum_{j=1,2,3,N-1} \lambda |B_{j+1}^n - B_j^n|\,|u_j^n| \\ &\le M_4 M_6 \Delta t \left(\sum_{j=4}^{N-2} \Delta x + \sum_{j=4}^{N-2} (|u_{j+2}^n - u_{j+1}^n| + |u_{j+1}^n - u_j^n|) \right) + 8\lambda M_4 \|B^n\|_\infty \\ &\le M_4 M_6 L \Delta t + 2 M_4 M_6 \Delta t TV(u^n) + 8 M_4 M_5 \Delta t.\end{aligned}$$

The term F_j^n can be estimated as

$$\begin{aligned}|F_j^n| &= \Delta t |m_{j+1}^n u_{j+1}^n - m_j^n u_{j+1}^n + m_j^n u_{j+1}^n - m_j^n u_j^n| \\ &\le M_3 M_4 \Delta t \Delta x + M_3 |u_{j+1}^n - u_j^n| \Delta t.\end{aligned}$$

Hence we have

$$\sum_{j=1}^{N-1} |F_j^n| \le M_3 M_4 L \Delta t + M_3 \Delta t TV(u^n).$$

Let

$$M_8 = \max\{M_4((M_3 + M_6)L + 8M_5),\ M_3 + M_5 + 2M_4 M_6\};$$

we have

$$TV(u^{n+1}) \le M_8 \Delta t + M_8 \Delta t\, TV(u^n) + \sum_{j=1}^{N-1} |u_j^n - u_{j-1}^n| + \lambda g_1^n\, |u_1^n - u_0^n| + |u_1^{n+1} - u_0^{n+1}|.$$

Next we discuss $|u_1^{n+1} - u_0^{n+1}|$. This boundary term has the form

$$\begin{aligned}|u_1^{n+1} - u_0^{n+1}| &= |(1 - \lambda g_1^n - \Delta t m_1^n)u_1^n + \lambda g_0^n u_0^n - u_0^{n+1}| \\ &= |(1 - \lambda g_1^n)(u_1^n - u_0^n) - m_1^n u_1^n \Delta t - \lambda(g_1^n - g_0^n)u_0^n - (u_0^{n+1} - u_0^n)| \\ &\le (1 - \lambda g_1^n)|u_1^n - u_0^n| + M_3 M_4 \Delta t + M_5 M_4 \Delta t + |u_0^{n+1} - u_0^n|.\end{aligned}$$

We then have

$$TV(u^{n+1}) \le M_9 \Delta t + M_9 \Delta t\, TV(u^n) + TV(u^n) + |u_0^{n+1} - u_0^n|,$$

where

$$M_9 = M_8 + M_4(M_3 + M_5).$$

Finally, we must estimate $|u_0^{n+1} - u_0^n|$. From (3.3), we have

$$\begin{aligned}g_0^{n+1}u_0^{n+1} - g_0^n u_0^n &= g_0^{n+1}(u_0^{n+1} - u_0^n) + (g_0^{n+1} - g_0^n)u_0^n \\ &= C^{n+1} - C^n + \sum_{j=0}^{N}{}' \left(\beta_j^{n+1}u_j^{n+1} - \beta_j^n u_j^n\right)\Delta x \\ &= C^{n+1} - C^n + \sum_{j=0}^{N}{}' \left(\beta_j^{n+1}(u_j^{n+1} - u_j^n) + (\beta_j^{n+1} - \beta_j^n)u_j^n\right)\Delta x.\end{aligned}$$

Rearranging terms and using (3.16) and the results of Lemma 3.2, we obtain

$$\begin{aligned}&\left|\left(g_0^{n+1} - \frac{1}{2}\beta_0^{n+1}\Delta x\right)(u_0^{n+1} - u_0^n)\right| \\ &\le |C^{n+1} - C^n| + |g_0^{n+1} - g_0^n|u_0^n + \frac{1}{2}u_0^n \Delta x|\beta_0^{n+1} - \beta_0^n| \\ &\qquad + \sum_{j=1}^{N}(\beta_j^{n+1}|u_j^{n+1} - u_j^n| + |\beta_j^{n+1} - \beta_j^n|u_j^n)\Delta x \\ &\le M_3\Delta t + M_4 M_7 \Delta t(1 + TV(u^n)) + \frac{1}{2}M_4 M_7 \Delta x \Delta t(1 + TV(u^n)) \\ &\quad + M_3 \sum_{j=1}^{N}\left(2\lambda M_3|u_j^n - u_{j-1}^n| + M_4(M_3 + M_5)\Delta t\right)\Delta x + M_4 M_7 L \Delta t(1 + TV(u^n)) \\ &\le \left(M_4 M_7(2 + L) + 2M_3^2\right)\Delta t\, TV(u^n) + \left(M_3 + M_4 M_7(2 + L) + M_3 M_4(M_3 + M_5)L\right)\Delta t,\end{aligned}$$

where in the last inequality we have assumed $\Delta x \le 2$. Notice that, by (3.7), $g_0^{n+1} \ge \mu > 0$. Hence if $\Delta x \le \frac{\mu}{M_3}$, we have $g_0^{n+1} - \frac{1}{2}\beta_0^{n+1}\Delta x \ge \frac{\mu}{2} > 0$. Hence

$$|u_0^{n+1} - u_0^n| \le M_{10} TV(u^n)\Delta t + M_{10}\Delta t$$

with

$$M_{10} = \frac{1}{\mu}\max\left(M_4 M_7(2 + L) + 2M_3^2,\ M_3 + M_4 M_7(2 + L) + M_3 M_4(M_3 + M_5)L\right).$$

Now, with $M_{11} = M_9 + M_{10}$, we have

$$TV(u^{n+1}) \le (1 + M_{11}\Delta t)TV(u^n) + M_{11}\Delta t,$$

which implies the boundedness of $TV(u^n)$. ∎

Next, we show the Lipschitz stability in t.

PROPOSITION 3.5. *There exists a positive constant M such that for any $q > p$, we have*

$$\sum_{j=1}^{N} \left| \frac{u_j^q - u_j^p}{\Delta t} \right| \Delta x \le M(q-p).$$

Proof. Using (3.6), (3.17), and the definition of A_j^n and B_j^n, we obtain

$$\begin{aligned}
\sum_{j=1}^{N} \left| \frac{u_j^{n+1} - u_j^n}{\Delta t} \right| \Delta x &= \sum_{j=1}^{N} \left| \left(\frac{B_j^n}{\Delta x} + m_j^n \right) u_j^n + (A_j^n - B_j^n) D^-(u_j^n) \right| \Delta x \\
&\le \sum_{j=1}^{N} \max_i |g_i^n - g_{i-1}^n| u_j^n + M_3 \sum_{j=1}^{N} u_j^n \Delta x + 3M_3 \sum_{j=1}^{N} |u_j^n - u_{j-1}^n| \\
&\le M_4 M_5 L + M_3 M_4 L + 3M_3 TV(u^n) \le M.
\end{aligned}$$

Thus,

$$\sum_{j=1}^{N} \left| \frac{u_j^q - u_j^p}{\Delta t} \right| \Delta x \le \sum_{n=p}^{q} \sum_{j=1}^{N} \left| \frac{u_j^{n+1} - u_j^n}{\Delta t} \right| \Delta x \le M(q-p). \qquad ∎$$

If we again define a family of functions $\{U_{\Delta x, \Delta t}\}$ by

$$U_{\Delta x, \Delta t}(x,t) = u_j^n$$

for $x \in [x_{j-1}, x_j)$, $t \in [t^{n-1}, t^n)$, $j = 1, \ldots, N$, and $n = 1, \ldots, l$, then we have the following proposition. The proof is the same as that for Proposition 2.6.

PROPOSITION 3.6. *Under the time step restriction for the validity of previous propositions in this section, the numerical solution $\{U_{\Delta x, \Delta t}\}$ converges to the unique $BV([0,L] \times [0,T])$ solution $u(x,t)$ for (1.1) when $\Delta x \to 0$.*

Finally, we remark that the scheme (3.1) is second order in space but only first order in time. We should use the following second order TVD Runge–Kutta time discretization [17]:

$$u^{(1)} = u^n + \Delta t L(u^n), \qquad u^{n+1} = \frac{1}{2} \left(u^n + u^{(1)} + \Delta t L(u^{(1)}) \right), \tag{3.18}$$

where L is the spatial operator. This will yield a second order (in space and time) scheme which shares the same stability and convergence properties as the scheme (3.1). See also [9, 10].

4. Numerical examples. In this section we perform numerical experiments to demonstrate the properties of the schemes developed in previous sections. We take the initial condition as $u^0(x) = -x^2 + x + 1$, with the parameters and functions in (1.1) and (1.2) taken as $L = 1$, $\alpha = 0.5$, $w(x) = 1$, $g(x,Q) = (1-x)(5-x+x^2/2-Q)$, $m(x,Q) = 4 + 2Q + (1-x)^2/2$, $\beta(x,Q) = (1+x)(2-Q)$. As in [1], the choice of these particular functions and initial boundary conditions is simply to demonstrate the accuracy and high resolution properties of our schemes. However, as proven in the previous sections, our schemes will be stable and convergent for all population models (1.1) satisfying assumptions (H1)–(H6).

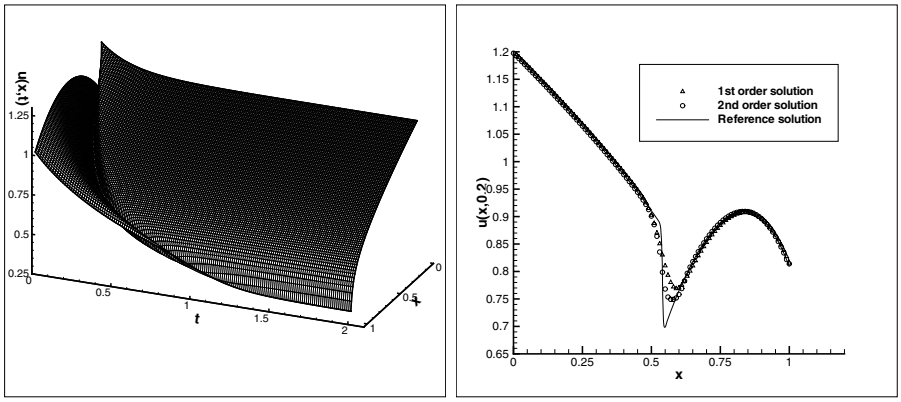


FIG. 4.1. *Left: The evolution of the solution to $t = 2$. Right: Numerical solutions using $N = 100$ uniform grid points using the first order scheme (triangles) and using the second order scheme (circles), versus the reference solution (solid line) obtained by the second order scheme using $N = 2000$ grid points.*

For the second order scheme, based on a local truncation error analysis, it is more accurate to adjust the mesh size for the second interval $x_2 - x_1$ from Δx to $\frac{3}{2}\Delta x$, and the mesh size for the second last interval $x_{N-1} - x_{N-2}$ from Δx to $\frac{1}{2}\Delta x$ (not the actual mesh sizes in the physical space—just that used in the scheme); hence we have made this adjustment in the computation. This apparently does not affect the stability and convergence analysis as the analysis does not require uniform meshes. The time step is chosen as $\Delta t_n = 0.8\Delta x / \|g^n(x,Q) + m^n(x,Q)\Delta x\|_\infty$ for the first order scheme, according to (2.5), and as $\Delta t_n = 0.8\Delta x / \|\frac{3}{2}g^n(x,Q) + m^n(x,Q)\Delta x\|_\infty$ for the second order scheme, according to (3.9).

First we demonstrate that the schemes are nonoscillatory in the presence of solution discontinuities. For this purpose we take $C(t) = 3$, which causes an incompatibility of the boundary data and the initial condition at the origin. The solution then has a discontinuity emitted from the left boundary and traveling to the right, until it moves outside the right boundary. See Figure 4.1, left, for the evolution of the solution until $t = 2$. When $t = 0.5$, the solution still contains a discontinuity. The numerical solutions using $N = 100$ uniformly spaced grid points for both the first order scheme and the second order scheme are plotted in Figure 4.1, right, against a reference solution which is obtained by the second order scheme with $N = 2000$ grid points. We can see clearly that both schemes can resolve the discontinuity without oscillation, and the second order scheme resolves the discontinuity much better without introducing spurious numerical oscillations. This verifies the high resolution property of the second order scheme. The solution for this problem changes very little after $t = 2$. We plot the rather smooth and monotone solution at $t = 20$ for both the first order scheme and the second order scheme in Figure 4.2. For such simple solutions there is no noticeable difference between the two schemes. Both schemes are stable for long time simulation.

Next, we demonstrate that the schemes can achieve their designed accuracy for smooth solutions. For this purpose we take $C(t) = \frac{38}{21} + t$, which ensures the compatibility of the boundary data and the initial condition at the origin. The solution then is continuous but has a discontinuous derivative (a kink) emitted from the left boundary and traveling to the right, until it moves outside the right boundary. When $t = 2$, the kink has already moved out of the right boundary and solution becomes

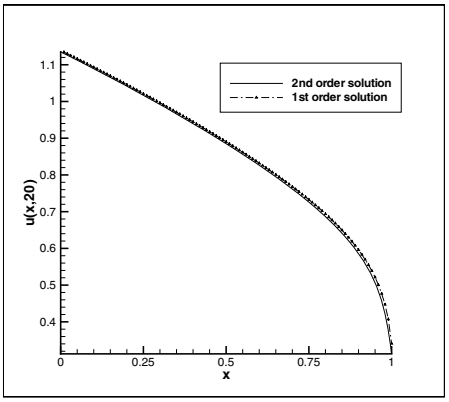


FIG. 4.2. *Solution at $t = 20$. $N = 100$ uniform grid points using the first order scheme (dashed line) and using the second order scheme (solid line).*

TABLE 4.1
L^1 errors and numerical order of accuracy of the first and second order schemes using N uniformly spaced mesh points.

N	First order scheme		Second order scheme	
N	L^1 error	order	L^1 error	order
10	9.49E-02		3.60E-02	
20	4.61E-02	1.04	1.19E-02	1.60
40	2.25E-02	1.03	3.88E-03	1.62
80	1.10E-02	1.03	1.18E-03	1.71
160	5.44E-03	1.02	3.03E-04	1.97

smooth. Since we do not know the exact solution, we use the second order scheme with $N = 10240$ grid points to produce a reference solution and then compute the L^1 errors of the first and second order schemes using coarser meshes; see Table 4.1. We can see that the designed orders of accuracy are obtained by the first and second order schemes for this smooth solution.

5. Concluding remarks. We have developed a first order explicit upwind scheme and a second order explicit high resolution scheme for solving a hierarchical size-structured population model with nonlinear growth, mortality, and reproduction rates, which contains global terms both for the boundary condition and for the coefficients in the equations. Stability and convergence are proved for both schemes for solutions with bounded total variation, which include discontinuous solutions. Numerical results are provided to demonstrate the capability of these schemes in resolving smooth as well as discontinuous solutions. Future work will include the design of higher order WENO schemes [12] with suitable treatment for boundary conditions and global constraints, and the study of schemes for the asymptotic behavior of the solution with techniques such as upwinding of the source, well-balancedness, etc., along the lines of, e.g., [7].

REFERENCES

[1] A. S. ACKLEH, K. DENG, AND S. HU, *A quasilinear hierarchical size structured model: Well-posedness and approximation*, Appl. Math. Optim., 51 (2005), pp. 35–59.

[2] K. W. Blayneh, *Hierarchical size-structured population model*, Dynam. Systems Appl., 9 (2002), pp. 527–540.
[3] A. Calsina and J. Saldana, *Asymptotic behavior of a model of hierarchically structured population dynamics*, J. Math. Biol., 35 (1997), pp. 967–987.
[4] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comput., 34 (1980), pp. 1–21.
[5] J. M. Cushing, *The dynamics of hierarchical age-structured population*, J. Math. Biol., 32 (1994), pp. 705–729.
[6] J. M. Cushing and J. Li, *Juvenile versus adult competition*, J. Math. Biol., 29 (1991), pp. 457–473.
[7] F. Filbet and C.-W. Shu, *Approximation of hyperbolic models for chemosensitive movement*, SIAM J. Sci. Comput., 27 (2005), pp. 850–872.
[8] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.
[9] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comput., 67 (1998), pp. 73–85.
[10] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.
[11] S. M. Henson and J. M. Cushing, *Hierarchical models of intra-specific competition: Scramble versus contest*, J. Math. Biol., 34 (1996), pp. 755–772.
[12] G. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[13] E. A. Kraev, *Existence and uniqueness for height structured hierarchical population models*, Natur. Resource Modeling, 14 (2001), pp. 45–70.
[14] R. J. LeVeque, *Numerical Methods for Conservation Laws*, Birkhäuser Verlag, Basel, 1990.
[15] S. Osher, *Convergence of generalized MUSCL schemes*, SIAM J. Numer. Anal., 22 (1985), pp. 947–961.
[16] C.-W. Shu, *TVB uniformly high-order schemes for conservation laws*, Math. Comp., 49 (1987), pp. 105–121.
[17] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[18] J. Smoller, *Shock Waves and Reaction-Diffusion Equations*, Springer-Verlag, New York, 1994.
[19] J. Weiner and S. X. Thomas, *Size variability and competition in plant monocultures*, Oikos, 47 (1986), pp. 211–222.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 371–388

AN ADAPTIVE LEAST-SQUARES MIXED FINITE ELEMENT METHOD FOR ELASTO-PLASTICITY*

GERHARD STARKE†

Abstract. A least-squares mixed finite element method for the incremental formulation of elasto-plasticity using a plastic flow rule of von Mises type with isotropic hardening is presented. This approach is based on the use of the stress tensor, in addition to the displacement field, as independent process variables. The nonlinear least-squares functional is shown to constitute an a posteriori error estimator on which an adaptive refinement strategy may be based. For the finite element implementation under plane strain conditions, quadratic (i.e., next-to-lowest-order) Raviart–Thomas elements are used for the stress approximation, while the displacement is represented by standard quadratic conforming elements. Computational results for a benchmark problem of elasto-plasticity under plane strain conditions are presented in order to illustrate the effectiveness of the least-squares approach.

Key words. least-squares mixed finite element method, elasto-plasticity, isotropic hardening, a posteriori error estimator

AMS subject classifications. 65N30, 65N50, 74C05

DOI. 10.1137/060652609

1. Introduction.

In this paper, a least-squares mixed finite element method for the incremental formulation of elasto-plastic deformation models is studied. This approach works with the stress tensor as an independent process variable, in addition to the displacement field. It is based on a first-order system modelling the elasto-plastic deformation process. The method studied in this paper constitutes an extension of the least-squares mixed finite element approach for linear elasticity presented in [12]. The closely related least-squares approaches investigated in [13, 14] could similarly be generalized to the elasto-plastic case. The main result of this paper is that, under the assumption of a plastic flow rule of von Mises type with isotropic hardening, the nonlinear least-squares functional associated with elasto-plasticity is elliptic with respect to an appropriate product space for the stresses and displacements. Our computational results suggest that the approximation properties actually deteriorate in the perfectly plastic case. This implies that it is not possible to extend our ellipticity result to perfect plasticity. Despite this deterioration of the approximation order in the perfectly plastic case, the adaptive implementation of the least-squares finite element method provides remarkably accurate results, in particular, for the stresses.

Finite element methods of least-squares type have been the object of many studies recently (see, e.g., the survey [7]). These methods may be viewed as an alternative to mixed finite element methods of saddle point structure whenever accurate approximations of the stress tensor are desired. Among its advantages is the greater flexibility in combining finite element spaces for the different process variables which are not restricted by an inf-sup condition. Moreover, if the least-squares functional is elliptic with respect to some norm on the underlying function spaces, then its local eval-

*Received by the editors February 21, 2006; accepted for publication (in revised form) September 21, 2006; published electronically February 13, 2007. Ths work was supported by the Deutsche Forschungsgemeinschaft (DFG) under grant STA 402/8-2.

http://www.siam.org/journals/sinum/45-1/65260.html

†Institut für Angewandte Mathematik, Universität Hannover, Welfengarten 1, 30167 Hannover, Germany (starke@ifam.uni-hannover.de).

uation provides an a posteriori error estimator. This may be used in an adaptive refinement technique; see [5] for a detailed study of such strategies in the context of least-squares formulations. The most appropriate combination for the elasto-plasticity models treated in this paper consists of Raviart–Thomas elements for the stresses coupled with conforming finite element spaces of the same polynomial degree for the displacement components. This is due to the fact that the same order of approximation is achieved for the individual variables. In particular, next-to-lowest-order Raviart–Thomas spaces are combined with piecewise quadratic conforming finite elements in our computations.

The numerical simulation of elasto-plastic deformation processes has been an intensive area of research for several decades. Two monographs which appeared at the end of the last century cover the state of the art from a more engineering-oriented perspective [22] and a more abstract mathematical view [15]. Error estimation and adaptive refinement strategies for elasto-plasticity based on duality techniques were proposed and studied in [19, 20]. Other approaches to adaptive finite element computations for elasto-plastic deformation processes are described in [1]. Even earlier, several error indicators were investigated about their suitability for problems of elasto-plasticity from an engineering point of view in [4]. The solution of the nonlinear algebraic systems associated with finite element discretizations of elasto-plastic models was the subject of [3, 6, 24]. As in [24], multigrid methods were applied in [16] to the solution of elasto-plastic deformation models discretized by finite elements. The efficiency was tested for benchmark test problems defined in [23], which constitutes another contribution to the same book resulting from a larger project on adaptive finite element methods in computational mechanics in which several research groups were involved. A detailed comparison of our least-squares finite element method with the adaptive approaches mentioned above is beyond the scope of this paper. Such a comparison would certainly depend on the choice of norm in which the approximation of the different variables is desired. At the very least, our least-squares method can be expected to provide more accurate stress approximations, in terms of computational effort, than a displacement-based approach.

The issue of time discretization is omitted almost completely in this paper by restricting ourselves to the implicit Euler scheme. Issues of the time discretization are important, however, in order to obtain accurate simulations of elasto-plastic deformation processes; see [10, 11] for details on this subject. A general framework for the numerical approximation of different models of elasto-plasticity was recently provided in [17].

In section 2, the first-order system model of incremental elasto-plasticity and the corresponding least-squares variational formulation are derived. The equivalence of the least-squares functional to a certain error norm on the product space of stresses and displacements is shown in section 3. This establishes ellipticity of the least-squares variational formulation and implies that the local evaluation of the least-squares functional constitutes an a posteriori error estimator to be used in adaptive refinement strategies. Section 4 contains the reduction to plane strain conditions and the specific finite element spaces appropriate under these circumstances. Finally, in section 5 the numerical results obtained with our adaptive least-squares method for a benchmark problem of elasto-plasticity are presented.

2. Least-squares formulation of incremental elasto-plasticity. Elasto-plastic deformation processes are usually modelled by a first-order system of the form

$$\begin{aligned} \operatorname{div} \boldsymbol{\sigma} &= 0\,, \\ \boldsymbol{\sigma} &= \mathcal{C}(\boldsymbol{\varepsilon}(\mathbf{u}) - \mathbf{p}) \end{aligned} \tag{2.1}$$

for the stress tensor $\boldsymbol{\sigma} : \Omega \to \mathbb{R}^{3\times 3}$ and the displacement field $\mathbf{u} : \Omega \to \mathbb{R}^3$. In (2.1), $\operatorname{div} \boldsymbol{\sigma}$ means row-wise application of the divergence operator, and $\nabla \mathbf{u}$ contains the gradient vectors of the components of $\mathbf{u}$ in each row. Similarly to the model of linear elasticity,

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \frac{1}{2}(\nabla \mathbf{u} + (\nabla \mathbf{u})^T) \tag{2.2}$$

denotes the strain tensor, and

$$\mathcal{C}\boldsymbol{\varepsilon} = 2\mu\boldsymbol{\varepsilon} + \lambda\,(\operatorname{tr}\boldsymbol{\varepsilon})\mathbf{I} \tag{2.3}$$

represents the linear material law. The difference from the elastic case lies in the term $\mathbf{p}$, which stands for the plastic strains satisfying additional constraints. To this end, we need to define the deviatoric stress part

$$\operatorname{dev}(\boldsymbol{\sigma}) = \boldsymbol{\sigma} - \frac{1}{3}\,(\operatorname{tr}\boldsymbol{\sigma})\,\mathbf{I}\,. \tag{2.4}$$

The system (2.1) is extended by the constraint

$$|\operatorname{dev}(\boldsymbol{\sigma})| \le \sqrt{\frac{2}{3}} K(\alpha) \tag{2.5}$$

with a hardening function $K(\alpha)$ and the evolution equations

$$\dot{\mathbf{p}} = \gamma \frac{\operatorname{dev}(\boldsymbol{\sigma})}{|\operatorname{dev}(\boldsymbol{\sigma})|}\,, \quad \dot{\alpha} = \gamma\sqrt{\frac{2}{3}}\,. \tag{2.6}$$

The parameter γ acts as a Lagrange multiplier associated with the constraint (2.5) and therefore satisfies

$$\gamma \ge 0 \quad \text{and} \quad \gamma\left(|\operatorname{dev}(\boldsymbol{\sigma})| - \sqrt{\frac{2}{3}}K(\alpha)\right) = 0\,. \tag{2.7}$$

The hardening parameter $\alpha : \Omega \to \mathbb{R}$ constitutes an additional process variable in the case of elasto-plasticity with hardening. Due to (2.6), elasto-plasticity models become time-dependent with the need to employ an appropriate time-discretization scheme. The model for elasto-plastic deformation processes described above is taken from [22, Chap. 2]. We restrict our exposition to these basic relations required for the derivation of the system arising in each step of a time-discretized model. Details on the mechanical background of elasto-plasticity models and different variational formulations suitable for numerical treatment may be found in [22, 15] or in [21, Chap. 6].

Discretization in time by an implicit Euler scheme leads to a first-order system for the increments $\boldsymbol{\sigma}^{\text{inc}}$ and $\mathbf{u}^{\text{inc}}$ in the representations $\boldsymbol{\sigma} = \boldsymbol{\sigma}^{\text{old}} + \boldsymbol{\sigma}^{\text{inc}}$ and $\mathbf{u} = \mathbf{u}^{\text{old}} + \mathbf{u}^{\text{inc}}$, respectively. The system associated with one time-step in an incremental formulation of elasto-plasticity may be written as

$$\begin{aligned} \operatorname{div}(\boldsymbol{\sigma}^{\text{old}} + \boldsymbol{\sigma}^{\text{inc}}) &= \mathbf{0}\,, \\ \boldsymbol{\sigma}^{\text{inc}} - \mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u}^{\text{inc}}); \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}}) &= \mathbf{0}\,. \end{aligned} \tag{2.8}$$

The stress operator $\mathcal{R}(\boldsymbol{\varepsilon}; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}})$ in (2.8) depends, in general, nonlinearly and non-smoothly on $\boldsymbol{\varepsilon}$ as soon as plastic deformation occurs.

For notational convenience the increments $\boldsymbol{\sigma}^{\text{inc}}$ and $\mathbf{u}^{\text{inc}}$ are simply denoted by $\boldsymbol{\sigma}$ and $\mathbf{u}$ (which had a different meaning in (2.1)) throughout the rest of this paper. For simplicity, we will also omit the dependence on $\boldsymbol{\sigma}^{\text{old}}$ and α^{old} in the stress operator and simply write $\mathcal{R}(\boldsymbol{\varepsilon})$ instead of $\mathcal{R}(\boldsymbol{\varepsilon}; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}})$. We introduce the Sobolev spaces

$$H(\operatorname{div}, \Omega) = \{\mathbf{s} \in L^2(\Omega)^3 : \operatorname{div} \mathbf{s} \in L^2(\Omega)\}\,,$$
$$H^1(\Omega) = \{p \in L^2(\Omega) : \nabla p \in L^2(\Omega)^3\}$$

and associated subspaces

$$H_{\Gamma_N}(\operatorname{div}, \Omega) = \{\mathbf{s} \in H(\operatorname{div}, \Omega) : \mathbf{s} \cdot \mathbf{n} = 0 \text{ on } \Gamma_N\}\,,$$
$$H^1_{\Gamma_D}(\Omega) = \{p \in H^1(\Omega) : p = 0 \text{ on } \Gamma_D\}$$

where homogeneous boundary conditions are imposed. The solution of (2.8) for $\boldsymbol{\sigma} : \Omega \to \mathbb{R}^{3\times 3}$ is then sought in $\boldsymbol{\sigma}^N + H_{\Gamma_N}(\operatorname{div}, \Omega)^3$, where $\boldsymbol{\sigma}^N \in H(\operatorname{div}, \Omega)^3$ satisfies the boundary conditions $\boldsymbol{\sigma}^N \cdot \mathbf{n} = \mathbf{g}$ on Γ_N. In connection with the incremental formulation (2.8), $\mathbf{g}$ stands for the increment of the boundary traction. The solution space for $\mathbf{u} : \Omega \to \mathbb{R}^3$ is $H^1_{\Gamma_D}(\Omega)^3$.

For the case of von Mises plasticity with isotropic hardening, the stress response is given by

$$\mathcal{R}(\boldsymbol{\varepsilon}) = \mathcal{C}\left(\boldsymbol{\varepsilon} - \frac{1}{2\mu}\,\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}))\,\frac{\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})}{|\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})|}\right)\,, \tag{2.9}$$

where the return parameter $\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}))$ is implicitly defined as the solution of the equation

$$\begin{aligned}&\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}))\\ &\quad = |\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})| - \sqrt{\frac{2}{3}}K\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\,\frac{\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}))}{2\mu}\right)\,,\end{aligned} \tag{2.10}$$

if $|\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\boldsymbol{\varepsilon})| > \sqrt{2/3}K(\alpha^{\text{old}})$ and $\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\boldsymbol{\varepsilon})) = 0$ otherwise. The return parameter γ_R in (2.10) plays the same role as the Lagrange multiplier γ in (2.6) and satisfies $\gamma_R = 2\mu\,\Delta t\,\gamma$, if Δt denotes the time-step size. The hardening parameter is updated by

$$\alpha = \alpha^{\text{old}} + \sqrt{\frac{2}{3}}\,\frac{\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}))}{2\mu}\,. \tag{2.11}$$

For the theoretical study in the next section, the following conditions on $K(\alpha)$ are assumed to hold for all $\alpha > 0$ (cf. [6]):

$$\begin{aligned}K(\alpha) &\geq K_0 > 0\,,\\ K'(\alpha) &\geq K_1 > 0\,.\end{aligned} \tag{2.12}$$

This is satisfied, for example, for exponential hardening where

$$K(\alpha) = K_0 + H\alpha + (K_\infty - K_0)(1 - e^{-\omega\alpha})$$

with given parameters $K_\infty \geq K_0 > 0$, $H > 0$, and $\omega > 0$ which are denoted as saturation stress, initial yield stress, hardening modulus, and hardening exponent, respectively.

In the case of perfect plasticity, we have $K(\alpha) \equiv K_0$, which means that (2.12) is not satisfied and the theoretical results in section 3 do not hold. Nevertheless we will present computational results for the perfectly plastic case in this paper. Since, for $|\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})| > \sqrt{2/3}K_0$, the return parameter is simply given by

$$\gamma_R(\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})) = |\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})| - \sqrt{\frac{2}{3}}K_0\,,$$

we end up with

$$\mathcal{R}(\boldsymbol{\varepsilon}) = \frac{1}{3}\,\mathrm{tr}(\mathcal{C}\boldsymbol{\varepsilon})\,\mathbf{I} - \mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}}) + \sqrt{\frac{2}{3}}K_0\frac{\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})}{|\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon})|} \tag{2.13}$$

for the stress response. This means that different strain increments $\boldsymbol{\varepsilon}$ and $\bar{\boldsymbol{\varepsilon}}$ lead to the same stress response as long as $\mathrm{tr}(\boldsymbol{\varepsilon}) = \mathrm{tr}(\bar{\boldsymbol{\varepsilon}})$ and $\mathrm{dev}(\boldsymbol{\varepsilon})$ is aligned with $\mathrm{dev}(\bar{\boldsymbol{\varepsilon}})$.

We close this section with the least-squares formulation of the first-order system (2.8). Throughout this paper, $\|\cdot\|$ will simply denote the $L^2(\Omega)$ (or, if applicable, $L^2(\Omega)^d$, $L^2(\Omega)^{d\times d}$) norm. The least-squares functional, associated with (2.8), is given by

$$\mathcal{F}(\boldsymbol{\sigma}, \mathbf{u}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}}) = \|\mathrm{div}(\boldsymbol{\sigma}^{\mathrm{old}} + \boldsymbol{\sigma})\|^2 + \|\mathcal{C}^{-1/2}(\boldsymbol{\sigma} - \mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u}))\|^2\,. \tag{2.14}$$

The weighting of the second term in the above functional is motivated from our earlier work on linear elasticity in [13, 12]. Its implementation is straightforward using the explicit formula

$$\mathcal{C}^{-1}\boldsymbol{\sigma} = \frac{1}{2\mu}\boldsymbol{\sigma} - \frac{\lambda}{2\mu(3\lambda + 2\mu)}(\mathrm{tr}\,\boldsymbol{\sigma})\mathbf{I}$$

for the inverse of the operator defined in (2.3). The corresponding least-squares formulation consists in minimizing (2.14) among all suitable $(\boldsymbol{\sigma}, \mathbf{u}) \in H(\mathrm{div}, \Omega)^3 \times H^1(\Omega)^3$. More precisely, our aim is to find $\boldsymbol{\sigma} \in \boldsymbol{\sigma}^N + H_{\Gamma_N}(\mathrm{div}, \Omega)^3$ and $\mathbf{u} \in H^1_{\Gamma_D}(\Omega)^3$ such that

$$\mathcal{F}(\boldsymbol{\sigma}, \mathbf{u}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}}) \leq \mathcal{F}(\boldsymbol{\sigma}^N + \boldsymbol{\tau}, \mathbf{v}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}}) \tag{2.15}$$

holds for all $\boldsymbol{\tau} \in H_{\Gamma_N}(\mathrm{div}, \Omega)^3$ and $\mathbf{v} \in H^1_{\Gamma_D}(\Omega)^3$. For hardening laws which satisfy (2.12), the well-posedness of the system (2.8) is studied in [15, sect. 8]. If the first-order system (2.8) is guaranteed to possess a unique solution, then it is also the unique minimizer of (2.15).

The analysis carried out in the next section will be based on a Korn inequality of the form

$$\left(\|\mathbf{v}\|^2 + \|\nabla\mathbf{v}\|^2\right) \leq C_K\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 \tag{2.16}$$

to hold for all $\mathbf{v} \in H^1_{\Gamma_D}(\Omega)^3$ with a constant C_K. Korn's inequality (2.16) is known to hold, e.g., if Γ_D does not vanish (cf. [8, sect. VI.3]). In fact, the constant in (2.16) satisfies $C_K \geq 2\mu$, since, if $\mathrm{div}\,\mathbf{v} = 0$,

$$\begin{aligned}
\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2 = 2\mu\|\boldsymbol{\varepsilon}(\mathbf{v})\|^2 &= 2\mu\left(\|\partial_1 v_1\|^2 + \|\partial_2 v_2\|^2 + 2\|\frac{1}{2}(\partial_2 v_1 + \partial_1 v_2)\|^2\right)\\
&\leq 2\mu\left(\|\partial_1 v_1\|^2 + \|\partial_2 v_2\|^2 + \|\partial_2 v_1\|^2 + \|\partial_1 v_2\|^2\right) = 2\mu\|\nabla\mathbf{v}\|^2\\
&\leq 2\mu\left(\|\mathbf{v}\|^2 + \|\nabla\mathbf{v}\|^2\right)\,.
\end{aligned}$$

3. The nonlinear least-squares functional as an error estimator. In this section, the equivalence of the nonlinear least-squares functional in (2.14) to the natural norm of the error is established. To this end, the estimate given in the following lemma is required.

LEMMA 3.1. *Under the assumptions* (2.12)*, there exists a constant* $C_R \in [0,1)$ *such that*

$$\|\mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{v}))) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v})\| \le C_R\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v})\| \tag{3.1}$$

holds for all $\mathbf{u},\mathbf{v} \in H^1_{\Gamma_D}(\Omega)^3$.

Proof. The special form (2.9) of the stress operator $\mathcal{R}(\boldsymbol{\varepsilon})$ implies

$$\begin{aligned}
&\mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{v}))) - \mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v}) \\
&\quad = \frac{1}{\sqrt{2\mu}}\left(\gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{v})))\frac{\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{v}))}{|\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{v}))|}\right. \\
&\qquad\qquad \left. - \gamma_R(\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u})))\frac{\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}))}{|\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}))|}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v})\| &\ge \|\operatorname{dev}(\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v}))\| = \|\sqrt{2\mu}\operatorname{dev}(\boldsymbol{\varepsilon}(\mathbf{u}-\mathbf{v}))\| \\
&= \left\|\frac{1}{\sqrt{2\mu}}\left(\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + 2\mu\,\boldsymbol{\varepsilon}(\mathbf{u})) - \operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + 2\mu\,\boldsymbol{\varepsilon}(\mathbf{v}))\right)\right\| \\
&= \left\|\frac{1}{\sqrt{2\mu}}\left(\operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u})) - \operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{v}))\right)\right\|.
\end{aligned}$$

With the abbreviations $\boldsymbol{\xi} = \operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}))$ and $\boldsymbol{\eta} = \operatorname{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{v}))$ this means that it is sufficient to show

$$\left\|\gamma_R(\boldsymbol{\eta})\frac{\boldsymbol{\eta}}{|\boldsymbol{\eta}|} - \gamma_R(\boldsymbol{\xi})\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|}\right\| \le C_R\|\boldsymbol{\eta}-\boldsymbol{\xi}\|.$$

This inequality certainly holds if

$$\left|\gamma_R(\boldsymbol{\eta})\frac{\boldsymbol{\eta}}{|\boldsymbol{\eta}|} - \gamma_R(\boldsymbol{\xi})\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|}\right| \le C_R|\boldsymbol{\eta}-\boldsymbol{\xi}| \tag{3.2}$$

is satisfied for all $x \in \Omega$.

In order to prove (3.2), we fix $x \in \Omega$ and investigate the function

$$\mathcal{S} : \mathbb{R}^{3\times 3} \to \mathbb{R}^{3\times 3}, \quad \mathcal{S}(\boldsymbol{\xi}) = \gamma_R(\boldsymbol{\xi})\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|} \tag{3.3}$$

more closely. $\mathcal{S}$ is differentiable at all $\boldsymbol{\xi} \in \mathbb{R}^{3\times 3}$ with $|\boldsymbol{\xi}| \ne \sqrt{2/3}K(\alpha^{\mathrm{old}})$. For $|\boldsymbol{\xi}| < \sqrt{2/3}K(\alpha^{\mathrm{old}})$, obviously, $\mathcal{S}'(\boldsymbol{\xi})[\boldsymbol{\chi}] = \mathbf{0}$, while, for $|\boldsymbol{\xi}| > \sqrt{2/3}K(\alpha^{\mathrm{old}})$,

$$\mathcal{S}'(\boldsymbol{\xi})[\boldsymbol{\chi}] = \gamma_R'(\boldsymbol{\xi})[\boldsymbol{\chi}]\frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|} + \gamma_R(\boldsymbol{\xi})\frac{\boldsymbol{\chi}}{|\boldsymbol{\xi}|} - \gamma_R(\boldsymbol{\xi})\frac{(\boldsymbol{\xi}:\boldsymbol{\chi})\,\boldsymbol{\xi}}{|\boldsymbol{\xi}|^3}$$

holds. $\gamma_R'(\boldsymbol{\xi})$ may be computed from differentiating the defining equation (2.10). This leads to

$$\gamma_R'(\boldsymbol{\xi})[\boldsymbol{\chi}] = \frac{\boldsymbol{\xi}:\boldsymbol{\chi}}{|\boldsymbol{\xi}|} - \frac{1}{3\mu}\gamma_R'(\boldsymbol{\xi})[\boldsymbol{\chi}]\,K'\left(\alpha^{\mathrm{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right)$$

or, equivalently,

$$\gamma_R'(\boldsymbol{\xi})[\boldsymbol{\chi}] = \left(1 + \frac{1}{3\mu}\, K'\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right)\right)^{-1} \frac{\boldsymbol{\xi} : \boldsymbol{\chi}}{|\boldsymbol{\xi}|}\,. \tag{3.4}$$

This implies

$$\begin{aligned}|S'(\boldsymbol{\xi})[\boldsymbol{\chi}]|^2 = \left(1 + \frac{1}{3\mu}\, K'\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right)\right)^{-2} \frac{(\boldsymbol{\xi} : \boldsymbol{\chi})^2}{|\boldsymbol{\xi}|^2} \\ + \gamma_R(\boldsymbol{\xi})^2 \left(\frac{|\boldsymbol{\chi}|^2}{|\boldsymbol{\xi}|^2} - \frac{(\boldsymbol{\xi} : \boldsymbol{\chi})^2}{|\boldsymbol{\xi}|^4}\right),\end{aligned}$$

which leads to

$$\begin{aligned}\frac{|S'(\boldsymbol{\xi})[\boldsymbol{\chi}]|^2}{|\boldsymbol{\chi}|^2} &= \left(1 + \frac{1}{3\mu}\, K'\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right)\right)^{-2} \frac{(\boldsymbol{\xi} : \boldsymbol{\chi})^2}{|\boldsymbol{\xi}|^2|\boldsymbol{\chi}|^2} + \frac{\gamma_R(\boldsymbol{\xi})^2}{|\boldsymbol{\xi}|^2}\left(1 - \frac{(\boldsymbol{\xi} : \boldsymbol{\chi})^2}{|\boldsymbol{\xi}|^2|\boldsymbol{\chi}|^2}\right) \\ &\le \max\left\{\left(1 + \frac{1}{3\mu}\, K'\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right)\right)^{-2}, \frac{\gamma_R(\boldsymbol{\xi})^2}{|\boldsymbol{\xi}|^2}\right\}\end{aligned}$$

for all $\boldsymbol{\chi} \neq \mathbf{0}$. This may be rewritten as

$$\sup_{\boldsymbol{\chi}\neq 0} \frac{|S'(\boldsymbol{\xi})[\boldsymbol{\chi}]|}{|\boldsymbol{\chi}|} \le \max\left\{\left(1 + \frac{1}{3\mu}\, K'\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right)\right)^{-1}, \frac{\gamma_R(\boldsymbol{\xi})}{|\boldsymbol{\xi}|}\right\}. \tag{3.5}$$

For the first of the two terms on the right-hand side in (3.5), (2.12) implies that it is bounded by $(1 + K_1/(3\mu))^{-1}$. For the second term let C_T denote an upper bound for the largest strain increment in the sense that $|\text{dev}(\boldsymbol{\varepsilon}(\mathbf{u}))| \le C_T$ holds. This implies that

$$|\text{dev}(\boldsymbol{\sigma}^{\text{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}))| \le |\text{dev}(\boldsymbol{\sigma}^{\text{old}})| + 2\mu|\text{dev}(\boldsymbol{\varepsilon}(\mathbf{u}))| \le \sqrt{\frac{2}{3}}K(\alpha^{\text{old}}) + 2\mu C_T$$

is satisfied. Using (2.10) and the fact that K is monotonically increasing (which follows from (2.12)), this leads to

$$\begin{aligned}\frac{\gamma_R(\boldsymbol{\xi})}{|\boldsymbol{\xi}|} &= 1 - \frac{1}{|\boldsymbol{\xi}|}\sqrt{\frac{2}{3}}K\left(\alpha^{\text{old}} + \sqrt{\frac{2}{3}}\frac{\gamma_R(\boldsymbol{\xi})}{2\mu}\right) \le 1 - \frac{1}{|\boldsymbol{\xi}|}\sqrt{\frac{2}{3}}K(\alpha^{\text{old}}) \\ &\le 1 - \frac{\sqrt{\frac{2}{3}}K(\alpha^{\text{old}})}{\sqrt{\frac{2}{3}}K(\alpha^{\text{old}}) + 2\mu C_T} = \frac{1}{1 + \sqrt{\frac{2}{3}}\frac{K(\alpha^{\text{old}})}{2\mu C_T}} \le \frac{1}{1 + \sqrt{\frac{2}{3}}\frac{K_0}{2\mu C_T}}\,.\end{aligned}$$

Therefore, (3.2) holds with

$$C_R = \max\left\{\left(1 + \frac{K_1}{3\mu}\right)^{-1}, \left(1 + \sqrt{\frac{2}{3}}\frac{K_0}{2\mu C_T}\right)^{-1}\right\} < 1\,. \qquad \square$$

Remark. In the case of perfect plasticity, different displacements $\mathbf{u}$ and $\mathbf{v}$ may lead to the same stress states $\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) = \mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{v}))$ as can be seen from (2.13). Therefore, (3.1) does not hold with $C_R < 1$ in that case.

THEOREM 3.2. *Let $\boldsymbol{\sigma} \in \boldsymbol{\sigma}^N + H_{\Gamma_N}(\mathrm{div}, \Omega)^3$, $\mathbf{u} \in H^1_{\Gamma_D}(\Omega)^3$ be the solution of the first-order system* (2.8)*. Then, under the assumptions* (2.12)*, there exist positive constants $\underline{\beta}$, $\overline{\beta}$ (which do not depend on the Lamé parameter λ) such that*

$$
\begin{aligned}
&\underline{\beta}\left(\|\mathrm{div}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2 + \|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})\|^2\right)\\
&\le \mathcal{F}(\bar{\boldsymbol{\sigma}}, \bar{\mathbf{u}}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}})\\
&\le \overline{\beta}\left(\|\mathrm{div}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2 + \|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})\|^2\right)
\end{aligned}
\tag{3.6}
$$

holds for all $\bar{\boldsymbol{\sigma}} \in \boldsymbol{\sigma}^N + H_{\Gamma_N}(\mathrm{div}, \Omega)^3$ and $\bar{\mathbf{u}} \in H^1_{\Gamma_D}(\Omega)^3$.

Proof. For simplicity we set $\mu = 1$ and observe that the equivalence is invariant with respect to the scaling of μ. Using the fact that $(\boldsymbol{\sigma}, \mathbf{u})$ is the exact solution of (2.8), we obtain

$$
\begin{aligned}
\mathcal{F}(\bar{\boldsymbol{\sigma}}, \bar{\mathbf{u}}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}}) &= \|\mathrm{div}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2\\
&\quad + \|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}}) - \mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\bar{\mathbf{u}})))\|^2\,.
\end{aligned}
\tag{3.7}
$$

For the lower bound in (3.6), we use the decomposition of an arbitrary matrix-valued function $\boldsymbol{\tau} \in L^2(\Omega)^{d\times d}$ into its symmetric and antisymmetric parts,

$$
\boldsymbol{\tau} = \mathrm{sy}\,\boldsymbol{\tau} + \mathrm{as}\,\boldsymbol{\tau} \text{ with } \mathrm{sy}\,\boldsymbol{\tau} = \frac{\boldsymbol{\tau}+\boldsymbol{\tau}^T}{2}\,, \quad \mathrm{as}\,\boldsymbol{\tau} = \frac{\boldsymbol{\tau}-\boldsymbol{\tau}^T}{2}\,.
$$

Obviously, $(\mathrm{sy}\,\boldsymbol{\tau}, \mathrm{as}\,\boldsymbol{\tau})_{0,\Omega} = 0$, which implies

$$
\|\boldsymbol{\tau}\|^2 = \|\mathrm{sy}\,\boldsymbol{\tau}\|^2 + \|\mathrm{as}\,\boldsymbol{\tau}\|^2 \ge \|\mathrm{as}\,\boldsymbol{\tau}\|^2\,.
$$

If this estimate is applied with $\boldsymbol{\tau} = \mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}}) - \mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\bar{\mathbf{u}})))$, we obtain

$$
\begin{aligned}
\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}}) - \mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\bar{\mathbf{u}})))\|^2 &\ge \|\mathrm{as}(\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}}))\|^2\\
&= \frac{1}{2}\|\mathrm{as}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2
\end{aligned}
\tag{3.8}
$$

(note that $\mathrm{as}(\mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\bar{\mathbf{u}})))) = \mathbf{0}$). The combination of (3.7) and (3.8) leads to

$$
\begin{aligned}
\mathcal{F}(\bar{\boldsymbol{\sigma}}, \bar{\mathbf{u}}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}}) \ge \frac{1}{3}\Big(&\|\mathrm{div}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2 + \|\mathrm{as}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2\\
&+\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}}) - \mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) - \mathcal{R}(\boldsymbol{\varepsilon}(\bar{\mathbf{u}})))\|^2\Big)\,.
\end{aligned}
$$

Inserting $C^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})$ and using the result of Lemma 3.1, we are further led to

$$\begin{aligned}\mathcal{F}(\bar{\boldsymbol{\sigma}},\bar{\mathbf{u}};\boldsymbol{\sigma}^{\mathrm{old}},\alpha^{\mathrm{old}}) &\geq \frac{1}{3}\Big(\|\mathrm{div}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2+\|\mathrm{as}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2\\ &\quad+(1-\rho)\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})-\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})\|^2\\ &\quad-\left(\frac{1}{\rho}-1\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})-\mathcal{C}^{-1/2}(\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u}))-\mathcal{R}(\boldsymbol{\varepsilon}(\bar{\mathbf{u}})))\|^2\Big)\\ &\geq \frac{1}{3}\Big(\|\mathrm{div}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2+\|\mathrm{as}\,(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})\|^2\\ &\quad+(1-\rho)\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}})-\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})\|^2\\ &\quad-\left(\frac{1}{\rho}-1\right)C_R^2\,\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u}-\bar{\mathbf{u}})\|^2\Big)=:\frac{1-\rho}{3}\mathcal{G}(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}},\mathbf{u}-\bar{\mathbf{u}})\,,\end{aligned}\tag{3.9}$$

where $\rho\in(0,1)$ is still free to be chosen appropriately below.

If we set $(\boldsymbol{\tau},\mathbf{v})=(\boldsymbol{\sigma}-\bar{\boldsymbol{\sigma}},\mathbf{u}-\bar{\mathbf{u}})\in H_{\Gamma_N}(\mathrm{div},\Omega)^3\times H^1_{\Gamma_D}(\Omega)^3$, then we are left with estimating the quadratic functional

$$\mathcal{G}(\boldsymbol{\tau},\mathbf{v})=\frac{\|\mathrm{div}\,\boldsymbol{\tau}\|^2+\|\mathrm{as}\,\boldsymbol{\tau}\|^2}{1-\rho}+\|\mathcal{C}^{-1/2}\boldsymbol{\tau}-\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2-\frac{C_R^2}{\rho}\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2\tag{3.10}$$

from below. The decomposition of $\boldsymbol{\tau}$ into its symmetric and antisymmetric parts and integration by parts lead to

$$\begin{aligned}(\boldsymbol{\tau},\boldsymbol{\varepsilon}(\mathbf{v}))_{0,\Omega}&=(\mathrm{sy}\,\boldsymbol{\tau},\boldsymbol{\varepsilon}(\mathbf{v}))_{0,\Omega}+(\mathrm{as}\,\boldsymbol{\tau},\boldsymbol{\varepsilon}(\mathbf{v}))_{0,\Omega}=(\mathrm{sy}\,\boldsymbol{\tau},\boldsymbol{\varepsilon}(\mathbf{v}))_{0,\Omega}\\ &=(\mathrm{sy}\,\boldsymbol{\tau},\nabla\mathbf{v})_{0,\Omega}=(\boldsymbol{\tau},\nabla\mathbf{v})_{0,\Omega}-(\mathrm{as}\,\boldsymbol{\tau},\nabla\mathbf{v})_{0,\Omega}\\ &=-(\mathrm{div}\,\boldsymbol{\tau},\mathbf{v})_{0,\Omega}-(\mathrm{as}\,\boldsymbol{\tau},\nabla\mathbf{v})_{0,\Omega}\,.\end{aligned}\tag{3.11}$$

Inserting (3.11) into (3.10) we obtain

$$\begin{aligned}\mathcal{G}(\boldsymbol{\tau},\mathbf{v})&=\frac{1}{1-\rho}\left(\|\mathrm{div}\,\boldsymbol{\tau}\|^2+\|\mathrm{as}\,\boldsymbol{\tau}\|^2\right)+\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2\\ &\quad+\left(1-\frac{C_R^2}{\rho}\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2-2(\boldsymbol{\tau},\boldsymbol{\varepsilon}(\mathbf{v}))\\ &=\frac{1}{1-\rho}\left(\|\mathrm{div}\,\boldsymbol{\tau}\|^2+\|\mathrm{as}\,\boldsymbol{\tau}\|^2\right)+\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2\\ &\quad+\left(1-\frac{C_R^2}{\rho}\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2+2\left((\mathrm{div}\,\boldsymbol{\tau},\mathbf{v})+(\mathrm{as}\,\boldsymbol{\tau},\nabla\mathbf{v})\right)\,.\end{aligned}\tag{3.12}$$

For the last term in (3.12), Korn's inequality (2.16) may be used to obtain

$$\begin{aligned}2\left((\mathrm{div}\,\boldsymbol{\tau},\mathbf{v})+(\mathrm{as}\,\boldsymbol{\tau},\nabla\mathbf{v})\right)&\leq\frac{1}{\delta}\left(\|\mathrm{div}\,\boldsymbol{\tau}\|^2+\|\mathrm{as}\,\boldsymbol{\tau}\|^2\right)+\delta\left(\|\mathbf{v}\|^2+\|\nabla\mathbf{v}\|^2\right)\\ &\leq\frac{1}{\delta}\left(\|\mathrm{div}\,\boldsymbol{\tau}\|^2+\|\mathrm{as}\,\boldsymbol{\tau}\|^2\right)+C_K\delta\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2\,,\end{aligned}$$

where $\delta\in(0,1)$ is still free to be chosen appropriately. Inserting this into (3.12) yields

$$\begin{aligned}\mathcal{G}(\boldsymbol{\tau},\mathbf{v})&\geq\left(\frac{1}{1-\rho}-\frac{1}{\delta}\right)\left(\|\mathrm{div}\,\boldsymbol{\tau}\|^2+\|\mathrm{as}\,\boldsymbol{\tau}\|^2\right)+\|\mathcal{C}^{-1/2}\boldsymbol{\tau}\|^2\\ &\quad+\left(1-\frac{C_R^2}{\rho}-C_K\delta\right)\|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{v})\|^2\,.\end{aligned}\tag{3.13}$$

If ρ is restricted to the interval $(C_R^2, 1)$, then we may choose

$$\delta = \left(\frac{1-\rho}{C_K} \left(1 - \frac{C_R^2}{\rho} \right) \right)^{1/2}$$

($\delta < 1$ is satisfied due to $C_K \geq 2$). If we insert this into (3.13), we see that

$$\begin{aligned} \mathcal{G}(\boldsymbol{\tau}, \mathbf{v}) \geq \frac{1}{1-\rho} \left(1 - \left(\frac{C_K(1-\rho)}{1 - \frac{C_R^2}{\rho}} \right)^{1/2} \right) \left(\|\operatorname{div} \boldsymbol{\tau}\|^2 + \|\operatorname{as} \boldsymbol{\tau}\|^2 \right) + \|\mathcal{C}^{-1/2} \boldsymbol{\tau}\|^2 \\ + \left(1 - \frac{C_R^2}{\rho} \right) \left(1 - \left(\frac{C_K(1-\rho)}{1 - \frac{C_R^2}{\rho}} \right)^{1/2} \right) \|\mathcal{C}^{1/2} \boldsymbol{\varepsilon}(\mathbf{v})\|^2 \end{aligned} \tag{3.14}$$

holds. Finally, $\rho \in (C_R^2, 1)$ may be chosen such that it satisfies

$$\frac{C_K(1-\rho)}{1 - \frac{C_R^2}{\rho}} < 1 \,.$$

(This is clearly possible, since the left-hand side tends to 0 as ρ approaches 1 and depends continuously on ρ.) We have therefore shown that

$$\mathcal{G}(\boldsymbol{\tau}, \mathbf{v}) \geq \hat{\beta} \left(\|\operatorname{div} \boldsymbol{\tau}\|^2 + \|\mathcal{C}^{-1/2} \boldsymbol{\tau}\|^2 + \|\mathcal{C}^{1/2} \boldsymbol{\varepsilon}(\mathbf{v})\|^2 \right) \tag{3.15}$$

holds with

$$\hat{\beta} = \left(1 - \frac{C_R^2}{\rho} \right) \left(1 - \left(\frac{C_K(1-\rho)}{1 - \frac{C_R^2}{\rho}} \right)^{1/2} \right) > 0,$$

which, combined with (3.9), implies the lower bound in (3.6) with $\underline{\beta} = \hat{\beta}(1-\rho)/3$.

The upper bound in (3.6) follows directly from (3.7) and (3.1), which gives

$$\begin{aligned} \mathcal{F}(\bar{\boldsymbol{\sigma}}, \bar{\mathbf{u}}; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}}) \leq \|\operatorname{div} (\boldsymbol{\sigma} - \bar{\boldsymbol{\sigma}})\|^2 + 2\|\mathcal{C}^{-1/2}(\boldsymbol{\sigma} - \bar{\boldsymbol{\sigma}})\|^2 \\ + 2(1 + C_R)^2 \|\mathcal{C}^{1/2} \boldsymbol{\varepsilon}(\mathbf{u} - \bar{\mathbf{u}})\|^2 \,. \end{aligned}$$

□

Remark. In the case of perfect plasticity, it is no longer possible to show the lower bound in (3.6) along the lines in the above proof. In fact, our numerical results documented in section 5 suggest that the equivalence (3.6) is actually lost in the case of perfect plasticity.

The practical implication of Theorem 3.2 is that, under the assumptions (2.12), the least-squares functional $\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}})$ constitutes an a posteriori estimator for any approximation $(\boldsymbol{\sigma}_h, \mathbf{u}_h)$. By its very definition, for any triangulation $\mathcal{T}_h$ of Ω,

$$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}}) = \sum_{T \in \mathcal{T}_h} \mathcal{F}_T(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}}) \,. \tag{3.16}$$

This means that the local evaluation of the functional, $\mathcal{F}_T(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\sigma}^{\text{old}}, \alpha^{\text{old}})$, can be used in an adaptive refinement strategy.

Of course, the approximation $(\boldsymbol{\sigma}_h, \mathbf{u}_h)$ to be used in practice comes from the solution of the least-squares minimization problem (2.15) with respect to finite element spaces. It can already be observed from the definition (2.9) that $\mathcal{R}(\varepsilon(\mathbf{u}))$ is not differentiable with respect to $\mathbf{u}$ everywhere. In the proof of Lemma 3.1, it becomes apparent that

$$\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u})) = \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}) - \mathcal{S}(\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\varepsilon(\mathbf{u})))$$

is not smooth for those $\mathbf{u} \in H^1_{\Gamma_D}(\Omega)^3$ with

$$|\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\varepsilon(\mathbf{u}))| = \sqrt{\frac{2}{3}} K(\alpha^{\mathrm{old}})\,, \tag{3.17}$$

since $\mathcal{S}(\boldsymbol{\xi})$, defined in (3.3), is not smooth for $|\boldsymbol{\xi}| = \sqrt{2/3}K(\alpha^{\mathrm{old}})$. The nonsmoothness of $\mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u}))$ implies that the least-squares functional $\mathcal{F}(\boldsymbol{\sigma}, \mathbf{u}; \boldsymbol{\sigma}^{\mathrm{old}}, \alpha^{\mathrm{old}})$ is also not differentiable for displacements which satisfy (3.17). This causes the Gauss–Newton iteration with a line search strategy (cf. [18, Chap. 10]) commonly used in least-squares finite element computations to slow down as plastic deformations become dominant. The issue of efficiently solving the nonlinear algebraic least-squares problems resulting from the discretization of (2.15) will be discussed elsewhere.

4. Plane strain model and finite element approximation. We restrict our computations in this paper to two-dimensional domains by assuming plane strain conditions, i.e.,

$$\boldsymbol{\varepsilon}(\mathbf{u}) = \begin{bmatrix} \partial_1 u_1 & (\partial_2 u_1 + \partial_1 u_2)/2 & 0 \\ (\partial_2 u_1 + \partial_1 u_2)/2 & \partial_2 u_2 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

This implies that

$$\mathrm{dev}(\boldsymbol{\varepsilon}(\mathbf{u})) = \begin{bmatrix} (2\partial_1 u_1 - \partial_2 u_2)/3 & (\partial_2 u_1 + \partial_1 u_2)/2 & 0 \\ (\partial_2 u_1 + \partial_1 u_2)/2 & (2\partial_2 u_2 - \partial_1 u_1)/3 & 0 \\ 0 & 0 & -(\partial_1 u_1 + \partial_2 u_2)/3 \end{bmatrix},$$

and therefore $\boldsymbol{\sigma} = \mathcal{R}(\boldsymbol{\varepsilon}(\mathbf{u}))$ is of the general form

$$\boldsymbol{\sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & 0 \\ \sigma_{21} & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}. \tag{4.1}$$

If we denote our two-dimensional domain again as Ω, then $\boldsymbol{\sigma}_1 = (\sigma_{11}, \sigma_{12}) \in H(\mathrm{div}, \Omega)$ and $\boldsymbol{\sigma}_2 = (\sigma_{21}, \sigma_{22}) \in H(\mathrm{div}, \Omega)$. Moreover, σ_{33} is constant in the x_3-direction and may be assumed in $L^2(\Omega)$. For the two remaining displacement components we still have $u_1, u_2 \in H^1_{\Gamma_D}(\Omega)$.

The choice of appropriate finite element spaces $\boldsymbol{\Sigma}_h$ and $\mathbf{U}_h$ for the approximation of $\boldsymbol{\sigma}$ and $\mathbf{u}$, respectively, is done with the aim of achieving a certain approximation order with respect to the norm in (3.6). Suitable for the stress approximation is a product space of Raviart–Thomas elements (of degree $k \geq 0$) for $\boldsymbol{\sigma}_1$ and $\boldsymbol{\sigma}_2$ and discontinuous piecewise polynomials (of the same degree $k \geq 0$) for σ_{33}. The interpolation estimate for Raviart–Thomas elements (cf. [9, Prop. III.3.9]) yields

$$\|\mathrm{div}\,(\boldsymbol{\sigma}_i - \boldsymbol{\Pi}_h \boldsymbol{\sigma}_i)\|^2 + \|\boldsymbol{\sigma}_i - \boldsymbol{\Pi}_h \boldsymbol{\sigma}_i\|^2 \leq \overline{C}^2 h^{2(k+1)} \left(|\mathrm{div}\,\boldsymbol{\sigma}_i|^2_{k+1,\Omega} + |\boldsymbol{\sigma}_i|^2_{k+1,\Omega}\right) \tag{4.2}$$

for $i = 1, 2$ with a suitable interpolation operator $\mathbf{\Pi}_h$. Standard piecewise polynomial interpolation, separately on each element $T \in \mathcal{T}_h$, leads to

$$\|\sigma_{33} - Q_h\sigma_{33})\|^2 \leq \overline{C}^2 h^{2(k+1)} |\sigma_{33}|^2_{k+1,\Omega} \tag{4.3}$$

with the corresponding interpolation operator Q_h. Finally, H^1-conforming finite elements which consist of piecewise polynomials of degree $k+1$ lead to

$$\|\nabla(\mathbf{u} - \mathbf{\Psi}_h\mathbf{u})\|^2 \leq \overline{C}^2 h^{2(k+1)} |\mathbf{u}|^2_{k+2,\Omega} \tag{4.4}$$

for the interpolation error. Combined with the result of Theorem 3.2, (4.2), (4.3), and (4.4) imply the error estimate

$$\begin{aligned}\Big(\|\mathrm{div}\,(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|^2 + \|\mathcal{C}^{-1/2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u} - \mathbf{u}_h)\|^2\Big)^{1/2} \\ \leq \left(\frac{\overline{\beta}}{\underline{\beta}}\right)^{1/2} \overline{C} h^{k+1} \left(|\mathrm{div}\,\boldsymbol{\sigma}_1|_{k+1,\Omega} + |\mathrm{div}\,\boldsymbol{\sigma}_2|_{k+1,\Omega} + |\boldsymbol{\sigma}|_{k+1,\Omega} + |\mathbf{u}|_{k+2,\Omega}\right)\end{aligned}$$

for the least-squares finite element approximation.

In particular, for $k = 1$, i.e, using next-to-lowest-order Raviart–Thomas elements combined with discontinuous piecewise linear elements for $\boldsymbol{\sigma}$ and continuous piecewise quadratic elements for $\mathbf{u}$, we obtain

$$\begin{aligned}\Big(\|\mathrm{div}\,(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|^2 + \|\mathcal{C}^{-1/2}(\boldsymbol{\sigma} - \boldsymbol{\sigma}_h)\|^2 + \|\mathcal{C}^{1/2}\boldsymbol{\varepsilon}(\mathbf{u} - \mathbf{u}_h)\|^2\Big)^{1/2} \\ \leq \left(\frac{\overline{\beta}}{\underline{\beta}}\right)^{1/2} \overline{C} h^2 \left(|\mathrm{div}\,\boldsymbol{\sigma}_1|_{2,\Omega} + |\mathrm{div}\,\boldsymbol{\sigma}_2|_{2,\Omega} + |\boldsymbol{\sigma}|_{2,\Omega} + |\mathbf{u}|_{3,\Omega}\right)\end{aligned} \tag{4.5}$$

for the least-squares finite element approximation. This is actually the combination of finite element spaces that we used in our computations presented in the next section. However, the regularity assumptions $\mathrm{div}\,\boldsymbol{\sigma}_i \in H^2(\Omega)$, $\boldsymbol{\sigma}_i \in H^2(\Omega)^2$ for $i = 1, 2$, $\sigma_{33} \in H^2(\Omega)$, and $\mathbf{u} \in H^3(\Omega)^2$ are rarely fulfilled in applications of practical relevance. The approximation estimate (4.5) therefore serves only as a guideline for the properties of the finite element spaces. In our actual computations the least-squares finite element method is implemented in an adaptive fashion based on (3.16) for a posteriori error estimation.

The implementation of the least-squares finite element method is done by evaluating the integrals in (2.14) with an appropriate quadrature rule. Since the finite element spaces used in our computations include piecewise polynomials up to degree 2, the integrands in the least-squares functional involve polynomials up to degree 4. A 7-point quadrature rule which is exact for polynomials of degree 5 on triangles (see [2, sect. 5.1]) is therefore used in our implementation. The return parameter $\gamma_R(\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\boldsymbol{\varepsilon}(\mathbf{u})))$ is computed at all quadrature nodes. The hardening parameter

$$\alpha = \alpha^{\mathrm{old}} + \sqrt{\frac{2}{3}} \frac{\gamma_R(\mathrm{dev}(\boldsymbol{\sigma}^{\mathrm{old}} + \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u})))}{2\mu}$$

(cf. (2.11)) is approximated by piecewise linear, not necessarily continuous, functions on the triangulation $\mathcal{T}_h$. This leads to the same order of approximation for the hardening parameter α as for the other process variables.

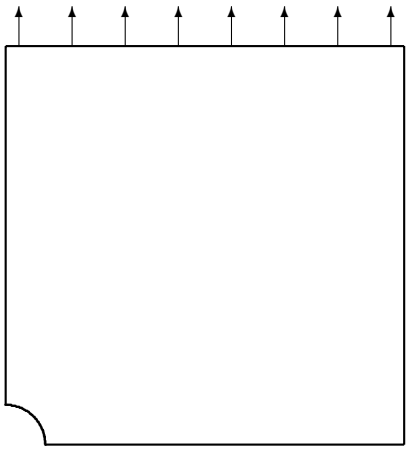


FIG. 5.1. *Computational domain and boundary conditions.*

5. Computational tests. In this section, numerical results for a benchmark problem of elasto-plasticity taken from [23] are presented. The problem to be considered is given by a quadratic plate of an elasto-plastic isotropic material with a circular hole in the center under plane strain conditions. At the upper and lower edges of the plate, traction forces pointing outwards are applied. Because of the symmetry of the domain, it suffices to discretize only a fourth of the total geometry. The computational domain is then given by

$$\Omega = \{\mathbf{x} \in \mathbb{R}^2 : 0 < x_1 < 10, 0 < x_2 < 10, x_1^2 + x_2^2 > 1\}$$

(see Figure 5.1). The boundary conditions on the top edge of the computational domain ($x_2 = 10, 0 < x_1 < 10$) are set to $\boldsymbol{\sigma} \cdot \mathbf{n} = (0, t)^T$, while on the right edge ($x_1 = 10, 0 < x_2 < 10$) and on the circular arc ($x_1^2 + x_2^2 = 1$) the boundary conditions are set to $\boldsymbol{\sigma} \cdot \mathbf{n} = (0, 0)$. Symmetry boundary conditions are prescribed on the rest of the boundary, i.e., $(\sigma_{11}, \sigma_{12}) \cdot \mathbf{n} = 0, u_2 = 0$ on the bottom ($x_2 = 0, 1 < x_1 < 10$), and $u_1 = 0, (\sigma_{21}, \sigma_{22}) \cdot \mathbf{n} = 0$ on the left ($x_1 = 0, 1 < x_2 < 10$). The Poisson ratio is $\nu = 0.29$, which implies for the Lamé constants $\lambda = 1.381\,\mu$. Actually $\mu = 1$ is set in our computations for simplicity, since the stress values do not depend on the size of μ.

Example 1. Our first set of computational experiments uses a combination of linear and exponential isotropic hardening of the form

$$K(\alpha) = K_0 + H\alpha + (K_\infty - K_0)(1 - e^{-\omega\alpha})$$

with $K_0 = 450$, $K_\infty = 750$, $H = 129$, and $\omega = 16.93$ taken from [23]. The load is increased starting from $t = 0$ in steps of $\Delta t = 2.5$. These rather small load steps were chosen in order to rule out artifacts caused by the first-order time discretization used in our computations. For each load step, an initial triangulation consisting of 52 elements is successively refined based on the local evaluation of the least-squares functional. Tables 5.1 to 5.3 show the error reduction, measured in terms of the functional, at different stages of the simulation. For $t = 150$, the results are still well within the elastic domain, which means that the results in Table 5.1 simply correspond to a linear elasticity problem. Inelastic deformation starts around $t = 170$ so that

TABLE 5.1
Plasticity with hardening: Reduction of the least-squares functional for $t = 150$.

l	1	2	3	4	5	6
# elements	113	224	452	917	1829	3610
dim $\mathbf{U}_h$	474	926	1858	3738	7440	14628
dim $\mathbf{\Sigma}_h$	1447	2882	5826	11851	23653	46742
$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h)$	3.04 e-6	5.77 e-7	1.37 e-7	4.04 e-8	1.05 e-8	3.00 e-9
$\|\mathrm{as}\,\boldsymbol{\sigma}_h\|^2$	7.48 e-3	1.43 e-3	3.32 e-4	9.36 e-5	2.37 e-5	6.57 e-6
$\|\mathrm{div}\,\boldsymbol{\sigma}_h\|^2$	7.45 e-8	9.92 e-9	7.68 e-10	1.43 e-10	1.64 e-11	2.39 e-12

TABLE 5.2
Plasticity with hardening: Reduction of the least-squares functional for $t = 300$.

l	1	2	3	4	5	6
# elements	113	224	456	923	1905	3953
dim $\mathbf{U}_h$	474	926	1874	3760	7742	15998
dim $\mathbf{\Sigma}_h$	1447	2882	5878	11931	24643	51203
$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h)$	4.02 e-6	2.65 e-6	1.62 e-6	3.19 e-7	3.45 e-8	6.39 e-9
$\|\mathrm{as}\,\boldsymbol{\sigma}_h\|^2$	2.90 e-2	5.55 e-3	1.32 e-3	4.02 e-4	1.10 e-4	3.33 e-5
$\|\mathrm{div}\,\boldsymbol{\sigma}_h\|^2$	8.91 e-8	2.34 e-8	2.58 e-9	2.06 e-10	2.54 e-11	3.06 e-12

TABLE 5.3
Plasticity with hardening: Reduction of the least-squares functional for $t = 400$.

l	1	2	3	4	5	6
# elements	113	226	459	923	1932	4002
dim $\mathbf{U}_h$	474	936	1884	3764	7834	16166
dim $\mathbf{\Sigma}_h$	1447	2906	5919	11927	25010	51868
$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h)$	5.78 e-5	1.87 e-5	5.60 e-6	8.20 e-7	9.76 e-8	2.22 e-8
$\|\mathrm{as}\,\boldsymbol{\sigma}_h\|^2$	4.49 e-2	1.07 e-2	3.31 e-3	9.25 e-4	3.40 e-4	1.09 e-4
$\|\mathrm{div}\,\boldsymbol{\sigma}_h\|^2$	2.18 e-6	3.37 e-7	2.52 e-8	3.01 e-10	8.15 e-11	1.02 e-11

the results in Table 5.2 already correspond to elasto-plastic computations for $t = 300$. Further increase of the load to $t = 400$ in Table 5.3 leads to a spreading of the zone in which inelastic deformations occur.

In Tables 5.1 to 5.3, the computational results show that the reduction rate of the functional does not deteriorate much as the load is increased and inelastic deformations become more dominant. The optimal convergence behavior achievable with the finite element spaces used here would result in a reduction of the least-squares functional proportional to $(\dim \mathbf{U}_h + \dim \mathbf{\Sigma}_h)^{-2}$. This behavior would be achieved with uniformly refined triangulations under sufficient regularity conditions (see (4.5); note that $\dim \mathbf{U}_h + \dim \mathbf{\Sigma}_h \approx h^{-2}$ in two space dimensions). In Tables 5.1 to 5.3, the number of degrees of freedom is approximately doubled with each refinement. This corresponds to a reduction of the functional by a factor 4 with each refinement, which is approximately achieved in our numerical results, at least on the finer levels.

The antisymmetry, measured by $\|\mathrm{as}\,\boldsymbol{\sigma}_h\|^2_{0,\Omega}$, is also shown in Tables 5.1 to 5.3. Note that (3.8) implies

$$\|\mathrm{as}(\boldsymbol{\sigma}_h - \boldsymbol{\sigma}_h^{\mathrm{old}})\|^2 \le 2\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h; \boldsymbol{\sigma}_h^{\mathrm{old}}, \alpha^{\mathrm{old}})$$

but that the antisymmetry actually accumulates with time. In the elastic case, for $t = 150$, the reduction occurs at about the same rates for the antisymmetric stress and for the least-squares functional. However, in the presence of plastic deformations, for $t = 300$ and $t = 400$, the antisymmetry is actually reduced at a faster rate than the functional. This is due to the fact that the antisymmetric stress is actually an

TABLE 5.4
Perfect plasticity: Reduction of the least-squares functional for $t = 300$.

l	1	2	3	4	5	6
# elements	120	253	502	967	1847	3667
dim $\mathbf{U}_h$	502	1044	2052	3936	7484	14832
dim $\mathbf{\Sigma}_h$	1538	3257	6482	12503	23915	47507
$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h)$	1.26 e-4	3.40 e-5	8.02 e-5	2.29 e-6	5.31 e-7	1.23 e-7
$\|\text{as } \boldsymbol{\sigma}_h\|^2$	3.62 e-2	7.82 e-3	2.22 e-3	5.99 e-4	2.41 e-4	8.57 e-5
$\|\text{div } \boldsymbol{\sigma}_h\|^2$	8.00 e-6	7.34 e-7	6.12 e-8	6.76 e-9	6.24 e-10	3.74 e-11

TABLE 5.5
Perfect plasticity: Reduction of the least-squares functional for $t = 400$.

l	1	2	3	4	5	6
# elements	117	256	548	1249	2779	6153
dim $\mathbf{U}_h$	492	1054	2228	5046	11206	24752
dim $\mathbf{\Sigma}_h$	1497	3298	7088	16187	36037	79849
$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h)$	5.04 e-4	1.88 e-4	6.26 e-5	1.94 e-5	5.68 e-6	3.07 e-6
$\|\text{as } \boldsymbol{\sigma}_h\|^2$	8.39 e-2	2.48 e-2	1.50 e-2	8.72 e-3	2.31 e-3	8.29 e-4
$\|\text{div } \boldsymbol{\sigma}_h\|^2$	1.30 e-4	3.19 e-5	3.36 e-6	7.24 e-7	5.48 e-8	2.39 e-8

TABLE 5.6
Perfect plasticity: Reduction of the least-squares functional for $t = 450$.

l	1	2	3	4	5	6
# elements	115	258	558	1220	2649	5638
dim $\mathbf{U}_h$	482	1062	2276	4940	10676	22668
dim $\mathbf{\Sigma}_h$	1473	3324	7210	15800	34357	73178
$\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h)$	3.21 e-3	3.01 e-3	1.34 e-3	5.91 e-4	2.71 e-4	1.17 e-4
$\|\text{as } \boldsymbol{\sigma}_h\|^2$	2.33 e-1	9.89 e-2	3.78 e-2	1.36 e-2	4.02 e-3	2.51 e-3
$\|\text{div } \boldsymbol{\sigma}_h\|^2$	1.85 e-3	1.63 e-3	6.45 e-4	2.56 e-4	7.05 e-5	1.88 e-5

accumulated quantity, while the least-squares functional shows the full deterioration of the functional with increasing load. It can also be observed from Tables 5.1 to 5.3 that the divergence error, measured by $\|\text{div } \boldsymbol{\sigma}_h\|^2$, decreases slightly faster than the overall functional.

Example 2. We also include numerical results for perfect plasticity, since tabulated benchmark values are available from [23] for this case. This allows us to verify the least-squares approach by a comparison of our results with the benchmark values which will be done further below. In this setting,

$$K(\alpha) \equiv K_0$$

with $K_0 = 450$, which means that the internal hardening variable α is obsolete. For perfect plasticity, the conditions (2.12) are not valid, and therefore the theoretical results from section 3 are not established.

With the same load steps as in Example 1 we obtain the numerical results shown in Tables 5.4 to 5.6. The results for $t = 150$ are, of course, identical to those in Table 5.1, since this still constitutes the same elastic problem as in Example 1. For $t = 300$, a reduction rate which nearly reaches the optimal asymptotic behavior of $\mathcal{F}(\boldsymbol{\sigma}_h, \mathbf{u}_h) \eqsim (\dim \mathbf{U}_h + \dim \mathbf{\Sigma}_h)^{-2}$ is attained. The reduction of the functional is much slower at the load step $t = 400$ and slows down even more for $t = 450$.

Figure 5.2 shows on the left the size of the deviatoric stress, scaled as $|\text{dev}(\boldsymbol{\sigma})|/K_0$, for the load steps $t = 300$, 400, and 450. The zone where plastic deformation occurs is clearly visible and expands with increasing load. The fact that the reduction of the

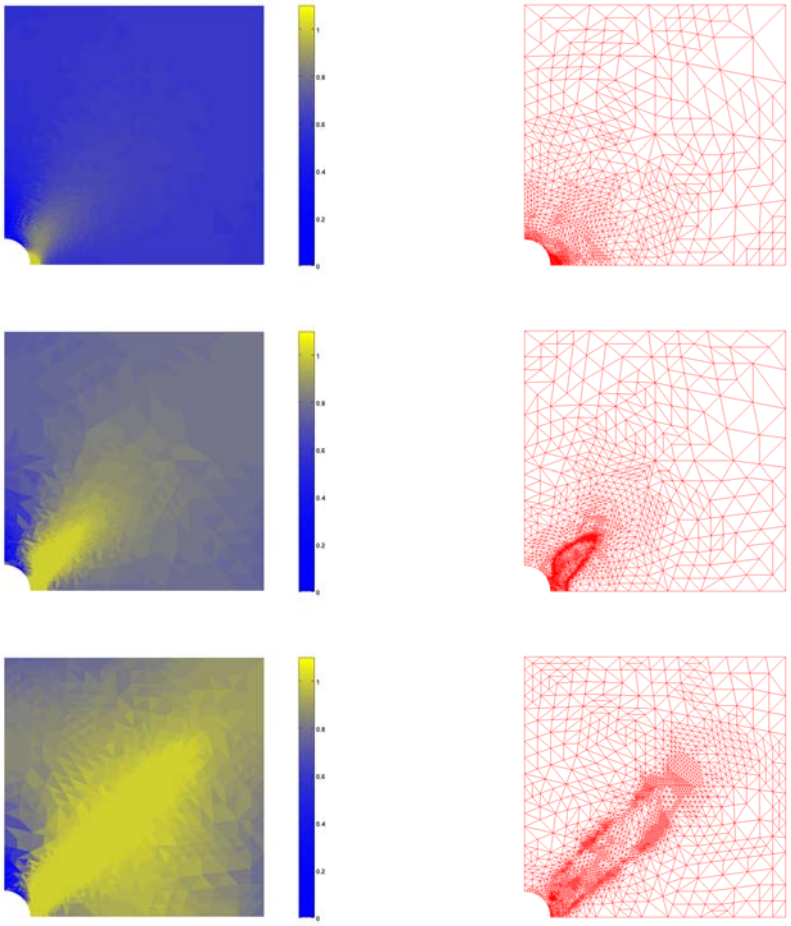


FIG. 5.2. *Deviatoric stress (left) and triangulation after six adaptive refinement steps (right) for* $t = 300$, 400, *and* 450.

functional slows down significantly as the plastic zone occupies most of the computational domain supports our speculation that this causes the ellipticity of the least-squares functional in the sense of Theorem 3.2 to deteriorate. Shown on the right in Figure 5.2 are the triangulations which result after six steps of adaptive refinement based on the elementwise evaluation of the least-squares functional.

Despite this deterioration of the convergence behavior, our results obtained with the least-squares method agree remarkably well with the benchmark results tabulated in [23]. In order to illustrate this, the benchmark results for a selected stress value, $s_{22}(1, 0)$, are shown for a full load cycle in Figure 5.3. The load cycle starts by increasing the traction forces t from 0 to 450 (pointing outwards), then decreasing from 450 to -450, and finally increasing from -450 to 0 again. The solid curve is the result of our computations, and the circles represent the values for the reference solution taken from Table 11.8 in [23]. Throughout the load cycle the difference between our results and the benchmark values is marginally small. Even after the completion of the cycle, our least-squares method gives a value of 514.38 for $s_{22}(1, 0)$

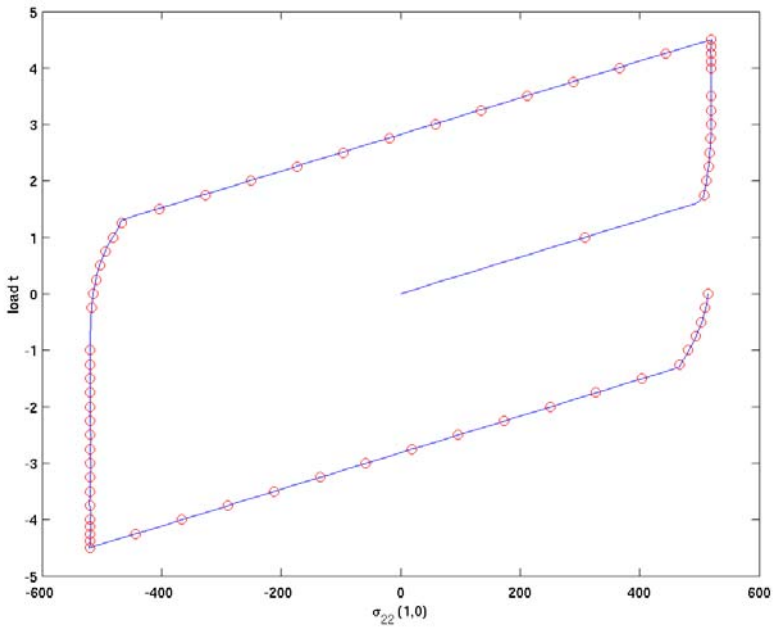


FIG. 5.3. *Stress values for one complete load cycle.*

compared to the reference solution of 513.93.

REFERENCES

[1] J. ALBERTY, C. CARSTENSEN, AND D. ZARRABI, *Adaptive numerical analysis in primal elasto-plasticity with hardening*, Comput. Methods Appl. Mech. Engrg., 171 (1999), pp. 175–204.
[2] K. E. ATKINSON, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.
[3] O. AXELSSON, R. BLAHETA, AND R. KOHUT, *Inexact Newton solvers in plasticity: Theory and experiments*, Numer. Linear Algebra Appl., 4 (1997), pp. 133–152.
[4] F.-J. BARTHOLD, M. SCHMIDT, AND E. STEIN, *Error indicators and mesh refinements for finite element computations of elastoplastic deformations*, Comput. Mech., 22 (1998), pp. 225–238.
[5] M. BERNDT, T. A. MANTEUFFEL, AND S. F. MCCORMICK, *Local error estimates and adaptive refinement for first-order system least squares*, Electron Trans. Numer. Anal., 6 (1997), pp. 35–43.
[6] R. BLAHETA, *Convergence of Newton-type methods in incremental return mapping analysis of elasto-plastic problem*, Comput. Methods Appl. Mech. Engrg., 147 (1997), pp. 167–185.
[7] P. B. BOCHEV AND M. D. GUNZBURGER, *Finite element methods of least-squares type*, SIAM Review, 40 (1998), pp. 789–837.
[8] D. BRAESS, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.
[9] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
[10] J. BÜTTNER AND B. SIMEON, *Runge–Kutta methods in elastoplasticity*, Appl. Numer. Math., 41 (2002), pp. 443–458.
[11] J. BÜTTNER AND B. SIMEON, *Time integration of the dual problem of elastoplasticity by Runge–Kutta methods*, SIAM J. Numer. Anal., 41 (2003), pp. 1564–1584.
[12] Z. CAI, J. KORSAWE, AND G. STARKE, *An adaptive least squares mixed finite element method*

for the stress-displacement formulation of linear elasticity, Numer. Methods Partial Differential Equations, 21 (2005), pp. 132–148.

[13] Z. Cai and G. Starke, *First-order system least squares for the stress-displacement formulation: Linear elasticity*, SIAM J. Numer. Anal., 41 (2003), pp. 715–730.

[14] Z. Cai and G. Starke, *Least-squares methods for linear elasticity*, SIAM J. Numer. Anal., 42 (2004), pp. 826–842.

[15] W. Han and B. D. Reddy, *Plasticity: Mathematical Theory and Numerical Analysis*, Springer, New York, 1999.

[16] S. Lang, C. Wieners, and G. Wittum, *The applications of adaptive parallel multigrid methods to problems in nonlinear solid mechanics*, in Error-Controlled Adaptive Finite Elements in Solid Mechanics, E. Stein, ed., John Wiley and Sons, New York, 2002, pp. 347–384.

[17] P. Neff and C. Wieners, *Comparison of models for finite plasticity: A numerical study*, Comput. Vis. Sci., 6 (2003), pp. 23–35.

[18] J. Nocedal and S. J. Wright, *Numerical Optimization*, Springer, New York, 1999.

[19] R. Rannacher and F.-T. Suttmeier, *A posteriori error estimation and mesh adaptation for finite element models in elasto-plasticity*, Comput. Methods Appl. Mech. Engrg., 176 (1999), pp. 333–361.

[20] R. Rannacher and F.-T. Suttmeier, *Error estimation and adaptive mesh design for fe models in elasto-plasticity theory*, in Error-Controlled Adaptive Finite Elements in Solid Mechanics, E. Stein, ed., John Wiley and Sons, New York, 2002, pp. 5–52.

[21] M. Rappaz, M. Bellet, and M. Deville, *Numerical Modeling in Materials Science and Engineering*, Springer, Berlin, 2003.

[22] J. C. Simo and T. J. R. Hughes, *Computational Inelasticity*, Springer, New York, 1998.

[23] E. Stein, P. Wriggers, A. Rieger, and M. Schmidt, *Benchmarks*, in Error-Controlled Adaptive Finite Elements in Solid Mechanics, E. Stein, ed., John Wiley and Sons, New York, 2002, pp. 385–404.

[24] C. Wieners, *Multigrid methods for Prandtl–Reuss plasticity*, Numer. Linear Algebra Appl., 6 (1999), pp. 457–478.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 389–420

FINITE ELEMENT METHODS FOR THE SIMULATION OF WAVES IN COMPOSITE SATURATED POROVISCOELASTIC MEDIA*

JUAN E. SANTOS† AND DONGWOO SHEEN‡

Abstract. This work presents and analyzes a collection of finite element procedures for the simulation of wave propagation in a porous medium composed of two weakly coupled solids saturated by a single-phase fluid. The equations of motion, formulated in the space-frequency domain, include dissipation due to viscous interaction between the fluid and solid phases with a correction factor in the high-frequency range and intrinsic anelasticity of the solids modeled using linear viscoelasticity. This formulation leads to the solution of a Helmholtz-type boundary value problem for each temporal frequency. For the spatial discretization, nonconforming finite element spaces are employed for the solid phases, while for the fluid phase the vector part of the Raviart–Thomas–Nedelec mixed finite element space is used. Optimal a priori error estimates for *global* standard and *hybridized* Galerkin finite element procedures are derived. An iterative nonoverlapping domain decomposition procedure is also presented and convergence results are derived. Numerical experiments showing the application of the numerical procedures to simulate wave propagation in partially frozen porous media are presented.

Key words. poroviscoelasticity, finite element method, error estimate, domain decomposition

AMS subject classifications. 65C20, 65N30, 65N55, 86-08

DOI. 10.1137/050629069

1. Introduction.

Wave propagation in composite porous materials has applications in many branches of science and technology, such as seismic methods in the presence of shaley sandstones [8], frozen or partially frozen sandstones [29, 10, 11], gas-hydrates in ocean-bottom sediments [12], and evaluation of the freezing conditions of foods by ultrasonic techniques [26]. A recent review of the theory of wave propagation in fluid-saturated porous media can be found in [7].

A theory to describe wave propagation in frozen porous media was first presented by Leclaire, Cohen-Tenoudji, and Aguirre Puente [24]. This model, valid for uniform porosity, predicts the existence of three compressional and two shear waves; the verification that additional (slow) waves can be observed in laboratory experiments was published by Leclaire, Cohen-Tenoudji, and Aguirre Puente [25]. Later, Carcione and Tinivella [12] generalized this theory to include the interaction between the solid and ice particles and grain cementation with decreasing temperature, used as a parameter to determine the bulk water content. Also, Carcione, Gurevich, and Cavallini [8] applied this theory to study the acoustic properties of shaley sandstones, assuming that sand and clay are *nonwelded* and form a continuous and interpenetrating porous composite skeleton. Both frozen porous media and shaley sandstones are examples of porous materials where the two solid phases are *weakly coupled* or *nonwelded*, i.e., both solids form a continuous and interacting composite structure, interchanging mechan-

*Received by the editors April 12, 2005; accepted for publication (in revised form) May 26, 2006; published electronically February 15, 2007.

http://www.siam.org/journals/sinum/45-1/62906.html

†CONICET, Departamento de Geofísica Aplicada, Facultad de Ciencias Astronómicas y Geofísicas, UNLP, Paseo del Bosque S/N, La Plata, 1900, Argentina, and Purdue University, West Lafayette, IN 47907 (santos@fcaglp.unlp.edu.ar).

‡Department of Mathematics, Seoul National University, Seoul 151–747, Korea (sheen@snu.ac.kr).
The research of this author was supported in part by Korea Research Foundation grant KRF-2004-C00007 and Korea Science and Engineering Foundation grant KOSEF R14-2003-019-01000-0.

ical energy. Similar *weakly coupled* formulations have previously been proposed. For instance, McCoy [28] has proposed a mixture theory appropriate for the combination of two *acoustic phases*.

Later, Santos, Ravazzoli, and Carcione [37] generalized to the nonuniform porosity case the models of Leclaire, Cohen-Tenoudji, and Aguirre Puente [24] and Carcione and Tinivella [12] valid only for uniform porosity. The formulation presented in [37] enabled us to identify the generalized coordinates of the system, which are the two solid displacement vectors and a new variable (denoted by $u^{(2)}$ in this paper) associated with the fluid displacement relative to the solid composite matrix, whose divergence is the change in the fluid content, in formal analogy with the classical Biot theory for a single solid-phase matrix. It also allowed us to identify the generalized forces of the system, which are the fluid pressure p_f and the stress tensors denoted by $\sigma^{(1)}$ and $\sigma^{(3)}$ in this paper.

This article presents a differential and numerical model to describe wave propagation in a heterogeneous poroviscoelastic frame consisting of two weakly coupled solid phases saturated by a single-phase fluid. The equations of motion, stated in the space-frequency domain, generalize those presented in [37, 9] by the inclusion of solid matrix dissipation using a linear viscoelastic model and introducing a frequency dependent correction factor in the mass and viscous coupling coefficients in the high-frequency range [4, 35].

The numerical procedures presented employ the nonconforming rectangular element defined in [17] to approximate the displacement vector in the solid phases. The dispersion analysis presented in [38] shows that employing this nonconforming element allows for a reduction in the number of points per wavelength necessary to reach a desired accuracy. On the other hand, the displacement in the fluid phase is approximated by using the vector part of the Raviart–Thomas–Nedelec mixed finite element space of zero order, which is a conforming space [34, 30]. The error analysis yields optimal a priori error estimates for the *global* standard and *hybridized* Galerkin methods.

Numerical simulation of waves in porous media is computationally expensive due to a large number of degrees of freedom needed to calculate wave fields accurately; the use of a domain decomposition iteration is a convenient approach to overcome this difficulty. Here we define a nonoverlapping domain decomposition iterative scheme and derive convergence results similar to those presented in [14] for solving second-order elliptic problems. This iterative procedure was used for the simulation of waves in a sample of water-saturated partially frozen Berea sandstone [9, 12], perturbed by a point source at seismic frequencies. The sample has an interior plane interface defined by a change in ice content in the pores, and the snapshots of the generated wave fields show clearly the events associated with the different types of waves.

2. The differential model. In this section we review and generalize a model recently presented by one of the authors and some of his colleagues [37] to describe the propagation of waves in a poroviscoelastic domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, in which the matrix consists of two different solids indicated by the superindices (1) and (3), saturated by a single-phase fluid indicated by the superindex (2). Thus, for any reference element E of bulk material we have

$$E = E^{(1)} \cup E^{(2)} \cup E^{(3)}.$$

Let $V^{(i)}$ denote the volumetric measure of the phase $E^{(i)}$ and let $V^{(b)}$ and $V^{(sm)}$ denote the volumetric measures of E and the solid matrix $E^{(sm)} = E^{(1)} \cup E^{(3)}$,

respectively, so that

$$V^{(sm)} = V^{(1)} + V^{(3)}, \qquad V^{(b)} = V^{(1)} + V^{(2)} + V^{(3)}.$$

We introduce the bulk volumetric fractions of the different components in the form

$$\phi = \frac{V^{(2)}}{V^{(b)}}, \quad \phi^{(1)} = \frac{V^{(1)}}{V^{(b)}}, \quad \phi^{(3)} = \frac{V^{(3)}}{V^{(b)}},$$

and the solid fractions of the composite matrix

$$S^{(1)} = \frac{V^{(1)}}{V^{(sm)}}, \quad S^{(3)} = \frac{V^{(3)}}{V^{(sm)}}, \quad \text{with} \quad S^{(1)} + S^{(3)} = 1.$$

For some practical applications it is convenient to define the *absolute* or *effective porosity* $\phi^{(a)}$ of the medium, defined as the ratio of the volume of the interconnected pores $V^{(p)}$ and the total volume of the sample, i.e.,

$$\phi^{(a)} = \frac{V^{(p)}}{V^{(b)}}.$$

These sets of fractions can have different meanings depending on the physical model considered. For example, in the case of a sandstone or soil at very low temperature, it is reasonable to consider that a part of the fluid which saturates the pore space is at a liquid state and the rest is frozen. If $E^{(1)}$ represents the mineral grains and $E^{(3)}$ the ice, for a given porosity $\phi^{(a)}$ and *bulk water content* ϕ, the following relations hold:

$$\phi^{(1)} = 1 - \phi^{(a)}, \quad \phi^{(3)} = \phi^{(a)} - \phi, \quad S^{(3)} = \frac{\phi^{(3)}}{1 - \phi}. \tag{2.1}$$

It is useful to introduce an additional fraction $S^{(3)'}$ to account for the *ice content in the pores*, given by

$$S^{(3)'} = \frac{V^{(3)}}{V^{(p)}} = \frac{\phi^{(3)}}{1 - \phi^{(1)}}.$$

A different application of this model would be the case of a shaley sandstone, that is, a porous rock mainly composed of quartz grains and clay particles, saturated by a fluid (such as water, brine, gas, or oil). In this case we assume that the fluid completely saturates the pore space of the composite rock so that $V^{(2)} \equiv V^{(p)}$. Then, if $E^{(1)}$ represents the grains of the rock and $E^{(3)}$ the clay part, for a given *matrix clay content* $S^{(3)}$ and water content ϕ, instead of (2.1) the following hold:

$$\phi = \phi^{(a)}, \quad \phi^{(1)} = S^{(1)}(1 - \phi), \quad \phi^{(3)} = S^{(3)}(1 - \phi).$$

Let us now consider a unit cube $\Omega = \Omega^{(1)} \cup \Omega^{(2)} \cup \Omega^{(3)} \subset \mathbb{R}^d$ of our fluid-saturated poroviscoelastic material with boundary $\Gamma = \partial\Omega$. Since by hypothesis the two solids are nonwelded (or weakly coupled), we assume that they can move independently and consequently we can distinguish three different particle displacement fields for this model. Let $u^{(m)} \equiv u^{(m)}(x,\omega) = (u_1^{(m)}(x,\omega), \ldots, u_d^{(m)}(x,\omega))^t$, $m = 1, 3$, be the averaged solid displacements over the bulk material Ω at the angular frequency ω, and let $\tilde{u}^{(2)} \equiv \tilde{u}^{(2)}(x,\omega) = (\tilde{u}_1^{(2)}(x,\omega), \ldots, \tilde{u}_d^{(2)}(x,\omega))^t$ denote the absolute fluid

displacement. Also, let the relative displacement of the fluid phase with respect to the composite solid matrix be defined by

$$u^{(2)} = \phi(\tilde{u}^{(2)} - S^{(1)}u^{(1)} - S^{(3)}u^{(3)})$$

and set $u = (u^{(1)}, u^{(2)}, u^{(3)})^t$. As explained in [37], the variable

$$\zeta = -\nabla \cdot u^{(2)}$$

represents the change in fluid content. Next we introduce the local stress tensors $\sigma_{jk}^{(1,s)}$ and $\sigma_{jk}^{(3,s)}$ in the solid parts $\Omega^{(1)}$ and $\Omega^{(3)}$, averaged over the bulk material and the fluid pressure p_f. Following [37], we define the second-order tensors

$$\sigma_{jk}^{(1)} = \sigma_{jk}^{(1,s)} - S^{(1)}\phi p_f \delta_{jk}, \quad \sigma_{jk}^{(3)} = \sigma_{jk}^{(3,s)} - S^{(3)}\phi p_f \delta_{jk}$$

associated with the total stresses in $\Omega^{(1)}$ and $\Omega^{(3)}$, respectively. Then the constitutive equations, stated in the space-frequency domain, are as follows [37]:

$$\sigma_{jk}^{(1)}(u) = \big[K_G^{(1)} e^{(1)} - B^{(1)}\zeta + B^{(3)} e^{(3)}\big]\delta_{jk} + 2\mu^{(1)} d_{jk}^{(1)} + \mu^{(13)} d_{jk}^{(3)}, \tag{2.2a}$$

$$\sigma_{jk}^{(3)}(u) = \big[K_G^{(3)} e^{(3)} - B^{(2)}\zeta + B^{(3)} e^{(1)}\big]\delta_{jk} + 2\mu^{(3)} d_{jk}^{(3)} + \mu^{(13)} d_{jk}^{(1)}, \tag{2.2b}$$

$$p_f(u) = -B^{(1)} e^{(1)} - B^{(2)} e^{(3)} + K_{av}\zeta, \tag{2.2c}$$

where

$$d_{jk}^{(m)} = \epsilon_{jk}(u^{(m)}) - \frac{1}{d}\, e^{(m)}\delta_{jk}, \ m = 1, 3, \text{ in } \mathbb{R}^d,$$

denotes the deviatoric tensor in $\Omega^{(m)}$, with $\epsilon_{jk}(u^{(m)})$ being the strain tensor with linear invariant $e^{(m)}$. In [37] the constitutive relations (2.2) were stated in the space-time domain with real coefficients in terms of the bulk and shear moduli of the two solid (dry) frames (denoted by $K_m^{(s1)}, K_m^{(s3)}, \mu_m^{(s1)}$, and $\mu_m^{(s3)}$, respectively), the bulk and shear moduli of the grains in the two solid phases (denoted by $K^{(s1)}, \mu^{(s1)}, K^{(s3)}$, and $\mu^{(s3)}$, respectively), and K_f, the bulk modulus of the fluid phase.

To introduce viscoelasticity we use the correspondence principle stated by M. Biot [3, 5]; i.e., we replace the real poroelastic coefficients $K_G^{(m)}, \mu^{(m)}$, $m = 1, 3$, and K_{av} in the constitutive relations (2.3a)–(2.3c) by complex frequency dependent poroviscoelastic moduli satisfying the same relations as in the elastic case. In this work the linear viscoelastic model presented in [27] is used to make these moduli complex and frequency dependent by using the following formula:

$$M(\omega) = \frac{M_{re}}{R_M(\omega) - iT_M(\omega)},$$

where M represents any of the five moduli mentioned above and the coefficient M_{re} is the relaxed elastic modulus associated with M [6]. The functions $R_M(\omega)$ and $T_M(\omega)$, associated with a continuous spectrum of relaxation times, are given by [27]

$$R_M(\omega) = 1 - \frac{1}{\pi \widehat{Q}_M} \ln \frac{1 + \omega^2 T_{1,M}^2}{1 + \omega^2 T_{2,M}^2}, \qquad T_M(\omega) = \frac{2}{\pi \widehat{Q}_M} \tan^{-1} \frac{\omega(T_{1,M} - T_{2,M})}{1 + \omega^2 T_{1,M} T_{2,M}}.$$

The model parameters $\widehat{Q}_M$, $T_{1,M}$, and $T_{2,M}$ are taken such that the quality factors $Q_M(\omega) = T_M/R_M$ are approximately equal to the constant $\widehat{Q}_M$ in the range of

frequencies where the equations are solved, which makes this model convenient for geophysical applications.

Next, by writing

$$\lambda^{(m)} = K_G^{(m)} - \frac{2}{d}\mu^{(m)}, \quad D^{(3)} = B^{(3)} - \frac{1}{d}\mu^{(13)} \text{ in } \mathbb{R}^d,$$

the constitutive relations (2.2) are then stated in the following equivalent form, which will be used in the analysis that follows:

$$\sigma_{jk}^{(1)}(u) = \left[\lambda^{(1)}e^{(1)} - B^{(1)}\zeta + D^{(3)}e^{(3)}\right]\delta_{jk} + 2\mu^{(1)}\epsilon_{jk}(u^{(1)}) + \mu^{(13)}\epsilon_{jk}(u^{(3)}), \tag{2.3a}$$

$$\sigma_{jk}^{(3)}(u) = \left[\lambda^{(3)}e^{(3)} - B^{(2)}\zeta + D^{(3)}e^{(1)}\right]\delta_{jk} + 2\mu^{(3)}\epsilon_{jk}(u^{(3)}) + \mu^{(13)}\epsilon_{jk}(u^{(1)}), \tag{2.3b}$$

$$p_f(u) = -B^{(1)}e^{(1)} - B^{(2)}e^{(3)} + K_{av}\zeta. \tag{2.3c}$$

Let the positive definite mass matrix $\mathcal{P} = \mathcal{P}(\omega)$ and the nonnegative dissipation matrix $\mathcal{B} = \mathcal{B}(\omega)$ be defined by

$$\mathcal{P} = \begin{bmatrix} p_{11}I & p_{12}I & p_{13}I \\ p_{12}I & p_{22}I & p_{23}I \\ p_{13}I & p_{23}I & p_{33}I \end{bmatrix}, \qquad \mathcal{B} = \begin{bmatrix} b_{11}I & -b_{12}I & -b_{11}I \\ -b_{12}I & b_{22}I & b_{12}I \\ -b_{11}I & b_{12}I & b_{11}I \end{bmatrix},$$

where I denotes the identity matrix in $\mathbb{R}^{d\times d}$. The nonnegative coefficients $p_{jk} = p_{jk}(\omega), b_{jk} = b_{jk}(\omega)$ in the definition of the matrices $\mathcal{P}$ and $\mathcal{B}$ are given by the formulae

$$p_{11}(\omega) = m_{11} + \frac{F_I(\theta)\,(f_{11} - b_{13})}{\omega}, \quad p_{12}(\omega) = m_{12} - \frac{F_I(\theta)f_{12}}{\omega}, \tag{2.4a}$$

$$p_{13}(\omega) = m_{13} - \frac{F_I(\theta)\,(f_{11} - b_{13})}{\omega}, \quad p_{22}(\omega) = m_{22} + \frac{F_I(\theta)f_{22}}{\omega}, \tag{2.4b}$$

$$p_{23}(\omega) = m_{23} + \frac{F_I(\theta)f_{12}}{\omega}, \quad p_{33}(\omega) = p_{33} + \frac{F_I(\theta)\,(f_{11} - b_{13})}{\omega}, \tag{2.4c}$$

$$b_{11}(\omega) = F_R(\theta)\,(f_{11} - b_{13}) + b_{13}, \quad b_{12}(\omega) = F_R(\theta)f_{12}, \quad b_{22}(\omega) = F_R(\theta)f_{22}, \tag{2.4d}$$

with the m_{ij}'s and f_{11}, f_{12}, and f_{22} computed as in [37] in terms of the mass densities $\rho^{(m)}, m = 1, 2, 3$, of each solid and fluid constituent, the fluid viscosity η, and the absolute permeabilities κ_1, κ_3 of the two solid frames. The coefficient b_{13} is a friction coefficient between the two solid phases and is left as a free parameter chosen so that

$$b_{11}b_{22} - b_{12}^2 > 0, \quad \omega > 0, \tag{2.5}$$

which is needed in order that the dissipation function be positive in the variables $u^{(2)}$ and $u^{(1)} - u^{(3)}$.

The complex valued frequency dependent function $F(\theta) = F_R(\theta) + iF_I(\theta)$ is the frequency correction function defined by Biot [4]:

$$F(\theta) = \frac{1}{4}\,\frac{\theta T(\theta)}{1 - \frac{2}{i\theta}T(\theta)}, \quad T(\theta) = \frac{\mathrm{ber}'(\theta) + i\mathrm{bei}'(\theta)}{\mathrm{ber}(\theta) + i\mathrm{bei}(\theta)},$$

with $\mathrm{ber}(\theta)$ and $\mathrm{bei}(\theta)$ being the Kelvin functions of the first kind and zero order. The frequency dependent argument $\theta = \theta(\omega)$ is given in terms of the pore size parameter a_p by the following equations:

$$\theta = a_p\sqrt{\omega\,\rho^{(2)}/\eta}, \quad a_p = 2\sqrt{\kappa A_0/\phi},$$

where $\frac{1}{\kappa} = \frac{1}{\kappa_1} + \frac{1}{\kappa_3}$ and A_0 is the Kozeny–Carman constant [2, 22]. This frequency correction is needed to include the departure of the relative flow from laminar type above a certain critical frequency depending on the pore radius, as explained in [4, 35].

Next, let $\mathcal{L}(u)$ be the second-order differential operator defined by

$$\mathcal{L}(u) = \big(\nabla \cdot \sigma^{(1)}(u), -\nabla p_f(u), \nabla \cdot \sigma^{(3)}(u)\big)^t.$$

Then the equations of motion in Ω, stated in the space-frequency domain, are given as follows [37]:

$$-\omega^2 \mathcal{P} u(x,\omega) + i\omega \mathcal{B} u(x,\omega) - \mathcal{L}(u(x,\omega)) = F(x,\omega), \quad (x,\omega) \in \Omega \times (0,\omega^*), \tag{2.6}$$

where $F(x,\omega) = (F^{(1)}(x,\omega), F^{(2)}(x,\omega), F^{(3)}(x,\omega))^t$ denotes the external source and ω^* is an upper temporal frequency of interest.

A plane wave analysis shows that three different compressional waves (P1, P2, and P3) and two shear waves (S1, S2) can propagate [24, 37]. The P1 and S1 waves correspond to the classical fast P and S waves propagating in elastic or viscoelastic isotropic solids. The additional slow waves are related to motions out of phase of the different phases. The experimental observation of the additional (slow) waves was reported by Leclaire, Cohen-Tenoudji, and Aguirre Puente [25].

Let us denote by ν the unit outer normal on Γ. In the two dimensional (2D) case let χ be a unit tangent on Γ so that $\{\nu, \chi\}$ is an orthonormal system on Γ. In the 3D case let χ^1 and χ^2 be two unit tangents on Γ so that $\{\nu, \chi^1, \chi^2\}$ is an orthonormal system on Γ.

Then, in the 2D case set

$$\mathcal{G}_\Gamma(u) = \big(\sigma^{(1)}(u)\nu \cdot \nu, \sigma^{(1)}(u)\nu \cdot \chi, p_f(u), \sigma^{(3)}(u)\nu \cdot \nu, \sigma^{(3)}(u)\nu \cdot \chi\big)^t, \tag{2.7a}$$

$$S_\Gamma(u) = \big(u^{(1)} \cdot \nu, u^{(1)} \cdot \chi, u^{(2)} \cdot \nu, u^{(3)} \cdot \nu, u^{(3)} \cdot \chi\big)^t, \tag{2.7b}$$

and in the 3D case set

$$\begin{aligned}\mathcal{G}_\Gamma(u) = \big(&\sigma^{(1)}(u)\nu \cdot \nu, \sigma^{(1)}(u)\nu \cdot \chi^1, \sigma^{(1)}(u)\nu \cdot \chi^2, p_f(u),\\ &\sigma^{(3)}(u)\nu \cdot \nu, \sigma^{(3)}(u)\nu \cdot \chi^1, \sigma^{(3)}(u)\nu \cdot \chi^2\big)^t,\end{aligned} \tag{2.8a}$$

$$S_\Gamma(u) = \big(u^{(1)} \cdot \nu, u^{(1)} \cdot \chi^1, u^{(1)} \cdot \chi^2, u^{(2)} \cdot \nu, u^{(3)} \cdot \nu, u^{(3)} \cdot \chi^1, u^{(3)} \cdot \chi^2\big)^t. \tag{2.8b}$$

Let us consider the solution of (2.3) with the following absorbing boundary condition:

$$-\mathcal{G}_\Gamma(u(x,\omega)) = i\omega \mathcal{D} S_\Gamma(u(x,\omega)), \qquad (x,\omega) \in \Gamma \times (0,\omega^*). \tag{2.9}$$

The matrix $\mathcal{D}$ in (2.9) is positive definite and given by

$$\mathcal{D} = \mathcal{M}_c{}^{\frac{1}{2}} \mathcal{S}_c{}^{\frac{1}{2}} \mathcal{M}_c{}^{\frac{1}{2}} = \big(\mathcal{E}_c \mathcal{M}_c{}^{-1}\big)^{\frac{1}{2}} \mathcal{M}_c, \tag{2.10}$$

with $\mathcal{S}_c = \mathcal{M}_c{}^{-\frac{1}{2}} \mathcal{E}_c \mathcal{M}_c{}^{-\frac{1}{2}}$, and in the 3D case,

$$\mathcal{M}_c = \begin{bmatrix} m_{11} & 0 & 0 & m_{12} & m_{13} & 0 & 0 \\ 0 & q_1 & 0 & 0 & 0 & q_2 & 0 \\ 0 & 0 & q_1 & 0 & 0 & 0 & q_2 \\ m_{12} & 0 & 0 & m_{22} & m_{23} & 0 & 0 \\ m_{13} & 0 & 0 & m_{23} & m_{33} & 0 & 0 \\ 0 & q_2 & 0 & 0 & 0 & q_3 & 0 \\ 0 & 0 & q_2 & 0 & 0 & 0 & q_3 \end{bmatrix},$$

$$\mathcal{E}_c = \begin{bmatrix} \lambda_{re}^{(1)} + 2\mu_{re}^{(1)} & 0 & 0 & B_{re}^{(1)} & D_{re}^{(3)} + \mu_{re}^{(13)} & 0 & 0 \\ 0 & \mu_{re}^{(1)} & 0 & 0 & 0 & \frac{1}{2}\mu_{re}^{(13)} & 0 \\ 0 & 0 & \mu_{re}^{(1)} & 0 & 0 & 0 & \frac{1}{2}\mu_{re}^{(13)} \\ B_{re}^{(1)} & 0 & 0 & K_{av,re} & B_{re}^{(2)} & 0 & 0 \\ D_{re}^{(3)} + \mu_{re}^{(13)} & 0 & 0 & B_{re}^{(2)} & \lambda_{re}^{(3)} + 2\mu_{re}^{(3)} & 0 & 0 \\ 0 & \frac{1}{2}\mu_{re}^{(13)} & 0 & 0 & 0 & \mu_{re}^{(3)} & 0 \\ 0 & 0 & \frac{1}{2}\mu_{re}^{(13)} & 0 & 0 & 0 & \mu_{re}^{(3)} \end{bmatrix},$$

with obvious modifications for the 2D case. The boundary condition (2.9) can be derived with a similar argument to that presented in [36] starting from the conservation of momentum equation on Γ and using the fact that the interaction energy among the different types of waves is small compared to the total energy involved.

3. A weak formulation. For $X \subset \mathbb{R}^d$ with boundary ∂X, let $(\cdot,\cdot)_X$ and $\langle \cdot,\cdot \rangle_{\partial X}$ denote the complex $L^2(X)$ and $L^2(\partial X)$ inner products for scalar, vector, or matrix valued functions. Also, for $s \in \mathbb{R}$, $\|\cdot\|_{s,X}$ and $|\cdot|_{s,X}$ will denote the usual norm and seminorm for the Sobolev space $H^s(X)$. In addition, if $X = \Omega$ or $X = \Gamma$, the subscript X may be omitted such that $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$ or $\langle \cdot,\cdot \rangle = \langle \cdot,\cdot \rangle_\Gamma$. Also, set

$$H(\mathrm{div};\Omega) = \{v \in [L^2(\Omega)]^d : \nabla\cdot v \in L^2(\Omega)\}, \quad H^1(\mathrm{div};\Omega) = \{v \in [H^1(\Omega)]^d : \nabla\cdot v \in H^1(\Omega)\},$$

with the norms

$$\|v\|_{H(\mathrm{div};\Omega)} = \left[\|v\|_0^2 + \|\nabla\cdot v\|_0^2\right]^{1/2}; \quad \|v\|_{H^1(\mathrm{div};\Omega)} = \left[\|v\|_1^2 + \|\nabla\cdot v\|_1^2\right]^{1/2}.$$

We will assume that the solution of (2.6) with the boundary condition (2.9) exists and satisfies the regularity assumption

$$\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_1 + \|\nabla\cdot u^{(2)}\|_1 \le C(w)\|F\|_0. \tag{3.1}$$

Let us introduce the space $\mathcal{V} = [H^1(\Omega)]^d \times H(\mathrm{div};\Omega) \times [H^1(\Omega)]^d$. Then multiply (2.3) by $v \in \mathcal{V}$, use integration by parts in the $(\mathcal{L}(u), v)$ term, and apply the boundary condition (2.9) to see that the solution u of (2.6) and (2.9) satisfies *the weak form,*

$$-\omega^2 (\mathcal{P}u, v) + i\omega (\mathcal{B}u, v) + \mathcal{A}(u,v) + i\omega \langle \mathcal{D}\ S_\Gamma(u), S_\Gamma(v)\rangle = (F, v), \qquad v = (v^{(1)}, v^{(2)}, v^{(3)})^t \in \mathcal{V}, \tag{3.2}$$

where $\mathcal{A}(u,v)$ is the bilinear form defined as follows:

$$\mathcal{A}(u,v) = \left(\sigma_{jk}^{(1)}(u), \varepsilon_{jk}(v^{(1)})\right) + \left(\sigma_{jk}^{(3)}(u), \varepsilon_{jk}(v^{(3)})\right) - \left(p_f(u), \nabla\cdot v^{(2)}\right), \qquad u, v \in \mathcal{V}. \tag{3.3}$$

In (3.3), and the rest of the paper, Einstein's convention of sum on repeated indices is used. Note that the bilinear form $\mathcal{A}(u,v)$ can be written in the form

$$\mathcal{A}(u,v) = (\mathbf{E}\ \tilde{\epsilon}(u), \tilde{\epsilon}(v)) = (\mathbf{E}_r\ \tilde{\epsilon}(u), \tilde{\epsilon}(v)) + i\,(\mathbf{E}_i\ \tilde{\epsilon}(u), \tilde{\epsilon}(v))\,,\ u, v \in \mathcal{V},$$

where $\mathbf{E} = \mathbf{E}_r + i\mathbf{E}_i$ is a complex matrix. Furthermore, we assume that the real part $\mathbf{E}_r$ is positive definite since in the elastic limit it is associated with the strain energy density. On the other hand, the imaginary part $\mathbf{E}_i$ is assumed to be positive definite because of the restriction imposed on our system by the first and second laws

of thermodynamics. A similar assumption was used in [33] to obtain restrictions on the imaginary parts of the coefficients in the constitutive relations for the case of a poroviscoelastic matrix saturated by a two-phase fluid. In the 2D case the matrix $\mathbf{E}$ is defined as follows, with the obvious extension to the 3D case:

$$\mathbf{E} = \begin{pmatrix} \widehat{\mathbf{E}} & 0 \\ 0 & \widehat{\mathbf{S}} \end{pmatrix}, \quad \widehat{\mathbf{S}} = \begin{pmatrix} 2\mu^{(1)} & \mu^{(13)} \\ \mu^{(13)} & 2\mu^{(3)} \end{pmatrix},$$

$$\widehat{\mathbf{E}} = \begin{pmatrix} \lambda^{(1)}+2\mu^{(1)} & \lambda^{(1)} & D^{(3)}+\mu^{(13)} & D^{(3)} & B^{(1)} \\ \lambda^{(1)} & \lambda^{(1)}+2\mu^{(1)} & D^{(3)} & D^{(3)}+\mu^{(13)} & B^{(1)} \\ D^{(3)}+\mu^{(13)} & D^{(3)} & \lambda^{(3)}+2\mu^{(3)} & \lambda^{(3)} & B^{(2)} \\ D^{(3)} & D^{(3)}+\mu^{(13)} & \lambda^{(3)} & \lambda^{(3)}+2\mu^{(3)} & B^{(2)} \\ B^{(1)} & B^{(1)} & B^{(2)} & B^{(2)} & K_{av} \end{pmatrix},$$

$$\tilde{\varepsilon}(u) = \begin{pmatrix} \varepsilon_{11}(u^{(1)}) \\ \varepsilon_{22}(u^{(1)}) \\ \varepsilon_{11}(u^{(3)}) \\ \varepsilon_{22}(u^{(3)}) \\ \nabla \cdot u^{(2)} \\ \varepsilon_{12}(u^{(1)}) \\ \varepsilon_{12}(u^{(3)}) \end{pmatrix}.$$

Let us analyze the uniqueness of the solution of our differential model for the case of a unit square $\Omega = (0,1)^2$ in the (x_1, x_2) plane to shorten the argument; the 3D case follows with the same argument. Then, set $F = 0$ and choose $v = u$ in (3.2). Taking the imaginary part in the resulting equation, we obtain

$$\omega\,(\mathcal{B}u, u) + (\mathbf{E}_i\ \tilde{\epsilon}(u), \tilde{\epsilon}(u)) + \omega\,\langle \mathcal{D}\ S_\Gamma(u), S_\Gamma(u)\rangle = 0.$$

Using (2.5) and that $\mathbf{E}_i$ and $\mathcal{D}$ are positive definite and $\mathcal{B}$ is nonnegative, we conclude that

$$u^{(2)} = 0, \quad u^{(1)} - u^{(3)} = 0, \qquad \Omega, \tag{3.4a}$$

$$u^{(1)} = 0, \quad u^{(3)} = 0, \qquad \Gamma, \tag{3.4b}$$

$$u^{(2)} \cdot \nu = 0, \qquad \Gamma. \tag{3.4c}$$

Consider the part Γ_1 of the boundary Γ defined by $\Gamma_1 = \{x = (x_1, x_2) \in \Gamma : x_1 = 1, 0 < x_2 < 1\}$. Notice that (3.4b) and (3.4c) imply that

$$\frac{\partial u_1^{(1)}}{\partial x_2} = \frac{\partial u_2^{(1)}}{\partial x_2} = \frac{\partial u_1^{(3)}}{\partial x_2} = \frac{\partial u_2^{(3)}}{\partial x_2} = 0, \quad \Gamma. \tag{3.5}$$

Owing to (2.9) $\mathcal{G}_\Gamma(u) = 0$ leads to the following relations on Γ_1:

$$\sigma_{11}^{(1)}(u) = (\lambda^{(1)} + 2\mu^{(1)})\,\frac{\partial u_1^{(1)}}{\partial x_1} + (D^{(3)} + \mu^{(13)})\frac{\partial u_1^{(3)}}{\partial x_1} + B^{(1)}\nabla\cdot u^{(2)} = 0, \tag{3.6a}$$

$$\sigma_{11}^{(3)}(u) = (\lambda^{(3)} + 2\mu^{(3)})\frac{\partial u_1^{(3)}}{\partial x_1} + (D^{(3)} + \mu^{(13)})\frac{\partial u_1^{(1)}}{\partial x_1} + B^{(2)}\nabla\cdot u^{(2)} = 0, \tag{3.6b}$$

$$\sigma_{12}^{(1)}(u) = \mu^{(1)}\frac{\partial u_2^{(1)}}{\partial x_1} + \frac{1}{2}\mu^{(13)}\frac{\partial u_2^{(3)}}{\partial x_1} = 0, \tag{3.6c}$$

$$\sigma_{12}^{(3)}(u) = \mu^{(3)}\frac{\partial u_2^{(3)}}{\partial x_1} + \frac{1}{2}\mu^{(13)}\frac{\partial u_2^{(1)}}{\partial x_1} = 0, \tag{3.6d}$$

$$-p_f(u) = B^{(1)}\frac{\partial u_1^{(1)}}{\partial x_1} + B^{(2)}\frac{\partial u_1^{(3)}}{\partial x_1} + K_{av}\nabla\cdot u^{(2)} = 0. \tag{3.6e}$$

Next we observe that (3.6c) and (3.6d) form a homogeneous 2×2 linear system of equations with coefficient matrix $2\,\widehat{\mathbf{S}}$, while (3.6a), (3.6b), and (3.6e) is another homogeneous linear system of equations with matrix coefficients

$$\mathcal{E}^{(\mathbf{p})} = \begin{bmatrix} \lambda^{(1)} + 2\mu^{(1)} & D^{(3)} + \mu^{(13)} & B^{(1)} \\ D^{(3)} + \mu^{(13)} & \lambda^{(3)} + 2\mu^{(3)} & B^{(2)} \\ B^{(1)} & B^{(2)} & K_{av} \end{bmatrix}.$$

We make the assumption (valid in any physically meaningful situation) that the coefficients in the matrix $\mathbf{E}_i$ fulfill

$$\mathrm{Im}(det(\mathcal{E}^{(\mathbf{p})})) > 0, \tag{3.7a}$$

$$\mathrm{Im}(det(\widehat{\mathbf{S}})) > 0. \tag{3.7b}$$

For example, a calculation shows that (3.7) is satisfied if the coefficients $\lambda^{(m)}, \mu^{(m)}$, $m = 1, 3$, and K_{av} are complex with nonzero imaginary parts and the imaginary part of K_{av} is chosen sufficiently small. Thus, under the condition (3.7), from (3.6) we conclude that

$$\frac{\partial u_2^{(1)}}{\partial x_1} = \frac{\partial u_2^{(3)}}{\partial x_1} = 0, \quad \Gamma_1, \tag{3.8a}$$

$$\frac{\partial u_1^{(1)}}{\partial x_1} = \frac{\partial u_1^{(3)}}{\partial x_1} = \nabla\cdot u^{(2)} = 0, \quad \Gamma_1. \tag{3.8b}$$

The same argument applies for the validity of (3.5) and (3.8) in the rest of the boundary. Thus by the Cauchy–Kovalevsky theorem $u^{(1)} = 0,\ u^{(3)} = 0$ in a neighborhood of any point on Γ where the coefficients are analytic and with the possible exception at the corners. Then the unique continuation principle [31] implies

$$u^{(1)} = u^{(3)} = 0, \quad \Omega. \tag{3.9}$$

Now from (3.4a) and (3.9) we have uniqueness. The 3D case follows with the identical argument.

We summarize the result in the following theorem.

THEOREM 3.1. *Under the assumption made in the above argument concerning the validity of* (3.7)*, problem* (2.6) *with the boundary condition* (2.9) *has a unique solution for any* $\omega \neq 0$.

For the analysis that follows a similar result can be demonstrated for the adjoint problem to (2.6) and (2.9). Thus, the solution $\psi = (\psi^{(1)}, \psi^{(2)}, \psi^{(3)})^t$ of the problem

$$-\omega^2 \mathcal{P}\psi - i\omega \mathcal{B}\psi - \mathcal{L}^*(\psi) = F, \quad \Omega \times (0, \omega^*), \tag{3.10a}$$

$$\mathcal{G}_\Gamma^*(\psi) - i\omega \mathcal{D} S_\Gamma(\psi) = 0, \quad \Gamma \times (0, \omega^*), \tag{3.10b}$$

is unique and satisfies the regularity assumption

$$\|\psi^{(1)}\|_2 + \|\psi^{(3)}\|_2 + \|\psi^{(2)}\|_1 + \|\nabla \cdot \psi^{(2)}\|_1 \leq C(\omega)\|F\|_0. \tag{3.11}$$

In (3.10a),

$$\mathcal{L}^*(\psi) = \big(\nabla \cdot \sigma^{(1,*)}(\psi), -\nabla p_f^*(\psi), \nabla \cdot \sigma^{(3,*)}(\psi)\big)^t,$$

where $\sigma^{(m,*)}(\psi), m = 1, 3$, and $p_f^*(\psi)$ are defined as in (2.3) but using the complex conjugates of the coefficients. Similarly, $\mathcal{G}_\Gamma^*(\psi)$ is defined as in (2.7) but using $\sigma^{(m,*)}(\psi), m = 1, 3$, and $p_f^*(\psi)$ in those definitions. As before, existence for (3.10) will be assumed.

4. The global finite element procedure. The numerical procedures will be defined and analyzed in detail in two dimensions and for rectangular elements. The changes for triangular elements and the 3D case will be described in section 9.

Let $\mathcal{T}^h(\Omega)$ be a nonoverlapping partition of Ω into rectangles Q_j of diameter bounded by h such that $\overline{\Omega} = \cup_{j=1}^J \overline{Q}_j$. Denote by ξ_j and ξ_{jk} the midpoints of $\partial Q_j \cap \Gamma$ and $\partial Q_j \cap \partial Q_k$, respectively. Let $\langle\langle \cdot, \cdot \rangle\rangle_{\Gamma_{jk}}$ denote the approximation to the (complex) inner product $\langle \cdot, \cdot \rangle_{\Gamma_{jk}}$ in $L^2(\Gamma_{jk})$ computed using the midpoint quadrature rule; more precisely,

$$\langle\langle u, v \rangle\rangle_{\Gamma_{jk}} = (u\overline{v})(\xi_{jk})|\Gamma_{jk}|,$$

where $|\Gamma_{jk}|$ denotes the measure of Γ_{jk}.

Let us denote by ν_{jk} the unit outer normal on $\partial Q_j \cap \partial Q_k$ from Q_j to Q_k and by ν_j the unit outer normal to ∂Q_j. Let χ_j and χ_{jk} be unit tangents on $\partial Q_j \cap \Gamma$ and $\partial Q_j \cap \partial Q_k$ so that $\{\nu_j, \chi_j\}$ and $\{\nu_{jk}, \chi_{jk}\}$ are orthonormal systems on $\partial Q_j \cap \Gamma$ and $\partial Q_j \cap \partial Q_k$, respectively.

To approximate each component of the solid displacement vector we employ the nonconforming finite element space as in [17], while to approximate the fluid displacement vector we choose the vector part of the Raviart–Thomas–Nedelec space [34, 30] of zero order. More specifically, set

$$\widehat{R} = [-1, 1]^2, \qquad \widehat{\mathcal{NC}}(\widehat{R}) = \mathrm{Span}\{1, \widehat{x}_1, \widehat{x}_2, \alpha(\widehat{x}_1) - \alpha(\widehat{x}_2)\}, \quad \alpha(\widehat{x}_1) = \widehat{x}_1^2 - \frac{5}{3}\widehat{x}_1^4,$$

with the degrees of freedom being the values at the midpoint of each edge of $\widehat{R}$. Also, if $\psi^L(\widehat{x}_1) = \frac{-1+\widehat{x}_1}{2}$, $\psi^R(\widehat{x}_1) = \frac{1+\widehat{x}_1}{2}$, $\psi^B(\widehat{x}_2) = \frac{-1+\widehat{x}_2}{2}$, $\psi^T(\widehat{x}_2) = \frac{1+\widehat{x}_2}{2}$, we have that

$$\widehat{\mathcal{W}}(\widehat{R}) = \mathrm{Span}\big\{(\psi^L(\widehat{x}_1), 0)^t, (\psi^R(\widehat{x}_1), 0)^t, (0, \psi^B(\widehat{x}_2))^t, (0, \psi^T(\widehat{x}_2))^t\big\}.$$

For each Q_j, let $F_{Q_j} : \widehat{R} \to Q_j$ be an invertible affine mapping such that $F_{Q_j}(\widehat{R}) = Q_j$, and define

$$\mathcal{NC}_j^h = \{v = (v_1, v_2)^t : \ v_i = \widehat{v}_i \circ F_{Q_j}^{-1}, \ \widehat{v}_i \in \widehat{\mathcal{NC}}(\widehat{R}), \ i = 1, 2\},$$

$$\mathcal{W}_j^h = \{w : \ w = \widehat{w} \circ F_{Q_j}^{-1}, \ \widehat{w} \in \widehat{\mathcal{W}}(\widehat{R})\}.$$

Setting

$$\mathcal{N}C^h = \{v : v_j = v|_{Q_j} \in \mathcal{N}C_j^h,\ v_j(\xi_{jk}) = v_k(\xi_{jk})\ \forall (j,k)\},$$
$$\mathcal{W}^h = \{w \in H(\mathrm{div};\Omega) :\ w_j = w|_{Q_j} \in \mathcal{W}_j^h\},$$

the global finite element space to approximate the solution u of (3.2) is defined by

$$\mathcal{V}^h = \mathcal{N}C^h \times \mathcal{W}^h \times \mathcal{N}C^h.$$

In order to state the approximation properties of $\mathcal{V}^h$ let us introduce the space

$$\tilde{\Lambda}_s^h = \big\{\tilde{\lambda}_s^h : \tilde{\lambda}_s^h|_{\partial Q_j \cap \partial Q_k} = \tilde{\lambda}_{s,jk}^h \in [P_0(\partial Q_j \cap \partial Q_k)]^2 \equiv \tilde{\Lambda}_{s,jk}^h, \quad \tilde{\lambda}_{s,jk}^h + \tilde{\lambda}_{s,kj}^h = 0\big\},$$

where $P_0(\partial Q_j \cap \partial Q_k)$ denotes the constant functions defined on $\partial Q_j \cap \partial Q_k$. Also, define the projections $\Pi_h : [H^2(\Omega)]^2 \to \mathcal{N}\mathcal{C}^h$ and $P_h^{(m)} : [H^2(\Omega)]^2 \times H^1(\mathrm{div};\Omega) \times [H^2(\Omega)]^2 \to \tilde{\Lambda}_s^h, m = 1,3$, associated with the two solid phases by

$$(\varphi^{(m)} - \Pi_h\varphi^{(m)})(\xi) = 0, \quad \xi = \xi_{jk} \text{ or } \xi_j,$$
$$\Big\langle \sigma^{(m)}(\psi_j)\nu - P_h^{(m)}(\psi_j), 1\Big\rangle_B = 0, \quad B = \partial Q_j \cap \partial Q_k \text{ or } \partial Q_j \cap \Gamma,$$

for all $\varphi \in [H^2(\Omega)]^2$ and $\psi \in [H^2(\Omega)]^2 \times H^1(\mathrm{div};\Omega) \times [H^2(\Omega)]^2$. Then, standard approximation theory implies that, for all $\varphi = (\varphi^{(1)}, \varphi^{(2)}, \varphi^{(3)})^t \in [H^2(\Omega)]^2 \times H^1(\mathrm{div};\Omega) \times [H^2(\Omega)]^2$,

$$\begin{aligned}(4.1)\ \sum_{m=1,3} \Bigg[&\|\varphi^{(m)} - \Pi_h\varphi^{(m)}\|_0 + h\left(\sum_j \|\varphi^{(m)} - \Pi_h\varphi^{(m)}\|_{1,Q_j}^2\right)^{\frac{1}{2}} \\ &+ h^2\left(\sum_j \|\varphi^{(m)} - \Pi_h\varphi^{(m)}\|_{2,Q_j}^2\right)^{\frac{1}{2}} + h^{\frac{1}{2}}\left(\sum_j |\varphi^{(m)} - \Pi_h\varphi^{(m)}|_{0,\partial Q_j}^2\right)^{\frac{1}{2}} \\ &+ h^{\frac{3}{2}}\left(\sum_j |\sigma^{(m)}(\varphi_j)\nu_j - P_h^{(m)}\varphi_j|_{0,\partial Q_j}^2\right)^{\frac{1}{2}}\Bigg] \\ &\le Ch^2\left(\|\varphi^{(1)}\|_2 + \|\varphi^{(3)}\|_2 + \|\nabla\cdot\varphi^{(2)}\|_1\right).\end{aligned}$$

We also notice the orthogonality to constants of the difference $\varphi_j^{(m)} - \varphi_k^{(m)}$ on the interfaces $\partial Q_j \cap \partial Q_k$ of Q_j and Q_k; that is,

$$\big\langle \varphi_j^{(m)} - \varphi_k^{(m)}, 1\big\rangle_{\partial Q_j \cap \partial Q_k} = 0 \text{ for all interfaces } \partial Q_j \cap \partial Q_k, \quad \varphi^{(m)} \in \mathcal{N}\mathcal{C}^h,\ m = 1,3.$$

Next, let us define the projection $\mathbf{Q}_h$ associated with the displacement vector of the fluid phase as follows:

$$\mathbf{Q}_h : [H^1(\Omega)]^2 \to \mathcal{W}^h : \big\langle \big(\mathbf{Q}_h\varphi^{(2)} - \varphi^{(2)}\big)\cdot\nu, 1\big\rangle_B = 0, \quad B = \partial Q_j \cap \partial Q_k \text{ or } B = \partial Q_j \cap \Gamma.$$

Then, it follows from [30, 34] that

$$\|\varphi^{(2)} - \mathbf{Q}_h\varphi^{(2)}\|_0 \le Ch\|\varphi^{(2)}\|_1, \tag{4.2a}$$
$$\|\varphi^{(2)} - \mathbf{Q}_h\varphi^{(2)}\|_{H(\mathrm{div};\Omega)} \le Ch\big(\|\varphi^{(2)}\|_1 + \|\nabla\cdot\varphi^{(2)}\|_1\big). \tag{4.2b}$$

Set

$$\mathcal{A}_h(u,v) = \sum_j \Big[\big(\sigma_{jk}^{(1)}(u), \varepsilon_{jk}(v^{(1)})\big)_{Q_j} + \big(\sigma_{jk}^{(3)}(u), \varepsilon_{jk}(v^{(3)})\big)_{Q_j} - \big(p_f(u), \nabla \cdot v^{(2)}\big)_{Q_j} \Big] \tag{4.3}$$

and

$$\Theta_h(u,v) = -\omega^2 (\mathcal{P}u, v) + i\omega (\mathcal{B}u, v) + \mathcal{A}_h(u,v) + i\omega \langle \mathcal{D}\ S_\Gamma(u), S_\Gamma(v)\rangle .$$

Then the *global* finite element procedure is defined as follows: find $u^h = (u^{(1,h)}, u^{(2,h)}, u^{(3,h)})^t \in \mathcal{V}^h$ such that

$$\Theta_h(u^h, v) = (F, v), \quad v = \big(v^{(1)}, v^{(2)}, v^{(3)}\big)^t \in \mathcal{V}^h. \tag{4.4}$$

Let us denote by $u_j^{(m,h)},\ j = 1, 2$, the components of the vector $u^{(m,h)},\ m = 1, 2, 3$. The following theorem analyzes the uniqueness of the solution of (4.4).

THEOREM 4.1. *Problem* (4.4) *has a unique solution for any* $\omega \neq 0$.

Proof. Set $F = 0$, choose $v = u^h$ in (4.4), and take the imaginary part in the resulting equation to obtain

$$\omega\big(\mathcal{B}u^h, u^h\big) + \sum_{Q_j} \big(\mathbf{E}_i\ \tilde{\epsilon}(u^h), \tilde{\epsilon}(u^h)\big)_{Q_j} + \omega\big\langle \mathcal{D}\ S_\Gamma(u^h), S_\Gamma(u^h)\big\rangle = 0. \tag{4.5}$$

Since each term in the left-hand side of (4.5) is nonnegative, in particular we have that $(\mathcal{B}u^h, u^h) = 0$, and the argument in the proof of Theorem 3.1 can be repeated to show that

$$u^{(2,h)} = 0, \quad u^{(1,h)} = u^{(3,h)}, \quad \Omega. \tag{4.6}$$

To show that $u^{(1,h)} = u^{(3,h)} = 0$, take an element, say Q_1, among the four elements which intersect Γ at the vertices of Ω; two faces of Q_1 are contained in Γ. After a proper transformation, without loss of generality we can assume that $Q_1 = (-1, 1)^2$ with the faces $\Gamma^R = \{(x_1, x_2) \in \Gamma : x_1 = 1\}$ and $\Gamma^T = \{(x_1, x_2) \in \Gamma : x_2 = 1\}$ contained in Γ. Set

$$\begin{aligned} u_1^{(1,h)} &= a_1 + b_1 x_1 + c_1 x_2 + d_1(\alpha(x_1) - \alpha(x_2)), \\ u_2^{(1,h)} &= a_2 + b_2 x_1 + c_2 x_2 + d_2(\alpha(x_1) - \alpha(x_2)). \end{aligned}$$

Since the boundary term in (4.5) must vanish and the matrix $\mathcal{D}$ is positive definite, we conclude that $S_\Gamma(u^h) = 0$ and consequently $u^{(m,h)}(x_1, x_2), m = 1, 3$, must vanish on $\Gamma^R \cup \Gamma^T$. In particular, at the midpoint of $\Gamma^R \cup \Gamma^T$ we have

$$\begin{aligned} u_1^{(1,h)}(1,0) &= a_1 + b_1 - \frac{2}{3} d_1 = 0, \quad u_1^{(1,h)}(0,1) = a_1 + c_1 + \frac{2}{3} d_1 = 0, \\ u_2^{(1,h)}(1,0) &= a_2 + b_2 - \frac{2}{3} d_2 = 0, \quad u_2^{(1,h)}(0,1) = a_2 + c_2 + \frac{2}{3} d_2 = 0. \end{aligned} \tag{4.7}$$

Next, since the second term in the left-hand side of (4.5) is nonnegative and the matrix

$\mathbf{E}_i$ is positive definite, for $(x_1, x_2) \in Q_1$ we must have

$$\varepsilon_{11}(u^{(1,h)}) = b_1 + 2d_1\left(x_1 - \frac{10}{3}x_1^3\right) = 0, \tag{4.8a}$$

$$\varepsilon_{22}(u^{(1,h)}) = c_2 - 2d_2\left(x_2 - \frac{10}{3}x_2^3\right) = 0, \tag{4.8b}$$

$$\varepsilon_{12}(u^{(1,h)}) = \frac{1}{2}\left[c_1 + b_2 - 2d_1\left(x_1 - \frac{10}{3}x_1^3\right) + 2d_2\left(x_2 - \frac{10}{3}x_2^3\right)\right] = 0. \tag{4.8c}$$

From (4.7) and (4.8) it follows that $u_1^{(1,h)}|_{Q_1} = u_2^{(1,h)}|_{Q_1} = 0$, and using (4.6) we also have $u_1^{(3,h)}|_{Q_1} = u_2^{(3,h)}|_{Q_1} = 0$. Let us take an element Q_2 adjacent to Q_1 that intersects Γ and has a common face Γ_{12} with Q_1. Then $u_1^{(1,h)}$ and $u_2^{(1,h)}$ vanish at the midpoints of Γ_2 and Γ_{12} and $\varepsilon_{11}(u^{(1,h)}), \varepsilon_{22}(u^{(1,h)})$, and $\varepsilon_{22}(u^{(1,h)})$ vanish identically on Q_2, so that repeating the above argument we verify that

$$u_1^{(m,h)}|_{Q_2} = u_2^{(m,h)}|_{Q_2} = 0, \quad m = 1, 3. \tag{4.9}$$

Repeating the argument, one can show that (4.9) holds for all elements with a face contained in Γ. Next stripping out such boundary elements, take a boundary element with two faces common with the corner of stripped out domain and repeat the argument to show the validity of (4.9) for those elements. Then continue the process until the domain is exhausted. This completes the proof. ∎

5. A priori error estimates for the global procedure. In this section, we derive an error estimate between the solutions u and u^h defined by (3.2) and (4.4), respectively. The argument in this section is close to that given in [21], which uses a boot-strapping argument similar to [15] for nonconforming finite element methods for Helmholtz-type problems. Also, see [16] for such a boot-strapping argument for conforming finite element methods for the Helmholtz equation.

Set

$$\mathbf{Z_h} = \left(\Pi_h u^{(1)}, \mathbf{Q}_h u^{(2)}, \Pi_h u^{(3)}\right)^t, \quad \delta = u - u^h, \quad \gamma = \mathbf{Z_h} u - u^h.$$

Our first goal is to derive an estimate for $\|\gamma\|_0$, and for that purpose we will solve the adjoint problem (3.10) to (2.6) and (2.9) with γ as a source term. It is convenient to define the following broken norms and seminorms:

$$\|v\|_{s,h}^2 = \sum_j \|v\|_{s,Q_j}^2, \quad |v|_{s,h}^2 = \sum_j |v|_{s,Q_j}^2, \quad |v|_{s,h,\Gamma}^2 = \sum_j |v|_{s,\partial Q_j \cap \Gamma}^2.$$

First note that for $v = (v^{(1)}, v^{(2)}, v^{(3)})^t \in [L^2(\Omega)]^6$ such that $v^{(1)}, v^{(3)} \in [H^1(Q_j)]^2$, $v^{(2)} \in H(\mathrm{div}; Q_j)$. Using integration by parts on each Q_j, we obtain

$$\begin{aligned}\Theta_h(u, v) = &\sum_j (-\omega^2 \mathcal{P}u + i\omega \mathcal{B}u - \mathcal{L}(u), v)_{Q_j} \\ &+ \sum_j \left\langle (\sigma^{(1)}(u)\nu, -p_f(u)\nu, \sigma^{(3)}(u)\nu)^t, (v^{(1)}, v^{(2)}, v^{(3)})^t \right\rangle_{\partial Q_j \setminus \Gamma}.\end{aligned} \tag{5.1}$$

Thus from (4.4) and (5.1) we see that for $v \in \mathcal{V}^h$,

$$\Theta_h(\delta, v) = \sum_j \left[\sum_{m=1,3} \left\langle \sigma^{(m)}(u)\nu, v^{(m)} \right\rangle_{\partial Q_j \setminus \Gamma} - \left\langle p_f(u), v^{(2)} \cdot \nu \right\rangle_{\partial Q_j \setminus \Gamma} \right]. \tag{5.2}$$

Notice that the regularity assumption (3.1) implies that $p_f(u) \in H^{1/2}(\partial Q_j \cap \partial Q_k)$ which, together with the fact that $v_j^{(2)} \cdot \nu_{jk} + v_k^{(2)} \cdot \nu_{kj} = 0$ in the sense of $H^{-1/2}(\partial Q_j \cap \partial Q_k)$, leads to

$$\sum_j \left\langle p_f(u), v^{(2)} \cdot \nu \right\rangle_{\partial Q_j \setminus \Gamma} = 0. \tag{5.3}$$

Hence, thanks to (5.3) and the fact that $v^{(1)}$ and $v^{(3)}$ are orthogonal to constants, (5.2) can be rewritten in the form

$$\Theta_h(\delta, v) = \sum_j \sum_{m=1,3} \left\langle \sigma^{(m)}(u)\nu - P_h^{(m)}(u), v^{(m)} \right\rangle_{\partial Q_j \setminus \Gamma}, \quad v \in \mathcal{V}^h. \tag{5.4}$$

Let $\psi = (\psi^{(1)}, \psi^{(2)}, \psi^{(3)})^t$ be the solution of the adjoint problem to (2.6) and (2.9):

$$-\omega^2 \mathcal{P}\psi - i\omega \mathcal{B}\psi - \mathcal{L}^*(\psi) = \gamma, \quad \Omega \times (0, \omega^*), \tag{5.5a}$$

$$\mathcal{G}_\Gamma^*(\psi) - i\omega \mathcal{D} S_\Gamma(\psi) = 0, \quad \Gamma \times (0, \omega^*). \tag{5.5b}$$

According to (3.11), ψ satisfies the regularity assumption

$$\|\psi^{(1)}\|_2 + \|\psi^{(3)}\|_2 + \|\psi^{(2)}\|_1 + \|\nabla \cdot \psi^{(2)}\|_1 \le C(w)\|\gamma\|_0. \tag{5.6}$$

Using integration by parts on each Q_j and applying the boundary condition (5.5b), we get

$$\begin{aligned} -(\gamma, \mathcal{L}^*(\psi)) &= \mathcal{A}_h(\gamma, \psi) + i\omega \langle \mathcal{D}\ S_\Gamma(\gamma), S_\Gamma(\psi) \rangle \\ &\quad - \sum_j \left[\sum_{m=1,3} \left\langle \gamma^{(m)}, \sigma^{(m,*)}(\psi)\nu \right\rangle_{\partial Q_j \setminus \Gamma} - \left\langle \gamma^{(2)} \cdot \nu, p_f^*(\psi) \right\rangle_{\partial Q_j \setminus \Gamma} \right]. \end{aligned} \tag{5.7}$$

Next, the argument used to show the validity of (5.3) can be applied to see that the last term in the right-hand side of (5.7) vanishes. Thus (5.7) implies that

$$\begin{aligned} \|\gamma\|_0^2 &= (\gamma, -\omega^2 \mathcal{P}\psi - i\omega \mathcal{B}\psi - \mathcal{L}^*(\psi)) \\ &= \Theta_h(\gamma, \psi) - \sum_j \sum_{m=1,3} \left\langle \gamma^{(m)}, \sigma^{(m,*)}(\psi)\nu \right\rangle_{\partial Q_j \setminus \Gamma}. \end{aligned} \tag{5.8}$$

Next, since $\sigma^{(m,*)}(\psi)\nu - P_h^{(m,*)}(\psi)$ has average value zero on $\partial Q_j \setminus \Gamma$, we have that for any $q^{(m)} \in [P_0(Q_j)]^2, m = 1, 3$,

$$\left\langle q^{(m)}, \sigma^{(m,*)}(\psi)\nu - P_h^{(m,*)}(\psi) \right\rangle_{\partial Q_j \setminus \Gamma} = 0, \quad m = 1, 3,$$

so that (5.8) can be stated in the form

$$\|\gamma\|_0^2 = \Theta_h(\gamma, \psi) - \sum_j \sum_{m=1,3} \left\langle \gamma^{(m)} - q^{(m)}, \sigma^{(m,*)}(\psi)\nu - P_h^{(m,*)}(\psi) \right\rangle_{\partial Q_j \setminus \Gamma}. \tag{5.9}$$

Next use (5.4) to see that for $v \in \mathcal{V}^h$,

$$\begin{aligned} \Theta_h(\gamma, v) &= \Theta_h(\delta, v) - \Theta_h(u - \mathbf{Z_h} u, v) \\ &= \sum_j \sum_{m=1,3} \left\langle \sigma^{(m)}(u)\nu - P_h^{(m)}(u), v^{(m)} \right\rangle_{\partial Q_j \setminus \Gamma} - \Theta_h(u - \mathbf{Z_h} u, v). \end{aligned} \tag{5.10}$$

Then use (5.10) in (5.9) to obtain

$$\begin{aligned}\|\gamma\|_0^2 &= \Theta_h(\gamma,\psi-v)-\Theta_h(u-\mathbf{Z_h}u,v)\\ &\quad+\sum_j\sum_{m=1,3}\big\langle\sigma^{(m)}(u)\nu-P_h^{(m)}(u),v^{(m)}\big\rangle_{\partial Q_j\setminus\Gamma}\\ &\quad-\sum_j\sum_{m=1,3}\big\langle\gamma^{(m)}-q^{(m)},\sigma^{(m,*)}(\psi)\nu-P_h^{(m,*)}(\psi)\big\rangle_{\partial Q_j\setminus\Gamma}.\end{aligned}\tag{5.11}$$

Next, since $\psi^{(m)}\in[H^2(\Omega)]^2, m=1,3$, (5.11) can be put in the equivalent form

$$\begin{aligned}\|\gamma\|_0^2 &= \Theta_h(\gamma,\psi-v)-\Theta_h(u-\mathbf{Z_h}u,v)\\ &\quad+\sum_j\sum_{m=1,3}\big\langle\sigma^{(m)}(u)\nu-P_h^{(m)}(u),v^{(m)}-\psi^{(m)}\big\rangle_{\partial Q_j\setminus\Gamma}\\ &\quad-\sum_j\left[\sum_{m=1,3}\big\langle\gamma^{(m)}-q^{(m)},\sigma^{(m,*)}(\psi)\nu-P_h^{(m,*)}(\psi)\big\rangle_{\partial Q_j\setminus\Gamma}\right].\end{aligned}\tag{5.12}$$

Let us bound each term in the right-hand side of (5.12). First, choose $v=(v^{(1)},v^{(2)},v^{(3)})^t=\mathbf{Z_h}\psi\in\mathcal{V}^h$ such that

$$\begin{aligned}&\sum_{m=1,3}\|\psi^{(m)}-v^{(m)}\|_0+h\|\psi^{(m)}-v^{(m)}\|_{1,h}+h^2\|v^{(m)}\|_{2,h}\\ &\qquad\le Ch^2\big(\|\psi^{(1)}\|_2+\|\psi^{(3)}\|_2\big)\le Ch^2\|\gamma\|_0,\end{aligned}\tag{5.13a}$$

$$\|\psi^{(2)}-v^{(2)}\|_0\le Ch\|\psi^{(2)}\|_1\le Ch\|\gamma\|_0,\tag{5.13b}$$

$$\begin{aligned}&\|\nabla\cdot(\psi^{(2)}-v^{(2)})\|_0+h\|\nabla\cdot(\psi^{(2)}-v^{(2)})\|_{1,h}\\ &\qquad\le Ch\|\nabla\cdot\psi^{(2)}\|_1\le Ch\|\gamma\|_0.\end{aligned}\tag{5.13c}$$

For the first term in the right-hand side of (5.12), using (5.13) we see that

$$\begin{aligned}|\Theta_h(\gamma,\psi-v)|&\le C(\omega)\Bigg[\|\gamma\|_0\|\psi-v\|_0+\sum_{m=1,3}\|\gamma^{(m)}\|_{1,h}\|\psi^{(m)}-v^{(m)}\|_{1,h}\\ &\qquad+\|\nabla\cdot\gamma\|_0\|\nabla\cdot(\psi-v)\|_0+|\langle S_\Gamma(\gamma),S_\Gamma(\psi-v)\rangle|\Bigg]\\ &\le C(\omega)h\|\gamma\|_0\big[\|\gamma^{(1)}\|_{1,h}+\|\gamma^{(3)}\|_{1,h}+\|\nabla\cdot\gamma^{(2)}\|_0\\ &\qquad+|\langle S_\Gamma(\gamma),S_\Gamma(\psi-v)\rangle|\big].\end{aligned}\tag{5.14}$$

The boundary integral in the right-hand side of (5.14) can be bounded using (5.6) and the trace inequality as follows:

$$|\langle S_\Gamma(\gamma),S_\Gamma(\psi-v)\rangle|\le C\|\gamma\|_0h^{3/2}\big[\|\gamma^{(1)}\|_{1,h}+|\gamma^{(3)}\|_{1,h}\big],\tag{5.15}$$

where we have used that

$$\sum_j\big\langle\big(\psi^{(2)}-\mathbf{Q}_h\psi^{(2)}\big)\cdot\nu,\gamma^{(2)}\cdot\nu\big\rangle_{\partial Q_j\setminus\Gamma}=0.$$

Hence, using (5.15) in (5.14), we get

$$|\Theta_h(\gamma,\psi-v)|\le C(\omega)h\|\gamma\|_0\left[\|\gamma^{(1)}\|_{1,h}+\|\gamma^{(3)}\|_{1,h}+\|\nabla\cdot\gamma^{(2)}\|_0\right].\tag{5.16}$$

By choosing $q_j^{(m)} = q^m|_{Q_j},\ m = 1, 3$, to be the average value of $\gamma^{(m)}$ on Q_j and using the trace inequality, (4.1) and (5.6), the last term in (5.12) is bounded as follows:

$$\begin{aligned}
&\left| \sum_{m=1,3} \sum_j \big\langle \gamma^{(m)} - q^{(m)}, \sigma^{(m,*)}(\psi)\nu - P_h^{(m,*)}(\psi) \big\rangle_{\partial Q_j \setminus \Gamma} \right| \\
&\quad \le \sum_{m=1,3} \left(\sum_j |\gamma^{(m)} - q^{(m)}|^2_{0,\partial Q_j \setminus \Gamma} \right)^{1/2} \left(\sum_j |\sigma^{(m,*)}(\psi)\nu - P_h^{(m,*)}(\psi)|^2_{0,\partial Q_j \setminus \Gamma} \right)^{1/2} \\
&\quad \le \sum_{m=1,3} \left(\sum_j h \|\gamma^{(m)}\|^2_{1,Q_j} \right)^{1/2} h^{1/2} \big(\|\psi^{(1)}\|_2 + \|\psi^{(3)}\|_2 + \|\nabla \cdot \psi^{(2)}\|_1 \big) \\
&\quad \le Ch\|\gamma\|_0 \big(\|\gamma^{(1)}\|_{1,h} + \|\gamma^{(3)}\|_{1,h} \big). \qquad (5.17)
\end{aligned}$$

Next, using integration by parts in the $\mathcal{A}_h(u - \mathbf{Z_h}u, v)$ term and the boundary condition (5.5b), the second term in the right-hand side of (5.12) can be written in the form

$$\begin{aligned}
\Theta_h(u - \mathbf{Z_h}u, v) &= \sum_j (u - \mathbf{Z_h}u, -\omega^2 \mathcal{P}v - i\omega \mathcal{B}v - \mathcal{L}^*(v))_{Q_j} \\
&\quad + \sum_j \langle S_\Gamma(u - \mathbf{Z_h}u), \mathcal{G}_\Gamma^*(v) \rangle_{\partial Q_j \setminus \Gamma} + \sum_j \langle S_\Gamma(u - \mathbf{Z_h}u), \mathcal{G}_\Gamma^*(v) \rangle_{\partial Q_j \cap \Gamma} \\
&\quad + i\omega \langle \mathcal{D} S_\Gamma(u - \mathbf{Z_h}u), S_\Gamma(v) \rangle \\
&= \sum_j (u - \mathbf{Z_h}u, -\omega^2 \mathcal{P}v - i\omega \mathcal{B}v - \mathcal{L}^*(v))_{Q_j} \\
&\quad + \sum_j \langle S_\Gamma(u - \mathbf{Z_h}u), \mathcal{G}_\Gamma^*(v) - \mathcal{G}_\Gamma^*(\psi) \rangle_{\partial Q_j \cap \Gamma} \\
&\quad + \sum_j \langle S_\Gamma(u - \mathbf{Z_h}u), \mathcal{G}_\Gamma^*(v) \rangle_{\partial Q_j \setminus \Gamma} \\
&\quad + i\omega \langle \mathcal{D} S_\Gamma(u - \mathbf{Z_h}u), S_\Gamma(v - \psi) \rangle \\
&\equiv T_1 + T_2 + T_3 + T_4. \qquad (5.18)
\end{aligned}$$

Let us bound each term in the right-hand side of (5.18). First, using (4.1), (4.2), and (5.13) we see that

$$|T_1| \le Ch\|\gamma\|_0 (\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_1 + \|\nabla \cdot u^{(2)}\|_1).$$

For the T_2 term, applying the trace inequality, (4.1), (4.2), (3.11), and (5.13), one has

$$\begin{aligned}
|T_2| &\le \sum_{m=1,3} \sum_j |u^{(m)} - \Pi_h u^{(m)}|_{0,\partial Q_j \cap \Gamma} |(\sigma^{(m,*)}(v) - \sigma^{(m,*)}(\psi)) \cdot \nu|_{0,\partial Q_j \cap \Gamma} \\
&\quad + \sum_j |(u^{(2)} - \mathbf{Q}_h u^{(2)}) \cdot \nu|_{-1/2,\partial Q_j \cap \Gamma} |p_f^*(v) - p_f^*(\psi)|_{1/2,\partial Q_j \cap \Gamma} \\
&\le C\|\gamma\|_0 [h^2 (\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + h(\|u^{(2)}\|_1 + \|\nabla \cdot u^{(2)}\|_1)]. \qquad (5.19)
\end{aligned}$$

Next, we decompose T_3 as follows:

$$\begin{aligned} T_3 &= \sum_j \langle S_\Gamma(u - \mathbf{Z_h}u), \mathcal{G}^*_\Gamma(v) - \mathcal{G}^*_\Gamma(\psi)\rangle_{\partial Q_j \setminus \Gamma} \\ &\quad + \sum_j \langle S_\Gamma(u - \mathbf{Z_h}u), \mathcal{G}^*_\Gamma(\psi)\rangle_{\partial Q_j \setminus \Gamma} \\ &\equiv T_{3,1} + T_{3,2}. \end{aligned} \tag{5.20}$$

Then, as in (5.19), the first term is bounded as follows:

$$|T_{3,1}| \le C\|\gamma\|_0[h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + h(\|u^{(2)}\|_1 + \|\nabla \cdot u^{(2)}\|_1)].$$

The other term in (5.20) can be bounded by again using the fact that $\Pi_h u_j^{(m)} - \Pi_h u_k^{(m)}$ is orthogonal to constants

$$\begin{aligned} |T_{3,2}| \le \Bigg| & \sum_{m=1,3} \sum_j \langle (u^{(m)} - \Pi_h u^{(m)}) \cdot \nu, \sigma^{(m,*)}(\psi)\nu \cdot \nu\rangle_{\partial Q_j \setminus \Gamma} \\ & + \langle (u^{(m)} - \Pi_h u^{(m)}) \cdot \chi, \sigma^{(m,*)}(\psi)\nu \cdot \chi\rangle_{\partial Q_j \setminus \Gamma} \\ & - \sum_j \langle (u^{(2)} - \mathbf{Q}_h u^{(2)}) \cdot \nu, p_f^*(\psi)\rangle_{\partial Q_j \setminus \Gamma} \Bigg| \\ \le C\|\gamma\|_0 & [h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2)], \end{aligned}$$

where we have again used the argument in (5.3) to cancel out the terms involving $u^{(2)}$ in the inequality above. Finally, in order to bound T_4, applying the trace inequality, (4.1), (4.2), and (5.13), we obtain

$$\begin{aligned} |T_4| &\le C \Bigg[\sum_{m=1,3} \sum_j |u^{(m)} - \Pi_h u^{(m)}|_{0,\partial Q_j \cap \Gamma} |v^{(m)} - \psi^{(m)}|_{0,\partial Q_j \cap \Gamma} \\ &\qquad + \sum_j |(u^{(2)} - \mathbf{Q}_h u^{(2)}) \cdot \nu|_{0,\partial Q_j \cap \Gamma} |(v^{(2)} - \psi^{(2)}) \cdot \nu|_{0,\partial Q_j \cap \Gamma} \Bigg] \\ &\le \sum_{m=1,3} \sum_j \|u^{(m)} - \Pi_h u^{(m)}\|^{\frac{1}{2}}_{0,Q_j} \|u^{(m)} - \Pi_h u^{(m)}\|^{\frac{1}{2}}_{1,Q_j} \\ &\quad \times \|\psi^{(m)} - v^{(m)}\|^{\frac{1}{2}}_{0,Q_j} \|\psi^{(m)} - v^{(m)}\|^{\frac{1}{2}}_{1,Q_j} \\ &\quad + \sum_j h^{\frac{1}{2}} |u^{(2)} \cdot \nu|_{\frac{1}{2},\partial Q_j \cap \Gamma} h^{\frac{1}{2}} |\psi^{(2)} \cdot \nu|_{\frac{1}{2},\partial Q_j \cap \Gamma} \\ &\le C\|\gamma\|_0[h^3(\|u^{(1)}\|_2 + \|u^{(3)}\|_2)] + Ch\|u^{(2)}\|_1 \|\psi^{(2)}\|_1 \\ &\le C\|\gamma\|_0[h^3(\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + Ch\|u^{(2)}\|_1]. \end{aligned}$$

Collecting the estimates for T_1, T_2, T_3, and T_4, we conclude that

$$|\Theta_h(u - \mathbf{Z_h}u, v)| \le C\|\gamma\|_0[h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + h(\|u^{(2)}\|_1 + \|\nabla \cdot u^{(2)}\|_1)]. \tag{5.21}$$

Next, use the trace inequality, (4.1), and (5.16) to bound the third term in the

right-hand side of (5.12) as follows:

$$\begin{aligned}
&\left|\sum_j \sum_{m=1,3} \langle \sigma^{(m)}(u)\nu - P_h^m(u), v^{(m)} - \psi^{(m)}\rangle_{\partial Q_j \setminus \Gamma}\right| \\
&\quad \le \sum_{m=1,3} \left(\sum_j |\sigma^{(m)}(u)\nu - P_h^{(m)}(u)|^2_{0,\partial Q_j\setminus\Gamma}\right)^{1/2} \left(\sum_j |v^{(m)} - \psi^{(m)}|^2_{0,\partial Q_j\setminus\Gamma}\right)^{1/2} \\
&\quad \le Ch^{1/2}(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|\nabla\cdot u^{(2)}\|_1) h^{3/2}(\|\psi^{(1)}\|_2 + \|\psi^{(3)}\|_2) \\
&\quad \le Ch^2\|\gamma\|_0(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|\nabla\cdot u^{(2)}\|_1). \qquad (5.22)
\end{aligned}$$

Thus, collecting the bounds in (5.16), (5.17), (5.21), and (5.22), we obtain

$$\begin{aligned}
\|\gamma\|_0 \le C(\omega)[&h(\|\gamma^{(1)}\|_{1,h} + \|\gamma^{(3)}\|_{1,h} + \|\nabla\cdot\gamma^{(2)}\|_0) \\
&+ h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + h(\|u^{(2)}\|_1 + \|\nabla\cdot u^{(2)}\|_1)]. \qquad (5.23)
\end{aligned}$$

Using the triangle inequality, the last estimate (5.23), and the approximation properties of Π_h and $\mathbf{Q}_h$ in (4.1) and (4.2), we get

$$\begin{aligned}
\|\delta\|_0 \le \|\gamma\|_0 + \|\mathbf{Z}_h u - u\|_0 &\le C(\omega)[h(\|\delta^{(1)}\|_{1,h} + \|\delta^{(3)}\|_{1,h} + \|\nabla\cdot\delta^{(2)}\|_0) \\
&+ h(\|u^{(1)} - \Pi_h u^{(1)}\|_{1,h} + \|u^{(3)} - \Pi_h u^{(3)}\|_{1,h} + \|\nabla\cdot(u^{(2)} - \mathbf{Q}_h u^{(2)})\|_0) \\
&+ h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + h(\|u^{(2)}\|_1 + \|\nabla\cdot u^{(2)}\|_1)] \\
&\le C(\omega)[h(\|\delta^{(1)}\|_{1,h} + \|\delta^{(3)}\|_{1,h} + \|\nabla\cdot\delta^{(2)}\|_0) \qquad (5.24) \\
&+ h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2) + h(\|u^{(2)}\|_1 + \|\nabla\cdot u^{(2)}\|_1)].
\end{aligned}$$

We next use a Gärding-type inequality to bound the δ terms in (5.24) in terms of the u terms in that inequality. Using Korn's second inequality [18, 32] and noting that $\mathbf{E}_i$ is positive definite, we get

$$\begin{aligned}
|\mathrm{Im}(\Theta_h(\delta,\delta))| &= \omega(\mathcal{B}\delta,\delta) + \sum_j (\mathbf{E}_i\ \tilde{\epsilon}(\delta), \tilde{\epsilon}(\delta))_{Q_j} + \omega\langle \mathcal{D}\ S_\Gamma(\delta), S_\Gamma(\delta)\rangle \\
&\ge C_1(\omega)[\|\delta^{(1)}\|^2_{1,h} + \|\delta^{(3)}\|^2_{1,h} + \|\nabla\cdot\delta^{(2)}\|^2_0 + \langle S_\Gamma(\delta), S_\Gamma(\delta)\rangle] \\
&\quad - C_2(\omega)\|\delta\|^2_0.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\|\delta^{(1)}\|^2_{1,h} + \|\delta^{(3)}\|^2_{1,h} + \|\nabla\cdot\delta^{(2)}\|^2_0 + \langle S_\Gamma(\delta), S_\Gamma(\delta)\rangle \\
&\quad \le C_3(\omega)|\Theta_h(\delta,\delta)| + C_2(\omega)\|\delta\|^2_0 \\
&\quad \le C_3(\omega)[\|\delta\|^2_0 + |\Theta_h(\delta, u - \mathbf{Z}_h u)| + |\Theta_h(\delta,\gamma)|]. \qquad (5.25)
\end{aligned}$$

Since $\gamma \in \mathcal{V}^h$, the expression for $\Theta_h(\delta,\gamma)$ given in (5.4) can be replaced by using (5.25)

so that

$$
\begin{aligned}
&\|\delta^{(1)}\|_{1,h}^2 + \|\delta^{(3)}\|_{1,h}^2 + \|\nabla\cdot\delta^{(2)}\|_0^2 + \langle S_\Gamma(\delta), S_\Gamma(\delta)\rangle \\
&\le C_3(\omega)\Bigg[\|\delta\|_0^2 - \omega^2(\mathcal{P}\delta, u-\mathbf{Z}_h u) + i\omega(\mathcal{B}\delta, u-\mathbf{Z}_h u) + \mathcal{A}_h(\delta, u-\mathbf{Z}_h u) \\
&\quad + i\omega\langle \mathcal{D}\ S_\Gamma(\delta), S_\Gamma(u-\mathbf{Z}_h u)\rangle + \sum_j \sum_{m=1,3} \langle \sigma^{(m)}(u)\nu - P_h^{(m)}(u), \gamma^{(m)}\rangle_{\partial Q_j\setminus\Gamma}\Bigg].
\end{aligned} \tag{5.26}
$$

Let us bound the last five terms in the right-hand side of (5.26). First, thanks to the approximation properties of Π_h and $\mathbf{Q}_h$ given in (4.1) and (4.2), it follows that

$$
\begin{aligned}
&|-\omega^2(\mathcal{P}\delta, u-\mathbf{Z}_h u) + i\omega(\mathcal{B}\delta, u-\mathbf{Z}_h u)| \\
&\qquad \le C(\omega)[\|\delta\|_0^2 + h^4(\|u^{(1)}\|_2^2 + \|u^{(3)}\|_2^2) + h^2\|u^{(2)}\|_1^2].
\end{aligned} \tag{5.27}
$$

Again, due to (4.1) and (4.2),

$$
\begin{aligned}
|\mathcal{A}_h(\delta, u-\mathbf{Z}_h u)| &\le C(\omega)\Bigg[\sum_{m=1,3}(\|\delta^{(m)}\|_{1,h}\|u^{(m)} - \Pi_h u^{(m)}\|_{1,h}) \\
&\qquad + \|\nabla\cdot\delta^{(2)}\|_0\|\nabla\cdot(u^{(2)} - \mathbf{Q}_h u^{(2)})\|_0\Bigg] \\
&\le \epsilon(\|\delta^{(1)}\|_{1,h}^2 + \|\delta^{(3)}\|_{1,h}^2 + \|\nabla\cdot\delta^{(2)}\|_0^2) \\
&\quad + C(\omega)[h^2(\|u^{(1)}\|_2^2 + \|u^{(3)}\|_2^2) + h^2\|\nabla\cdot u^{(2)}\|_1^2].
\end{aligned} \tag{5.28}
$$

Next, using the trace inequality and approximation properties (4.1) and (4.2) again, we have

$$
\begin{aligned}
&|\omega\langle \mathcal{D}\ S_\Gamma(\delta), S_\Gamma(u-\mathbf{Z}_h u)\rangle| \\
&\quad \le \epsilon\langle \mathcal{D}S_\Gamma(\delta), S_\Gamma(\delta)\rangle + C(\omega)\langle \mathcal{D}S_\Gamma(u-\mathbf{Z}_h u), S_\Gamma(u-\mathbf{Z}_h u)\rangle \\
&\quad \le \epsilon\langle \mathcal{D}S_\Gamma(\delta), S_\Gamma(\delta)\rangle \\
&\qquad + C(\omega)\Bigg[\sum_{m=1,3}\sum_j |u^{(m)} - \Pi_h u^{(m)}|_{0,\partial Q_j\cap\Gamma}^2 + \sum_j |(u^{(2)} - \mathbf{Q}_h u^{(2)})\cdot\nu|_{0,\partial Q_j\cap\Gamma}^2\Bigg] \\
&\quad \le \epsilon\langle \mathcal{D}S_\Gamma(\delta), S_\Gamma(\delta)\rangle + C(\omega)\Bigg[h^3(\|u^{(1)}\|_2^2 + \|u^{(3)}\|_2^2) + \sum_j h^2|u^{(2)}\cdot\nu|_{1,\partial Q_j\cap\Gamma}^2\Bigg] \\
&\quad \le \epsilon\langle \mathcal{D}S_\Gamma(\delta), S_\Gamma(\delta)\rangle + C(\omega)\Bigg[h^3(\|u^{(1)}\|_2^2 + \|u^{(3)}\|_2^2) + \sum_j h^2\|u^{(2)}\|_{\frac{3}{2},Q_j}^2\Bigg] \\
&\quad \le \epsilon\langle \mathcal{D}S_\Gamma(\delta), S_\Gamma(\delta)\rangle + C(\omega)[h^3(\|u^{(1)}\|_2^2 + \|u^{(3)}\|_2^2) + h^2\|u^{(2)}\|_{\frac{3}{2}}^2].
\end{aligned} \tag{5.29}
$$

Finally, owing to the orthogonality property of $\gamma^{(m)}$ to constants on $\partial Q_j\setminus\Gamma$, the trace

inequality, and (4.1), it follows that

$$\begin{aligned}
&\left|\sum_j \sum_{m=1,3} \langle \sigma^{(m)}(u)\nu - P_h^{(m)}(u), \gamma^{(m)}\rangle_{\partial Q_j\setminus\Gamma}\right| \\
&\quad = \left|\sum_j \sum_{m=1,3} \langle \sigma^{(m)}(u)\nu - P_h^{(m)}(u), \gamma^{(m)} - q^{(m)}\rangle_{\partial Q_j\setminus\Gamma}\right| \\
&\quad \le C \sum_{m=1,3} \left(\sum_j |\sigma^{(m)}(u)\nu - P_h^{(m)}(u)|^2_{0,\partial Q_j\setminus\Gamma}\right)^{1/2} \left(\sum_j |\gamma^{(m)} - q^{(m)}|^2_{0,\partial Q_j\setminus\Gamma}\right)^{1/2} \\
&\quad \le Ch^{1/2}(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|\nabla\cdot u^{(2)}\|_1) \sum_{m=1,3} \left(\sum_j h\|\gamma^{(m)}\|_{1,Q_j}\right)^{1/2} \\
&\quad \le Ch \sum_{m=1,3} \|\gamma^{(m)}\|_{1,h}(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|\nabla\cdot u^{(2)}\|_1) \\
&\quad \le Ch\left(\sum_{m=1,3} \|\delta^{(m)}\|_{1,h} + \|u^{(m)} - \Pi_h u^{(m)}\|_{1,h}\right)(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|\nabla\cdot u^{(2)}\|_1) \\
&\quad \le \epsilon(\|\delta^{(1)}\|^2_{1,h} + \|\delta^{(3)}\|^2_{1,h}) + Ch^2(\|u^{(1)}\|_2^2 + \|u^{(3)}\|_2^2 + \|\nabla\cdot u^{(2)}\|_1^2). \qquad (5.30)
\end{aligned}$$

Hence using (5.27), (5.28), (5.29), and (5.30) in (5.26), we have the following estimate:

$$\begin{aligned}
&\|\delta^{(1)}\|_{1,h} + \|\delta^{(3)}\|_{1,h} + \|\nabla\cdot\delta^{(2)}\|_0 + \langle S_\Gamma(\delta), S_\Gamma(\delta)\rangle^{\frac{1}{2}} \\
&\qquad \le C(\omega)[\|\delta\|_0 + h(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_{\frac{3}{2}} + \|\nabla\cdot u^{(2)}\|_1)]. \qquad (5.31)
\end{aligned}$$

Next, use (5.31) in (5.24) to obtain

$$\begin{aligned}
\|\delta\|_0 \le C(\omega)[&h\|\delta\|_0 + h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_{\frac{3}{2}}) \\
&+ h(\|u^{(2)}\|_1 + \|\nabla\cdot u^{(2)}\|_1)]. \qquad (5.32)
\end{aligned}$$

Then, for sufficiently small $h > 0$ such that $0 < C(\omega)h < 1$, the term $\|\delta\|_0$ in the right-hand side of (5.32) is absorbed in the left-hand side, and therefore

$$\|\delta\|_0 \le C(\omega)[h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_{\frac{3}{2}}) + h(\|u^{(2)}\|_1 + \|\nabla\cdot u^{(2)}\|_1)]. \qquad (5.33)$$

Finally, using (5.33) in (5.31), we arrive at the following error estimate:

$$\begin{aligned}
&\|\delta^{(1)}\|_{1,h} + \|\delta^{(3)}\|_{1,h} + \|\nabla\cdot\delta^{(2)}\|_0 + \langle S_\Gamma(\delta), S_\Gamma(\delta)\rangle \le C(\omega)h[\|u^{(1)}\|_2 \\
&\quad + \|u^{(3)}\|_2 + \|u^{(2)}\|_{\frac{3}{2}} + \|\nabla\cdot u^{(2)}\|_1].
\end{aligned}$$

We summarize the above in the following theorem.

THEOREM 5.1. *Let $u \in \mathcal{V}$ and $u^h \in \mathcal{V}^h$ be the solutions of* (3.2) *and* (4.4)*, respectively. We then have the following energy-norm error estimate: for sufficiently*

small $h > 0$,

$$\sum_{m=1,3} \|u^{(m)} - u^{(m,h)}\|_{1,h} + \|\nabla \cdot (u^{(2)} - u^{(2,h)})\|_0 + \sum_{m=1,3} |u^{(m)} - u^{(m,h)}|_{0,\Gamma} + |(u^{(2)} - u^{(2,h)}) \cdot \nu|_{0,\Gamma} \le C(\omega) h \big[\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_{\frac{3}{2}} + \|\nabla \cdot u^{(2)}\|_1\big].$$

Also, we have the $[L^2(\Omega)]^6$*-error estimate as follows: for sufficiently small* $h > 0$,

$$\|u - u^h\|_0 \le C(\omega)[h^2(\|u^{(1)}\|_2 + \|u^{(3)}\|_2 + \|u^{(2)}\|_{\frac{3}{2}}) + h(\|u^{(2)}\|_1 + \|\nabla \cdot u^{(2)}\|_1)].$$

6. A global hybridized nonconforming finite element procedure. Let us decompose $\Omega \in \mathbb{R}^2$ into nonoverlapping subdomains $\Omega_1, \ldots, \Omega_N$ such that each Ω_j is composed of the union of disjoint rectangles $Q \in \mathcal{T}^h(\Omega)$, with the interfaces $\Gamma_{jk} = \partial\Omega_j \cap \partial\Omega_k$. Also, let $\Gamma_j = \partial\Omega_j \cap \Gamma$. Set

$$\begin{aligned}
\mathcal{T}^h(\Omega_j) &= \{Q \in \mathcal{T}^h(\Omega) : Q \in \Omega_j\},\\
\mathcal{NC}^h(\Omega_j) &= \{v_j : \Omega_j \to \mathbb{C}^2,\ v_j|_Q \in \mathcal{NC}_j^h\ \forall Q \in \mathcal{T}^h(\Omega_j), v_j|_{Q_k}(\xi_{kl}) = v_j|_{Q_l}(\xi_{kl})\ \forall (k,l)\},\\
\mathcal{W}^h(\Omega_j) &= \{w \in H(\mathrm{div};\Omega_j) : w_k = w|_{Q_k} \in \mathcal{W}_k^h\},\\
\mathcal{V}^h(\Omega_j) &= \mathcal{NC}^h(\Omega_j) \times \mathcal{W}^h(\Omega_j) \times \mathcal{NC}^h(\Omega_j).
\end{aligned}$$

Our global hybridized finite element space is then defined by

$$\mathcal{V}_{-1}^h = \{v \in [L^2(\Omega)]^6 : v|_{\Omega_j} \in \mathcal{V}^h(\Omega_j)\}.$$

In order to define a hybridized procedure, we follow the ideas in [1, 19, 20, 14] to impose the continuity constraints across interior interfaces using Lagrange multipliers. Thus we introduce the space $\tilde{\Lambda}_{-1,j}^h$, with $\tilde{\lambda}_{jk}^h \in \tilde{\Lambda}_{-1,j}^h$ associated with $\mathcal{G}_{\Gamma_{jk}}(u_j)$ on Γ_{jk}:

$$\tilde{\Lambda}_{-1,j}^h = \{\tilde{\lambda}_j^h : \tilde{\lambda}_{jk}^h = \tilde{\lambda}_j^h|_{\partial Q \cap \Gamma_{jk}} \in \tilde{\Lambda}_{-1,jk}^h\ \forall Q \in \Omega_j \text{ such that } \overline{Q} \cap \Gamma_{jk} \neq \emptyset\},$$

where

$$\tilde{\Lambda}_{-1,jk}^h = \{\tilde{\lambda}_{jk}^h : \tilde{\lambda}_{jk}^h \in [P_0(\partial Q \cap \Gamma_{jk})]^5\ \forall Q \in \Omega_j \text{ such that } \overline{Q} \cap \Gamma_{jk} \neq \emptyset, \tilde{\lambda}_{jk}^h = \tilde{\lambda}_{kj}^h\}\ \forall j,k.$$

Set

$$\tilde{\Lambda}_{-1}^h = \cup_j \tilde{\Lambda}_{-1,j}^h.$$

The global hybridized nonconforming procedure is defined in the following fashion: find $(\tilde{u}^h, \tilde{\lambda}^h) \in \mathcal{V}_{-1}^h \times \tilde{\Lambda}_{-1}^h$ such that

$$\text{(6.1a)}\quad \sum_j \sum_{Q \in \mathcal{T}^h(\Omega_j)} [-\omega^2(\mathcal{P}\tilde{u}_j^h, v)_Q + i\omega(\mathcal{B}\tilde{u}_j^h, v)_Q + \mathcal{A}_{h,Q}(\tilde{u}_j^h, v)] - \sum_{j,k} \langle\langle \tilde{\lambda}_{jk}^h, S_{\Gamma_{jk}}(v)\rangle\rangle_{\Gamma_{jk}} + i\omega \sum_j \langle \mathcal{D} S_{\Gamma_j}(\tilde{u}_j^h), S_{\Gamma_j}(v)\rangle_{\Gamma_j} = (F, v), \quad v \in \mathcal{V}_{-1}^h,$$

$$\text{(6.1b)}\quad \sum_{j,k} \langle\langle \theta, S_{\Gamma_{jk}}(\tilde{u}_j^h)\rangle\rangle_{\Gamma_{jk}} = 0, \quad \theta \in \tilde{\Lambda}_{-1}^h,$$

where $\mathcal{A}_{h,Q}$ indicates the restriction to Q of the bilinear form $\mathcal{A}_h$ defined in (4.3) and $S_{\Gamma_{jk}}, S_{\Gamma_j}$ are defined as in (2.7)–(2.8). The following theorem gives an existence and uniqueness result for the procedure (6.1).

THEOREM 6.1. *Problem* (6.1) *has a unique solution.*

Proof. It is enough to show uniqueness due to finite dimensionality. For this, set $F = 0$ and add (6.1a) with the choice of $v = \tilde{u}^h$ and (6.1b) with the choice $\theta = \tilde{\lambda}^h$. Then the imaginary part in the resulting equation reduces to

$$\sum_j \sum_{Q\in\mathcal{T}^h(\Omega_j)} [\omega(\mathcal{B}\tilde{u}_j^h, \tilde{u}_j^h)_Q + (\mathbf{E}_i\tilde{\varepsilon}(\tilde{u}_j^h), \tilde{\varepsilon}(\tilde{u}_j^h))_Q] + \omega\langle \mathcal{D}S_{\Gamma_j}(\tilde{u}_j^h), S_{\Gamma_j}(\tilde{u}_j^h)\rangle_{\Gamma_j} = 0. \tag{6.2}$$

Now an argument similar to that given in the proof of Theorem 4.1 shows that

$$\tilde{u}^h = 0 \quad \text{in} \quad \Omega.$$

Thus (6.1a) reduces to

$$\sum_{j,k} \langle\langle \tilde{\lambda}^h, S_{\Gamma_{jk}}(v)\rangle\rangle_{\Gamma_{jk}} = 0, \quad v \in \mathcal{NC}_{-1}^h. \tag{6.3}$$

Now, for each Ω_j and each $Q \in \Omega_j$ with Q facing the boundary Γ, we can choose $v \in \mathcal{V}^h(\Omega_j)$ with the degrees of freedom chosen such that $S_{\Gamma_{jk}}(v)$ is equal to $\tilde{\lambda}^h$ at the midpoint m of one edge of Q and zero degrees of freedom at the other three midpoints of Q to show that $\tilde{\lambda}^h = 0$ at the midpoint m. Repeating the argument for all midpoints of Q and all $Q \in \Omega_j$ whose faces meet $\partial\Omega_j$ for each j yields that $\tilde{\lambda}^h = 0$. This completes the proof. ∎

We next notice the validity of the following lemma, whose obvious proof is omitted.

LEMMA 6.1. *If* $\tilde{u}^h \in \mathcal{V}_{-1}^h$, *then* $\tilde{u}^h \in \mathcal{V}^h$ *if and only if*

$$\sum_{j,k} \langle\langle \theta, S_{\Gamma_{jk}}(\tilde{u}^h)\rangle\rangle_{\Gamma_{jk}} = 0, \quad \theta \in \tilde{\Lambda}_{-1}^h.$$

Remark 6.1. As a consequence of Theorem 6.1 and Lemma 6.1, $\tilde{u}^h$ solves problem (4.4).

7. The domain decomposition iterative procedures. Consider the decomposition of problem (2.6) and (2.9) over Ω_j as follows: for $j = 1, \ldots, N$, find $u_j(x, \omega)$ satisfying

$$-\omega^2\mathcal{P}u_j + i\omega\mathcal{B}u_j - \mathcal{L}(u_j) = F, \quad \Omega_j, \tag{7.1a}$$

$$\mathcal{G}_{\Gamma_{jk}}(u_j) + i\omega\beta_{jk}S_{\Gamma_{jk}}(u_j) = \mathcal{G}_{\Gamma_{kj}}(u_k) - i\omega\beta_{jk}S_{\Gamma_{kj}}(u_k), \qquad \Gamma_{jk}, \tag{7.1b}$$

$$-\mathcal{G}_{\Gamma_j}(u_j) = i\omega\mathcal{D}S_{\Gamma_j}(u_j), \qquad \Gamma_j, \tag{7.1c}$$

where $\mathcal{G}_{\Gamma_{jk}}$ and $\mathcal{G}_{\Gamma_j}$ are defined as in (2.7)–(2.8). Notice that (7.1b) is equivalent to imposing the two consistency conditions

$$\mathcal{G}_{\Gamma_{jk}}(u_j) = \mathcal{G}_{\Gamma_{kj}}(u_k), \qquad \Gamma_{jk},$$
$$\beta_{jk}(S_{\Gamma_{jk}}(u_j) + S_{\Gamma_{kj}}(u_k)) = 0, \qquad \Gamma_{jk}.$$

A weak form of (7.1) at the differential level may be stated as follows: for all j, find $u_j \in [H^1(\Omega_j)]^2 \times H(\mathrm{div};\Omega_j) \times [H^1(\Omega_j)]^2$ such that

$$\begin{aligned} &-\omega^2(\mathcal{P}u_j, v)_{\Omega_j} + i\omega(\mathcal{B}u_j, v)_{\Omega_j} + \mathcal{A}_j(u_j, v) + i\omega\langle \mathcal{D}\ S_{\Gamma_j}(u_j), S_{\Gamma_j}(v)\rangle \\ &\quad + \sum_k \big\langle i\omega\beta_{jk}(\mathcal{S}_{\Gamma_{jk}}(u_j) + \mathcal{S}_{\Gamma_{kj}}(u_k)) - \mathcal{G}_{\Gamma_{jk}}(u_k), v\big\rangle_{\Gamma_{jk}} = (F, v)_{\Omega_j}, \\ &\quad v = (v^{(1)}, v^{(2)}, v^{(3)})^t \in [H^1(\Omega_j)]^2 \times H(\mathrm{div};\Omega_j) \times [H^1(\Omega_j)]^2, \end{aligned} \tag{7.2}$$

where $\mathcal{A}_j$ is the restriction to Ω_j of the bilinear form $\mathcal{A}$ defined in (3.3). Then a Jacobi-type iteration at the differential level may be defined by changing u_j and u_k in (7.2) into $u_j^{\{n\}}$ and $u_k^{\{n-1\}}$, respectively. This motivates the definition of our hybridized nonconforming domain decomposition procedure. For that purpose, we introduce a new set of Lagrange multipliers $\tilde{\tilde{\lambda}}_{jk}^h$ associated with $\mathcal{S}_{\Gamma_{jk}}(u_j)$ at the midpoints ξ_{jk} of the face of element $Q \in \Omega_j$ such that $\overline{Q} \cap \Gamma_{jk} \neq \emptyset$ for all the interior interfaces Γ_{jk}. Let

$$\tilde{\tilde{\Lambda}}_{-1,j}^h = \{\tilde{\tilde{\lambda}}_j^h : \tilde{\tilde{\lambda}}_{jk}^h = \tilde{\tilde{\lambda}}_j^h|_{\partial Q \cap \Gamma_{jk}} \in [P_0(\partial Q \cap \Gamma_{jk})]^5\ \forall Q \in \Omega_j \text{ such that } \overline{Q} \cap \Gamma_{jk} \neq \emptyset\},$$

and set $\tilde{\tilde{\Lambda}}_{-1}^h = \cup_j \tilde{\tilde{\Lambda}}_{-1,j}^h$.

Remark 7.1. Note that we have two copies of $[P_0(\Gamma_{jk})]^5$ on each Γ_{jk}, one from Ω_j to Ω_k and another from Ω_k to Ω_j.

An iterative procedure corresponding to (7.2) is defined as follows: for all $j = 1, \ldots, N$, choose an initial guess $(u_j^{\{h,0\}}, \tilde{\tilde{\lambda}}_j^{\{h,0\}}) \in \mathcal{V}^h(\Omega_j) \times \tilde{\tilde{\Lambda}}_{-1}^h$. Then, for $n = 1, 2, 3, \ldots$, compute $(u_j^{\{h,n\}}, \tilde{\tilde{\lambda}}_j^{\{h,n\}}) \in \mathcal{V}^h(\Omega_j) \times \tilde{\tilde{\Lambda}}_{-1,j}^h$ as the solution of the equations

$$\begin{aligned} &\sum_{Q \in \mathcal{T}^h(\Omega_j)} \big[-\omega^2\big(\mathcal{P}u_j^{\{h,n\}}, v\big)_Q + i\omega\big(\mathcal{B}u_j^{\{h,n\}}, v\big)_Q + \mathcal{A}_{h,Q}(u_j^{\{h,n\}}, v)\big] \\ &\quad + i\omega\big\langle \mathcal{D}\ S_{\Gamma_j}(u_j^{\{h,n\}}), S_{\Gamma_j}(v)\big\rangle_{\Gamma_j} + \sum_k \big\langle i\omega\beta_{jk}\mathcal{S}_{\Gamma_{jk}}(u_k^{\{h,n*\}}), \mathcal{S}_{\Gamma_{jk}}(v)\big\rangle_{\Gamma_{jk}} \\ &= (F, v)_{\Omega_j} - \sum_k \big\langle i\omega\beta_{jk}\mathcal{S}_{\Gamma_{jk}}(u_k^{\{h,n*\}}), \mathcal{S}_{\Gamma_{jk}}(v)\big\rangle_{\Gamma_{jk}} \\ &\quad + \sum_k \big\langle\big\langle \tilde{\tilde{\lambda}}_{kj}^{\{h,n*\}}, \mathcal{S}_{\Gamma_{jk}}(v)\big\rangle\big\rangle_{\Gamma_{jk}}, \quad v \in \mathcal{V}^h(\Omega_j), \end{aligned} \tag{7.3a}$$

$$\tilde{\tilde{\lambda}}_{jk}^{\{h,n\}} = \tilde{\tilde{\lambda}}_{kj}^{\{h,n*\}} - i\omega\beta_{jk}[S_{\Gamma_{jk}}(u_j^{\{h,n\}}) + S_{\Gamma_{kj}}(u_k^{\{h,n*\}})](\xi_{jk}) \quad \text{on } \Gamma_{jk} \forall k, \tag{7.3b}$$

for all $j = 1, \ldots, N$, where n^* is defined according to the iteration type as follows:

Jacobi type: $n^* = n - 1$.

Seidel type: $n^* = \begin{cases} n-1, & j < k, \\ n, & j > k, \end{cases}$

red-black type: $n^* = \begin{cases} n-1, & \Omega_j \text{ is red, i.e., } j \in I_R, \\ n, & \Omega_j \text{ is black, i.e., } j \in I_B. \end{cases}$

Here for the red-black type, the red and black parts of subdomains are given alternatively such that $\overline{\Omega} = \overline{[\cup_{j \in I_R} \Omega_j]} \cup \overline{[\cup_{j \in I_B} \Omega_j]}$. If, for $\{j,k\} \subset I_R$ or $\{j,k\} \subset I_B$, $\overline{\Omega}_j \cap \overline{\Omega}_k \neq \emptyset$, then $\overline{\Omega}_j \cap \overline{\Omega}_k$ consists of a common vertex (in 2D) or a common edge (in 3D) of Ω_j and Ω_k.

8. Convergence of the iterative procedure. Next, we analyze the convergence of the iterative procedure (7.3). For simplicity in the notation we consider the case $\beta_{jk} = \beta I$ with $\beta = \beta_R > 0$ and I being the identity matrix of suitable size.

It follows immediately from (6.1) that for j, k, $(\tilde{u}_j^h, \tilde{\lambda}_{jk}^h) \in \mathcal{V}^h(\Omega_j) \times \tilde{\Lambda}_{-1,j}^h$ satisfy the local equations

$$\sum_{Q \in \mathcal{T}^h(\Omega_j)} \left[-\omega^2(\mathcal{P}\tilde{u}^h, v)_Q + i\omega(\mathcal{B}\tilde{u}^h, v)_Q + \mathcal{A}_{h,Q}(\tilde{u}^h, v)\right]$$

$$-\sum_k \left\langle \left\langle \tilde{\lambda}_{jk}^h, v \right\rangle \right\rangle_{\Gamma_{jk}} + i\omega \left\langle \mathcal{D} S_{\Gamma_j}(\tilde{u}^h), S_{\Gamma_j}(v) \right\rangle_{\Gamma_j} = (F, v)_{\Omega_j},\ v \in \mathcal{V}^h(\Omega_j). \tag{8.1}$$

Also, since $\tilde{\lambda}_{jk}^h = \tilde{\lambda}_{kj}^h$, (6.1b) is equivalent to

$$\tilde{\lambda}_{jk}^h = \tilde{\lambda}_{kj}^h - i\omega\beta \left[S_{\Gamma_{jk}}(\tilde{u}_j^h) + S_{\Gamma_{kj}}(\tilde{u}_k^h)\right](\xi_{jk}) \quad \text{on } \Gamma_{jk} \forall k. \tag{8.2}$$

Since $\tilde{u}^h$ satisfies the error estimates given in Theorem 5.1, in order to show the convergence of the iteration procedure (7.3) it is sufficient to demonstrate that $u_j^{\{h,n\}} \to \tilde{u}_j^h$ and $\tilde{\tilde{\lambda}}_{jk}^{\{h,n\}} \to \tilde{\lambda}_{jk}^h$ as $n \to \infty$ for all j, k. For this, set

$$d_j^n = u_j^{\{h,n\}} - \tilde{u}_j^h, \quad x \in \Omega_j, \qquad \eta_{jk}^n = \tilde{\tilde{\lambda}}_{jk}^{\{h,n\}} - \tilde{\lambda}_{jk}^h \quad \text{on } \Gamma_{jk}.$$

Then, from (7.3)–(8.2), we obtain the following iteration error equations:

$$\sum_{Q \in \mathcal{T}^h(\Omega_j)} \left[-\omega^2 \left(\mathcal{P} d_j^n, v\right)_Q + i\omega \left(\mathcal{B} d_j^n, v\right)_Q + \mathcal{A}_{h,Q}(d_j^n, v)\right]$$

$$+ i\omega \left\langle \mathcal{D}\ S_{\Gamma_j}(d_j^n), S_{\Gamma_j}(v) \right\rangle_{\Gamma_j} - \sum_k \left\langle \left\langle \eta_{jk}^n, \mathcal{S}_{\Gamma_{jk}}(v) \right\rangle \right\rangle_{\Gamma_{jk}} = 0,\ v \in \mathcal{V}^h(\Omega_j), \tag{8.3a}$$

$$\eta_{jk}^n = \eta_{kj}^{n*} - i\omega\beta \left[S_{\Gamma_{jk}}(d_j^n) + S_{\Gamma_{kj}}(d_k^{n*})\right](\xi_{jk}) \quad \text{on } \Gamma_{jk} \forall k. \tag{8.3b}$$

Let us define the pseudoenergy R^n at the nth iteration step as follows:

$$R^n = R^n(d^n, \eta^n) = \sum_{j,k} |\eta_{jk}^n + i\omega\beta S_{\Gamma_{jk}}(d_j^n)(\xi_{jk})|_{0,\Gamma_{jk}}^2. \tag{8.4}$$

A similar argument to that in [21] shows that $d^n \to 0$ in $L^2(\Omega_j)$ and $\eta^n \to 0$ as n goes to ∞, so that the procedures (7.3) converge.

Let us turn to analyze the actual convergence rate by using a fixed point argument. Let $T_F : \mathcal{V}_{-1}^h \times \tilde{\tilde{\Lambda}}_{-1}^h \to \mathcal{V}_{-1}^h \times \tilde{\tilde{\lambda}}_{-1}^h$ be defined as follows: for any $(p, \theta) \in \mathcal{V}_{-1}^h \times \tilde{\lambda}_{-1}^h$, $(u, \eta) = T_F(p, \theta)$ is the solution of the equations

$$\sum_{Q\in\mathcal{T}^h(\Omega_j)} \big[-\omega^2 \left(\mathcal{P}u_j, v\right)_Q + i\omega \left(\mathcal{B}u_j, v\right)_Q + \mathcal{A}_{h,Q}(u_j, v)\big]$$

$$+ i\omega \left\langle \mathcal{D}\ S_{\Gamma_j}(u_j), S_{\Gamma_j}(v)\right\rangle_{\Gamma_j} + \sum_k \left\langle i\omega\beta_{jk}\mathcal{S}_{\Gamma_{jk}}(u_j), \mathcal{S}_{\Gamma_{jk}}(v)\right\rangle_{\Gamma_{jk}} \tag{8.5a}$$

$$= (F, v)_{\Omega_j} - \sum_k \left\langle i\omega\beta_{jk}\mathcal{S}_{\Gamma_{jk}}(p_k), \mathcal{S}_{\Gamma_{jk}}(v)\right\rangle_{\Gamma_{jk}} + \sum_k \left\langle\left\langle \theta_{kj}, \mathcal{S}_{\Gamma_{jk}}(v)\right\rangle\right\rangle_{\Gamma_{jk}},$$

$$v \in \mathcal{V}^h(\Omega_j),$$

$$\eta_{jk} = \theta_{kj} - i\omega\beta_{jk}\left[S_{\Gamma_{jk}}(u_j) + S_{\Gamma_{kj}}(p_k)\right](\xi_{jk}) \quad \text{on } \Gamma_{jk}\forall k, \tag{8.5b}$$

for all $j = 1, \ldots, N$.

Notice that $T_F(p,\theta) = T_0(p,\theta) + T_F(0,0)$ and (p,θ) is a fixed point of T_F if and only if

$$T_F(p,\theta) = (p,\theta) = T_0(p,\theta) + T_F(0,0).$$

Therefore, a fixed point of T_F is a solution of the equation

$$(I - T_0)(p,\theta) = T_F(0,0).$$

An estimate on the spectral radius of the operator T_0 can be obtained using similar arguments to those in [17, 21], as follows.

THEOREM 8.1. *Let $\rho(T_0)$ be the spectral radius of T_0. Then $\rho(T_0) < 1$ and consequently the iterative procedure* (7.3) *is convergent.*

9. The triangular and the three-dimensional cases.

9.1. The triangular element case. Let $\overline{\Omega} = \cup_{j=1}^J \overline{Q}_j$ be a quasi-regular partition of Ω into triangles Q_j's; here, Ω can be a convex polygon. Let us change the definition of the set $\mathcal{NC}_j^h$ in section 4 to $\mathcal{NC}_j^h = [P_1(Q_j)]^2$, with the degrees of freedom being the midpoint values of the edges of Q_j. Also, change the definition of the space $\mathcal{W}_j$ to be the vector part of the Raviart–Thomas–Nedelec mixed finite element space of zero order based on triangles [34, 30], with the degrees of freedom being the values of the normal component of the fluid displacement vector at the midpoints of the edges of Q_j.

An inspection of the analysis shows that all the conclusions presented for the rectangular case in Theorem 4.1 about the existence and uniqueness of the solution u^h of (4.4), the a priori error estimates in derived in Theorem 5.1, and the convergence of the iterative domain decomposition method in Theorem 8.1 remain valid for the new definition of the space $\mathcal{V}^h$.

9.2. The three-dimensional case. Let $Q_j, j = 1, \ldots, J$, be a nonoverlapping partition of Ω. If the Q_j's are tetrahedrons we take $\mathcal{NC}_j^h = [P_1(Q_j)]^3$. If the Q_j's are cubic elements, we set $\widehat{R} = (-1,1)^3$ and

$$\widehat{S}(\widehat{R}) = \mathrm{Span}\big\{1, \widehat{x}, \widehat{y}, \widehat{z}, \alpha(\widehat{x}) - \alpha(\widehat{y}), \alpha(\widehat{x}) - \alpha(\widehat{z})\big\} \tag{9.1}$$

$$= \mathrm{Span}\left\{\frac{1}{2}\widehat{x} \pm \frac{\alpha(\widehat{x})}{2\alpha(1)}, \frac{1}{2}\widehat{y} \pm \frac{\alpha(\widehat{y})}{2\alpha(1)}, \frac{1}{2}\widehat{z} \pm \frac{\alpha(\widehat{z})}{2\alpha(1)}\right\},$$

and choose $\mathcal{NC}_j^h = [S(Q_j)]^3$. The four and six degrees of freedom associated with the tetrahedron case and (9.1) are the values at the centers of the faces. Also, take $\mathcal{W}_j$ to

TABLE 1
Material properties of the frozen sandstone model.

Solid grain	bulk modulus, $K^{(s1)}$	38.7 GPa
	shear modulus, $\mu^{(s1)}$	39.6 GPa
	density, $\rho^{(1)}$	2650 kg/m^3
	permeability $\kappa^{(1),0}$	1.07 10^{-13} m^2
Ice	bulk modulus, $K^{(s3)}$	8.58 GPa
	shear modulus, $\mu^{(s3)}$	3.32 GPa
	density, $\rho^{(3)}$	920 kg/m^3
	permeability $\kappa^{(3),0}$	5 10^{-4} m^2
Fluid	bulk modulus, $K^{(f)}$	2.25 GPa
	density, $\rho^{(2)}$	1000 kg/m^3
	viscosity, η	10^{-6} cP
Air	bulk modulus, $K^{(a)}$	1.5 10^{-4} GPa
	shear modulus, $\mu^{(a)}$	0 GPa

be the Raviart–Thomas–Nedelec space of order zero over either tetrahedrons or cubic elements depending on Q_j[30].

Next, change the definitions of the spaces $\mathcal{V}^h$, $\mathcal{V}^h_{-1}$, $\tilde{\Lambda}^h_s, \tilde{\Lambda}^h_{-1,j}$, and $\tilde{\tilde{\Lambda}}^h_{-1,j}$ in the obvious fashion. With these changes in the definitions, all the results derived for the 2D case remain unchanged.

10. Numerical experiments. We performed wave propagation simulation in a sample of water-saturated partially frozen Berea sandstone, with an interior plane interface Γ defined by a change in ice content in the pores. In this case $\Omega^{(1)}$ and $\Omega^{(3)}$ correspond to the sandstone and ice, respectively. The material properties of the system, taken from [9, 12], are given in Table 1. Since we would like to run an experiment in which the slow waves can actually be observed in the low-frequency range, the water viscosity value was taken to be of 10^{-6} centipoise. The computational domain Ω is a square of side length $L = 3$ km with a uniform partition of Ω into squares of side length $h = L/261$. The absolute porosity is $\phi^{(a)} = .18$, with the ice content in the pores changing from $S^{(3)'} = 20$ percent in the lower layer to $S^{(3)'} = 82$ percent in the upper layer.

The source function is a point source representing a force applied to the rock frame in the vertical z direction, located at ($x_s = 1.5$ km, $z_s = 1.88$ km). It has the form $F = (F^{(1)}, F^{(2)}, F^{(3)})^t = (F^{(1)}, 0, 0)^t$,

$$F^{(1)} = \left(0, \frac{\partial \delta_{(x_s,z_s)}}{\partial z}\right)^t g(\omega),$$

where $\delta_{(x_s,z_s)}$ denotes the Dirac distribution and $g(\omega)$ is the Fourier transform of the waveform of central (dominant) frequency $f_0 = 12$ Hz given by

$$g(t) = -2\xi(t - t_0)e^{-\xi(t-t_0)^2},$$

with $\xi = 8\ f_0^2$, $t_0 = 1.25/f_0$.

For the calculation of the elastic coefficients we need values for the bulk and shear moduli of the two solid (dry) frames, denoted by $K_m^{(s1)}, K_m^{(s3)}, \mu_m^{(s1)}$, and $\mu_m^{(s3)}$, respectively. Following [24, 12, 37], it is assumed that $K_m^{(s1)} = 14.4$ GPa and that the

TABLE 2
Wave speeds and attenuation factors for all waves at frequency 12 Hz.

	Ice content 0.82		Ice content 0.20	
Wave	Phase velocity (km/s)	Attenuation (dB)	Phase velocity (km/s)	Attenuation (dB)
Fast P1 wave	4.316	$1.872\,10^{-3}$	3.723	$4.282\,10^{-2}$
Slow P2 wave	1.463	1.825	$7.281\,10^{-1}$	$1.151\,10^{1}$
Slow P3 wave	$9.577\,10^{-2}$	$4.053\,10^{1}$	$1.192\,10^{-1}$	6.562
Fast S1 wave	2.946	1.281	2.384	$2.202\,10^{-1}$
Slow S2 wave	$7.104\,10^{-1}$	$5.582\,10^{-2}$	$1.013\,10^{-1}$	$4.605\,10^{-1}$

modulus $\mu_m^{(s1)}$, $\mu_m^{(s3)}$, and $K_m^{(s3)}$ can be computed using a percolation-type model with critical exponent 3.8 [13] using the relations

$$\mu_m^{(sj)} = \left[\mu_m^{(sj),max} - \mu_m^{(sj),0}\right]\left[\frac{\phi^{(3)}}{1-\phi^{(1)}}\right]^{3.8} + \mu_m^{(sj),0}, \quad j = 1, 3,$$

$$K_m^{(s3)} = \left[K_m^{(s3),max} - K_m^{(s3),0}\right]\left[\frac{\phi^{(3)}}{1-\phi^{(1)}}\right]^{3.8} + K_m^{(s3),0},$$

where $\mu_m^{(s1),max}, \mu_m^{(s3),max}$, and $K_m^{(s3),max}$ are computed using the Kuster and Toksöz model [23], taking the known values of $K^{(s1)}, \mu^{(s1)}, K^{(s3)}, \mu^{(s3)}$ for the background medium with inclusions of air, with properties $K^{(a)}, \mu^{(a)}$ (see Table 1). The moduli $\mu_m^{(s1),0}$, $\mu_m^{(s3),0}$, and $K_m^{(s3),0}$ are appropriate reference values, which we take as

$$\mu_m^{(s1),0} = 13.3 \text{ GPa}, \quad K_m^{(s3),0} = \mu_m^{(s3),0} = 0.$$

The viscoelastic parameters describing the dissipative behavior of the saturated sandstone are given as follows: $T_{1,M} = (2\pi\,10)^{-1}$ms, $T_{2,M} = (2\pi\,10^9)^{-1}$ms, and the mean quality factors are taken to be $\widehat{Q}_M = 300$ for $M = K_G^{(1)}, K_G^{(3)}, \mu^{(1)}, \mu^{(3)}, K_{av}$. The value of the Kozeny–Carman constant was taken to be 5 [22]. Also, the coefficient b_{13} in the definition (2.4) of the mass and viscous coupling coefficients was taken to be zero.

Table 2 displays values of the phase velocity and attenuation factors at 12 Hz for the five different types of waves for the two-layer model used in this experiment.

The following figures present snapshots of the wave fields for this experiment, generated after solving (7.3) for 110 equally spaced temporal frequencies in the interval $(0, 12\,\text{Hz})$ and using an approximate inverse Fourier transform as explained in [16].

Figures 1, 2, and 3 show, respectively, snapshots of the vertical component of the particle velocity of the three phases at $t = 410$ ms, where we can observe that after arriving at the interface Γ, the direct P1 wave labeled P1D has generated the transmitted fast P1 wave labeled P1T-P1D and the slow P2 transmitted and reflected waves labeled P2R-P1D and P2T-P1D, respectively. Also, after arriving at Γ, the direct fast shear wave labeled S1D has generated the transmitted and reflected fast shear waves labeled S1T-S1D and S1R-S1D, respectively. In the snapshots for the ice and fluid phases in Figures 2 and 3 we can also observe the direct slow P2 wave front labeled P2D. The relative amplitudes among the snapshots in Figures 1, 2, and 3 are 1, 0.56873, and 0.023708, respectively. We observe that the slow P2 wave is observed better in the ice and fluid phases than in the solid matrix phase.

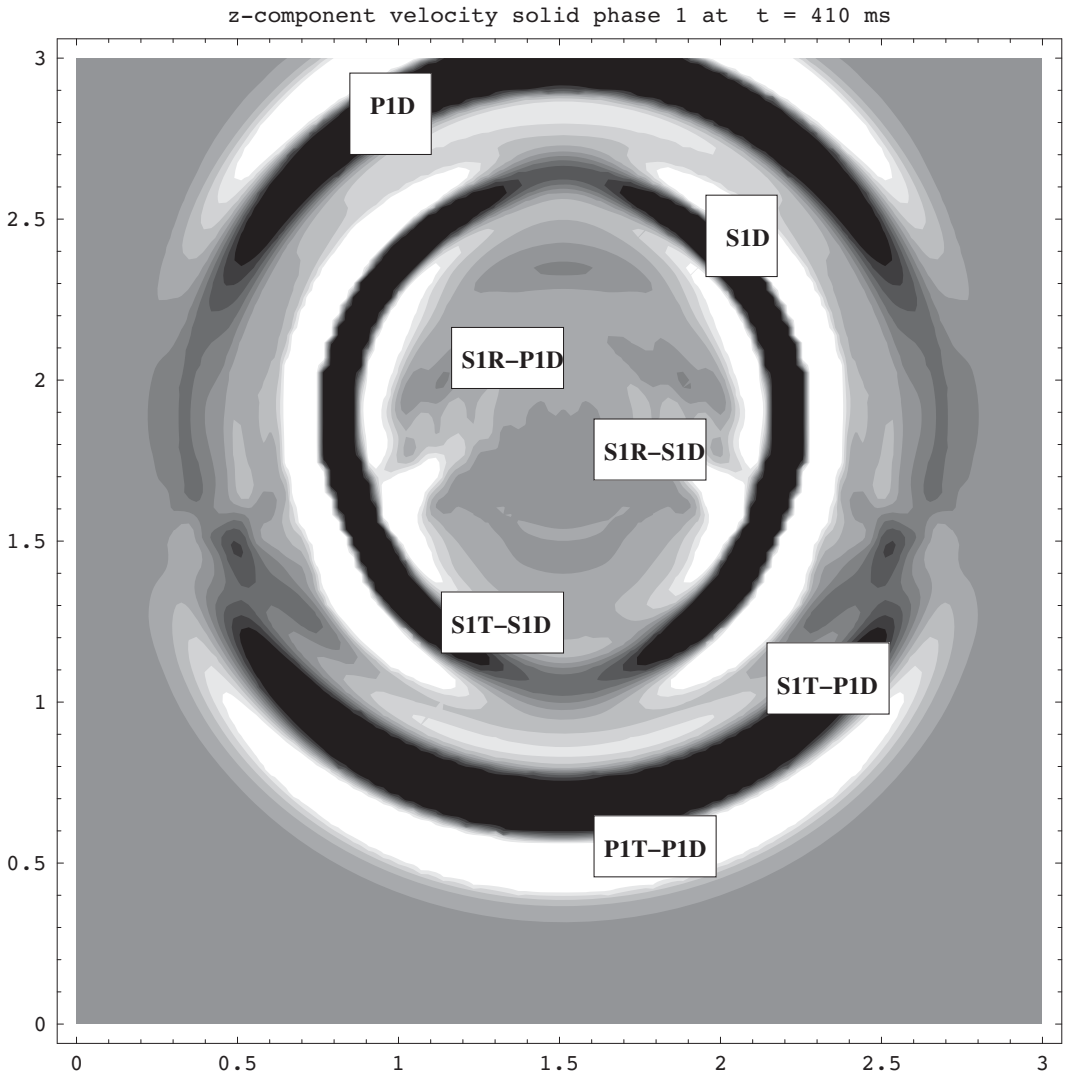


FIG. 1. *Snapshot of the vertical particle velocity of the solid matrix phase at* $t = 410$ *ms.*

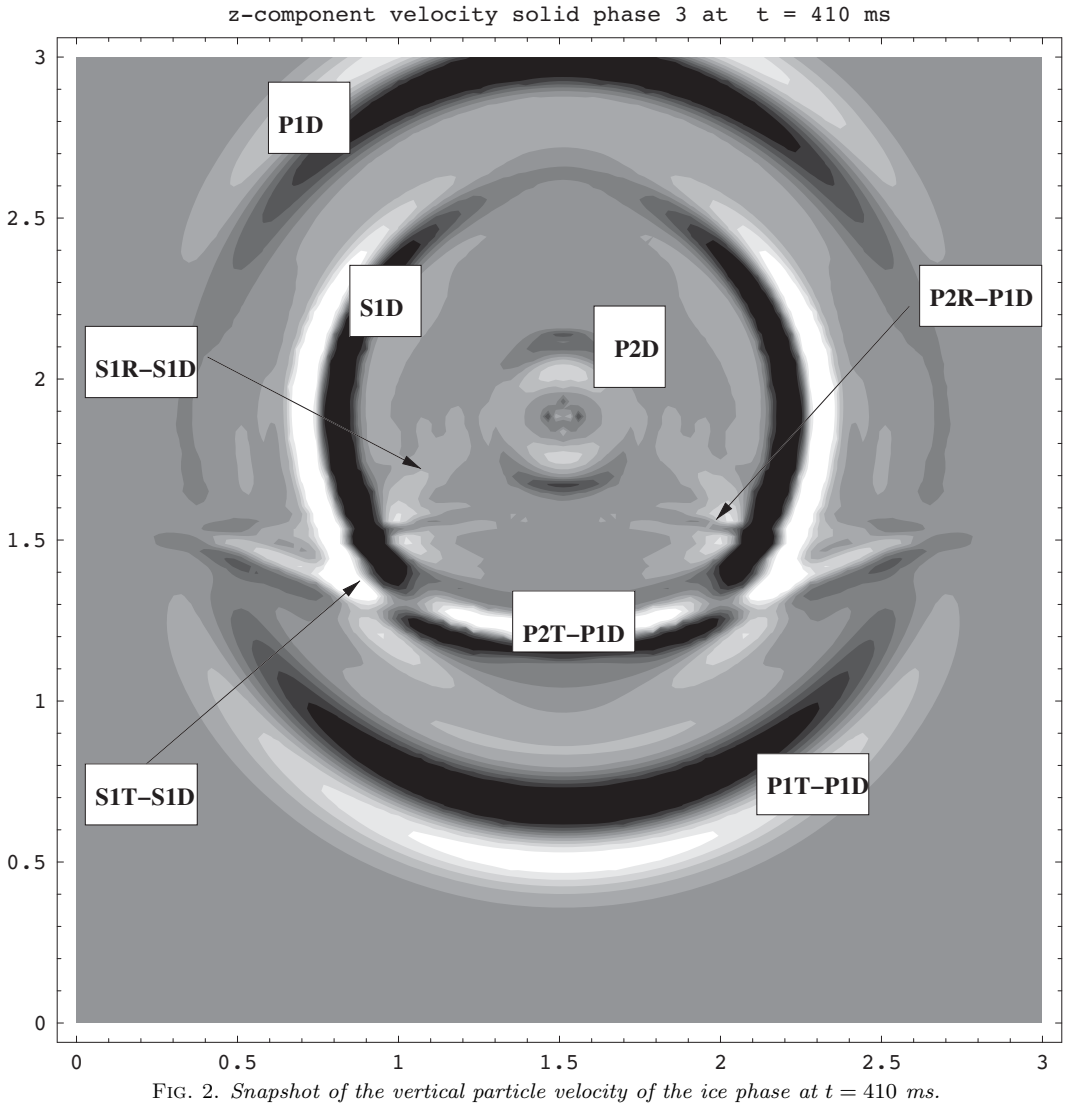


FIG. 2. *Snapshot of the vertical particle velocity of the ice phase at* $t = 410$ *ms.*

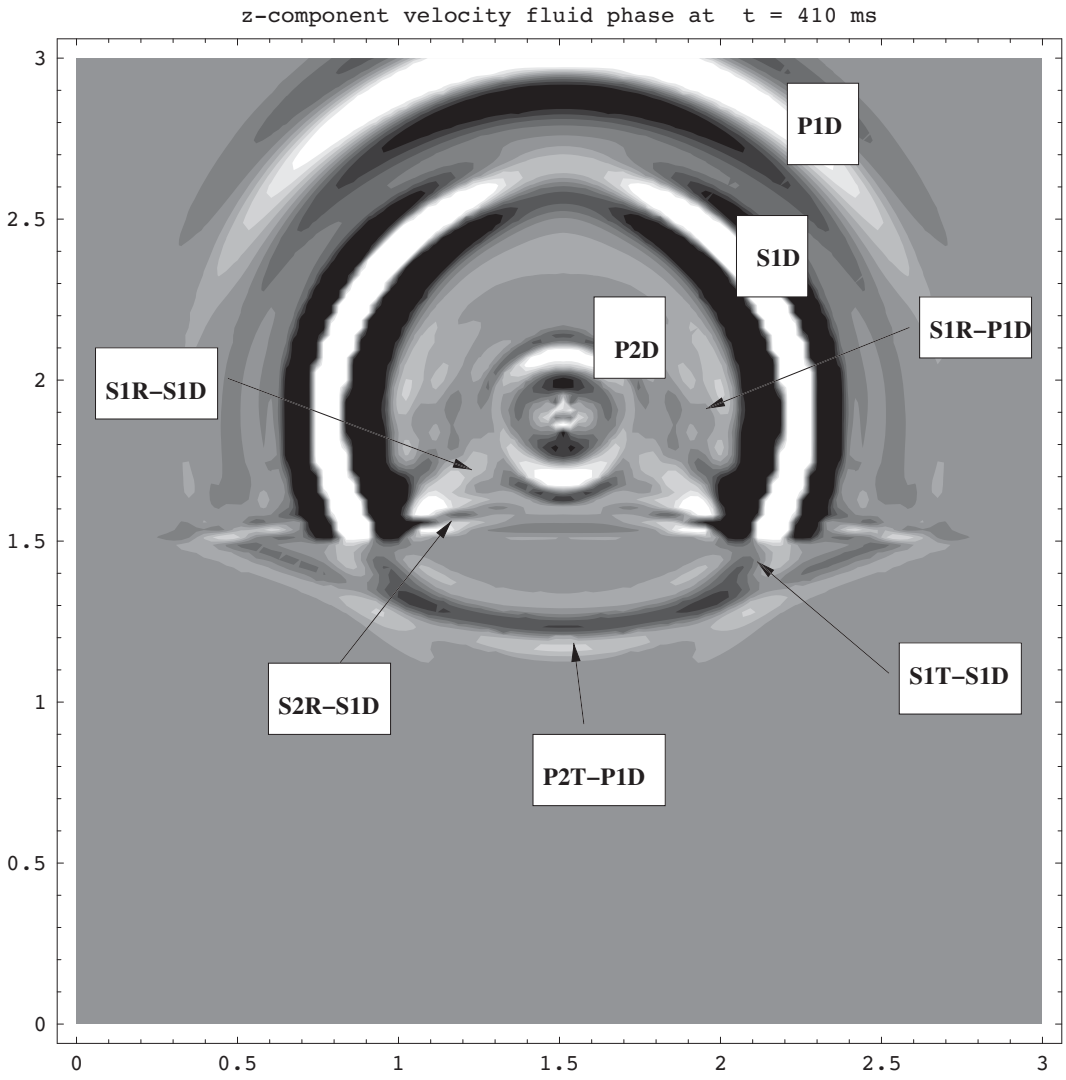


FIG. 3. *Snapshot of the vertical particle velocity of the fluid phase at* $t = 410$ *ms.*

REFERENCES

[1] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Anal. Numer., 19 (1985), pp. 7–32.

[2] J. Bear, *Dynamics of Fluids in Porous Media*, Dover Publications, New York, 1972.

[3] M. A. Biot, *Theory of deformation of a porous viscoelastic anisotropic solid*, J. Appl. Phys., 27 (1956), pp. 459–467.

[4] M. A. Biot, *Theory of propagation of elastic waves in a fluid-saturated porous solid.* II. *Higher frequency range*, J. Acoust. Soc. Amer., 28 (1956), pp. 179–191.

[5] M. A. Biot, *Mechanics of deformation and acoustic propagation in porous media*, J. Appl. Phys., 33 (1962), pp. 1482–1498.

[6] T. Bourbie, O. Coussy, and B. Zinszner, *Acoustics of Porous Media*, Editions Technip, Paris, 1987.

[7] J. M. Carcione, *Wave Fields in Real Media: Wave Propagation in Anisotropic, Anelastic, and Porous Media*, Handbook of Geophysical Exploration 31, Pergamon Press, Amsterdam, 2001.

[8] J. M. Carcione, B. Gurevich, and F. Cavallini, *A generalized Biot-Gassmann model for the acoustic properties of shaley sandstones*, Geophys. Prospecting, 48 (2000), pp. 539–557.

[9] J. M. Carcione, J. E. Santos, C. L. Ravazzoli, and H. B. Helle, *Wave simulation in partially frozen porous media with fractal freezing conditions*, J. Appl. Phys., 94 (2003), pp. 7839–7847.

[10] J. M. Carcione and G. Seriani, *Seismic velocities in permafrost*, Geophys. Prospecting, 46 (1998), pp. 441–454.

[11] J. M. Carcione and G. Seriani, *Wave simulation in frozen porous media*, J. Comput. Phys., 170 (2001), pp. 676–695.

[12] J. M. Carcione and U. Tinivella, *Bottom-simulating reflectors: Seismic velocities and AVO effects*, Geophysics, 65 (2000), pp. 54–67.

[13] D. Deptuck, J. P. Harrison, and P. Zawadzki, *Measurement of elasticity and conductivity in a three-dimensional percolation system*, Phys. Rev. Lett., 54 (1985), pp. 913–916.

[14] J. Douglas, Jr., P. L. Paes Leme, J. E. Roberts, and J. Wang, *A parallel iterative procedure applicable to the approximate solution of second order partial differential equations by mixed finite element methods*, Numer. Math., 65 (1993), pp. 95–108.

[15] J. Douglas, Jr., J. E. Santos, and D. Sheen, *Nonconforming Galerkin methods for the Helmholtz equation*, Numer. Methods Partial Differential Equations, 17 (2001), pp. 475–494.

[16] J. Douglas, Jr., J. E. Santos, D. Sheen, and L. Bennethum, *Frequency domain treatment of one-dimensional scalar waves*, Math. Models Methods Appl. Sci., 3 (1993), pp. 171–194.

[17] J. Douglas, Jr., J. E. Santos, D. Sheen, and X. Ye, *Nonconforming Galerkin methods based on quadrilateral elements for second order elliptic problems*, ESAIM Math. Model. Numer. Anal., 33 (1999), pp. 747–770.

[18] G. Duvaut and J.-L. Lions, *Inequalities in Mechanics and Physics*, Springer-Verlag, Berlin, Heidelberg, 1976.

[19] B. X. Fraeijs de Veubeke, *Displacement and equilibrium models in the finite element method*, in Stress Analysis, O. C. Zienkiewicz and G. Holister, eds., Wiley, New York, 1965, pp. 145–197.

[20] B. X. Fraeijs de Veubeke, *Stress function approach*, in International Congress on the Finite Element Method in Structural Mechanics, Bournemouth, UK, 1975, pp. 321–332.

[21] T. Ha, J. E. Santos, and D. Sheen, *Nonconforming finite element methods for the simulation of waves in viscoelastic solids*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 5647–5670.

[22] J. M. Hovem and G. D. Ingram, *Viscous attenuation of sound in saturated sand*, J. Acoust. Soc. Amer., 66 (1979), pp. 1807–1812.

[23] G. T. Kuster and M. N. Toksöz, *Velocity and attenuation of seismic waves in two-phase media: Part* 1, *Theoretical formulations*. Geophysics, 39 (1974), pp. 587–606.

[24] P. Leclaire, F. Cohen-Tenoudji, and J. Aguirre Puente, *Extension of Biot's theory of wave propagation to frozen porous media*, J. Acoust. Soc. Amer., 96 (1994), pp. 3753–3767.

[25] P. Leclaire, F. Cohen-Tenoudji, and J. Aguirre Puente, *Observation of two longitudinal and two transverse waves in a frozen porous medium*, J. Acoust. Soc. Amer., 97 (1995), pp. 2052–2055.

[26] S. Lee, P. Cornillon, and O. Campanella, *Propagation of ultrasound waves through frozen foods*, in Proceedings of the Annual Meeting and Food Expo., Anaheim, CA, 2002.

[27] H. P. Liu, D. L. Anderson, and H. Kanamori, *Velocity dispersion due to anelasticity: Implications for seismology and mantle composition*, Geophys. J. R. Astr. Soc., 147 (1976), pp. 41–58.
[28] J. J. McCoy, *Conditionally averaged response formulation for two-phase random mixtures*, J. Appl. Mech., 58 (1991), pp. 973–981.
[29] J. L. Morack and J. C. Rogers, *Seismic evidence of shallow permafrost beneath the islands in the Beafort Sea*, Arctic, 3 (1981), pp. 166–174.
[30] J. C. Nedelec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[31] L. Nirenberg, *Uniqueness in Cauchy problems for differential equations with constant leading coefficients*, Comm. Pure Appl. Math., 10 (1957), pp. 89–105.
[32] J. A. Nitsche, *On Korn's second inequality*, RAIRO Anal. Numer., 15 (1981), pp. 237–248.
[33] C. L. Ravazzoli and J. E. Santos, *A theory for wave propagation in porous rocks saturated by two-phase fluids under variable pressure conditions*, Bollettino di Geofisica Teorica ed Applicata, 46 (2005), pp. 261–285.
[34] P. A. Raviart and J. M. Thomas, *A mixed finite element method for* 2^{nd} *order elliptic problems*, in Mathematical Aspects of the Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.
[35] J. E. Santos, J. M. Corberó, C. L. Ravazzoli, and J. L. Hensley, *Reflection and transmission coefficients in fluid-saturated porous media*, J. Acoust. Soc. Amer., 91 (1992), pp. 1911–1923.
[36] J. E. Santos, J. Corberó, J. Douglas, Jr., and O. M. Lovera, *Finite element methods for a model for full waveform acoustic logging*, IMA J. Numer. Anal., 8 (1988), pp. 415–433.
[37] J. E. Santos, C. L. Ravazzoli, and J. M. Carcione, *A model for wave propagation in a composite solid matrix saturated by a single-phase fluid*, J. Acoust. Soc. Amer., 115 (2004), pp. 2749–2760.
[38] F. I. Zyserman, P. M. Gauzellino, and J. E. Santos, *Dispersion analysis of a nonconforming finite element method for the Helmholtz and elastodynamic equations*, Internat. J. Numer. Meth. Engrg., 58 (2003), pp. 1381–1395.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 421–442

AN ADAPTIVE FINITE ELEMENT METHOD FOR THE LAPLACE–BELTRAMI OPERATOR ON IMPLICITLY DEFINED SURFACES*

ALAN DEMLOW† AND GERHARD DZIUK‡

Abstract. We present an adaptive finite element method for approximating solutions to the Laplace–Beltrami equation on surfaces in $\mathbb{R}^3$ which may be implicitly represented as level sets of smooth functions. Residual-type a posteriori error bounds which show that the error may be split into a "residual part" and a "geometric part" are established. In addition, implementation issues are discussed and several computational examples are given.

Key words. Laplace–Beltrami operator, adaptive finite element methods, a posteriori error estimation, boundary value problems on surfaces

AMS subject classifications. 58J32, 65N15, 65N30

DOI. 10.1137/050642873

1. Introduction.

In this paper we derive residual-based a posteriori error estimates for piecewise linear finite element approximations to solutions of the Laplace–Beltrami equation

$$\begin{aligned} -\Delta_\Gamma u &= f \text{ on } \Gamma, \\ u &= 0 \text{ on } \partial\Gamma. \end{aligned} \tag{1.1.1}$$

Here Γ is a connected two-dimensional surface embedded in $\mathbb{R}^3$, and $-\Delta_\Gamma$ is the Laplace–Beltrami operator on Γ. $\partial\Gamma$ is required to be "piecewise curvilinear" in a sense which we will make precise below. We also allow $\partial\Gamma = \emptyset$, in which case the conditions $\int_\Gamma f \, d\sigma = \int_\Gamma u \, d\sigma = 0$ are required to guarantee existence and uniqueness of u. Here $d\sigma$ is the surface measure on Γ.

A finite element method for (1.1.1) was introduced in [Dz88]. Let Γ_h be a polyhedral approximation to Γ having triangular faces, and let S_h be the continuous functions which are affine on each face of Γ_h. We then let $u_h \in S_h$ solve

$$\int_{\Gamma_h} \nabla_{\Gamma_h} u_h \nabla_{\Gamma_h} v_h \, d\sigma_h = \int_{\Gamma_h} v_h f_h \, d\sigma_h \quad \forall \, v_h \in S_h. \tag{1.1.2}$$

Here ∇_{Γ_h} is the tangential derivative on Γ_h, σ_h is the surface measure on Γ_h, and f_h is an approximation to f on Γ_h. As above, we require the side conditions $\int_{\Gamma_h} f_h \, d\sigma_h = \int_{\Gamma_h} u_h \, d\sigma_h = 0$ if $\partial\Gamma_h = \emptyset$.

A key feature of our theoretical development is that Γ is represented as the 0 level set of a signed distance function d with $|d(x)| = dist(x, \Gamma)$. Our approach requires

*Received by the editors October 17, 2005; accepted for publication (in revised form) July 14, 2006; published electronically February 15, 2007. This research is based upon work partially supported by a National Science Foundation postdoctoral research fellowship and a grant of the Deutsche Forschungsgemeinschaft.

http://www.siam.org/journals/sinum/45-1/64287.html

†Department of Mathematics, University of Kentucky, Patterson Office Tower 715, Lexington, KY 40506-0027 (demlow@ms.uky.edu).

‡Abteilung für Angewandte Mathematik, Hermann-Herder-Str. 10, 79104 Freiburg, Germany (gerd@mathematik.uni-freiburg.de).

access to the derivatives of d (the normal vector and curvature tensor) and also requires that Γ_h lie in a strip about Γ on which a unique orthogonal projection $a(x)$ onto Γ is defined. This projection is instrumental in suitably defining the discrete data f_h and also in carrying out both a priori and a posteriori error analysis. Also, if $\partial\Gamma$ is nonempty we shall require that $\partial\Gamma = a(\partial\Gamma_h)$ so that $\partial\Gamma$ is in a sense piecewise curvilinear. This is similar to requiring polygonal boundaries when performing finite element calculations on domains in $\mathbb{R}^n$ in that it rules out "variational crimes" resulting from boundary approximations.

In practice, Γ often is defined as a level set of a function ζ which is not a distance function. In this situation one must approximate the projection $a(x)$ numerically, and the other necessary geometric information may then be computed in a straightforward fashion. In practical terms, the resulting finite element code requires the user to supply the data f, the level set function ζ and its first and second derivatives, and an initial mesh which lies in a sufficiently narrow strip about Γ to guarantee that the projection a is a bijection between Γ_h and Γ. In what follows we shall discuss some details of our implementation in addition to providing a posteriori error estimates.

Optimal-order $H^1(\Gamma)$ and $L_2(\Gamma)$ a priori estimates for the method (1.1.2) were proved in [Dz88]. Roughly speaking, the finite element error may be broken into an *almost-best-approximation* term typical of finite element methods in $\mathbb{R}^n$, a *geometric error term* which is due to the discretization of Γ, and a *data approximation term* due to the approximation of f on Γ by f_h on Γ_h. On a mesh whose elements have diameter h, the latter two terms are of order h^2 for typical choices of f_h and are thus of higher order when the error is measured in the H^1-norm.

In this paper we provide a posteriori error control in the $H^1(\Gamma)$-norm via residual-type estimators. As in the a priori analysis, the error is split into three terms: a *residual indicator term*, a *geometric error term*, and a *data approximation term*. Computation of these error terms requires pointwise access to geometric information, in particular to the projection a and the normal vector and curvature tensor on Γ. However, the asymptotically dominant term requires no explicit geometric quantities except those which are necessary to compute the discrete data f_h.

A relatively simple setting is assumed here in order to concentrate on effects arising from the discretization of Γ. In particular, we do not consider problems with nonconstant coefficients, the case where $a(\Gamma_h) \neq \Gamma$, lower-order terms, or nonhomogeneous Dirichlet or Neumann boundary conditions. These additional complexities may be handled in much the same way as for problems on polygonal domains in $\mathbb{R}^2$, so we refer, for example, to the works [DR98], [DW00], [BCD04], [MN05], and [AO00], where many of these issues are considered. Under suitable assumptions, our development also holds largely unchanged for surfaces of codimension 1 which are immersed in $\mathbb{R}^n$, $n \geq 2$.

In order to conclude the introduction, we briefly describe other strategies for performing adaptive finite element calculations on surfaces. One possibility is the use of a global parametrization to represent Γ and define a suitable mesh. This approach was taken in [AP05] to perform adaptive finite element calculations on the sphere. The key to this method is a global parametrization which maps a triangulated planar domain onto the sphere in such a way that the resulting curved "triangles" are isotropic (shape-regular). A more general approach using local parametrizations (charts) to represent 2- and 3-manifolds is described in [Ho01].

The technique we present here has several advantages when compared with the two described above. Extending the use of global parametrizations to surfaces other than the sphere is relatively difficult because a new parametrization must be found for

every surface on which computations are to be performed. In addition, the analysis of the Clemént-type interpolant used to prove reliability of a posteriori estimates in [AP05] is specific to the sphere and would have to be redone for other surfaces. In contrast, implementation of our method is quite straightforward for the sphere and may be carried out in a fairly general way for a large class of surfaces. The analysis we give here also is not restricted to any particular surface. The use of local parametrizations described in [Ho01] provides a framework for computations on manifolds which is in some ways more general than that which we propose here. However, the use of overlapping local charts adds to the complexity of both the resulting finite element code and the theoretical analysis. Indeed, the issue of approximation theory when using local charts is not addressed rigorously in [Ho01]. A final advantage of our approach is that it provides rigorous theoretical background for adaptive methods in certain situations in which no parametrization is available, such as implicit computations of surfaces evolving, for example, by mean curvature flow (cf. [Dz91], [BMN05], [CDDRR04]).

This paper is organized as follows: In section 2 we give a number of preliminaries and assumptions necessary for our theoretical development. In section 3 we then prove global a posteriori upper bounds and local lower bounds. In section 4 our implementation is described. In section 5, we demonstrate the flexibility of our approach by describing computational experiments on three different surfaces: a spherical subdomain with a nonempty boundary, a torus (which is nonconvex and has a topological type different than that of the sphere), and an ellipsoid (which requires numerical approximation of the distance function).

2. Preliminaries and assumptions.

2.1. The continuous surface Γ. We assume that Γ is a connected C^2 compact hypersurface which is the zero level set of a signed distance function $|d(x)| = dist(x, \Gamma)$ defined on an open subset U_0 of $\mathbb{R}^3$. If $\partial\Gamma = \emptyset$ we also assume for simplicity that $d < 0$ on the interior of Γ and $d > 0$ on the exterior. $\vec{\nu} = \nabla d$ is then the outward-pointing unit normal on Γ. Note that $|\vec{\nu}| = 1$ wherever d is defined. Let also $\mathbf{H} : \mathbb{R}^3 \to \mathbb{R}^{3\times 3}$ be the Weingarten map defined by

$$\mathbf{H}_{ij}(x) = \vec{\nu}_{i,x_j}(x) = \vec{\nu}_{j,x_i}(x), \tag{2.2.1}$$

that is, $\mathbf{H}(x) = D^2 d(x)$, and let $\kappa_i(x)$, $i = 1, 2$, and 0 be the eigenvalues of $\mathbf{H}(x)$. For $x \in \Gamma$, κ_1 and κ_2 are the principal curvatures.

Next we define the projection

$$a(x) = x - d(x)\vec{\nu}(x). \tag{2.2.2}$$

We then let $U \subset \mathbb{R}^3$ be a strip of width δ about Γ, where $\delta > 0$ is sufficiently small to ensure that the decomposition

$$x = a(x) + d(x)\vec{\nu}(x) \tag{2.2.3}$$

is unique for $x \in U$. We require that

$$\delta < \left[\max_{i=1,2} \|\kappa_i\|_{L_\infty(\Gamma)}\right]^{-1}. \tag{2.2.4}$$

For $x \in U$, we also note the useful formula

$$\kappa_i(x) = \frac{\kappa_i(a(x))}{1 + d(x)\kappa_i(a(x))} \tag{2.2.5}$$

for the curvature of parallel surfaces, cf. Lemma 14.17 of [GT98].

The condition (2.2.4) is sufficient to ensure that the decomposition (2.2.3) is *locally* unique (cf. [GT98, section 14.6]), but we require that it be globally unique. This global requirement is a simplifying assumption which restricts our presentation to embedded surfaces. Immersed surfaces (including surfaces with self-intersections) could also be considered with slight changes to our presentation.

We may uniquely extend a function ψ defined on Γ to U by

$$\psi^\ell(x) = \psi(a(x)) \tag{2.2.6}$$

for $x \in U$. Let

$$\mathbf{P} = \mathbf{I} - \vec{\nu} \otimes \vec{\nu}, \tag{2.2.7}$$

where $\otimes$ is the tensor or outer product $\vec{a} \otimes \vec{b} = \vec{a}\vec{b}^T$ (vectors here are in column form). We then define the tangential gradient

$$\nabla_\Gamma \psi = \nabla \psi^\ell - (\vec{\nu} \cdot \nabla \psi^\ell)\vec{\nu} = \mathbf{P} \nabla \psi^\ell \tag{2.2.8}$$

for ψ defined on Γ and extended to U via (2.2.6). Note that $\nabla_\Gamma \psi$ depends only on the values of ψ on Γ even though its definition formally involves the extension of ψ to Γ. Note also that $-\Delta_\Gamma = -\nabla_\Gamma \cdot \nabla_\Gamma$. Finally, we denote by $H^1(\Gamma)$ the functions on Γ having a tangential gradient in $L^2(\Gamma)$.

2.2. The discrete surface Γ_h and mesh $\mathcal{T}_h$. Let $\Gamma_h \subset U$ be a polyhedron consisting of a set $\mathcal{T}_h$ of triangular faces, that is, $\Gamma_h = \cup_{T \in \mathcal{T}_h} \overline{T}$. Let also $\vec{\nu}_h$ denote the (piecewise constant) unit outer normal on Γ_h, and let $\mathcal{N}$ denote the set of nodes of triangles in $\mathcal{T}_h$. We assume that $a : \Gamma_h \to \Gamma$ is bijective and that $\vec{\nu} \cdot \vec{\nu}_h > 0$ everywhere on Γ_h. We note that it is often simplest to define Γ_h so that $\mathcal{N} \subset \Gamma$, but this is not theoretically required in any way. Also denote by h_T the diameter of $T \in \mathcal{T}_h$. Given $z \in \mathcal{N}$, we define the patch $\omega_z = interior(\cup_{\overline{T} \ni z} \overline{T})$ and let $h_z = \max_{T \subset \omega_z} h_T$. Also, let $\mathcal{E}$ denote the set of edges of triangles in $\mathcal{T}_h$. Finally, $\varphi_z \in S_h$ denotes the canonical basis function associated to z, that is, $\varphi_{z_i}(z_j) = \delta_{ij}$ for $z_i, z_j \in \mathcal{N}$.

Analyses of a posteriori estimates for finite element methods on domains in $\mathbb{R}^n$ typically assume that the underlying mesh is shape-regular, that is, all elements in $\mathcal{T}_h$ have a bounded aspect ratio. Under this assumption, constants depending on the aspect ratio of the elements of the mesh are then bounded and may be absorbed into a global constant of moderate size. This approach is reasonable because typical mesh refinement algorithms preserve shape-regularity.

The situation is somewhat more complicated in the current context of finite element methods on surfaces. The first issue which arises is that the mesh is perturbed after each refinement by projecting newly created nodes onto the continuous surface Γ via a. While these perturbations are asymptotically negligible, we are not aware of a proof that the refinement/perturbation procedure described here maintains shape regularity beginning from an arbitrary shape-regular mesh with nodes on Γ and lying in U. A second problem is that, in contrast to the situation in $\mathbb{R}^n$, shape regularity does not automatically imply that the number of triangles sharing a given node is bounded. However, if the number of elements in the patches of the initial coarse mesh used to begin the refinement algorithm is bounded, we may guarantee that such a bound will hold for all subsequent meshes by applying a suitable refinement algorithm. This is, in particular, the case for the newest-node subdivision algorithm.

As we have not been able to theoretically guarantee that a family of meshes maintains shape regularity under mesh refinement, we take the following approach.

We first prove a posteriori bounds which do not assume shape regularity. In these estimates, lack of shape regularity is penalized by a single factor. In our computational examples we do not include this penalty factor in our estimator but instead monitor it to ensure that mesh quality remains acceptable. In our examples, the penalty term remains of moderate size when refining meshes which initially lie within U.

2.3. Lifts and extensions of functions. Given a function v_h defined on Γ_h, we define the lift $\tilde{v}_h$ by $v_h(x) = \tilde{v}_h(a(x))$ for $x \in \Gamma_h$. We may then extend $\tilde{v}_h$ to U by (2.2.6). For v_h defined on Γ_h and $x \in U$, we thus define

$$v_h^\ell(x) = \tilde{v}_h(a(x)). \tag{2.2.9}$$

The overall effect of (2.2.9) is to extend v_h defined on Γ_h to U. Formally, however, this operation consists of a lift to Γ followed by extension to U. We emphasize that *all* extentions of functions to U referred to in this paper are constant along normals to Γ. Thus for our purposes extensions of functions defined on Γ and of functions defined on Γ_h have essentially the same properties.

Letting $\vec{\nu}_h$ denote the normal on Γ_h, we define for $x \in \Gamma_h$

$$\mathbf{P}_h(x) = \mathbf{I} - \vec{\nu}_h(x) \otimes \vec{\nu}_h(x) \tag{2.2.10}$$

so that, for V defined on U and $x \in \Gamma_h$,

$$\nabla_{\Gamma_h} V(x) = \mathbf{P}_h \nabla V(x). \tag{2.2.11}$$

We see from (2.2.2) and (2.2.9) that, for $x \in \Gamma_h$ and v_h defined on Γ_h,

$$\nabla v_h^\ell(x) = [(\mathbf{P} - d\mathbf{H})(x)] \nabla v_h^\ell(a(x)). \tag{2.2.12}$$

Since $\vec{\nu} \cdot \vec{\nu} \equiv 1$, we have $\vec{\nu}\mathbf{H} = \mathbf{H}\vec{\nu} = 0$ and $\mathbf{P}\mathbf{H} = \mathbf{H}\mathbf{P} = \mathbf{H}$ so that, for $x \in \Gamma_h$,

$$\nabla v_h^\ell(x) = [(\mathbf{I} - d\mathbf{H})(x)][\mathbf{P}(x)] \nabla v_h^\ell(a(x)) = [(\mathbf{I} - d\mathbf{H})(x)] \nabla_\Gamma v_h^\ell(a(x)). \tag{2.2.13}$$

Thus

$$\nabla_{\Gamma_h} v_h(x) = \nabla_{\Gamma_h} v_h^\ell(x) = [\mathbf{P}_h(x)][(\mathbf{I} - d\mathbf{H})(x)][\mathbf{P}(x)] \nabla_\Gamma v_h^\ell(a(x)). \tag{2.2.14}$$

Correspondingly, for $\psi \in H^1(\Gamma)$ (2.2.6) yields

$$\nabla_{\Gamma_h} \psi^\ell(x) = [\mathbf{P}_h(x)][(\mathbf{I} - d\mathbf{H})(x)][\mathbf{P}(x)] \nabla_\Gamma \psi(a(x)). \tag{2.2.15}$$

For $x \in \Gamma_h$, (2.2.13) yields

$$\nabla_\Gamma v_h^\ell(a(x)) = [(\mathbf{I} - d\mathbf{H})(x)]^{-1} \nabla v_h^\ell(x). \tag{2.2.16}$$

The invertibility of $[(\mathbf{I} - d\mathbf{H})(x)]$ on U may be derived from (2.2.4) and (2.2.5). Indeed, if e_1 and e_2 are the eigenvectors of $\mathbf{H}$ corresponding to $\kappa_1(x)$ and $\kappa_2(x)$, then $[(\mathbf{I} - d\mathbf{H})(x)]^{-1} = \vec{\nu} \otimes \vec{\nu} + (1 + d(x)\kappa_1(a(x)))e_1 \otimes e_1 + (1 + d(x)\kappa_2(a(x)))e_2 \otimes e_2$. We shall need to compute $\nabla_\Gamma v_h^\ell$ when $v_h \in S_h$. In such cases we initially have access only to the tangential derivative $\nabla_{\Gamma_h} v_h$ and not to ∇v_h^ℓ, which according to (2.2.16) is necessary to compute $\nabla_\Gamma v_h^\ell$. Since $\nabla_{\Gamma_h} v_h(x) = [\mathbf{P}_h(x)] \nabla v_h^\ell(x)$, we have $0 = \nabla v_h^\ell(x) \cdot \vec{\nu} = \nabla_{\Gamma_h} v_h(x) \cdot \vec{\nu} + (\vec{\nu}_h \cdot \vec{\nu}) \nabla v_h^\ell(x) \cdot \vec{\nu}_h$. Thus

$$\nabla v_h^\ell(x) \cdot \vec{\nu}_h = -\frac{\nabla_{\Gamma_h} v_h(x) \cdot \vec{\nu}}{\vec{\nu}_h \cdot \vec{\nu}}, \tag{2.2.17}$$

and for $x \in \Gamma_h$

$$\nabla v_h^\ell(x) = \left[\mathbf{I} - \frac{\vec{\nu}_h \otimes \vec{\nu}}{\vec{\nu}_h \cdot \vec{\nu}}\right] \nabla_{\Gamma_h} v_h(x). \tag{2.2.18}$$

Combining (2.2.16) and (2.2.18), we thus find that, for $x \in \Gamma_h$,

$$\nabla_\Gamma v_h^\ell(a(x)) = [(\mathbf{I} - d\mathbf{H})(x)]^{-1} \left[\mathbf{I} - \frac{\vec{\nu}_h \otimes \vec{\nu}}{\vec{\nu}_h \cdot \vec{\nu}}\right] \nabla_{\Gamma_h} v_h(x). \tag{2.2.19}$$

Next we state an integral equality which we shall use repeatedly. For $x \in \Gamma_h$, let

$$\mu_h(x)\, \mathrm{d}\sigma_h(x) = \mathrm{d}\sigma(a(x)), \tag{2.2.20}$$

and also let

$$\mathbf{A}_h(x) = \mathbf{A}_h^\ell(a(x)) = \frac{1}{\mu_h(x)} [\mathbf{P}(x)][(\mathbf{I} - d\mathbf{H})(x)][\mathbf{P}_h(x)][(\mathbf{I} - d\mathbf{H})(x)][\mathbf{P}(x)]. \tag{2.2.21}$$

Then from (2.2.14) and (2.2.15), we have

$$\int_{\Gamma_h} \nabla_{\Gamma_h} v_h \nabla_{\Gamma_h} \psi_h \, \mathrm{d}\sigma_h = \int_\Gamma \mathbf{A}_h^\ell \nabla_\Gamma v_h^\ell \nabla_\Gamma \psi_h^\ell \, \mathrm{d}\sigma. \tag{2.2.22}$$

Note that this equality holds without regard to the original domain of definition of v_h and ψ; that is, we may for example replace ψ_h and ψ_h^ℓ with ψ^ℓ and ψ, respectively, where $\psi \in H^1(\Gamma)$. We also emphasize that the quantities d and $\mathbf{H}$ in (2.2.21) are always evaluated on the discrete surface Γ_h, even though $\mathbf{A}_h^\ell$ often appears in integrals over the continuous surface Γ.

Finally we give an explicit formula for the quantity μ_h defined above. The proof of this formula is tedious but elementary, and we sketch it in Appendix A.

PROPOSITION 2.1. *Assume that $x \in \Gamma_h$. Then*

$$\mu_h(x) = (1 - d(x)\kappa_1(x))(1 - d(x)\kappa_2(x))\vec{\nu} \cdot \vec{\nu}_h. \tag{2.2.23}$$

2.4. Interpolation and Poincaré inequality. In this section we define an interpolant and prove error bounds for it. Given $\psi \in L_1(\Gamma)$ and $z \in \mathcal{N}$, we let

$$\psi_z^\ell = \frac{1}{\int_{\omega_z} \varphi_z \, \mathrm{d}\sigma_h} \int_{\omega_z} \varphi_z \psi^\ell \, \mathrm{d}\sigma_h \tag{2.2.24}$$

and define

$$I_h \psi^\ell = \sum_{z \in \mathcal{N}} \psi_z^\ell \varphi_z. \tag{2.2.25}$$

A similar interpolant is used, for example, in [FV06] to prove a posteriori bounds on a domain in $\mathbb{R}^2$. Noting that $\{\varphi_z\}_{z \in \mathcal{N}}$ is a partition of unity, we then have the following relationship:

$$\int_{\Gamma_h} (\psi^\ell - I_h \psi^\ell) \, \mathrm{d}\sigma_h = \sum_{z \in \mathcal{N}} \int_{\omega_z} (\psi^\ell - \psi_z^\ell) \varphi_z \, \mathrm{d}\sigma_h = 0. \tag{2.2.26}$$

LEMMA 2.2 (Poincaré inequality). *Let $\psi \in H^1(\Gamma)$. Let m_z be the number of elements sharing the node z, and let $\tilde{\omega}_z$ be the lift of the patch ω_z onto Γ . Then for each $z \in \mathcal{N}$,*

$$\|\psi^\ell - \psi_z^\ell\|_{L_2(\omega_z)} \le C \max_{T\subset\omega_z} \sqrt{|T|} m_z \max_{T\subset\omega_z} \frac{h_T}{\sqrt{|T|}} \|\mathbf{A}_h\|^{\frac{1}{2}}_{\ell_2,L_\infty(\omega_z)} \|\nabla_\Gamma \psi\|_{L_2(\tilde{\omega}_z)}. \tag{2.2.27}$$

Let also $z \in \overline{e} \in \mathcal{E}$. Then

$$\|\psi^\ell - \psi_z^\ell\|_{L_2(e)} \le C\sqrt{|e|} m_z \max_{T\subset\omega_z} \frac{h_T}{\sqrt{|T|}} \|\mathbf{A}_h\|^{\frac{1}{2}}_{\ell_2,L_\infty(\omega_z)} \|\nabla_\Gamma \psi\|_{L_2(\tilde{\omega}_z)}. \tag{2.2.28}$$

Here $\|\mathbf{A}_h\|_{\ell_2,L_\infty(\omega_z)} = \|\|\mathbf{A}_h\|_{\ell_2\to\ell_2}\|_{L_\infty(\omega_z)}$, where $\|\cdot\|_{\ell_2\to\ell_2}$ is the standard matrix 2-norm, and C does not depend on any essential quantities.

Remark 2.3. The terms in (2.2.27) and (2.2.28) may be classified as follows: In shape-regular meshes the quantities $\max_{T\subset\omega_z}\sqrt{|T|}$ and $\sqrt{|e|}$ may be reduced to h_z and $h_z^{1/2}$, respectively, where h_z is the maximum element diameter in ω_z. The factor m_z accounts for the number of elements sharing the vertex z if this number is not known to be bounded, and the factor $\max_{T\subset\omega_z} h_T/\sqrt{|T|}$ accounts for the aspect ratio of T. If $\mathcal{T}_h$ is shape-regular and m_z is bounded, we thus have

$$\|\psi^\ell - \psi_z^\ell\|_{L_2(\omega_z)} \le C h_z \|\mathbf{A}_h\|^{\frac{1}{2}}_{\ell_2,L_\infty(\omega_z)} \|\nabla_{\Gamma_h} \psi^\ell\|_{L_2(\omega_z)}, \tag{2.2.29}$$

$$\|\psi^\ell - \psi_z^\ell\|_{L_2(e)} \le C h_z^{\frac{1}{2}} \|\mathbf{A}_h\|^{\frac{1}{2}}_{\ell_2,L_\infty(\omega_z)} \|\nabla_\Gamma \psi\|_{L_2(\tilde{\omega}_z)}, \tag{2.2.30}$$

where C does not depend on essential quantities.

Proof. We first show that I_h is locally L_2-stable independent of mesh properties. Note first that

$$\|\psi_z^\ell\|_{L_2(\omega_z)} = |\omega_z|^{1/2}|\psi_z^\ell| \le |\omega_z|^{1/2} \frac{\|\varphi_z\|_{L_2(\omega_z)}}{\|\varphi_z\|_{L_1(\omega_z)}} \|\psi^\ell\|_{L_2(\omega_z)}. \tag{2.2.31}$$

If z is an interior node, we let $\hat{\omega}_z$ be a regular m_z-gon with vertices lying on the unit circle. If z is a boundary node, then we let $\hat{\omega}_z$ be one half of a regular $2m_z$-gon with vertices lying on the unit circle. In either case the reference domain $\hat{\omega}_z$ is convex and may be broken into m_z congruent triangles with the origin being a vertex of each. There is a natural piecewise-affine transformation $F_z : \hat{\omega}_z \to \omega_z$. We denote by $\hat{T}$ the inverse image of $T \subset \omega_z$ and by $\hat{u}$ the inverse image of $u \in H^1(\Gamma_h)$ under this transformation. For $p = 1$ or $p = 2$ and any $\hat{T} \subset \hat{\omega}_z$,

$$\|\varphi_z\|^p_{L_p(\omega_z)} = \sum_{T\subset\omega_z} \int_T \varphi_z^p \,\mathrm{d}\sigma_h = \sum_{T\subset\omega_z} \frac{|T|}{|\hat{T}|} \int_{\hat{T}} \hat{\varphi}_z^p \,\mathrm{d}\hat{x} = |\omega_z| \frac{\int_{\hat{T}} \hat{\varphi}_z^p \,\mathrm{d}\hat{x}}{|\hat{T}|}. \tag{2.2.32}$$

An elementary calculation yields $\sqrt{|\hat{T}|}\frac{\|\hat{\varphi}_z\|_{L_2(\hat{T})}}{\|\hat{\varphi}_z\|_{L_1(\hat{T})}} = \sqrt{\frac{3}{2}}$, which when combined with (2.2.31) and (2.2.32) yields

$$\|\psi_z^\ell\|_{L_2(\omega_z)} \le \sqrt{\frac{3}{2}} \|\psi^\ell\|_{L_2(\omega_z)}. \tag{2.2.33}$$

Thus for any $K \in \mathbb{R}$,

$$\|\psi^\ell - \psi_z^\ell\|_{L_2(\omega_z)} \le \left(1 + \sqrt{\frac{3}{2}}\right) \|\psi^\ell - K\|_{L_2(\omega_z)}. \tag{2.2.34}$$

Choosing $K = \frac{1}{|\hat{\omega}_z|}\int_{\hat{\omega}_z} \hat{\psi}^\ell \,\mathrm{d}\hat{x}$ and noting that $\nabla\hat{\psi}^\ell = \nabla_{\Gamma_h}\psi^\ell \nabla F_z$, we next find that

$$
\begin{aligned}
\|\psi^\ell - K\|^2_{L_2(\omega_z)} &= \sum_{T\subset\omega_z} \frac{|T|}{|\hat{T}|}\int_{\hat{T}} (\hat{\psi}^\ell - K)^2 \,\mathrm{d}\hat{x} \\
&= \frac{1}{|\hat{T}|}\max_{T\subset\omega_z}|T| \int_{\hat{\omega}_z} (\hat{\psi}^\ell - K)^2 \,\mathrm{d}\hat{x} \\
&\le C_P(\hat{\omega}_z)^2 \frac{1}{|\hat{T}|}\max_{T\subset\omega_z}|T| \int_{\hat{\omega}_z} |\nabla\hat{\psi}^\ell|^2 \,\mathrm{d}\hat{x} \\
&= C_P(\hat{\omega}_z)^2 \frac{1}{|\hat{T}|}\max_{T\subset\omega_z}|T| \sum_{T\subset\omega_z}\frac{|\hat{T}|}{|T|}\int_T |\nabla_{\Gamma_h}\psi^\ell \nabla F_z|^2 \,\mathrm{d}\sigma_h \\
&\le C_P(\hat{\omega}_z)^2 \max_{T\subset\omega_z}|T| \max_{T\subset\omega_z}\frac{\|\nabla F_z|_T\|^2_{\ell_2\to\ell_2}}{|T|}\|\nabla_{\Gamma_h}\psi^\ell\|^2_{L_2(\omega_z)}.
\end{aligned}
\tag{2.2.35}
$$

Here C_P is the Poincaré constant for $\hat{\omega}_z$. It is not hard to compute that

$$
\|\nabla F_z|_T\|_{\ell_2\to\ell_2} \le C m_z h_T,
\tag{2.2.36}
$$

where C does not depend on any essential quantities. Combining (2.2.34), (2.2.35), and (2.2.36) and finally applying (2.2.22) yields (2.2.27).

The proof of (2.2.28) is accomplished by employing a trace inequality and slightly modifying the preceding proof. Assume that $e \subset \overline{T} \subset \overline{\omega_z}$. Let $\check{T}$ be the unit simplex in $\mathbb{R}^2$ (note that this is not the same as the reference element employed above), with $\hat{e}$ denoting the transformation of e to $\check{T}$. Letting $\hat{F}_T$ denote the affine transformation of $\check{T}$ to T, we note that $\|\hat{F}_T\|_{\ell_2\to\ell_2} \le h_T$. Employing a trace inequality on $\check{T}$ then yields

$$
\begin{aligned}
\|\psi^\ell - \psi^\ell_z\|_{L_2(e)} &\le \sqrt{|e|}\|\hat{\psi}^\ell - \psi^\ell_z\|_{L_2(\hat{e})} \\
&\le C\sqrt{|e|}(\|\hat{\psi}^\ell - \psi^\ell_z\|_{L_2(\check{T})} + \|\nabla\hat{\psi}^\ell\|_{L_2(\check{T})}) \\
&\le C\sqrt{\tfrac{|e|}{|T|}}(\|\psi^\ell - \psi^\ell_z\|_{L_2(T)} + h_T\|\nabla_{\Gamma_h}\psi^\ell\|_{L_2(T)}).
\end{aligned}
\tag{2.2.37}
$$

$\sqrt{|e|/|T|}h_T\|\nabla_{\Gamma_h}\psi^\ell\|_{L_2(T)}$ is clearly bounded by the right-hand side of (2.2.28), so we must consider only the first term in the last line above.

Letting $\hat{T}$, $\hat{\omega}_z$, and K be as defined as before, we first procede as in (2.2.35) to find that

$$
\begin{aligned}
\sqrt{\tfrac{|e|}{|T|}}\|\psi^\ell - K\|_{L_2(T)} &\le \sqrt{\tfrac{|e|}{|\hat{T}|}}\|\hat{\psi}^\ell - K\|_{L_2(\hat{T})} \\
&\le \sqrt{\tfrac{|e|}{|\hat{T}|}}\|\hat{\psi}^\ell - K\|_{L_2(\hat{\omega}_z)} \\
&\le C\sqrt{|e|}m_z \max_{T\subset\omega_z}\frac{h_T}{\sqrt{|T|}}\|\nabla_{\Gamma_h}\psi^\ell\|_{L_2(\omega_z)}.
\end{aligned}
\tag{2.2.38}
$$

Proceeding as in (2.2.31) through (2.2.34), we next find that

$$
\begin{aligned}
\sqrt{\tfrac{|e|}{|T|}}\|(\psi^\ell - K)_z\|_{L_2(T)} &\le \sqrt{|e|}|(\psi_\ell - K)_z| \\
&\le \sqrt{\tfrac{3}{2}}\sqrt{\tfrac{|e|}{|\omega_z|}}\|\psi^\ell - K\|_{L_2(\omega_z)}.
\end{aligned}
\tag{2.2.39}
$$

Combining (2.2.39) and (2.2.35) yields

$$\begin{aligned}\sqrt{\frac{|e|}{|T|}}\|(\psi^\ell - K)_z\|_{L_2(T)} &\le C\sqrt{|e|}\max_{T\subset\omega_z}\sqrt{\tfrac{|T|}{|\omega_z|}}m_z\max_{T\subset\omega_z}\frac{h_T}{\sqrt{|T|}}\|\nabla_{\Gamma_h}\psi^\ell\|_{L_2(\omega_z)}\\ &\le C\sqrt{|e|}m_z\max_{T\subset\omega_z}\frac{h_T}{\sqrt{|T|}}\|\nabla_{\Gamma_h}\psi^\ell\|_{L_2(\omega_z)}.\end{aligned}\tag{2.2.40}$$

Since $\|\psi^\ell-\psi_z^\ell\|_{L_2(T)} \le \|\psi^\ell-K\|_{L_2(T)}+\|(\psi^\ell-K)_z\|_{L_2(T)}$, combining (2.2.37), (2.2.38), and (2.2.40) with (2.2.22) completes the proof of (2.2.28). ∎

3. The estimator. In this section we develop a computable and reliable estimator for $\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\Gamma)}$.

3.1. Residual equation. We first derive a residual equation. Let $\psi \in H_0^1(\Gamma)$, where $H_0^1(\Gamma)$ is the set of functions in $H^1(\Gamma)$ having a vanishing trace if $\partial\Gamma \neq \emptyset$ and having a vanishing mean value if $\partial\Gamma = \emptyset$. Following [Dz88] and applying (2.2.22), we find that, for $\psi \in H^1(\Gamma)$ and $\psi_h \in S_h$,

$$\begin{aligned}\int_\Gamma \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma\psi\,\mathrm{d}\sigma = &\int_{\Gamma_h} f^\ell\mu_h\psi^\ell\,\mathrm{d}\sigma_h - \int_\Gamma[\mathbf{P}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\nabla_\Gamma\psi\,\mathrm{d}\sigma\\ &- \int_{\Gamma_h}\nabla_{\Gamma_h}u_h\nabla_{\Gamma_h}\psi^\ell\,\mathrm{d}\sigma_h\end{aligned}\tag{3.3.1}$$

and

$$\begin{aligned}\int_\Gamma \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma\psi_h^\ell\,\mathrm{d}\sigma &= \int_{\Gamma_h} f^\ell\mu_h\psi_h\,\mathrm{d}\sigma_h - \int_\Gamma[\mathbf{P}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\nabla_\Gamma\psi_h^\ell\,\mathrm{d}\sigma\\ &\quad - \int_{\Gamma_h}\nabla_{\Gamma_h}u_h\nabla_{\Gamma_h}\psi_h\,\mathrm{d}\sigma_h\\ &= \int_{\Gamma_h}(f^\ell\mu_h - f_h)\psi_h\,\mathrm{d}\sigma_h - \int_\Gamma[\mathbf{P}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\nabla_\Gamma\psi_h^\ell\,\mathrm{d}\sigma.\end{aligned}\tag{3.3.2}$$

Combining (3.3.1) and (3.3.2), we find that

$$\begin{aligned}\int_\Gamma \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma\psi\,\mathrm{d}\sigma &= \int_\Gamma \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma\psi\,\mathrm{d}\sigma - \int_\Gamma \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma\psi_h^\ell\,\mathrm{d}\sigma\\ &\quad + \int_\Gamma \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma\psi_h^\ell\,\mathrm{d}\sigma\\ &= \int_{\Gamma_h} f^\ell\mu_h(\psi^\ell-\psi_h)\,\mathrm{d}\sigma_h - \int_{\Gamma_h}\nabla_{\Gamma_h}u_h\nabla_{\Gamma_h}(\psi^\ell-\psi_h)\,\mathrm{d}\sigma_h\\ &\quad - \int_\Gamma[\mathbf{P}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\nabla_\Gamma\psi\,\mathrm{d}\sigma + \int_{\Gamma_h}(f^\ell\mu_h - f_h)\psi_h\,\mathrm{d}\sigma_h.\end{aligned}\tag{3.3.3}$$

Next we note that

$$
\begin{aligned}
&-\int_{\Gamma_h} \nabla_{\Gamma_h} u_h \nabla_{\Gamma_h}(\psi^\ell - \psi_h)\,\mathrm{d}\sigma_h \\
&\qquad = \sum_{T\in\mathcal{T}_h} \int_T \Delta_{\Gamma_h} u_h(\psi^\ell - \psi_h)\,\mathrm{d}\sigma_h - \int_{\partial T} \nabla_{\Gamma_h} u_h \cdot \vec{n}(\psi^\ell - \psi_h)\,\mathrm{d}s \\
&\qquad = \int_{\Gamma_h} \Delta_{\Gamma_h} u_h(\psi^\ell - \psi_h)\,\mathrm{d}\sigma_h - \frac{1}{2}\sum_{T\in\mathcal{T}} \int_{\partial T} [[\nabla_{\Gamma_h} u_h]](\psi^\ell - \psi_h)\,\mathrm{d}s,
\end{aligned} \tag{3.3.4}
$$

where $\Delta_{\Gamma_h} u_h$ is a piecewise polynomial and $\vec{n}$ is the conormal vector to the triangle T (that is, $\vec{n}\cdot\vec{\nu}_h = 0$). In the current situation $\Delta_{\Gamma_h} u_h$ is identically 0, but we include it to make clear how the corresponding term would appear in other situations. Also, let e be an edge shared by elements T_1 and T_2 which have normals $\vec{n}_1$ and $\vec{n}_2$, respectively. Then $[[\nabla_{\Gamma_h} u_h]] = \nabla_{\Gamma_h} u_h|_{T_1}\cdot\vec{n}_1 - \nabla_{\Gamma_h} u_h|_{T_2}\cdot\vec{n}_2$ is the jump in the normal derivative across e. If $e\subset\partial\Gamma_h$ we set $[[\nabla_{\Gamma_h} u_h]]|_e = 0$. Note that $\vec{n}_1$ lies in the plane of T_1 and $\vec{n}_2$ lies in the plane of T_2, so in contrast to the situation which arises on domains in $\mathbb{R}^n$, we generally have $\vec{n}_1 \neq -\vec{n}_2$. Finally, we insert (3.3.4) into (3.3.3) to find

$$
\begin{aligned}
\int_\Gamma \nabla_\Gamma(u - u_h^\ell)\nabla_\Gamma\psi\,\mathrm{d}\sigma &= \int_{\Gamma_h} (f^\ell\mu_h + \Delta_{\Gamma_h} u_h)(\psi^\ell - \psi_h)\,\mathrm{d}\sigma_h \\
&\quad -\frac{1}{2}\sum_{T\in\mathcal{T}} \int_{\partial T} [[\nabla_{\Gamma_h} u_h]](\psi^\ell - \psi_h)\,\mathrm{d}s - \int_\Gamma [\mathbf{P} - \mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell \nabla_\Gamma\psi\,\mathrm{d}\sigma \\
&\quad + \int_{\Gamma_h} (f^\ell\mu_h - f_h)\psi_h\,\mathrm{d}\sigma_h \\
&\quad \equiv \mathrm{I} + \mathrm{II} + \mathrm{III} + \mathrm{IV}.
\end{aligned} \tag{3.3.5}
$$

3.2. A posteriori upper bound (reliability). We begin by bounding term I of (3.3.5). Let $\psi_h = I_h\psi^\ell$, and let $s_z = m_z \max_{T\subset\omega_z} h_T/\sqrt{|T|}$. Also let $R = f^\ell\mu_h + \Delta_{\Gamma_h} u_h$, and let $\{R_z\}_{z\in\mathcal{N}}$ be constants. Recalling that $\{\varphi_z\}_{z\in\mathcal{N}}$ is a partition of unity, recalling (2.2.25) and (2.2.26), and applying Lemma 2.2, we then have

$$
\begin{aligned}
\mathrm{I} &= \sum_{z\in\mathcal{N}} \int_{\omega_z} R(\psi^\ell - \psi_z^\ell)\varphi_z\,\mathrm{d}\sigma_h = \sum_{z\in\mathcal{N}} \int_{\omega_z} (R - R_z)(\psi^\ell - \psi_z^\ell)\varphi_z\,\mathrm{d}\sigma_h \\
&\le C\sum_{z\in\mathcal{N}} \max_{T\subset\omega_z} \sqrt{|T|} s_z \|\mathbf{A}_h\|_{\ell_2, L_\infty(\omega_z)}^{\frac{1}{2}} \cdot \|\varphi_z(R - R_z)\|_{L_2(\omega_z)} h_z \|\nabla_\Gamma\psi\|_{L_2(\tilde{\omega}_z)}.
\end{aligned} \tag{3.3.6}
$$

Next we turn to bounding the term II. Applying Lemma 2.2, we find

$$
\begin{aligned}
\mathrm{II} &= -\frac{1}{2}\sum_{z\in\mathcal{N}} \sum_{\overline{e}\ni z} \int_e \varphi_z [[\nabla_{\Gamma_h} u_h]](\psi^\ell - \psi_z^\ell)\,\mathrm{d}s \\
&\le C\sum_{z\in\mathcal{N}} \sum_{\overline{e}\ni z} \sqrt{|e|} s_z \|\mathbf{A}_h\|_{\ell_2, L_\infty(\omega_z)}^{\frac{1}{2}} \|\varphi_z [[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)} \|\nabla_\Gamma\psi\|_{L_2(\tilde{\omega}_z)}.
\end{aligned} \tag{3.3.7}
$$

Let

$$
\eta_z = s_z \left(\max_{T\subset\omega_z} \sqrt{|T|} \|\varphi_z(R - R_z)\|_{L_2(\omega_z)} + \sum_{\overline{e}\ni z} \sqrt{|e|} \|\varphi_z [[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)} \right). \tag{3.3.8}
$$

Combining (3.3.6) and (3.3.7) and noting that each element T has only three nodes, we thus find that

$$\begin{aligned}\mathrm{I}+\mathrm{II} &\leq C\sum_{z\in\mathcal{N}}\|\mathbf{A}_h\|^{\frac{1}{2}}_{\ell_2,L_\infty(\omega_z)}\eta_z\|\nabla_\Gamma\psi\|_{L_2(\tilde{\omega}_z)}\\ &\leq C\left(\sum_{z\in\mathcal{N}}\|\mathbf{A}_h\|_{\ell_2,L_\infty(\omega_z)}\eta_z^2\right)^{\frac{1}{2}}\left(\sum_{z\in\mathcal{N}}\|\nabla_\Gamma\psi\|^2_{L_2(\tilde{\omega}_z)}\right)^{\frac{1}{2}}\\ &\leq C\left(\sum_{z\in\mathcal{N}}\|\mathbf{A}_h\|_{\ell_2,L_\infty(\omega_z)}\eta_z^2\right)^{\frac{1}{2}}\|\nabla_\Gamma\psi\|_{L_2(\Gamma)},\end{aligned}\tag{3.3.9}$$

where C does not depend on $\mathcal{T}_h$ or any other essential quantities.

In order to bound the term III, we use (2.2.19) to compute

$$\begin{aligned}\mathrm{III} &= -\sum_{z\in\mathcal{N}}\int_{\tilde{\omega}_z}\varphi_z^\ell[\mathbf{P}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\nabla_\Gamma\psi\,\mathrm{d}\sigma\\ &\leq \sum_{z\in\mathcal{N}}\left\|\sqrt{\varphi_z^\ell}[\mathbf{P}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\right\|_{L_2(\tilde{\omega}_z)}\left\|\sqrt{\varphi_z^\ell}\nabla_\Gamma\psi\right\|_{L_2(\tilde{\omega}_z)}\\ &= \sum_{z\in\mathcal{N}}\left\|\sqrt{\mu_h}\sqrt{\varphi_z}[\mathbf{P}-\mathbf{A}_h][\mathbf{I}-d\mathbf{H}]^{-1}\left[\mathbf{I}-\frac{\vec{\nu}_h\otimes\vec{\nu}}{\vec{\nu}_h\cdot\vec{\nu}}\right]\nabla_{\Gamma_h}u_h\right\|_{L_2(\omega_z)}\\ &\quad\cdot\left\|\sqrt{\varphi_z^\ell}\nabla_\Gamma\psi\right\|_{L_2(\tilde{\omega}_z)}.\end{aligned}\tag{3.3.10}$$

Defining

$$\mathbf{B}_h=\sqrt{\mu_h}[\mathbf{P}-\mathbf{A}_h][\mathbf{I}-d\mathbf{H}]^{-1}[\mathbf{I}-\frac{\vec{\nu}_h\otimes\vec{\nu}}{\vec{\nu}_h\cdot\vec{\nu}}]\tag{3.3.11}$$

and recalling that $\sum_{z\in\mathcal{N}}\varphi_z=\sum_{z\in\mathcal{N}}\varphi_z^\ell\equiv 1$, we finally compute

$$\begin{aligned}\mathrm{III} &\leq \left(\sum_{z\in\mathcal{N}}\|\sqrt{\varphi_z}\mathbf{B}_h\nabla_{\Gamma_h}u_h\|^2_{L_2(\omega_z)}\right)^{1/2}\left(\sum_{z\in\mathcal{N}}\|\sqrt{\varphi_z^\ell}\nabla_\Gamma\psi\|^2_{L_2(\tilde{\omega}_z)}\right)^{1/2}\\ &= \|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|_{L_2(\Gamma_h)}\|\nabla_\Gamma\psi\|_{L_2(\Gamma)}.\end{aligned}\tag{3.3.12}$$

Finally we bound the term IV. First we note that, for $z\in\mathcal{N}$ and with ψ_z^ℓ defined as in (2.2.24),

$$\begin{aligned}\|\sqrt{\varphi_z}\psi_z^\ell\|_{L_2(\Omega_z)} &= \sqrt{\int_{\omega_z}\varphi_z\,\mathrm{d}\sigma_h}\frac{1}{\int_{\omega_z}\varphi_z\,\mathrm{d}\sigma_h}\left|\int_{\omega_z}\varphi_z\psi_z^\ell\,\mathrm{d}\sigma_h\right|\\ &\leq \|\sqrt{\varphi_z}\psi^\ell\|_{L_2(\omega_z)}.\end{aligned}\tag{3.3.13}$$

Since $\psi\in H_0^1(\Gamma)$ has either a vanishing trace on $\partial\Omega$ or a vanishing mean value over

Ω, we may apply (3.3.13) and a Poincaré inequality to compute

$$\begin{aligned}
\mathrm{IV} &= \int_{\Gamma_h} (f^\ell \mu_h - f_h)\psi_h \,\mathrm{d}\sigma_h \\
&= \sum_{z\in\mathcal{N}} \int_{\omega_z} (f^\ell \mu_h - f_h)\varphi_z \psi_z^\ell \,\mathrm{d}\sigma_h \\
&\le \sum_{z\in\mathcal{N}} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|_{L_2(\omega_z)} \|\sqrt{\varphi_z}\psi_z^\ell\|_{L_2(\omega_z)} \\
&\le \sum_{z\in\mathcal{N}} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|_{L_2(\omega_z)} \|\sqrt{\varphi_z}\psi^\ell\|_{L_2(\omega_z)} \\
&\le \sum_{z\in\mathcal{N}} \left\| \frac{1}{\sqrt{\mu_h}} \right\|_{L_\infty(\omega_z)} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|_{L_2(\Gamma_h)} \left\| \sqrt{\varphi_z^\ell}\psi \right\|_{L_2(\bar{\omega_z})} \\
&\le \left(\sum_{z\in\mathcal{N}} \left\| \frac{1}{\mu_h} \right\|_{L_\infty(\omega_z)} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|^2_{L_2(\Gamma_h)} \right)^{1/2} \|\psi\|_{L_2(\Gamma)} \\
&\le C_P(\Gamma) \left(\sum_{z\in\mathcal{N}} \left\| \frac{1}{\mu_h} \right\|_{L_\infty(\omega_z)} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|^2_{L_2(\Gamma_h)} \right)^{1/2} \|\nabla_\Gamma \psi\|_{L_2(\Gamma)}.
\end{aligned} \tag{3.3.14}$$

Making the substitution $\psi = u - u_h^\ell$ if $\partial\Gamma \neq \emptyset$ or $\psi = u - u_h^\ell - \frac{1}{|\Gamma|}\int_\Gamma (u - u_h^\ell)\,\mathrm{d}\sigma$ if $\partial\Gamma = \emptyset$ while combining (3.3.3), (3.3.12), and (3.3.14) yields

$$\begin{aligned}
\|\nabla_\Gamma(u - u_h^\ell)\|^2_{L_2(\Gamma)} \le \Bigg[& C \left(\sum_{z\in\mathcal{N}} \|\mathbf{A}_h\|_{\ell_2, L_\infty(\omega_z)} \eta_z^2 \right)^{1/2} + \|\mathbf{B}_h \nabla_{\Gamma_h} u_h\|_{L_2(\Gamma_h)} \\
& + C_P(\Gamma) \left(\sum_{z\in\mathcal{N}} \left\| \frac{1}{\mu_h} \right\|_{L_\infty(\omega_z)} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|^2_{L_2(\Gamma_h)} \right)^{1/2} \Bigg] \\
& \cdot \|\nabla_\Gamma(u - u_h^\ell)\|_{L_2(\Gamma)}.
\end{aligned} \tag{3.3.15}$$

Dividing (3.3.15) through by $\|\nabla_\Gamma(u - u_h^\ell)\|_{L_2(\Gamma)}$ then yields the following theorem.

THEOREM 3.1. *Under the assumptions in section* 2,

$$\|\nabla_\Gamma(u - u_h^\ell)\|_{L_2(\Gamma)} \le \mathcal{R} + \mathcal{G} + \mathcal{D}, \tag{3.3.16}$$

where

$$\mathcal{R} = C \left(\sum_{z\in\mathcal{N}} \|\mathbf{A}_h\|_{\ell_2, L_\infty(\omega_z)} \eta_z^2 \right)^{1/2}, \tag{3.3.17}$$

$$\mathcal{G} = \|\mathbf{B}_h \nabla_{\Gamma_h} u_h\|_{L_2(\Gamma_h)}, \tag{3.3.18}$$

and

$$\mathcal{D} = C_P(\Gamma) \left(\sum_{z\in\mathcal{N}} \left\| \frac{1}{\mu_h} \right\|_{L_\infty(\omega_z)} \|\sqrt{\varphi_z}(f^\ell \mu_h - f_h)\|^2_{L_2(\Gamma_h)} \right)^{1/2}. \tag{3.3.19}$$

Here $\eta_z = s_z(\max_{T\subset\omega_z} \sqrt{|T|}\|\varphi_z(R - R_z)\|_{L_2(\omega_z)} + \sum_{\bar{e}\ni z} \sqrt{|e|}\|\varphi_z[[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)})$ *as in* (3.3.8), *the constants* R_z *in* η_z *may be freely chosen, and* C *does not depend on* $\mathcal{T}_h$ *or* Γ.

We make a few brief remarks concerning Theorem 3.1, beginning with the *residual term* $\mathcal{R}$. First we note that if the nodes of Γ_h lie on Γ, then $\|[(\mathbf{P}-\mathbf{A}_h)(x)]\|_{\ell_2\to\ell_2} \le Ch_T^2$ for $x \in T$ (cf. [Dz88]). Thus up to a higher-order term, $\mathcal{R}$ is bounded by $C(\sum_{z\in\mathcal{N}} \eta_z^2)^{1/2}$. Next we consider the *geometric error term* $\mathcal{G}$. Note first that unlike the residual term $\mathcal{R}$, it contains no unknown constants. Secondly, $\mathcal{G}$ is heuristically of higher order since $\|\mathbf{B}_h\|_{\ell_2\to\ell_2} \le C\|\mathbf{P}-\mathbf{A}_h\|_{\ell_2\to\ell_2} \le Ch_T^2$. The *data approximation term* $\mathcal{D}$ is 0 if we let $f_h = \mu_h f^\ell$ and assume exact quadrature, both of which we shall do in our numerical tests. In [Dz88] the definition $f_h(x) = f^\ell(x) - \frac{1}{|\Gamma_h|}\int_{\Gamma_h} f^\ell \,\mathrm{d}\sigma_h$ is made. This choice has the advantage of not requiring the computation of the ratio μ_h of the continuous to the discrete measure and still leads to optimal-order H^1 and L_2 estimates. However, computation of $\mathcal{R}$ and $\mathcal{G}$ requires access to μ_h in any case, so we shall use the definition $f_h = \mu_h f^\ell$ and thereby exclude $\mathcal{D}$. A final note concerning $\mathcal{D}$ is that it includes the global Poincaré constant $C_P(\Gamma)$. In contrast to the terms $\mathcal{R}$ and $\mathcal{G}$, $\mathcal{D}$ is thus not entirely built up of quantities which are locally determined.

Finally, we note that the dominant term in (3.3.16) does not depend explicitly on geometric information about Γ. Since $\|\mathbf{A}_h-\mathbf{P}\|_{\ell_2,L_\infty(\omega_z)} \le C(\tilde\omega_z)h_z^2$, we may compute $\mathcal{G} \le (\sum_{z\in\mathcal{N}} C(\tilde\omega_z^2)h_z^4\|\sqrt{\varphi_z}\nabla_{\Gamma_h}u_h\|^2_{L_2(\omega_z)})^{1/2}$. Also, $\mathcal{R} \le C(\sum_{z\in\mathcal{N}}\eta_z^2 + C(\Gamma)h_z^2\eta_z^2)^{1/2}$. Finally, as shown in [Dz88], $\mathcal{D}$ is of higher order even if the choice $f_h(x) = f^\ell(x) - \frac{1}{|\Gamma_h|}\int_{\Gamma_h} f^\ell \,\mathrm{d}\sigma_h$ is made. Thus the dominant part of the a posteriori upper bound is $C(\sum_{z\in\mathcal{N}}\eta_T^2)^{1/2}$, exactly as for problems in planar domains.

The estimator given in Theorem 3.1 could in principle be implemented, but it is possible to define a more convenient estimator for practical use. Recalling the comments of section 2.2, we may simplify it by assuming shape-regularity. In addition, residual estimators are typically calculated elementwise instead of patchwise, so we define an alternate estimator which allows mostly elementwise calculations (the only exception is the term involving $\|A_h\|$, which must be patch-based). In our computations we shall apply the estimator naturally derived from the following corollary.

COROLLARY 3.2. *Assume that $f_h = \mu_h f^\ell$, that $\mathcal{T}_h$ is shape-regular, and that m_z is bounded. Then*

(3.3.20)
$$\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\Gamma)} \le \sqrt{2}\left(\sum_{T\in\mathcal{T}_h} C\|\mathbf{A}_h\|_{\ell_2,L_\infty(\omega_T)}\eta_T^2 + \|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|^2_{L_2(T)}\right)^{1/2} \equiv \Theta.$$

Here $\omega_T = \cup_{z\in\overline{T}}\omega_z$, $\eta_T = h_T\|R\|_{L_2(T)} + h_T^{1/2}\|[[\nabla_{\Gamma_h}u_h]]\|_{L_2(\partial T)}$, and C depends on $\max_{z\in\mathcal{N}} m_z$ and the minimum angle over all elements of $\mathcal{T}_h$.

The proof of Corollary 3.2 follows by setting $R_z = 0$ in (3.3.8) and noting that, under the assumptions that $\mathcal{T}_h$ is shape-regular and m_z is bounded, h_T is equivalent to h_z for all vertices z of T and to $|e|$ for all $e \subset \partial T$. ∎

3.3. A posteriori lower bound (efficiency). In this section we prove a local a posteriori lower bound which is a counterpart to the upper bound in Corollary 3.2. Such lower bounds verify (up to higher-order terms) that the stated a posteriori estimate does not overestimate the actual error and also are an essential ingredient in proving the convergence of adaptive methods; cf. [MNS02].

PROPOSITION 3.3. *Assume that $f_h = \mu_h f^\ell$, that $\mathcal{T}_h$ is shape-regular, and that m_z is bounded. Then for $T \in \mathcal{T}_h$,*

$$\begin{aligned}\eta_T \le{}& C\|\mathbf{A}_h\|^{1/2}_{\ell_2,L_\infty(\omega_T)}(\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\tilde{\omega}_T)} + \|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|_{L_2(\omega_T)})\\ &+Ch_T\|R-R_T\|_{L_2(\omega_T)}.\end{aligned} \tag{3.3.21}$$

Here C depends on $\max_{z\in\mathcal{N}:z\in\overline{T}} m_z$ and the minimum angle of the elements in ω_T, and R_T is an arbitrary piecewise linear function.

Proof. We shall follow the well-known proof of Verfürth (cf. [Ver89]). First let $z \in \mathcal{N}$ and $T \subset \omega_z$. Letting z_i, $1 \le i \le 3$, be the nodes of T, we define the bubble function $\phi_T = \prod_{i=1}^3 \varphi_{z_i}$. In addition, let R_T be an arbitrary piecewise linear approximation to R on T. Let also $\tilde{T}$ denote the natural lift of T to Γ. Then using (3.3.5) with $\psi = R_T^\ell\phi_T^\ell$ and $\psi_h = 0$ and noting that $\phi_T = 0$ on $\partial\tilde{T}$, we have

$$\begin{aligned}\int_T RR_T\phi_T\,\mathrm{d}\sigma_h &= \int_{\tilde{T}} \nabla_\Gamma(u-u_h^\ell)\nabla_\Gamma(R_T^\ell\phi_T^\ell)\,\mathrm{d}\sigma\\ &\quad+\int_{\tilde{T}}[\mathbf{I}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\nabla_\Gamma(R_T^\ell\phi_T^\ell)\,\mathrm{d}\sigma\\ &\le (\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\tilde{T})} + \|[\mathbf{I}-\mathbf{A}_h^\ell]\nabla_\Gamma u_h^\ell\|_{L_2(\tilde{T})})\|\nabla_\Gamma(R_T^\ell\phi_T^\ell)\|_{L_2(\tilde{T})}\\ &\le (\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\tilde{T})} + \|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|_{L_2(T)})\\ &\quad\cdot\|\mathbf{A}_h\|^{1/2}_{L_\infty(T)}\|\nabla_{\Gamma_h}(R_T\phi_T)\|_{L_2(T)}.\end{aligned} \tag{3.3.22}$$

Since $R_T\phi_T$ is a polynomial, we may apply an inverse inequality to find

$$\|\nabla_{\Gamma_h}(R_T\phi_T)\|_{L_2(T)} \le Ch_T^{-1}\|R_T\phi_T\|_{L_2(T)} \le Ch_T^{-1}\|R_T\|_{L_2(T)}, \tag{3.3.23}$$

where C depends only on the shape-regularity of T. Thus

$$\begin{aligned}&\int_T R\,R_T\phi_T\,\mathrm{d}\sigma_h\\ &\quad\le Ch_T^{-1}\|\mathbf{A}_h\|^{1/2}_{L_\infty(T)}(\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\tilde{T})} + \|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|_{L_2(T)})\|R_T\|_{L_2(T)}.\end{aligned} \tag{3.3.24}$$

Applying Theorem 2.2 of [AO00], we next note that

$$\begin{aligned}\|R_T\|^2_{L_2(T)} &\le \|\sqrt{\phi_T}R_T\|^2_{L_2(T)}\\ &\le \left(\|\sqrt{\phi_T}(R-R_T)\|_{L_2(T)} + \left(\int_T RR_T\phi_T\,\mathrm{d}\sigma_h\right)^{1/2}\right)\|R_T\|_{L_2(T)}.\end{aligned} \tag{3.3.25}$$

Combining the previous inequalities, we thus find

$$\begin{aligned}\|R_T\|^2_{L_2(T)} \le{}& C[\|R-R_T\|_{L_2(T)} + h_T^{-1}\|\mathbf{A}_h\|^{1/2}_{L_\infty(T)}(\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\tilde{T})}\\ &+\|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|_{L_2(T)})]\|R_T\|_{L_2(T)}.\end{aligned} \tag{3.3.26}$$

Thus

$$\begin{aligned}h_T\|R\|_{L_2(T)} \le{}& C[\|\mathbf{A}_h\|^{1/2}_{L_\infty(T)}(\|\nabla_\Gamma(u-u_h^\ell)\|_{L_2(\tilde{T})} + \|\mathbf{B}_h\nabla_{\Gamma_h}u_h\|_{L_2(T)})\\ &+h_T\|R-R_T\|_{L_2(T)}].\end{aligned} \tag{3.3.27}$$

Next we bound the edge residual $\|[[\nabla_{\Gamma_h} u_h]]\|_{L_2(\partial T)}$. Let e be an edge which is shared by elements $T_1 = T$ and T_2 and whose closure contains the nodes z_1 and z_2. Let $\lambda_{i,j}$, $i,j = 1,2$, be the barycentric coordinate on triangle i corresponding to the vertex z_j, and define $\phi_e|_{T_i} = \lambda_{i,1}\lambda_{i,2}$. Thus $\phi_e \in H_0^1(T_1 \cup T_2)$, and $\phi_e > 0$ on e. Then

$$\|[[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)} \le C\|\sqrt{\phi_e}[[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)}. \tag{3.3.28}$$

Noting that $[[\nabla_{\Gamma_h} u_h]]_e$ is a constant, we employ (3.3.5) with $\psi = (|[[\nabla_{\Gamma_h} u_h]]|_e \phi_e^\ell)$ to find

(3.3.29)

$$\begin{aligned}
\int_e |[[\nabla_{\Gamma_h} u_h]]|^2 \phi_e \,\mathrm{d}s &= \int_{\tilde{T}_1 \cup \tilde{T}_2} \nabla_\Gamma (u - u_h^\ell) \nabla_\Gamma ([[\nabla_{\Gamma_h} u_h]]_e \phi_e^\ell) \,\mathrm{d}\sigma \\
&\quad - \int_{T_1 \cup T_2} R |[[\nabla_{\Gamma_h} u_h]]|_e \phi_e \,\mathrm{d}\sigma_h + \int_{\tilde{T}_1 \cup \tilde{T}_2} [\mathbf{I} - \mathbf{A}_h^\ell] \nabla_\Gamma u_h^\ell \nabla_\Gamma ([[\nabla_{\Gamma_h} u_h]]_e \phi_e) \,\mathrm{d}\sigma \\
&\le |[[\nabla_{\Gamma_h} u_h]]|_e (\|\nabla_\Gamma (u - u_h^\ell)\|_{L_2(\tilde{T}_1 \cup \tilde{T}_2)} \|\mathbf{A}_h\|_{L_\infty(T_1 \cup T_2)}^{1/2} \|\nabla_{\Gamma_h} \phi_e\|_{L_2(T_1 \cup T_2)} \\
&\quad + \|R\|_{L_2(T_1 \cup T_2)} \|\phi_e\|_{L_2(T_1 \cup T_2)} \\
&\quad + \|\mathbf{B}_h \nabla_{\Gamma_h} u_h\|_{L_2(T_1 \cup T_2)} \|\mathbf{A}_h\|_{L_\infty(T_1 \cup T_2)}^{1/2} \|\nabla_{\Gamma_h} \phi_e\|_{L_2(T_1 \cup T_2)}).
\end{aligned}$$

A simple scaling argument yields $\|\phi_e\|_{L_2(T_1 \cup T_2)} \le C h_T$ and $\|\nabla_{\Gamma_h} \phi_e\|_{L_2(T_1 \cup T_2)} \le C$, so that

$$\begin{aligned}
\int_e |[[\nabla_{\Gamma_h} u_h]]|^2 \,\phi_e \,\mathrm{d}s &\le C h_T^{-1/2} \|[[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)} \\
&\quad \cdot [\|\mathbf{A}_h\|_{L_\infty(T_1 \cup T_2)}^{1/2} \|\nabla_\Gamma (u - u_h^\ell)\|_{L_2(\tilde{T}_1 \cup \tilde{T}_2)} \\
&\quad + h_T \|R\|_{L_2(T_1 \cup T_2)} + \|\mathbf{A}_h\|_{L_\infty(T_1 \cup T_2)}^{1/2} \|\mathbf{B}_h \nabla_{\Gamma_h} u_h\|_{L_2(T_1 \cup T_2)}].
\end{aligned} \tag{3.3.30}$$

Combining the previous three inequalities, we find that

$$\begin{aligned}
h_T^{1/2} \|[[\nabla_{\Gamma_h} u_h]]\|_{L_2(e)} &\le C(\|\mathbf{A}_h\|_{L_\infty(T_1 \cup T_2)}^{1/2} \|\nabla_\Gamma (u - u_h^\ell)\|_{L_2(\tilde{T}_1 \cup \tilde{T}_2)} \\
&\quad + h_T \|R\|_{L_2(T_1 \cup T_2)} + \|\mathbf{A}_h\|_{L_\infty(T_1 \cup T_2)}^{1/2} \|\mathbf{B}_h \nabla_{\Gamma_h} u_h\|_{L_2(T_1 \cup T_2)}).
\end{aligned} \tag{3.3.31}$$

Summing (3.3.31) over the three edges of T and combining (3.3.31) with (3.3.27) completes the proof of (3.3.21). ∎

4. Implementation details. In this section we provide some details concerning implementation.

4.1. Computation of geometric quantities. We assume that $\Gamma = \{x \in \mathbb{R}^3 : \zeta(x) = 0\}$, where ζ is sufficiently smooth with a nonzero gradient in a large enough neighborhood of Γ. In addition, we assume that ζ, its gradient, and its Hessian matrix are available and that for $x \in U$ we can approximate $a(x)$ with sufficient accuracy. In the next subsection we describe a simple approach for approximating $a(x)$.

First we note that if $x \in \Gamma$, $\vec{\nu}(x) = \frac{\nabla \zeta}{|\nabla \zeta|}$. Thus for $x \in U$,

$$\vec{\nu}(x) = \frac{\nabla \zeta(a(x))}{|\nabla \zeta(a(x))|}. \tag{4.4.1}$$

In addition, we have for $x \in \Gamma$

$$\mathbf{H}(x) = \nabla_\Gamma \vec{\nu}(x) = \mathbf{P} \nabla \frac{\nabla \zeta(x)}{|\nabla \zeta(x)|}. \tag{4.4.2}$$

For the sake of concreteness, we note that $\nabla \frac{\nabla\zeta(x)}{|\nabla\zeta(x)|}$ is not necessarily symmetric and that $[\nabla \frac{\nabla\zeta}{|\nabla\zeta|}]_{ij} = \frac{\partial}{\partial x_i}\frac{\zeta_{x_j}}{|\nabla\zeta|}$. The eigenvalues κ_1 and κ_2 of H in the directions orthogonal to $\vec{\nu}$ may then be approximated numerically. We also recall the relationship $\kappa_i(x) = \frac{\kappa_i(a(x))}{1+d(x)\kappa_i(a(x))}$ from (2.2.5). Finally, we emphasize that d is the *signed* distance function, that is, $d(x) = sign(\zeta(x))|a(x) - x|$ for $x \in U \setminus \Gamma$.

The above information is sufficient to implement the adaptive method described above. In particular, for $x \in \Gamma_h$ we use (2.2.23) to define the discrete data

$$f_h(x) = \mu_h(x) f(a(x)) = (1 - d(x)\kappa_1(x))(1 - d(x)\kappa_2(x))\vec{\nu}(x) \cdot \vec{\nu}_h(x) f(a(x)). \tag{4.4.3}$$

Here $\vec{\nu}(x)$ is computed via (4.4.1), κ_1 and κ_2 are computed via (2.2.5), and $\vec{\nu}_h$ must be computed from mesh information. Next we note that

$$\|\mathbf{A}_h(x)\|_{\ell_2\to\ell_2} \le \frac{\max(1 - d(x)\kappa_1(x), 1 - d(x)\kappa_2(x))}{|\vec{\nu}(x) \cdot \vec{\nu}_h(x)| \min(1 - d(x)\kappa_1(x), 1 - d(x)\kappa_2(x))} \equiv \overline{A}_h(x) \tag{4.4.4}$$

and

$$\begin{aligned} \|\mathbf{B}_h(x)\|_{\ell_2\to\ell_2} &\le \tfrac{1}{\mu_h(x)}[|d(x)(\kappa_1(x) - \kappa_2(x))| \\ &\quad + |1 - \vec{\nu}(x) \cdot \vec{\nu}_h(x)|(1 + 4\max(1 - d(x)\kappa_1(x), 1 - d(x)\kappa_2(x)))] \\ &\equiv \overline{B}_h(x). \end{aligned} \tag{4.4.5}$$

The expressions on the right-hand sides of (4.4.4) and (4.4.5) may be computed using (4.4.1) and (2.2.5) as before. Since $\mathbf{P} - \mathbf{A}_h$ and $\mathbf{B}_h$ are of higher order, using the above approximations for the norms of $\mathbf{A}_h$ and $\mathbf{B}_h$ should lead to at most a slight overestimation of the overall error while yielding nontrivial computational savings.

4.2. Computation of d and a. The efficient computation of the projection a and distance function d are central to implementing the finite element method and a posteriori estimators described here. In a very few cases, d is available explicitly (for example, $d(x) = |x| - r$ for a sphere of radius r). Even for relatively simple surfaces such as ellipsoids, however, an explicit expression for d is not available and a and d must be approximated. Since d is assumed to be smooth and we need to be concerned only about starting points sufficiently close to Γ, standard methods of nonlinear optimization are, in principle, applicable.

We have tested two different algorithms for computing a: one being Newton's method and the other being an ad hoc first-order method. Before describing the methods we note a relationship which we shall use in our algorithms. For $x \in U$, $\zeta(x) = \int_0^{d(x)} \nabla\zeta(a(x) + t\vec{\nu}(x)) \cdot \vec{\nu}(x)\,\mathrm{d}t = d|\nabla\zeta(x)| + O(d^2)$. Thus

$$d(x) \approx \frac{\zeta(x)}{|\nabla\zeta(x)|}. \tag{4.4.6}$$

Next we describe our implementation of Newton's method. Assume that $x_0 \in U$ and that we wish to compute $a(x_0)$. In order to employ Newton's method, we seek a stationary point of the function $F(x,\lambda) = |x - x_0| + \lambda\zeta(x)$. Note that $\nabla F(x,\lambda) = (2(x - x_0) + \lambda\nabla\zeta(x), \zeta(x))$. Thus $\nabla F(x,\lambda) = 0$ implies that $x \in \Gamma$ and $(x - x_0)$ is parallel to $\nabla\zeta(x)$, that is, $x = a(x_0)$. In order to choose a starting point, we note that $2(x-x_0)+\lambda\nabla\zeta(x) = 0$ implies that $\lambda = 2d(x_0)/|\nabla\zeta(x)|$. Using (4.4.6), we thus choose the starting value $(x_0, \lambda_0) = (x_0, 2\phi(x_0)/|\nabla\phi(x_0)|^2)$ for Newton's method. Given a

tolerance *tol*, we iterate Newton's method until

$$\left(\frac{\zeta(x)^2}{|\nabla\zeta(x)|^2}+\left|\frac{\nabla\zeta(x)}{|\nabla\zeta(x)|}-\frac{x-x_0}{|x-x_0|}\right|^2\right)^{1/2}<tol. \tag{4.4.7}$$

Fulfillment of this stopping criteria guarantees that the returned value $x\approx a(x_0)$ lies in the correct direction from x_0 to within *tol* and that, because of (4.4.6), $d(x)<tol$ up to higher-order terms.

The first-order algorithm which we employed may be described as follows: Since $a(x)=x-d(x)\vec{\nu}(x)$, we may use (4.4.6) and $\vec{\nu}(x)\approx\frac{\nabla\zeta(x)}{|\nabla\zeta(x)|}$ to approximate a by $a(x)\approx x-\frac{\zeta(x)\nabla\zeta(x)}{|\nabla\zeta(x)|^2}$. Iterating this relationship leads to an algorithm which converges to some point on Γ but not generally to $a(x)$. We thus correct the direction $x-x_0$ at each step, yielding the following algorithm.

1. Stipulate *tol* and x_0, and initialize $x=x_0$.
2. While (4.4.7) is not satisfied, iterate the following steps:
   (a) Calculate $\tilde{x}=x-\frac{\zeta(x)\nabla\zeta(x)}{|\nabla\zeta(x)|^2}$ and $dist=sign(\zeta(x_0))|\tilde{x}-x_0|$.
   (b) Set $x=x_0-dist\frac{\nabla\zeta(\tilde{x})}{|\nabla\zeta(\tilde{x})|}$.

In practice, the second of the two algorithms was more efficient than Newton's method. While Newton's method converged in less steps as one would expect, each step is relatively expensive. We also note that we have not rigorously analyzed the error in either of these methods which results from using the stopping criterion (4.4.7). A more rigorous analysis of robust algorithms for approximating a would thus be desirable.

5. Computational examples. In this section we describe several computational examples. All computations were performed using the finite element toolbox ALBERTA [SS05], and graphics were processed using the software GMV [Or05]. Also, the constant C appearing in the estimator Θ in (3.3.20) was taken to 0.25 in all calculations.

5.1. Example 1: Computation on a spherical subdomain. In our first test we consider a problem which was used as an example in the paper [AP05]. This problem demonstrates the ease with which our method handles problems in which the distance function is explicitly available and also provides a convenient place to consider surfaces with boundaries.

Let S^2 be the unit sphere with angular spherical coordinates (ϕ,θ), where ϕ $(0\le\phi<2\pi)$ is the azimuthal angle in the xy-plane and $\theta=\cos^{-1}z$ $(0\le\theta\le\pi)$ is the polar angle from the z-axis. Following [AP05], we let Γ consist of points in S^2 such that $0\le\phi\le\frac{5\pi}{3}$, and let $u(\phi,\theta)=(\sin\theta)^\lambda\sin\lambda\phi$ for $\lambda=.6$. Then u satisfies $-\Delta_\Gamma u=f$ in Γ and $u=0$ on $\partial\Gamma$ with $f=\lambda(\lambda+1)(\sin\theta)^\lambda\sin\lambda\phi$. Note that u is singular at the poles, so we may expect an adaptive algorithm to refine more heavily there.

Computation of the geometric quantities necessary to implement our method is quite straightforward. We employ the distance function $d(x)=|x|-1$ for the sphere (note that we do not actually require access to the distance function for Γ here). In addition, we may easily compute that $\vec{\nu}(x)=\frac{x}{|x|}$, $a(x)=\frac{x}{|x|}$, and $\mathbf{H}_{ij}(x)=\frac{\delta_{ij}}{|x|}-\frac{x_ix_j}{|x|^3}$. The eigenvalues of $\mathbf{H}(x)$ are the principle curvatures of the sphere of radius $|x|$, that is, $\kappa_1=\kappa_2=\frac{1}{|x|}$. Computation of μ_h, $\overline{A}_h$, and $\overline{B}_h$ is similarly straightforward.

In Figure 5.1, the initial mesh of six elements is displayed along with an adaptively refined mesh colored with the solution u_h. A blowup showing refinement near the

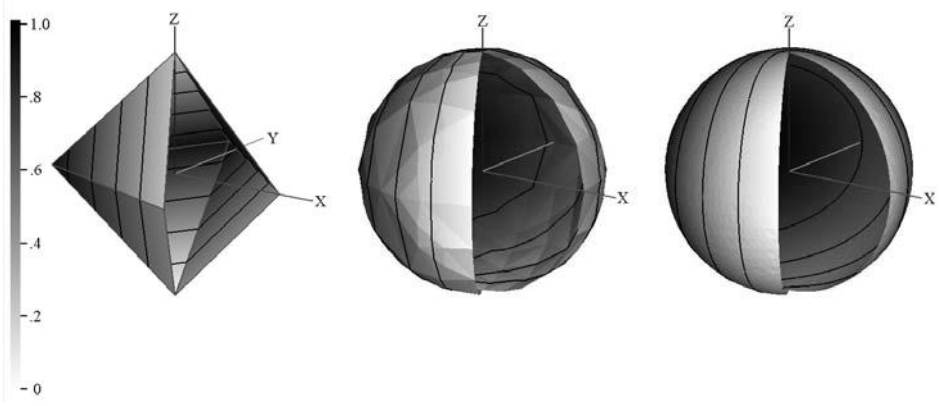


FIG. 5.1. *Experiment* 1: *The initial mesh with* 6 *nodes (left) and adaptively refined meshes with* 151 *(center) and* 5559 *(right) DOF displaying* u_h.

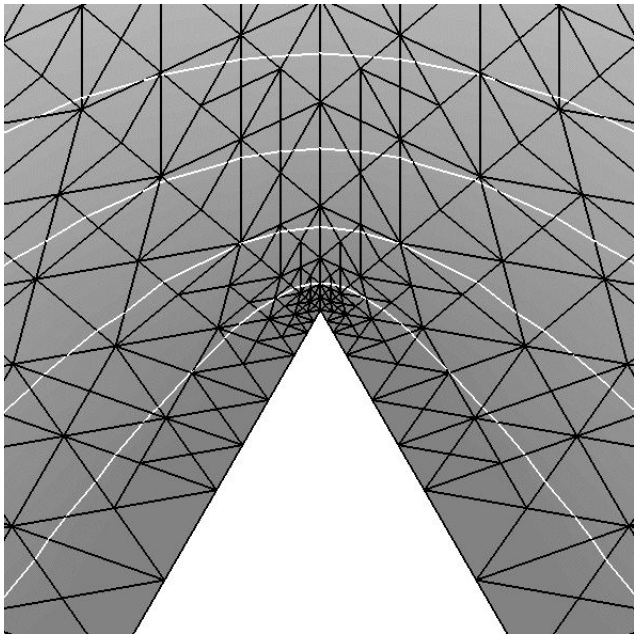


FIG. 5.2. *Experiment* 1: *View of the refined mesh along the* z*-axis, magnified* $80\times$*, with contour lines of* u.

positive z-pole is displayed in Figure 5.2. Finally, a graph displaying the error, the residual estimator Θ defined in (3.3.20), and various geometric quantities is given in Figure 5.3. First note that the quantity $\max_{T\in\mathcal{T}_h} h_T/\sqrt{|T|}$ appears to reach a maximum value of about 3. Thus our assumption in Corollary 3.2 that the mesh is shape-regular is justified for this example. Also, the error $\|\nabla_\Gamma(u - u_h^\ell)\|_{L_2(\Omega)}$ and the residual estimator Θ converge with optimal order and appear to have a constant ratio as the mesh is refined. Finally, the quantities $\|\overline{B}_h|\nabla_{\Gamma_h} u_h|\|_{L_2(\Gamma_h)}$ and $\|1 - \overline{A}_h\|_{L_\infty(\Gamma_h)}$ are plotted and show second-order convergence, confirming experimentally our theoretical observation that these geometric contributions to the error are of

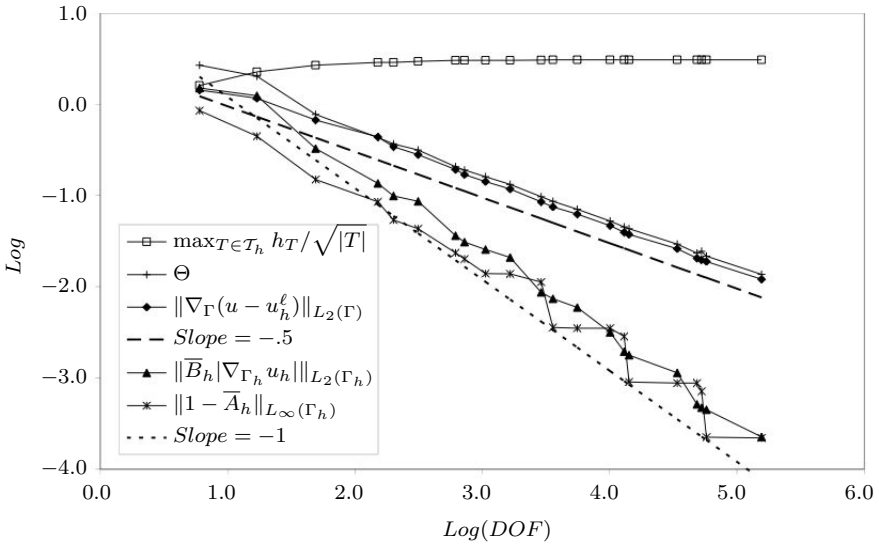


FIG. 5.3. *Experiment* 1: *Error, estimator, and various geometric quantities.*

higher order. It is worth noting here that the quantity $\overline{A}_h$ appears in the estimator as a multiplicative factor. Since it converges to 1, it would thus be reasonable and computationally more efficient to omit it entirely once $\|\overline{A}_h\|_{L_\infty(\Gamma_h)}$ is observed to reach a given tolerance.

5.2. Example 2: Computation on a torus. In our second test we performed a computation on a torus. $d(x) = \sqrt{(r_0 - \sqrt{x^2+y^2})^2 + z^2} - r_1$ is the signed distance function for a torus whose axis of revolution is the z-axis, whose radius of revolution is r_0, and which has thickness $2r_1$. The other necessary geometric quantities may be computed from this formula. We took $r_0 = 1$ and $r_0 = 0.25$. As a test solution we took the function

$$u(x,y,z) = e^{\frac{1}{1.85-x^2}} \sin y, \tag{5.5.1}$$

which has exponential peaks on the outer portions of the torus which lie near the x-axis.

In Figure 5.4 we display $\overline{A}_h$ on the the initial 24-node mesh; note that here $\overline{A}_h$ is about 5 on the outer edge of the torus, so it enters into the calculation in a significant way. Also displayed in Figure 5.4 is a refined mesh having 1248 nodes and displaying the discrete solution u_h. In Figure 5.5 we display the local H_1 error contributions along with the three components $\overline{A}_h$, η_T, and $\overline{B}_h|\nabla_{\Gamma_h}u_h|$ of the estimator Θ. The local residual indicator η_T reflects reasonably well the local error distribution, while the contributions from $\overline{A}_h$ and $\overline{B}_h|\nabla_{\Gamma_h}u_h|$ are relatively insignificant. Also, the maximum ratio $h_T/\sqrt{|T|}$ observed during this calculation was 5.28. This relatively large number reflects the fact that the triangles in the initial mesh displayed in Figure 5.4 already have relatively high aspect ratios.

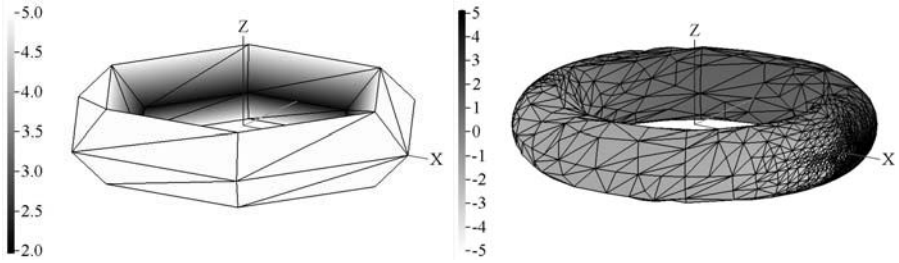


FIG. 5.4. *Experiment* 2: *The initial mesh displaying* $\overline{A}_h$ *(left) and the refined mesh with* 1248 *DOF displaying* u_h *(right).*

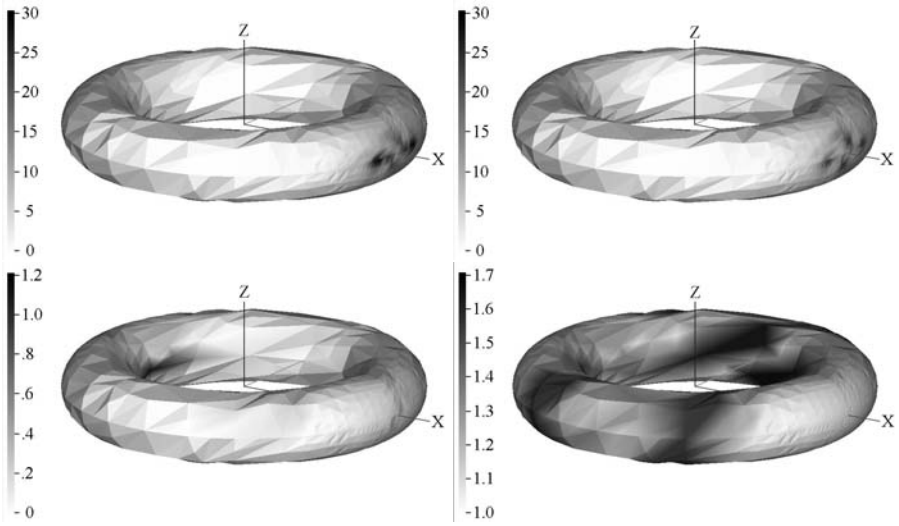


FIG. 5.5. *Experiment* 2: *The local residual* η_T *(top left),* H_1 *error (top right),* $\overline{B}_h|\nabla_{\Gamma_h} u_h|$ *(bottom left), and* $\overline{A}_h$ *(bottom right).*

5.3. Example 3: Computation on an ellipsoid. In our third computational example we let Γ be an ellipsoid satisfying the level set equation

$$x^2 + y^2 + \frac{z^2}{400} = 1. \tag{5.5.2}$$

As a test solution we took $u(x, y, z) = \sin y$ so that u and its derivatives were of moderate size.

In Figure 5.6 we display the local residual contribution η_T and the local geometric error $\overline{B}_h|\nabla_{\Gamma_h} u_h|$ on an adaptively refined mesh having 10383 nodes. Note that the maximum values of η_T and $\overline{B}_h|\nabla_{\Gamma_h} u_h|$ are approximately equal. Thus the geometric error plays a role in the marking of some elements even on a refined mesh. As in Figure 5.3, however, the overall geometric error $\|\overline{B}_h|\nabla_{\Gamma_h}|\|_{L_2(\Gamma_h)}$ declines approximately as DOF^{-1}, while the residual error $(\sum_{T\in\mathcal{T}_h} \eta_T^2)^{1/2}$ declines as $\mathrm{DOF}^{-1/2}$ (we do not display a chart for the current situation as it is entirely analogous to Figure 5.3). Finally, the maximum ratio $h_T/\sqrt{|T|}$ observed in this adaptive calculation (up 64521 DOF) was 3.41, so that mesh quality remained reasonable.

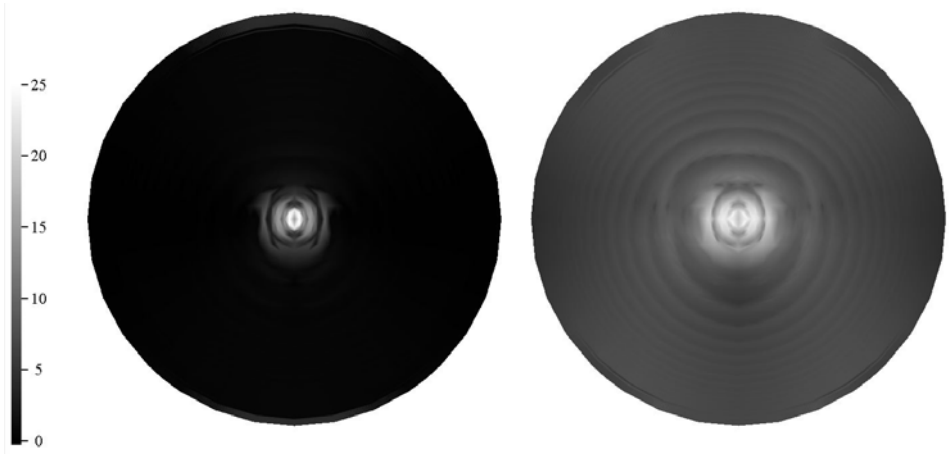


FIG. 5.6. *Experiment* 3: *Mesh with* 10383 *nodes displaying the relative sizes of* $\overline{B}_h|\nabla_{\Gamma_h} u_h|$ *(left) and the local residual* η_T *(right). The view is along the z-axis.*

Appendix. Proof of Proposition 2.1. Proposition 2.1 is potentially of interest in other situations (e.g., when Γ_h is a higher-order polynomial approximation to Γ), so we begin by stating a more general version.

PROPOSITION A.1. *Let* $\hat{T}$ *be the unit simplex in* $\mathbb{R}^2$*, and suppose that* $F : \hat{T} \to U$ *is a* C^1 *mapping whose gradient has two nonzero singular values at each point in* $\hat{T}$*. Suppose* $x \in T := F(\hat{T})$*, let* $\vec{\nu}_h$ *be the normal to* T *at* x*, and let* $\mathrm{d}\sigma_h$ *be a surface measure on* T*. Assume also that* $\vec{\nu} \cdot \vec{\nu}_h > 0$*. Letting* $\mathrm{d}\sigma_h(x)\mu_h(x) = \mathrm{d}\sigma(x)$*, we then have*

$$\mu_h(x) = \vec{\nu} \cdot \vec{\nu}_h (1 - d(x)\kappa_1(x))(1 - d(x)\kappa_2(x)). \tag{A.A.1}$$

Proof. We fix a point $\hat{x} \in \hat{T}$ and let $x = F(\hat{x})$ and $\mathbb{R}^{3\times 2} \ni \mathbf{A} = \nabla F(\hat{x})$. Let $\{\vec{e}_1, \vec{e}_2\}$, $\{\vec{u}_1, \vec{u}_2\}$, and $\{\vec{v}_1, \vec{v}_2\}$ be orthonormal bases for $\mathbb{R}^2$ and the tangent spaces to Γ_h and Γ at x and $a(x)$, respectively. We assume also that $\{\vec{v}_1, \vec{v}_2, \vec{\nu}\}$ are the eigenvectors of $(\mathbf{I} - d\mathbf{H})(x)$ corresponding to the eigenvalues $\lambda_1 = 1 - d(x)\kappa_1(x)$, $\lambda_2 = 1 - d(x)\kappa_2(x)$, and $\lambda_3 = 1$. Letting $\times$ denote the cross product, we have $\mathrm{d}\sigma_h = |\mathbf{A}\vec{e}_1 \times \mathbf{A}\vec{e}_2|\,\mathrm{d}\hat{x}$ and $\mathrm{d}\sigma = |([\mathbf{I} - d\mathbf{H}][\mathbf{P}][\mathbf{A}]\vec{e}_1) \times ([\mathbf{I} - d\mathbf{H}][\mathbf{P}][\mathbf{A}]\vec{e}_2)|\,\mathrm{d}\hat{x}$.

Next we recall the formula $(\mathbf{B}\vec{x}_1) \times (\mathbf{B}\vec{x}_2) = \mathbf{B}_{adj}(\vec{x}_1 \times \vec{x}_2)$, where $\mathbf{B}$ is symmetric and nonsingular and $\mathbf{B}_{adj} = (\det \mathbf{B})\mathbf{B}^{-1}$. Noting that $(\mathbf{I} - d\mathbf{H})_{adj}$ has eigenvectors $\{\vec{v}_1, \vec{v}_2, \vec{\nu}\}$ with eigenvalues $\{\lambda_2\lambda_3, \lambda_1\lambda_3, \lambda_1\lambda_2\}$, we calculate

$$\begin{aligned}([\mathbf{I} - d\mathbf{H}][\mathbf{P}][\mathbf{A}]\vec{e}_1) \times ([\mathbf{I} - d\mathbf{H}][\mathbf{P}][\mathbf{A}]\vec{e}_2) &= \lambda_2\lambda_3[([\mathbf{P}][\mathbf{A}]\vec{e}_1 \times [\mathbf{P}][\mathbf{A}]\vec{e}_2) \cdot \vec{v}_1]\vec{v}_1 \\ + \lambda_1\lambda_3[([\mathbf{P}][\mathbf{A}]\vec{e}_1 \times [\mathbf{P}][\mathbf{A}]\vec{e}_2) \cdot \vec{v}_2]\vec{v}_2 &+ \lambda_1\lambda_2[([\mathbf{P}][\mathbf{A}]\vec{e}_1 \times [\mathbf{P}][\mathbf{A}]\vec{e}_2) \cdot \vec{\nu}]\vec{\nu}.\end{aligned} \tag{A.A.2}$$

But $([\mathbf{P}][\mathbf{A}]\vec{e}_1 \times [\mathbf{P}][\mathbf{A}]\vec{e}_2) \perp \vec{v}_i$, $i = 1, 2$, $([\mathbf{P}][\mathbf{A}]\vec{e}_1 \times [\mathbf{P}][\mathbf{A}]\vec{e}_2) \cdot \vec{\nu} = (\mathbf{A}\vec{e}_1 \times \mathbf{A}\vec{e}_2) \cdot \vec{\nu}$, and $\mathbf{A}\vec{e}_1 \times \mathbf{A}\vec{e}_2 \parallel \vec{\nu}_h$. Thus

$$\begin{aligned}\mathrm{d}\sigma &= |([\mathbf{I} - d\mathbf{H}][\mathbf{P}][\mathbf{A}]\vec{e}_1) \times ([\mathbf{I} - d\mathbf{H}][\mathbf{P}][\mathbf{A}]\vec{e}_2)|\,\mathrm{d}\hat{x} \\ &= \lambda_1\lambda_2[(\mathbf{A}\vec{e}_1 \times \mathbf{A}\vec{e}_2) \cdot \vec{\nu}]\vec{\nu}\,\mathrm{d}\hat{x} = \lambda_1\lambda_2|\mathbf{A}\vec{e}_1 \times \mathbf{A}\vec{e}_2|\vec{\nu}_h \cdot \vec{\nu}\,\mathrm{d}\hat{x}.\end{aligned} \tag{A.A.3}$$

Recalling that $\mathrm{d}\sigma_h = |\mathbf{A}\vec{e}_1 \times \mathbf{A}\vec{e}_2|\,\mathrm{d}\hat{x}$ completes the proof. ∎

REFERENCES

[AO00] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Pure Appl. Math. (NY), Wiley-Intersci. Hoboken, NJ, 2000.

[AP05] T. Apel and C. Pester, *Clement-type interpolation on spherical domains—interpolation error estimates and application to a posteriori error estimation*, IMA J. Numer. Anal., 25 (2005), pp. 310–336.

[BMN05] E. Bänsch, P. Morin, and R. H. Nochetto, *A finite element method for surface diffusion: The parametric case*, J. Comput. Phys., 203 (2005), pp. 321–343.

[BCD04] S. Bartels, C. Carstensen, and G. Dolzmann, *Inhomogeneous Dirichlet conditions in a priori and a posteriori finite element error analysis*, Numer. Math., 99 (2004), pp. 1–24.

[CDDRR04] U. Clarenz, U. Diewald, G. Dziuk, M. Rumpf, and R. Rusu, *A finite element method for surface restoration with smooth boundary conditions*, Comput. Aided Geom. Design, 21 (2004), pp. 427–445.

[DR98] W. Dörfler and M. Rumpf, *An adaptive strategy for elliptic problems including a posteriori controlled boundary approximation*, Math. Comp., 67 (1998), pp. 1361–1382.

[DW00] W. Dörfler and O. Wilderotter, *An adaptive finite element method for a linear elliptic equation with variable coefficients*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 481–491.

[Dz88] G. Dziuk, *Finite elements for the Beltrami operator on arbitrary surfaces*, in Partial Differential Equations and Calculus of Variations, Lecture Notes in Math. 1357, Springer, Berlin, 1988, pp. 142–155.

[Dz91] G. Dziuk, *An algorithm for evolutionary surfaces*, Numer. Math., 58 (1991), pp. 603–611.

[FV06] F. Fierro and A. Veeser, *A posteriori error estimates, gradient recovery by averaging, and superconvergence*, Numer. Math., 103 (2006), pp. 267–298.

[GT98] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, Berlin, 1998.

[Ho01] M. Holst, *Adaptive numerical treatment of elliptic systems on manifolds*, Adv. Comput. Math., 15 (2001), pp. 139–191.

[MN05] K. Mekchay and R. H. Nochetto, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827.

[MNS02] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658. Revised reprint of *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[Or05] F. A. Ortega, *GMV Version* 3.8, Technical rep. LA-UR-95-2986, Los Alamos National Laboratory, Los Alamos, NM, 2005.

[SS05] A. Schmidt and K. G. Siebert, *Design of adaptive finite element software*, in The Finite Element Toolbox ALBERTA, Lec. Notes Comput. Sci. Eng. 42, CD-ROM, Springer, Berlin, 2005.

[Ver89] R. Verfürth, *A posteriori error estimators for the Stokes equations*, Numer. Math., 55 (1989), pp. 309–325.

SIAM J. NUMER. ANAL.
Vol. 45, No. 1, pp. 443–455

CONVERGENCE OF FOURTH ORDER COMPACT DIFFERENCE SCHEMES FOR THREE-DIMENSIONAL CONVECTION-DIFFUSION EQUATIONS*

GIVI BERIKELASHVILI†, MURLI M. GUPTA‡, AND MANANA MIRIANASHVILI§

Abstract. We consider a Dirichlet boundary-value problem for the three-dimensional convection-diffusion equations with constant coefficients in the unit cube. A high order compact finite difference scheme is constructed on a 19-point stencil using the Steklov averaging operators. We prove that the finite difference scheme converges in discrete $W_2^m(\omega)$-norm with the convergence rate $O(h^{s-m})$, where the real parameter s satisfies the condition $\max(1.5, m) < s \le m + 4$, $m = 0, 1, 2$, and the exact solution belongs to the Sobolev space $W_2^s(\Omega)$.

Key words. convection-diffusion equation, convergence estimates, three-dimensions, high accuracy, compact approximations, finite differences

AMS subject classifications. 65N06, 65N15, 76D05

DOI. 10.1137/050622833

1. Introduction.

Let $\Omega = \{x = (x_1, x_2, x_3) : 0 < x_\alpha < 1, \alpha = 1, 2, 3\}$ be the unit cube with boundary Γ. Let D^ν denote the differential operator $D^\nu = \partial^{|\nu|}/(\partial x_1^{\nu_1} \partial x_2^{\nu_2} \partial x_3^{\nu_3})$, where $\nu = (\nu_1, \nu_2, \nu_3)$ are multi-indices with nonnegative integer components, $|\nu| = \nu_1 + \nu_2 + \nu_3$. By $W_2^s(\Omega)$, $s \ge 0$, we denote a Sobolev space with the norm defined by

$$||u||^2_{W_2^s(\Omega)} = \sum_{j=0}^{s} |u|^2_{W_2^j(\Omega)}, \qquad |u|_{W_2^j(\Omega)} = \sum_{|\nu|=j} ||D^\nu u||_{L_2(\Omega)}$$

when s is an integer. If s is a noninteger, let $s = \bar{s} + \varepsilon$, where $\bar{s}$ is the integer part of s and $0 < \varepsilon < 1$. In this case, the norm is defined by

$$||u||^2_{W_2^s(\Omega)} = ||u||^2_{W_2^{\bar{s}}(\Omega)} + |u|^2_{W_2^s(\Omega)},$$

where

$$|u|^2_{W_2^s(\Omega)} = \sum_{|\nu|=\bar{s}} \int_\Omega \int_\Omega \frac{|D^\nu u(x) - D^\nu u(y)|^2}{|x - y|^{3+2\varepsilon}} dx\, dy.$$

In particular, for $s = 0$, we have $W_2^0 = L_2$.

Let $\bar{\omega}$ be the uniform grid in $\bar{\Omega}$ with mesh size h, $\omega = \bar{\omega} \cap \Omega$, $\gamma = \bar{\omega} \backslash \omega$. We define the difference quotients (forward, backward, and central, respectively) in x_α direction as follows:

$$v_{x_\alpha} = \frac{(I^{(+\alpha)} - I)v}{h}, \quad v_{\bar{x}_\alpha} = \frac{(I - I^{(-\alpha)})v}{h}, \quad \partial v = \frac{(I^{(+\alpha)} - I^{(-\alpha)})v}{2h},$$

where $Iv = v$, $I^{(\pm\alpha)}v = v(x \pm hr_\alpha)$, and r_α is the unit vector on the x_α axis.

*Received by the editors January 17, 2005; accepted for publication (in revised form) October 25, 2006; published electronically February 15, 2007.
http://www.siam.org/journals/sinum/45-1/62283.html

†A. Razmadze Mathematical Institute, Georgian Academy of Sciences, 1, M.Aleksidze str.,Tbilisi 0193, Georgia (bergi@rmi.acnet.ge).

‡Department of Mathematics, The George Washington University, Washington, DC 20052 (mmg@gwu.edu).

§N. Muskhelishvili Institute of Computational Mathematics, Georgian Academy of Sciences, 8, Akuri str., Tbilisi 0193, Georgia (pikriag@yahoo.com).

Let $L_2(\omega)$ be the Hilbert space of all discrete functions $y = y(x)$, defined on the grid ω and vanishing on γ, with the inner product and norm defined by

$$(y, v) = \sum_{x\in\omega} h^3 y(x)v(x), \quad ||y|| = (y, y)^{1/2}.$$

Further, we define the following norms:

$$||y||_{W_2^0(\omega)} = ||y||, \qquad ||y||^2_{W_2^1(\omega)} = \sum_{\alpha=1}^{3} ||y_{\bar{x}_\alpha}||^2_{(\alpha)},$$

$$||y||^2_{W_2^2(\omega)} = ||y_{\bar{x}_1 x_1}||^2 + ||y_{\bar{x}_2 x_2}||^2 + ||y_{\bar{x}_3 x_3}||^2 + 2||y_{\bar{x}_1 \bar{x}_2}||^2_{(1,2)} + 2||y_{\bar{x}_1 \bar{x}_3}||^2_{(1,3)} + 2||y_{\bar{x}_2 \bar{x}_3}||^2_{(2,3)}.$$

In the definitions of the norms $||\cdot||_{(\alpha)}$, $||\cdot||_{(\alpha,\beta)}$ the sums run not only over all interior grid points $x \in \omega$, but also over the boundary points $x \in \gamma$ with the coordinates $x_\alpha = 1$ for $||\cdot||_{(\alpha)}$ and over the boundary points $x \in \gamma$ with the coordinates $x_\alpha = 1$, $x_\beta = 1$ for $||\cdot||_{(\alpha,\beta)}$. The inner product $(\cdot,\cdot)_{(\alpha)}$ is defined in a similar manner.

In this paper, we investigate certain high order compact finite difference schemes for the Dirichlet boundary value problem for three-dimensional convection-diffusion equations with constant coefficients:

$$\Delta u + \sum_{\alpha=1}^{3} \lambda_\alpha \frac{\partial u}{\partial x_\alpha} = f(x),\ x \in \Omega, \quad u(x) = 0,\ \ x \in \Gamma,\ \lambda_\alpha = \text{const}. \tag{1}$$

We obtain discretization error estimates of up to the fourth order that are consistent with the smoothness of the solution sought. By definition (see [1]), these error estimates have the form

$$||y - u||_{W_2^m(\omega)} \le ch^{s-m}||u||_{W_2^s(\Omega)}, \qquad s > m \ge 0, \tag{2}$$

where y is the solution of the finite difference scheme and c denotes a positive generic constant, independent of h and u.

Fourth order finite difference schemes for this problem were considered in [2, 3, 4], and it was numerically exhibited, through a variety of test examples, that the discrete solutions converge to the exact solutions of class $C^6(\bar{\Omega})$ in discrete norm $C(\omega)$, and the rate of convergence was exhibited to be $O(h^4)$ (see also [18]). The case of (1) with variable coefficients has been considered by various authors (see, e.g., [16, 17] for three-dimensional and [19] two-dimensional convection-diffusion equations) who described fourth order compact finite difference schemes and exhibited the fourth order convergence through numerical examples. In [18], an attempt was made to carry out theoretical analysis for the convection diffusion equation (1) with constant coefficients—these authors used an eigenvalue analysis to prove that the coefficient matrix arising from the 19-point discretization of (1) is positive definite when the cell Reynolds number exceeds a critical value and that the discrete solution remains oscillation free in such cases. To our knowledge, no theoretical error estimates have so far been published for (1).

In this paper, we first present the derivation of the 19-point compact finite difference scheme for (1); the resulting finite difference scheme is the same as that introduced and used in previous papers [4, 16]. Next, we derive discretization error estimates of type (2) for the real parameter s satisfying $\max(1.5, \mathrm{m}) < \mathrm{s} \le \mathrm{m}+4$, $m = 0, 1, 2$, under the assumption that the solution of the original boundary-value problem

(1) belongs to the Sobolev space $W_2^s(\Omega)$. These error estimates are derived using certain well-known techniques (see, e.g., [5, 6]) that employ the generalized Bramble–Hilbert lemma. Similar error estimates were previously obtained by Berikelashvili for two-dimensional convection-diffusion equations with constant coefficients in [7].

2. Construction of finite difference schemes. In the Hilbert space $L_2(\omega)$ we define the difference operators

$$\Lambda_\alpha = y_{\bar{x}_\alpha x_\alpha}, \quad \Lambda_{(\alpha)} = \sum_{\substack{k=1 \\ k\neq\alpha}}^{3} \Lambda_k, \quad \alpha = 1, 2, 3.$$

We need the following averaging operators for functions defined on Ω:

$$S_1^- v(x) = \frac{1}{h} \int_{x_1-h}^{x_1} v(t, x_2, x_3)\, dt,$$

$$T_1 v(x) = \frac{1}{h^2} \int_{x_1-h}^{x_1+h} (h - |x_1 - t|) v(t, x_2, x_3)\, dt.$$

The operators S_α^-, T_α are defined in a similar manner for $\alpha = 2, 3$. Notice that these operators commute in the case of different indices and

$$T_\alpha \frac{\partial^2 u}{\partial x_\alpha^2} = \Lambda_\alpha u, \quad T_\alpha \frac{\partial u}{\partial x_\alpha} = (S_\alpha^- u)_{x_\alpha}.$$

Let

$$T = \prod_{k=1}^{3} T_k, \quad T_{(\alpha)} = \prod_{\substack{k=1 \\ k\neq\alpha}}^{3} T_k, \quad \Lambda_{(\alpha)} = \sum_{\substack{k=1 \\ k\neq\alpha}}^{3} \Lambda_k.$$

We assume that the solution u of the boundary-value problem (1) belongs to the Sobolev space $W_2^s(\Omega), s > 1.5$. Applying operator T to (1) we obtain

$$\sum_{\alpha=1}^{3} \Lambda_\alpha T_{(\alpha)} u + \sum_{\alpha=1}^{3} \lambda_\alpha S_\alpha^- T_{(\alpha)} u_{x_\alpha} = Tf. \tag{3}$$

It can be easily verified that, on the set of sufficiently smooth functions, the following operators are equivalent (with errors of order $O(h^4)$):

$$T_\alpha \sim I + \frac{h^2}{12} \Lambda_\alpha$$

and consequently

$$T_{(\alpha)} \sim I + \frac{h^2}{12} \Lambda_{(\alpha)}.$$

Therefore we denote

$$\eta_\alpha = T_{(\alpha)} u - \big(I + \frac{h^2}{12} \Lambda_{(\alpha)}\big) u,$$

which implies

$$\Lambda_\alpha T_{(\alpha)} u = \Lambda_\alpha \left(I + \frac{h^2}{12} \Lambda_{(\alpha)} \right) u + \Lambda_\alpha \eta_\alpha. \tag{4}$$

Similarly,

$$S_\alpha^- T_{(\alpha)} \sim \frac{1}{2} (I + I^{(-\alpha)}) \left(I + \frac{h^2}{6} \Lambda_{(\alpha)} \right) - \frac{h^2}{12} S_\alpha^- T_{(\alpha)} \Delta.$$

The approximation error of this relation is

$$\eta^\alpha = S_\alpha^- T_{(\alpha)} u - \frac{1}{2} (I + I^{(-\alpha)}) \left(I + \frac{h^2}{6} \Lambda_{(\alpha)} \right) u + \frac{h^2}{12} S_\alpha^- T_{(\alpha)} \Delta u,$$

from which it follows that

$$S_\alpha^- T_{(\alpha)} u_{x_\alpha} = \partial_\alpha \left(I + \frac{h^2}{6} \Lambda_{(\alpha)} \right) u - \frac{h^2}{12} T \frac{\partial}{\partial x_\alpha} \Delta u + (\eta^\alpha)_{x_\alpha}. \tag{5}$$

From (3), (4) and (5), we obtain the following equality:

$$\begin{aligned} &\sum_{\alpha=1}^{3} \Lambda_\alpha \left(I + \frac{h^2}{12} \Lambda_{(\alpha)} \right) u + \sum_{\alpha=1}^{3} \lambda_\alpha \partial_\alpha \left(I + \frac{h^2}{6} \Lambda_{(\alpha)} \right) u \\ &+ \frac{h^2}{6} \sum_{1 \le \alpha < \beta \le 3} \lambda_\alpha \lambda_\beta T \frac{\partial^2 u}{\partial x_\alpha \partial x_\beta} + \frac{h^2}{12} \sum_{\alpha=1}^{3} \lambda_\alpha^2 T \frac{\partial^2 u}{\partial x_\alpha^2} + \sum_{\alpha=1}^{3} (\Lambda_\alpha \eta_\alpha + \lambda_\alpha \eta_{x_\alpha}^\alpha) = \varphi, \end{aligned} \tag{6}$$

where

$$\varphi = Tf + \frac{h^2}{12} \sum_{\alpha=1}^{3} \lambda_\alpha T \frac{\partial f}{\partial x_\alpha}. \tag{7}$$

Using

$$\eta^{\alpha\beta} = \frac{h^2}{6} \left(S_\alpha^- S_\beta^- T_\gamma u - \frac{1}{4} (I + I^{(-\alpha)} + I^{(-\beta)} + I^{(-\alpha)} I^{(-\beta)}) u \right),$$

$$\gamma = 6 - \alpha - \beta, \quad \eta_\alpha^\alpha = \frac{h^2}{12} (T_{(\alpha)} u - u),$$

we obtain

$$\frac{h^2}{6} T \frac{\partial^2 u}{\partial x_\alpha \partial x_\beta} = \frac{h^2}{6} \partial_\alpha \partial_\beta u + (\eta^{\alpha\beta})_{x_\alpha x_\beta}, \quad \frac{h^2}{12} T \frac{\partial^2 u}{\partial x_\alpha^2} = \frac{h^2}{12} \Lambda_\alpha u + \Lambda_\alpha \eta_\alpha^\alpha,$$

and from (6) we get

$$\begin{aligned} &\sum_{\alpha=1}^{3} \Lambda_\alpha \left(I + \frac{h^2}{12} \Lambda_{(\alpha)} \right) u + \sum_{\alpha=1}^{3} \lambda_\alpha \partial_\alpha \left(I + \frac{h^2}{6} \Lambda_{(\alpha)} \right) u \\ &+ \frac{h^2}{6} \sum_{1 \le \alpha < \beta \le 3} \lambda_\alpha \lambda_\beta \partial_\alpha \partial_\beta u + \frac{h^2}{12} \sum_{\alpha=1}^{3} \lambda_\alpha^2 \Lambda_\alpha u = \varphi + \psi, \end{aligned} \tag{8}$$

where the remainder term (truncation error) ψ is given by

$$\psi = -\sum_{\alpha=1}^{3}(\Lambda_\alpha \eta_\alpha + \lambda_\alpha \eta^\alpha_{x_\alpha} + \lambda^2_\alpha \Lambda_\alpha \eta^\alpha_\alpha) - \sum_{1\le\alpha<\beta\le 3} \lambda_\alpha \lambda_\beta (\eta^{\alpha\beta})_{x_\alpha x_\beta}. \tag{9}$$

By dropping the remainder term on the right-hand side of (8) and replacing the continuous solution $u(x)$ by the grid function $y(x)$, we obtain the finite difference scheme

$$-\mathcal{L}_h y = \varphi, \quad x \in \omega, \quad y \in L_2(\omega), \tag{10}$$

where the right-hand side φ is defined in (7) and

$$\mathcal{L}_h y \equiv (\mathcal{A}+\mathcal{B}+\mathcal{C})y, \quad \mathcal{A} = -\sum_{\alpha=1}^{3} \Lambda_\alpha \left(I + \frac{h^2}{12}\Lambda_{(\alpha)}\right), \quad \mathcal{C} = -\sum_{\alpha=1}^{3} \lambda_\alpha \partial_\alpha \left(I + \frac{h^2}{6}\Lambda_{(\alpha)}\right),$$

$$\mathcal{B} = -\frac{h^2}{6} \sum_{1\le\alpha<\beta\le 3} \lambda_\alpha \lambda_\beta \partial_\alpha \partial_\beta - \frac{h^2}{12} \sum_{\alpha=1}^{3} \lambda^2_\alpha \Lambda_\alpha, \quad \Lambda_{(\alpha)} = \sum_{\substack{\beta=1 \\ \beta\ne\alpha}}^{3} \Lambda_\beta.$$

3. A priori estimate of discretization error. We start with a few preliminary results.

The operators Λ_α are self-adjoint and negative definite in the Hilbert space $L_2(\omega)$ with respect to the inner product $(.,.)$, and (see, e.g., [8])

$$8I \le -\Lambda_\alpha \le (4/h^2)I, \quad \Lambda_\alpha \Lambda_\beta = \Lambda_\beta \Lambda_\alpha, \ \alpha,\beta = 1,2,3, \ \beta \ne \alpha. \tag{11}$$

Consequently, the operator

$$\Lambda = -\sum_{\alpha=1}^{3} \Lambda_\alpha$$

is self-adjoint and positive definite in $L_2(\omega)$, and

$$24I \le \Lambda \le (12/h^2)I.$$

One can easily verify that

$$||y||_{W_2^1(\omega)} = (\Lambda y, y)^{1/2}, \quad ||y||_{W_2^2(\omega)} = ||\Lambda y||.$$

LEMMA 3.1. *$\mathcal{L}_h$ is a positive definite operator in space $L_2(\omega)$, and for any $y \in L_2(\omega)$, we have the following estimate:*

$$3(\mathcal{L}_h y, y) \ge ||y||^2_{W_2^1(\omega)}. \tag{12}$$

Proof. Taking into account inequality (11), we have

$$(\mathcal{A}y, y) \ge (1/3)(\Lambda y, y) = (1/3)||y||^2_{W_2^1(\omega)} \tag{13}$$

as $I + (h^2/12)\Lambda_{(\alpha)} \ge (1/3)I$. It is easy to verify that $(\partial_\alpha y, v) = -(y, \partial_\alpha v)$ and consequently

$$(\mathcal{C}y, y) = 0. \tag{14}$$

Further, we have

$$(\mathcal{B}y,y)=\frac{h^2}{12}\sum_{\alpha=1}^{3}\lambda_\alpha^2||y_{\bar{x}_\alpha}||_{(\alpha)}^2+\frac{h^2}{6}\sum_{1\le\alpha<\beta\le3}\lambda_\alpha\lambda_\beta(\partial_\alpha y,\partial_\beta y).$$

However,

$$||\partial_\alpha y||^2=\sum_\omega h^3\left(\frac{y_{\bar{x}_\alpha}+y_{x_\alpha}}{2}\right)^2\le\sum_\omega\frac{h^3}{2}(y_{\bar{x}_\alpha}^2+y_{x_\alpha}^2)\le||y_{\bar{x}_\alpha}||_{(\alpha)}^2,$$

which yields

$$(\mathcal{B}y,y)\ge\frac{h^2}{12}\sum_{\alpha=1}^{3}\lambda_\alpha^2||\partial_\alpha y||^2+\frac{h^2}{6}\sum_{1\le\alpha<\beta\le3}\lambda_\alpha\lambda_\beta(\partial_\alpha y,\partial_\beta y)$$

$$=\frac{h^2}{12}||\sum_{\alpha=1}^{3}\lambda_\alpha\partial_\alpha y||^2\ge0. \tag{15}$$

The relations (13)–(15) complete the proof of Lemma 3.1. □

LEMMA 3.2. *For any $y\in L_2(\omega)$ the following estimates are valid:*

$$||\Lambda y||\le c_0||\mathcal{L}_h y||, \tag{16}$$

$$||\mathcal{L}_h^{-1}y||\le c_0||\Lambda^{-1}y||, \tag{17}$$

where $c_0=3+(9\sqrt{2}/4)\lambda$, $\lambda=max(|\lambda_1|,|\lambda_2|,|\lambda_3|)$.

Proof. We have $(\Lambda_1 y,\Lambda_2 y)=-(\Lambda_2 y_{\bar{x}_1},y_{\bar{x}_1})_{(1)}\ge(\partial_2 y_{\bar{x}_1},\partial_2 y_{\bar{x}_1})_{(1)}=(-\Lambda_1\partial_2 y,\partial_2 y)$, and $(\Lambda_1 y,\Lambda_3 y)\ge(-\Lambda_1\partial_3 y,\partial_3 y)$, which yields

$$(\Lambda_1 y,\lambda_2^2\Lambda_2 y+\lambda_3^2\Lambda_3 y+2\lambda_2\lambda_3\partial_2\partial_3 y)\ge||\partial_1(\lambda_2\partial_2+\lambda_3\partial_3)y||^2.$$

Using this inequality we obtain

$$(\mathcal{B}y,-\Lambda_1 y)=(h^2/12)\big(\lambda_1^2||\Lambda_1 y||^2+(\Lambda_1 y,\lambda_2^2\Lambda_2 y+\lambda_3^2\Lambda_3 y+2\lambda_2\lambda_3\partial_2\partial_3 y)$$

$$+2\lambda_1\lambda_2(\partial_1\partial_2 y,\Lambda_1 y)+2\lambda_1\lambda_3(\partial_1\partial_3 y,\Lambda_1 y)\big)$$

$$\ge(h^2/12)||(\lambda_1\Lambda_1+\lambda_2\partial_1\partial_2+\lambda_3\partial_1\partial_3)y||^2\ge0.$$

Analogously, $(\mathcal{B}y,-\Lambda_k y)\ge0$, for $k=2,3$ which yields $(\mathcal{B}y,\Lambda y)\ge0$. From (13), we have $(\mathcal{A}y,\Lambda y)\ge\frac{1}{3}||\Lambda y||^2$, which yields $((\mathcal{A}+\mathcal{B})y,\Lambda y)\ge\frac{1}{3}||\Lambda y||^2$, or

$$||\Lambda y||\le3||(\mathcal{A}+\mathcal{B})y||. \tag{18}$$

It is clear that

$$||\mathcal{C}y||\le\lambda\sum_{\alpha=1}^{3}\left\|\left(I+\frac{h^2}{6}\Lambda_{(\alpha)}\right)\partial_\alpha y\right\|\le\lambda\sum_{\alpha=1}^{3}||\partial_\alpha y||$$

$$\le\sqrt{3}\lambda\left(\sum_{\alpha=1}^{3}||\partial_\alpha y||^2\right)^{1/2}\le\sqrt{3}\lambda||y||_{W_2^1(\omega)}. \tag{19}$$

Using Lemma 3.1 and the discrete analogue of Friedrichs inequality $||y||^2_{W_2^1(\omega)} \geq 24||y||^2$, we obtain

$$||y||_{W_2^1(\omega)} \leq \frac{\sqrt{6}}{4}||\mathcal{L}_h y||, \tag{20}$$

and therefore the relation

$$||\mathcal{C}y|| \leq \frac{3\sqrt{2}}{4}\lambda||\mathcal{L}_h y|| \tag{21}$$

follows from inequality (19). Substituting $\mathcal{A}+\mathcal{B} = \mathcal{L}_h - \mathcal{C}$ into (18) and using (21), we obtain the estimate in (16). The estimate

$$||\Lambda y|| \leq c||\mathcal{L}_h^* y|| \ \ \forall y \in L_2(\omega) \tag{22}$$

can be obtained in a similar manner. Further, $\mathcal{L}_h^*\mathcal{L}_h$ is a self-adjoint positive definite operator in $L_2(\omega)$, and

$$||\mathcal{L}_h^{-1}y|| = ||(\mathcal{L}_h^*\mathcal{L}_h)^{-1}\mathcal{L}_h^* y|| = \sup_{v\neq 0}\frac{|(\mathcal{L}_h^* y, v)|}{||\mathcal{L}_h^*\mathcal{L}_h y||} = \sup_{v\neq 0}\frac{|(\Lambda^{-1}y, \Lambda\mathcal{L}_h v)|}{||\mathcal{L}_h^*\mathcal{L}_h y||}$$
$$\leq ||\Lambda^{-1}y||\sup_{v\neq 0}\frac{||\Lambda\mathcal{L}_h v)||}{||\mathcal{L}_h^*\mathcal{L}_h v||}.$$

Using this result and (22), we obtain (17). This completes the proof of Lemma 3.2. □

Let $z = y - u$, where u is the continuous solution of the boundary-value problem (1) and y is the solution of the finite difference scheme (10). Substituting $y = u+z$ into (10) and taking into account (8), we obtain the following problem for the discretization error z:

$$\mathcal{L}_h z = \psi, \ \ z \in L_2(\omega), \tag{23}$$

where the truncation error ψ is defined in (9).

LEMMA 3.3. *For the solution of problem* (23) *the following estimates hold:*

$$||z||_{W_2^m(\omega)} \leq c_m J_m(u), \ \ m = 0,1,2, \ c_1 = 3, \ c_2 = c_0, \tag{24}$$

where

$$J_0(u) = \sum_{\alpha=1}^{3}\left(||\eta_\alpha|| + \lambda||\eta^\alpha||_{(\alpha)} + \lambda^2||\eta^\alpha_\alpha||\right) + \sum_{1\leq\alpha<\beta\leq 3}\lambda^2||\eta^{\alpha\beta}||_{(\alpha,\beta)},$$
$$J_1(u) = \sum_{\alpha=1}^{3}\left(||\eta_{\alpha\bar{x}_\alpha}||_{(\alpha)} + \lambda||\eta^\alpha||_{(\alpha)} + \lambda^2||\eta^\alpha_{\alpha\bar{x}_\alpha}||_{(\alpha)}\right) + \sum_{1\leq\alpha<\beta\leq 3}\lambda^2||\eta^{\alpha\beta}_{x_\beta}||_{(\alpha)},$$
$$J_2(u) = \sum_{\alpha=1}^{3}\left(||\Lambda_\alpha\eta_\alpha|| + \lambda||\eta^\alpha_{x_\alpha}|| + \lambda^2||\Lambda_\alpha\eta^\alpha_\alpha||\right) + \sum_{1\leq\alpha<\beta\leq 3}\lambda^2||\eta^{\alpha\beta}_{x_\alpha x_\beta}||.$$

Proof. For $m = 0$, (24) can be established using the estimate (17), taking into account that $||\Lambda_\alpha v|| \leq ||\Lambda v||$, $||v_{\bar{x}_\alpha}||_{(\alpha)} \leq ||\Lambda v||$, and $||v_{x_\alpha x_\beta}|| \leq ||\Lambda v||$.

For $m = 1$, we have from (12) and (23):

$$||z||^2_{W_2^1(\omega)} \le 3|(\psi, z)|. \tag{25}$$

Using the definition of ψ in (9) and summing by parts, we obtain

$$(\psi, z) = -\sum_{\alpha=1}^{3} \left((\eta_{\alpha\bar{x}_\alpha}, z_{\bar{x}_\alpha})_{(\alpha)} + \lambda_\alpha (\eta^\alpha, z_{\bar{x}_\alpha})_{(\alpha)} + \lambda_\alpha^2 (\eta^\alpha_{\alpha\bar{x}_\alpha}, z_{\bar{x}_\alpha})_{(\alpha)} \right)$$
$$- \sum_{1\le\alpha<\beta\le 3} \lambda_\alpha \lambda_\beta (\eta^{\alpha\beta}_{x_\beta}, z_{\bar{x}_\alpha})_{(\alpha)}.$$

As $||z_{\bar{x}_\alpha}||_{(\alpha)} \le ||z||_{W_2^1(\omega)}$, using the Cauchy inequality we get $(\psi, z)| \le J_1(u)\, ||z||_{W_2^1(\omega)}$. Using (25) we obtain the desired result (24) for the case $m = 1$.

For $m = 2$, the estimate (24) follows immediately from (23) using (16). Thus, Lemma 3.3 is proved. □

To determine the rate of convergence of the finite difference scheme (10) with the help of Lemma 3.3, it is sufficient to estimate the corresponding norms of the expressions $\eta_\alpha, \eta^\alpha, \eta^\alpha_\alpha$, and $\eta^{\alpha\beta}$ on the right-hand side $J_k(u)$ of (24).

For this we need the next lemma.

LEMMA 3.4. *Assume that the linear functional $l(u)$ is bounded in $W_2^s(E)$, where $s = \bar{s} + \epsilon$, $\bar{s}$ is an integer, $0 < \epsilon \le 1$, and $l(P) = 0$ for every polynomial P of degree $\le \bar{s}$ in three variables. Then, there exists a constant c, independent of u, such that $|l(u)| \le c\ |u|_{W_2^s(\Omega)}$.*

This lemma is a particular case of the Dupont–Scott approximation theorem [9] and represents a generalization of the Bramble–Hilbert lemma [10] (see also [11]).

4. Estimate of the convergence rate. By π_k let us denote the set of all polynomials of degree $\le k$ in three variables.

We assert that the following inequalities hold for $\alpha = 1, 2, 3$:

$$||\eta_\alpha|| \le ch^s |u|_{W_2^s(\Omega)}, \quad s \in (1.5, 4],$$
$$||\eta_{\alpha\bar{x}_\alpha}||_{(\alpha)} \le ch^{s-1} |u|_{W_2^s(\Omega)}, \quad s \in (1.5, 5], \tag{26}$$
$$||\Lambda_\alpha \eta_\alpha|| \le ch^{s-2} |u|_{W_2^s(\Omega)}, \quad s \in (1.5, 6].$$

First we consider the expression η_1. Let $e = e(x) = \{\xi = (\xi_1, \xi_2, \xi_3) :\ |\xi_\alpha - x_\alpha| \le h,\ \alpha = 1, 2, 3\}$. By $\widetilde{u}(t)$ we denote a function obtained from $u(\xi)$ by changing the variables $\xi_\alpha = x_\alpha + t_\alpha h, \alpha = 1, 2, 3$, and mapping the function $e(x)$ onto $\widetilde{e} = \{t = (t_1, t_2, t_3) :\ |t_\alpha| \le 1,\ \alpha = 1, 2, 3\}$. Since $u(\xi) = u(x_1 + t_1 h, x_2 + t_2 h, x_3 + t_3 h) = \widetilde{u}(t)$, the expression

$$\eta_1 = T_2 T_3 u - \left(I + \frac{h^2}{12}\Lambda_2 + \frac{h^2}{12}\Lambda_3 \right) u$$

turns into

$$\eta_1 = \int_{-1}^{1}\int_{-1}^{1} (1 - |t_2|)(1 - |t_3|)\widetilde{u}(0, t_2, t_3)\, dt_2 dt_3$$
$$-\frac{1}{12}\big(\widetilde{u}(0,1,0) + \widetilde{u}(0,-1,0) + \widetilde{u}(0,0,1) + \widetilde{u}(0,0,-1) + 8\widetilde{u}(0,0,0)\big).$$

Consequently, $|\eta_1| \le c||\widetilde{u}||_{C(\widetilde{e})} \le c||\widetilde{u}||_{W_2^s(\widetilde{e})}$ as $W_2^s \subset C$ when $s > 1.5$. Utilizing the fact that η_1, as a functional of $\widetilde{u}$, vanishes on π_3 (which can be verified directly) and using Lemma 3.4, we obtain $|\eta_1| \le c|\widetilde{u}|_{W_2^s(\widetilde{e})}, s \in (1.5, 4]$. Reverting to the old variables, this yields $|\eta_1| \le ch^{s-1.5}|u|_{W_2^s(e)}, s \in (1.5, 4]$. Consequently, we have

$$||\eta_1||^2 = \sum_\omega h^3|\eta_1|^2 \le ch^{2s}\sum_\omega |u|^2_{W_2^s(e)} \le ch^{2s}|u|^2_{W_2^s(\Omega)}.$$

The other estimates in (26) can be obtained analogously, using the fact that the functionals $\eta_{\alpha\bar{x}_\alpha}$ and $\Lambda_\alpha\eta_\alpha$ vanish on π_4 and π_5, respectively.

The boundedness of the other error functionals is evident when $u \in W_2^s(\Omega)$, $s > 3/2$. Only the term $S_\alpha^- T_{(\alpha)}\Delta u$ involved in η^α needs a special consideration. For instance, for $\alpha = 1$ we have

$$S_1^- T_2T_3\Delta u = \left(T_2T_3\frac{\partial u}{\partial x_1}\right)_{\bar{x}_1} + \Lambda_2 S_1^- T_3 u + \Lambda_3 S_1^- T_2 u.$$

It follows from $u \in W_2^s$ that $(\partial u/\partial x_1) \in W_2^{s-1}$ and for fixed x_1 this derivative, as a function of variables x_2, x_3, belongs to $W_2^{s-3/2}$. Therefore the averaging $T_2T_3(\partial u/\partial x_1)$ makes sense and can be estimated by $||u||_{W_2^s(\Omega)}$.

The expressions η^α and $\eta^\alpha_{\bar{x}_\alpha}$ vanish on π_3 and π_4, respectively, and the following estimates can be obtained in a similar manner:

$$||\eta^\alpha|| \le ch^s|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 4], \tag{27}$$

$$||\eta^\alpha_{\bar{x}_\alpha}||_{(\alpha)} \le ch^{s-1}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 5]. \tag{28}$$

From (27), it follows that

$$||\eta^\alpha|| \le ch^{s-1}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 5]. \tag{29}$$

Similarly, for η^α_α, $\eta^\alpha_{\alpha\bar{x}_\alpha}$, and $\Lambda_\alpha\eta^\alpha_\alpha$, which vanish, respectively, on π_1, π_2, and π_3, we have the following estimates:

$$||\eta^\alpha_\alpha|| \le ch^{s+2}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 2],$$

$$||\eta^\alpha_{\alpha\bar{x}_\alpha}||_{(\alpha)} \le ch^{s+1}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 3],$$

$$||\Lambda_\alpha\eta^\alpha_\alpha|| \le ch^s|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 4].$$

Consequently, we obtain

$$||\eta^\alpha_\alpha|| \le ch^s|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 4],$$

$$||\eta^\alpha_{\alpha\bar{x}_\alpha}||_{(\alpha)} \le ch^{s-1}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 5], \tag{30}$$

$$||\Lambda_\alpha\eta^\alpha_\alpha|| \le ch^{s-2}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 6].$$

The expressions $\eta^{\alpha\beta}$, $\eta^{\alpha\beta}_{\bar{x}_\beta}$, and $\eta^{\alpha\beta}_{x_\alpha x_\beta}$ vanish, respectively, on π_1, π_2, and π_3. Consequently,

$$||\eta^{\alpha\beta}|| \le ch^{s+2}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 2],$$
$$||\eta^{\alpha\beta}_{\bar{x}_\beta}||_{(\alpha)} \le ch^{s+1}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 3].$$
$$||\eta^{\alpha\beta}_{x_\alpha x_\beta}|| \le ch^{s}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 4].$$

From this we obtain

$$\begin{aligned} &||\eta^{\alpha\beta}|| \le ch^{s}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 4], \\ &||\eta^{\alpha\beta}_{\bar{x}_\beta}||_{(\alpha)} \le ch^{s-1}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 5], \\ &||\eta^{\alpha\beta}_{x_\alpha x_\beta}|| \le ch^{s-2}|u|_{W_2^s(\Omega)}, \quad s \in (1.5, 6]. \end{aligned} \tag{31}$$

As a result, using estimates (26)–(31) and Lemma 3.3, we arrive at the following proposition.

THEOREM 4.1. *Let the exact solution of the boundary-value problem* (1) *belong to* $W_2^s(\Omega)$, $s > 1.5$. *Then, the discretization error of the finite difference scheme* (10) *in the discrete* W_2^m*-norm is determined by the estimate* (2) *for* s *satisfying* $\max(1.5, m) < s \le m + 4$, $m = 0, 1, 2$.

5. Comparisons with other methods. Let $\sigma = (\sigma_1, \sigma_2, \sigma_3)$ be a multi-index with components $-1, 0$, or 1, $|\sigma| = |\sigma_1| + |\sigma_2| + |\sigma_3|$, and $y_\sigma = y_{i+\sigma_1, j+\sigma_2, k+\sigma_3}$. At the node (ih, jh, kh) the finite difference scheme (10) may be represented in the form

$$\sum_{|\sigma|=0}^{2} a_\sigma y_\sigma = 6h^2\varphi, \tag{32}$$

where

$$\begin{aligned} a_\sigma &= -24 - h^2(\lambda_1^2 + \lambda_2^2 + \lambda_3^2) \quad \text{if } \sigma = (0,0,0), \\ a_\sigma &= 2 + h\sum_{k=1}^{3}\sigma_k\lambda_k + \frac{h^2}{2}\sum_{k=1}^{3}(\sigma_k\lambda_k)^2 \quad \text{if } |\sigma| = 1, \\ a_\sigma &= 1 + \frac{h}{2}\sum_{k=1}^{3}\sigma^k\lambda_k + \frac{h^2}{4}\sum_{k,l=1}^{3}\sigma_k\sigma_l\lambda_k\lambda_l \quad \text{if } |\sigma| = 2. \end{aligned}$$

The left-hand side of (32) is the same as that of the 19-point finite difference scheme defined by Gupta and Zhang [4].

In the case when $u \in W_2^s(\Omega)$, $s > 3.5$, and thus $f \in W_2^{s-2}(\Omega)$ is continuous, the right-hand side of the finite difference scheme (32) can be written as

$$F = 3h^2 f + \frac{h^4}{2}\sum_{\alpha=1}^{3}\lambda_\alpha\partial_\alpha f + \frac{h^2}{2}\sum_{|\sigma|=1} f_\sigma,$$

and the resulting scheme,

$$\sum_{|\sigma|=0}^{2} a_\sigma y_\sigma = F, \tag{33}$$

is exactly the same as in [4].

TABLE 5.1
Maximum errors for (34).

	Re = 1	Re = 1000
$h = 1/16$	1.50×10^{-5}	9.04×10^{-3}
$h = 1/32$	9.45×10^{-7}	1.15×10^{-3}
$h = 1/64$	5.96×10^{-8}	8.36×10^{-5}

Using the imbedding $W_2^2(\omega)$ into $C(\omega)$ for discrete functions of three variables, we obtain for both schemes (32) and (33) the discretization error estimate in uniform metric

$$||y - u||_{C(\omega)} \le ch^{s-2}||u||_{W_2^s(\Omega)}$$

with $s \in (2, 6]$ and $s \in (3.5, 6]$, respectively. This proves the fourth order discretization error estimates for both of the compact difference schemes for three-dimensional convection-diffusion equations.

As an illustration, we present numerical data from Gupta and Zhang [4] where the following convection diffusion equation was solved for various values of the parameter Re:

$$\Delta u - \mathrm{Re}\left[\cos\alpha\cos\beta\frac{\partial u}{\partial x_1} + \cos\alpha\sin\beta\frac{\partial u}{\partial x_2} + \sin\alpha\frac{\partial u}{\partial x_3}\right] = f(x). \tag{34}$$

In Table 5.1, we give data for the parameters $\alpha = 35°$, $\beta = 45°$ and the exact solution given by $u = \sin\pi x_1 \sin\pi x_2 \sin\pi x_3$. The maximum errors of the numerical solutions are exhibited in Table 5.1 for $h = 1/16$, $h = 1/32$, and $h = 1/64$ and for Re = 1 and Re = 1000. As shown in [4], the maximum norms of the numerical errors decay according to $O(h^4)$, and this rate of convergence has now been theoretically established by the results of this paper.

Remark 1. Our results are also valid in the case when the grid $\bar{\omega}$ is uniform in each direction x_1, x_2, and x_3 with steps h_1, h_2, and h_3, respectively. High accuracy compact finite difference schemes for problem (1) cannot, generally, be defined on irregular grids.

Remark 2. Equations of the type (1) often arise in problems of fluid dynamics as a linearized version of the momentum equation, and it is desirable for the corresponding finite difference schemes to have high accuracy, especially for large values of λ_α. Therefore it is important to represent the convergence estimates in terms of the parameter λ. Such estimates, e.g., in $W_2^1(\omega)$-norm for the difference scheme (10) have the form

$$\|z\|_{W_2^1(\omega)} \le cM(\lambda)h^{s-1}\|u\|_{W_2^s(\Omega)}, \quad s \in (1.5, 5],$$

where c is independent of h, λ, and $u(x)$ and

$$M(\lambda) = 1 + \lambda h + \lambda^2 h^2 \quad \text{if } s \in (1.5, 3],$$

$$M(\lambda) = 1 + \lambda h + \lambda^2 h^{5-s} \quad \text{if } s \in (3, 4],$$

$$M(\lambda) = 1 + \lambda^2 h^{5-s} \quad \text{if } s \in (4, 5].$$

Remark 3. As noted above, a few finite difference schemes for (1) (with variable coefficients λ_α) are known for approximation of the considered problem (see, e.g.,

[4, 16, 17] for three-dimensional problems and [19] for two-dimensional problems). However, the rate of convergence (of any order) can be exhibited only computationally in such cases as has been done through a number of test examples in the cited papers.

In some special cases, the fourth order convergence can be proved only when $u \in C^6(\bar{\Omega})$ whereas we have obtained convergence estimate of order h^s for $s \in (1.5, 4]$:

$$\|y - u\| \le ch^s \|u\|_{W_2^s(\Omega)}, \quad s \in (1.5, 4].$$

At present we do not have sufficient mathematical tools to establish estimates such as (12) for variable coefficient operators $\mathcal{L}_h$, and we plan to work on the error estimates for variable coefficient problems in the future.

Remark 4. Certain streamline-diffusion finite element methods (SDFEM) for problem (1) are also known in the literature (see, e.g., [12, 13, 14, 15]). For such methods, typically the convergence is obtained in the so-called streamline-diffusion norm. This norm weakens as the diffusion parameter tends to zero (in our case $\lambda \to \infty$). However, unlike our finite difference scheme (10), theoretical convergence estimates better than $O(h^2)$ are not available for SDFEM.

Acknowledgment. The authors wish to thank the referees for their helpful comments and for pointing out the SDFEM schemes. We hope that the confluence of these two methods (SDFEM and ours) will lead to strong and useful results in the future.

REFERENCES

[1] R. D. Lazarov, V. L. Makarov, and A. A. Samarskii, *Applying exact difference schemes to the construction and analysis of difference schemes on generalized solutions*, Mat. Sb., 117 (1982), pp. 469–480 (in Russian).

[2] V. V. Badagadze, *On the construction of difference schemes for a second order elliptic differential equation*, U.S.S.R. Comput. Math and Math Phys., 6 (1966), pp. 139–151 (in English); Zh. Vichisl. Mat. i. Mat. Fiz., 6 (1966), pp. 512–520 (in Russian).

[3] U. Ananthakrishnaiah, R. Manohar, and J. W. Stephenson, *Fourth-order finite difference methods for three-dimensional elliptic problems with variable coefficients*, Numer. Methods Partial Differential Equations, 3 (1987), pp. 229–240.

[4] M. M. Gupta and J. Zhang, *High accuracy multigrid solution of the 3D convection-diffusion equation*, Appl. Math. Comput., 113 (2000), pp. 249–274.

[5] A. A. Samarskii, R. D. Lazarov, and V. L. Makarov, *Difference Schemes for Differential Equations with Generalized Solutions*, Visshaya Shkola, Moscow, 1987 (in Russian).

[6] R. D. Lazarov, V. L. Makarov, and W. Weinelt, *On the convergence of difference schemes for the approximation of solutions $u \in W_2^m (m > 0.5)$ of elliptic equations with mixed derivatives*, Numer. Math., 44 (1984), pp. 223–232.

[7] G. Berikelashvili, *The difference schemes of high order accuracy for elliptic equations with lower derivatives*, Proc. A. Razmadze Math. Inst., 117 (1998), pp. 1–6.

[8] A. A. Samarskii, *Theory of Difference Schemes*, Nauka, Moscow, 1983 (in Russian).

[9] T. Dupont and R. Scott, *Polynomial approximation of functions in Sobolev spaces*, Math. Comp., 34 (1987), pp. 441–463.

[10] J. H. Bramble and S. R. Hilbert, *Bounds for a class of linear functionals with application to Hermite interpolation*, Numer. Math., 16 (1971), pp. 362–369.

[11] E. E. Süli, B. S. Jovanović, and L. D. Iovanović, *On the construction of finite difference schemes approximating generalized solutions*, Publ. Inst Math. (Beograd) (N.S.), 37 (1985), pp. 123–128.

[12] M. Stynes and L. Tobiska, *The SDFEM for a convection-diffusion problem with a boundary layer: Optimal error analysis and enhancement of accuracy*, SIAM J. Numer. Anal., 41 (2003), pp. 1620–1642.

[13] G. Matthies and L. Tobiska, *The streamline-diffusion method for conforming and nonconforming finite elements of lowest order applied to convection-diffusion problems*, Computing, 66 (2001), pp. 343–364.

[14] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli, *A priori error analysis of residual-free bubbles for advection-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 1933–1948.
[15] R. D. Lazarov, L. Tobiska, and P. S. Vassilevski, *Streamline diffusion least-squares mixed finite element methods for convection-diffusion problems*, East-West J. Numer. Math., 5 (1997), pp. 249–264.
[16] J. Zhang, *An explicit fourth-order compact finite difference scheme for three dimensional convection diffusion equation*, Comm. Numer. Methods Engrg., 14 (1998), pp. 263–280.
[17] J. Zhang, L. Ge, and M. M. Gupta, *Fourth order compact finite difference scheme for* $3D$ *convection diffusion equation with boundary layers on nonuniform grids*, Neural Parallel Sci. Comput., 8 (2000), pp. 373–392.
[18] A. Gopaul and M. Bhuruth, *private communication*, 2004.
[19] M. M. Gupta, R. Manohar, and J. W. Stephenson, *A single cell high order scheme for the convection-diffusion equation with variable coefficients*, Internat. J. Numer. Methods Fluids, 4 (1984), pp. 641–651.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 457–472

OPTIMAL CONVERGENCE FOR THE IMPLICIT SPACE-TIME DISCRETIZATION OF PARABOLIC SYSTEMS WITH *p*-STRUCTURE*

LARS DIENING†, CARSTEN EBMEYER‡, AND MICHAEL RŮŽIČKA†

Abstract. Parabolic systems with p-structure are considered on convex polyhedral domains under Dirichlet boundary conditions. A fully discrete scheme is studied using C^0-piecewise linear finite elements in space and the backward Euler difference scheme in time. A priori error estimates in quasi norms are proved, and optimal convergence rates are obtained.

Key words. degenerate parabolic system, time discretization, finite element methods, weak and strong solution, error analysis

AMS subject classifications. 65M15, 65M60

DOI. 10.1137/05064120X

1. Introduction.

Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a convex polyhedral domain, and let $T \in (0,\infty)$, $I = [0,T]$, and Q_T be the time-space cylinder $I \times \Omega$. For a given right-hand side $\mathbf{f} : Q_T \to \mathbb{R}^d$ and a given initial value $\mathbf{u}_0 : \Omega \to \mathbb{R}^d$ we seek $\mathbf{u} = (u_1, \dots, u_d)^\top : Q_T \to \mathbb{R}^d$ solving the system

$$\begin{aligned} \partial_t \mathbf{u} - \operatorname{div} \mathbf{S}(\nabla \mathbf{u}) &= \mathbf{f} && \text{in } Q_T\,, \\ \mathbf{u}(0) &= \mathbf{u}_0 && \text{in } \Omega\,, \\ \mathbf{u} &= 0 && \text{on } I \times \partial\Omega\,, \end{aligned} \tag{1.1}$$

where $\mathbf{S}$ has p-structure (cf. (2.1), (2.2)). Typical prototypes are

$$\mathbf{S}(\nabla \mathbf{u}) = \big|\nabla \mathbf{u}\big|^{p-2} \nabla \mathbf{u} \quad \text{and} \quad \mathbf{S}(\nabla \mathbf{u}) = \big(1 + \big|\nabla \mathbf{u}\big|^2\big)^{\frac{p-2}{2}} \nabla \mathbf{u}, \tag{1.2}$$

where $1 < p < \infty$.

In this paper we deal with a fully discrete scheme using continuous piecewise linear finite elements in space and the backward Euler time discretization. In the case of the heat equation, which corresponds to the case $p = 2$, it is well known that

$$\|u - U\|^2_{L^\infty(I,L^2(\Omega))} + \int_0^T \|\nabla u - \nabla U\|_2^2 \, dt \leq c\left(k^2 + h^2\right)$$

if the data are suitable. Here U denotes the space-time discretization of u, k is the size of each time step, and h is the mesh size; cf. [24]. The aim of this paper is to generalize this result to problem (1.1). Therefore, we estimate the approximation error in suitable quasi norms. Moreover, we give assumptions on the data that provide the regularity required for $\mathbf{u}$ for deriving these optimal convergence rates.

*Received by the editors September 26, 2005; accepted for publication (in revised form) August 28, 2006; published electronically February 26, 2007.
http://www.siam.org/journals/sinum/45-2/64120.html

†Mathematisches Institut, Abteilung für Angewandte Mathematik, Universität Freiburg, Eckerstraße 1, D-79104 Freiburg, Germany (diening@mathematik.uni-freiburg.de, rose@mathematik.uni-freiburg.de).

‡Mathematisches Seminar, Universität Bonn, Nussallee 15, D-53115 Bonn, Germany (ebmeyermsl@uni-bonn.de).

Extensive research has been carried out for time discretizations and finite element approximations of problems with p-structure; cf. [1, 2, 3, 4, 7, 10, 13, 16, 19, 20, 21, 25]. However, with the exception of [1, 10, 21], all results are suboptimal in the sense that either the order of the error estimate is not optimal or the assumed regularity of the solution is too high and not realistic for general situations. For instance, in the case of the elliptic degenerate p-Laplace equation it is well known that sharp error estimates cannot be obtained if the error is estimated in Sobolev or weighted Sobolev norms. A significant development in the error estimation for such degenerate problems is the quasi-norm approach of Barrett and Liu; cf. [2, 3, 19]. The key idea is to introduce a quasi norm $\|\cdot\|_{(\nabla \mathbf{u})}$ satisfying

$$\|\nabla \mathbf{u} - \nabla \mathbf{U}\|^2_{(\nabla \mathbf{u})} \cong \int_\Omega \big(\mathbf{S}(\nabla \mathbf{u}) - \mathbf{S}(\nabla \mathbf{U})\big) \cdot (\nabla \mathbf{u} - \nabla \mathbf{U})\, dx,$$

where $1 < p < \infty$. Utilizing this quasi norm, sharp error estimates can be established; e.g., for elliptic problems with p-structure it was shown in [2] that

$$\|\nabla u - \nabla u_h\|_{(\nabla u)} \le c \inf_{v_h \in V_h} \|\nabla u - \nabla v_h\|_{(\nabla u),}$$

where u is the weak solution, u_h is the finite element approximation of u, and V_h is the finite element space of continuous piecewise linear functions vanishing on $\partial\Omega$. Moreover, in [10] the finite element interpolation error estimation theory in the quasi norms was established. It was proved that

$$\inf_{v_h \in V_h} \|\nabla u - \nabla v_h\|^2_{(\nabla u)} \le c\, h^2 \int_\Omega |\nabla u|^{p-2} |\nabla^2 u|^2,$$

where h is the mesh size. Since regularity results for problems with p-structure are available (cf. [15, 6, 12, 11]), the integral on the right-hand side is finite and thus the optimal convergence rate for elliptic problems

$$\|\nabla u - \nabla u_h\|^2_{(\nabla u)} \le c\, h^2$$

was obtained. In the case of time discretizations most results again are only suboptimal. Usually, error estimates of the form

$$\|u - u_k\|^2_{l^\infty(I_k, L^2(\Omega))} + \|\nabla u - \nabla u_k\|^2_{l^p(I_k, L^p(\Omega))} \le c\, k^{2\alpha(p)},$$

where u is the solution of the parabolic problem with p-structure, u_k is the backward Euler time discretized solution, and $\alpha(p) \in (0, 1)$, are proved. In [21] optimal estimates for the time discretization by means of the Yoshida approximation are proved; i.e., one has $\alpha(p) = 1$. However, no spatial discretization is treated there and it seems to be difficult to include further nonmonotone nonlinearities in this approach. Recently, abstract error estimates for a full space-time discretization with $\alpha(p) = 1$ were given in [1]. Under additional regularity assumptions explicit error estimates in terms of the time-step size and mesh size are derived there. The approach presented here is completely different.

In this paper we treat parabolic systems with p-structure for all $\frac{2d}{d+2} < p < \infty$. We will estimate the approximation error between the solution $\mathbf{u}$ of problem (1.1) and the solution $\mathbf{U}$ of the fully discrete scheme using continuous piecewise linear finite elements in space and the backward Euler time discretization in quasi norms and show the *optimal* convergence rate

$$\|\mathbf{u} - \mathbf{U}\|^2_{L^\infty(I, L^2(\Omega))} + \int_0^T \|\nabla \mathbf{u} - \nabla \mathbf{U}\|^2_{(\nabla \mathbf{u})}\, dt \le c\,(k^2 + h^2),$$

where k is the time-step size and h the mesh size, which are related by the mesh ratio condition $h^{\alpha(p,d)} \le c\,k$ (cf. (5.1)). This result is derived for the natural regularity of the problem, which is available under rather general assumptions on the data. Note that the condition (5.1) is present even for $p = 2$. We believe that this condition is only of technical character and hope that it can be removed by an appropriate argument.

2. Preliminaries. For matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ with components A_{ij} and B_{ij} we denote $\mathbf{A}\cdot\mathbf{B} = \sum_{i,j=1}^{d} A_{ij}B_{ij}$. We write $f \cong g$ if and only if there exist constants $c_0, c_1 > 0$, such that

$$c_0 f \le g \le c_1 f\,,$$

where we always indicate on what the constants may depend. We use the usual notation of Lebesgue-, Sobolev-, and Bochner-spaces, respectively, namely $\big(L^p(\Omega), \|\cdot\|_p\big)$, $\big(W^{k,p}(\Omega), \|\cdot\|_{k,p}\big)$, and $\big(L^p(I,X), \|\cdot\|_{L^p(I,X)}\big)$, where X is some Banach space, respectively. We denote $\langle f, g\rangle := \int_\Omega f(x)\,g(x)\,dx$. Moreover, for $f : I\times\Omega \to \mathbb{R}$ we often write $f(t)$ instead of $f(t,\cdot)$.

Concerning the structure of the system (1.1) we assume that the operator induced by $-\operatorname{div}\mathbf{S}(\nabla\mathbf{u})$ has *p-structure*; i.e., there exist $p > 1$, $\kappa \ge 0$, and $\gamma_1, \gamma_2 > 0$ such that

$$\sum_{i,j,k,l=1}^{d} \frac{\partial S_{ij}(\mathbf{A})}{\partial A_{kl}} B_{ij}B_{kl} \ge \gamma_1 \big(\kappa + |\mathbf{A}|\big)^{p-2}|\mathbf{B}|^2 \tag{2.1}$$

is satisfied for all $\mathbf{B}, \mathbf{A} \in \mathbb{R}^{d\times d}$, and that

$$\left|\frac{\partial S_{ij}(\mathbf{A})}{\partial A_{kl}}\right| \le \gamma_2 \big(\kappa + |\mathbf{A}|\big)^{p-2} \tag{2.2}$$

is satisfied for all $i, j, k, l = 1, \dots, d$. Note that the above prototypes (1.2) satisfy these assumptions. Closely related to the operator $\mathbf{S}$ with p-structure is the function $\mathbf{F} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ defined by

$$\mathbf{F}(\mathbf{B}) := (\kappa + |\mathbf{B}|)^{\frac{p-2}{2}}\mathbf{B}\,.$$

This is clarified by the following algebraic lemma.

LEMMA 2.1. *For all $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ there holds*

$$\big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big)\cdot\big(\mathbf{A} - \mathbf{B}\big) \cong |\mathbf{A} - \mathbf{B}|^2\big(\kappa + |\mathbf{B}| + |\mathbf{A}|\big)^{p-2} \tag{2.3a}$$

$$\cong |\mathbf{F}(\mathbf{A}) - \mathbf{F}(\mathbf{B})|^2 \tag{2.3b}$$

with constants depending on p only.

The proof can be found in the appendix. For $\mathbf{A}, \mathbf{B} : \Omega \to \mathbb{R}^{d\times d}$ we introduce the quasi norm (cf. [2, 3, 19])

$$\|\mathbf{A}\|_{(\mathbf{B})} := \left(\int_\Omega \big(\kappa + |\mathbf{B}(x)| + |\mathbf{A}(x)|\big)^{p-2}|\mathbf{A}(x)|^2\,dx\right)^{\frac{1}{2}}.$$

Lemma 2.1 and the fact that $|\mathbf{B}| + |\mathbf{A} - \mathbf{B}| \cong |\mathbf{B}| + |\mathbf{A}|$ imply the following lemma.

LEMMA 2.2. *For all* $\mathbf{v}, \mathbf{w} \in W^{1,p}(\Omega)$ *there holds*

$$\begin{aligned} \|\nabla \mathbf{v} - \nabla \mathbf{w}\|^2_{(\nabla \mathbf{w})} &\cong \big\langle \mathbf{S}(\nabla \mathbf{v}) - \mathbf{S}(\nabla \mathbf{w}), \nabla \mathbf{v} - \nabla \mathbf{w} \big\rangle \\ &\cong \|\mathbf{F}(\nabla \mathbf{v}) - \mathbf{F}(\nabla \mathbf{w})\|_2^2 \end{aligned} \tag{2.4}$$

with constants depending on p *only.*

Furthermore, for $\delta > 0$ there is a constant $c_\delta > 0$ such that for all $\lambda, \mu, \nu \geq 0$ (cf. [3])

$$(\lambda + \mu)^{p-2} \mu \, \nu \leq \delta (\lambda + \mu)^{p-2} \mu^2 + c_\delta (\lambda + \nu)^{p-2} \nu^2 \, .$$

Due to this inequality, (2.3a), and (6.8) there holds the following result.

LEMMA 2.3. *For* $\delta > 0$ *there exists* $c_\delta > 0$ *such that for all* $\mathbf{A}, \mathbf{B}, \mathbf{C} \in \mathbb{R}^{d \times d}$ *there holds*

$$\begin{aligned} &\big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big) \cdot \big(\mathbf{A} - \mathbf{C}\big) \\ &\quad \leq \delta \, \big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big) \cdot \big(\mathbf{A} - \mathbf{B}\big) + c_\delta \, \big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{C})\big) \cdot \big(\mathbf{A} - \mathbf{C}\big) \end{aligned}$$

and

$$\big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big) \cdot \big(\mathbf{A} - \mathbf{C}\big) \leq \delta \, \big|\mathbf{F}(\mathbf{A}) - \mathbf{F}(\mathbf{B})\big|^2 + c_\delta \, \big|\mathbf{F}(\mathbf{A}) - \mathbf{F}(\mathbf{C})\big|^2 \, .$$

Especially for $\mathbf{v}, \mathbf{w}_1, \mathbf{w}_2 \in W^{1,p}(\Omega)$ we easily deduce from this that

$$\begin{aligned} &\big\langle \mathbf{S}(\nabla \mathbf{v}) - \mathbf{S}(\nabla \mathbf{w}_1), \nabla \mathbf{v} - \nabla \mathbf{w}_2 \big\rangle \\ &\qquad \leq \delta \, \|\nabla \mathbf{v} - \nabla \mathbf{w}_1\|^2_{(\nabla \mathbf{v})} + c_\delta \, \|\nabla \mathbf{v} - \nabla \mathbf{w}_2\|^2_{(\nabla \mathbf{v})} \end{aligned}$$

and

$$\begin{aligned} &\big\langle \mathbf{S}(\nabla \mathbf{v}) - \mathbf{S}(\nabla \mathbf{w}_1), \nabla \mathbf{v} - \nabla \mathbf{w}_2 \big\rangle \\ &\qquad \leq \delta \, \|\mathbf{F}(\nabla \mathbf{v}) - \mathbf{F}(\nabla \mathbf{w}_1)\|_2^2 + c_\delta \, \|\mathbf{F}(\nabla \mathbf{v}) - \mathbf{F}(\nabla \mathbf{w}_2)\|_2^2 \, . \end{aligned} \tag{2.5}$$

3. The continuous problem. In this section we investigate the regularity of solutions of system (1.1). Throughout the remainder of the paper we assume that

$$\begin{aligned} \mathbf{u}_0 &\in W^{1,2}_0(\Omega) \cap W^{1,p}_0(\Omega), \\ \operatorname{div}\big(\mathbf{S}(\nabla \mathbf{u}_0)\big) &\in L^2(\Omega), \\ \mathbf{f} &\in L^{p'}(I, L^{p'}(\Omega)) \cap C(I, L^2(\Omega)) \cap L^2(I, W^{1,2}_0(\Omega)), \\ \partial_t \mathbf{f} &\in L^2(I, L^2(\Omega)), \end{aligned} \tag{3.1}$$

where p' is the dual exponent of p. We call $\mathbf{u}$ a weak solution of problem (1.1) if $\mathbf{u} \in L^\infty(I, L^2(\Omega)) \cap L^p(I, W^{1,p}_0(\Omega))$ with $\partial_t \mathbf{u} \in L^{p'}(I, (W^{1,p}_0(\Omega))^*)$ satisfies for almost all $t \in I$

$$\begin{aligned} \big\langle \partial_t \mathbf{u}, \mathbf{w} \big\rangle_{W^{1,p}_0(\Omega)} + \big\langle \mathbf{S}(\nabla \mathbf{u}), \nabla \mathbf{w} \big\rangle &= \langle \mathbf{f}, \mathbf{w} \rangle \qquad \forall \mathbf{w} \in W^{1,p}_0(\Omega), \\ \mathbf{u}(0) &= \mathbf{u}_0, \end{aligned} \tag{3.2}$$

where $\langle \mathbf{v}, \mathbf{w} \rangle_{W^{1,p}_0(\Omega)}$ denotes the duality pairing in $W^{1,p}_0(\Omega)$. It is well known (cf. [18, 14]) that for $p \geq \frac{2d}{d+2}$ there exists a unique weak solution of (1.1) with

$$\|\mathbf{u}\|^2_{L^\infty(I, L^2(\Omega))} + \|\mathbf{u}\|^p_{L^p(I, W^{1,p}(\Omega))} \leq C \, \big(\, \|\mathbf{f}\|^{p'}_{L^{p'}(I, L^{p'}(\Omega))} + \|\mathbf{u}_0\|^2_2 \big) \, .$$

LEMMA 3.1. *The solution* $\mathbf{u}$ *of* (1.1) *satisfies under the assumptions* (3.1)

$$\|\partial_t \mathbf{u}\|^2_{C(I,L^2(\Omega))} + \int_0^T \left\|\partial_t\big(\mathbf{F}(\nabla \mathbf{u})\big)\right\|_2^2 \, dt \le c(\mathbf{f}, \mathbf{u}_0).$$

Proof. In [14, Theorem 6.2.1][1] (cf. [18] for $p \ge 2$) it is shown that problem (1.1) possesses under the above assumptions a solution $\mathbf{u} \in C^1_w(I, L^2(\Omega)) \cap C_w(I, W^{1,p}_0(\Omega))$; i.e., $\partial_t \mathbf{u}$ is weakly continuous in $L^2(\Omega)$, and $\mathbf{u}$ is weakly continuous in $W^{1,p}_0(\Omega)$. In particular we have

$$\sup_{t\in[0,T]} \|\partial_t \mathbf{u}(t)\|_{L^2(\Omega)} \le c\,, \tag{3.3}$$

which is the first part of the assertion. In order to prove the second part we use the difference quotient with respect to time. Let $D^\tau \mathbf{u}(t,x) := \frac{1}{\tau}(\mathbf{u}(t+\tau,x) - \mathbf{u}(t,x))$, $\tau > 0$. We apply D^τ to (3.2) and set $\mathbf{w} := D^\tau \mathbf{u}$. Then for all t

$$\big\langle \partial_t D^\tau \mathbf{u}, D^\tau \mathbf{u}\big\rangle + \big\langle D^\tau\big(\mathbf{S}(\nabla \mathbf{u})\big), D^\tau \nabla \mathbf{u}\big\rangle = \langle D^\tau \mathbf{f}, D^\tau \mathbf{u}\rangle.$$

This and (2.4) yield

$$\frac{1}{2}\partial_t \|D^\tau \mathbf{u}\|_2^2 + c\left\|D^\tau\big(\mathbf{F}(\nabla \mathbf{u})\big)\right\|_2^2 \le \|D^\tau \mathbf{f}\|_2 \|D^\tau \mathbf{u}\|_2.$$

Integration over $t \in [0, t^*]$, $0 \le t^* \le T$, and taking the supremum over $t^* \in [0,T]$ imply

$$\begin{aligned} &\|D^\tau \mathbf{u}\|^2_{C(I,L^2(\Omega))} + \int_0^T \left\|D^\tau\big(\mathbf{F}(\nabla \mathbf{u})\big)\right\|_2^2 \, dt \\ &\le c\left(\|D^\tau \mathbf{u}(0)\|_2^2 + \|\partial_t \mathbf{f}\|^2_{L^2(I,L^2(\Omega))} + \|\partial_t \mathbf{u}\|^2_{L^2(I,L^2(\Omega))}\right) \\ &\le c\,, \end{aligned}$$

where we used (3.1) and (3.3). The second assertion now follows from the properties of the difference quotient. ∎

4. Backward time discretization. We introduce the notation $t_n := nk$, $N := \lfloor T/k \rfloor$, $I_n := (t_{n-1}, t_n)$,

$$\mathbf{u}^n(x) := \mathbf{u}(t_n, x), \quad \text{and} \quad \overline{\mathbf{u}}^n(x) := k^{-1}\int_{I_n} \mathbf{u}(s,x)\, ds.$$

We define $\mathbf{y}^0 := \mathbf{u}_0$ and successively let $\mathbf{y}^n$, $n = 1, \dots, N$, be the solutions of

$$\begin{aligned} d_t \mathbf{y}^n(x) - \operatorname{div}\big(\mathbf{S}(\nabla \mathbf{y}^n(x))\big) &= \overline{\mathbf{f}}^n(x) \quad \text{in } \Omega, \\ \mathbf{y}^n &= 0 \qquad \text{on } \partial\Omega, \end{aligned} \tag{4.1}$$

where $d_t \mathbf{y}^n := \frac{1}{k}(\mathbf{y}^n - \mathbf{y}^{n-1})$. Since each $\mathbf{y}^n$ is the solution of a stationary problem, $\mathbf{y}^n$ is well defined and unique. Using the test function $\mathbf{y}^n$ and the properties of $\mathbf{S}$ yields

$$\sup_n \|\mathbf{y}^n\|_2^2 + k\sum_n \|\mathbf{y}^n\|^p_{1,p} \le C\left(1 + k\sum_n \left\|\overline{\mathbf{f}}^n\right\|^{p'}_{p'} + \|\mathbf{u}_0\|_2^2\right).$$

[1] Note that in [14] only $\mathbf{f} = \mathbf{0}$ is treated. However, under the above assumptions the proof continues to hold without any changes.

Furthermore, we have the following regularity results.

LEMMA 4.1. *The solutions* $\mathbf{y}^n$ *of* (4.1) *satisfy*

$$\sup_{1\le n\le N}\|\nabla \mathbf{y}^n\|_2^2 + k\sum_{n=1}^{N}\big\|\nabla\big(\mathbf{F}(\nabla \mathbf{y}^n)\big)\big\|_2^2 \le C(\mathbf{f},\mathbf{u}_0).$$

Proof. We only sketch the idea and refer the reader to [11, 12] for details. Let us define $D^\sigma \mathbf{y}^n(x) := \frac{1}{|\sigma|}(\mathbf{y}^n(x+\sigma)-\mathbf{y}^n(x))$, $\sigma\in\mathbb{R}^d\setminus\{0\}$. We multiply (4.1) with $D^{-\sigma}D^\sigma \mathbf{y}^n$ and find

$$\langle d_t D^\sigma \mathbf{y}^n, D^\sigma \mathbf{y}^n\rangle + \big\langle D^\sigma\big(\mathbf{S}(\nabla \mathbf{y}^n)\big), D^\sigma\nabla \mathbf{y}^n\big\rangle = \langle D^\sigma \overline{\mathbf{f}}^n, D^\sigma \mathbf{y}^n\rangle.$$

In view of (2.4) we obtain

$$\frac{1}{2}d_t\,\|D^\sigma \mathbf{y}^n\|_2^2 + c\,\big\|D^\sigma\big(\mathbf{F}(\nabla \mathbf{y}^n)\big)\big\|_2^2 \le \big\|D^\sigma \overline{\mathbf{f}}^n\big\|_2\|D^\sigma \mathbf{y}^n\|_2.$$

Summation over $n=1,\dots,M$ implies

$$\begin{aligned}&\big\|D^\sigma \mathbf{y}^M\big\|_2^2 + k\sum_{n=1}^{M}\big\|D^\sigma\big(\mathbf{F}(\nabla \mathbf{y}^n)\big)\big\|_2^2\\ &\quad\le c\left(\|D^\sigma \mathbf{u}_0\|_2^2 + k\sum_{n=1}^{M}\left(\big\|D^\sigma\overline{\mathbf{f}}^n\big\|_2^2 + \big\|D^\sigma \mathbf{y}^n\big\|_2^2\right)\right).\end{aligned}$$

We take the supremum over $M=1,\dots,N$ and get

$$\begin{aligned}&\sup_n\|D^\sigma \mathbf{y}^n\|_2^2 + k\sum_{n=1}^{N}\big\|D^\sigma\big(\mathbf{F}(\nabla \mathbf{y}^n)\big)\big\|_2^2\\ &\quad\le c\left(\|\nabla \mathbf{u}_0\|_2^2 + \|\nabla \mathbf{f}\|_{L^2(I,L^2(\Omega))}^2 + k\sum_{n=1}^{N}\|D^\sigma \mathbf{y}^n\|_2^2\right).\end{aligned}$$

The discrete Gronwall lemma yields for $k<1/(2c)$

$$\sup_n\|D^\sigma \mathbf{y}^n\|_2^2 + k\sum_{n=1}^{N}\big\|D^\sigma\big(\mathbf{F}(\nabla \mathbf{y}^n)\big)\big\|_2^2 \le C(\mathbf{f},\mathbf{u}_0).$$

The properties of the difference quotient prove the lemma. ∎

LEMMA 4.2. *The solutions* $\mathbf{y}^n$ *of* (4.1) *satisfy*

$$k\sum_{n=1}^{N}\|\mathbf{y}^n\|_{W^{2,\frac{4}{4-p}}(\Omega)}^2 \le C(\mathbf{f},\mathbf{u}_0)\qquad \text{if } p\le 2$$

and

$$k\sum_{n=1}^{N}\|\mathbf{y}^n\|_{\mathcal{N}^{1+\frac{2}{p},p}(\Omega)}^p \le C(\mathbf{f},\mathbf{u}_0)\qquad \text{if } p>2,$$

where $\mathcal{N}^{1+\frac{2}{p},p}(\Omega)$ *denotes a Nikol′skiĭ space (cf.* [17]*).*

Proof. First, we consider the case that $p \le 2$. Hölder's inequality yields

$$
\begin{aligned}
(4.2)\qquad & k\sum_{n=1}^{N} \left\|\nabla^2 \mathbf{y}^n\right\|_{\frac{4}{4-p}}^2 \\
&= k\sum_{n=1}^{N} \left[\int_\Omega (\kappa + |\nabla \mathbf{y}^n|)^{\frac{4-2p}{4-p}} \, (\kappa + |\nabla \mathbf{y}^n|)^{\frac{2p-4}{4-p}} \, |\nabla^2 \mathbf{y}^n|^{\frac{4}{4-p}}\right]^{\frac{4-p}{2}} \\
&\le k\sum_{n=1}^{N} \left[\left(\int_\Omega (\kappa + |\nabla \mathbf{y}^n|)^2\right)^{\frac{2-p}{4-p}} \left(\int_\Omega (\kappa + |\nabla \mathbf{y}^n|)^{p-2} \, |\nabla^2 \mathbf{y}^n|^2\right)^{\frac{2}{4-p}}\right]^{\frac{4-p}{2}} \\
&\le \sup_n \|\kappa + |\nabla \mathbf{y}^n|\|_2^{2-p} \left[k\sum_{n=1}^{N} \int_\Omega (\kappa + |\nabla \mathbf{y}^n|)^{p-2} \, |\nabla^2 \mathbf{y}^n|^2\right].
\end{aligned}
$$

Note that in the case $\kappa = 0$ one carries out the above calculation with some $\widetilde{\kappa} > 0$ and bounds in the last line the term $(\widetilde{\kappa} + |\nabla \mathbf{y}^n|)^{p-2} \, |\nabla^2 \mathbf{y}^n|^2$ by $|\nabla \mathbf{y}^n|^{p-2} \, |\nabla^2 \mathbf{y}^n|^2$. A straightforward calculation shows that $(\kappa + |\nabla \mathbf{y}^n|)^{p-2} \, |\nabla^2 \mathbf{y}^n|^2 \cong \left|\nabla\left(\mathbf{F}(\nabla \mathbf{y}^n)\right)\right|^2$, with constants depending only on p. Thus it follows from Lemma 4.1 that the right-hand side of (4.2) is finite.

Now let us treat the case that $p > 2$. The proof of Lemma 4.1 entails

$$
k\sum_{n=1}^{N} \langle D^\sigma(\mathbf{S}(\nabla \mathbf{y}^n)), D^\sigma \nabla \mathbf{y}^n\rangle \cong k\sum_{n=1}^{N} \left\|D^\sigma\left(\mathbf{F}(\nabla \mathbf{y}^n)\right)\right\|_2^2 \le C(\mathbf{f}, \mathbf{u}_0).
$$

Noting that

$$
\begin{aligned}
\left(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\right) \cdot \left(\mathbf{A} - \mathbf{B}\right) &\cong |\mathbf{A} - \mathbf{B}|^2 \left(\kappa + |\mathbf{A}| + |\mathbf{A} - \mathbf{B}|\right)^{p-2} \\
&\ge |\mathbf{A} - \mathbf{B}|^p,
\end{aligned}
$$

it follows that

$$
k\sum_{n=1}^{N} \int_\Omega \sigma^{-2} \, |\nabla \mathbf{y}^n(x + \sigma) - \nabla \mathbf{y}^n(x)|^p \, dx \le C(\mathbf{f}, \mathbf{u}_0).
$$

Thus, difference quotients of order $\frac{2}{p}$ of $\nabla \mathbf{y}^n$ are bounded in $L^p(\Omega)$. This yields the assertion. ∎

Remark 4.3. Note that one can improve for $p \le 2$ the regularity stated above by using methods from [8], [7, Prop. 3.7], [22, Prop. 3.23]. Essentially one uses $d_t^2 \mathbf{y}^n$ and $-\Delta \mathbf{y}^n$ (treating the term with $d_t \mathbf{y}^n$ as a right-hand side) as test functions to derive better regularity properties for $\mathbf{y}^n$. This implies, with the help of a parabolic embedding theorem, that $\mathbf{y}^n \in l^\infty(I_k, W^{1,r}(\Omega))$ with $r = r(p, d) > 2$. Thus (4.2) can be improved. The improved Lemma 4.2 would lead to an improvement of the mesh ratio condition (5.1). Due to the complicated dependence on p and d we do not proceed this way here, since it does not lead to optimal results.

We will now estimate the error between $\mathbf{u}$ and $\mathbf{y}^n$.

LEMMA 4.4. *There is a constant c independent of k such that*

$$
\begin{aligned}
&\sup_{1 \le n \le N} \|\mathbf{u}^n - \mathbf{y}^n\|_2^2 + \sum_{n=1}^{N} \int_{I_n} \left\|\mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{y}^n)\right\|_2^2 \, dt \\
&\quad \le c \sum_{n=1}^{N} \int_{I_n} \left\|\mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{u}^n)\right\|_2^2 \, dt.
\end{aligned}
$$

Proof. Averaging (1.1) over I_n we find

$$d_t \mathbf{u}^n - \operatorname{div}(\overline{\mathbf{S}(\nabla \mathbf{u})}^n) = \overline{\mathbf{f}}^n. \tag{4.3}$$

Let $\mathbf{e}^n := \mathbf{u}^n - \mathbf{y}^n$. Taking the difference between (4.3) and (4.1) and multiplying by $\mathbf{e}^n$ we get

$$\left\langle d_t \mathbf{e}^n, \mathbf{e}^n \right\rangle + \left\langle \overline{\mathbf{S}(\nabla \mathbf{u})}^n - \mathbf{S}(\nabla \mathbf{y}^n), \nabla \mathbf{e}^n \right\rangle = 0.$$

Let $1 \le M \le N$. Multiplying by k and summing over n from 1 to M we obtain

$$\frac{1}{2}\langle \mathbf{e}^M, \mathbf{e}^M \rangle + \frac{k^2}{2} \sum_{n=1}^{M} \|d_t \mathbf{e}^n\|_2^2 + \sum_{n=1}^{M} \int_{I_n} \left\langle \mathbf{S}(\nabla \mathbf{u}(t)) - \mathbf{S}(\nabla \mathbf{y}^n), \nabla \mathbf{e}^n \right\rangle dt = 0.$$

Hence,

$$\begin{aligned} &\frac{1}{2}\langle \mathbf{e}^M, \mathbf{e}^M \rangle + \sum_{n=1}^{M} \int_{I_n} \left\langle \mathbf{S}(\nabla \mathbf{u}(t)) - \mathbf{S}(\nabla \mathbf{y}^n), \nabla \mathbf{u}(t) - \nabla \mathbf{y}^n \right\rangle dt \\ &\quad \le \sum_{n=1}^{M} \int_{I_n} \left\langle \mathbf{S}(\nabla \mathbf{u}(t)) - \mathbf{S}(\nabla \mathbf{y}^n), \nabla \mathbf{u}(t) - \nabla \mathbf{u}^n \right\rangle dt. \end{aligned}$$

Inequality (2.5) yields

$$\begin{aligned} &\left\langle \mathbf{S}(\nabla \mathbf{u}(t)) - \mathbf{S}(\nabla \mathbf{y}^n), \nabla \mathbf{u}(t) - \nabla \mathbf{u}^n \right\rangle \\ &\quad \le \delta \left\| \mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{y}^n) \right\|_2^2 + c_\delta \left\| \mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{u}^n) \right\|_2^2. \end{aligned}$$

We absorb the first term of the right-hand side and utilize (2.4) to prove the lemma. ∎

Corollary 4.5. *There is a constant c independent of k such that*

$$\begin{aligned} &\sup_{1 \le n \le N} \|\mathbf{u}^n - \mathbf{y}^n\|_2^2 + k \sum_{n=1}^{N} \left\| \mathbf{F}(\nabla \mathbf{u}^n) - \mathbf{F}(\nabla \mathbf{y}^n) \right\|_2^2 \\ &\qquad \le c \sum_{n=1}^{N} \int_{I_n} \|\mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{u}^n)\|_2^2 \, dt. \end{aligned} \tag{4.4}$$

Proof. The left-hand side of (4.4) is bounded by

$$\sup_{1 \le n \le N} \|\mathbf{u}^n - \mathbf{y}^n\|_2^2 + 2 \sum_{n=1}^{N} \int_{I_n} \|\mathbf{F}(\nabla \mathbf{u}^n) - \mathbf{F}(\nabla \mathbf{u}(t))\|_2^2 + \|\mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{y}^n)\|_2^2 \, dt \,.$$

The statement now follows from Lemma 4.4. ∎

Lemma 4.6. *There holds*

$$\sum_{n=1}^{N} \int_{I_n} \left\| (\mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{u}^n) \right\|_2^2 dt \le c\, k^2 \, \left\| \partial_t \big(\mathbf{F}(\nabla \mathbf{u}) \big) \right\|_{L^2(I, L^2(\Omega))}^2 .$$

Proof. We estimate

$$\begin{aligned}\int_{I_n}\int_\Omega |\mathbf{F}(\nabla \mathbf{u}^n)-\mathbf{F}(\nabla \mathbf{u})|^2\,dx\,dt &= \int_{I_n}\int_\Omega \left|\int_t^{t_n} \partial_\tau\Big(\mathbf{F}\big(\nabla \mathbf{u}(\tau)\big)\Big)\,d\tau\right|^2 dx\,dt\\ &\le k\int_{I_n}\int_\Omega\int_t^{t_n}\left|\partial_\tau\Big(\mathbf{F}\big(\nabla \mathbf{u}(\tau)\big)\Big)\right|^2 d\tau\,dx\,dt\\ &\le k^2\int_{I_n}\int_\Omega |\partial_t\big(\mathbf{F}(\nabla \mathbf{u})\big)|^2\,dx\,dt.\end{aligned}$$

Summing over n, this yields the assertion. ∎

From Lemmas 3.1, 4.4, and 4.6 and Corollary 4.5 we now deduce the following theorem.

THEOREM 4.7. *There is a constant c independent of k such that*

$$\begin{aligned}&\sup_{1\le n\le N}\|\mathbf{u}^n-\mathbf{y}^n\|_2^2 + k\sum_{n=1}^N \big\|\mathbf{F}(\nabla \mathbf{u}^n)-\mathbf{F}(\nabla \mathbf{y}^n)\big\|_2^2\\ &\quad+\sum_{n=1}^N\int_{I_n}\|\mathbf{F}(\nabla \mathbf{u}(t))-\mathbf{F}(\nabla \mathbf{y}^n)\|_2^2\,dt \le c\,k^2.\end{aligned}$$

5. Finite element discretization. Let $\mathcal{T}_h$ be a decomposition of Ω into closed d-simplices, where h denotes the mesh size. We suppose that $\mathcal{T}_h$ is a regular triangulation in the sense of [5, section 3.2]. Moreover, let

$$h^{\alpha(p,d)} \le c\,k\,, \tag{5.1}$$

where $\alpha(p,d)=2-d(1-\frac{p}{2})$ if $p\in(1,2]$ and $\alpha(p,d)=2+(d-2)(1-\frac{2}{p})=d+\frac{2(2-d)}{p}$ if $p\in[2,\infty)$. We define the space of continuous piecewise linear finite elements,

$$V_h := \big\{\boldsymbol{\chi}\in C^0(\Omega;\mathbb{R}^d)\,:\,\boldsymbol{\chi}|_K \text{ is linear for all } K\in\mathcal{T}_h \text{ and } \boldsymbol{\chi}|_{\partial\Omega}=0\big\}\,.$$

Further, let $P_h\mathbf{u}\in V_h$ be some appropriate interpolant of $\mathbf{u}$.[2]

For $\mathbf{U}^0 := P_h\mathbf{u}_0$ let the functions $\mathbf{U}^n\in V_h$, $n=1,\dots,N$, be the solutions of the algebraic equations

$$\big\langle d_t\mathbf{U}^n,\boldsymbol{\chi}^n\big\rangle + \big\langle \mathbf{S}(\nabla\mathbf{U}^n),\nabla\boldsymbol{\chi}^n\big\rangle = \big\langle \overline{\mathbf{f}}^n,\boldsymbol{\chi}^n\big\rangle \qquad \forall\boldsymbol{\chi}^n\in V_h. \tag{5.2}$$

The solvability of (5.2) follows easily from Brouwer's fixed point theorem and the properties of $\mathbf{S}$. We define $\mathbf{U}: I\times\Omega\to\mathbb{R}^d$ by

$$\mathbf{U}(t,x) := \begin{cases}\mathbf{U}^0(x) & \text{for } t=0,\\ \mathbf{U}^n(x) & \text{for } t_{n-1}<t\le t_n.\end{cases}$$

The aim of the this section is to prove the following theorem.

[2] In the case $p>\frac{d}{2}$ one can take the Lagrange interpolation operator (cf. [10])and for smaller values of p the interpolation operator from [23] (cf. [9]).

THEOREM 5.1. *Under the assumptions* (3.1) *and* (5.1) *there is for any* $p > \frac{2d}{d+2}$ *a constant* c *independent of* h *and* k *such that*

$$\sup_{1\le n\le N} \|\mathbf{u}^n - \mathbf{U}^n\|_2^2 + k\sum_{n=1}^{N} \left\|\mathbf{F}(\nabla \mathbf{u}^n) - \mathbf{F}(\nabla \mathbf{U}^n)\right\|_2^2 + \sum_{n=1}^{N}\int_{I_n} \|\mathbf{F}(\nabla \mathbf{u}(t)) - \mathbf{F}(\nabla \mathbf{U}^n)\|_2^2 \, dt \le c\left(h^2 + k^2\right).$$

In terms of quasi norms this reads

$$\|\mathbf{u} - \mathbf{U}\|^2_{L^\infty(0,T;L^2(\Omega))} + \int_0^T \|\nabla \mathbf{u} - \nabla \mathbf{U}\|^2_{(\nabla \mathbf{u}(t))} \, dt \le c\left(h^2 + k^2\right).$$

COROLLARY 5.2. *Under the assumptions* (3.1) *and* (5.1) *there exists for any* $p \in (\frac{2d}{d+2}, 2]$ *a constant* c *independent of* h *and* k *such that*

$$\|\nabla \mathbf{u} - \nabla \mathbf{U}\|_{L^p(I;L^p(\Omega))} \le c\,(h + k).$$

We proceed in several steps. First we estimate $\mathbf{y}^n - \mathbf{U}^n$.

LEMMA 5.3. *There is a constant* c *independent of* h *and* k *such that*

$$\sup_{1\le n\le N} \|\mathbf{y}^n - \mathbf{U}^{\mathbf{n}}\|_2^2 + k\sum_{n=1}^{N} \|\mathbf{F}(\nabla \mathbf{y}^n) - \mathbf{F}(\nabla \mathbf{U}^n)\|_2^2 \le c\left[\sum_{n=0}^{N} \|\mathbf{y}^n - P_h\mathbf{y}^n\|_2^2 + k\sum_{n=1}^{N} \|\mathbf{F}(\nabla \mathbf{y}^n) - \mathbf{F}(\nabla P_h\mathbf{y}^n)\|_2^2\right].$$

Proof. Let $\mathbf{E}^n = \mathbf{y}^n - \mathbf{U}^n$. Taking the difference between (4.1) and (5.2) and multiplying by $P_h\mathbf{E}^n$ we get

$$\left\langle d_t\mathbf{E}^n, P_h\mathbf{E}^n\right\rangle + \left\langle \mathbf{S}(\nabla \mathbf{y}^n) - \mathbf{S}(\nabla \mathbf{U}^n), \nabla P_h\mathbf{E}^n\right\rangle = 0.$$

Notice that $\langle d_t\mathbf{E}^n, \mathbf{E}^n\rangle = \langle d_t\mathbf{E}^n, P_h\mathbf{E}^n\rangle + \langle d_t\mathbf{E}^n, \mathbf{y}^n - P_h\mathbf{y}^n\rangle$ and

$$k\sum_{n=1}^{M} \left\langle d_t\mathbf{E}^n, \mathbf{E}^n\right\rangle = \frac{1}{2}\left\|\mathbf{E}^M\right\|_2^2 - \frac{1}{2}\left\|\mathbf{E}^0\right\|_2^2 + \frac{k^2}{2}\sum_{n=1}^{M} \left\|d_t\mathbf{E}^n\right\|_2^2.$$

Thus, we find

$$\begin{aligned}\frac{1}{2}\left\|\mathbf{E}^M\right\|_2^2 &+ \frac{k^2}{2}\sum_{n=1}^{M} \|d_t\mathbf{E}^n\|_2^2 + k\sum_{n=1}^{M} \left\langle \mathbf{S}(\nabla \mathbf{y}^n) - \mathbf{S}(\nabla \mathbf{U}^n), \nabla \mathbf{y}^n - \nabla \mathbf{U}^n\right\rangle \\ &\le k\sum_{n=1}^{M} \left\langle \mathbf{S}(\nabla \mathbf{y}^n) - \mathbf{S}(\nabla \mathbf{U}^n), \nabla \mathbf{y}^n - \nabla P_h\mathbf{y}^n\right\rangle + \frac{1}{2}\left\|\mathbf{y}^0 - P_h\mathbf{y}^0\right\|_2^2 \\ &\quad + k\sum_{n=1}^{M} \left\langle d_t\mathbf{E}^n, \mathbf{y}^n - P_h\mathbf{y}^n\right\rangle.\end{aligned}$$

Inequality (2.5) yields

$$\begin{aligned}&\left\langle \mathbf{S}(\nabla \mathbf{y}^n) - \mathbf{S}(\nabla \mathbf{U}^n), \nabla \mathbf{y}^n - \nabla P_h\mathbf{y}^n\right\rangle \\ &\quad \le \delta\,\|\mathbf{F}(\nabla \mathbf{y}^n) - \mathbf{F}(\nabla \mathbf{U}^n)\|_2^2 + c_\delta\,\|\mathbf{F}(\nabla \mathbf{y}^n) - \mathbf{F}(\nabla P_h\mathbf{y}^n)\|_2^2.\end{aligned}$$

Furthermore,

$$k\sum_{n=1}^{M}\left\langle d_t\mathbf{E}^n,\mathbf{y}^n-P_h\mathbf{y}^n\right\rangle\le\delta k^2\sum_{n=1}^{M}\|d_t\mathbf{E}^n\|_2^2+c_\delta\sum_{n=1}^{M}\|\mathbf{y}^n-P_h\mathbf{y}^n\|_2^2.$$

We absorb the terms with δ into the left-hand side and use (2.4). Taking the supremum over $M=1,\dots,N$, the assertion follows. □

THEOREM 5.4. *Under the assumptions* (3.1) *and* (5.1) *there is for any* $p>\frac{2d}{d+2}$ *a constant* c *independent of* h *and* k *such that*

$$\sup_{1\le n\le N}\|\mathbf{y}^n-\mathbf{U}^{\mathbf{n}}\|_2^2+k\sum_{n=1}^{N}\|\mathbf{F}(\nabla\mathbf{y}^n)-\mathbf{F}(\nabla\mathbf{U}^n)\|_2^2\le c\,h^2.$$

Proof. Below we will show that under the assumption (5.1) there holds

$$\sum_{n=0}^{N}\|\mathbf{y}^n-P_h\mathbf{y}^n\|_2^2\le c\,h^2. \tag{5.3}$$

Further, it holds that

$$\|\mathbf{F}(\nabla\mathbf{y}^n)-\mathbf{F}(\nabla P_h\mathbf{y}^n)\|_2^2\le c\,\|\nabla\mathbf{y}^n-\nabla P_h\mathbf{y}^n\|_{(\nabla\mathbf{y}^n)}^2\le c\,h^2\,\|\nabla\mathbf{F}(\nabla\mathbf{y}^n)\|_2^2. \tag{5.4}$$

The proof of this inequality can be found in [10] for the cases $d=2,3$ and $p>\frac{d}{2}$. The method there, however, works for any $d\ge 2$. The case $p>1$ will be treated in [9]. Due to Lemma 4.1 we obtain

$$k\sum_{n}\|\mathbf{F}(\nabla\mathbf{y}^n)-\mathbf{F}(\nabla P_h\mathbf{y}^n)\|_2^2\le c\,h^2. \tag{5.5}$$

In view of (5.3), (5.5), and Lemma 5.3 the assertion follows.

Next we prove (5.3). Let us note that $\mathbf{y}^0\in W^{1,2}(\Omega)$ and thus

$$\left\|\mathbf{y}^0-P_h\mathbf{y}^0\right\|_2^2\le c\,h^2.$$

Now let us distinguish two cases.

Case 1. $p\le 2$. It is well known that (cf. [5, Theorem 3.1.5])

$$\|\mathbf{y}^n-P_h\mathbf{y}^n\|_2^2\le c\,h^{2\beta}\,\left\|\nabla^2\mathbf{y}^n\right\|_{\frac{4}{4-p}}^2,$$

where $\beta=2+d\big(\frac12-\frac{4-p}{4}\big)$. Notice that due to (5.1) we have $h^{2-d(1-\frac{p}{2})}\le c\,k$. Thus, $h^{2\beta}=h^2\,h^{2-d(1-\frac{p}{2})}\le c\,h^2\,k$. We get

$$\sum_{n=1}^{N}\|\mathbf{y}^n-P_h\mathbf{y}^n\|_2^2\le c\,h^2\,k\sum_{n=1}^{N}\left\|\nabla^2\mathbf{y}^n\right\|_{\frac{4}{4-p}}^2.$$

In view of Lemma 4.2, estimate (5.3) follows.

Case 2. $p>2$. Since the proof of Theorem 3.1.5 in [5] is based on the chain rule and the formula of change of variables, the result can be easily generalized to Nikol′skiĭ spaces. Thus we have

$$\|\mathbf{y}^n-P_h\mathbf{y}^n\|_2^2\le c\,h^{2\beta}\,\|\mathbf{y}^n\|_{\mathcal{N}^{1+\frac{2}{p},p}(\Omega)}^2,$$

where $\beta = (1+\frac{2}{p}) + d\big(\frac{1}{2}-\frac{1}{p}\big)$. Recalling (5.1), that is, $h^{d+\frac{2(2-d)}{p}} \le c\,k$, we have $h^{2\beta} = h^2\, h^{d+\frac{2(2-d)}{p}} \le c\,h^2\,k$. Hence,

$$\sum_{n=1}^{N} \|\mathbf{y}^n - P_h\mathbf{y}^n\|_2^2 \le c\,h^2\,k \sum_{n=1}^{N} \|\mathbf{y}^n\|^2_{\mathcal{N}^{1+\frac{2}{p},p}(\Omega)}.$$

Using Lemma 4.2 we obtain estimate (5.3). ∎

We now get to the proof of our main result.

Proof of Theorem 5.1. For $t \in I_n$ there holds

$$\|\mathbf{F}(\nabla\mathbf{u}) - \mathbf{F}(\nabla\mathbf{U}^n)\|_2^2 \le c\Big(\|\mathbf{F}(\nabla\mathbf{u}) - \mathbf{F}(\nabla\mathbf{y}^n)\|_2^2 + \|\mathbf{F}(\nabla\mathbf{y}^n) - \mathbf{F}(\nabla\mathbf{U}^n)\|_2^2\Big).$$

Moreover, we have

$$\sup_{t\in I_n} \|\mathbf{u} - \mathbf{U}^n\|_2^2 \le c\Big(\sup_{t\in I_n} \|\mathbf{u} - \mathbf{u}^n\|_2^2 + \|\mathbf{u}^n - \mathbf{y}^n\|_2^2 + \|\mathbf{y}^n - \mathbf{U}^n\|_2^2\Big)$$

and

$$\sup_{t\in I_n} \|\mathbf{u} - \mathbf{u}^n\|^2_{L^2(\Omega)} \le c\,k^2\,\|\partial_t\mathbf{u}\|^2_{L^\infty(I_n,L^2(\Omega))}.$$

From Theorems 4.7 and 5.4 and the calculations above we conclude the assertion. ∎

Proof of Corollary 5.2. Let $\omega := \kappa + |\nabla\mathbf{u}| + |\nabla(\mathbf{u}-\mathbf{U})|$. Hölder's inequality (with $q_1 = \frac{2}{p}$ and $q_2 = \frac{2}{2-p}$) entails

$$\begin{aligned}\int_0^T\!\!\int_\Omega |\nabla(\mathbf{u}-\mathbf{U})|^p\,dx\,dt &= \int_0^T\!\!\int_\Omega \omega^{\frac{(2-p)p}{2}}\,\omega^{\frac{(p-2)p}{2}}\,|\nabla(\mathbf{u}-\mathbf{U})|^p\,dx\,dt\\ &\le \int_0^T\Big[\int_\Omega \omega^p\,dx\Big]^{\frac{2-p}{2}}\Big[\int_\Omega \omega^{p-2}\,|\nabla(\mathbf{u}-\mathbf{U})|^2\,dx\Big]^{\frac{p}{2}}\,dt\\ &\le c\int_0^T \|\nabla(\mathbf{u}-\mathbf{U})\|^2_{(\nabla\mathbf{u})}\,dt.\end{aligned}$$

Here we have used the fact that $\|\omega\|^{2-p}_{L^\infty(0,T;L^p(\Omega)} \le c$. In the case of $\kappa = 0$ we proceed as in the proof of Lemma 4.2. Utilizing Theorem 5.1 we obtain

$$\|\nabla(\mathbf{u}-\mathbf{U})\|^2_{L^p(0,T;L^p(\Omega)} \le c\int_0^T \|\nabla(\mathbf{u}-\mathbf{U})\|^2_{(\nabla\mathbf{u})}\,dt \le c\,(h^2+k^2).$$

This yields the assertion. ∎

6. Appendix. We will prove Lemma 2.1 in several steps.

LEMMA 6.1. *Let* $\alpha > -1$ *and* $\mathbf{A},\mathbf{B} \in \mathbb{R}^{d\times d}$ *with* $|\mathbf{A}| + |\mathbf{B}| > 0$*; then*

$$c_0(\alpha)\,\big(|\mathbf{A}| + |\mathbf{B}|\big)^\alpha \le \int_0^1 |\theta\mathbf{A} + (1-\theta)\mathbf{B}|^\alpha\,d\theta \le c_1(\alpha)\,\big(|\mathbf{A}| + |\mathbf{B}|\big)^\alpha \tag{6.1}$$

with

$$c_0(\alpha) := \min\Big\{\frac{1}{\alpha+1}, \frac{2^{-\alpha}}{\alpha+1}, 2^{-\alpha}\Big\}, \qquad c_1(\alpha) := \max\Big\{\frac{1}{\alpha+1}, \frac{2^{-\alpha}}{\alpha+1}, 2^{-\alpha}\Big\}.$$

The constants c_0 and c_1 are optimal.

Proof. Let $f(\mathbf{A},\mathbf{B})$ denote the middle expression of (6.1). We have

$$\big|(1-\theta)|\mathbf{A}|-\theta|\mathbf{B}|\big| \le |(1-\theta)\mathbf{A}+\theta\mathbf{B}| \le (1-\theta)|\mathbf{A}|+\theta|\mathbf{B}|, \tag{6.2}$$

and for all $\alpha > -1$, $|\mathbf{A}|+|\mathbf{B}| > 0$ it holds that

$$\begin{aligned} \int_0^1 \big|(1-\theta)|\mathbf{A}|-\theta|\mathbf{B}|\big|^\alpha \, d\theta &= \frac{|\mathbf{A}|^{\alpha+1}+|\mathbf{B}|^{\alpha+1}}{(\alpha+1)(|\mathbf{A}|+|\mathbf{B}|)}\,, \\ \int_0^1 \big|(1-\theta)|\mathbf{A}|+\theta|\mathbf{B}|\big|^\alpha \, d\theta &= \frac{|\mathbf{A}|^{\alpha+1}-|\mathbf{B}|^{\alpha+1}}{(\alpha+1)(|\mathbf{A}|-|\mathbf{B}|)}\,, \end{aligned} \tag{6.3}$$

where the last expression can be continuously extended for $|\mathbf{A}| = |\mathbf{B}|$ by $|\mathbf{A}|^\alpha$. Define $g, h : \mathbb{R}^{\ge 0} \times \mathbb{R}^{\ge 0} \to \mathbb{R}$ by

$$g_1(s,t) := \frac{s^{\alpha+1}+t^{\alpha+1}}{(\alpha+1)(s+t)^{\alpha+1}}, \qquad g_2(s,t) := \frac{s^{\alpha+1}-t^{\alpha+1}}{(\alpha+1)(s-t)(s+t)^\alpha},$$

with g_2 continuously extended on $\{s=t\}$ by $g_2(s,s) = 2^{-\alpha}$. Then (6.2) and (6.3) imply

$$\min\big\{g_1(|\mathbf{A}|,|\mathbf{B}|), g_2(|\mathbf{A}|,|\mathbf{B}|)\big\} \le f(\mathbf{A},\mathbf{B}) \le \max\big\{g_1(|\mathbf{A}|,|\mathbf{B}|), g_2(|\mathbf{A}|,|\mathbf{B}|)\big\}. \tag{6.4}$$

Since the g_j are scaling invariant under $(s,t) \mapsto \lambda(s,t)$ it suffices to estimate g_j on the set $N := \{s+t=1, s\ge 0, t\ge 0\}$. On this set N the functions g_j simplify to

$$h_1(s,t) := \frac{s^{\alpha+1}+t^{\alpha+1}}{\alpha+1}, \qquad h_2(s,t) := \frac{s^{\alpha+1}-t^{\alpha+1}}{(\alpha+1)(s-t)};$$

i.e., $h_j(s,t) = g_j(s,t)$ for $(s,t) \in N$. We will see that the $h_j|_N$ assumes its extrema at $(1,0),(0,1),(\frac12,\frac12)$. At these points

$$\begin{aligned} h_1(0,1) = h_1(1,0) &= 1/(\alpha+1)\,, & h_2(0,1) = h_2(1,0) &= 1/(\alpha+1)\,, \\ h_1\big(\tfrac12,\tfrac12\big) &= 2^{-\alpha}/(\alpha+1)\,, & h_2\big(\tfrac12,\tfrac12\big) &= 2^{-\alpha}\,. \end{aligned}$$

Let $(s_0,t_0) \in N \setminus \{(1,0),(0,1),(1/2,1/2)\}$ be another extremum of $h_j|_N$. Then by the method of Langrange multipliers $(\nabla h_j)(s_0,t_0) = \lambda_0\,(1,1)$ for some $\lambda_0 \in \mathbb{R}$. For $(s,t) \in N \setminus \{(1,0),(0,1),(1/2,1/2)\}$ it holds that

$$\begin{aligned} (\nabla h_1)(s,t) &= (s^\alpha, t^\alpha), \\ (\nabla h_2)(s,t) &= \left(\frac{\alpha s^{\alpha+1}+(\alpha+1)s^\alpha t - t^{\alpha+1}}{(\alpha+1)(s-t)^2}, \frac{\alpha t^{\alpha+1}+(\alpha+1)t^\alpha s - s^{\alpha+1}}{(\alpha+1)(s-t)^2}\right). \end{aligned}$$

If $(\nabla h_1)(s_0,t_0) = \lambda_0\,(1,1)$, then $s_0 = t_0$. If $(\nabla h_2)(s_0,t_0) = \lambda_0\,(1,1)$, then $\alpha s_0^{\alpha+1} + (\alpha+1)s_0^\alpha t_0 - t_0^{\alpha+1} = \alpha t_0^{\alpha+1} + (\alpha+1)t_0^\alpha s_0 - s_0^{\alpha+1}$, which also implies $s_0 = t_0$. In both cases $s_0 = t_0$, which contradicts the choice of (s_0,t_0). Therefore, $(1,0),(0,1),(1/2,1/2)$ are the only extrema of h_j on N. Thus the extreme values of g_1 are $1/(\alpha+1)$ and $2^{-\alpha}/(\alpha+1)$, and the extreme values of g_2 are $1/(\alpha+1)$ and $2^{-\alpha}$. This and (6.4) prove (6.1). The optimality of the constants in (6.2) follows from

$$\begin{aligned} f(\mathbf{A},\mathbf{A}) = g_2(|\mathbf{A}|,|\mathbf{A}|) &= 2^{-\alpha}\,, \\ f(\mathbf{A},0) = g_2(|\mathbf{A}|,0) \quad &= 1/(\alpha+1)\,, \\ f(\mathbf{A},-\mathbf{A}) = g_1(|\mathbf{A}|,|\mathbf{A}|) &= 2^{-\alpha}/(\alpha+1)\,. \end{aligned}$$

The assertion follows. ∎

LEMMA 6.2. *For all* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ *with* $|\mathbf{A}| + |\mathbf{B}| > 0$, *all* $\alpha > -1$, *and* $\kappa \geq 0$, *there holds*

$$\int_0^1 \big(\kappa + |\theta\,\mathbf{A} + (1-\theta)\mathbf{B}|\big)^{\alpha}\, d\theta \cong \big(\kappa + |\mathbf{B}| + |\mathbf{A}|\big)^{\alpha}, \tag{6.5}$$

with constants depending on p *only.*

Proof. Let $\mathbf{A}_\theta := \theta\mathbf{A} + (1-\theta)\mathbf{B}$. Using the convexity of $t \mapsto (\kappa+t)^{\alpha+2}$ and (6.1) we deduce

$$\begin{aligned}
\int_0^1 (\kappa + |\mathbf{A}_\theta|)^{\alpha}\, d\theta &\geq \int_0^1 \frac{(\kappa + |\mathbf{A}_\theta|)^{\alpha+2}}{(\kappa + |\mathbf{A}| + |\mathbf{B}|)^2}\, d\theta \\
&\geq \frac{(\kappa + \int_0^1 |\mathbf{A}_\theta|\, d\theta)^{\alpha+2}}{(\kappa + |\mathbf{A}| + |\mathbf{B}|)^2} \\
&\geq \frac{(\kappa + \frac{1}{4}(|\mathbf{A}| + |\mathbf{B}|))^{\alpha+2}}{(\kappa + |\mathbf{A}| + |\mathbf{B}|)^2} \\
&\geq 4^{-(\alpha+2)}(\kappa + |\mathbf{A}| + |\mathbf{B}|)^{\alpha}.
\end{aligned}$$

Since $\alpha > -1$, there exists $r > 1$ with $r\alpha > -1$, and the mapping $t \mapsto (\kappa+t)^{r\alpha}t$ is nondecreasing on $[0,\infty)$. We estimate with (6.1)

$$\begin{aligned}
\int_0^1 (\kappa + |\mathbf{A}_\theta|)^{\alpha}\, d\theta &= \int_0^1 \Big((\kappa + |\mathbf{A}_\theta|)^{r\alpha}|\mathbf{A}_\theta|\Big)^{\frac{1}{r}} |\mathbf{A}_\theta|^{-\frac{1}{r}}\, d\theta \\
&\leq \int_0^1 \Big((\kappa + |\mathbf{A}| + |\mathbf{B}|)^{r\alpha}(|\mathbf{A}| + |\mathbf{B}|)\Big)^{\frac{1}{r}} |\mathbf{A}_\theta|^{-\frac{1}{r}}\, d\theta \\
&\leq \frac{2^{1/r}}{1-\frac{1}{r}}\,(\kappa + |\mathbf{A}| + |\mathbf{B}|)^{\alpha}.
\end{aligned}$$

This proves the lemma. ∎

LEMMA 6.3. *Let* $\mathbf{S}$ *satisfy* (2.1), (2.2). *Then for all* $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d\times d}$ *it holds that*

$$\mathbf{S}(\mathbf{A}) \cdot \mathbf{A} \cong |\mathbf{A}|^2\big(\kappa + |\mathbf{A}|\big)^{p-2}, \tag{6.6}$$

$$\big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big) \cdot \big(\mathbf{A} - \mathbf{B}\big) \cong \big|\mathbf{A} - \mathbf{B}\big|^2\big(\kappa + |\mathbf{B}| + |\mathbf{A}|\big)^{p-2}, \tag{6.7}$$

$$\big|\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big| \cong \big|\mathbf{A} - \mathbf{B}\big|\big(\kappa + |\mathbf{B}| + |\mathbf{A}|\big)^{p-2} \tag{6.8}$$

with constants depending on p *but not on* κ.

Proof. Note that statement (6.6) is a special case of (6.7) by setting $\mathbf{B} = \mathbf{0}$. In the case $\mathbf{B} = \mathbf{A} = \mathbf{0}$, nothing has to be proved. In the case $|\mathbf{A}| + |\mathbf{B}| > 0$, Lemma 6.2 and the assumptions (2.1), (2.2) imply that

$$\begin{aligned}
&\big(\mathbf{S}(\mathbf{A}) - \mathbf{S}(\mathbf{B})\big) \cdot (\mathbf{A} - \mathbf{B}) \\
&\quad = \sum_{i,j,k,l=1}^{d} \int_0^1 \frac{\partial S_{ij}(\theta\mathbf{A} + (1-\theta)\mathbf{B})}{\partial D_{kl}} (A-B)_{ij}(A-B)_{kl}\, d\theta \\
&\quad \cong |\mathbf{A} - \mathbf{B}|^2 \int_0^1 \big(\kappa + |\theta\mathbf{A} + (1-\theta)\mathbf{B}|\big)^{p-2}\, d\theta \\
&\quad \cong |\mathbf{A} - \mathbf{B}|^2\big(\kappa + |\mathbf{B}| + |\mathbf{A}|\big)^{p-2}.
\end{aligned}$$

In the same manner we get

$$\begin{aligned}\big|\mathbf{S}(\mathbf{A})-\mathbf{S}(\mathbf{B})\big| &= \left(\sum_{i,j=1}^{d}\left|\sum_{k,l=1}^{d}\int_0^1 \frac{\partial S_{ij}(\theta\mathbf{A}+(1-\theta)\mathbf{B})}{\partial D_{kl}}\,d\theta\,(A-B)_{kl}\right|^2\right)^{\frac12}\\ &\le c\,|\mathbf{A}-\mathbf{B}|\big(\kappa+|\mathbf{B}|+|\mathbf{A}|\big)^{p-2}.\end{aligned}$$

This yields the upper estimate of (6.8). The lower estimate follows from (6.7). ∎

LEMMA 6.4. *For all* $\mathbf{A},\mathbf{B}\in\mathbb{R}^{d\times d}$ *there holds*

$$\big|\mathbf{F}(\mathbf{A})-\mathbf{F}(\mathbf{B})\big| \cong \big(\kappa+|\mathbf{B}|+|\mathbf{A}|\big)^{\frac{p-2}{2}}|\mathbf{A}-\mathbf{B}|$$

with constants depending on p *but not on* κ.

Proof. For $q:=\frac{p+2}{2}$ we have

$$\mathbf{F}(\mathbf{B}) = (\kappa+|\mathbf{B}|)^{\frac{p-2}{2}}\mathbf{B} = (\kappa+|\mathbf{B}|)^{q-2}\mathbf{B}\,.$$

Thus $\mathbf{F}$ satisfies (2.1), (2.2) with p replaced by q. Now (6.8) reads as follows:

$$\big|\mathbf{F}(\mathbf{A})-\mathbf{F}(\mathbf{B})\big| \cong \big|\mathbf{A}-\mathbf{B}\big|\big(\kappa+|\mathbf{B}|+|\mathbf{A}|\big)^{q-2}.$$

This proves the lemma. ∎

Now, Lemmas 6.3 and 6.4 immediately imply Lemma 2.1.

REFERENCES

[1] J. W. BARRETT AND R. BERMEJO, *An Improved Error Bound for the Discretization of the Parabolic p-Laplacian and Related Degenerate Quasilinear Equations and Variational Inequalities*, manuscript, Imperial College, London, 2005.

[2] J. W. BARRETT AND W. B. LIU, *Finite element approximation of some degenerate quasilinear elliptic and parabolic problems*, in Numerical Analysis 1993, Pitman Res. Notes Math. Ser. 303, Longman Scientific and Technical, Harlow, UK, 1994, pp. 1–16.

[3] J. W. BARRETT AND W. B. LIU, *Finite element approximation of the parabolic p-Laplacian*, SIAM J. Numer. Anal., 31 (1994), pp. 413–428.

[4] S. S. CHOW, *Finite element error estimates for nonlinear elliptic equations of monitone type*, Numer. Math., 54 (1988), pp. 373–393.

[5] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[6] E. DIBENEDETTO, *Degenerate Parabolic Equations*, Springer-Verlag, New York, 1993.

[7] L. DIENING, A. PROHL, AND M. RŮŽIČKA, *On time-discretizations for generalized Newtonian fluids*, in Nonlinear Problems in Mathematical Physics and Related Topics, II. Int. Math. Ser. (N.Y.) 2, Kluwer Academic Publishers, New York, 2002, pp. 89–118.

[8] L. DIENING AND M. RŮŽIČKA, *Strong solutions for generalized Newtonian fluids*, J. Math. Fluid Mech., 7 (2005), pp. 413–450.

[9] L. DIENING AND M. RŮŽIČKA, *Error Estimates for Interpolation Operators in Orlicz–Sobolev Spaces and Quasi Norms*, manuscript, University Freiburg, Freiburg, Germany, 2006.

[10] C. EBMEYER AND W. B. LIU, *Quasi-norm interpolation error estimates for the piecewise linear finite element approximation of p-Laplacian problems*, Numer. Math., 100 (2005), pp. 233–258.

[11] C. EBMEYER, W. B. LIU, AND M. STEINHAUER, *Global regularity in fractional order Sobolev spaces for the p-Laplace equation on polyhedral domains*, Z. Anal. Anwendungen, 24 (2005), pp. 353–374.

[12] C. EBMEYER, *Regularity in Sobolev spaces of steady flows of fluids with shear-dependent viscosity*, Math. Methods Appl. Sci., 29 (2006), pp. 1687–1707.

[13] M. FARHLOUL, *A mixed finite element method for a nonlinear Dirichlet problem*, IMA J. Numer. Anal., 18 (1998), pp. 121–132.

[14] H. Gajewski, K. Gröger, and K. Zacharias, *Nichtlineare Operatorgleichungen und Operatordifferentialgleichungen*, Akademie-Verlag, Berlin, 1974.
[15] E. Giusti, *Direct Methods in the Calculus of Variations*, World Scientific, River Edge, NJ, 2003.
[16] R. Glowinski and A. Marrocco, *Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité, d'une class problèmes de Dirichlet non linéaires*, RAIRO Anal. Numér., 9 (1975), pp. 41–76.
[17] A. Kufner, O. John, and S. Fučik, *Function Spaces*, Academia, Prague, 1977.
[18] J. L. Lions, *Quelques Méthodes de Résolution des Problèmes aux Limites Non Linéaires*, Dunod, Paris, 1969.
[19] W. Liu and N. Yan, *Quasi-norm local error estimates for p-Laplacian*, SIAM J. Numer. Anal., 39 (2001), pp. 100–127.
[20] A. Prohl and M. Růžička, *On fully implicit space-time discretization for motions of incompressible fluids with shear-dependent viscosities: The case* $p \leq 2$, SIAM J. Numer. Anal., 39 (2001), pp. 214–249.
[21] J. Rulla, *Error analysis for implicit approximations to solutions to Cauchy problems*, SIAM J. Numer. Anal., 33 (1996), pp. 68–87.
[22] M. Růžička, *Modeling, mathematical, and numerical analysis of electrorheological fluids*, Appl. Math., 49 (2004), pp. 565–609.
[23] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[24] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Springer Ser. Comput. Math. 25, Springer-Verlag, Berlin, 1997.
[25] D. Wei, *Existence, uniqueness, and numerical analysis of solutions of a quasilinear parabolic problem*, SIAM J. Numer. Anal., 29 (1992), pp. 484–497.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 473–483

SAMPLING EIGENVALUES IN HARDY SPACES*

AMIN BOUMENIR† AND VU KIM TUAN†

Abstract. In this work we extend the sampling method to compute eigenvalues of singular non–self-adjoint Sturm–Liouville problems in the presence of a continuous spectrum. We first show that the characteristic function, whose zeros are the eigenvalues, belongs to a Hardy space, and then develop a new sampling formula for its reconstruction. We estimate the truncation error, obtain computable error bounds, and test the method with a few numerical experiments.

Key words. Sturm–Liouville operators, sampling, interpolation, Hardy spaces

AMS subject classifications. 65D05, 34L16

DOI. 10.1137/050647335

1. Introduction.

In this work, we would like to develop a new sampling technique to compute eigenvalues in the presence of a continuous spectrum. Consider the following singular Sturm–Liouville problem, which is a widely used model in scattering theory:

$$\begin{cases} Ly(x) := -y''(x,\lambda) + q(x)y(x,\lambda) = \lambda^2 y(x,\lambda), & x \in [0,\infty), \\ y(0,\lambda) = 0, \end{cases} \tag{1.1}$$

where q is complex valued and satisfies $\int_0^\infty (1+x)\,|q(x)|\,dx < \infty$.

Observe that the differential operator L, which is acting in the Hilbert space $L^2_{dx}(0,\,\infty)$, is regular at $x = 0$ but singular at $x = \infty$. It also follows from the asymptotics of the solutions [12, p. 221] that L is in the limit point case at $x = \infty$, and so no boundary condition is required at $x = \infty$. If q is real valued, then L is self-adjoint, the positive part of its spectrum is continuous, and there are at most a finite number of negative isolated eigenvalues which are known as the bound states in scattering theory. The case $\Im(q(x)) \neq 0$ leads to complex eigenvalues, and our work is concerned with their computational aspects. Recall that classical eigenvalue solvers [2, 3] cannot handle the singular non–self-adjoint case due to the presence of a continuous spectrum. For example, when $L - \lambda I$ has a compact inverse and thus no continuous spectrum is present, finite element or Galerkin-type methods can be used. In the case when the inverse operator is bounded but not compact, then a spectral approximation for bounded operators is possible, as shown in [11]. Also, for closed not necessarily bounded operators in Banach spaces, spectral projectors and analytic properties of the resolvent operator are used to locate the spectrum; see [5, 6, 7]. These methods, which usually either are direct or use the inverse operator, detect the spectrum by contour integration. The next type of numerical methods is shooting-type methods, which integrate numerically the solution of (1.1) and then find the values λ for which the boundary condition $y(0,\lambda) = 0$ is satisfied. For instance, if $y(0,\lambda)$ and all the eigenvalues are real, then a zero crossing argument may help detect simple roots. However, when numerical integration is over an infinite interval and

*Received by the editors December 12, 2005; accepted for publication (in revised form) September 12, 2006; published electronically February 26, 2007.
http://www.siam.org/journals/sinum/45-2/64733.html
†Department of Mathematics, University of West Georgia, Carrollton, GA 30118 (boumenir@westga.edu, vu@westga.edu).

there are no boundary conditions at infinity or the roots are off the real line, then the shooting method is extremely hard to implement, if not impossible. To overcome these difficulties, the sampling method reconstructs the function $y(0,\lambda)$ for $\lambda \in \mathbb{C}$ and then finds its roots. For example, in the regular non–self-adjoint case, it interpolates the characteristic function by a rational function and then computes its zeros in the complex plane [1]. A crucial step is the fact that the characteristic function is an entire function and belongs to a Paley–Wiener space, which allows its recovery by the Shannon theorem; see [15]. Unfortunately in (1.1) the presence of a continuous or essential spectrum signals that the characteristic function $y(0,\lambda)$ cannot be an entire function of λ. The main difficulty in any reconstruction problem is uniqueness, which is settled by finding the right space. For example, in a Paley–Wiener space of type π, one must use all values $\{F(n)\}_{n\in\mathbb{Z}}$ in order to recover the function F uniquely. The key to determining the analytic properties of $y(0,\lambda)$ is provided by the Gelfand–Levitan–Marchenko integral representation of the solution [4, p. 77] for $x \geq 0$,

$$y(x,\lambda) = e^{i\lambda x} + \int_x^\infty K(x,t)e^{i\lambda t}dt, \qquad \Im(\lambda) \geq 0. \tag{1.2}$$

We now give a brief outline of the paper. In the second section, we show that if $y(x,\lambda) \in L^2_{dx}(0,\infty)$, then $y(0,\lambda)$, up to a constant, must be a Hardy function. In the third section, we show how to reconstruct or interpolate a function in a Hardy space from a sequence of its values, and in the fourth section, we outline the main steps in the algorithm and convergence results. This leads to the approximation of $y(0,\lambda)$ by rational functions, whose zeros yield the sought eigenvalue approximation of problem (1.1). In the fifth section, we estimate the truncation error and show that it leads to computable error bounds. The last section covers a few examples which illustrate how the method is implemented numerically.

2. The Hardy space. Denote the Hardy space of complex valued functions defined in the right half-plane by

$$\mathcal{H}^2\left(\mathbb{R}^2_+\right) = \Big\{ F(s) : \quad F(\sigma+i\tau) \text{ is analytic } \text{ for } \sigma > 0 \text{ and } \sup_{\sigma>0} \int_{-\infty}^{\infty} |F(\sigma+i\tau)|^2 d\tau < \infty \Big\}.$$

It is also well known that F is a Hardy function if and only if it is the Laplace transform of a function $f(t) \in L^2_{dt}(0,\infty)$:

$$F \in \mathcal{H}^2\left(\mathbb{R}^2_+\right) \quad \Longleftrightarrow \quad F(s) = \mathcal{L}(f)(s) := \int_0^\infty f(t)\, e^{-st}\, dt, \quad \Re(s) > 0. \tag{2.1}$$

In all that follows, $y(x,\lambda)$ is a solution of (1.1) represented by (1.2). Since K is a continuous function, then we have from (1.2)

$$y(0,\lambda) = 1 + \int_0^\infty K(0,t)e^{i\lambda t}dt, \tag{2.2}$$

and thus $y(0,\lambda) - 1$ is a Fourier transform. To say more on the analytic properties of $y(0,\lambda)$ we need the following result.

PROPOSITION 1. *Assume that* $\int_0^\infty (1+x)|q(x)|dx < \infty$*; then* $K(0,t) \in L^1_{dt}(0,\infty) \cap L^2_{dt}(0,\infty)$.

Proof. The proof follows from the estimate [9, p. 16]

$$|K(x,t)| \leq \frac{1}{2}\int_{\frac{x+t}{2}}^{\infty} |q(\eta)|d\eta \; e^{\int_x^\infty \eta|q(\eta)|d\eta} \tag{2.3}$$

and the fact that as $t \to \infty$,

$$\begin{aligned}\int_{\frac{t}{2}}^{\infty} |q(\eta)|d\eta &= \int_{\frac{t}{2}}^{\infty} \frac{(1+\eta)}{(1+\eta)}|q(\eta)|d\eta \\ &\leq \sup_{\frac{t}{2}\leq\eta}\left(\frac{1}{1+\eta}\right)\int_{\frac{t}{2}}^{\infty}(1+\eta)|q(\eta)|d\eta \\ &\leq \frac{2}{2+t}\int_{\frac{t}{2}}^{\infty}(1+\eta)|q(\eta)|d\eta.\end{aligned}$$

In other words, $K(0,t) = O(\frac{1}{2+t})$, and thus $K(0,t) \in L^2_{dt}(0,\infty)$. The Hardy inequality

$$\int_0^\infty \left|\frac{1}{t}\int_t^\infty f(x)\,dx\right| dt \leq \int_0^\infty |f(x)|\,dx$$

also guarantees that $K(0,t) \in L^1_{dt}(0,\infty)$. ∎

PROPOSITION 2. *If* $\int_0^\infty (1+x)\,|q(x)|\,dx < \infty$, *then* $y(0,is) - 1 \in \mathcal{H}^2(\mathbb{R}^2_+)$.

Proof. From (2.2), if we replaced λ by is, then

$$\begin{aligned}y(0,is) - 1 &= \int_0^\infty K(0,t)e^{-st}dt \\ &= \mathcal{L}\left(K(0,t)\right)(s),\end{aligned} \tag{2.4}$$

while by Proposition 1, $K(0,.) \in L^2_{dt}(0,\infty)$. Thus its Laplace transform $y(0,is) - 1$ belongs to the Hardy space $\mathcal{H}^2\left(\mathbb{R}^2_+\right)$; see [8]. ∎

In order to find the eigenvalues of (1.1) by sampling, we first need to reconstruct $y(0,is)$ by interpolation and then solve $y(0,is) = 0$ for $\Re(s) > 0$. As far as the authors are aware, there is no sampling formula in Hardy spaces; see [13]. Therefore we develop a new sampling formula for functions in Hardy spaces in the next section.

3. Sampling in Hardy spaces. Recall the shifted factorial

$$(a)_k = a\,(a+1)\ldots(a+k-1) = \frac{\Gamma\,(a+k)}{\Gamma\,(a)}. \tag{3.1}$$

THEOREM 1. *Let* $F \in \mathcal{H}^2\left(\mathbb{R}^2_+\right)$. *Then*

$$F(s) = \sum_{k=0}^{\infty} \frac{(2k+1)\left(\frac{1}{2}-s\right)_k}{(s+\frac{1}{2})_{k+1}} \sum_{n=0}^{k} \frac{(-k)_n\,(k+1)_n}{(n!)^2} F\left(n+\frac{1}{2}\right), \tag{3.2}$$

where the series converges uniformly on any compact subset of the right half-plane.

Proof. First, if we set $e^{-t} = x$, then (2.1) becomes

$$\begin{aligned}F(s) &= \int_0^\infty e^{-st}f(t)dt \\ &= \int_0^1 g(x)x^{s-1/2}dx, \quad \Re(s) > 0,\end{aligned} \tag{3.3}$$

where $g(x) := f(-\ln x)x^{-1/2}$. It is very easy to see that $F \in \mathcal{H}^2(\mathbb{R}^2_+)$, i.e., $f(t) \in L^2_{dt}(0,\infty)$ if and only if $g(x) \in L^2_{dx}(0,1)$.

Next we need to expand $x^{s-1/2}$ in terms of $x^{n-1/2}$ to bring out the sampled values $F(n)$. Since $\{x^n\}_{n=0}^{\infty}$ is not an orthogonal family in $L^2_{dx}(0,1)$, we use the set $P_n^*(x) = \sqrt{2n+1}\,P_n(1-2x)$, the normalized Legendre orthogonal polynomials on the interval [0, 1]. From their connection with the hypergeometric function

$$P_k(1-2x) = F(-k, k+1; 1, x)$$

we obtain

$$P_k^*(x) = \sqrt{2k+1}\sum_{n=0}^{k}\frac{(-k)_n\,(k+1)_n}{(n!)^2}x^n = \sum_{n=0}^{k} a_{kn}x^n,$$

where

$$a_{kn} := \sqrt{2k+1}\,\frac{(-k)_n\,(k+1)_n}{(n!)^2}.$$

We start with the expansion formula of the power function in a series of Legendre polynomials [14],

$$(a-1)\left(\frac{1-x}{2}\right)^{a-2} = \sum_{k=0}^{\infty}(2k+1)\,\frac{(2-a)_k}{(a)_k}\,P_k(x),$$

where $-1 < x < 1$ and $a > 3/2$. Replacing $\frac{1-x}{2}$ and $a - 5/2$ by x and s, respectively, leads to

$$x^{s-1/2} = \sum_{k=0}^{\infty}\frac{(2k+1)\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}}P_k(1-2x), \qquad 0 < x < 1. \tag{3.4}$$

So

$$x^{s-1/2} = \sum_{k=0}^{\infty} c_k(s)P_k^*(x) = \sum_{k=0}^{\infty} c_k(s)\sum_{n=0}^{k} a_{kn}x^n,$$

where

$$c_k(s) = \frac{\sqrt{2k+1}\,\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}}. \tag{3.5}$$

Since

$$c_k(s) = \int_0^1 x^{s-1/2}P_k^*(x)dx$$

is the Fourier coefficient of $x^{s-1/2} \in L^2_{dx}(0,1)$, $\Re(s) > 0$, and $P_k^*(x) \in L^2_{dx}(0,1)$, we must have $c_k(s) \in \mathcal{H}^2(\mathbb{R}^2_+)$. Now let g be any function from $L^2_{dx}(0,1)$; then

$$\int_0^1 g(x)x^{s-1/2}dx = \int_0^1 g(x)\sum_{k=0}^{\infty} c_k(s)P_k^*(x)\,dx. \tag{3.6}$$

We can interchange the order of integration and infinite summation in the last expression to get

$$
\begin{aligned}
F(s) &= \sum_{k=0}^{\infty} c_k(s) \int_0^1 g(x) P_k^*(x)\, dx \\
&= \sum_{k=0}^{\infty} c_k(s) \int_0^1 g(x) \sum_{n=0}^{k} a_{kn} x^n\, dx \\
&= \sum_{k=0}^{\infty} c_k(s) \sum_{n=0}^{k} a_{kn} \int_0^1 g(x) x^n\, dx \\
&= \sum_{k=0}^{\infty} c_k(s) \sum_{n=0}^{k} a_{kn} F\left(n+\frac{1}{2}\right).
\end{aligned} \tag{3.7}
$$

Thus we can interpolate a function $F \in \mathcal{H}^2\left(\mathbb{R}_+^2\right)$ from a sequence of its values $\{F(n+1/2)\}_{n\geq 0}$.

To see that the series converges uniformly in any compact of $\Re(s) > 0$, let $s \in \Omega$, where Ω is compact, and use (3.7) to write

$$
\begin{aligned}
\left| F(s) - \sum_{k=0}^{N} c_k(s) \int_0^1 g(x) P_k^*(x)\, dx \right| &= \left| \int_0^1 g(x) \sum_{k=N+1}^{\infty} c_k(s) P_k^*(x)\, dx \right| \\
&\leq \|g\|_2 \left\| \sum_{k=N+1}^{\infty} c_k(s) P_k^*(x) \right\|_2 \\
&= \|g\|_2 \sqrt{\sum_{k=N+1}^{\infty} |c_k(s)|^2}, \qquad (3.8)
\end{aligned}
$$

since $P_k^*(x)$ form an orthonormal system in $L^2_{dx}(0,1)$. All that we need to show uniform convergence is

$$
\sup_{s\in\Omega} \sum_{k=N+1}^{\infty} |c_k(s)|^2 \to 0 \quad \text{as} \quad N \to \infty. \tag{3.9}
$$

To estimate the remainder we first express the c_k through the gamma function

$$
\begin{aligned}
c_k(s) &= \frac{\sqrt{2k+1}\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}} \\
&= \sqrt{2k+1} \frac{\Gamma\left(\frac{1}{2}-s+k\right)}{\Gamma\left(s+\frac{1}{2}+k+1\right)} \frac{\Gamma\left(s+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}-s\right)} \qquad (3.10)
\end{aligned}
$$

and use its asymptotics

$$
\frac{\Gamma(k+a)}{\Gamma(k+b)} \approx k^{a-b} \quad \text{as } k \to \infty.
$$

We next use the fact that we have

$$
\sup_{s\in\Omega} \left| \frac{\Gamma\left(s+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}-s\right)} \right| < \infty
$$

to deduce from (3.10) that

$$|c_k(s)| = O\left(k^{\frac{1}{2}} k^{\frac{1}{2}-\Re(s)-\Re(s)-\frac{3}{2}}\right),$$

$$|c_k(s)|^2 = O\left(k^{-1-4\Re(s)}\right) \tag{3.11}$$

uniformly in Ω, which implies (3.9), and the proof is complete. □

Combining Proposition 2 and Theorem 1, we can write

$$y(0,is) = 1 + \sum_{k=0}^{\infty} c_k(s) \sum_{n=0}^{k} a_{kn} \left[y\left(0, in + \frac{i}{2}\right) - 1 \right]. \tag{3.12}$$

4. Algorithm. We now describe the algorithm that would allow us to use (3.12). Recall that for $-\lambda^2 \in \mathbb{C}$ to be an eigenvalue, we need

$$y(x,i\lambda) \in L^2_{dx}(0,\infty) \quad \text{and} \quad y(0,i\lambda) = 0.$$

Use (1.2) and (2.3) to see that for each fixed λ, with $\Im(\lambda) > 0$, there are two possible solutions whose asymptotic behavior is either $e^{i\lambda x}$ or $e^{-i\lambda x}$, the so-called Jost solutions

$$\phi_{\pm}(x,\lambda) = e^{\pm i\lambda x} + o\,(1) \quad \text{as } x \to \infty, \tag{4.1}$$

and any solution of (1.1) is their combination. Obviously only $\phi_+(x,i\lambda) \in L^2_{dx}(0,\infty)$, and so

$$y(x,i\lambda) = e^{-\lambda x} + \int_x^{\infty} K(x,t) e^{-\lambda t} dt, \qquad \Re(\lambda) > 0. \tag{4.2}$$

Thus we need to solve the following boundary value problem at $x = \infty$, namely,

$$\begin{cases} -y''(x,i\lambda) + q(x) y(x,i\lambda) = -\lambda^2 y(x,i\lambda), \\ \lim\limits_{x\to\infty} y(x,i\lambda) e^{x\lambda} = 1, \end{cases} \tag{4.3}$$

to ensure that the solution is $L^2_{dx}(0,\infty)$. From the computational point of view, the best we can hope for is to replace integration over $(0,\infty)$ by $(0,L)$, where L is large enough. The asymptotic behavior of the solution is used to bound the difference. To this end, integration by parts in (4.2) and the fact $2K(x,x) = \int_x^{\infty} q(\eta)d\eta$ [4] reduces (4.2) to

$$y(x,i\lambda) = e^{-\lambda x} + \frac{e^{-\lambda x}}{2\lambda} \int_x^{\infty} q(\eta)d\eta + \int_x^{\infty} K_t(x,t) \frac{e^{-\lambda t}}{\lambda} dt, \qquad \Re(\lambda) > 0. \tag{4.4}$$

Thus since the solution of (4.3) decays rapidly, we can start integration from $x = L < \infty$, instead of $x = \infty$, and from (4.4) we deduce the initial value problem

$$\begin{cases} -y_L''(x,i\lambda\,) + q(x) y_L(x,i\lambda\,) = -\lambda^2 y_L(x,i\lambda\,), & x \in [0,L], \\ y_L(L,i\lambda) = e^{-\lambda L} + \frac{e^{-\lambda L}}{2\lambda} \int_L^{\infty} q(\eta)d\eta, \\ y_L'(L,i\lambda) = -\lambda e^{-\lambda L} - \frac{e^{-\lambda L}}{2} \int_L^{\infty} q(\eta)d\eta. \end{cases} \tag{4.5}$$

Equation (4.5) allows us to compute the values $y_L(0,i\lambda\,)$ for $\lambda = \frac{1}{2},\ \frac{3}{2},\ \frac{5}{2}, \ldots, n + \frac{1}{2}, \ldots$, which are needed for sampling by formula (3.12).

It is easy to see that the error function

$$\theta(x,i\lambda) = y(x,i\lambda) - y_L(x,i\lambda)$$

satisfies, by linearity,

$$\text{(4.6)}\qquad \begin{cases} -\theta''(x,i\lambda) + q(x)\theta(x,i\lambda\ ) = -\lambda^2\theta(x,i\lambda\ ), & x\in[0,L],\\ \theta(L,i\lambda) = \frac{1}{\lambda}\int_L^\infty K_t(L,t)e^{-\lambda t}dt \quad \text{and} \quad \theta'(L,i\lambda) = \int_L^\infty K_x(L,t)e^{-\lambda t}dt \end{cases}$$

or the equivalent Volterra integral equation, with $0\le x\le L$,

$$\theta(x,i\lambda) = \psi(x,i\lambda) + \int_L^x \frac{\sinh\lambda(x-t)}{\lambda}q(t)\theta(t,i\lambda)dt,$$

where

$$\text{(4.7)}\qquad \psi(x,i\lambda) = \theta(L,i\lambda)\cosh\left(\lambda(x-L)\right) + \theta'(L,i\lambda)\frac{1}{\lambda}\sinh\left(\lambda(x-L)\right).$$

Thus from (4.7) it follows that

$$|\psi(x,i\lambda)| \le \Xi(\lambda)e^{\Re(\lambda)(L-x)},$$

where $\Re(\lambda) > 0$ and

$$\Xi(\lambda) = |\theta(L,i\lambda)| + \frac{1}{|\lambda|}\left|\theta'(L,i\lambda)\right|.$$

Now define the successive iterations by

$$\begin{aligned}\theta_0(x,i\lambda) &= \psi(x,i\lambda),\\ \theta_n(x,i\lambda) &= \int_L^x \frac{\sinh\lambda(x-t)}{\lambda}q(t)\theta_{n-1}(t,i\lambda)dt,\end{aligned}$$

from which it follows that, for $0\le x\le L$,

$$\text{(4.8)}\qquad |\theta_n(x,i\lambda)| \le \frac{1}{n!}\left(\frac{1}{|\lambda|}\int_x^L |q(t)|\,dt\right)^n \Xi(\lambda)e^{\Re(\lambda)(L-x)}.$$

Indeed, (4.8) is true for $n=0$, and if it holds true for $n-1$, then

$$\begin{aligned}|\theta_n(x,i\lambda)| &= \left|\int_L^x \frac{\sinh\left(\lambda(x-t)\right)}{\lambda}q(t)\theta_{n-1}(t,i\lambda)dt\right|\\ &\le \int_x^L e^{\Re(\lambda)(t-x)}\left|\frac{q(t)}{\lambda}\right||\theta_{n-1}(t,i\lambda)|\,dt\\ &\le \Xi(\lambda)\int_x^L e^{\Re(\lambda)(t-x)}e^{\Re(\lambda)(L-t)}\frac{1}{(n-1)!}\left(\frac{1}{|\lambda|}\int_t^L |q(\eta)|\,d\eta\right)^{n-1}\left|\frac{q(t)}{\lambda}\right|dt\\ &\le \frac{1}{n!}\left(\frac{1}{|\lambda|}\int_x^L |q(\eta)|\,d\eta\right)^n \Xi(\lambda)e^{\Re(\lambda)(L-x)}.\end{aligned}$$

Thus for $\Re(\lambda) > 0$ the series $\theta(x,i\lambda) = \sum_{n\ge0}\theta_n(x,i\lambda)$ converges, and we have

$$\text{(4.9)}\qquad |\theta(x,i\lambda)| \le e^{\frac{1}{|\lambda|}\int_x^L |q(\eta)|d\eta}\Xi(\lambda)e^{\Re(\lambda)(L-x)}.$$

We now have the following lemma.

LEMMA 1. *For any $x \geq 0$ and $\Re(\lambda) > 0$ we have $\Xi(\lambda)e^{\Re(\lambda)(L-x)} \to 0$ as $L \to \infty$.*

Proof. Recall that when q satisfies $\int_0^\infty (1+x)\,|q(x)|\,dx < \infty$, then [10, Lemma 3.1.2, p. 178]

$$|K_t(x,t)| \leq \frac{1}{4}\left|q\left(\frac{x+t}{2}\right)\right| + Ce^{\int_0^\infty \zeta|q(\zeta)|d\zeta}\int_x^\infty |q(\zeta)|d\zeta \int_{\frac{x+t}{2}}^\infty |q(\zeta)|d\zeta, \tag{4.10}$$

from which it follows that K_t is integrable at ∞. Thus for $x < L$ we have

$$\begin{aligned}\left|\theta(L,i\lambda)e^{\Re(\lambda)(L-x)}\right| &\leq \frac{1}{|\lambda|}\int_L^\infty |K_t(L,t)|\,e^{-\Re(\lambda)t}dte^{\Re(\lambda)(L-x)}\\ &\leq \frac{1}{|\lambda|}e^{\Re(\lambda)x}\int_L^\infty |K_t(L,t)|\,dt,\end{aligned}$$

which yields that $|\theta(L,i\lambda)e^{\Re(\lambda)(L-x)}| \to 0$ as $L \to \infty$. Since K_x satisfies the same estimates in (4.10) as K_t, a similar argument leads to $\theta'(L,i\lambda)\frac{1}{\lambda}e^{\Re(\lambda)(L-x)} \to 0$ as $L \to \infty$. ∎

Combining the lemma and (4.9), we obtain the convergence result.

PROPOSITION 3. *We have, for $\Re(\lambda) > 0$, $y_L(0,i\lambda) \to y(0,i\lambda)$ as $L \to \infty$.*

Thus we can approximate $y(0,i\lambda)$ by using the values $y_L(0,is)$ for $s = \frac{1}{2},\ \frac{3}{2},\ \frac{5}{2},\ldots,$ in the sampling formula (3.12).

5. Truncation error. In this section, we show that it takes only a few values, N say, to recover a good approximation of $y(0,is)-1$. The truncation error has already been estimated in (3.8), and combining it with (3.11) leads to

$$\begin{aligned}\sqrt{\sum_{k=N+1}^\infty |c_k(s)|^2} &= O\left(\sqrt{\sum_{k=N+1}^\infty k^{-1-4\Re(s)}}\right)\\ &= O\left(N^{-2\Re(s)}\right).\end{aligned}$$

We recall that by (2.4) and (3.3) we can recast (3.8) as

$$\begin{aligned}&\left|y(0,is) - 1 - \sum_{k=0}^N \frac{(2k+1)\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}}\sum_{n=0}^k \frac{(-k)_n\,(k+1)_n}{n!n!}\left(y\left(0,i\left(n+\frac{1}{2}\right)\right)-1\right)\right|\\ &\quad\leq \sqrt{\int_0^1 \left|\frac{1}{\sqrt{x}}K\left(0,-\ln(x)\right)\right|^2 dx}\sqrt{\sum_{k=N+1}^\infty |c_k(s)|^2}\\ &\quad\leq \sqrt{\int_0^\infty |K\left(0,t\right)|^2\,dt}\sqrt{\sum_{k=N+1}^\infty |c_k(s)|^2},\end{aligned}$$

which is valid for $\Re(s) > 0$. Since, by Proposition 1, $K(0,t) \in L^2_{dt}(0,\infty)$ if $\int_0^\infty (1+x)\cdot|q(x)|\,dx < \infty$, we have the following proposition.

PROPOSITION 4. *Assume that* $\int_0^\infty (1+x)\,|q(x)|\,dx < \infty$*; then the truncation error for* $\Re(s) > 0$ *is given by*

$$\left| y(0,is) - 1 - \sum_{k=0}^{N} \frac{(2k+1)\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}} \sum_{n=0}^{k} \frac{(-k)_n\,(k+1)_n}{n!n!} \left(y\left(0, i\left(n+\frac{1}{2}\right)\right) - 1 \right) \right| = O\left(\frac{1}{N^{-2\Re(s)}}\right).$$

Thus we can recover the boundary function $y(0,is)$ by using the $N+1$ values $\left\{y(0, i\left(n+\frac{1}{2}\right)\right\}_{n=0}^{N}$ or their approximations $y_L\left(0, i\left(n+\frac{1}{2}\right)\right)$ for $n = 0, \ldots, N$. Using the previous notation, an error on the sampled values

$$y\left(0, i\left(n+\frac{1}{2}\right)\right) = y_L\left(0, i\left(n+\frac{1}{2}\right)\right) + \theta\left(0, i\left(n+\frac{1}{2}\right)\right)$$

leads to an error $\varepsilon_N(s)$ on the recovered function,

$$y(0,is) = y_\varepsilon(0,is) + \varepsilon_N(s),$$

where

$$y_\varepsilon(0,is) = 1 + \sum_{k=0}^{N} \frac{(2k+1)\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}} \sum_{n=0}^{k} \frac{(-k)_n\,(k+1)_n}{n!n!} \left(y_L(0, i\left(n+\frac{1}{2}\right) - 1 \right).$$

Both truncation errors are included in $\varepsilon_N(s)$, which can be estimated for $\Re(s) > 0$ by

$$|\varepsilon_N(s)| \le \left| \sum_{k=0}^{N} \frac{(2k+1)\left(\frac{1}{2}-s\right)_k}{\left(s+\frac{1}{2}\right)_{k+1}} \sum_{n=0}^{k} \frac{(-k)_n\,(k+1)_n}{n!n!} \theta_n \right| + \sqrt{\int_0^\infty |K(0,t)|^2\,dt} \sqrt{\sum_{k=N+1}^{\infty} |c_k(s)|^2}.$$

Note that $\int_0^\infty |K(0,t)|^2\,dt$ can be estimated in terms of q only by (2.3) and yields a computable error bound.

6. Examples. We use simple examples where the exact values are available, so that a comparison with our numerical results is possible. In Examples 1 and 3 below, we sample with $N = 10$ and $L = 100$, while $N = 5$ only in Example 2.

Example 1. Consider the singular Sturm–Liouville problem

$$\begin{cases} -y''(x,\lambda) - 16H(3-x)y(x,\lambda) = \lambda^2 y(x,\lambda), & 0 \le x, \\ y(0,\lambda) = 0. \end{cases}$$

Its exact solution is

$$y(0,i\lambda) = e^{-3\lambda}\lambda \frac{\sin\left(3\sqrt{16-\lambda^2}\right)}{\sqrt{16-\lambda^2}} + e^{-3\lambda}\cos\left(3\sqrt{16-\lambda^2}\right).$$

The eigenvalues λ^2 obtained by sampling compare well with the "exact" ones:

	Exact	Sampling
λ_1	−1.7341473181761	−1.74403065715701
λ_2	−7.7381824457812	−7.73145171424856
λ_3	−12.287216856980	−12.2871029333224
λ_4	−15.067032974543	−15.0694822812548

Example 2. Consider the singular Sturm–Liouville problem

$$\begin{cases} -y''(x,\lambda) - xH(4-x)y(x,\lambda) = \lambda^2 y(x,\lambda), & 0 \le x, \\ y(0,\lambda) = 0, \end{cases}$$

whose boundary $y(0,\lambda)$ is in terms of the Airy functions

$$\begin{aligned} &y(0,\lambda) \\ &= \frac{\exp(-4\lambda)\left(AiryBi(-4+\lambda^2)\lambda - AiryBi(1,-4+\lambda^2)\right) AiryAi(\lambda^2)}{(AiryBi(-4+\lambda^2)AiryAi(1,-4+\lambda^2) - AiryBi(1,-4+\lambda^2)AiryAi(-4+\lambda^2))} \\ &\quad - \frac{\exp(-4\lambda)(\lambda AiryAi(-4+\lambda^2) - AiryAi(1,-4+\lambda^2))AiryBi(\lambda^2)}{(AiryBi(-4+\lambda^2)AiryAi(1,-4+\lambda^2) - AiryBi(1,-4+\lambda^2)AiryAi(-4.+\lambda^2))}. \end{aligned}$$

The operator has two negative eigenvalues, and a sampling at five points gives the following values:

	Exact	Sampling
λ_1^2	$-0.408\,556\,101$	$-0.407\,847\,606$
λ_2^2	$-2.199\,310\,808$	$-2.199\,274\,885$

Example 3. Here we consider a complex potential, i.e., the non–self-adjoint case. For simplicity we take $q(x) = (-3+4I)\,H(2-x)$, which satisfies the integral condition, since its support is finite. The "exact" eigenvalues are

$$\begin{aligned} \lambda_1^2 &= -2.366244570422974900 + 2.77959479317720 * I, \\ \lambda_2^2 &= 2.331242100943824884 + 1.444609392874578 * I. \end{aligned}$$

Sampling returns the characteristic function $y(0,\lambda)$ whose roots are

$$\begin{aligned} &-24.478660334593254119478447594810 + 1.120663932982268597901489419162 9 * I, \\ &-11.880762137293417570328509302643 - 12.256428855012470888743443305746 * I, \\ &-10.773223452406265088447046330037 + 11.183480685054562021984436680059 * I, \\ &-3.851978749364907686196649687849 6 + 6.820442155522640744736778476141 2 * I, \\ &-3.393876162198283248485872314760 6 - 7.765724107521692527238126019009 5 * I, \\ &-1.892146494123336769367207540188 5 + 3.380811623590107681266342100745 4 * I, \\ &-1.163475053230193440020205285847 4 + .934728826662452880407214691265 73 * I, \\ &-0.963619216785758091591199908854 40 - 3.945514752416201480449973014290 1 * I, \\ &-0.407624028851342922839455836953 23 - 0.000025337806145023868811196176 * I, \\ &0.417362878943329935909122614903 37 - 1.591169234219253978484738460266 3 * I, \\ &1.734562222252534111936564894393 6 - 0.801293335470853715211071161385 91 * I. \end{aligned}$$

Discarding the first nine, since their $\Re(\lambda) < 0$, and keeping the last two yields

$$\begin{aligned} \lambda_1^2 &= -2.366635093395643431 + 2.779786297300938674 * I, \\ \lambda_2^2 &= 2.3576277592060224472 + 1.328189944959602990 * I. \end{aligned}$$

Future work will deal with making the code more automatic in its selection of roots, so as to keep only the ones with a positive real part. We shall also investigate the case of imbedded eigenvalues in the continuous spectrum, which would have interesting applications in quantum mechanics and in scattering theory.

Acknowledgment. The authors sincerely thank the referee for his constructive comments.

REFERENCES

[1] A. Boumenir, *Sampling and eigenvalues of non-self-adjoint Sturm–Liouville problems*, SIAM J. Sci. Comput., 23 (2001), pp. 219–229.
[2] A. Boumenir and B. Chanane, *The computation of negative eigenvalues of singular Sturm-Liouville problems*, IMA J. Numer. Anal., 21 (2001), pp. 489–501.
[3] J. H. Bramble and J. E. Osborn, *Rate of convergence estimates for nonselfadjoint eigenvalue approximations*, Math. Comp., 27 (1973), pp. 525–549.
[4] K. Chadan and P. C. Sabatier, *Inverse Problems in Quantum Scattering Theory*, Springer-Verlag, New York, Berlin, 1989.
[5] F. Chatelin, *Spectral Approximation of Linear Operators*, Comput. Sci. Appl. Math., Academic Press, New York, 1983.
[6] J. Descloux, *Error bounds for an isolated eigenvalue obtained by the Galerkin method*, Z. Angew. Math. Phys., 30 (1979), pp. 167–176.
[7] J. Descloux, M. Luskin, and J. Rappaz, *Approximation of the spectrum of closed operators: The determination of normal modes of a rotating basin*, Math. Comp., 36 (1981), pp. 137–154.
[8] V. Ditkine and A. Proudnikov, *Transformations Integrales et Calcul Operationnel*, MIR Publisher, Moscow, 1978 (French translation of the Russian original).
[9] B. M. Levitan, *Inverse Sturm-Liouville Problems*, VNU Science Press, Utrecht, The Netherlands, 1987.
[10] V. Marchenko, *Sturm-Liouville Operators and Applications*, Oper. Theory Adv. Appl. 22, Birkhäuser Boston, Cambridge, MA, 1986.
[11] W. H. Mills, Jr., *The resolvent stability condition for spectra convergence with application to the finite element approximation of noncompact operators*, SIAM J. Numer. Anal., 16 (1979), pp. 695–703.
[12] M. A. Naimark, *Linear Differential Operators, Part* II, Ungar, New York, 1969.
[13] K. Seip, *Interpolation and Sampling in Sampling Spaces of Analytic Functions*, Univ. Lect. Series 33, AMS, Providence, RI, 2004.
[14] P. K. Suetin, *Classical Orthogonal Polynomials*, Nauka, Moscow, 1979.
[15] A. Zayed, *Advances in Shannon's Sampling Theory*, CRC Press, Boca Raton, FL, 1993.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 484–499

MAXIMUM L^2-CONVERGENCE RATES OF THE CRANK–NICOLSON SCHEME TO THE STOKES INITIAL VALUE PROBLEM*

JUERGEN RODENKIRCHEN†

Abstract. Let A denote the Stokes operator and D_{A^α} the domain of its fractional powers A^α. We consider the homogeneous Stokes initial value problem with initial data $u(0) = u_0 \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$. For Stokes-like equations the range $\varepsilon \in (0, \frac{1}{4})$ is of special interest, as any solution derived from $\varepsilon \geq \frac{1}{4}$ would necessarily have to satisfy an additional, in practice unverifiable compatibility condition at time $t = 0$. Approximating any strong solution $u \in C^0([0,\infty), D_{A^{1+\varepsilon}})$ in time direction on a finite time interval $[0,T]$ with a Crank–Nicolson scheme, we show convergence of order $O(\frac{\tau^2}{t^{1-\varepsilon}})$ which is maximal for the assumed regularity of the data and reflects the loss of regularity as $t \to 0$. The error estimates are derived by energy and semigroup methods combined with a parabolic duality argument.

Key words. Stokes equation, approximation, Crank–Nicolson scheme, convergence, compatibility condition

AMS subject classifications. 35Q30, 76D07, 65J10

DOI. 10.1137/060653834

1. Introduction.

We consider a viscous incompressible flow at time $t \geq 0$ in a bounded open domain $\Omega \subset \mathbb{R}^3$, $\partial\Omega$ being sufficiently smooth (e.g., a compact 2-dimensional C^3-submanifold of $\mathbb{R}^3$; see [23, p. 1082]), described by the Stokes initial and boundary value problem

$$\left.\begin{array}{l} \frac{\partial}{\partial t}u - \Delta u + \nabla p = 0 \ \text{ in } \Omega,\ t > 0, \\ \nabla \cdot u = 0\,, \quad u(\cdot,x)_{|_{\partial\Omega}} = 0, \quad u(0,\cdot) = u_0. \end{array}\right\} \tag{1.1}$$

Here we assume without loss of generality the constant kinematic viscosity $\nu > 0$ to equal 1. The unknown $u = (u_1, u_2, u_3)$, $u_i = u_i(t,x)$ and $p = p(t,x) \geq 0$ denote the velocity field and the scalar kinematic pressure, respectively. Let $H(\Omega)$ be the space of solenoidal L^2-vector fields on Ω (see section 2). Applying Weyl's orthogonal projection $P : L^2(\Omega) \to H(\Omega)$ to (1.1), which in virtue of the Helmholtz decomposition of $L^2(\Omega)$ maps into zero the gradients $\nabla q \in L^2(\Omega)$, we obtain the homogeneous[1] Stokes initial value problem in the form of the evolution system

$$\left.\begin{array}{l} \partial_t u + Au = 0,\ t > 0, \\ u(0) = u_0 \end{array}\right\} \tag{1.2}$$

for the unknown function $u : [0,\infty) \to D_A$. Here ∂_t means the time derivative and $A = -P\Delta$ denotes the Stokes operator on Ω (see section 2).

Let u be a strong solution of (1.2) on a finite time interval $[0,T]$, i.e., a function u differentiable a.e. on $[0,T]$ such that $\partial_t u \in L^1(0,T;H)$, $u(0) = u_0$ and $\partial_t u = -Au(t)$

*Received by the editors March 7, 2006; accepted for publication (in revised form) September 21, 2006; published electronically February 26, 2007.
http://www.siam.org/journals/sinum/45-2/65383.html
†Hella KgaA, Rixbecker Strasse 75, 59552 Lippstadt, Germany (juergen.rodenkirchen@t-online.de, juergen.rodenkirchen@hella.com).
[1]The inhomogeneous case with outer force density f being a gradient field $f = \nabla q \in L^2(\Omega)$ is thus included in the considered homogeneous case.

a.e. on $[0,T]$ (see [18, Definition 2.8, p. 109], [14, Definition 6.1.2(i), p. 174]). Then, approximating u on an equidistant time grid of $[0,T]$ (with fixed but arbitrary time step of size $\tau = \frac{T}{N} > 0$, $N \in \mathbb{N}$) by a modified Crank–Nicolson sequence $(U_k^\tau)_{k=1,\dots,N}$, we give explicit τ-dependent estimates of the error $u(t_k) - U_k^\tau$ at time $t = t_k = k \cdot \tau$ in $L^2(\Omega)$. Here the modification is due to two initial steps of Rothe's scheme (see section 3).

The convergence rate of the formally second order Crank–Nicolson scheme is known to depend on the regularity properties of the solution to be approximated (see [2], [8], [11], [17], [31], [32, pp. 110–120]). To obtain full second order convergence in $L^2(\Omega)$, the solution u of (1.2) has to be at least strong H^4-continuous [31], [32]. However, for Stokes-like equations such regularity assumptions cannot realistically be assumed, as any H^3-continuous solution already has to satisfy an additional compatibility condition at time $t = 0$, which turns out to be virtually uncheckable for given data (see [29], [15, p. 91], [27, p. 97], [21, p. 134], [10, p. 281], [25, p. 254]).

The question of how smooth a solution can be in practice has been answered by Rautmann [21, Theorem 4.1, p. 147]: Any solution u of (1.2) which is strongly D_{A^η}-continuous has to satisfy a compatibility condition if $\eta > \frac{5}{4}$. And, conversely (see [21, Theorem 3.1, p. 143]), for $\eta < \frac{5}{4}$ assume $u_0 \in D_{A^\eta}$. Then (1.2) yields a strong D_{A^η}-continuous solution without any nonrealistic compatibility condition to be satisfied at time $t = 0$.

For realistically assumable initial data $u_0 \in D_A$, Heywood and Rannacher proved optimal convergence of order $O(\frac{\tau^2}{t})$ in L^2 [11]. Thus, our aim was to provide error estimates in L^2 of order $O(\frac{\tau^2}{t^{1-\varepsilon}})$ for initial data $u_0 \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$, with special emphasis on the realistic range $\varepsilon \in (0, \frac{1}{4})$.[2]

2. Preliminaries. Let $H^0(\Omega) = L^2(\Omega)$ and $H^m(\Omega)$, $m \in \mathbb{N}$, denote the Hilbert spaces

$$H^m(\Omega) = \{u \in L^2(\Omega) \mid \partial_x^\alpha u \in L^2(\Omega),\ |\alpha| \le m,\ m \in \mathbb{N}\}$$

equipped with the norms

$$\|u\|_{H^m(\Omega)} = \left(\sum_{|\alpha| \le m} \int_\Omega |\partial_x^\alpha u(x)|^2 dx \right)^{\frac{1}{2}}, \quad \partial_x^\alpha u(x) = \frac{\partial^{|\alpha|}}{\partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2} \partial_{x_3}^{\alpha_3}} u(x),$$

where $\alpha = (\alpha_1, \alpha_2, \alpha_3)$, $\alpha_i \ge 0$, $|\alpha| = \alpha_1 + \alpha_2 + \alpha_3$, and $|\cdot|$ is the Euclidean norm in $\mathbb{R}^3$ [1]. For abbreviation in the following we let $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ denote norm and inner product in $L^2(\Omega)$:

$$\langle f, g \rangle = \int_\Omega f(x) \cdot g(x)\, dx, \quad \|f\| = \langle f, f \rangle^{\frac{1}{2}}.$$

Let $D(\Omega) = C^\infty_{0,\sigma}(\Omega)$ be the set of all real divergence-free C^∞ vector functions having compact support in Ω. Then $H(\Omega)$ and $V(\Omega)$ denote the closure of $D(\Omega)$ in $L^2(\Omega)$ and $H^1(\Omega)$, respectively:

$$H(\Omega) = \overline{D(\Omega)}^{\|\cdot\|}, \quad V(\Omega) = \overline{D(\Omega)}^{\|\cdot\|_{H^1(\Omega)}}.$$

[2] Corresponding maximal convergence rates can be derived in higher order spaces including H^2 and $D_{A^{1+\varepsilon}}$ up to the order $\varepsilon < \frac{1}{4}$. This will be studied separately.

In virtue of the Helmholtz decomposition

$$L^2 = H \oplus G, G = \left\{ v \in L^2 \mid \exists\, q \in H^1 \ : \ v = \nabla q \right\} \tag{2.1}$$

of L^2 (see [6], [7], [25, pp. 81–89]), let P be Weyl's orthogonal projection $P : L^2 \to H$ [34]. Then the Stokes operator A is defined as the closure in H of the operator $-P\Delta$, which is positive definite and symmetric on the dense subset $D \subset H$, hence positive and self-adjoint; its domain is $D_A = H^2 \cap V$ (see [5, pp. 270, 275–276], [15, pp. 44–45], [3], [26], [25, pp. 127–132]). For the Stokes resolvent $(A+\lambda)^{-1} : H \to D_A$ there holds the following.

LEMMA 2.1. *The resolvent set of the Stokes operator A contains the origin and there exist positive constants c and λ_0 such that*

$$\|(A+\lambda)^{-1}\| \leq \frac{c}{1+|\lambda|} \quad \forall \lambda \in \mathbb{C} \ : \ \ Re(\lambda) \geq -\lambda_0 \tag{2.2}$$

See [33, p. 74], which leads to the following.

LEMMA 2.2. *$-A$ generates the analytic strictly contractive semigroup $\{e^{-tA};\ t \geq 0\}$ on H, $e^{-tA} : H \to H$ is uniformly bounded for $t \geq 0$, $e^{-tA} H \subset D_A$ for $t > 0$, and*

$$\frac{d}{dt} e^{-tA} u + A e^{-tA} u = 0 \quad \forall\ u \in H. \tag{2.3}$$

See [5, pp. 279–280], [4, pp. 101–108], [12], [13], [19], [35], [25, p. 203].

Let $\alpha \in \mathbb{R}$, $\alpha > 0$. Then the fractional powers $A^{-\alpha}$ exist as bounded operators by means of the spectral representation

$$A^{-\alpha} = \frac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-tA} dt, \tag{2.4}$$

where Γ denotes Euler's Gamma function (see [28, p. 10 and Theorem 2.3.2, p. 44]). These operators are invertible, and thus $A^\alpha = (A^{-\alpha})^{-1}$ define the fractional powers of A for positive exponents (see [25, pp. 134, 203]). In addition, let D_{A^α} be the domain of A^α. Then $D_{A^\alpha} \subset H$. For $\alpha > \beta$, the imbedding $D_{A^\alpha} \hookrightarrow D_{A^\beta}$ is compact and D_{A^α} is dense in D_{A^β} (see [30, p. 98], [4, p. 158]).

For Lemmas 2.3 and 2.4 see [5, Lemma 2.10, p. 280] and [22, Lemma 1.3], respectively.

LEMMA 2.3. *Let $u \in D_{A^\beta}$ and $0 < \alpha \leq e$. Then*

$$\|A^{\alpha+\beta} e^{-tA} u\| \leq t^{-\alpha} \cdot \|A^\beta u\| \quad \forall t > 0. \tag{2.5}$$

LEMMA 2.4. *For each $f \in H$ and each real $\tau > 0$ let*

$$f^* = (1+\tau A)^{-1} f \tag{2.6}$$

denote the Yosida approximation of f. Then $f^* \in D_A$ and

$$\|A^\alpha f^*\| \leq c \cdot \tau^{\beta-\alpha} \cdot \|A^\beta f\| \tag{2.7}$$

for each $f \in D_{A^\beta}$ with $0 \leq \beta \leq \alpha \leq 1$.

Additionally, we will make frequent use of the *Cauchy–Young inequality*: Let $a, b \in \mathbb{R}$ with $a, b > 0$. Then

$$a \cdot b \leq \eta \cdot a^q + c_\eta \cdot b^{q'} \tag{2.8}$$

with $\eta = \frac{r^q}{q}$, $c_\eta = \frac{r^{-q'}}{q'}$, $r > 0$, and $\frac{1}{q} + \frac{1}{q'} = 1$, $q > 1$.

3. A modified Crank–Nicolson scheme. Let $\varepsilon \in (0,1)$, $u_0 \in D_{A^{1+\varepsilon}}$, and consider a unique strong solution $u \in C^0([0,\infty), D_{A^{1+\varepsilon}})$ of (1.2), which exists in virtue of Lemma 2.2 due to the representation $u(t) = e^{-tA}u_0$. Here $A^{1+\varepsilon}e^{-tA}u_0 = e^{-tA}A^{1+\varepsilon}u_0 \in H$ for all $t > 0$ because the fractional powers $A^{1+\varepsilon}$ commute with e^{-tA} on $D_{A^{1+\varepsilon}}$ (see [23, p. 1087]). Consider a finite time interval $J = [0,T]$, $T > 0$, and let

$$\tau = \frac{T}{N} > 0, \quad t_k = k \cdot \tau, \quad k = 0, 1, \dots, N, \quad N \in \mathbb{N},$$

be an equidistant time grid on J. On J we approximate u by a Crank–Nicolson sequence $(U_k^\tau)_{k=1,\dots,N}$ with step size $\tau > 0$:

$$\frac{U_k^\tau - U_{k-1}^\tau}{\tau} + \frac{1}{2} \cdot A(U_k^\tau + U_{k-1}^\tau) = 0, \quad k = 1, \dots, N, \tag{3.1}$$

where $U_0^\tau = u_0$ is the given initial value. To obtain optimum convergence we modify the Crank–Nicolson scheme (3.1) with two steps of the locally second order Rothe scheme:

$$\frac{U_k^\tau - U_{k-1}^\tau}{\tau} + AU_k^\tau = 0, \quad k = 1, \dots, N, \tag{3.2}$$

at initial times $t = t_1$, t_2. Thus, our final approximation scheme reads

$$\left.\begin{array}{ll} U_k^\tau - U_{k-1}^\tau + \frac{\tau}{2} \cdot A(U_k^\tau + U_{k-1}^\tau) = 0, & k = 3, \dots, N, \\ U_k^\tau = (1 + \tau A)^{-1} U_{k-1}^\tau,\ U_0^\tau = u_0, & k = 1, 2. \end{array}\right\} \tag{3.3}$$

The approximation error at time $t = t_k > 0$ is then measured by means of the error function

$$E_k^\tau = U_k^\tau - u(t_k). \tag{3.4}$$

LEMMA 3.1. *Let $U_0^\tau \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$. Then the solution $(U_k^\tau)_{k=1,\dots,N}$ of (3.3) is unique and $U_k^\tau \in D_{A^{1+\varepsilon}}$ for each $k = 1, \dots, N$.*

Proof. Let $U_0^\tau \in D_{A^{1+\varepsilon}} \hookrightarrow D_A$. Then, in virtue of Lemmas 2.1 and 2.4, we obtain by rewriting (3.3)

$$U_k^\tau = (1 + \tau \cdot A)^{-k} U_0^\tau \in D_A$$

exist uniquely for $k = 1, 2$. Thus,

$$U_k^\tau = \left(1 + \frac{\tau}{2} \cdot A\right)^{-1} \left(1 - \frac{\tau}{2} \cdot A\right) U_{k-1}^\tau \in D_A \tag{3.5}$$

exist uniquely for $k \geq 3, \dots, N$ by successively calculating U_k^τ starting from the last Rothe step $U_2^\tau \in D_A$. Moreover, in virtue of Lemma 2.4 and because the operators A^ε commute with $(1 + \tau \cdot A)^{-1}$ on D_{A^ε} (see [24, p. 65], [12]), we obtain by observing $A^{1+\varepsilon}U_0^\tau \in H$

$$A^\varepsilon U_k^\tau = A^\varepsilon (1 + \tau \cdot A)^{-k} U_0^\tau = (1 + \tau \cdot A)^{-k} A^\varepsilon U_0^\tau \in D_A$$

for $k = 1, 2$. Thus,

$$\begin{aligned} A^\varepsilon U_k^\tau &= A^\varepsilon \left(1 + \frac{\tau}{2} \cdot A\right)^{-1} \left(1 - \frac{\tau}{2} \cdot A\right) U_{k-1}^\tau \\ &= \left(1 + \frac{\tau}{2} \cdot A\right)^{-1} \left[\, A^\varepsilon \left(1 - \frac{\tau}{2} \cdot A\right) U_{k-1}^\tau \,\right] \ \in\ D_A \end{aligned}$$

for $k = 3, \dots, N$ by induction, which implies $U_k^\tau \in D_{A^{1+\varepsilon}}$, $k = 1, \dots, N$. ∎

4. Convergence in $L^2(\Omega)$. Our main result reads as follows.

THEOREM 4.1. *Assume $u_0 \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$, and let $u \in C^0([0,\infty), D_{A^{1+\varepsilon}})$ be a strong solution of the Stokes initial value problem* (1.2). *Let $(U_k^\tau)_{k=1,\dots,N} \subset D_{A^{1+\varepsilon}}$ with $\tau = \frac{T}{N}$, $T > 0$, $N \in \mathbb{N}$, be the solution of the modified Crank–Nicolson scheme* (3.3). *Then the error estimate*

$$\|U_m^\tau - u(t_m)\| \le c \cdot \frac{\tau^2}{t_m^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \tag{4.1}$$

holds for each $1 \le m \le N$. The constant $c = c(\varepsilon)$ is independent of τ and t_m, but $c(\varepsilon) \to \infty$ as $\varepsilon \to 0^+$.

For the proof a parabolic duality argument to be introduced later on is essential. As a preparatory step we need a couple of a priori estimates of the error function (3.4), which we collect in Lemmas 4.2–4.4. Note that all constants appearing are positive and generic; i.e., they may have different values in different places.

LEMMA 4.2. *Assume $u_0 \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$, and let $u \in C^0([0,\infty), D_{A^{1+\varepsilon}})$ be a strong solution of the Stokes initial value problem* (1.2). *Let the modified Crank–Nicolson scheme* (3.3) *be started at any time $t_n > 0$ with initial value $U_n^\tau \in D_{A^{1+\varepsilon}}$. Then the estimates*

$$\|A^{-(1+\varepsilon)} E_m^\tau\| \le c_1 \cdot \left(\|A^{-(1+\varepsilon)} E_n^\tau\| + \frac{\tau^2}{t_n^{1-\varepsilon}} \cdot \|u_0\| \right), \tag{4.2}$$

$$\|E_m^\tau\| \le c_2 \cdot \left(\|E_n^\tau\| + \frac{\tau^2}{t_n^{1-\varepsilon}} \cdot \|A^{(1+\varepsilon)} u_0\| \right) \tag{4.3}$$

hold for each $m, n \in \mathbb{N}$ with $0 < n \le m \le N$. The constants $c_i = c_i(\varepsilon)$, $i = 1, 2$, are independent of τ and t_m.

Proof. We first show that the estimates (4.2), (4.3) are valid for the nonmodified Crank–Nicolson scheme (3.1): We integrate (1.2) over $[t_{k-1}, t_k]$ and add

$$Z_k := \frac{\tau}{2} \cdot A(u(t_k) + u(t_{k-1}))$$

on both sides to get

$$u(t_k) - u(t_{k-1}) + \int_{t_{k-1}}^{t_k} Au(t)dt + Z_k = Z_k. \tag{4.4}$$

Thus, subtracting (4.4) from (3.1) gives

$$E_k^\tau - E_{k-1}^\tau + \frac{\tau}{2} \cdot A(E_k^\tau + E_{k-1}^\tau) = \phi_k, \tag{4.5}$$

where

$$\phi_k = \int_{t_{k-1}}^{t_k} Au(t)dt - Z_k = \int_{t_{k-1}}^{t_k} Au(t)dt - \frac{\tau}{2} \cdot A(u(t_k) + u(t_{k-1})). \tag{4.6}$$

Equation (4.6) can be rewritten in the equivalent form

$$\phi_k = \int_{t_{k-1}}^{t_k} A \left[\frac{t_k - t}{\tau} \cdot (u(t) - u(t_{k-1})) + \frac{t - t_{k-1}}{\tau} \cdot (u(t) - u(t_k)) \right] dt. \tag{4.7}$$

Observing that u can be differentiated twice (see Lemma 2.2), we can expand u near t in the Taylor series

$$u(t) = u(t_{k-1}) + (t - t_{k-1}) \cdot \partial_t u(t) - \int_{t_{k-1}}^{t} (s - t_{k-1}) \cdot \partial_s^2 u(s) ds \tag{4.8}$$

and

$$u(t) \;=\; u(t_k) - (t_k - t) \cdot \partial_t u(t) - \int_{t}^{t_k} (t_k - s) \cdot \partial_s^2 u(s) ds. \tag{4.9}$$

Using the expansions (4.8), (4.9) in (4.7), we find in virtue of the linearity of A

$$\phi_k \;=\; \phi_{k1} \;+\; \phi_{k2}, \tag{4.10}$$

where

$$\phi_{k1} = -\int_{t_{k-1}}^{t_k} \left[\frac{t_k - t}{\tau} \cdot \int_{t_{k-1}}^{t} (s - t_{k-1}) \cdot A \partial_s^2 u(s) ds \right] dt, \tag{4.11}$$

$$\phi_{k2} = -\int_{t_{k-1}}^{t_k} \left[\frac{t - t_{k-1}}{\tau} \cdot \int_{t}^{t_k} (t_k - s) \cdot A \partial_s^2 u(s) ds \right] dt. \tag{4.12}$$

Note that $A\partial_t^2 u$ exists for $t \geq \tau > 0$ (see [9, Theorem 3, pp. 660–661 and Theorem 3′, p. 672]). Because $E_k^\tau \in D_{A^{1+\varepsilon}}$ for all $k \geq 1$ (see Lemma 3.1) and because of the compact imbedding $D_{A^\alpha} \hookrightarrow D_{A^\beta}$ for $\alpha > \beta$, we can take the inner product of (4.5) with $A^{-2(1+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)$ in H^0, which in virtue of the symmetry of A gives

$$\|A^{-(1+\varepsilon)} E_k^\tau\|^2 - \|A^{-(1+\varepsilon)} E_{k-1}^\tau\|^2 + \frac{\tau}{2} \cdot \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|^2 \leq g_k^\tau \tag{4.13}$$

with right-hand side

$$g_k^\tau = g_{k1}^\tau \;+\; g_{k2}^\tau, \tag{4.14}$$

where

$$g_{ki}^\tau = |\langle \phi_{ki}, A^{-2(1+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\rangle|, \quad i = 1, 2. \tag{4.15}$$

We estimate g_k^τ with the help of the representation (4.11) and (4.12) of ϕ_{ki}, $i = 1, 2$, in virtue of the boundedness of $A^{-(\frac{3}{2}+\varepsilon)}$, recalling the symmetry of A, as follows:

$$\begin{aligned} g_{k1}^\tau = |\langle \phi_{k1}, A^{-2(1+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\rangle| \;&=\; |\langle A^{-(\frac{3}{2}+\varepsilon)}\phi_{k1}, A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\rangle| \\ &\leq \tau^{-1} \cdot \left\| \int_{t_{k-1}}^{t_k} (t_k - t) \cdot \int_{t_{k-1}}^{t} (s - t_{k-1}) \cdot A^{-(\frac{3}{2}+\varepsilon)} A \partial_s^2 u(s) ds\, dt \right\| \cdot T_k, \end{aligned}$$

where

$$T_k^\tau = \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|. \tag{4.16}$$

Using the integral inequality

$$\left\| \int_a^b \varphi(t) dt \right\| \leq (b - a)^{\frac{1}{2}} \cdot \left| \int_a^b \|\varphi(t)\|^2 dt \right|^{\frac{1}{2}}, \tag{4.17}$$

which follows from the Cauchy–Schwarz inequality for each $\varphi \in L^2(a,b;H^0(\Omega))$, we conclude that

$$\begin{aligned} g_{k1}^\tau &\le \tau^{-\frac{1}{2}} \cdot \left| \int_{t_{k-1}}^{t_k} (t_k - t)^2 \cdot \left\| \int_{t_{k-1}}^{t} |(s - t_{k-1})| \cdot A^{-(\frac{1}{2}+\varepsilon)} \partial_s^2 u(s) ds \right\|^2 dt \right|^{\frac{1}{2}} \cdot T_k^\tau \\ &\le \left| \int_{t_{k-1}}^{t_k} (t_k - t)^2 \cdot \int_{t_{k-1}}^{t_k} (s - t_{k-1})^2 \cdot \|A^{-(\frac{1}{2}+\varepsilon)} \partial_s^2 u(s)\|^2 ds\, dt \right|^{\frac{1}{2}} \cdot T_k^\tau \\ &= \frac{\tau^{\frac{3}{2}}}{\sqrt{3}} \cdot \left| \int_{t_{k-1}}^{t_k} (s - t_{k-1})^2 \cdot \|A^{-(\frac{1}{2}+\varepsilon)} \partial_s^2 u(s)\|^2 ds \right|^{\frac{1}{2}} \cdot T_k^\tau \\ &\le \frac{\tau^{\frac{5}{2}}}{\sqrt{3}} \cdot \left| \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_s^2 u(s)\|^2 ds \right|^{\frac{1}{2}} \cdot T_k^\tau. \end{aligned}$$

Due to an analog estimate of g_{k2}^τ we arrive at

$$g_k^\tau = g_{k1}^\tau + g_{k2}^\tau \;\le\; 2\frac{\tau^{\frac{5}{2}}}{\sqrt{3}} \cdot \left| \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_s^2 u(s)\|^2 ds \right|^{\frac{1}{2}} \cdot T_k^\tau. \tag{4.18}$$

Applying the Cauchy–Young inequality (2.8) to (4.18) we conclude by definition (4.16) of T_k^τ that

$$g_k^\tau \le c_\eta \cdot \tau^4 \cdot \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_t^2 u(t)\|^2 dt + \eta \cdot \tau \cdot \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|^2, \tag{4.19}$$

which in virtue of (4.13) gives

$$\begin{aligned} \|A^{-(1+\varepsilon)} E_k^\tau\|^2 \;-\; \|A^{-(1+\varepsilon)} E_{k-1}^\tau\|^2 \;+\; \tau \cdot \left(\frac{1}{2} - \eta\right) \cdot \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|^2 \\ \le\; c_\eta \cdot \tau^4 \cdot \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_t^2 u(t)\|^2 dt. \end{aligned}$$

Thus, summing up from $k = n+1$ to m and neglecting

$$\tau \cdot \left(\frac{1}{2} - \eta\right) \cdot \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|^2 \ge 0$$

by assuming $\eta \le \frac{1}{2}$, we obtain the a priori estimate

$$\|A^{-(1+\varepsilon)} E_m^\tau\|^2 - \|A^{-(1+\varepsilon)} E_n^\tau\|^2 \le c_\eta \cdot \tau^4 \cdot \int_{t_n}^{t_m} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_t^2 u(t)\|^2 dt. \tag{4.20}$$

Recalling the semigroup representation $u(t) = e^{-tA} u_0$ for the solution of the homogeneous Stokes initial value problem (1.2), we differentiate $\partial_t u + Au = 0$, which in virtue of Lemma 2.2 and the closedness of A implies $\partial_t^2 u(t) = A^2 e^{-tA} u_0$ (see [24, pp. 75–76]). Then Lemma 2.3 can be applied to find

$$\|A^{-(\frac{1}{2}+\varepsilon)} \partial_t^2 u(t)\| \;=\; \|A^{\frac{3}{2}-\varepsilon} e^{-tA} u_0\| \;\le\; c \cdot t^{-\frac{3}{2}+\varepsilon} \cdot \|u_0\|,\; t > 0. \tag{4.21}$$

Finally, estimating the integrand of the right-hand side of (4.20) by (4.21), we are led to

$$\begin{aligned}\|A^{-(1+\varepsilon)}E_m^\tau\|^2 &= \|A^{-(1+\varepsilon)}E_n^\tau\|^2 - c_0\cdot\tau^4\cdot(t_m^{-2+2\varepsilon}-t_n^{-2+2\varepsilon})\cdot\|u_0\|^2\\ &\le \|A^{-(1+\varepsilon)}E_n^\tau\|^2 + c_1\cdot\tau^4\cdot t_n^{-2+2\varepsilon}\cdot\|u_0\|^2\\ &\le c\cdot\Big(\|A^{-(1+\varepsilon)}E_n^\tau\| + \tau^2\cdot t_n^{-1+\varepsilon}\cdot\|u_0\|\Big)^2\end{aligned}$$

with $c=\max\{1,c_1\}$. This proves (4.2) for the nonmodified Crank–Nicolson scheme (3.1).

For (4.3), taking the inner product of (4.5) with $E_k^\tau + E_{k-1}^\tau$ in H^0, a similar procedure as above leads to

$$\|E_k^\tau\|^2 - \|E_{k-1}^\tau\|^2 \;+\; \tau\cdot\left(\frac{1}{2}-\eta\right)\cdot\|A^{\frac12}(E_k^\tau+E_{k-1}^\tau)\|^2 \;\le\; c_\eta\cdot\tau^4\cdot\int_{t_{k-1}}^{t_k}\|A^{\frac12}\partial_t^2u(t)\|^2dt.$$

This gives

$$\|E_m^\tau\|^2 - \|E_n^\tau\|^2 \le c_\eta\cdot\tau^4\cdot\int_{t_n}^{t_m}\|A^{\frac12}\partial_t^2u(t)\|^2dt \tag{4.22}$$

by choosing $\eta\le\frac12$, neglecting $\tau\cdot(\frac12-\eta)\cdot\|A^{\frac12}(E_k^\tau+E_{k-1}^\tau)\|^2\le 0$, and summing up from $k=n+1$ to m. Estimating the integrand of the right-hand side of (4.22) with the help of Lemma 2.3 by

$$\|A^{\frac12}\partial_t^2u(t)\| \;=\; \|A^{\frac32-\varepsilon}e^{-tA}A^{1+\varepsilon}u_0\| \;\le\; c\cdot t^{-(\frac32-\varepsilon)}\cdot\|A^{1+\varepsilon}u_0\| \tag{4.23}$$

for $t>0$ proves (4.3) for the nonmodified Crank–Nicolson scheme (3.1) similarly as above by integrating the right-hand side of (4.22) using (4.23). Next, we show that (4.2), (4.3) hold locally for Rothe's scheme (3.2): Again, we integrate (1.2) over $[t_{k-1},t_k]$, but this time adding

$$\tilde Z_k := \tau\cdot Au(t_k)$$

on both sides yields

$$E_k^\tau - E_{k-1}^\tau + \tau\cdot AE_k^\tau = \varphi_k, \tag{4.24}$$

where

$$\varphi_k \;=\; \int_{t_{k-1}}^{t_k}Au(t)dt - \tau\cdot Au(t_k) \;=\; \int_{t_{k-1}}^{t_k}\int_{t_k}^{t}A\partial_su(s)ds\,dt. \tag{4.25}$$

Similarly as above, taking the inner product of (4.24) with $A^{-2(1+\varepsilon)}(E_k^\tau+E_{k-1}^\tau)$ in H^0, we obtain by substituting (4.25) for the right-hand side of (4.24)

$$\|A^{-(1+\varepsilon)}E_k^\tau\|^2 - \|A^{-(1+\varepsilon)}E_{k-1}^\tau\|^2 + \tau\cdot\|A^{-(\frac12+\varepsilon)}(E_k^\tau+E_{k-1}^\tau)\|^2 \le G_k^\tau \tag{4.26}$$

with

$$\begin{aligned}G_k^\tau \;&=\; |\langle\varphi_k, A^{-2(1+\varepsilon)}(E_k^\tau+E_{k-1}^\tau)\rangle| \;=\; |\langle A^{-(\frac32+\varepsilon)}\varphi_k, A^{-(\frac12+\varepsilon)}(E_k^\tau+E_{k-1}^\tau)\rangle|\\ &\le\; \|\int_{t_{k-1}}^{t_k}\int_{t_{k-1}}^{t}A^{-(\frac32+\varepsilon)}A\partial_su(s)dsdt\|\cdot T_k^\tau,\end{aligned}$$

where T_k^τ is again defined by (4.16). Thus, applications of the integral inequality (4.17) and the Cauchy–Young inequality (2.8) to G_k^τ yield

$$\begin{aligned}
G_k^\tau &\le \tau^{\frac{1}{2}} \cdot \left| \int_{t_{k-1}}^{t_k} \left\| \int_{t_{k-1}}^{t} A^{-(\frac{1}{2}+\varepsilon)} \partial_s u(s) ds \right\|^2 dt \right|^{\frac{1}{2}} \cdot T_k^\tau \\
&\le \tau^{\frac{1}{2}} \cdot \left| \int_{t_{k-1}}^{t_k} (t - t_{k-1}) \cdot \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_s u(s)\|^2 ds\, dt \right|^{\frac{1}{2}} \cdot T_k^\tau \\
&= \frac{\tau^{\frac{3}{2}}}{\sqrt{2}} \cdot \left| \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_s u(s)\|^2 ds \right|^{\frac{1}{2}} \cdot T_k^\tau \\
&\le c_\eta \cdot \tau^2 \cdot \int_{t_{k-1}}^{t_k} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_t u(t)\|^2 dt + \eta \cdot \tau \cdot (T_k^\tau)^2.
\end{aligned}$$

From this we find the a priori error bound

$$\|A^{-(1+\varepsilon)} E_m^\tau\|^2 - \|A^{-(1+\varepsilon)} E_n^\tau\|^2 \le c \cdot \tau^2 \cdot \int_{t_n}^{t_m} \|A^{-(\frac{1}{2}+\varepsilon)} \partial_t u(t)\|^2 dt \tag{4.27}$$

by absorbing $\eta \cdot \tau \cdot (T_k^\tau)^2 = \eta \cdot \tau \cdot \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|^2$ for $\eta \le 1$ into the left-hand side of (4.26) and subsequently neglecting the resulting term

$$(1-\eta) \cdot \tau \cdot (T_k^\tau)^2 = (1-\eta) \cdot \tau \cdot \|A^{-(\frac{1}{2}+\varepsilon)}(E_k^\tau + E_{k-1}^\tau)\|^2 \ge 0$$

and summing up from $k = n+1$ to $k = m$.

Observing $\partial_t u(t) = -Ae^{-tA} u_0$ in virtue of Lemma 2.2 and recalling the boundedness of $A^{-(\frac{1}{2}+\varepsilon)}$, we estimate the integrand of the right-hand side of (4.27) with the help of Lemma 2.3 by

$$\|A^{-(\frac{1}{2}+\varepsilon)} \partial_t u(t)\| \;=\; \|A^{\frac{1}{2}-\varepsilon} e^{-tA} u_0\| \le c \cdot t^{-\frac{1}{2}+\varepsilon} \cdot \|u_0\|. \tag{4.28}$$

Thus, integrating (4.27) using (4.28), we are led to

$$\begin{aligned}
\|A^{-(1+\varepsilon)} E_m^\tau\|^2 - \|A^{-(1+\varepsilon)} E_n^\tau\|^2 \;&\le\; c \cdot \tau^2 \cdot \int_{t_n}^{t_m} t^{-1+2\varepsilon} dt \cdot \|u_0\|^2 \\
=\; c \cdot \tau^2 \cdot \int_{t_n}^{t_m} t^{-2+2\varepsilon} \cdot t\, dt \cdot \|u_0\|^2 \;&\le\; c \cdot \tau^2 \cdot t_n^{-2+2\varepsilon} \cdot (t_m^2 - t_n^2) \cdot \|u_0\|^2 \\
=\; c \cdot \tau^4 \cdot t_n^{-2+2\varepsilon} \cdot (m^2 - n^2) \cdot \|u_0\|^2 \;&\le\; c \cdot \tau^4 \cdot t_n^{-2+2\varepsilon} \cdot (m-n)^2 \cdot \|u_0\|^2
\end{aligned}$$

for fixed $n < m$. Therefore,

$$\|A^{-(1+\varepsilon)} E_m^\tau\| \le c \cdot \left(\|A^{-(1+\varepsilon)} E_n^\tau\| \;+\; (m-n) \cdot \frac{\tau^2}{t_n^{1-\varepsilon}} \cdot \|u_0\| \right), \tag{4.29}$$

which is locally (4.2). Similarly there hold

$$\begin{aligned}
\|E_m^\tau\|^2 - \|E_n^\tau\|^2 &\le c \cdot \tau^2 \cdot \int_{t_n}^{t_m} \|A^{\frac{1}{2}} \partial_t u(t)\|^2 dt = c \cdot \tau^2 \cdot \int_{t_n}^{t_m} \|A^{\frac{1}{2}-\varepsilon} e^{-tA} A^{1+\varepsilon} u_0\|^2 dt \\
&\le c \cdot \tau^2 \cdot \int_{t_n}^{t_m} t^{-1+2\varepsilon} dt \cdot \|A^{1+\varepsilon} u_0\|^2 \\
&\le c \cdot \tau^4 \cdot t_n^{-2+2\varepsilon} \cdot (m-n)^2 \cdot \|A^{1+\varepsilon} u_0\|^2,
\end{aligned}$$

and thus

$$\|E_m^\tau\| \le c \cdot \left(\|E_n^\tau\| + (m-n) \cdot \frac{\tau^2}{t_n^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \right), \tag{4.30}$$

which is locally (4.3).

Finally, combining the global estimates (4.2) (resp., (4.3)) for the nonmodified Crank–Nicolson scheme (3.1) with the local estimates, (4.29) (resp., (4.30)) yields the desired estimates for the modified Crank–Nicolson scheme (3.3). ∎

LEMMA 4.3. *Assume $u_0 \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$, and let $u \in C^0([0,\infty), D_{A^{1+\varepsilon}})$ be a strong solution of the Stokes initial value problem* (1.2). *Let the modified Crank–Nicolson scheme* (3.3) *be started at time $t_0 = 0$ with initial value $U_0^\tau = u_0 \in D_{A^{1+\varepsilon}}$. Then the estimate*

$$\|E_m^\tau\| \le c \cdot \tau^{1+\varepsilon} \cdot \|A^{1+\varepsilon} u_0\| \tag{4.31}$$

holds for $1 \le m \le N$. The constant $c = c(\varepsilon)$ is independent of τ, but $c(\varepsilon) \to \infty$ as $\varepsilon \to 0^+$.

Proof. For Rothe's scheme, we recall

$$\|E_m^\tau\|^2 - \|E_n^\tau\|^2 \le c \cdot \tau^2 \cdot \int_{t_n}^{t_m} \|A^{\frac{1}{2}-\varepsilon} e^{-tA} A^{1+\varepsilon} u_0\|^2 dt \tag{4.32}$$

from the proof of Lemma 4.2. Thus, setting $n = 0$ and observing $E_0^\tau = 0$, we conclude that

$$\begin{aligned} \|E_m^\tau\|^2 &\le c \cdot \tau^2 \cdot \int_0^{t_m} t^{-1+2\varepsilon} dt \cdot \|A^{1+\varepsilon} u_0\|^2 \\ &= \frac{c}{2\varepsilon} \cdot \tau^2 \cdot t_m^{2\varepsilon} \cdot \|A^{1+\varepsilon} u_0\|^2 = \frac{c}{2\varepsilon} \cdot m^{2\varepsilon} \cdot \tau^{2+2\varepsilon} \cdot \|A^{1+\varepsilon} u_0\|^2. \end{aligned}$$

Therefore,

$$\|E_m^\tau\| \le \frac{c}{\sqrt{2\varepsilon}} \cdot m^\varepsilon \cdot \tau^{1+\varepsilon} \cdot \|A^{1+\varepsilon} u_0\| \ \le \ \frac{c}{\sqrt{2\varepsilon}} \cdot m \cdot \tau^{1+\varepsilon} \cdot \|A^{1+\varepsilon} u_0\| \tag{4.33}$$

because $0 < \varepsilon < 1$. Finally, combining the local estimate (4.33) for Rothe's scheme (3.2) and (4.3) of Lemma 4.2 gives the desired estimate (4.31) for the modified Crank–Nicolson scheme (3.3) as follows. For $0 \le m \le 2$ is nothing to prove, just take (4.33) for the Rothe scheme. For $m > 2$ we have in virtue of (4.3)

$$\begin{aligned} \|E_m^\tau\| &\le c_0 \cdot \left(\|E_2^\tau\| + \tau^2 \cdot t_2^{-1+\varepsilon} \cdot \|A^{1+\varepsilon} u_0\| \right) \\ &\le c_0 \cdot \left(2 \cdot c_1 \cdot \tau^{1+\varepsilon} + 2^{-1+\varepsilon} \cdot \tau^{1+\varepsilon} \right) \cdot \|A^{1+\varepsilon} u_0\| \le c_2 \cdot \tau^{1+\varepsilon} \cdot \|A^{1+\varepsilon} u_0\| \end{aligned}$$

because E_2^τ is calculated by Rothe's scheme and $E_0^\tau = 0$ by assumption. ∎

Additionally, the following lower order "smoothing" estimate holds (see [20]; see also [17] for the more general case of time-dependent operators).

LEMMA 4.4. *Assume $u_0 \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$, and let $u \in C^0([0,\infty), D_{A^{1+\varepsilon}})$ be a strong solution of the Stokes initial value problem* (1.2). *Let the modified Crank–Nicolson scheme* (3.3) *be started at time $t_0 = 0$ with initial value $U_0^\tau = u_0 \in D_{A^{1+\varepsilon}}$. Then the estimate*

$$\|E_m^\tau\| \le c \cdot \frac{\tau^{1-\varepsilon}}{t_m^{1-\varepsilon}} \cdot \|u_0\| \tag{4.34}$$

holds for $1 \leq m \leq N$. *The constant* c *is independent of* τ *and* t_m.

Proof. Let $u_0 \in D_{A^{1+\varepsilon}} \hookrightarrow H$. Then

$$\|E_m^\tau\| \leq c \cdot \frac{\tau}{t_m} \cdot \|u_0\| \tag{4.35}$$

for $m \geq 1$ (see [20, p. 347], [17, Lemma 3.6]). Thus,

$$\|E_m^\tau\| \leq c \cdot \frac{\tau^{1-\varepsilon}}{t_m^{1-\varepsilon}} \cdot \frac{\tau^\varepsilon}{t_m^\varepsilon} \cdot \|u_0\| \;\leq\; c \cdot \frac{\tau^{1-\varepsilon}}{t_m^{1-\varepsilon}} \cdot \|u_0\| \tag{4.36}$$

for $m \geq 1$. ∎

We now introduce the parabolic duality argument (see [16], [17]). For that we recall the self-adjointness of A, i.e., $A = A'$ (see section 2). Consider the "backward evolution system" to the Stokes initial value problem (1.2)

$$\left.\begin{array}{ll} \partial_t v - Av = 0, & T \geq t_m > t \geq 0, \\ v(t_m) = v_m. & \end{array}\right\} \tag{4.37}$$

Its corresponding backward discrete analogon of the modified Crank–Nicolson scheme (3.3) with given initial values $v(t_m) = V_m^\tau$ at time t_m with $0 < m \leq N$ reads

$$\left.\begin{array}{ll} V_k^\tau - V_{k-1}^\tau - \frac{\tau}{2} \cdot A(V_k^\tau + V_{k-1}^\tau) = 0, & k = m, \ldots, 3, \\ V_{k-1}^\tau = (1 + \tau \cdot A)^{-1} V_k^\tau, & k = 2, 1. \end{array}\right\} \tag{4.38}$$

Lemma 4.5. *Let* $v_m \in D_{A^{1+\varepsilon}}$, $\varepsilon \in (0,1)$, $0 < m \leq N$. *Then the backward problems* (4.37) *and* (4.38) *have unique solutions* v *and* $(V_k^\tau)_{k=m-1,\ldots,0}$, *with*

$$\begin{array}{l} v \in C^0([0,t_m], D_{A^{1+\varepsilon}}), \;\; v(t) = e^{-(t_m - t)A} v_m, \;\; t_m > t \geq 0, \\ V_k^\tau \in D_{A^{1+\varepsilon}}, \;\; m > k \geq 0. \end{array}$$

Proof. Let $v_m \in D_{A^{1+\varepsilon}}$. Then $v \in C^0([0,t_m], D_{A^{1+\varepsilon}})$ in virtue of Lemma 2.2 with time reversed. Let $s(t) = t_m - t > 0$. Then $\frac{d}{ds} e^{-sA} v_m + A e^{-sA} v_m = 0$ for $v_m \in D_{A^{1+\varepsilon}} \hookrightarrow H$ by Lemma 2.2. Let $v(t) = e^{-(t_m - t)A} v_m$. Then $\frac{d}{dt} v(t) = -\frac{d}{ds} e^{-sA} v_m$ and $v(t_m) = v_m$. Let $m \geq k \geq 3$. Rewriting (4.38) we obtain

$$V_{k-1}^\tau = \left(1 + \frac{\tau}{2} \cdot A\right)^{-1} \left(1 - \frac{\tau}{2} \cdot A\right) V_k^\tau = \left(1 - \frac{\tau}{2} \cdot A\right) \left(1 + \frac{\tau}{2} \cdot A\right)^{-1} V_k^\tau \in D_{A^{1+\varepsilon}}$$

by successively calculating V_k^τ starting from the given initial value $V_m^\tau \in D_{A^{1+\varepsilon}}$ (see the proof of Lemma 3.1). Thus

$$V_{k-1}^\tau = (1 + \tau \cdot A)^{-(3-k)} V_2^\tau \in D_{A^{1+\varepsilon}}$$

for $k = 2, 1$. ∎

For abbreviation in the following we let

$$u_\mu := u(t_\mu), \;\; v_\mu := v(t_\mu) \tag{4.39}$$

for the solution u and v of (1.2) and (4.37), respectively. Lemma 4.6 provides decisive inner product identities connecting the solutions of (1.2) and (3.3) with the corresponding solutions of the backward problems (4.37) and (4.38), respectively.

LEMMA 4.6. *Let u, $(U_k^\tau)_{k=1,\dots,m}$ be the solutions of* (1.2), (3.3)*, respectively, and v, $(V_k^\tau)_{k=m-1,\dots,0}$ be the solutions of* (4.37), (4.38)*, respectively. Then*

$$\langle u_p, v_p\rangle - \langle u_q, v_q\rangle = 0, \tag{4.40}$$

$$\langle U_p^\tau, V_p^\tau\rangle - \langle U_q^\tau, V_q^\tau\rangle = 0 \tag{4.41}$$

hold for each p, q with $0 \le p \le q \le m \le N$.

Proof. Let u and v be the solutions of (1.2) and (4.37), respectively. Then

$$\begin{aligned}\frac{d}{dt}\langle u(t), v(t)\rangle &= \langle \partial_t u(t), v(t)\rangle + \langle u(t), \partial_t v(t)\rangle\\ &= \langle -Au(t), v(t)\rangle + \langle u(t), Av(t)\rangle\\ &= \langle u(t), -Av(t)\rangle + \langle u(t), Av(t)\rangle = 0\end{aligned}$$

in virtue of Lemmas 2.2 and 4.5 because of the self-adjointness of A. Thus, the identity (4.40) follows by integration over $[t_p, t_q]$.

For the discrete solutions, for $3 \le \mu \le m \le N$, rewriting (3.3) and (4.38) in the equivalent forms

$$U_\mu^\tau = \left(1 + \frac{\tau}{2}\cdot A\right)^{-1}\left(1 - \frac{\tau}{2}\cdot A\right)U_{\mu-1}^\tau, V_{\mu-1}^\tau = \left(1 + \frac{\tau}{2}\cdot A\right)^{-1}\left(1 - \frac{\tau}{2}\cdot A\right)V_\mu^\tau,$$

we find in virtue of the commutativity of $(1 + \frac{\tau}{2}\cdot A)^{-1}$ and $(1 - \frac{\tau}{2}\cdot A)$ on D_A (see [24, p. 65]) and due to the symmetry of the Stokes resolvent (see [12, p. 279])

$$\begin{aligned}\langle U_\mu^\tau, V_\mu^\tau\rangle &= \langle (1+\tfrac{\tau}{2}\cdot A)^{-1}(1-\tfrac{\tau}{2}\cdot A)U_{\mu-1}^\tau, V_\mu^\tau\rangle\\ &= \langle U_{\mu-1}^\tau, (1+\tfrac{\tau}{2}\cdot A)^{-1}(1-\tfrac{\tau}{2}\cdot A)V_\mu^\tau\rangle = \langle U_{\mu-1}^\tau, V_{\mu-1}^\tau\rangle.\end{aligned}$$

Similarly, the same holds true for the corresponding resolvent equations of the Rothe steps ($1 \le \mu \le 2$), and thus

$$\langle U_\mu^\tau, V_\mu^\tau\rangle - \langle U_{\mu-1}^\tau, V_{\mu-1}^\tau\rangle = 0 \tag{4.42}$$

for $1 \le \mu \le m \le N$. The desired identity (4.41) follows by summing up (4.42) from $\mu = p+1$ to q , $0 \le p \le q \le m \le N$. □

In view of Lemmas 4.2–4.6 we are now prepared to finish the following proof.

Proof of Theorem 4.1. For estimating the error $\|E_m^\tau\|$, we will use that

$$\|E_m^\tau\| = \|E_m^\tau\|_H \;\le\; \sup\{|\langle E_m^\tau, v_m\rangle| \;:\; v_m \in H,\ \|v_m\| = 1\}$$

holds for $E_m^\tau \in D_{A^{1+\varepsilon}} \hookrightarrow H$. Thus,

$$\|E_m^\tau\| \le \sup\{|\langle E_m^\tau, v_m\rangle| \;:\; v_m \in D_{A^{1+\varepsilon}},\ \|v_m\| = 1\} \tag{4.43}$$

because D_{A^α} is dense in H for $\alpha > 0$. Therefore, our aim is to derive an identity for the inner product $\langle E_m^\tau, v_m\rangle$, which we then estimate with the help of Lemmas 4.2–4.4 and Lemma 4.6. For that let v and $(V_\mu^\tau)_{\mu=m-1,\dots,0}$ be the solutions of (4.37) and (4.38), respectively, corresponding to the given initial values $v_m = V_m^\tau$ at initial time

t_m with $1 \le m \le N$. In addition, let $M := [\frac{m}{2}]$ be the largest integer less than or equal to $\frac{m}{2}$. Then, applying Lemma 4.6, we obtain by observing $v_m - V_m^\tau = 0$

$$\begin{aligned}\langle E_m^\tau, v_m\rangle &= \langle U_m^\tau, v_m\rangle - \langle u_m, v_m\rangle \\ &= \langle U_m^\tau, v_m - V_m^\tau\rangle + \langle U_m^\tau, V_m^\tau\rangle - \langle u_m, v_m\rangle \\ &= \langle U_m^\tau, V_m^\tau\rangle - \langle u_m, v_m\rangle \\ &= \langle U_M^\tau, V_M^\tau\rangle - \langle u_M, v_M\rangle \\ &= \langle E_M^\tau, v_M\rangle + \langle U_M^\tau, V_M^\tau - v_M\rangle.\end{aligned}$$

Therefore, substituting $U_M^\tau = E_M^\tau + u_M$ into the latter equation, we will estimate the right-hand side of

$$\langle E_m^\tau, v_m\rangle \;=\; \langle E_M^\tau, v_M\rangle \;+\; \langle E_M^\tau, V_M^\tau - v_M\rangle \;+\; \langle u_M, V_M^\tau - v_M\rangle \tag{4.44}$$

in view of (4.43) as follows:

A bound for $|\langle E_M^\tau, v_M\rangle|$. Let $(\tilde V_\mu^\tau)_{\mu=M-1,\dots,0}$ be the solution of (4.38) corresponding to the given initial value $\tilde V_M^\tau = v_M$ at initial time $t_M \le t_m$. Then, using Lemma 4.6, we find

$$\begin{aligned}\langle E_M^\tau, v_M\rangle &= \langle U_M^\tau, v_M\rangle - \langle u_M, v_M\rangle \\ &= \langle U_M^\tau, v_M - \tilde V_M^\tau\rangle - \langle u_M, v_M\rangle + \langle U_M^\tau, \tilde V_M^\tau\rangle \\ &= \langle U_M^\tau, \tilde V_M^\tau\rangle - \langle u_M, v_M\rangle \\ &= \langle U_0^\tau, \tilde V_0^\tau\rangle - \langle u_0, v_0\rangle \\ &= \langle u_0, \tilde V_0^\tau\rangle - \langle u_0, v_0\rangle \\ &= \langle u_0, \tilde V_0^\tau - v_0\rangle \\ &= \langle A^{1+\varepsilon} u_0, A^{-(1+\varepsilon)}(\tilde V_0^\tau - v_0)\rangle\end{aligned}$$

because $U_0^\tau = u_0$ by the assumption of Theorem 4.1. Here the latter equation is due to the symmetry of A. Thus, application of the Cauchy–Schwarz inequality yields

$$|\langle E_M^\tau, v_M\rangle| \le \|A^{1+\varepsilon} u_0\| \cdot \|A^{-(1+\varepsilon)}(\tilde V_0^\tau - v_0)\|. \tag{4.45}$$

Next, in virtue of Lemma 4.5, we can apply (4.2) of Lemma 4.2 with time reversed to the right-hand side of (4.45) to obtain

$$\|A^{-(1+\varepsilon)}(\tilde V_0^\tau - v_0)\| \le c_0 \cdot \frac{\tau^2}{(t_m - t_M)^{1-\varepsilon}} \cdot \|v_m\| \tag{4.46}$$

because $\tilde V_M^\tau - v_M = 0$. Let $m \ge 1$. Then $M = [\frac{m}{2}] \le \frac{m}{2}$, i.e., $t_m - t_M \ge t_{\frac{m}{2}}$ or $\frac{1}{t_m - t_M} \le \frac{2}{t_m}$. Thus, combining (4.45) and (4.46) gives

$$|\langle E_M^\tau, v_M\rangle| \le c \cdot \frac{\tau^2}{t_m^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \cdot \|v_m\|, \tag{4.47}$$

where $c = 2^{1-\varepsilon} \cdot c_0$.

A bound for $|\langle E_M^\tau, V_M^\tau - v_M\rangle|$. Let $M = [\frac{m}{2}] \ge 0$. Then, applying (4.31) of Lemma 4.3 to E_M^τ yields

$$\|E_M^\tau\| \le c_0 \cdot \tau^{1+\varepsilon} \cdot \|A^{1+\varepsilon} u_0\|. \tag{4.48}$$

Let $M \geq 1$, i.e., $m \geq 2$. Then, with similar arguments as above, in virtue of Lemma 4.5, applying (4.34) of Lemma 4.4 with time reversed to $V_M^\tau - v_M$, we conclude that

$$\|v_M - V_M^\tau\| \leq c_1 \cdot \frac{\tau^{1-\varepsilon}}{(t_m - t_M)^{1-\varepsilon}} \cdot \|v_m\| \;\leq\; c_2 \cdot \frac{\tau^{1-\varepsilon}}{t_m^{1-\varepsilon}} \cdot \|v_m\|, \tag{4.49}$$

where $c_2 = 2^{1-\varepsilon} \cdot c_1$. Therefore, applying the Cauchy–Schwarz inequality and combining (4.48) and (4.49), we are led to

$$|\langle E_M^\tau, V_M^\tau - v_M\rangle| \leq c_0 \cdot c_2 \cdot \frac{\tau^2}{t_m^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \cdot \|v_m\| \tag{4.50}$$

for $m \geq 2$.

A bound for $|\langle u_M, V_M^\tau - v_M\rangle|$. Let $(\tilde{U}_\mu^\tau)_{\mu=M+1,\dots,m}$ be the solution of the modified Crank–Nicolson scheme (3.3) corresponding to the given initial value $\tilde{U}_M^\tau = u_M$ at initial time $t_M \leq t_m$. Then, in virtue of Lemma 4.6, observing $V_m^\tau = v_m$, we find

$$\begin{aligned}
\langle u_M, V_M^\tau - v_M\rangle &= \langle u_M, V_M^\tau\rangle - \langle u_M, v_M\rangle \\
&= \langle u_M - \tilde{U}_M^\tau, V_M^\tau\rangle + \langle \tilde{U}_M^\tau, V_M^\tau\rangle - \langle u_M, v_M\rangle \\
&= \langle \tilde{U}_M^\tau, V_M^\tau\rangle - \langle u_M, v_M\rangle \\
&= \langle \tilde{U}_m^\tau, V_m^\tau\rangle - \langle u_m, v_m\rangle \\
&= \langle \tilde{U}_m^\tau - u_m, v_m\rangle.
\end{aligned}$$

Thus, in virtue of the Cauchy–Schwarz inequality, it remains to estimate the right-hand side of

$$|\langle u_M, V_M^\tau - v_M\rangle| \leq \|\tilde{U}_m^\tau - u_m\| \cdot \|v_m\|. \tag{4.51}$$

Applying (4.3) of Lemma 4.2 with $n = M$ to the right-hand side of (4.51) yields

$$\|\tilde{U}_m^\tau - u_m\| \leq c_0 \cdot \frac{\tau^2}{t_M^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \tag{4.52}$$

because $\tilde{U}_M^\tau - u_M = 0$ by construction of $(\tilde{U}_\mu^\tau)$. Next, let $m \geq 2$. Then $M = [\frac{m}{2}] \geq \frac{m-1}{2} \geq \frac{m}{4}$, i.e., $\frac{1}{t_M} \leq \frac{4}{t_m}$. Then, from (4.51) and (4.52), we conclude that

$$|\langle u_M, V_M^\tau - v_M\rangle| \leq c_1 \cdot \frac{\tau^2}{t_m^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \cdot \|v_m\|, \tag{4.53}$$

where $c_1 = 4^{1-\varepsilon} \cdot c_0$. Finally, combining the bounds (4.47), (4.50), and (4.53) with (4.44), we arrive at

$$|\langle E_m^\tau, v_m\rangle| \leq c \cdot \frac{\tau^2}{t_m^{1-\varepsilon}} \cdot \|A^{1+\varepsilon} u_0\| \cdot \|v_m\|, \tag{4.54}$$

which gives the desired estimate (4.1) by applying (4.43) to (4.54). This completes the proof of Theorem 4.1. ∎

REFERENCES

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

[2] W. Borchers, *Zur Stabilität und Faktorisierungsmethode für die Navier–Stokes Gleichungen inkompressibler viskoser Flüssigkeiten* Habilitationsschrift für das Fach Mathematik im Fachbereich Mathematik-Informatik der Universität-GH Paderborn, Paderborn, Germany, 1992.

[3] L. Cattabriga, *Su un problema al contorno relativo al sistema de equazioni di Stokes*, Rend. Sem. Mat. Univ. Padova, 31 (1961), pp. 308–340.

[4] A. Friedman, *Partial Differential Equations*, rev. ed., Holt, Rinehart, and Winston, New York, 1976.

[5] H. Fujita and T. Kato, *On the Navier–Stokes initial value problem* I., Arch. Rational Mech. Anal., 16 (1964), pp. 269–315.

[6] D. Fujiwara and H. Morimoto, *An L_r-theorem of the Helmholtz decomposition of vector fields*, J. Fac. Sci. Univ. Tokyo, Sect. IA Math., 24 (1977), pp. 685–700.

[7] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier–Stokes Equations*, Vol. I, Springer, New York, 1994.

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

[9] J. G. Heywood, *The Navier–Stokes equations: On the existence, regularity, and decay of solutions*, Indiana Univ. Math. J., 29 (1980), pp. 639–681.

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

[11] J. G. Heywood and R. Rannacher, *Finite-element approximation of the nonstationary Navier–Stokes problem Part* IV: *Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[12] T. Kato, *Perturbation Theory for Linear Operators*, 2nd ed., Springer, Berlin, 1976.

[13] S. G. Krein, *Linear Differential Equations in Banach Spaces*, Trans. Amer. Math. Soc., AMS, Providence, RI, 1971.

[14] G. E. Ladas and V. Lakshmikantham, *Differential Equations in Abstract Spaces*, Academic Press, New York, London, 1972.

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

[16] M. Luskin and R. Rannacher, *On the smoothing property of the Galerkin method for parabolic equations*, SIAM J. Numer. Anal., 19 (1981), pp. 93–113.

[17] M. Luskin and R. Rannacher, *On the smoothing property of the Crank–Nicolson scheme*, Applicable Anal., 14 (1982), pp. 117–135.

[18] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1983.

[19] R. S. Phillips, *Dissipative operators and hyperbolic systems of partial differential equations*, Trans. Amer. Math. Soc., 90 (1959), pp. 193–254.

[20] R. Rannacher, *Discretization of the heat equation with singular initial data*, Z. Angew. Math. Mech., 62 (1982), pp. 346–348.

[21] R. Rautmann, *On optimum regularity of Navier–Stokes solutions at time $t = 0$*, Math. Z., 184 (1983), pp. 141–149.

[22] R. Rautmann, *H^2-convergent linearizations to the Navier–Stokes initial value problem*, in Proceedings of the International Conference on New Developments in Partial Differential Equations and Applications to Mathematical Physics (Ferrara, 1991), G. Butazzo, G. P. Galdi, and L. Zanghirati, eds., Plenum Press, New York, 1992, pp. 135–156.

[23] R. Rautmann, *H^2-convergence of Rothe's scheme to the Navier–Stokes equations*, Nonlinear Anal., 24 (1995), pp. 1081–1102.

[24] R. Rautmann and K. Masuda, *H^2-convergent approximation schemes to the Navier–Stokes Equations*, Comment. Math. Univ. St. Paul, 43 (1994), pp. 55–108.

[25] H. Sohr, *The Navier-Stokes Equations*, Birkhäuser, Basel, 2001.

[26] V. A. Solonnikov, *On the differential properties of the solution of the first boundary-value problem for the non-stationary system of Navier–Stokes equations*, Trudy Mat. Inst. Steklov, 73 (1964), pp. 221–291.

[27] V. A. Solonnikov, *Estimates of solutions of a nonstationary linearized system of Navier–Stokes equations*, Amer. Math. Soc. Transl. Ser. 2, 75 (1968), pp. 1–116.

[28] H. Tanabe, *Equations of Evolution*, Pitman, London, 1979.

[29] R. Temam, *Behavior at time $t = 0$ of the solutions of semi-linear evolution equations*, J.

Differential Equations, 43 (1982), pp. 73–92.
[30] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland, Amsterdam, 1978.
[31] W. Varnhorn, *Time stepping procedures for the nonstationary Stokes equations*, Math. Methods Appl. Sci., 15 (1992), pp. 39–55.
[32] W. Varnhorn, *The Stokes Equations*, Akademie Verlag, Berlin, 1994.
[33] W. Von Wahl, *The Equations of Navier–Stokes and Abstract Parabolic Equations*, Vieweg, Braunschweig, 1985.
[34] H. Weyl, *The method of orthogonal projection*, Duke Math. J., 7 (1940), pp. 441–444.
[35] K. Yosida, *Functional Analysis*, 4th ed., Springer, New York, Heidelberg, 1974.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 500–513

THEORETICALLY SUPPORTED SCALABLE FETI FOR NUMERICAL SOLUTION OF VARIATIONAL INEQUALITIES*

ZDENĚK DOSTÁL† AND DAVID HORÁK‡

Abstract. The FETI method with a natural coarse grid is combined with recently proposed optimal algorithms for the solution of bound and/or equality constrained quadratic programming problems in order to develop a scalable solver for elliptic boundary variational inequalities such as those describing equilibrium of a system of bodies in mutual contact. A discretized model problem is first reduced by the duality theory of convex optimization to the quadratic programming problem with bound and equality constraints. The latter is then modified by means of orthogonal projectors to the natural coarse grid introduced by Farhat, Mandel, and Roux [*Comput. Methods Appl. Mech. Engrg.*, 115 (1994), pp. 365–385]. Finally, the classical results on linear scalability for linear problems are extended to boundary variational inequalities. The results are validated by numerical experiments. The experiments also confirm that the algorithm enjoys the same parallel scalability as its linear counterpart.

Key words. domain decomposition, variational inequality, scalability, parallel algorithms, FETI

AMS subject classifications. 65N55, 65K10, 73V25

DOI. 10.1137/050639454

1. Introduction.

The FETI (finite element tearing and interconnecting) domain decomposition method was originally proposed by Farhat and Roux [26] for parallel solving of the linear problems described by elliptic partial differential equations. Its key ingredient is decomposition of the spatial domain into nonoverlapping subdomains that are "glued" by Lagrange multipliers, so that, after eliminating the primal variables, the original problem is reduced to a small, relatively well conditioned, typically equality constrained quadratic programming problem that is solved iteratively. The time that is necessary for both the elimination and iterations can be reduced nearly proportionally to the number of the processors, so that the algorithm enjoys parallel scalability. Observing that the equality constraints may be used to define so-called natural coarse grid, Farhat, Mandel, and Roux [25] modified the basic FETI algorithm so that they were able to adapt the results by Bramble, Pasciak, and Schatz [5] to prove its numerical scalability, i.e., asymptotically linear complexity. A comprehensive review of the mathematical results related to the FETI methods may be found in the monograph by Tosseli and Widlund [38].

If the FETI procedure is applied to an elliptic variational inequality, the resulting quadratic programming problem has not only the equality constraints but also the nonnegativity constraints. Even though the latter is a considerable complication as compared with linear problems, it seems that the FETI procedure should be even more powerful for the solution of variational inequalities than for linear problems. The reason is that FETI not only reduces the original problem to a smaller and

*Received by the editors September 2, 2005; accepted for publication (in revised form) August 17, 2006; published electronically March 2, 2007. This research was supported by grant 101/04/1145 of the Grant Agency of the Czech Republic, projects of the Ministry of Education MSM6198910027, 1ET400300415, and ME641, and AV ČR S3086102.

http://www.siam.org/journals/sinum/45-2/63945.html

†FEI VŠB-Technical University of Ostrava, CZ-70833 Ostrava, Czech Republic, and Institute of Geonics of AS CR, Ostrava, Czech Republic (zdenek.dostal@vsb.cz).

‡FEI VŠB-Technical University of Ostrava, CZ-70833 Ostrava, Czech Republic (david.horak@vsb.cz).

better conditioned one, but it also replaces for free all the inequalities by the bound constraints. Promising experimental results by Dureisseix and Farhat [22] supported this claim and even indicated numerical scalability of their method. Similar results were achieved also for the FETI–DP (dual–primal) method introduced by Farhat et al. [24]. The FETI–DP method is very similar to the original FETI; the only difference is that it enforces the continuity of displacements at corners on primal level. A new Lagrange multipliers algorithm, FETI–C, based on FETI–DP and on active set strategies with additional planning steps and preconditioning, was introduced by Avery et al. [1] and Dureisseix and Farhat [22]. Its scalability was demonstrated experimentally.

Another approach yielding experimental evidence of scalability was proposed by Dostál et al. [12, 13, 14]. The algorithm combined FETI with a special variant of the augmented Lagrangian method [10]. Scalability was later proved for an algorithm that enforced the equality constraints by the optimal dual penalty [15, 16] and solved the resulting bound constrained problem by recent, in a sense optimal, algorithms [7, 21]. Using the same algorithms, Dostál, Horák, and Stefanica then proved numerical scalability for a FETI–DP algorithm applied to two-dimensional (2D) coercive model problems discretized by means of either nodal [18] or mortar [19] Lagrange multipliers. Most recently, the scalability results were extended to include semicoercive problems [20]. The results used the effective condition number of the dual Schur complement of the stiffness matrix which was proved to be bounded by CH^2/h^2, where C is a constant independent of the discretization and decomposition parameters h and H, respectively. The results did not assume any preconditioning. Indeed, numerical experiments by the present authors, V. Vondrák, and M. Lesoinne indicated that the performance of our FETI–DP based algorithms may be considerably improved by preconditioning.

It should be noted that the effort to develop scalable solvers for variational inequalities was not restricted to FETI. For example, using ideas related to Mandel [34], Kornhuber, Krause, Sander, and Wohlmuth [30, 31, 40, 32, 33] gave experimental evidence of numerical scalability of the algorithm based on monotone multigrid. Probably the first theoretical results concerning development of scalable algorithms were proved by Schöberl [36, 37].

In this paper, we use the FETI method with a natural coarse grid to develop a scalable algorithm for numerical solution of both coercive and semicoercive variational inequalities. The result exploits the classical FETI1 upper bound CH/h [25] on the condition number of the regular part of the corresponding Hessian and remains valid for more general elliptic variational inequalities, including those describing equilibrium of a system of elastic bodies in mutual contact.

The paper is organized as follows. After describing a model problem, we briefly review the FETI methodology [12] that turns the variational inequality into a well conditioned quadratic programming problem with bound and equality constraints. Then we review our algorithms for the solution of the resulting bound and equality constrained quadratic programming problem whose rate of convergence may be expressed in terms of bounds on the spectrum of the dual Schur complement matrix [21, 8, 9]. Finally, we present the main results about optimality of our method and give results of numerical experiments with parallel implementation of the algorithm in PETSc [3].

2. Model problem. For the sake of simplicity, we shall reduce our analysis to a simple model problem, but our reasoning is valid also in more general cases,

including contact problems of 2D and three-dimensional (3D) elasticity, provided that the conditions exploited in the proof of the results by Farhat, Mandel, and Roux [25] are satisfied. Let $\Omega = \Omega^1 \cup \Omega^2$, $\Omega^1 = (0,1) \times (0,1)$, and $\Omega^2 = (1,2) \times (0,1)$ denote open domains with boundaries Γ^1, Γ^2 and their parts $\Gamma^i_u, \Gamma^i_f, \Gamma^i_c$ formed by the sides of Ω^i, $i = 1,2$, as in Figure 1(a) or 1(b). Let $H^1(\Omega^i)$, $i = 1,2$, denote the Sobolev space of the first order in the space $L^2(\Omega^i)$ of the functions on Ω^i whose squares are integrable in the sense of Lebesgue. Let

$$V^i = \left\{ v^i \in H^1(\Omega^i) : v^i = 0 \quad \text{on} \quad \Gamma^i_u \right\}$$

denote the closed subspaces of $H^1(\Omega^i)$, $i = 1,2$, and let

$$V = V^1 \times V^2 \qquad \text{and} \qquad \mathcal{K} = \left\{ (v^1, v^2) \in V : v^2 - v^1 \geq 0 \quad \text{on} \quad \Gamma_c \right\}$$

denote the closed subspace and the closed convex subset of $\mathcal{H} = H^1(\Omega^1) \times H^1(\Omega^2)$, respectively. The relations on the boundaries are in terms of traces. On $\mathcal{H}$ we shall define a symmetric bilinear form

$$a(u,v) = \sum_{i=1}^{2} \int_{\Omega^i} \left(\frac{\partial u^i}{\partial x} \frac{\partial v^i}{\partial x} + \frac{\partial u^i}{\partial y} \frac{\partial v^i}{\partial y} \right) d\Omega$$

and a linear form

$$\ell(v) = \sum_{i=1}^{2} \int_{\Omega^i} f^i v^i d\Omega,$$

where $f^i \in L^2(\Omega^i)$, $i = 1,2$, are the restrictions of

$$f(x,y) = \begin{cases} -1 & \text{for} \quad (x,y) \in (0,1) \times [0.75,1), \\ 0 & \text{for} \quad (x,y) \in (0,1) \times [0,0.75) \quad \text{and} \quad (x,y) \in (1,2) \times [0.25,1), \\ -3 & \text{for} \quad (x,y) \in (1,2) \times [0,0.25) \end{cases}$$

for a coercive problem and

$$f(x,y) = \begin{cases} -3 & \text{for} \quad (x,y) \in (0,1) \times [0.75,1), \\ 0 & \text{for} \quad (x,y) \in (0,1) \times [0,0.75) \quad \text{and} \quad (x,y) \in (1,2) \times [0.25,1), \\ -1 & \text{for} \quad (x,y) \in (1,2) \times [0,0.25) \end{cases}$$

for a semicoercive problem. Thus we can define a problem to find

$$\min \quad q(u) = \frac{1}{2} a(u,u) - \ell(u) \quad \text{subject to} \quad u \in \mathcal{K}. \tag{2.1}$$

We shall consider two variants of the Dirichlet data. In the first case, both the membranes are fixed on the outer edges as in Figure 1(a), so that

$$\Gamma^1_u = \{(0,y) \in \mathbb{R}^2 : y \in [0,1]\}, \qquad \Gamma^2_u = \{(2,y) \in \mathbb{R}^2 : y \in [0,1]\}.$$

Since the Dirichlet conditions are prescribed on parts Γ^i_u, $i = 1,2$, of the boundaries of both the membranes with positive measure, the quadratic form a is coercive, which guarantees both existence and uniqueness of the solution [28, 27]. In the second case, only the left membrane is fixed on the outer edge and the right membrane has no prescribed displacement as in Figure 1(b), so that

$$\Gamma^1_u = \{(0,y) \in \mathbb{R}^2 : y \in [0,1]\}, \qquad \Gamma^2_u = \emptyset.$$

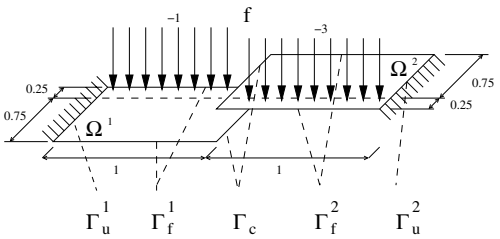


Fig. 1a: Coercive model problem

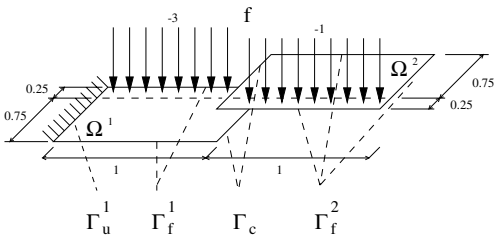


Fig. 1b: Semicoercive model problem

FIG. 1. *Model problems.*

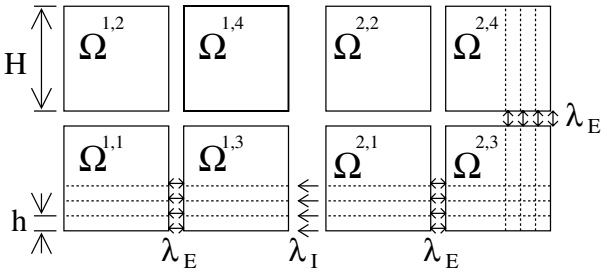


FIG. 2. *Domain decomposition and discretization.*

Even though a is in this case only semidefinite, the form q is still coercive due to the choice of f so that it has again the unique solution [28, 27].

More details about this particular model problem may be found in [12]. The solution of the model problem may be interpreted as the displacement of two membranes under the traction f. The left edge of the right membrane is not allowed to penetrate below the right edge of the left membrane.

3. Domain decomposition and discretization. In our definition of the problem, we have so far used only the natural decomposition of the spatial domain Ω into Ω^1 and Ω^2. However, to enable efficient application of the domain decomposition methods, we can optionally decompose each Ω^i into subdomains $\Omega^{i1}, \ldots, \Omega^{ip}$, $p > 1$, as in Figure 2. The continuity in Ω^1 and Ω^2 of the global solution assembled from the local solutions u^{ij} will be enforced by the "gluing" conditions $u^{ij}(x) = u^{ik}(x)$ that should be satisfied for any x on the interface $\Gamma^{ij,ik}$ of Ω^{ij} and Ω^{ik}. After modifying appropriately the definition of problem (2.1), introducing regular grids in the subdomains Ω^{ij} that match across the interfaces $\Gamma^{ij,kl}$, indexing contiguously the nodes and entries of corresponding vectors in the subdomains, and using the Lagrangian finite element discretization, we get the discretized version of problem (2.1) with auxiliary domain decomposition that reads

$$\min \frac{1}{2} u^\top K u - f^\top u \quad \text{s.t.} \quad B_{\mathcal{I}} u \le 0 \quad \text{and} \quad B_{\mathcal{E}} u = 0. \tag{3.1}$$

In (3.1), K denotes a block diagonal positive semidefinite stiffness matrix, the full rank matrices $B_{\mathcal{I}}$ and $B_{\mathcal{E}}$ describe the discretized nonpenetration and gluing conditions, respectively, and f represents the discrete analogue of the linear term $\ell(u)$. The rows of $B_{\mathcal{E}}$ and $B_{\mathcal{I}}$ are filled with zeros except 1 and -1 in positions that correspond to the nodes with the same coordinates on the artificial or contact boundaries, respectively.

In particular, if b_i denotes a row of $B_{\mathcal{E}}$ or $B_{\mathcal{I}}$, then b_i will not have more than four nonzero entries, and for any displacement vector u, $b_i u$ will denote the difference or jump between the displacements on each side of the boundary. Some more details may be found in [12].

Our next step is to simplify the problem, in particular to replace the general inequality constraints $B_{\mathcal{I}} u \leq 0$ by the nonnegativity constraints using the duality theory. To this end, let us introduce the Lagrangian associated with problem (3.1) by

$$L(u, \lambda_{\mathcal{I}}, \lambda_{\mathcal{E}}) = \frac{1}{2} u^\top K u - f^\top u + \lambda_{\mathcal{I}}^\top B_{\mathcal{I}} u + \lambda_{\mathcal{E}}^\top B_{\mathcal{E}} u, \tag{3.2}$$

where $\lambda_{\mathcal{I}}$ and $\lambda_{\mathcal{E}}$ are the Lagrange multipliers associated with inequalities and equalities, respectively. Introducing the notation

$$\lambda = \begin{bmatrix} \lambda_{\mathcal{I}} \\ \lambda_{\mathcal{E}} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{\mathcal{I}} \\ B_{\mathcal{E}} \end{bmatrix},$$

we can observe that B is a full rank matrix and write the Lagrangian briefly as

$$L(u, \lambda) = \frac{1}{2} u^\top K u - f^\top u + \lambda^\top B u.$$

It is well known [4] that (3.1) is equivalent to the saddle point problem

$$\text{Find} \quad (\overline{u}, \overline{\lambda}) \quad \text{so that} \quad L(\overline{u}, \overline{\lambda}) = \sup_{\lambda_{\mathcal{I}} \geq 0} \inf_u L(u, \lambda). \tag{3.3}$$

For fixed λ, the Lagrange function $L(\cdot, \lambda)$ is convex in the first variable and the minimizer u of $L(\cdot, \lambda)$ satisfies

$$Ku - f + B^\top \lambda = 0. \tag{3.4}$$

Equation (3.4) has a solution iff

$$f - B^\top \lambda \in \mathrm{Im} K, \tag{3.5}$$

which can be expressed more conveniently by means of a matrix R whose columns span the null space of K as

$$R^\top (f - B^\top \lambda) = 0. \tag{3.6}$$

The matrix R may be formed directly so that each floating subdomain is assigned to a column of R with ones in positions of the nodal variables that belong to the subdomain, and zeros elsewhere. It may be checked that $R^\top B^\top$ is a full rank matrix. The matrix R may also be extracted from K [23]. Now assume that λ satisfies (3.5) and denote by $K^\dagger$ any matrix that satisfies

$$K K^\dagger K = K. \tag{3.7}$$

Let us note that a generalized inverse $K^\dagger$ that satisfies (3.7) may be evaluated at a cost comparable with the Cholesky decomposition of regularized K [23]. It may be verified directly that if u solves (3.4), then there is a vector α such that

$$u = K^\dagger (f - B^\top \lambda) + R\alpha. \tag{3.8}$$

After substituting expression (3.8) into problem (3.3) and changing signs, we shall get the minimization problem to find

$$\min \Theta(\lambda) \quad \text{s.t.} \quad \lambda_{\mathcal{I}} \geq 0 \quad \text{and} \quad R^\top (f - B^\top \lambda) = 0, \tag{3.9}$$

where

$$\Theta(\lambda) = \frac{1}{2} \lambda^\top B K^\dagger B^\top \lambda - \lambda^\top B K^\dagger f. \tag{3.10}$$

Once the solution $\overline{\lambda}$ of (3.9) is known, the vector $\overline{u}$ that solves (3.1) can be evaluated by (3.8) and the formula [12]

$$\alpha = -(R^\top \widetilde{B}^\top \widetilde{B} R)^{-1} R^\top \widetilde{B}^\top \widetilde{B} K^\dagger (f - B^\top \overline{\lambda}), \tag{3.11}$$

where $\widetilde{B} = [\widetilde{B}_{\mathcal{I}}^\top, B_{\mathcal{E}}^\top]^\top$, and the matrix $\widetilde{B}_{\mathcal{I}}$ is formed by the rows b_i of $B_{\mathcal{I}}$ that correspond to the positive components of the solution $\overline{\lambda}$ characterized by $\overline{\lambda}_i > 0$.

4. Natural coarse grid. Even though the problem (3.9) is much more suitable for computations than (3.1) and was used to solve the discretized variational inequalities efficiently [11], further improvement may be achieved by adapting some simple observations and the results of Farhat, Mandel, and Roux [25]. Let us denote

$$F = B K^\dagger B^\top, \qquad \widetilde{d} = B K^\dagger f,$$
$$\widetilde{G} = R^\top B^\top, \qquad \widetilde{e} = R^\top f,$$

and let T denote a regular matrix that defines orthonormalization of the rows of $\widetilde{G}$ so that the matrix

$$G = T \widetilde{G}$$

has orthonormal rows. After denoting

$$e = T \widetilde{e},$$

problem (3.9) reads

$$\min \quad \frac{1}{2} \lambda^\top F \lambda - \lambda^\top \widetilde{d} \quad \text{s.t.} \quad \lambda_{\mathcal{I}} \geq 0 \quad \text{and} \quad G\lambda = e. \tag{4.1}$$

Next we shall transform the problem of minimization on the subset of the affine space to that on the subset of the vector space by looking for the solution of (4.1) in the form $\lambda = \mu + \widetilde{\lambda}$, where $G\widetilde{\lambda} = e$. The following lemma shows that we can even find $\widetilde{\lambda}$ such that $\widetilde{\lambda}_{\mathcal{I}} = 0$.

LEMMA 4.1. *Let B be such that the negative entries of $B_{\mathcal{I}}$ are in the columns that correspond to the nodes in the floating subdomain Ω^2. Then there is $\widetilde{\lambda}_{\mathcal{I}} \geq 0$ such that $G\widetilde{\lambda} = \widetilde{e}$.*

Proof. See [15] (coercive problem) or [16] (semicoercive problem). ∎

To carry out the transformation, denote $\lambda = \mu + \widetilde{\lambda}$ so that

$$\frac{1}{2} \lambda^\top F \lambda - \lambda^\top \widetilde{d} = \frac{1}{2} \mu^\top F \mu - \mu^\top (\widetilde{d} - F\widetilde{\lambda}) + \frac{1}{2} \widetilde{\lambda}^\top F \widetilde{\lambda} - \widetilde{\lambda}^\top \widetilde{d}$$

and the problem (4.1) is, after returning to the old notation, equivalent to

$$\min \quad \frac{1}{2}\lambda^\top F\lambda - \lambda^\top d \quad \text{s.t.} \quad G\lambda = 0 \quad \text{and} \quad \lambda_{\mathcal{I}} \geq -\widetilde{\lambda}_{\mathcal{I}}, \tag{4.2}$$

where $d = \widetilde{d} - F\widetilde{\lambda}$ and $\widetilde{\lambda}_{\mathcal{I}} \geq 0$.

Our final step is based on observation that the problem (4.2) is equivalent to

$$\min \quad \frac{1}{2}\lambda^\top (PFP + \overline{\rho}Q)\lambda - \lambda^\top Pd \quad \text{s.t.} \quad G\lambda = 0 \quad \text{and} \quad \lambda_{\mathcal{I}} \geq -\widetilde{\lambda}_{\mathcal{I}}, \tag{4.3}$$

where $\overline{\rho}$ is arbitrary positive constant and

$$Q = G^\top G \qquad \text{and} \qquad P = I - Q$$

denote the orthogonal projectors on the image space of $G^\top$ and on the kernel of G, respectively. The regularization term is introduced in order to simplify the reference to the results of quadratic programming that assume regularity of the Hessian matrix of the quadratic form. The problem (4.3) turns out to be a suitable starting point for development of an efficient algorithm for variational inequalities due to the classical estimates of the extreme eigenvalues. To formulate them, we shall denote by $\alpha_{\min}(A)$ and $\alpha_{\max}(A)$ the smallest and the largest eigenvalue of a given symmetric matrix A, respectively.

THEOREM 4.2. *There are constants $C_1 > 0$ and $C_2 > 0$ independent of the discretization parameter h and the decomposition parameter H such that*

$$\alpha_{\min}(PFP|\mathrm{Im}P) \ \geq \ C_1 \ \ \textit{and} \ \ \alpha_{\max}(PFP|\mathrm{Im}P) \leq ||PFP|| \ \leq \ C_2\frac{H}{h}. \tag{4.4}$$

Proof. See Theorem 3.2 of Farhat, Mandel, and Roux [25]. □

Note: The statement of Theorem 3.2 in [25] gives only an upper bound on the spectral condition number $\kappa(PFP|\mathrm{Im}P)$. However, the reasoning that precedes and substantiates their estimate proves both bounds of (4.4).

5. Optimal solvers to bound and equality constrained problems. We shall now briefly review our, in a sense, optimal algorithms for the solution of the bound and equality constrained problem (4.3). They combine our semimonotonic augmented Lagrangian method [8] which generates approximations for the Lagrange multipliers for the equality constraints in the outer loop with the working set algorithm for bound constrained auxiliary problems in the inner loop [21]. If a new Lagrange multiplier vector μ is used for the equality constraints, the augmented Lagrangian for problem (4.3) can be written as

$$L(\lambda, \mu, \rho) \ = \ \frac{1}{2}\lambda^\top (PFP + \overline{\rho}Q)\lambda \ - \ \lambda^\top Pd \ + \ \mu^\top G\lambda + \rho\lambda^\top Q\lambda.$$

The gradient of $L(\lambda, \mu, \rho)$ is given by

$$g(\lambda, \mu, \rho) \ = \ (PFP + \overline{\rho}Q)\lambda - Pd + G^T(\mu + \rho G\lambda).$$

Let $\mathcal{I}$ denote the set of the indices of the bound constrained entries of $\lambda \geq -\widetilde{\lambda}$. The *projected gradient* $g^P = g^P(\lambda, \mu, \rho)$ of L at λ is given componentwise by

$$g_i^P \ = \ \begin{cases} g_i & \text{for} \quad \lambda_i > -\widetilde{\lambda}_i \quad \text{or} \quad i \notin \mathcal{I}, \\ g_i^- & \text{for} \quad \lambda_i = -\widetilde{\lambda}_i \quad \text{and} \quad i \in \mathcal{I}, \end{cases}$$

where $g_i^- = \min\{g_i, 0\}$. Our algorithm is a variant of that proposed by Conn, Gould, and Toint [6] for identifying stationary points of more general problems. Its modification by Dostál, Friedlander, and Santos [10] was used by Dostál and Horák to develop a scalable FETI based algorithm, as shown experimentally in [14]. The key to proving optimality results is to combine the adaptive precision control of auxiliary problems in Step 1 with the new update rule for the penalty parameter ρ in Step 4. All the necessary parameters are listed in Step 0, and typical values of these parameters for our model problem are given in brackets.

ALGORITHM 5.1. Semimonotonic augmented Lagrangian method for bound and equality constrained problems (SMALBE).

Step 0. {*Initialization of parameters.*}
Given $\eta > 0$ [$\eta = \|Pd\|$], $\beta > 1$ [$\beta = 10$], $M > 0$ [$M = 1$],
$\rho_0 > 0$ [$\rho_0 = 100$], and μ^0 [$\mu^0 = 0$], set $k = 0$.

Step 1. {*Inner iteration with adaptive precision control.*}
Find λ^k such that $\lambda^k_{\mathcal{I}} \geq -\widetilde{\lambda}_{\mathcal{I}}$
$||g^P(\lambda^k, \mu^k, \rho_k)|| \leq \min\{M\|G\lambda^k\|, \eta\}$.

Step 2. {*Stopping criterion.*}
If $||g^P(\lambda^k, \mu^k, \rho_k)||$ and $||G\lambda^k||$ are sufficiently small,
then λ^k is the solution.
end if.

Step 3. {*Update of the Lagrange multipliers.*}
$\mu^{k+1} = \mu^k + \rho_k G\lambda^k$

Step 4. {*Update the penalty parameter.*}
If $k > 0$ and $L(\lambda^k, \mu^k, \rho^k) < L(\lambda^{k-1}, \mu^{k-1}, \rho_{k-1}) + \rho_k\|G\lambda^k\|^2/2$
then $\rho_{k+1} = \beta\rho_k$
else $\rho_{k+1} = \rho_k$
end if.

Step 5. Increase k and return to Step 1.

Step 1 may be implemented by any algorithm for minimization of the augmented Lagrangian L with respect to λ subject to $\lambda_{\mathcal{I}} \geq -\widetilde{\lambda_{\mathcal{I}}}$, which guarantees convergence of the projected gradient to zero. More about the properties and implementation of SMALBE algorithm may be found in [8].

The unique feature of the SMALBE algorithm is its capability to find an approximate solution to problem (4.3) in a number of steps which is uniformly bounded in terms of the bounds on the spectrum of $PFP + \overline{\rho}Q$ [8]. To get a bound on the number of matrix multiplication, it is necessary to have an algorithm which can solve the problem

$$\text{minimize } L(\lambda, \mu, \rho) \quad \text{s.t.} \quad \lambda_{\mathcal{I}} \geq -\widetilde{\lambda}_{\mathcal{I}} \tag{5.1}$$

with the rate of convergence in terms of the bounds on the spectrum of the Hessian matrix of L.

To describe such an algorithm, let us recall that the unique solution $\overline{\lambda} = \overline{\lambda}(\mu, \rho)$ of (5.1) satisfies the Karush–Kuhn–Tucker (KKT) conditions

$$g^P(\overline{\lambda}, \mu, \rho) = 0. \tag{5.2}$$

Let $\mathcal{A}(\lambda)$ and $\mathcal{F}(\lambda)$ denote the *active set* and *free set* of indices of λ, respectively, i.e.,

$$\mathcal{A}(\lambda) = \{i \in \mathcal{I} : \lambda_i = -\widetilde{\lambda}_i\} \quad \text{and} \quad \mathcal{F}(\lambda) = \{i : \lambda_i > -\widetilde{\lambda}_i \ \text{ or } \ i \notin \mathcal{I}\}.$$

To enable an alternative reference to the KKT conditions [4], let us define the *free gradient* $\varphi(\lambda)$ and the *chopped gradient* $\beta(\lambda)$ by

$$\varphi_i(\lambda) = \begin{cases} g_i(\lambda) & \text{for} \quad i \in \mathcal{F}(\lambda), \\ 0 & \text{for} \quad i \in \mathcal{A}(\lambda) \end{cases} \quad \text{and} \quad \beta_i(\lambda) = \begin{cases} 0 & \text{for} \quad i \in \mathcal{F}(\lambda), \\ g_i^-(\lambda) & \text{for} \quad i \in \mathcal{A}(\lambda) \end{cases}$$

so that the KKT conditions are satisfied iff the *projected gradient* $g^P(\lambda) = \varphi(\lambda) + \beta(\lambda)$ is equal to zero. We call λ *feasible* if $\lambda_i \geq -\widetilde{\lambda}_i$ for $i \in \mathcal{I}$. The projector P to the set of feasible vectors is defined for any λ by

$$P(\lambda)_i = \max\{\lambda_i, -\widetilde{\lambda}_i\} \quad \text{for} \quad i \in \mathcal{I}, \qquad P(\lambda)_i = \lambda_i \quad \text{for} \quad i \notin \mathcal{I}.$$

Let A denote the Hessian of L with respect to λ. The *expansion step* is defined by

$$\lambda^{k+1} = P\left(\lambda^k - \overline{\alpha}\varphi(\lambda^k)\right) \tag{5.3}$$

with the steplength $\overline{\alpha} \in (0, \|A\|^{-1}]$. This step may expand the current active set. To describe it without P, let $\widetilde{\varphi}(\lambda)$ be the *reduced free gradient* for any feasible λ, with entries

$$\widetilde{\varphi}_i = \widetilde{\varphi}_i(\lambda) = \min\{\lambda_i/\overline{\alpha}, \varphi_i\} \quad \text{for} \quad i \in \mathcal{I}, \qquad \widetilde{\varphi}_i = \varphi_i \quad \text{for} \quad i \in \mathcal{E}$$

such that

$$P\left(\lambda - \overline{\alpha}\varphi(\lambda)\right) = \lambda - \overline{\alpha}\widetilde{\varphi}(\lambda). \tag{5.4}$$

If the inequality

$$||\beta(\lambda^k)||^2 \leq \Gamma^2 \widetilde{\varphi}(\lambda^k)^\top \varphi(\lambda^k) \tag{5.5}$$

holds, then we call the iterate λ^k *strictly proportional*. The test (5.5) is used to decide which component of the projected gradient $g^P(\lambda^k)$ will be reduced in the next step.

The *proportioning step* is defined by

$$\lambda^{k+1} = \lambda^k - \alpha_{cg}\beta(\lambda^k).$$

The steplength α_{cg} is chosen to minimize $L(\lambda^k - \alpha\beta(\lambda^k), \mu^k, \rho_k)$ with respect to α, i.e.,

$$\alpha_{cg} = \frac{\beta(\lambda^k)^\top g(\lambda^k)}{\beta(\lambda^k)^\top A \beta(\lambda^k)}.$$

The purpose of the proportioning step is to remove indices from the active set.

The *conjugate gradient step* is defined by

$$\lambda^{k+1} = \lambda^k - \alpha_{cg} p^k, \tag{5.6}$$

where p^k is the conjugate gradient direction [2] which is constructed recurrently. The recurrence starts (or restarts) with $p^s = \varphi(\lambda^s)$ whenever λ^s is generated by the expansion step or the proportioning step. If p^k is known, then p^{k+1} is given by the formulae [2]

$$p^{k+1} = \varphi(\lambda^k) - \gamma p^k, \qquad \gamma = \frac{\varphi(\lambda^k)^\top A p^k}{(p^k)^\top A p^k}. \tag{5.7}$$

The conjugate gradient steps are used to carry out the minimization in the face $\mathcal{W}_{\mathcal{J}} = \{\lambda : \lambda_i = 0 \text{ for } i \in \mathcal{J}\}$ given by $\mathcal{J} = \mathcal{A}(\lambda^s)$ efficiently. The algorithm that we use may now be described as follows.

ALGORITHM 5.2. Modified proportioning with reduced gradient projections (MPRGP).

Let λ^0 be an n-vector such that $\lambda_i \geq -\widetilde{\lambda}_i$ for $i \in \mathcal{I}$, $\overline{\alpha} \in (0, \|A\|^{-1}]$, and $\Gamma > 0$ be given. For $k \geq 0$ and λ^k known, choose λ^{k+1} by the following rules:

Step 1. If $g^P(\lambda^k) = 0$, set $\lambda^{k+1} = \lambda^k$.

Step 2. If λ^k is strictly proportional and $g^P(\lambda^k) \neq 0$, try to generate λ^{k+1} by the conjugate gradient step. If $\lambda_i^{k+1} \geq 0$ for $i \in \mathcal{I}$, then accept it, else generate λ^{k+1} by the expansion step.

Step 3. If λ^k is not strictly proportional, define λ^{k+1} by proportioning.

The MPRGP algorithm has an R-linear rate of convergence in terms of the spectral condition number of the Hessian A of L [21]. More about the properties and implementation of the SMALBE algorithm may be found in [21] and [9].

6. Optimality. To show that Algorithm 5.1 with the inner loop implemented by Algorithm 5.2 is optimal for the solution of problem (or a class of problems) (4.3), we shall introduce new notation that complies with that used in [9].

We shall use

$$\mathcal{T} = \{(H, h) \in \mathbb{R}^2 : H \leq 1,\ 2h \leq H \ \text{ and } \ H/h \in \mathbb{N}\}$$

as the set of indices. Given a constant $C \geq 2$, we shall define a subset $\mathcal{T}_C$ of $\mathcal{T}$ by

$$\mathcal{T}_C = \{(H, h) \in \mathbb{R}^2 : H \leq 1,\ 2h \leq H,\ H/h \in \mathbb{N} \ \text{ and } \ H/h \leq C\}.$$

For any $t \in \mathcal{T}$, we shall define

$$\begin{aligned} A_t &= PFP + \overline{\rho}Q, \quad & b_t &= Pd, \\ C_t &= G, & \ell_{t,\mathcal{I}} &= -\widetilde{\lambda}_{\mathcal{I}}, \ \text{ and } \ \ell_{t,\mathcal{E}} = -\infty \end{aligned}$$

by the vectors and matrices generated with the discretization and decomposition parameters H and h, respectively, so that the problem (4.3) is equivalent to the problem

$$\text{minimize } \ \Theta_t(\lambda_t) \ \text{ s.t. } \ C_t\lambda_t = 0 \ \text{ and } \ \lambda_t \geq \ell_t, \tag{6.1}$$

where $\Theta_t(\lambda) = \frac{1}{2}\lambda^\top A_t\lambda - b_t^\top\lambda$. Using these definitions, Lemma 4.1, and $GG^\top = I$, we obtain

$$\|C_t\| \leq 1 \ \text{ and } \ \|\ell_t^+\| = 0, \tag{6.2}$$

where for any vector v with the entries v_i, v^+ denotes the vector with the entries $v_i^+ = \max\{v_i, 0\}$. Moreover, it follows by Theorem 4.2 that for any $C \geq 2$ there are constants $a^C_{\max} > a^C_{\min} > 0$ such that

$$a^C_{\min} \leq \alpha_{\min}(A_t) \leq \alpha_{\max}(A_t) \leq a^C_{\max} \tag{6.3}$$

for any $t \in \mathcal{T}_C$. Moreover, there are positive constants C_1 and C_2 such that $a^C_{\min} \geq C_1$ and $a^C_{\max} \leq C_2 C$. In particular, it follows that the assumptions of Theorem 5 (i.e., the inequalities (6.2) and (6.3)) of [9] are satisfied for any set of indices $\mathcal{T}_C$, $C \geq 2$, and we have the following result.

THEOREM 6.1. *Let $C \geq 2$ denote a given constant; let $\{\lambda_t^k\}$, $\{\mu_t^k\}$, and $\{\rho_{t,k}\}$ be generated by Algorithm* 5.1 *(SMALBE) for* (6.1) *with $\|b_t\| \geq \eta_t > 0$, $\beta > 1$, $M > 0$, $\rho_{t,0} = \rho_0 > 0$, and $\mu_t^0 = 0$. Let $s \geq 0$ denote the smallest integer such that $\beta^s \rho_0 \geq M^2/a_{\min}$ and assume that Step* 1 *of Algorithm* 5.1 *is implemented by means of Algorithm* 5.2 *(MPRGP) with parameters $\Gamma > 0$ and $\overline{\alpha} \in (0, (a_{\max} + \beta^s \rho_0)^{-1}]$, so that it generates the iterates $\lambda_t^{k,0}, \lambda_t^{k,1}, \ldots, \lambda_t^{k,l} = \lambda_t^k$ for the solution of* (6.1) *starting from $\lambda_t^{k,0} = \lambda_t^{k-1}$ with $\lambda_t^{-1} = 0$, where $l = l_{t,k}$ is the first index satisfying*

$$\|g^P(\lambda_t^{k,l}, \mu_t^k, \rho_{t,k})\| \leq M \|C_t \lambda_t^{k,l}\| \tag{6.4}$$

or

$$\|g^P(\lambda_t^{k,l}, \mu_t^k, \rho_{t,k})\| \leq \epsilon \|b_t\| \min\{1, M^{-1}\}. \tag{6.5}$$

Then for any $t \in \mathcal{T}_C$ and problem (6.1)*, Algorithm* 5.1 *generates an approximate solution $\lambda_t^{k_t}$ which satisfies*

$$M^{-1} \|g^P(\lambda_t^{k_t}, \mu_t^{k_t}, \rho_{t,k_t})\| \leq \|C_t \lambda_t^{k_t}\| \leq \epsilon \|b_t\| \tag{6.6}$$

at $O(1)$ matrix-vector multiplications by the Hessian of the augmented Lagrangian L_t for (6.1) *and $\rho_{t,k} \leq \beta^s \rho_0$.*

7. Numerical experiments. In this section we report some results of numerical solution of the semicoercive model problem of section 2 in order to illustrate the performance of the algorithm, in particular its numerical and parallel scalability. To this end, we have implemented Algorithm 5.1 with the solution of auxiliary bound constraints by Algorithm 5.2 in C exploiting PETSc [3] to solve problem (4.3) with varying decomposition and discretization parameters.

The experiments were run on the Lomond 18-processor Sun HPC 6500 Ultra SPARC-II based SMP system with 400 MHz, 18 GB of shared memory, 90 GB disc space, nominal peak performance 14.4 GFlops, 16 kB level 1 and 8 MB level 2 cache in EPCC Edinburgh, and on the Turing Cray T3E 1200, 788 applications processors, each 1.2 GFlops with 256 MB, 209 GB memory, 28 command processors, 2TB disk space, high-speed network with low latency in the University of Manchester. All the computations were carried out with parameters

$$M = 1, \quad \rho_0 = 10, \quad \Gamma = 1, \quad \lambda^0 = \max\left\{-\widetilde{\lambda}, \frac{1}{2} Bf\right\}, \quad \epsilon = 10^{-4}.$$

The solutions of our benchmarks are in Figure 3. The results of computations are summarized in Tables 1–3.

Table 1 illustrates numerical scalability of Algorithm 5.1. In particular, for varying decompositions and discretization parameters, the upper row of each field of the table gives the corresponding primal dimension/dual dimension/times in seconds, while the number in the lower row gives a number of the conjugate gradient iterations that were necessary for the solution of the problem to the given precision. We can see that the number of the conjugate gradient iterations for a given ratio H/h (in rows) varies very moderately.

Table 2 indicates that the algorithm presented enjoys high parallel scalability. The results for the largest problems are in Table 3.

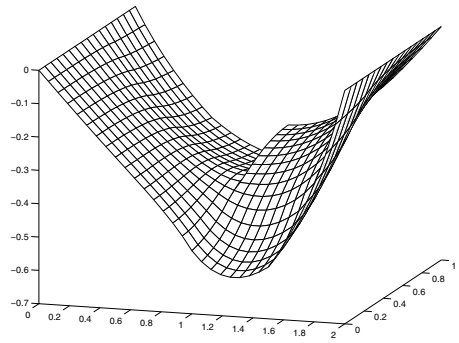


Fig. 3a: Coercive problem

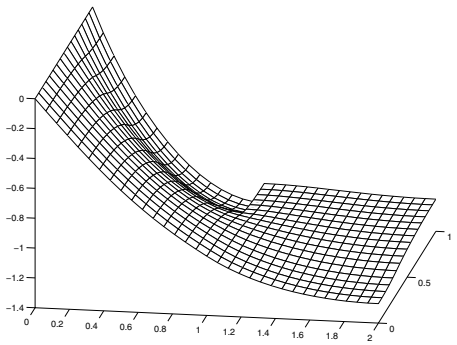


Fig. 3b: Semicoercive problem

FIG. 3. *Solution of model problems.*

TABLE 1
Performance for varying decomposition and discretization.

H	1	1/2	1/4	1/8
H/h \ procs	2	8	16	16
128	33282/129/41.95 28	133128/1287/89.50 59	532512/6687/74.9 36	2130048/29823/421.5 47
64	8450/65/2.04 22	33800/647/4.14 47	135200/3359/7.10 33	540800/14975/53.48 43
32	2178/33/0.20 17	8712/327/0.50 33	34848/1695/1.48 30	139392/7551/11.66 37
16	578/17/0.04 13	2312/167/0.18 29	9248/863/0.68 26	36992/3839/4.30 32
8	162/9/0.03 10	648/87/0.10 20	2592/447/0.39 23	10365/1983/2.06 27
4	50/5/0.01 7	200/47/0.04 19	800/239/0.28 22	3200/1055/1.30 25

TABLE 2
Parallel scalability for 128 *subdomains.*

Processors	1	2	4	8	16	32
Time[sec]	2907.13	1022.03	462.4	165.8	68.06	51.40

TABLE 3
Highlights.

h	H	Prim. dim.	Dual. dim.	Num. of subdom.	Procs	Out. iter.	Cg. iter.	Time [sec]
1/1024	1/8	2130048	29823	128	32 of Lomond	2	47	167
1/2048	1/8	8454272	59519	128	32 of Lomond	2	65	1202

8. Comments and conclusion. We have presented scalability results related to the application of the augmented Lagrangians with the FETI based domain decomposition method using the natural coarse grid to the solution of variational inequalities by recently developed algorithms for the solution of special quadratic programming problems. In particular, we have shown that the solution of the discretized elliptic variational inequality to a prescribed precision may be found in a number of matrix vector multiplications bounded independently of the discretization parameter provided the

ratio of the decomposition and the discretization parameters is kept bounded. Numerical experiments with a model variational inequality are in agreement with the theory and indicate that the algorithm may be efficient. The results remain valid also for the solution of frictionless 2D and 3D contact problems of elasticity and may be adapted to the solution of problems with Coulomb friction as indicated in [17]. The solution of auxiliary linear problems in the inner loop may be improved by standard preconditioners [29, 35, 38] and may be adapted to the mortar discretization [39]; however, since the preconditioning transforms the bound constraints to more general inequality constraints, it is nontrivial to get improved convergence results in this way. We shall discuss these topics elsewhere.

REFERENCES

[1] P. Avery, G. Rebel, M. Lesoinne, and C. Farhat, *A numerically scalable dual-primal substructuring method for the solution of contact problems.* I: *The frictionless case*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2403–2426.

[2] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.

[3] S. Balay, W. Gropp, L. C. McInnes, and B. Smith, *PETSc* 2.0 *Users Manual*, http://www.mcs.anl.gov/petsc/.

[4] D. P. Bertsekas, *Nonlinear Programming*, Athena Scientific, Cambridge, MA, 1999.

[5] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring* I, Math. Comp., 47 (1986), pp. 103–134.

[6] A. R. Conn, N. I. M. Gould, and P. L. Toint, *A globally convergent augmented Lagrangian algorithm for optimization with general constraints and simple bounds*, SIAM J. Numer. Anal., 28 (1991), pp. 545–572.

[7] Z. Dostál, *A proportioning based algorithm for bound constrained quadratic programming with the rate of convergence*, Numer. Algorithms, 34 (2003), pp. 293–302.

[8] Z. Dostál, *Inexact semi-monotonic augmented Lagrangians with optimal feasibility convergence for quadratic programming with simple bounds and equality constraints*, SIAM J. Numer. Anal., 43 (2005), pp. 96–115.

[9] Z. Dostál, *An optimal algorithm for bound and equality constrained quadratic programming problems with bounded spectrum*, Computing, 78 (2006), pp. 311–328.

[10] Z. Dostál, A. Friedlander, and S. A. Santos, *Augmented Lagrangians with adaptive precision control for quadratic programming with simple bounds and equality constraints*, SIAM J. Optim., 13 (2003), pp. 1120–1140.

[11] Z. Dostál, A. Friedlander, and S. A. Santos, *Solution of contact problems of elasticity by FETI domain decomposition*, in Domain Decomposition Methods 10, Contemp. Math. 218, AMS, Providence, RI, 1998, pp. 82–93.

[12] Z. Dostál, F. A. M. Gomes, and S. A. Santos, *Duality based domain decomposition with natural coarse space for variational inequalities*, J. Comput. Appl. Math., 126 (2000), pp. 397–415.

[13] Z. Dostál, F. A. M. Gomes, and S. A. Santos, *Solution of contact problems by FETI domain decomposition with natural coarse space projection*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1611–1627.

[14] Z. Dostál and D. Horák, *Scalability and FETI based algorithm for large discretized variational inequalities*, Math. Comput. Simulation, 61 (2003), pp. 347–357.

[15] Z. Dostál and D. Horák, *Scalable FETI with optimal dual penalty for a variational inequality*, Numer. Linear Algebra Appl., 11 (2004), pp. 455–472.

[16] Z. Dostál and D. Horák, *Scalable FETI with optimal dual penalty for semicoercive variational inequalities*, in Current Trends in Scientific Computing, Contemp. Math. 329, AMS, Providence, RI, 2003, pp. 79–88.

[17] Z. Dostál, D. Horák, R. Kučera, V. Vondrák, J. Haslinger, J. Dobiáš, and S. Pták, *FETI based algorithms for contact problems: Scalability, large displacements and* 3*D Coulomb friction*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 395–409.

[18] Z. Dostál, D. Horák, and D. Stefanica, *A scalable FETI-DP algorithm for coercive variational inequalities*, Appl. Numer. Math., 54 (2005), pp. 378–390.

[19] Z. Dostál, D. Horák, and D. Stefanica, *A scalable FETI–DP algorithm with non–penetration mortar conditions on contact interface*, submitted.

[20] Z. Dostál, D. Horák, and D. Stefanica, *A scalable FETI–DP algorithm for semi-coercive*

variational inequalities, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 1369–1379.
[21] Z. Dostál and J. Schöberl, *Minimizing quadratic functions over non-negative cone with the rate of convergence and finite termination*, Comput. Optim. Appl., 30 (2005), pp. 23–44.
[22] D. Dureisseix and C. Farhat, *A numerically scalable domain decomposition method for solution of frictionless contact problems*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 2643–2666.
[23] C. Farhat and M. Géradin, *On the general solution by a direct method of a large scale singular system of linear equations: Application to the analysis of floating structures*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 675–696.
[24] C. Farhat, M. Lesoinne, P. LeTallec, K. Pierson, and D. Rixen, *FETI-DP: A dual–prime unified FETI method.* I: *A faster alternative to the two–level FETI method*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 1523–1544.
[25] C. Farhat, J. Mandel, and F.-X. Roux, *Optimal convergence properties of the FETI domain decomposition method*, Comput. Methods Appl. Mech. Engrg., 115 (1994), pp. 365–385.
[26] C. Farhat and F.-X. Roux, *An unconventional domain decomposition method for an efficient parallel solution of large-scale finite element systems*, SIAM J. Sci. Comput., 13 (1992), pp. 379–396.
[27] R. Glowinski, *Variational Inequalities*, Springer-Verlag, Berlin, 1980.
[28] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek, *Solution of Variational Inequalities in Mechanics*, Springer-Verlag, Berlin, 1988.
[29] A. Klawonn and O. B. Widlund, *FETI and Neumann-Neumann iterative substructuring mathods: Connections and new results*, Commun. Pure Appl. Math., 54 (2001), pp. 57–90.
[30] R. Kornhuber, *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*, Teubner, Stuttgart, 1997.
[31] R. Kornhuber and R. Krause, *Adaptive multigrid methods for Signorini's problem in linear elasticity*, Comput. Vis. Sci., 4 (2001) pp. 9–20.
[32] R. Krause and O. Sander, *Fast solving of contact problems on complicated geometries*, in Lecture Notes in Computational Science and Engineering 40, R. Kornhuber et al., eds., Springer-Verlag, Berlin, 2005, pp. 495–502.
[33] R. H. Krause and B. I. Wohlmuth, *A Dirichlet-Neumann type algorithm for contact problems with friction*, Comput. Vis. Sci., 5 (2002), pp. 139–148.
[34] J. Mandel, *Étude algébrique d'une méthode multigrille pour quelques problèmes de frontière libre*, Comptes Rendus de l'Academie des Sciences Sr. I, 298, 1984, pp. 469–472 (in French).
[35] J. Mandel and R. Tezaur, *Convergence of substructuring method with Lagrange multipliers*, Numer. Math., 73 (1996), pp. 473–487.
[36] J. Schöberl, *Solving the Signorini problem on the basis of domain decomposition techniques*, Computing, 60 (1998), pp. 323–344.
[37] J. Schöberl, *Efficient contact solvers based on domain decomposition techniques*, Comput. Math., 42 (1998), pp. 1217–1228.
[38] A. Toselli and O. B. Widlund, *Domain Decomposition Methods: Algorithms and Theory*, Springer Series on Computational Mathematics 34, Springer-Verlag, Berlin, 2004.
[39] B. I. Wohlmuth, *Discretization Methods and Iterative Solvers Based on Domain Decomposition*, Springer-Verlag, Berlin, 2001.
[40] B. I. Wohlmuth and R. Krause, *Monotone methods on nonmatching grids for nonlinear contact problems*, SIAM J. Sci. Comput., 25 (2003), pp. 324–347.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 514–538

ERROR CONTROL FOR A CLASS OF RUNGE–KUTTA DISCONTINUOUS GALERKIN METHODS FOR NONLINEAR CONSERVATION LAWS*

ANDREAS DEDNER†, CHARALAMBOS MAKRIDAKIS‡, AND MARIO OHLBERGER†

Abstract. We propose an a posteriori error estimate for the Runge–Kutta discontinuous Galerkin method (RK-DG) of arbitrary order in arbitrary space dimensions. For stabilization of the scheme a general framework of projections is introduced. Finally it is demonstrated numerically how the a posteriori error estimate is used to design both an efficient grid adaption and gradient limiting strategy. Numerical experiments show the stability of the scheme and the gain in efficiency in comparison with computations on uniform grids.

Key words. discontinuous Galerkin, higher order, adaptive methods, error estimate, finite element

AMS subject classifications. 35L65, 65M60, 76M10

DOI. 10.1137/050624248

1. Introduction.

In this paper we study a generalized version of the Runge–Kutta discontinuous Galerkin (RK-DG) approximation of Cockburn and Shu (see [10, 7, 11]) for nonlinear scalar conservation laws in several space dimensions. As a prototype conservation law, consider the Cauchy initial value problem

$$\partial_t u + \nabla \cdot f(u) = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+ , \tag{1.1}$$

$$u(x,0) = u_0(x) \quad \text{in } \mathbb{R}^d. \tag{1.2}$$

Here $u : \mathbb{R}^d \times \mathbb{R}^+ \to \mathbb{R}$ denotes the dependent solution variable, $f \in C^1(\mathbb{R})$ denotes the flux function, and $u_0 \in \mathrm{BV}(\mathbb{R}^\mathrm{d}) \cap \mathrm{L}^\infty(\mathbb{R}^\mathrm{d})$ the initial data with $u_0 \in [A, B]$ a.e. It is well known (see, for example, [15, 12]) that (1.1)–(1.2) admits a unique entropy weak solution in the class of functions of bounded variation (BV). For later use let us briefly recall that an entropy weak solution is a weak solution of (1.1)–(1.2) which satisfies for all entropy pairs (S, F_S) and all test functions $\phi \in C_0^1(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^+)$:

$$-\int_{\mathbb{R}^d}\int_{\mathbb{R}^+} \left(S(u)\partial_t\phi + F_S(u)\cdot\nabla\phi\right)\, dt\, dx - \int_{\mathbb{R}^d} S(u_0)\phi(x,0)\, dx \le 0\,. \tag{1.3}$$

Recall that (S, F_S) is called an entropy-entropy flux pair or, more simply, an entropy pair for (1.1), iff S is convex and $F_S' = S' f'$.

Numerical methods for nonlinear hyperbolic conservation laws are usually rather complicated since they need to approximate a partial differential equation with nonstandard stability behavior. It turns out that in many successful computational methods the theoretical backup is very limited. This is partly because when constructing high-order methods, stabilization terms have to be added so that in the limit the

*Received by the editors February 11, 2005; accepted for publication (in revised form) September 1, 2006; published electronically March 2, 2007. This work was supported by the European RTN–network HYKE under contract number HPRN-CT-2002-00282.
http://www.siam.org/journals/sinum/45-2/62424.html
†Abteilung für Angewandte Mathematik, Universität Freiburg, Hermann-Herder-Str. 10, D-79104 Freiburg, Germany (dedner@mathematik.uni-freiburg.de, mario@mathematik.uni-freiburg.de).
‡Department of Applied Mathematics, University of Crete, GR-71409 Heraklion, Greece (makr@tem.uoc.gr).

solution satisfying (1.3) is computed. These mechanisms include shock capturing terms or limiters that result in complicated and highly nonlinear schemes; see [8] for a comprehensive review of high-order finite difference, finite volume, and finite element methods for hyperbolic conservation laws. The RK-DG approximation of Cockburn and Shu [11] is a very successful method that combines many desirable properties. It is based on totally discontinuous finite element spaces for the space discretization, while the time discretization is based on appropriate Runge–Kutta schemes. The available theory for RK-DG methods for nonlinear problems is limited to certain stability and TVD (total variation diminishing) properties proved in [9, 11] and to error estimates for one dimensional smooth solutions (see [35]). The problem of showing convergence towards the unique entropy solution for the high-order version of these methods seems rather difficult.

In this paper we consider a generalized version of RK-DG methods designed for use with dynamic mesh modification. We are interested in the following question: Is it possible to establish a rigorous error control for RK-DG methods in mesh adaptive computations? An answer to this question will be based on certain a posteriori estimates and does not necessarily depend on available a priori convergence results for the method. As a consequence, it provides a (nonstandard) way of theoretical backup for a method with no available convergence theory. In this context we also refer to [16], where this was done for MUSCL finite difference schemes.

First, we prove a posteriori error estimates for generalized DG methods. We then use these estimates to provide an *hp*-adaptive algorithm that is used together with a rigorous error control. The computational performance of the resulting methods and algorithms is tested in one dimensional examples.

The literature on a posteriori error control and adaptive solution algorithms for (RK-DG) approximations is scant. We refer, for instance, to Hartmann and Houston [18] and Larson and Barth [27], where duality techniques were used for designing adaptive schemes. We also refer to Süli and Houston and coworkers [20, 19, 33] for *hp*-adaptive DG methods for hyperbolic problems. Another approach towards error control for DG methods was introduced by Adjerid et al. [1], where asymptotically correct a posteriori estimates of spatial discretization errors were derived in one dimension for smooth solutions.

2. Formulation of the generalized DG methods and main results. Let $\mathcal{T}$ denote an element decomposition of $\mathbb{R}^d$ with control volumes $T_j \in \mathcal{T}$, $j \in J$ such that $\cup_{T\in\mathcal{T}}\overline{T} = \mathbb{R}^d$. Let h_T denote a length scale associated with each control volume T, e.g., $h_T \equiv \operatorname{diam}(T)$. For two distinct control volumes T_i and T_j in $\mathcal{T}$, the intersection is either an oriented edge ($d = 2$) or face ($d = 3$) S_{ij} with oriented normal ν_{ij}, or else a set of measure at most $d-2$. The set $N(j)$ denotes the index set of neighboring control volumes to T_j, and the index set of the oriented edges or faces of the grid is given by $\mathcal{E} = \{(j,l)|\, T_j \in \mathcal{T},\, l \in N(j),\, j > l\}$. The set of edges or faces of the element decomposition $\mathcal{T}$ will be denoted by Γ. On $\mathcal{T}$ we define the space of (possible) discontinuous piecewise polynomials of degree p by $V_h^p := \{v_h \in BV(\mathbb{R}^d)|\, v_T := v_h|_T \in \mathbb{P}_p \text{ for all } T \in \mathcal{T}\}$. Let us denote by $\Pi_{V_h^p}$ the L^2-projection into V_h^p. Furthermore, following standard notation, $[v_h]|_{S_{ij}} := (v_j|_{S_{ij}} - v_i|_{S_{ij}})\nu_{ij}$ is the jump of v_h on the edge S_{ij}, and $\{v_h\}_{S_{ij}} := 1/2(v_j|_{S_{ij}} + v_i|_{S_{ij}})$ denotes the mean of v_h at an interface.

DEFINITION 2.1 (space-discrete DG approximation). *$u_h \in C^1(0,T;V_h^p)$ is called a semidiscrete DG approximation of* (1.1)–(1.2) *iff*

$$u_h(0) = \Pi_{V_h^p}(u_0), \tag{2.1}$$

$$\frac{d}{dt}(u_h(t), v_h) - (f(u_h(t)), \nabla v_h) + (f_h(u_h(t)), [v_h])_\Gamma = 0 \quad \text{for all } v_h \in V_h^p. \tag{2.2}$$

Here $(\cdot,\cdot)$ denotes the L^2 inner product, $(\cdot,\cdot)_\Gamma$ denotes the L^2 inner product on the set of interfaces in Γ, and f_h denotes a given numerical flux function that is uniquely defined on the interfaces of the element decomposition (see Assumption 2.2).

Note that due to the fact that both u_h and the test space V_h^p are discontinuous, the global definition of the scheme (2.2) is equivalent to the following local one:

$$\frac{d}{dt}(u_j(t), v_j)_{T_j} - (f(u_j(t)), \nabla v_j)_{T_j} + \sum_{l \in N(j)} (f_{jl}(u_j(t), u_l(t)), v_j)_{S_{jl}} = 0 \tag{2.3}$$

$$\text{for all } v_j \in \mathbb{P}_p, T_j \in \mathcal{T}.$$

Here $(\cdot,\cdot)_{T_j}$, $(\cdot,\cdot)_{S_{jl}}$ denote the local inner product on T_j, S_{jl}, respectively, and $f_{jl}(u_j(t), u_l(t))$ is the restriction of $f_h(u_h)$ to S_{jl}. Note that the numerical fluxes f_{jl} are usually defined via a standard finite difference "upwind-type" one dimensional flux, and it is the only source of "artificial viscosity" in the scheme (2.2). We make the following standard assumptions on the numerical flux function.

ASSUMPTION 2.2 (numerical flux function). *The numerical fluxes are supposed to be functions $f_{jl} \in C^1(\mathbb{R}^2, \mathbb{R})$ which satisfy for all $u, v, u', v' \in [A, B]$ the following conditions (respectively, monotony, conservation, regularity, and consistency):*

$$\partial_u f_{jl}(u,v) \geq 0, \qquad \partial_v f_{jl}(u,v) \leq 0, \qquad f_{jl}(u,v) = -f_{lj}(v,u), \tag{2.4}$$

$$f_{jl}(u,u) = n_{jl}|S_{jl}|\mathbf{f}(u), \qquad |f_{jl}(u,v) - f_{jl}(u',v')| \leq LS_{jl}|(|u-u'| + |v-v'|). \tag{2.5}$$

In the literature of DG methods the stabilization due to the "upwinding" of the discrete fluxes is usually accompanied by extra artificial "shock capturing" terms as in [21, 22, 5] or limiting projections as in [11]. (Error estimates for the shock capturing DG method were obtained in [5].) The RK-DG methods introduced by Cockburn and Shu are based on a combination of limiting projections and Runge–Kutta discretization of the ODE (2.2). Therefore, in the next step we are going to introduce limiting projections in the discretization that will be chosen in section 5.

2.1. Generalized semidiscrete DG approximation. We introduce a hybrid scheme that incorporates all the characteristics of a RK-DG scheme used with mesh modification with time, but assumes that the ODE in time is solved exactly in each time step. To this end we introduce a partition of the time interval $(0, T_{\max})$, $\{0 = t^0, \ldots, t^N = T_{\max}\}$, and we define the time step $\Delta t^n := t^{n+1} - t^n$. With each time interval $(t^n, t^{n+1}]$ we associate a (possibly different) finite element space $V_{h,n}^p$ on a grid $\mathcal{T}_n$ defined as

$$V_{h,n}^p := \{v_h \in BV(\mathbb{R}^d) |\, v_h|_T \in \mathbb{P}_p \text{ for all } T \in \mathcal{T}_n\}. \tag{2.6}$$

The associated index set of the grid $\mathcal{T}_n$ is denoted by J^n. In what follows we might often drop the index n in objects related to the finite element space.

To define a local projection operator we proceed as follows: We define $\overline{v_h}$ through $\overline{v_j} := \Pi_{V_h^0}(v)|_{T_j}$ for any $v \in L^2(\Omega)$; i.e., $\overline{v_h}$ is the elementwise average of v. Furthermore, with each n we associate projections $\Lambda_h^{n,t}$ with the following properties.

ASSUMPTION 2.3 (projection operator). *The projection $\Lambda_h^{n,t}$ is supposed to be a continuous function with respect to t on the interval $[t^n, t^{n+1}]$. If $t \in (t^n, t^{n+1}]$, the operators act $\Lambda_h^{n,t} : V_{h,n}^p \to V_{h,n}^p$ and satisfy*

$$\overline{\Lambda_h^{n,t}(v_h(\cdot,t))} = \overline{v_h(t)}, \quad t \in (t^n, t^{n+1}]\,. \tag{2.7}$$

In addition, $\Lambda_h^{n,t^n} : V_{h,n-1}^p \to V_{h,n}^p$ is a projection to the new mesh, with the property

$$\overline{\Lambda_h^{n,t^n}(v_h(\cdot,t^n))} = \overline{v_h(t^n)}\,. \tag{2.8}$$

In the last equation the elementwise average is taken in the new mesh, i.e., corresponds to the projection $\Pi_{V_{h,n}^0}$. At t^n the two operators Λ_h^{n,t^n} and Λ_h^{n-1,t^n} satisfy

$$\|\Lambda_h^{n,t^n}(u_h) - \overline{u_h}\|_\infty \le \|\Lambda_h^{n-1,t^n}(u_h) - \overline{u_h}\|_\infty. \tag{2.9}$$

Properties (2.7), (2.8) lead to a conservation of mass, whereas assumption (2.9) guarantees that the gradients in the discrete solution are not increased between time steps. Note that $\Lambda_h^{n,t}$ accounts for both limiting projections and projections to the new spaces. We define the restriction of $\Lambda_j^{n,t}$ through $\Lambda_j^{n,t} \equiv \Lambda_h^{n,t}$ in $T_j \times [t^n, t^{n+1}]$, $j \in J^n$.

DEFINITION 2.4 (generalized semidiscrete DG approximation). *Suppose that $\Lambda_h^{n,t}$ with the above properties is given, and assume that the fluxes f_{ij} are monotone. u_h is called a generalized semidiscrete DG approximation of (1.1)–(1.2) if $u_h^{-1} := \Pi_{V_{h,0}^p}(u_0)$ and for $n = 0, \dots, N-1$, $u_h^n|_{[t^n,t^{n+1}]} \in C^1(t^n, t^{n+1}; V_{h,n}^p)$ is defined through*

$$u_h^n(t^n) := \Lambda_h^{n,t^n}(u_h^{n-1}(t^n))\,, \tag{2.10}$$

$$\begin{aligned}\frac{d}{dt}(u_j^n(t), v_j)_{T_j} \;=\; & -\sum_{l\in N(j)} (f_{jl}(\Lambda_j^{n,t}(u_h^n(t)), \Lambda_l^{n,t}(u_h^n(t))), v_j)_{S_{jl}} \\ & + (f(\Lambda_j^{n,t}(u_h^n(t))), \nabla v_j)_{T_j} \text{ for all } v_j \in \mathbb{P}_p, j \in J^n, t \in (t^n, t^{n+1})\,.\end{aligned} \tag{2.11}$$

We then define $u_h \in L^\infty(0, T_{\max}; V_{h,n}^p)$ as $u_h(0) := u_h^{-1}$, and $u_h|_{(t^n,t^{n+1}]} := u_h^n|_{(t^n,t^{n+1}]}$.

In section 5 we combine the above method with Runge–Kutta time discretizations to obtain the generalized class of fully discrete RK-DG methods. This class includes the method of Cockburn and Shu, but we consider alternative choices for the limiting projections motivated by the a posteriori result for (2.10)–(2.11), proved in what follows.

2.2. A posteriori error estimate for the semidiscrete DG method. We will show a posteriori estimates for the error $\|(u - u_h)(T_{\max})\|_{L^1}$. To do that we compare u and u_h with $\widetilde{u}_h$ defined as

$$\widetilde{u}_h(t) = \Lambda_h^{n,t}(u_h(t)) \quad \text{for } t \in (t^n, t^{n+1}], \quad n = 0, \dots, N-1. \tag{2.12}$$

Then $\|(\widetilde{u}_h - u_h)(T_{\max})\|_{L^1}$ is an a posteriori quantity, and the control of $\|(u - \widetilde{u}_h)(T_{\max})\|_{L^1}$ will be obtained in what follows by employing Kruzkov estimates.

Note that, by definition, $\widetilde{u}_h$ might be discontinuous at the time nodes t^n. This will be the case either when the spatial mesh is modified at this node, or when we decide to use different projections on $(t^{n-1}, t^n]$ and $(t^n, t^{n+1}]$. In fact, due to the definitions of u_h and the projections, we have

$$\widetilde{u}_h(t^{n+}) - \widetilde{u}_h(t^n) = \Lambda_h^{n,t^n} u_h(t^n) - \Lambda_h^{n-1,t^n} u_h(t^n) = (\Lambda_h^{n,t^n} - \Lambda_h^{n-1,t^n}) u_h(t^n)\,. \tag{2.13}$$

Before stating our main result we introduce the following notation:

$$\widetilde{u}_j = \widetilde{u}_h \quad \text{in } T_j, \qquad \widetilde{u}^n = \widetilde{u}_h \quad \text{in } (t^n, t^{n+1}], \qquad \widetilde{u}^n(t^n) = \widetilde{u}_h(t^{n+}),$$

with the obvious extension for combined indexes.

THEOREM 2.5 (a posteriori error estimate for the semidiscrete DG method). *Let u_h be given by the semidiscrete generalized DG method* (2.10)–(2.11). *For $\widetilde{u}_h$ given by* (2.12) *we have the following a posteriori error estimate:*

$$||(u-u_h)(T_{\max})||_{L^1(B_R(x_0))} \le ||(\widetilde{u}_h - u_h)(T_{\max})||_{L^1(B_R(x_0))} + \eta_h,$$

where $\eta_h := \eta_0 + \sqrt{K_1\eta_1} + \sqrt{K_2\eta_2}$, $\eta_0 := \sum_{j\in J^0}\eta_{0,j}$, $\eta_i := \sum_n \sum_{j\in J^n}\eta^n_{i,j}$, $i = 1, 2$, and the local contributions $\eta^n_{i,j}$ are given as

(2.14)

$$\eta_{0,j} := \int_{T_j} |u_0 - \widetilde{u}^0_j(0)|, \qquad \eta^n_{1,j} := h_j R^n_{T,j} + \frac{1}{2} h_{jl} R^n_{S,j} + h_j R^n_{\Lambda,j},$$

$$\eta^n_{2,j} := ||\overline{\widetilde{u}^n_j} - \widetilde{u}^n_j||_{L^\infty((t^n,t^{n+1})\times T_j)} R^n_{T,j} + \frac{1}{2} \max_{k\in\{j,l\}} ||\overline{\widetilde{u}^n_k} - \widetilde{u}^n_k||_{L^\infty((t^n,t^{n+1})\times S_{jl})} R^n_{S,j}$$

$$+ ||\overline{\widetilde{u}^{n-1}}(t^n) - \widetilde{u}^{n-1}(t^n)||_{L^\infty(T_j)} R^n_{\Lambda,j}. \tag{2.15}$$

Here, we used the notation

$$R^n_{T,j} := \int_{t^n}^{t^{n+1}} \int_{T_j} \left|\partial_t \widetilde{u}_j + \nabla\cdot f(\widetilde{u}_j)\right|, \quad R^n_{\Lambda,j} := \int_{T_j} |\widetilde{u}^n(t^{n+1}) - \widetilde{u}^{n+1}(t^{n+1})|, \tag{2.16}$$

$$R^n_{S,j} := \int_{t^n}^{t^{n+1}} \sum_{l\in N(j)} \int_{S_{jl}} Q_{jl}(\widetilde{u}_j, \widetilde{u}_l)|\widetilde{u}_j - \widetilde{u}_l|, \quad h_{jl} := \max_{l\in N(j)} \operatorname{diam}(T_j \cup T_l), \tag{2.17}$$

$$\text{and} \quad Q_{jl}(u,v) := \frac{2f_{jl}(u,v) - f_{jl}(u,u) - f_{jl}(v,v)}{u-v}. \tag{2.18}$$

K_1, K_2 are constants depending on the total variation of the initial data and on the maximal slope of the flux, but they are independent of the maximal time $T_{\max}$. The independence of $T_{\max}$ is due to the fact that we consider only the semidiscrete scheme. For a detailed definition of the constants, see the proof on page 525.

The error estimator in Theorem 2.5 is composed of the two parts η_1, η_2. The first part corresponds to the standard estimates known for first order schemes [6, 25, 30], and the second part of the estimate corresponds to error terms which are present only in higher order approximations. In the following Corollary 2.6 we have rearranged these terms so that the estimate is more suitable for designing an adaptive scheme.

COROLLARY 2.6. *With the assumptions and notations of Theorem* 2.5, *we have*

$$\eta_h \le \eta_0 + R_h, \quad \text{with} \quad R_h^2 = 2\sum_n \sum_{j\in J^n} \rho^n_j \left(R^n_{T,j} + R^n_{S,j} + R^n_{\Lambda,j}\right), \tag{2.19}$$

$$\text{and} \qquad \rho^n_j := K_1 h_j + K_2 \max_{k\in\{j,\, l\in N(j)\}} ||\overline{\widetilde{u}^n_k} - \widetilde{u}^n_k||_{L^\infty((t^n,t^{n+1})\times T_k)}. \tag{2.20}$$

Proof. The estimate $\eta_h \le \eta_0 + R_h$ is a direct consequence of Theorem 2.5 if we estimate $\sqrt{K_1\eta_1} + \sqrt{K_2\eta_2} \le \sqrt{2K_1\eta_1 + 2K_2\eta_2}$. By rearranging terms in $K_1\eta_1 + K_2\eta_2$ and using $h_{jl} \le h_j + h_l$, we obtain (2.19). ∎

We are now going to discuss some aspects of the error estimate.

Semidiscrete versus fully discrete estimates. The above a posteriori result is extended in a straightforward manner when the ODE (2.1) is discretized by Euler's method. RK-DG methods, though, use high-order Runge–Kutta schemes for time discretization. The proof of a result for high-order Runge–Kutta schemes is nontrivial and requires new ideas. Therefore, an analysis of the fully discrete case is left for future work.

The error bound of Theorem 2.5 is used to design our adaptive algorithm in section 5. To do that we introduce in section 4 the fully discrete generalized RK-DG method and express it as an ODE for each time slab $[t^n, t^{n+1}]$. This is done by using the *natural continuous extension* for Runge–Kutta schemes introduced by Zennaro [34].

First-order versus high-order estimates. In the case $p = 0$ the DG method reduces to a standard finite volume scheme that allows mesh modification with n. Then the first term in $\eta^n_{1,j}$ and the whole $\eta^n_{2,j}$ will be zero. The last term in $\eta^n_{1,j}$ will account for coarsening errors due to mesh modification. Such terms were not included in the previous a posteriori estimates for finite volume schemes [6, 25, 29]. Another implication due to higher-order polynomials used in the finite element spaces is the appearance of $\overline{\widetilde{u}^n_j} - \widetilde{u}^n_j$ in various norms in the term $\eta^n_{2,j}$. A comparison with $\eta^n_{1,j}$ leads to the conclusion that it would be desirable to have $\overline{\widetilde{u}^n_j} - \widetilde{u}^n_j = O(h_j)$. In general this is not guaranteed unless $\widetilde{u}^n_j$ is a result of certain limiting projections which restrict gradients or/and polynomial degrees of u_h. This observation is one of the main motivations for the choice of the limiting projections and the design of the adaptive algorithm in section 5.

Computational "convergence" of the estimators. Theorem 2.5 is a rather general result that covers any projection $\Lambda^{n,t}_h$ with the properties (2.7) and (2.8). In addition, due to the generality of Kruzkov's estimates used in the proof above, an a posteriori bound can be seen as a "worst case scenario" upper bound. It is clear that if in the computational runs the estimators converge to zero, then the error will do the same. In this sense Theorem 2.5 allows for the design of error control algorithms based on upper bound estimates. This issue is discussed in detail in sections 5 and 6. At this point we would like to note that, among many other choices presented in section 5, $h - p$ versions of RK-DG methods allow error control algorithms based on the estimators of Theorem 2.5. In addition, for the test problems discussed in this paper we examine the computational behavior for the RK-DG method with limiters from [10, 7] and the corresponding estimator of Theorem 2.5. Concluding, we are able to provide adaptive error control based algorithms for both $h - p$ and derivative restriction generalized versions of DG methods.

The rest of the paper is organized as follows: In section 3 we prove Theorem 2.5. The proof is based on an abstract Kruzkov estimate for approximations of the entropy solution of the conservation law (Theorem 3.3) and on a weak cell entropy inequality for the method (Lemma 3.4). In section 4 we present the fully discrete generalized RK-DG method and its continuous in time form with the help of the "continuous extension" for Runge–Kutta schemes. In section 5 we present the limiting projections and the adaptive error control-based algorithms for the corresponding DG methods. In section 6 we discuss the computational performance of the various methods in several test cases.

3. Proof of the a posteriori error estimate. In this section we establish an error estimate for approximations of conservation laws. It is an extension to smooth entropies of the corresponding results in [24, 23, 4]. The notation and the form of the

result follows [23, Lemma 4.1]. We start with the definition of the entropy residual.

DEFINITION 3.1 (entropy residual R_S). *Let $\tilde{u} \in L^\infty(\mathbb{R}^d \times \mathbb{R}^+)$ be an arbitrary function. Then, corresponding to the definition of an entropy weak solution, we define the entropy residual R_S by*

$$\langle R_S(\tilde{u}), \phi \rangle := \iint\limits_{\mathbb{R}^d \times \mathbb{R}^+} S(\tilde{u})\partial_t \phi + F_S(\tilde{u}) \cdot \nabla \phi + \int_{\mathbb{R}^d} S(u_0)\phi(\cdot, 0). \tag{3.1}$$

For the error estimate we require a regularization of the Kruzkov entropy $|v-k|$.

DEFINITION 3.2 (δ-regularized Kruzkov entropy). *Let $\bar{S} \in C^2(\mathbb{R}, \mathbb{R}^+)$ be given as $\bar{S}(v) = (6v^2 - v^4)/8$ if $|v| \le 1$ and $\bar{S}(v) = |v| - 3/8$ otherwise. For any $\delta > 0$, $v \in \mathbb{R}$ let us define $S_\delta : \mathbb{R} \to \mathbb{R}^+$ by $S_\delta(v) := \delta \bar{S}(\frac{v}{\delta})$. Furthermore, define $F_{S,\delta} : \mathbb{R}^2 \to \mathbb{R}$ for any $v, k \in \mathbb{R}$ by $F_{S,\delta}(v,k) := \int_\kappa^v f'(w) S'_\delta(w-k)dw$.*

In the following result, u_h stands for any approximation of problem (1.1)–(1.2).

THEOREM 3.3 (abstract Kruzkov estimate). *Let $u_h, u \in L^\infty_{loc}([0,\infty), L^1_{loc}(\mathbb{R}^d))$ be right continuous in t, with values in $L^1_{loc}(\mathbb{R}^d)$. Assume that u is the entropy solution of (1.1)–(1.2). Let $S(v) = S(v-k) = S_\delta(v-k)$ be the δ-regularized Kruzkov entropy and $F_S(v) = F_S(v\,,\,k) = F_{S,\delta}(v\,,\,k)$ the corresponding entropy flux. Let $\Psi \ge 0$ be a test function $\Psi \in C_c^\infty((0,\infty) \times \mathbb{R}^d)$, and assume that u_h satisfies*

$$\begin{aligned} -\langle R_S(u_h), \Psi \rangle &= -\iint_{(0,\infty)\times\mathbb{R}^d} \big(S(u_h - k)\partial_t \Psi + F_S(u_h\,,\,k) \cdot \nabla_x \Psi\big) dtdx \\ &\le \iint_{(0,\infty)\times\mathbb{R}^d} \left(\beta_O B_O(\Psi) + \alpha_G |\partial_t \Psi| + \sum_j \beta_H^j B_H^j \left(\frac{\partial \Psi}{\partial x_j} \right) \right) dxdt \quad \text{for all } k \in \mathbb{R}, \end{aligned}$$

where $\alpha_G, \beta_O, \beta_H^j$ are nonnegative k-independent but possibly δ-dependent *functions in $L^1_{loc}([0,\infty) \times \mathbb{R}^d)$ and $\alpha_G \in L^\infty_{loc}([0,\infty), L^1_{loc}(\mathbb{R}^d))$.*

For fixed $\Delta, \delta > 0$, let $\mathcal{T}_h = \{K\}$ be a given element decomposition of $[0,\infty) \times \mathbb{R}^d$ into elements K such that $\operatorname{diam}(K_t) \le \Delta$ in the case where B_O or B_H^j is not identically zero; here $K_t = \{x : (t,x) \in K\}$. If, in addition, for all $(t,x) \in K$, $1 \le i, j \le d$,

$$|B_O(\Psi)(t,x)| \le C \sup_{x' \in K_t} |\Psi(t,x')|\,, \quad \left| B_H^j \left(\frac{\partial \Psi}{\partial x_j} \right)(t,x) \right| \le C \sup_{x' \in K_t} \left| \frac{\partial \Psi}{\partial x_j}(t,x') \right|\,, \tag{3.2}$$

where C is a constant independent of Ψ and the decomposition $\mathcal{T}_h$, then the following estimate holds: for any $T_{\max} \ge 0$, $x_0 \in \mathbb{R}^d$, $R > 0$, $\rho > 0$ with $M = Lip(f)$, we have

$$\begin{aligned} \int_{|x-x_0|<R} |u_h(T_{\max}, x) - u(T_{\max}, x)| dx &\le \int_{B_0} |u_h(0,x) - u(0,x)| dx \\ &+ C\,(M\,+1) TV(u^0)\;\Delta + C\{k_1 TV(u^0) + k_0\,\chi_{\operatorname{supp}(u_h-u)(T_{\max})}\,(R+\Delta)^d\,\}\,\delta \\ &+ C\left(1 + \frac{T_{\max}(1+M)}{\Delta}\right) \sup_{0\le t\le T_{\max}+\rho} \int_{B_t} \alpha_G(t,x) dx \\ &+ C \iint_{0\le t\le T_{\max}\; x\in B_t^\Delta} \left(\beta_O(t,x) + \frac{1}{\Delta} \sum_{j=1}^d \beta_H^j(t,x) \right) dxdt\,, \end{aligned}$$

where $B_t = B(x_0, R + M(T_{\max} - t) + \Delta)$, $B_t^{\Delta} = B(x_0, R + M(T_{\max} - t) + 2\Delta)$, *and* χ_D *denotes the characteristic function of the set* D.

Proof. The proof follows the lines of [4, 24], where special attention has to be paid to the treatment of the smooth approximation S_δ of the Kruzkov entropies. This is done analogously to [6] and [28]. For a detailed version of the proof, we refer to [13]. ∎

3.1. Estimate on the entropy residual. To apply the abstract theorem of the previous subsection we need to estimate $\langle R_S(\tilde{u}_h), \phi \rangle$ for ϕ being a test function and $\tilde{u}_h$ defined in (2.12). This will be done in the following lemmas.

LEMMA 3.4 (weak cell entropy inequality). *Let* (S, F_S) *denote a smooth entropy pair. Then the following cell entropy inequality holds for* $\widetilde{u}_h$*:*

$$I_j^n := I_{1,j}^n + I_{2,j}^n + I_{3,j}^n + I_{4,j}^n = -D_j^n \le 0, \tag{3.3}$$

where, for $\phi \in C_0^1(\mathbb{R}^d \times \mathbb{R}^+, \mathbb{R}^+)$,

$$\begin{aligned}
I_{1,j}^n &= (\partial_t S(\widetilde{u}_j^n) + \nabla \cdot F_S(\widetilde{u}_j^n), \overline{\phi_j})_{T_j},\\
I_{2,j}^n &= \sum_{l \in N(j)} \Big(F_{jl}(\widetilde{u}_j^n, \widetilde{u}_l^n) - F_S(\widetilde{u}_j^n) \cdot \mathbf{n}_{jl}, \overline{\phi_j} \Big)_{S_{jl}},\\
I_{3,j}^n &= (\partial_t \widetilde{u}_j^n + \nabla \cdot f(\widetilde{u}_j^n), (S'(\overline{\widetilde{u}_j^n}) - S'(\widetilde{u}_j^n))\overline{\phi_j})_{T_j},\\
I_{4,j}^n &= \sum_{l \in N(j)} \Big(f_{jl}(\widetilde{u}_j^n, \widetilde{u}_l^n) - f(\widetilde{u}_j^n) \cdot \mathbf{n}_{jl}, (S'(\overline{\widetilde{u}_j^n}) - S'(\widetilde{u}_j^n))\overline{\phi_j} \Big)_{S_{jl}},\\
D_j^n &= \sum_{l \in N(j)} \left(\int_{\widetilde{u}_j^n}^{\widetilde{u}_l^n} \partial_w f_{jl}(\widetilde{u}_j^n, w) \int_w^{\widetilde{u}_j^n} S''(s)\, ds\, dw, \overline{\phi_j} \right)_{S_{jl}}.
\end{aligned}$$

Here $F_{jl}(\alpha, \beta) = \int_\alpha^\beta \partial_s f_{jl}(\alpha, s) S'(s) ds + F_S(\alpha)$ *is a discrete entropy flux that is consistent with* F_S.

Proof. Let $p \ge 0$. We start by choosing $v_h = S'(\overline{\widetilde{u}_h^n})\overline{\phi_h}$ in the local form, the scheme (2.11). This yields

$$(\partial_t u_j^n, S'(\overline{\widetilde{u}_j^n})\overline{\phi_j})_{T_j} + \sum_{l \in N(j)} (f_{jl}(\widetilde{u}_j^n, \widetilde{u}_l^n), S'(\overline{\widetilde{u}_j^n})\overline{\phi_j})_{S_{jl}} = 0.$$

Next, (2.7) implies $(\partial_t u_j^n, v_j)_{T_j} = (\partial_t \widetilde{u}_j^n, v_j)_{T_j}$ for all $v_h \in V_{h,n}^0$. Therefore

$$(\partial_t \widetilde{u}_j^n, S'(\overline{\widetilde{u}_j^n})\overline{\phi_j})_{T_j} + \sum_{l \in N(j)} (f_{jl}(\widetilde{u}_j^n, \widetilde{u}_l^n), S'(\overline{\widetilde{u}_j^n})\overline{\phi_j})_{S_{jl}} = 0.$$

We insert zeros to get

$$\begin{aligned}
0 = {} & (\partial_t \widetilde{u}_j^n + \nabla \cdot f(\widetilde{u}_j^n), S'(\widetilde{u}_j^n)\overline{\phi_j})_{T_j} + \sum_{l \in N(j)} \Big(f_{jl}(\widetilde{u}_j^n, \widetilde{u}_l^n) - f(\widetilde{u}_j^n) \cdot \mathbf{n}_{jl}, S'(\widetilde{u}_j^n)\overline{\phi_j} \Big)_{S_{jl}}\\
& + (\partial_t \widetilde{u}_j^n + \nabla \cdot f(\widetilde{u}_j^n), (S'(\overline{\widetilde{u}_j^n}) - S'(\widetilde{u}_j^n))\overline{\phi_j})_{T_j}\\
& + \sum_{l \in N(j)} \Big(f_{jl}(\widetilde{u}_j^n, \widetilde{u}_l^n) - f(\widetilde{u}_j^n) \cdot \mathbf{n}_{jl}, (S'(\overline{\widetilde{u}_j^n}) - S'(\widetilde{u}_j^n))\overline{\phi_j} \Big)_{S_{jl}}.
\end{aligned}$$

We complete the proof with the following identity:

$$
\begin{aligned}
&(f_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n)-f(\widetilde{u}_j^n)\cdot\mathbf{n}_{jl})S'(\widetilde{u}_j^n)=(f_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n)-f_{jl}(\widetilde{u}_j^n,\widetilde{u}_j^n))S'(\widetilde{u}_j^n)\\
&\quad=\int_{\widetilde{u}_j^n}^{\widetilde{u}_l^n}\partial_w f_{jl}(\widetilde{u}_j^n,w)S'(w)dw+\int_{\widetilde{u}_j^n}^{\widetilde{u}_l^n}\partial_w f_{jl}(\widetilde{u}_j^n,w)\big(S'(\widetilde{u}_j^n)-S'(w)\big)\mathrm{dw}\\
&\quad=F_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n)-F_{jl}(\widetilde{u}_j^n,\widetilde{u}_j^n)+\int_{\widetilde{u}_j^n}^{\widetilde{u}_l^n}\partial_w f_{jl}(\widetilde{u}_j^n,w)\int_w^{\widetilde{u}_j^n}S''(s)\mathrm{dsdw}\,. \qquad \square
\end{aligned}
$$

REMARK 3.5. (1) *Note that the dissipation term D_j^n in the cell entropy inequality* (3.3) *is positive because of the monotonicity of the numerical flux and the convexity of S.* (2) *If $p=0$, we have $I_{3,j}^n, I_{4,j}^n=0$.* (3) *As expected for a high-order scheme, the weak cell entropy inequality is, in general, not a real cell entropy inequality in the classical sense. However, its use is important to conclude the following estimates; compare to* [16].

LEMMA 3.6 (entropy residual for the semidiscrete DG approximation). *Let $u_h, \widetilde{u}_h$ as before. Then the following estimate holds true for all $\phi\in C_0^1(\mathbb{R}^d\times(0,T_{\max}),\mathbb{R}^+)$:*

$$
\langle R_S(u_h),\phi\rangle \geq T_1+T_2+T_3+T_4+T_5+T_6, \tag{3.4}
$$

where

$$
\begin{aligned}
T_1&:=\iint_{\mathbb{R}^d\times\mathbb{R}^+}\Big(\partial_t S(\widetilde{u}_h)+\nabla\cdot F_S(\widetilde{u}_h)\Big)\Big(\overline{\phi_h}-\phi\Big),\\
T_2&:=\int_0^T\sum_{j\in J^n}\sum_{l\in N(j)}\Big(F_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n)-F_S(\widetilde{u}_j^n),\overline{\phi_h}-\phi\Big)_{S_{jl}},\\
T_3&:=\sum_n\int_{t^n}^{t^{n+1}}\sum_{j\in J^n}\Big(\partial_t\widetilde{u}_j^n+\nabla\cdot f(\widetilde{u}_j^n),(S'(\overline{\widetilde{u}_j^n})-S'(\widetilde{u}_j^n))\overline{\phi_j}\Big)_{T_j},\\
T_4&:=\sum_n\int_{t^n}^{t^{n+1}}\sum_{j\in J^n}\sum_{l\in N(j)}\Big(f_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n)-f(\widetilde{u}_j^n)\cdot\mathbf{n}_{jl},(S'(\overline{\widetilde{u}_j^n})-S'(\widetilde{u}_j^n))\overline{\phi_j}\Big)_{S_{jl}},\\
T_5&:=\sum_n\sum_{j\in J^n}\Big(\widetilde{u}_j^n(t^n)-\widetilde{u}_j^{n-1}(t^n),\int_0^1 S'(u_j(\theta))d\theta\ (\overline{\phi_h}(t^n)-\phi(t^n))\Big)_{T_j},\\
T_6&:=\sum_n\sum_{j\in J^n}\Big(\widetilde{u}_j^n(t^n)-\widetilde{u}_j^{n-1}(t^n),(S'(\overline{\widetilde{u}_j^{n-1}(t^n)})-\int_0^1 S'(v^n(\theta))d\theta)\ \overline{\phi_h}(t^n)\Big)_{T_j},
\end{aligned}
$$

where in the definition of T_5 and T_6 we use the abbreviation

$$
v^n(\theta):=\widetilde{u}^{n-1}(t^n)+\theta(\widetilde{u}^n(t^n)-\widetilde{u}^{n-1}(t^n)).
$$

Note that T_1 is the element residual, and T_2 is the jump residual in space. T_3 and T_4 are to be seen as kinds of stability errors coming from the higher-order approximation, and T_5, T_6 account for possible discontinuities in time of the projected function $\widetilde{u}_h$.

Proof. A summation of the cell entropy-like inequality (3.3) on all elements $T_j\in\mathcal{T}_n$ and an integration in time leads to

$$
I_h:=\sum_n\int_{t^n}^{t^{n+1}}\sum_{j\in J^n}\Big(I_{1,j}^n+I_{2,j}^n+I_{3,j}^n+I_{4,j}^n\Big)\leq 0.
$$

Next, let us look at the entropy residual. Using integration by parts in time and locally in space, we get

$$\langle R_S(\widetilde{u}_h),\phi\rangle = -\sum_n \int_{t^n}^{t^{n+1}} \sum_{j\in J^n} \left[\int_{T_j} \partial_t S(\widetilde{u}_j^n)\phi + \nabla\cdot F_S(\widetilde{u}_j^n)\phi + \sum_{l\in N(j)} \int_{S_{jl}} F(\widetilde{u}_j^n)\cdot \mathbf{n}_{jl}\phi \right] + \sum_n \sum_{j\in J^n} \int_{T_j} (S(\widetilde{u}_j^n)(t^{n+1})\phi(t^{n+1}) - S(\widetilde{u}_j^n)(t^n))\phi(t^n)) \,.$$

Due to the conservation property of the numerical flux and since ϕ is continuous we have $\int_0^T \sum_{j\in J^n}\sum_{l\in N(j)} \int_{S_{jl}} F_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n)\phi = 0$. Thus, by rearranging the summation, we get

$$\begin{aligned}\langle R_S(\widetilde{u}_h),\phi\rangle = &-\sum_n \int_{t^n}^{t^{n+1}} \sum_{j\in J^n} \Big(\partial_t S(\widetilde{u}_j^n) + \nabla\cdot F_S(\widetilde{u}_j^n),\phi\Big)_{T_j} \\ &+ \sum_n \int_{t^n}^{t^{n+1}} \sum_{j\in J^n}\sum_{l\in N(j)} \Big(F_{jl}(\widetilde{u}_j^n,\widetilde{u}_l^n) - F(\widetilde{u}_j^n)\cdot\mathbf{n}_{jl},\phi\Big)_{S_{jl}} \\ &- \sum_n \sum_{j\in J^n} \Big(S(\widetilde{u}^n)(t^n) - S(\widetilde{u}^{n-1})(t^n),\phi(t^n)\Big)_{T_j}\,. \end{aligned} \tag{3.5}$$

Note that, regarding the last term, since the above sums are reduced to integrals over the spatial domain, we have decided to split them into sums over $T_j \in \mathcal{T}_n$, although $\widetilde{u}^{n-1}(t^n) \in V^p_{h,n-1}$. Next, using the property (2.8) of the projections, we obtain

$$\sum_n \sum_{j\in J^n} \Big(\widetilde{u}_j^n(t^n) - \widetilde{u}_j^{n-1}(t^n), S'(\overline{\widetilde{u}_j^{n-1}(t^n)})\overline{\phi_h}(t^n)\Big)_{T_j} = 0.$$

Using the definition of $v^n(\theta)$, we rewrite the last summand (3.5) of the residual as follows:

$$\begin{aligned} &-\sum_n \sum_{j\in J^n} \Big(S(\widetilde{u}^n)(t^n) - S(\widetilde{u}^{n-1})(t^n),\phi(t^n)\Big)_{T_j} \\ &= -\sum_n \sum_{j\in J^n} \Big(S(\widetilde{u}^n)(t^n) - S(\widetilde{u}^{n-1})(t^n),\phi(t^n)\Big)_{T_j} \\ &\quad + \sum_n \sum_{j\in J^n} \Big((\widetilde{u}^n - \widetilde{u}^{n-1})(t^n), S'(\overline{\widetilde{u}^{n-1}(t^n)})\overline{\phi_h}(t^n)\Big)_{T_j} \\ &= \sum_n \sum_{j\in J^n} \left((\widetilde{u}^n - \widetilde{u}^{n-1})(t^n), S'(\overline{\widetilde{u}^{n-1}(t^n)})\overline{\phi_h}(t^n) - \int_0^1 S'(v^n(\theta))\mathrm{d}\theta\phi(t^n)\right)_{T_j} \\ &= \sum_n \sum_{j\in J^n} \left((\widetilde{u}^n - \widetilde{u}^{n-1})(t^n), \int_0^1 S'(v^n(\theta))\mathrm{d}\theta\ (\overline{\phi_h}(t^n) - \phi(t^n))\right)_{T_j} \\ &\quad + \sum_n \sum_{j\in J^n} \left((\widetilde{u}^n - \widetilde{u}^{n-1})(t^n), \Big(S'(\overline{\widetilde{u}^{n-1}(t^n)}\Big) - \int_0^1 S'(v^n(\theta))\mathrm{d}\theta\ \overline{\phi_h}(t^n)\right)_{T_j}. \end{aligned}$$

Finally by comparing the residual with I_h, we arrive at the final result, noticing

$$\langle R_S(u_h),\phi\rangle \geq \langle R_S(u_h),\phi\rangle + I_h = T_1 + T_2 + T_3 + T_4 + T_5 + T_6. \qquad \square$$

We conclude by further estimating the T_i terms in (3.4).

LEMMA 3.7 (estimate on the residual). *The following estimates on the contributions to the residual hold true:*

$$|T_1 + T_2 + T_5| \leq ||S'||_{L^\infty} ||\nabla\phi||_{L^\infty} \sum_n \sum_{j \in J^n} \eta_{1,j}^n,$$
$$|T_3 + T_4 + T_6| \leq ||S''||_{L^\infty} ||\phi||_{L^\infty} \sum_n \sum_{j \in J^n} \eta_{2,j}^n,$$

where the local error indicators $\eta_{i,j}^n$ are defined in Theorem 2.5 above.

Proof. The goal of estimating the terms T_i is to get some power of the mesh size h from the differences in the test functions. On the other hand, every power of h that we might gain must be paid for by a higher derivative of either ϕ or S. Since we will later choose ϕ to approximate certain δ-functions and S to approximate the nonsmooth Kruzkov entropies, derivatives of ϕ and S' will blow up with a certain rate, depending on the corresponding approximation parameters. The goal is therefore to restrict to first derivatives of ϕ and second derivatives of S.

Estimate on T_1. To estimate this term, we need a local estimate on the difference of the test functions. As $\Pi_{V_h^0}$ is exact on polynomials of degree $p = 0$, we get $|(\Pi_{V_h^0}(\phi)(x) - \phi(x))|_{T_j}| \leq h_j ||\nabla\phi||_{L^\infty(T_j)}$, which finally leads to the estimate on T_1.

Estimate on T_2. We get, by rearranging the summation in space,

$$T_2 = \int_0^T \sum_{(j,l) \in \mathcal{E}^n} \int_{S_{jl}} (F_{jl}(\widetilde{u}_j, \widetilde{u}_l) - F_{jl}(\widetilde{u}_j, \widetilde{u}_j))(\overline{\phi_j} - \phi) - (F_{jl}(\widetilde{u}_j, \widetilde{u}_l) - F_{jl}(\widetilde{u}_l, \widetilde{u}_l))(\overline{\phi_l} - \phi).$$

From here we get the estimate

$$|T_2| \leq ||S'||_{L^\infty} \sum_n \int_{t^n}^{t^{n+1}} \sum_{(j,l) \in \mathcal{E}^n} h_{jl} ||\nabla\phi||_{L^\infty(T_j \cup T_l)} \int_{S_{jl}} Q_{jl}(\widetilde{u}_j, \widetilde{u}_l) |\widetilde{u}_j - \widetilde{u}_l|$$
$$\leq ||S'||_{L^\infty} ||\nabla\phi||_{L^\infty} \sum_n \sum_{j \in J^n} \frac{1}{2} \int_{t^n}^{t^{n+1}} \sum_{l \in N(j)} h_{jl} \int_{S_{jl}} Q_{jl}(\widetilde{u}_j, \widetilde{u}_l) |\widetilde{u}_j - \widetilde{u}_l|,$$

where we have used the monotonicity of the numerical fluxes f_{jl}.

Analogous to the estimates for T_1, T_2, we derive for T_3 and T_4

$$|T_3| \leq ||S''||_{L^\infty} ||\phi||_{L^\infty} \sum_n \sum_{j \in J^n} \int_{t^n}^{t^{n+1}} ||\overline{\widetilde{u}_j^n} - \widetilde{u}_j^n||_{L^\infty(T_j)} \int_{T_j} |\partial_t \widetilde{u}_h + \nabla \cdot f(\widetilde{u}_h)|,$$

$$|T_4| \leq ||S''||_{L^\infty} ||\phi||_{L^\infty} \sum_n \sum_{j \in J^n} \frac{1}{2} \int_{t^n}^{t^{n+1}} \sum_{l \in N(j)} \max_{k \in \{j,l\}} ||\overline{\widetilde{u}_k^n} - \widetilde{u}_k^n||_{L^\infty(S_{jl})} \cdot \int_{S_{jl}} Q_{jl}(\widetilde{u}_j, \widetilde{u}_l) |\widetilde{u}_j - \widetilde{u}_l|.$$

Estimate on T_5 and T_6.

$$|T_5| \le ||S'||_{L^\infty} ||\nabla\phi(t^n)||_{L^\infty(T_j)} \sum_n \sum_{j\in J^n} h_j \int_{T_j} |\widetilde{u}^n(t^n) - \widetilde{u}^{n-1}(t^n)|,$$

$$|T_6| \le ||S''||_{L^\infty} ||\phi(t^{n+1})||_{L^\infty(T_j)} \cdot \sum_n \sum_{j\in J^n} ||\overline{\widetilde{u}_j^{n-1}}(t^n) - \widetilde{u}_j^{n-1}(t^n)||_{L^\infty(T_j)} \int_{T_j} |\widetilde{u}^n(t^n) - \widetilde{u}^{n-1}(t^n)|.$$

For the last estimate we used

$$\left| S'(\overline{\widetilde{u}_j^{n-1}(t^n)}) - \int_0^1 S'(v^n(\theta))\mathrm{d}\theta \right| \le ||S''||_{L^\infty} \int_0^1 |\overline{\widetilde{u}^{n-1}(t^n)} - \widetilde{u}^{n-1}(t^n) + \theta(\widetilde{u}^n(t^n) - \widetilde{u}^{n-1}(t^n))|\mathrm{d}\theta,$$

which gives us the bound on T_6, since the function under the integral is monotone decreasing in θ due to (2.9). The estimate of the theorem now follows by introducing the notation from Theorem 2.5. ∎

We are ready now to complete the proof of Theorem 2.5.

Proof of Theorem 2.5. Lemma 3.7 shows that $\widetilde{u}_h$ satisfies the assumption of Theorem 3.3 with $\alpha_G := 0$ and $\beta_O, \sum_k \beta_H^k$ given by the following local contributions:

$$\beta_O|_{T_j\times[t^n,t^{n+1})} := \frac{1}{\Delta t\ |T_j|} ||S''||_{L^\infty} \eta_{2,j}^n, \qquad \sum_k \beta_H^k|_{T_j\times[t^n,t^{n+1})} := \frac{1}{\Delta t\ |T_j|}\ \eta_{1,j}^n.$$

Using the definition of the entropy $S = S_\delta$ (see Definition 3.2), we estimate $||S''||_{L^\infty} \le K_S \frac{1}{\delta}$. Theorem 2.5 now follows from Theorem 3.3 by choosing the regularization parameters Δ, δ as

$$\Delta := \sqrt{\frac{\sum_n \sum_{j\in J^n} \eta_{1,j}^n}{K_1}}, \qquad \delta := \sqrt{\frac{\sum_n \sum_{j\in J^n} \eta_{2,j}^n}{K_2}}, \tag{3.6}$$

where $K_1 := (M+1)TV(u^0)$, $K_2 := K_S^{-1}(k_1 TV(u^0) + k_0\, \chi_{\mathrm{supp}(u_h-u)(T_{\max})}\, (R+1)^d)$. ∎

4. Fully discrete RK-DG method and continuous in time extension. In this section we will present the generalized class of RK-DG methods that result from time discretization of the semidiscrete method of Definition 2.4 (see also [11]). Thus let us suppose that we can write the semidiscrete DG method as a system of ODEs for a vector valued function $U : (0, T_{\max}) \to \mathbb{R}^N$, where N corresponds to the degrees of freedom of u_h. Then (2.2) can be written in the general form

$$\frac{\mathrm{d}}{\mathrm{d}t} U(t) = L(U(t), t). \tag{4.1}$$

A general explicit m-stage Runge–Kutta method for (4.1) can be represented as

$$W^l := U^n + \Delta t \sum_{k=1}^{l-1} a_{lk} L^k, \quad L^l := L(W^l, t^n + c_l \Delta t), \quad l = 1, \dots, m, \tag{4.2}$$

$$U^{n+1} := U^n + \Delta t \sum_{k=1}^{m} b_k L^k. \tag{4.3}$$

To ensure consistency, the additional constraints $\sum_{k=1}^{m} b_k = 1$, $c_l = \sum_{k=1}^{l-1} a_{lk} \in [0,1]$ have to be imposed. The scheme is characterized by the values b_k, $k = 1, \dots, m$, and a lower triangular matrix a_{lk}, $l = 2, \dots, m$, $k < l$. For particular strongly stability preserving Runge–Kutta methods we refer to [32] and to the review articles [17, 31].

The Runge–Kutta method, as presented above, gives only approximations at the discrete time steps t^n. In order to obtain a continuous approximation in time, we seek a polynomial approximation U_h in time, such that in each interval $[t^n, t^{n+1}]$ the Runge–Kutta scheme can be written in the form

$$\frac{\mathrm{d}}{\mathrm{dt}} U_h(t) = L_h(U_h(t), t).$$

A way to construct such polynomials is given by the so-called natural continuous extension (NCE) of Runge–Kutta methods, introduced by Zennaro [34]. The main result of [34] can be summarized as follows. Each m-stage Runge–Kutta method of order $\tilde{m}$ has a natural continuous extension U_h of polynomial degree $\tilde{p}$ with $\frac{\tilde{m}+1}{2} \leq \tilde{p} \leq \min\{m^*, \tilde{m}\}$, where m^* is the number of distinct values of the coefficients c_l, in the sense that there exist m polynomials $b_l \in \mathbb{P}^{\tilde{p}}(0,1)$, $l = 1, \dots, m$, such that

$$\begin{aligned} U_h(t^n) &= U^n, \qquad U_h(t^{n+1}) = U^{n+1}, \\ U_h(t^n + s\Delta t) &:= U^n + \Delta t \sum_{k=1}^{m} b_k(s) L^k, \quad 0 \leq s \leq 1. \end{aligned} \tag{4.4}$$

Let us suppose that U_h is given by the m-stage NCE Runge–Kutta scheme (for an explicit construction, see [34]). From (4.4) we get

$$U_h(t^n) = U^n, \qquad \frac{\mathrm{d}}{\mathrm{dt}} U_h(t) = \sum_{k=1}^{m} b_k' \left(\frac{t - t^n}{\Delta t} \right) L^k, \quad t \in [t^n, t^{n+1}]. \tag{4.5}$$

Defining the discrete operator L_h as $L_h(U_h, t) := \sum_{k=1}^{m} b_k'(\frac{t-t^n}{\Delta t}) L^k$, we have reached our desired goal. The Runge–Kutta method can now be written

$$U_h(t^n) = U^n, \qquad \frac{\mathrm{d}}{\mathrm{dt}} U_h(t) = L_h(U_h, t). \tag{4.6}$$

Let us further define the fully discrete RK-DG method in a form equivalent to (4.6). Starting from Definition 2.4, we first define the local operators L_j^n by

$$\langle L_j^n(u_h(t)), v_j \rangle|_{T_j} := (f(u_j^n(t)), \nabla v_j)_{T_j} - \sum_{l \in N(j)} (f_{jl}(u_j^n(t), u_l^n(t)), v_j)_{S_{jl}} \tag{4.7}$$

for all $T_j \in \mathcal{T}_h$, $n = 0, \dots, N$, $v_h \in V_h^p$, and L_h^n through

$$\langle L^n(u_h(t)), v_h \rangle := \sum_{j \in J^n} \langle L_j^n(u_h(t)), v_j \rangle|_{T_j}.$$

Definition 4.1 (fully discrete generalized RK-DG approximation). *Let an m-stage Runge–Kutta method be given according to* (4.2), (4.3)*, and let us suppose that a projection $\Lambda_h^{n,t}$ with the properties* (2.7), (2.8) *is given. Furthermore, let the natural continuous extension of highest possible degree $\tilde{p}$ be given according to* (4.4)*. Let us denote $\Lambda_h^{n,k} := \Lambda_h^{n,t^n + c_k \Delta t}$ for $k = 1, \dots, m$. U_h is called a generalized fully discrete*

RK-DG approximation of (1.1)–(1.2) *if* $U_h^{-1} := \Lambda_h^{0,0}(u_0)$, *and for* $n = 0, \ldots, N-1$, $U_h^n := U_h|_{(t^n,t^{n+1}]} \in C^1(t^n, t^{n+1}; V_{h,n}^p)$ *is defined through*

$$U_h^n(t^n) := \Lambda_h^{n,t^n}(U_h^{n-1}(t^n)), \tag{4.8}$$

$$(W_j^{n,l}, v_j) = (U_j^n(t^n), v_j) + \Delta t \sum_{k=1}^{l-1} a_{lk} \langle L_j^n(\Lambda_h^{n,k}(W_h^{n,k})), v_j \rangle, \tag{4.9}$$

$$(U_j^n(t), v_j) = (U_j^n(t^n), v_j) + \Delta t \sum_{k=1}^{m} b_k \left(\frac{t-t^n}{\Delta t} \right) \langle L_j^n(\Lambda_h^{n,k}(W_h^{n,k})), v_j \rangle \tag{4.10}$$

$$\text{for all } v_j \in \mathbb{P}_p, j \in J^n, t \in [t^n, t^{n+1}].$$

Thus, with the definition of L_h^n, *the fully discrete generalized RK–DG approximation satisfies on each time slab* (t^n, t^{n+1}) *the ODE*

$$(\partial_t U_h^n(t), v_h) = \sum_{k=1}^{m} b_k' \left(\frac{t-t^n}{\Delta t} \right) \langle L_h^n(\Lambda_h^{n,k}(W_h^{n,k})), v_h \rangle \quad \textit{for all} \quad v_h \in V_{h,n}^p. \tag{4.11}$$

REMARK 4.2. *In our numerical experiments we used polynomial degree* $p = 1, 2, 3$ *for the space discretization combined with the NCE Runge–Kutta method of the same degree. In* [34] *extensions of Runge–Kutta methods are constructed with optimal order up to* $p = 4$, *but for* $p = 3$ *and* $p = 4$ *it is necessary to include a stage reuse procedure to obtain the desired order. In our examples we have included stage reuse since this does not increase the computational cost of the scheme.*

5. Choice of the projections and adaptive strategy. In Definition 2.1 we introduced a class of semidiscrete DG methods for arbitrary limiting projections $\Lambda_h^{n,t}$ and computational grids. In this subsection we describe specific choices of projection operators that are motivated by the a posteriori error estimate (Theorem 2.5). Furthermore, we give an adaptive strategy for local mesh refinement.

The numerical solution in the interval $(t^n, t^{n+1}]$ is defined in the following algorithm. We start with a guess $\tilde{\mathcal{T}}^n$ for the grid and $\tilde{\Lambda}_h^{n,t}$ for the limiting projection.

- **given:** grid $\tilde{\mathcal{T}}^n$, projection $\tilde{\Lambda}_h^{n,t}$ for $t \in [t^n, t^{n+1}]$, and $u^n(t^n, x)$
- **do**
  1. Let $\mathcal{T}^n = \tilde{\mathcal{T}}^n$ and $\Lambda_h^{n,t} = \tilde{\Lambda}_h^{n,t}$ for $t \in [t^n, t^{n+1}]$.
  2. Compute $u^n(t, x)$ for $t \in (t^n, t^{n+1}]$ on $\mathcal{T}^n$ using $\Lambda_h^{n,t}$.
  3. Compute indicators and new limiting projection $\tilde{\Lambda}_h^{n,t}$ on $\mathcal{T}^n$;
     i.e., for $j \in J^n$ compute the following:
     - $\rho_j^n, R_{T,j}^n, R_{S,j}^n$ (cf. Corollary 2.6),
     - $\tilde{\Lambda}_h^{n,t}$ for $t \in (t^n, t^{n+1}]$ (cf. section 5.1),
     - $\tilde{R}_{\Lambda,j}^n := \int_{T_j} |\widetilde{u}_j^{n+1} - \tilde{\Lambda}_h^{n,t^{n+1}} u_j^n(t^{n+1})|$.
  4. Compute error indicator for interval $(t^n, t^{n+1}]$ on $\mathcal{T}^n$:
     $R^n := 2 \sum_{j \in J^n} \rho_j^n (R_{T,j}^n + R_{S,j}^n + \tilde{R}_{\Lambda,j}^n)$.
  5. Refine grid $\mathcal{T}^n \to \tilde{\mathcal{T}}^n$, and project $\tilde{\Lambda}_h^{n,t}$ for $t \in (t^n, t^{n+1}]$ onto $\tilde{\mathcal{T}}^n$.
- **while** $R^n > \mathrm{TOL}^n$
- define $\tilde{\mathcal{T}}^{n+1}$ by coarsening $\tilde{\mathcal{T}}^n$ so that
  $R^n + 2 \sum_{j \in \tilde{J}^n} \rho_j^n (\int_{T_j} |\tilde{\Lambda}_h^{n,t^{n+1}} u_j^n(t^{n+1}) - \Pi^{n \to n+1} \tilde{\Lambda}_h^{n,t^{n+1}} u_j^n(t^{n+1})|) < \mathrm{TOL}^n$.
- define $\tilde{\Lambda}_h^{n+1,t^{n+1}} := \Pi^{n \to n+1} \tilde{\Lambda}_h^{n,t^{n+1}}$ and $\tilde{\Lambda}_h^{n+1,t}$ for $t \in (t^{n+1}, t^{n+2}]$ on $\tilde{\mathcal{T}}^{n+1}$ using $\tilde{\Lambda}_h^{n,t^n}$.

The algorithm is based on the assumption that

$$\begin{aligned} &K_1 h_j + K_2 \max_{k\in\{j,l\in N(j)\}} ||\overline{\tilde{\Lambda}_h^{n,t} u_k^n} - \tilde{\Lambda}_h^{n,t} u_k^n||_{L^\infty((t^n,t^{n+1})\times T_k)} \\ \le\, &K_1 h_j + K_2 \max_{k\in\{j,l\in N(j)\}} ||\overline{\Lambda_h^{n,t} u_k^n} - \Lambda_h^{n,t} u_k^n||_{L^\infty((t^n,t^{n+1})\times T_k)}. \end{aligned}$$

With the restrictive choice $\tilde{\Lambda}_h^{n,t}(v) = \overline{v}$ we have $\max_k ||\overline{\tilde{\Lambda}_h^{n,t} u_k^n} - \tilde{\Lambda}_h^{n,t} u_k^n||_{L^\infty} = 0$, and therefore $\rho_j^n = K_1 h_j$, as in the first-order case. In fact, with this limiting operator the DG scheme reduces to the first-order finite volume scheme for which the convergence of the error indicator can be shown rigorously for $h \to 0$. Therefore, with a suitable choice of $\mathcal{T}^n$ the iteration (1)–(5) always terminates, and in practice our scheme requires hardly any iterations. Note that from our strategy it follows that

$$\sum_{j\in J^n} \rho_j^n \left(R_{T,j}^n + R_{S,j}^n + R_{\Lambda,j}^n \right) \le \mathrm{TOL}^n.$$

Thus, the error $||(u - \widetilde{u}_h)(T_{\max})||_{L^1(B_R(x_0))}$ is bounded by some prescribed tolerance TOL which satisfies $\sum_n \mathrm{TOL}^n \le \mathrm{TOL}$. This is summarized in the following lemma.

LEMMA 5.1. *Let η_h denote the global error estimator from Theorem* 2.5*, and let a prescribed tolerance TOL be given. If the computational mesh is adapted due to the strategy described above using the methods described in the following subsections, then it follows that $\eta_h \le \mathrm{TOL}$.*

Thus, the adaptive strategy together with Theorem 2.5 yields a rigorous control on the error $||(u - \widetilde{u}_h)(T_{\max})||_{L^1(B_R(x_0))}$.

5.1. Choice of the projection operator in one space dimension. Our algorithm is based on an initial guess for the projection operator, which we denote with $\tilde{\Lambda}_h^{n,t}$. It is used to define the final projection operators $\Lambda_h^{n,t}$ in the generalized RK-DG method of Definition 2.4 or 4.1. We now introduce two different approaches for constructing $\tilde{\Lambda}_h^{n,t}$. The first approach is based on a restriction of the gradients of the approximate solution based on the error estimate in Corollary 2.6. The second approach is a p-adaptive projection where the local polynomial degree of the approximate solution is chosen in accordance with the error indicators in Corollary 2.6. Together with the local mesh adaption strategy that we will discuss in the next subsection, both methods are then used in an hp-adaptive manner. The operator $\tilde{\Lambda}_h^{n,t}$ is always constructed on a fixed mesh $\mathcal{T}^n$ and then prolonged/restricted onto a modified mesh in such a way that refinement of cells does not change the projected function.

The goal of the choice of the projection $\tilde{\Lambda}_h^{n,t}$ is twofold. On the one hand, we need a projection or limiting of the solution in order to stabilize the scheme. On the other hand, the factor ρ_j^n should be of the order of h_j. Together with a reasonable assumption on the boundedness of the residual terms $R_{T,j}^n, R_{S,j}^n, R_{\Lambda,j}^n$, this requirement guarantees the convergence of the error estimate η_h for $h \to 0$. We expect that in regions where the solution u is smooth the stated requirement is met even if we choose $\Lambda_j^{n,t} = id$, whereas near discontinuities the term $||\overline{u_h} - u_h||_{L^\infty}$ grows without bound. The projection should therefore be active only on mesh cells near discontinuities. Thus, we suggest defining a projection parameter λ_h as

$$\lambda_j^n(t) := \tilde{\lambda}_j^n + \frac{t - t^n}{\Delta t^n} \tilde{\lambda}_j^{n+1}, \qquad \tilde{\lambda}_j^n := \frac{h_j}{\left(h_j + \frac{1}{\Delta t^n} \left(R_{T,j}^n + R_{S,j}^n \right) \right)^{\frac{p+2}{p+1}}} \tag{5.1}$$

and ensuring that our projection operators yield a solution with the property

$$||\overline{\widetilde{u}_j^n(\cdot,t)} - \widetilde{u}_j^n(\cdot,t)||_{L^\infty(T_j)} \leq \lambda_j^n(t). \tag{5.2}$$

We expect that the upper bound λ_j^n is of order h_j near discontinuities, whereas it is of order $h_j^{-\frac{1}{p+1}}$ in smooth regions. As the error $||\overline{u_j^n(\cdot,t)} - u_j^n(\cdot,t)||_{L^\infty(T_j)}$ is expected to converge with order h_j in smooth regions and to remain constant near discontinuities, the upper restriction leads to a projection of the solution near discontinuities, and at the same time $||\overline{\widetilde{u}_j^n(\cdot,t)} - \widetilde{u}_j^n(\cdot,t)||_{L^\infty(T_j)}$ would be at least of order $\mathcal{O}(h_j)$. The bound (5.2) dictates how to construct the operator $\tilde{\Lambda}_j^{n,t}$ from a given projection $\Lambda_j^{n,t}$.

In what follows we propose two possible choices for the projection $\tilde{\Lambda}_h^{n,t}$ in one space dimension, which both satisfy the upper bound (5.2) for given limiter function λ_h. The resulting projections fall into the class of moment-limiters as introduced in [3]. In order to define the methods, let φ_l, $l = 0, \ldots, p_{\max}$, denote the orthogonal basis of Legendre polynomials on the cell $T_j := (x_{j-1/2}, x_{j+1/2})$ such that $\varphi_l \in \mathbb{P}_l(T_j)$. We then have $\varphi_0 = 1$ and thus $\overline{u_j^n(\cdot,t)} = u_{j,0}^n(t)$ with the local expansion $u_j^n(x,t) = \sum_{l=0}^{p_{\max}} u_{j,l}^n(t)\varphi_l(x)$.

5.1.1. P-adaptive method in one dimension. Let $1 \leq l^* \leq p_{\max}$ denote the maximal index such that

$$\sum_{l=1}^{l^*} u_{j,l}^n(t^n)\varphi_l(x) \leq \lambda_j^n(t^n) \quad \text{for all } x \in T_j.$$

Then, the p-adaptive projection on the cell T_j is defined through

$$\tilde{\Lambda}_j^{n,t}(u_h(\cdot,t)) := \sum_{l=0}^{l^*} u_{j,l}^n(t)\varphi_l(x). \tag{5.3}$$

5.1.2. Derivative-restriction method in one dimension. For fixed $t \in [t^n, t^{n+1}]$ let $1 \leq l^* \leq p_{\max}$ denote the maximal index such that

$$\sum_{l=1}^{l^*} u_{j,l}^n(t)\varphi_l(x) \leq \lambda_j^n(t) \quad \text{for all } x \in T_j.$$

In contrast to the p-adaptive strategy, we allow that the derivative of degree l^*+1 is not switched off completely but is reduced in such a way that the bound (5.2) still holds:

$$\tilde{\Lambda}_j^{n,t}(u_h(\cdot,t)) := \sum_{l=0}^{l^*} u_{j,l}^n(t)\varphi_l(x) + \widetilde{u}_{j,l^*+1}^n(t)\varphi_{l^*+1}(x), \tag{5.4}$$

where $\widetilde{u}_{j,l^*+1}^n(t)$ is given as

$$\widetilde{u}_{j,l^*+1}^n(t) := \operatorname{sgn}(u_{j,l^*+1}^n(t)) \ \min\left\{ |u_{j,l^*+1}^n(t)|, \lambda_j^n(t) - \left\| \sum_{l=0}^{l^*} u_{j,l}^n(t)\varphi_l \right\|_{L^\infty(T_j)} \right\}.$$

After the refinement of a cell T_j or the coarsening of a set of cells $(T_{j_k})_{k=1}^l$ the operator $\tilde{\Lambda}_h^{n,t}$ has to be modified to operate on the new grid. Both of the choices described above require the definition of λ_j^n on the new grid cells:

- *refinement* $(T_j \to (T_{j_k})_{k=1}^l)$: let $\lambda_{j_k}^n = \frac{1}{l}\lambda_j^n$;
- *coarsening* $((T_{j_k})_{k=1}^l \to T_j)$: let $\lambda_j^n = \sum_{k=0}^l \lambda_{j_k}^n$.

5.2. Adaptive strategy for local mesh refinement. In this subsection we describe an adaptive strategy for local mesh adaptation that is based on an equidistribution strategy of the error indicator η_h of Theorem 2.5. However, there are two significant modifications of the equidistribution strategy when compared with the strategy presented in [25] or [28]. The first modification is that we distribute the error only among those elements that significantly contribute to the error, and secondly we also incorporate the projection error from mesh coarsening into the adaptive strategy. These modifications are of minor importance for smooth solutions but result in quite different adaptive convergence behavior for problems with discontinuities. In detail, the new adaptive strategy is given as follows.

Using the notation of Corollary 2.6, let us define for a prescribed tolerance TOL the local error indicators η_j^n for some given $\Theta \in (0,1)$ as

$$\eta_j^0(M) := \frac{M}{(1-\Theta)\ \mathrm{TOL}}\eta_{0,j}, \qquad \eta_j^n(M) := \frac{2\ T_{\max}\ M}{\Delta t^n\ (\Theta\ \mathrm{TOL}^n)^2}\rho_j^n(R_{T,j}^n + R_{S,j}^n + \tilde{R}_{\Lambda,j}^n),$$

where we again have used the abbreviation $\tilde{R}_{\Lambda,j}^n := \int_{T_j} |\overline{\widetilde{u}_j^{n+1}}(t^{n+1}) - \tilde{\Lambda}_h^{n,t^{n+1}} u_j^n(t^{n+1})|$. The operator $\tilde{\Lambda}_h^{n,t^{n+1}}$ is again a suitable projection operator defined on the mesh $\mathcal{T}_n$ used to stabilize the scheme and to guarantee that ρ_j^n converges for $h_j \to 0$. The adaptive strategy at the time t^{n+1} is then given as follows. For $\alpha \in (0, 0.5)$, $M \in \mathbb{N}$ let us define the set of significant elements as

$$I_s^n(M) := \{j \in I^n |\ \eta_j^n(M) \geq \alpha\},$$

and let M^n implicitly be defined through $M^n = |I_s^n(M^n)|$, where $|\cdot|$ denotes the cardinality of the set. We define ε^n as

$$\varepsilon^n := \sum_{j \in I^n \setminus I_s^n(M^n)} \eta_j^n(1)$$

and suppose that α is chosen small enough to ensure $\varepsilon^n \in (0, 0.5)$. We then define for given $\beta \in (0,1)$ the sets

$$I_r := \{T_j |\ \eta_j^n(M^n) \geq (1-\varepsilon^n)\}, \qquad \tilde{I}_c := \{T_j |\ \eta_j^n(M^n) \leq \beta(1-\varepsilon^n)\}$$

and mark all elements of the set I_r for refinement and those in the set $\tilde{I}_c$ as candidates for coarsening. Coarsening of the mesh leads to an additional projection error of the approximate solution that contributes to the indicator $R_{\Lambda,j}^n$. We split this error into two parts according to $R_{\Lambda,j}^n \leq \tilde{R}_{\Lambda,j}^n + R_{c,j}^n$ with $R_{c,j}^n := \int_{T_j} |\tilde{\Lambda}_h^{n,t^{n+1}} u_j^n(t^{n+1}) - \Pi^{n\to n+1}\tilde{\Lambda}_h^{n,t^{n+1}} u_j^n(t^{n+1})|$, using the operator $\Pi^{n\to n+1}$ to denote the L^2-projection from one grid to another. We calculate the error terms $\eta_{c,j}^n(M^n) := \eta_j^n(M^n) + \frac{2\ T_{\max}\ M}{\Delta t^n\ (\Theta\ \mathrm{TOL}^n)^2}\rho_j^n R_{c,j}^n$ for all $T_j \in \tilde{I}_c$, define the updated set I_c as

$$I_c := \{T_j \in \tilde{I}_c |\ \eta_{c,j}^n(M^n) \leq \beta(1-\varepsilon^n)\},$$

and mark all elements of the set I_c for coarsening. Finally, all elements in the set I_r are refined, until the refinement set I_r is an empty set. Then, all elements of the set I_c are coarsened.

5.3. Evaluation of the semidiscrete error indicators. In Theorem 2.5 we give an a posteriori error estimate for the semidiscrete DG method from Definition 2.4. Conventionally, Runge–Kutta time discretizations that are used in practice provide values only for the approximate solution at the discrete time steps t^n. Thus, our error indicators could not be evaluated continuously in time. In order to give a suitable interpretation of the fully discrete Runge-Kutta solution (see Definition 4.1) we use the natural continuous extension as defined through (4.4). Thus, the approximate fully discrete solution is continuous in time on each time slab $[t^n, t^{n+1}]$, and all contributions of the error indicators $\eta_{i,j}^n$ of Theorem 2.5 are computable.

6. Adaptive numerical experiments in one space dimension. In this section we numerically examine the RK-DG methods defined in Definition 4.1 together with the projections from subsection 5.1 and the local adaptive grid refinement from subsection 5.2. We study the convergence behavior of the estimator η_h from Theorem 2.5, as well as the convergence of the error itself. As test problems, we look at a linear transport problem with smooth and discontinuous regions in the solution. This example is a scalar prototype for contact discontinuities. As a second very challenging example we choose the Buckley–Leverett equation. Here, the flux function is nonconvex, and thus the solution consists of compound waves. For such fluxes there exist several weak solutions that are compatible with a single entropy, but only one of those solutions is the unique entropy solution in the Kruzkov sense. It is well known that higher-order numerical schemes may have difficulties in selecting this unique Kruzkov entropy solution (see also [2] and [26]). In order to compare the efficiency of the selected RK-DG methods we are going to plot the error estimators and errors against the overall number of grid cells $M_{\text{tot}}(\mathcal{T}_h) := \sum_{n=1}^{N} \sum_{T_j \in \mathcal{T}_h^n} 1$. As M_{tot} is available for uniform refined grids, as well as for adaptively refined grids, and as M_{tot} is proportional to the degrees of freedom for fixed polynomial degree p, this is a good way to compare our adaptive method with standard approaches on uniform grids. Furthermore, let us define the experimental order of convergence of a grid-dependent quantity e_h as

$$EOC(e_{H\to h}) := \log\left(\frac{e(\mathcal{T}_H)}{e(\mathcal{T}_h)}\right) \log^{-1}\left(\frac{M_{\text{tot}}(\mathcal{T}_h)}{M_{\text{tot}}(\mathcal{T}_H)}\right). \tag{6.1}$$

Note that a convergence rate $\mathcal{O}(h^p)$ on uniform grids in one space dimension corresponds to $EOC(e_{2h\to h}) = \frac{p}{2}$, as a refinement from grids with cells of size $2h$ to grids with cells of size h leads to two times the number of grid cells per time step and two times the number of time steps. This yields $M_{\text{tot}}(\mathcal{T}_h) = 4M_{\text{tot}}(\mathcal{T}_{2h})$.

6.1. Linear transport equation. As a first numerical example we look at the linear transport equation

$$\partial_t u + a\partial_x u = 0, \qquad u(\cdot, 0) = u_0(\cdot)$$

with the constant transport velocity $a = 2$. For fixed initial data the solution u is then given by $u(x,t) = u_0(x - at)$. We study the setting on $[-1,1] \times [0,2]$ with periodic boundary conditions for the following nonsmooth initial data:

$$u_0(x) := \begin{cases} 1 - (x+1.5)^2, & \text{for } x < -0.5, \\ \sin((x+0.5)\pi), & \text{for } -0.5 \le x < 0.5, \\ 1 - (x-0.5)^2, & \text{for } 0.5 \le x. \end{cases}$$

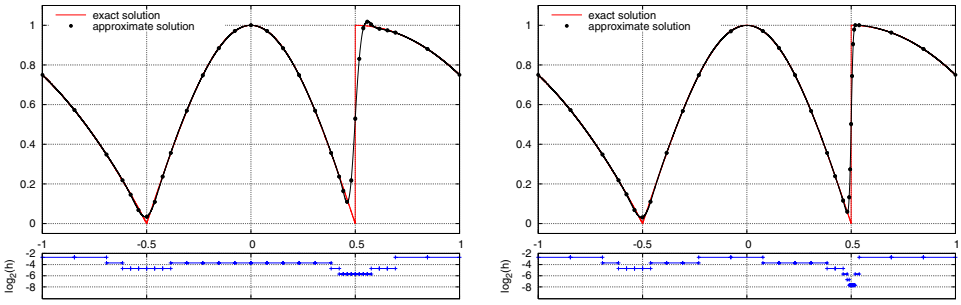


FIG. 6.1. *Comparison of the approximate solutions obtained with the* p-adaptive *method (left-hand side) and the* derivative-restriction *method (right-hand side) on adaptively refined grids with* $p_{\max} = 2$. *For both computations we used the prescribed tolerance* $TOL = 0.5$ *and* $T_{\max} = 2.0$.

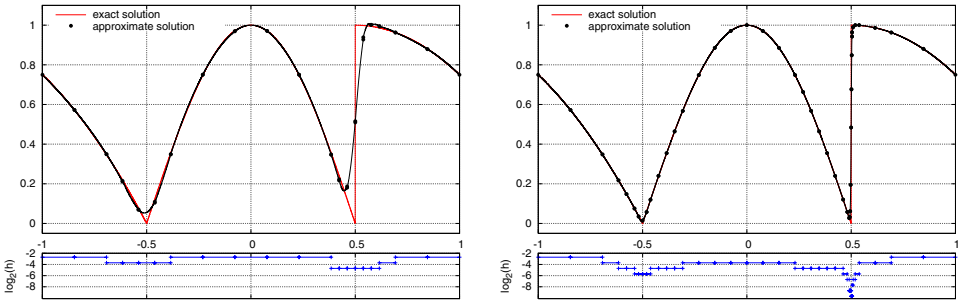


FIG. 6.2. *Comparison of the approximate solutions obtained with the* derivative-restriction *method for* $p_{\max} = 2$ *with* $TOL = 1$ *(right-hand side) and* $TOL = 0.25$ *(left-hand side) on adaptively refined grids at* $T_{\max} = 2.0$.

Since $x - 2a = x - 4$ is equal to x on a $[-1, 1]$ periodic domain it follows that $u(x, 2) = u_0(x)$. We first compare the two projection methods described in subsection 5.1 for $p_{\max} = 2$ (see Figure 6.1). All results are computed with the adaptation strategy from subsection 5.2 with TOL = 0.5. In Figure 6.1 both the exact solution and the approximate solution are shown together with the grid density function.

The comparison of the projection methods for fixed maximal polynomial degree shows that both methods lead to a good resolution of the smooth region as well as of the discontinuity. The *p-adaptive* method (Figure 6.1(left)) produces slight overshoots in front of the discontinuity, but these decrease on finer grids. The refinement strategy together with the *derivative-restriction* method produces a slightly finer grid in the region of the discontinuity, whereas the grid is coarser in the smooth regions.

Results with $p_{\max} = 2$ and the *derivative-restriction* method for different values of TOL are shown in Figure 6.2. It can be clearly seen that the grid is hardly refined in the smooth regions of the solution, whereas the fineness in the region of the discontinuity and also around the kink increases for smaller tolerance values. The coarsest grid level corresponds to a grid with 13 cells. With TOL = 1 only seven cells are added to the final grid—two in the region of the kink and five in the shock region. With TOL = 0.25, 30 cells are added—about 50% of which are located in the shock region.

A comparison of the efficiency of our new method on uniform grids for $p_{\max} = 0, 1, 2$ is shown in Figure 6.3. The increase in efficiency due to an increase of the

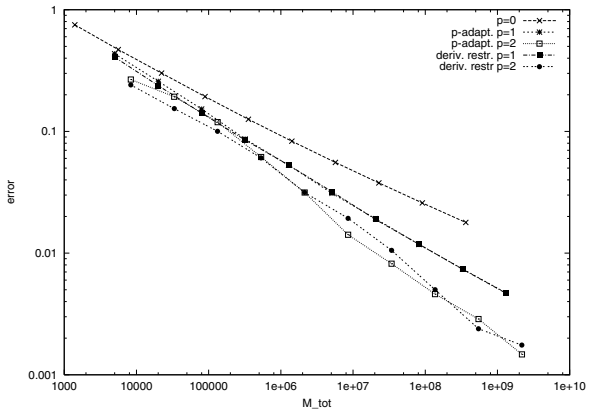


FIG. 6.3. *Convergence study for our new schemes for the linear transport problem on uniformly refined grids.*

TABLE 6.1
Experimental order of convergence for the error and the estimator of the new p-adaptive *method and the* derivative-restriction *method on adaptively refined grids for the linear transport problem.*

	derivative-restriction method		*p-adaptive* method	
$p_{\max}$	$EOC(e_{H\to h})$	$EOC(\eta_{H\to h})$	$EOC(e_{H\to h})$	$EOC(\eta_{H\to h})$
0	0.292	0.193	0.292	0.193
1	0.431	0.228	0.476	0.368
2	0.544	0.342	0.515	0.450

polynomial degree can be clearly seen. Furthermore, the difference between our two projection methods is hardly significant. For $p_{\max} = 1$ there is hardly any difference, and also for $p_{\max} = 2$ there is no clear indication of which method is the more efficient.

Next we compare our adaptive schemes for $p_{\max} = 0, 1, 2$. In Table 6.1 the convergence rates as defined in (6.1) are given for the error $EOC(e_{H\to h})$ and the estimator $EOC(\eta_{H\to h})$. The error and the estimator show better convergence rates for higher polynomial degree. The convergence rates of the error are even better than what we expect to be optimal for discontinuous solutions on uniform computational grids. On uniform grids with mesh size h the optimal rate is supposed to be $h^{\frac{p_{\max}+1}{p_{\max}+2}}$, which corresponds to $EOC(e_{H\to h}) = \frac{1}{2}\frac{p_{\max}+1}{p_{\max}+2}$ (i.e., $0.250, 0.333, 0.375$ for $p_{\max} = 0, 1, 2$). Although the convergence rate of the indicator differs from the convergence rate of the error, the ratio between the prescribed tolerance and the indicator is about constant. In the optimal case this ratio should be close to one. Our adaptive strategy leads to an efficiency index of about $0.5 - 0.8$.

6.2. Buckley–Leverett problem. As a second example we look at the Buckley–Leverett equation, which is a one dimensional model for two-phase flow in porous media where capillary pressure effects are neglected. The unknown variable $u : (-1, 1) \times (0, 0.4) \to \mathbb{R}$ is the saturation of the wetting phase within a two-phase mixture. It satisfies the nonlinear conservation law

$$u_t + \partial_x f(u) = 0 \text{ on } (-1, 1) \times (0, 0.4), \qquad u(\cdot, 0) = u_0 \text{ on } (-1, 1),$$

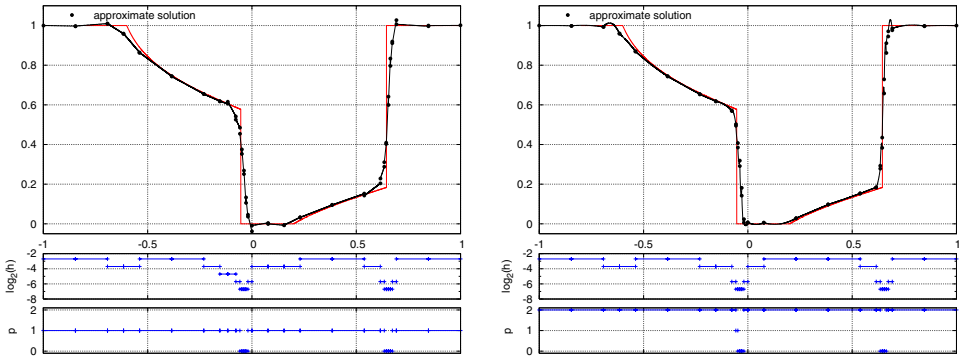


FIG. 6.4. *Comparison of the approximate solutions obtained with the* derivative-restriction *method on adaptively refined grids with* $p_{\max} = 1$ *(left) and* $p_{\max} = 2$ *(right) with* $TOL = 0.5$ *and* $T_{\max} = 0.4$.

where the fractional flow rate f is given as $f(s) = \frac{u^2}{u^2 + \frac{1}{2}(1-u)^2}$. We look at this problem for the following initial data:

$$u_0(x) := \begin{cases} 1 & \text{for } x < -0.6,\ 0.2 \le x, \\ 0 & \text{for } -0.6 \le x < 0.2. \end{cases}$$

Thus, the solution of our Buckley–Leverett problem consists of the solution of two distinct Riemann problems for t smaller than some critical time $T^* > 0.4$. The solution of each Riemann problem is a composed wave consisting of a rarefaction wave and an attached shock.

In Figure 6.4 we plot the exact solution together with the approximation using our adaptive strategy for $p_{\max} = 1, 2$. Since the structure of the solution away from the discontinuities is far simpler than in the advection problem studied above, the advantage of the quadratic ansatz functions is not evident. The grid density function hardly depends on the polynomial degree since almost all grid points are located in the shock regions. Only the kinks at the beginning of the rarefaction waves lead to additional slight refinement. Since the highest grid resolution produced by our refinement strategy is the same for $p_{\max} = 1$ and $p_{\max} = 2$ and the approximation error is dominated by the shocks, the $p_{\max} = 2$ version of the DG method does not lead to a more efficient scheme, as can be seen from Figure 6.5. This must be attributed to the smaller CFL stability restriction required in the higher-order schemes and the resulting smaller time steps. A more complicated structure of the solution—as can be found only in systems in higher space dimension—is required to demonstrate the advantage of an *hp*-adaptive strategy for nonlinear conservation laws with discontinuous solutions.

The results so far show that our adaptive strategy and our projection methods based on the error estimate from Theorem 2.5 lead to good schemes for both linear and nonlinear test problems. We conclude our numerical experiments with results demonstrating the advantage of including the coarsening error in the indicator. Results computed with and without using this "jump" indicator are shown in Figure 6.6. Including the error due to coarsening leads to a higher grid resolution around the kink at the left side of the rarefaction waves. Without this indicator the grid is coarsened to such a degree that the rarefaction wave is not sufficiently resolved and the convergence rate of the adaptive scheme is severely reduced.

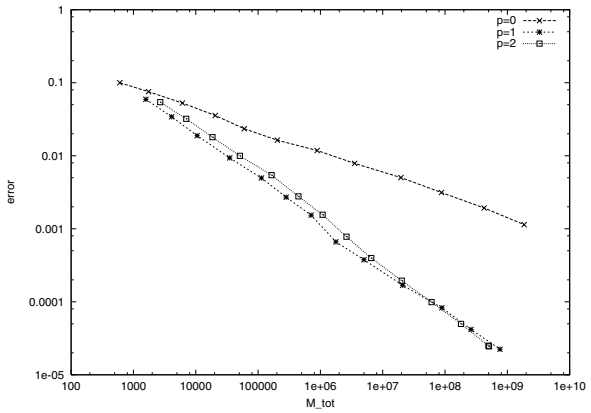


FIG. 6.5. *Convergence study for the* derivative-restriction *approximation of the Buckley–Leverett problem on adaptively refined grids.*

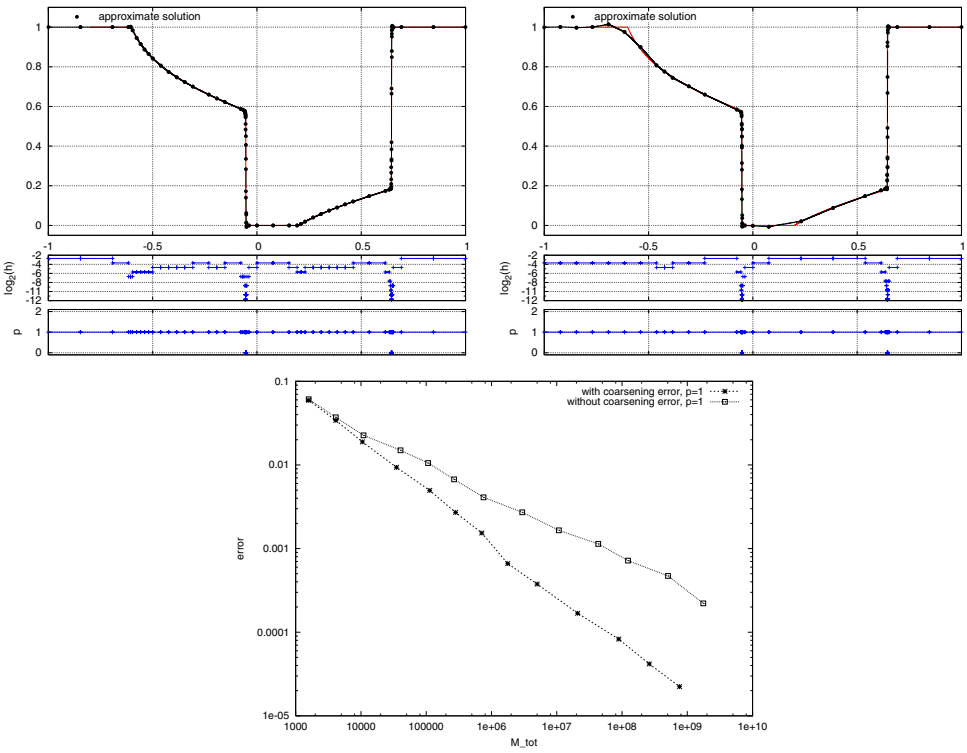


FIG. 6.6. *Comparison of the new adaptive* derivative-restriction *scheme with (top left) and without (top right) incorporated coarsening projection error. In the top row the solutions are plotted at* $T_{\max} = 0.4$ *with* $TOL = 0.125$. *In the bottom figure the error is plotted versus* M_{tot} *for the* derivative-restriction *scheme with and without incorporated coarsening projection error.*

6.3. Local comparison of error indicator and error. In this subsection we compare the local distribution of the error indicator values $(\rho_j^n(R_{T,j}^n + R_{S,j}^n))^{1/2}$, $j \in J^n$, from the a posteriori error estimate in Corollary 2.6 with the distribution of the local L^1-error $||u(t^n) - u_h^n||_{L^1(T_j)}$, $j \in J^n$. We compare the local distribution of

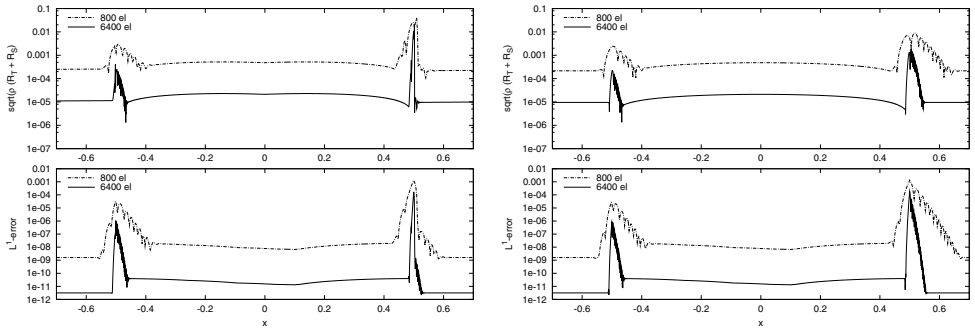


FIG. 6.7. *Comparison of the local indicator values for the limiter of Cockburn and Shu (left) and the p-*adaptive *method (right) for the linear transport problem. The diagrams on the top show the local distribution of the indicator values on uniform grids with* 800 *and* 6400 *elements, while the diagrams at the bottom give the corresponding distribution of the exact local* L^1*-error.*

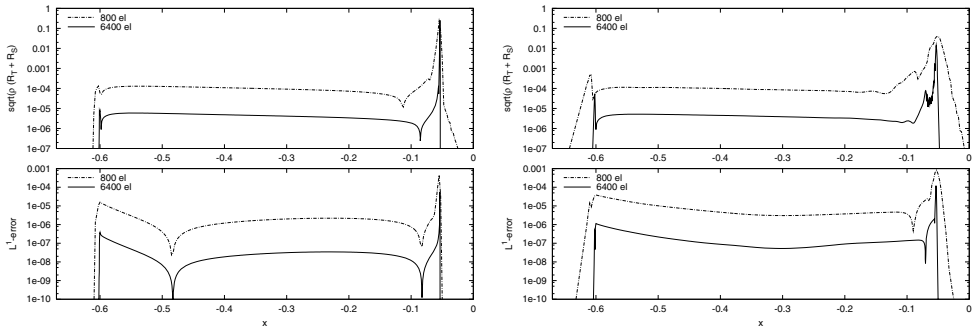


FIG. 6.8. *Comparison of the local indicator values for the limiter of Cockburn and Shu (left) and the p-* adaptive *method (right) for the Buckle–Leverett problem. The diagrams on the top show the local distribution of the indicator values on uniform grids with* 800 *and* 6400 *elements, while the diagrams at the bottom give the corresponding distribution of the exact local* L^1*-error.*

the indicator and exact error for the *p-adaptive* scheme and the method of Cockburn and Shu (see [10, 7]). In Figure 6.7 the comparison is given for the linear transport problem from subsection 6.1 on the interval $(-0.67, 0.67)$, while in Figure 6.8 the comparison for the Buckley–Leverett problem is shown on the interval $(-0.7, 0)$. The underlying computations were done with $p_{\max} = 2$.

From the comparison it can be seen that the local distribution of the error indicator captures very well the behavior of the exact error for both methods. In addition, it can be seen that the error and the indicator decrease with the expected higher-order rate within the smooth regions and show a reduction of the decrease rate near discontinuities or kinks. A more detailed analysis of the convergence behavior of the indicators within the regions of discontinuity reveals that the indicator for the method of Cockburn and Shu does not decrease within the few elements that form the discontinuity, while the indicator for the *p-adaptive* method does decrease significantly. As the local indicator quantities are L^1-quantities, we expect also that the global estimator is not asymptotically reduced for the method of Cockburn and Shu, and thus we have to use the indicators with great care to steer grid adaptivity for this method, as there would be no stopping criteria for the mesh refinement within this shock region. On the other hand, the indicator may be used outside the shock region

without problems and thus could be used in combination with a restriction on the maximal refinement level within the shock regions.

7. Conclusion. We have proved an a posteriori error estimate for a class of semidiscrete discontinuous Galerkin methods on adaptively refined computational meshes (see Theorem 2.5 and section 3). The estimate provides a rigorous error control and is used for the design of stabilizing limiting projection operators (see section 5.1) as well as for the design of a local grid adaptation strategy (see section 5.2). Numerical examples in one space dimension demonstrate that the resulting adaptive schemes converge with higher order compared with the standard first-order method. In addition, it was shown that the error estimator η_h from the a posteriori Theorem 2.5 converges with higher order for higher-order methods. The analysis of the more involved fully discrete case needs additional new ideas and is therefore left for further study. The application of the principle ideas to the multidimensional case and to systems of conservation laws is of special importance and will be the subject of future work. For first results in two space dimensions that show a very nice behavior of the resulting *hp*-adaptive schemes, we refer to [14].

Acknowledgment. We thank Eitan Tadmor for the hint about the existence of the work on NCE Runge–Kutta schemes.

REFERENCES

[1] S. Adjerid, K. D. Devine, J. E. Flaherty, and L. Krivodonova, *A posteriori error estimation for discontinuous Galerkin solutions of hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1097–1112.

[2] C. Arvanitis, C. Makridakis, and A. E. Tzavaras, *Stability and convergence of a class of finite element schemes for hyperbolic systems of conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 1357–1393.

[3] R. Biswas, K. D. Devine, and J. E. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Appl. Numer. Math., 14 (1994), pp. 255–283.

[4] F. Bouchut and B. Perthame, *Kruzkov's estimates for scalar conservation laws revisited*, Trans. Amer. Math. Soc., 350 (1998), pp. 2847–2870.

[5] B. Cockburn and P.-A. Gremaud, *Error estimates for finite element methods for scalar conservation laws*, SIAM J. Numer. Anal., 33 (1996), pp. 522–554.

[6] B. Cockburn and P. A. Gremaud, *A priori error estimates for numerical methods for scalar conservation laws. Part* I: *The general approach*, Math. Comput., 65 (1996), pp. 533–573.

[7] B. Cockburn, S. Hou, and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws.* IV. *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[8] B. Cockburn, C. Johnson, C.-W. Shu, and E. Tadmor, *Advanced Numerical Approximation of Nonlinear Hyperbolic Equations* (papers from the C.I.M.E. Summer School held in Cetraro, 1997), Lecture Notes in Math. 1697, A. Quarteroni, Fondazione C.I.M.E., eds., Springer-Verlag, Berlin, 1998.

[9] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws.* II. *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[10] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.

[11] B. Cockburn and C.-W. Shu, *Runge–Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput., 16 (2001), pp. 173–261.

[12] C. M. Dafermos, *Hyperbolic Conservation Laws in Continuum Physics*, Grundlehren Math. Wiss. 325, Springer-Verlag, Berlin, 2000.

[13] A. Dedner, C. Makridakis, and M. Ohlberger, *Error Control for a Class of Runge–Kutta Discontinuous Galerkin Methods for Nonlinear Conservation Laws*, Preprint 05-01, Mathematisches Institut, Freiburg, Germany, 2005.

[14] A. Dedner and M. Ohlberger, *A new hp-adaptive DG scheme for conservation laws based on error control*, in Proceedings of the 11th International Conference on Hyperbolic Problems: Theory, Numerics, and Applications, Lyon, France, 2006, to appear.
[15] T. Gallouët and R. Herbin, *A uniqueness result for measure-valued solutions of nonlinear hyperbolic equations*, Differential Integral Equations, 6 (1993), pp. 1383–1394.
[16] L. Gosse and C. Makridakis, *Two a posteriori error estimates for one-dimensional scalar conservation laws*, SIAM J. Numer. Anal., 38 (2000), pp. 964–988.
[17] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.
[18] R. Hartmann and P. Houston, *Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 979–1004.
[19] P. Houston, B. Senior, and E. Süli, *hp-discontinuous Galerkin finite element methods for hyperbolic problems: Error analysis and adaptivity*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 153–169.
[20] P. Houston and E. Süli, *hp-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems*, SIAM J. Sci. Comput., 23 (2001), pp. 1226–1252.
[21] J. Jaffré, C. Johnson, and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.
[22] C. Johnson and A. Szepessy, *Adaptive finite element methods for conservation laws based on a posteriori error estimates*, Comm. Pure Appl. Math., 48 (1995), pp. 199–234.
[23] T. Katsaounis and C. Makridakis, *Finite volume relaxation schemes for multidimensional conservation laws*, Math. Comput., 70 (2001), pp. 533–553.
[24] M. A. Katsoulakis, G. Kossioris, and Ch. Makridakis, *Convergence and error estimates of relaxation schemes for multidimensional conservation laws*, Comm. Partial Differential Equations, 24 (1999), pp. 395–424.
[25] D. Kröner and M. Ohlberger, *A-posteriori error estimates for upwind finite volume schemes for nonlinear conservation laws in multi dimensions*, Math. Comp., 69 (2000), pp. 25–39.
[26] A. Kurganov, G. Petrova, and B. Popov, *Adaptive semi-discrete central-upwind schemes for nonconvex hyperbolic conservation laws*, SIAM J. Sci. Comput., to appear.
[27] M. G. Larson and T. J. Barth, *A posteriori error estimation for adaptive discontinuous Galerkin approximations of hyperbolic systems*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lecture Notes in Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 363–368.
[28] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 355–387.
[29] M. Ohlberger, *Higher order finite volume methods on selfadaptive grids for convection dominated reactive transport problems in porous media*, Comput. Visual. Sci., 7 (2004), pp. 41–51.
[30] M. Ohlberger and J. Vovelle, *Error estimate for the approximation of non-linear conservation laws on bounded domains by the finite volume method*, Math. Comp., 75 (2006), pp. 113–150.
[31] C.-W. Shu, *A survey of strong stability preserving high order time discretizations*, in Collected Lectures on the Preservation of Stability under Discretization (Fort Collins, CO, 2001), Proc. Appl. Math. 109, SIAM, Philadelphia, 2002, pp. 51–61.
[32] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[33] E. Süli and P. Houston, *Adaptive finite element approximation of hyperbolic problems*, in Error Estimation and Adaptive Discretization Methods in Computational Fluid Dynamics, Lecture Notes in Comput. Sci. Eng. 25, Springer, Berlin, 2003, pp. 269–344.
[34] M. Zennaro, *Natural continuous extensions of Runge-Kutta methods*, Math. Comput., 46 (1986), pp. 119–133.
[35] Q. Zhang and C.-W. Shu, *Error estimates to smooth solutions of Runge–Kutta discontinuous Galerkin methods for scalar conservation laws*, SIAM J. Numer. Anal., 42 (2004), pp. 641–666.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 539–557

VARIATIONAL MULTISCALE ANALYSIS: THE FINE-SCALE GREEN'S FUNCTION, PROJECTION, OPTIMIZATION, LOCALIZATION, AND STABILIZED METHODS*

T. J. R. HUGHES† AND G. SANGALLI‡

Abstract. We derive an explicit formula for the fine-scale Green's function arising in variational multiscale analysis. The formula is expressed in terms of the classical Green's function and a projector which defines the decomposition of the solution into coarse and fine scales. The theory is presented in an abstract operator format and subsequently specialized for the advection-diffusion equation. It is shown that different projectors lead to fine-scale Green's functions with very different properties. For example, in the advection-dominated case, the projector induced by the H_0^1-seminorm produces a fine-scale Green's function which is highly attenuated and localized. These are very desirable properties in a multiscale method and ones that are not shared by the L^2-projector. By design, the coarse-scale solution attains optimality in the norm associated with the projector. This property, combined with a localized fine-scale Green's function, indicates the possibility of effective methods with local character for dominantly hyperbolic problems. The constructs lead to a new class of stabilized methods, and the relationship between H_0^1-optimality and the streamline-upwind Petrov-Galerkin (SUPG) method is described.

Key words. multiscale, advection-diffusion

AMS subject classification. 65N30

DOI. 10.1137/050645646

1. Introduction.

The variational multiscale (VMS) method [12, 13] was introduced as a framework for incorporating missing fine-scale effects into numerical problems governing coarse-scale behavior. It has provided a rationale for stabilized methods and a platform for the development of new methods (see, e.g., [11, 14, 15, 16, 18] for application to turbulence modeling). The fundamental mathematical object in the method is the so-called *fine-scale Green's function*, introduced in [13]. Although it is a simple matter to characterize coarse-scale and fine-scale subspaces, not much is known about the fine-scale Green's function. In this paper, we study the fine-scale Green's function and present a formula for explicitly computing it from the classical Green's function. This is accomplished by observing that the decomposition of a function into a sum of coarse-scale and fine-scale components is uniquely specified by identifying a projector from the space of all scales onto the coarse-scale subspace. Different projectors produce different decompositions. The problem for the fine-scale Green's function is then posed in terms of the fine-scale subspace. Compared with the problem for the classical Green's function, this amounts to a constrained formulation. The constraint can be released by invoking the Lagrange multiplier method and the unconstrained problem can be solved in terms of the classical Green's function and the projector. The fine-scale Green's function enjoys orthogonality relations with respect

*Received by the editors November 19, 2005; accepted for publication (in revised form) October 17, 2006; published electronically March 19, 2007.
http://www.siam.org/journals/sinum/45-2/64564.html

†Institute for Computational Engineering and Sciences, The University of Texas at Austin, Austin, TX 78712-0027 (hughes@ices.utexas.edu). The work of this author was supported by Sandia contract A0340.0 with the University of Texas and is gratefully acknowledged.

‡Dipartimento di Matematica, Università di Pavia, Via Ferrata 1, 27100 Pavia, Italy (sangalli@imati.cnr.it). This research of this author was supported by the J. Tinsley Oden Faculty Fellowship Research Program and PRIN 2004 project of the Italian MIUR.

to the projector. If a scalar product is introduced with a corresponding projector, the coarse-scale solution of the original problem is the optimal approximation in terms of the induced norm. The theory summarizing these ideas is presented in section 2 in an abstract operator format for a general linear isomorphism.

These ideas are applied to the advection-diffusion equation in section 3. The fine-scale Green's function is explicitly calculated in one dimension for linear, quadratic, and cubic finite elements when the projector is defined by the H_0^1-seminorm. In this case, the fine-scale Green's function is local in that it is confined to individual elements and is not coupled from one element to another, even in advection-dominated cases. This is a highly desirable property in multiscale analysis and in complete contrast with the classical Green's function which exhibits global support in advection-dominated cases. It also suggests that efficient, approximate, multiscale methods possessing local character may be possible for dominantly hyperbolic phenomena. On the other hand, selecting the L^2-projector results in a fine-scale Green's function with global coupling. These results show clearly that the choice of projector is of key importance in the development of a multiscale method.

The fine-scale Green's function becomes increasingly complicated as the order of the coarse-scale space is increased. However, it is observed that due to the orthogonality properties of the fine-scale Green's function, it only interacts with the highest-order polynomial term in the residual. This means that for a kth-order coarse-scale space, the fine-space Green's function modification to the coarse-scale equation can be replaced by an equivalent stabilization term involving a computable, elementwise constant (i.e., a "τ" in the notation of stabilized methods) and derivatives of the residual and weighting operator of order $k-1$. Remarkably, the modification reduces to elementwise constant terms requiring no quadrature despite the complexity of the fine-scale Green's function. This results in optimal higher-order methods of extraordinary simplicity.

To assess the situation in multiple dimensions, the two-dimensional advection-diffusion equation is studied. Here, rather than proceeding analytically, numerical procedures involving very fine meshes are utilized to determine Green's functions. As in the one-dimensional case, the classical Green's function exhibits global character with support in the form of a tail surrounding the upwind characteristic through the point of application of the Dirac mass. When advection-dominated, this tail is not attenuated with distance. However, the fine-scale Green's function for the H_0^1-projector is highly attenuated under the same circumstances and is essentially confined to a small number of elements (in the coarse-scale space) surrounding the point of application of the Dirac mass. The L^2-projector engenders a fine-scale Green's function which is not localized, and one concludes that the main observations made for the one-dimensional case are essentially true in two dimensions.

The H_0^1-projector produces a method which is highly localized and attains an optimal approximation in the H_0^1-seminorm, a combination of desirable properties. It is also noted that the modification it introduces to a classical Galerkin formulation involves an additional stabilization term in which the coarse-scale residual is weighted by the fine-scale Green's function convolved only with the advective part of the operator, i.e., the diffusive operator does not appear in the weighting. These are features that the H_0^1-optimal method has in common with streamline-upwind Petrov-Galerkin (SUPG) [9].

In section 4 we draw conclusions.

2. The abstract framework.

2.1. The abstract problem. Let V be a Hilbert space endowed with a norm $\|\cdot\|_V$ and a scalar product $(\cdot,\cdot)_V$. Let V^* be the dual of V, and let ${}_{V^*}\langle\cdot,\cdot\rangle_V$ be the pairing between them. Let $\mathcal{L}: V \to V^*$ be a linear isomorphism. Given $f \in V^*$, we consider the abstract problem of finding $u \in V$ such that

$$\mathcal{L}u = f. \tag{2.1}$$

The variational formulation of (2.1) is find $u \in V$ such that

$${}_{V^*}\langle \mathcal{L}u, v\rangle_V = {}_{V^*}\langle f, v\rangle_V \qquad \forall v \in V. \tag{2.2}$$

The solution u can be expressed as $u = \mathcal{G}f$, where $\mathcal{G}: V^* \to V$ is the Green's operator, i.e., $\mathcal{G} = \mathcal{L}^{-1}$.

2.2. The variational multiscale formulation. Let $\bar{V}$ be a closed subspace of V, and let $\mathcal{P}$ be a linear projector onto $\bar{V}$; i.e., $\mathcal{P}^2 = \mathcal{P}$ and $\mathrm{Range}(\mathcal{P}) = \bar{V}$. We assume $\mathcal{P}$ to be continuous in V. Since $\mathcal{P}\bar{v} = \bar{v}$ for all $\bar{v} \in \bar{V}$, we have the obvious inf-sup condition

$$\inf_{\bar{v}\in\bar{V}} \sup_{w\in V} \frac{(\mathcal{P}w, \bar{v})_V}{\|w\|_V \|\bar{v}\|_V} \geq 1. \tag{2.3}$$

We define $V' = \mathrm{Ker}(\mathcal{P})$, which is a closed subspace of V, thanks to the continuity of $\mathcal{P}$. As a consequence,

$$V = \bar{V} \oplus V'; \tag{2.4}$$

i.e., any $v \in V$ can be written uniquely as $v = \bar{v} + v'$, where $\bar{v} \in \bar{V}$ and $v' \in V'$: indeed, $\bar{v} = \mathcal{P}v$ and $v' = v - \mathcal{P}v$. In particular, we split the solution u of (2.1) as $u = \bar{u} + u'$. In the VMS approach, $\bar{V}$ represents the space of computable *coarse scales*, while V' contains the unresolved *fine scales*. The aim of the VMS approach is to obtain $\bar{u} = \mathcal{P}u$.

The variational formulation (2.2) splits into

$${}_{V^*}\langle \mathcal{L}\bar{u}, \bar{v}\rangle_V + {}_{V^*}\langle \mathcal{L}u', \bar{v}\rangle_V = {}_{V^*}\langle f, \bar{v}\rangle_V \qquad \forall \bar{v} \in \bar{V}, \tag{2.5}$$

$${}_{V^*}\langle \mathcal{L}\bar{u}, v'\rangle_V + {}_{V^*}\langle \mathcal{L}u', v'\rangle_V = {}_{V^*}\langle f, v'\rangle_V \qquad \forall v' \in V'. \tag{2.6}$$

We assume that (2.5) is a well-posed problem for $\bar{u}$ alone, meaning that it admits a unique solution $\bar{u} \in \bar{V}$ given u' and f. Analogously, we assume that (2.6) is well-posed for $u' \in V'$ given $\bar{u}$ and f. For that, we ask the inf-sup conditions for $\mathcal{L}$ on $\bar{V}$ and V'

$$\inf_{\bar{w}\in\bar{V}} \sup_{\bar{v}\in\bar{V}} \frac{{}_{V^*}\langle \mathcal{L}\bar{w}, \bar{v}\rangle_V}{\|\bar{w}\|_V \|\bar{v}\|_V} > 0 \quad \text{and} \quad \sup_{\bar{w}\in\bar{V}} \frac{{}_{V^*}\langle \mathcal{L}\bar{w}, \bar{v}\rangle_V}{\|\bar{w}\|_V} > 0 \quad \forall \bar{v} \in \bar{V}\backslash\{0\}, \tag{2.7}$$

$$\inf_{w'\in V'} \sup_{v'\in V'} \frac{{}_{V^*}\langle \mathcal{L}w', v'\rangle_V}{\|w'\|_V \|v'\|_V} > 0 \quad \text{and} \quad \sup_{w'\in V'} \frac{{}_{V^*}\langle \mathcal{L}w', v'\rangle_V}{\|w'\|_V} > 0 \quad \forall v' \in V'\backslash\{0\}. \tag{2.8}$$

If $\mathcal{L}$ is coercive on V, i.e., ${}_{V^*}\langle \mathcal{L}v, v\rangle_V \geq C\|v\|_V^2$ for $C > 0$ and for all $v \in V$, then (2.7)–(2.8) hold.

We associate with (2.6) the *fine-scale Green's operator* $\mathcal{G}': V^* \to V'$, which gives u' from the coarse-scale *residual* $f - \mathcal{L}\bar{u}$, i.e.,

$$u' = \mathcal{G}'(f - \mathcal{L}\bar{u}). \tag{2.9}$$

Having $\mathcal{G}'$, we can eliminate u' from (2.5), and we obtain the VMS formulation for $\bar{u}$:

$$ {}_{V^*}\langle \mathcal{L}\bar{u}, \bar{v}\rangle_V - {}_{V^*}\langle \mathcal{L}\mathcal{G}'\mathcal{L}\bar{u}, \bar{v}\rangle_V = {}_{V^*}\langle f, \bar{v}\rangle_V - {}_{V^*}\langle \mathcal{L}\mathcal{G}' f, \bar{v}\rangle_V \qquad \forall \bar{v} \in \bar{V}. \tag{2.10} $$

Because of (2.4), the formulation (2.10) admits a unique solution, which is precisely $\bar{u} = \mathcal{P}u$.

2.3. The fine-scale Green's operator. We denote by $\mathcal{P}^T : \bar{V}^* \to V^*$ the adjoint of $\mathcal{P}$, i.e.,

$$ {}_{V^*}\langle \mathcal{P}^T\bar{\mu}, v\rangle_V = {}_{\bar{V}^*}\langle \bar{\mu}, \mathcal{P}v\rangle_{\bar{V}} \qquad \forall v \in V,\ \bar{\mu} \in \bar{V}^*, $$

where $\bar{V}^*$ is the dual of $\bar{V}$ and ${}_{\bar{V}^*}\langle\cdot,\cdot\rangle_{\bar{V}}$ is the pairing between them.

In the next result we express $\mathcal{G}'$ in terms of $\mathcal{G}$ and $\mathcal{P}$.

THEOREM 2.1. *Under the assumptions of sections* 2.1 and 2.2*, we have*

$$ \mathcal{G}' = \mathcal{G} - \mathcal{G}\mathcal{P}^T(\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}\mathcal{P}\mathcal{G}, \tag{2.11} $$

$$ \mathcal{G}'\mathcal{P}^T = 0, \quad \textit{and} \quad \mathcal{P}\mathcal{G}' = 0. \tag{2.12} $$

Proof. Since (2.6) is a constrained problem, we can rephrase it making use of a Lagrange multiplier in mixed (unconstrained) form: Find $u' \in V$ and $\bar{\lambda} \in \bar{V}^*$ such that

$$ \mathcal{L}u' + \mathcal{P}^T\bar{\lambda} = r, \tag{2.13} $$

$$ \mathcal{P}u' = 0, \tag{2.14} $$

where $r = f - \mathcal{L}\bar{u}$. The well-posedness of (2.13)–(2.14), for any $r \in V^*$, is guaranteed by our previous assumptions. Indeed, following [2], we need (2.8) and the inf-sup condition for $\mathcal{P}$

$$ \inf_{\bar{\mu}\in\bar{V}^*} \sup_{w\in V} \frac{{}_{\bar{V}^*}\langle \bar{\mu}, \mathcal{P}w\rangle_{\bar{V}}}{\|\bar{\mu}\|_{\bar{V}^*}\|w\|_V}, $$

which is equivalent to (2.3) in this Hilbert space setting. From (2.13) we get $u' = \mathcal{G}(r - \mathcal{P}^T\bar{\lambda})$; substituting in (2.14) gives $\mathcal{P}\mathcal{G}r - \mathcal{P}\mathcal{G}\mathcal{P}^T\bar{\lambda} = 0$; the well-posedness of (2.13)–(2.14) guarantees the invertibility of $\mathcal{P}\mathcal{G}\mathcal{P}^T$; and hence we obtain $\bar{\lambda} = (\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}\mathcal{P}\mathcal{G}r$. Finally, using this in the expression for u' yields

$$ u' = (\mathcal{G} - \mathcal{G}\mathcal{P}^T(\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}\mathcal{P}\mathcal{G})r, \tag{2.15} $$

which gives (2.11).

From (2.11), we immediately have

$$ \mathcal{G}'\mathcal{P}^T = \mathcal{G}\mathcal{P}^T - \mathcal{G}\mathcal{P}^T(\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}(\mathcal{P}\mathcal{G}\mathcal{P}^T) = \mathcal{G}\mathcal{P}^T - \mathcal{G}\mathcal{P}^T = 0 $$

and

$$ \mathcal{P}\mathcal{G}' = \mathcal{P}\mathcal{G} - (\mathcal{P}\mathcal{G}\mathcal{P}^T)(\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}\mathcal{P}\mathcal{G} = \mathcal{P}\mathcal{G} - \mathcal{P}\mathcal{G} = 0, $$

which are (2.12). ∎

Using the expression (2.11) in (2.10), we see that the left-hand side of (2.10) is

$$ {}_{V^*}\langle \mathcal{L}\bar{u}, \bar{v}\rangle_V - {}_{V^*}\langle \mathcal{L}\mathcal{G}'\mathcal{L}\bar{u}, \bar{v}\rangle_V = {}_{\bar{V}^*}\langle (\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}\bar{u}, \bar{v}\rangle_{\bar{V}}\,. \tag{2.16} $$

As $(\mathcal{P}\mathcal{G}\mathcal{P}^T)^{-1}$ is obviously invertible, (2.16) confirms that (2.10) is a well-posed formulation.

In the cases of practical interest, $\bar{V}$ is a finite-dimensional subspace of V. If the dimension of $\bar{V}$ is N, then we can find a set of functionals $\{\mu_i\}_{i=1,\dots,N}$ such that for all $v \in V$

$$ {}_{V^*}\langle \mu_i, v\rangle_V = 0 \quad \forall i = 1,\dots,N \qquad \Leftrightarrow \qquad \mathcal{P}v = 0. \tag{2.17}$$

In other words, the equations ${}_{V^*}\langle \mu_i, v\rangle_V = 0$ for $1 \le i \le N$ characterize v as a fine-scale function, i.e., $v \in V'$. From the mathematical standpoint, $\{\mu_i\}_{i=1,\dots,N}$ is a basis for the image of $\mathcal{P}^T$. Therefore, it is clear that (2.12) is equivalent to

$$ \mathcal{G}'\mu_i = 0 \qquad \forall i = 1,\dots,N \tag{2.18}$$

and

$$ {}_{V^*}\langle \mu_i, \mathcal{G}'\nu\rangle_V = 0 \qquad \forall \nu \in V^* \ \forall i = 1,\dots,N. \tag{2.19}$$

Moreover, after introducing the vector $\boldsymbol{\mu} \in (V^*)^N$ and its transpose,

$$ \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_N \end{bmatrix} \qquad \text{and} \qquad \boldsymbol{\mu}^T = \begin{bmatrix} \mu_1 & \dots & \mu_N \end{bmatrix}; $$

the vector $\mathcal{G}\boldsymbol{\mu}^T \in V^N$,

$$ \mathcal{G}\boldsymbol{\mu}^T = \begin{bmatrix} \mathcal{G}\mu_1 & \dots & \mathcal{G}\mu_N \end{bmatrix}; $$

the matrix $\boldsymbol{\mu}\mathcal{G}\boldsymbol{\mu}^T \in \mathbb{R}^{N\times N}$,

$$ \boldsymbol{\mu}\mathcal{G}\boldsymbol{\mu}^T = \begin{bmatrix} {}_{V^*}\langle \mu_1, \mathcal{G}\mu_1\rangle_V & \dots & {}_{V^*}\langle \mu_1, \mathcal{G}\mu_N\rangle_V \\ \vdots & \ddots & \vdots \\ {}_{V^*}\langle \mu_N, \mathcal{G}\mu_1\rangle_V & \dots & {}_{V^*}\langle \mu_N, \mathcal{G}\mu_N\rangle_V \end{bmatrix}; $$

and the vector of functionals $\boldsymbol{\mu}\mathcal{G} : (V^*) \to \mathbb{R}^N$ (i.e., $\boldsymbol{\mu}\mathcal{G} \in (V^{**})^N$) such that

$$ \boldsymbol{\mu}\mathcal{G}(\nu) = \begin{bmatrix} {}_{V^*}\langle \mu_1, \mathcal{G}\nu\rangle_V \\ \vdots \\ {}_{V^*}\langle \mu_N, \mathcal{G}\nu\rangle_V \end{bmatrix} \quad \forall \nu \in V^*, $$

it is easy to see that (2.11) is equivalent to

$$ \mathcal{G}' = \mathcal{G} - \mathcal{G}\boldsymbol{\mu}^T \left[\boldsymbol{\mu}\mathcal{G}\boldsymbol{\mu}^T\right]^{-1} \boldsymbol{\mu}\mathcal{G}. \tag{2.20}$$

2.4. Orthogonal projectors and optimization. An interesting case, and the only one considered in what follows, is when $\mathcal{P}$ is an orthogonal projector.

Given a scalar product $(\cdot,\cdot)$ defined on $V \times V$, possibly different than $(\cdot,\cdot)_V$, the related orthogonal projector $\mathcal{P}$ is obviously defined by

$$ (\mathcal{P}w, \bar{v}) = (w, \bar{v}) \qquad \forall w \in V, \forall \bar{v} \in \bar{V}. \tag{2.21}$$

Recall that, in order to fit in the abstract framework of section 2.2, $\mathcal{P}$ must be a continuous operator in V. However, when $\bar{V}$ is a finite-dimensional space, this holds for any scalar product $(\cdot,\cdot)$ which is continuous on $V \times V$.

In this context, V' and $\bar{V}$ are orthogonal complements with respect to $(\cdot,\cdot)$, and the VMS formulation provides the optimal approximation $\bar{u} \in \bar{V}$ of u, with respect to the norm $\|\cdot\|$ induced by the scalar product $(\cdot,\cdot)$.

3. The advection-diffusion model problem. Let d be the space dimension ($d = 1$ and $d = 2$ will be taken into consideration in the examples), and let $\Omega \subset \mathbb{R}^d$ be a regular domain. We consider the advection-diffusion model problem

$$\mathcal{L}u = -\kappa\Delta u + \beta \cdot \nabla u = f \text{ in } \Omega \qquad \text{with } u_{|\partial\Omega} = 0, \tag{3.1}$$

where $f \in L^2(\Omega)$ is the source term, $\kappa > 0$ is the scalar diffusivity, and $\beta : \Omega \to \mathbb{R}^d$ is the advection field, for which we assume $\mathrm{div}(\beta) = 0$. For the variational formulation of (3.1), within the framework of section 2, we set $V = H^1_0 \equiv H^1_0(\Omega)$ whence $V^* = H^{-1}$. Typical finite element spaces will be considered as coarse spaces $\bar{V}$.

In this context, it is convenient to represent the Green's operator $\mathcal{G}$ through the Green's function $g : \Omega \times \Omega \to \mathbb{R}$ such that

$$u(y) = \int_\Omega g(x, y) f(x)\, dx \tag{3.2}$$

for almost every y in Ω. Note that in (3.2) and in what follows the integrals have to be interpreted in the sense of distributions. We refer to [22] for details. Some explicit representations are given in [13].

We recall that $g_{|\partial(\Omega\times\Omega)} = 0$ and, for all $y \in \Omega$, $\mathcal{L}^* g(\cdot, y) = \delta(\cdot - y)$, where δ is the Dirac mass at the origin and $\mathcal{L}^* = -\kappa\Delta - \beta \cdot \nabla$ denotes the dual of $\mathcal{L}$.

Furthermore, we introduce the fine-scale Green's function $g' : \Omega \times \Omega \to \mathbb{R}$, which represents the fine-scale Green's operator $\mathcal{G}'$ and gives the fine-scale component u' of u from the coarse-scale residual $r = f - \mathcal{L}\bar{u}$ by

$$u'(y) = \int_\Omega g'(x, y) r(x)\, dx. \tag{3.3}$$

Recall, however, that the space of fine scales V' as well as the fine-scale Green's function g' depend on the underlying projector $\mathcal{P}$. With an abuse of notation, in the next sections we shall write V' and g' without distinction among the different projectors taken into consideration. In particular, we will deal with the H^1_0-projector $\mathcal{P} = \mathcal{P}_{H^1_0}$, associated with the scalar product $(w, v) = (w, v)_{H^1_0} = \int_\Omega \nabla w(x) \cdot \nabla v(x)\, dx$, and the usual L^2-projector $\mathcal{P} = \mathcal{P}_{L^2}$. Having a set of functionals $\{\mu_i\}_{i=1,\dots,N}$ as in (2.17), i.e., giving

$$\int_\Omega \mu_i(x) v(x)\, dx = 0 \quad \forall i = 1, \dots, N \qquad \Leftrightarrow \qquad \mathcal{P}v = 0, \tag{3.4}$$

then g' is obtained straightforwardly by (2.20) as

(3.5)

$$\begin{aligned} g'(x, y) =& g(x, y) - \left[\int_\Omega g(\tilde{x}, y)\mu_1(\tilde{x})\, d\tilde{x} \quad \cdots \quad \int_\Omega g(\tilde{x}, y)\mu_N(\tilde{x})\, d\tilde{x}\right] \\ & \times \begin{bmatrix} \int_\Omega g(\tilde{x}, \tilde{y})\mu_1(\tilde{x})\mu_1(\tilde{y})\, d\tilde{x}d\tilde{y} & \cdots & \int_\Omega g(\tilde{x}, \tilde{y})\mu_N(\tilde{x})\mu_1(\tilde{y})\, d\tilde{x}d\tilde{y} \\ \vdots & \ddots & \vdots \\ \int_\Omega g(\tilde{x}, \tilde{y})\mu_1(\tilde{x})\mu_N(\tilde{y})\, d\tilde{x}d\tilde{y} & \cdots & \int_\Omega g(\tilde{x}, \tilde{y})\mu_N(\tilde{x})\mu_N(\tilde{y})\, d\tilde{x}d\tilde{y} \end{bmatrix}^{-1} \\ & \times \begin{bmatrix} \int_\Omega g(x, \tilde{y})\mu_1(\tilde{y})\, d\tilde{y} \\ \vdots \\ \int_\Omega g(x, \tilde{y})\mu_N(\tilde{y})\, d\tilde{y} \end{bmatrix}, \end{aligned}$$

while (2.18) and (2.19) mean that, for all $x, y \in \Omega$ and for all $i = 1, \dots, N$,

$$\int_\Omega g'(\tilde{x}, y)\mu_i(\tilde{x})\, d\tilde{x} = 0 \text{ and } \int_\Omega g'(x, \tilde{y})\mu_i(\tilde{y})\, d\tilde{y} = 0. \tag{3.6}$$

In this context, the VMS formulation (2.10) reads as follows: Find $\bar{u} \in \bar{V}$ such that

$$\begin{aligned}\int_\Omega (\kappa \nabla \bar{u}(x) - \beta \bar{u}) \cdot \nabla \bar{v}(x)\, dx - \int_\Omega \int_\Omega \mathcal{L}\bar{u}(x) g'(x,y) \mathcal{L}^* \bar{v}(y)\, dxdy \\ = \int_\Omega f(x)\bar{v}(x)\, dx - \int_\Omega \int_\Omega f(x) g'(x,y) \mathcal{L}^* \bar{v}(y)\, dxdy \qquad \forall \bar{v} \in \bar{V}.\end{aligned} \tag{3.7}$$

3.1. Linear elements and H_0^1-optimality in one dimension. Let $d = 1$ and $\Omega = (0, L)$. Consider a grid of nodes $0 = x_0 < x_1 < \cdots < x_{n_{el}-1} < x_{n_{el}} = L$ and the related subdivision of $(0, L)$ into n_{el} elements (x_{i-1}, x_i), $i = 1, \ldots, n_{el}$. Let $\bar{V} \subset H_0^1$ be the space of piecewise-linear (with respect to the subdivision) functions, which is of dimension $N = n_{el} - 1$.

In this context, the H_0^1-projector $\mathcal{P} = \mathcal{P}_{H_0^1}$ plays a special role; indeed, $(\mathcal{P}v)(x_i) = v(x_i)$ for all $i = 1, \ldots, N$. Then, the VMS approach provides in this case a nodally exact approximation $\bar{u}$ of the exact solution u (see [3, 8, 12, 13]).

In order to have (3.4), we set $\mu_i = \delta(x - x_i)$. The abstract property (3.6) becomes, in this case,

$$g'(x, x_i) = g'(x_i, y) = 0 \qquad \forall i = 1, \ldots, N, \quad 0 \le x, y \le L; \tag{3.8}$$

i.e., g' vanishes if one of its two arguments is a node of the grid. Moreover, (3.5) gives

$$\begin{aligned}g'(x,y) = g(x,y) - \begin{bmatrix} g(x_1,y) & \cdots & g(x_N,y)\end{bmatrix} \\ \times \begin{bmatrix} g(x_1,x_1) & \cdots & g(x_N,x_1) \\ \vdots & \ddots & \vdots \\ g(x_1,x_N) & \cdots & g(x_N,x_N)\end{bmatrix}^{-1} \times \begin{bmatrix} g(x,x_1) \\ \vdots \\ g(x,x_N)\end{bmatrix}.\end{aligned} \tag{3.9}$$

Recalling that $\mathcal{L}^* g(\cdot, y) = \delta(\cdot - y)$, from (3.9) we get

$$\begin{aligned}\mathcal{L}^* g'(\cdot,y) = \delta(\cdot - y) - \begin{bmatrix} g(x_1,y) & \cdots & g(x_N,y)\end{bmatrix} \\ \times \begin{bmatrix} g(x_1,x_1) & \cdots & g(x_N,x_1) \\ \vdots & \ddots & \vdots \\ g(x_1,x_N) & \cdots & g(x_N,x_N)\end{bmatrix}^{-1} \times \begin{bmatrix} \delta(\cdot - x_1) \\ \vdots \\ \delta(\cdot - x_N)\end{bmatrix}.\end{aligned}$$

If $x_{i-1} < y < x_i$, then

$$\mathcal{L}^* g'(\cdot, y) = \delta(\cdot - y) \text{ in } (x_{i-1}, x_i), \tag{3.10}$$

while when $y > x_i$ or $y < x_{i-1}$,

$$\mathcal{L}^* g'(\cdot, y) = 0 \text{ in } (x_{i-1}, x_i). \tag{3.11}$$

This, with (3.8), fully characterizes g': By (3.8) and (3.11), we see that $g'(x, y) = 0$ if x and y belong to two different elements; moreover, (3.8) and (3.10) say that g' is, on each $(x_{i-1}, x_i) \times (x_{i-1}, x_i)$, the so-called *element* Green's function g^{el}, i.e., the Green's function for the restriction of $\mathcal{L}$ to the element (x_{i-1}, x_i), with homogeneous Dirichlet boundary conditions at the endpoints x_{i-1} and x_i. Since $g'(x, y) \neq 0$ only when x and y belong to the same element, (3.3) can be localized within each element

$$u'(y) = \int_{x_{i-1}}^{x_i} g'(x,y) r(x)\, dx \qquad \forall y \in (x_{i-1}, x_i). \tag{3.12}$$

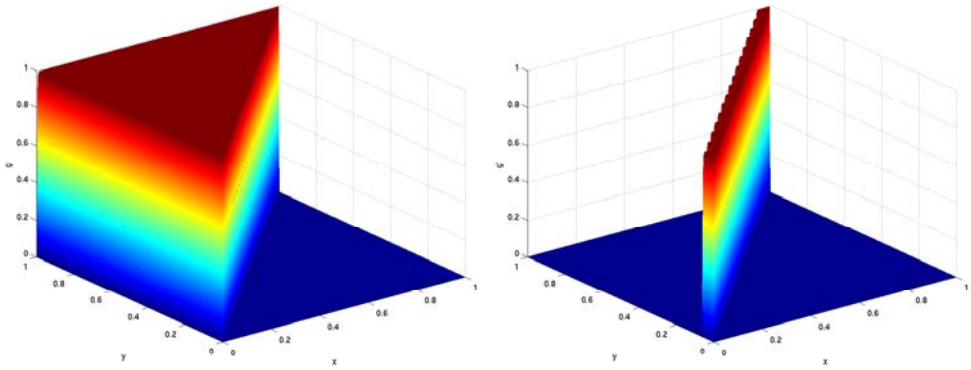


FIG. 3.1. *Comparison between the Green's function g (left) and the fine-scale Green's function g' (right) for the one-dimensional problem and linear elements, with $\mathcal{P} = \mathcal{P}_{H_0^1}$, $\kappa = 10^{-3}$, $\beta = 1$, $L = 1$, and a uniform grid of $n_{el} = 16$ elements. Note that the support of g' is local in that there is no coupling between elements. This is an advantage of $\mathcal{P} = \mathcal{P}_{H_0^1}$.*

See a plot of g' in Figure 3.1 where we consider the case of a uniform mesh of 16 elements (for $\kappa = 10^{-3}$, $\beta = 1$, and $L = 1$) and we compare with the plot of the Green's function g.

As stated above, the structure of g' for this case is well known in the literature [3, 8, 12, 13]. Indeed, recognizing that V' is the space of *bubbles*

$$V' = \bigoplus_{i=1,\dots,n_{el}} H_0^1(x_{i-1}, x_i), \tag{3.13}$$

the fine-scale variational equation (2.6) splits element by element and admits the strong form

$$\mathcal{L}u' = f - \mathcal{L}\bar{u} \text{ on } (x_{i-1}, x_i) \qquad \text{with } u'(x_{i-1}) = u'(x_i) = 0 \tag{3.14}$$

for each $i = 1, \dots, n_{el}$; u' is the solution of the advection-diffusion problem at the element level with the coarse-scale residual acting as the right-hand side. This is why $g' = g^{el}$ at the element level.

Moreover, assuming piecewise-constant coefficients κ, β and source term f, the fine-scale effect on the coarse-scale variational equation is

$$\begin{aligned} \int_0^L \int_0^L \mathcal{L}^* \bar{v}(y) g'(x,y) r(x)\, dxdy &= \sum_{i=1}^{n_{el}} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \mathcal{L}^* \bar{v}(y) g'(x,y) r(x)\, dxdy \\ &= \sum_{i=1}^{n_{el}} \frac{\int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} g'(x,y)\, dxdy}{x_i - x_{i-1}} \int_{x_{i-1}}^{x_i} r(x) \mathcal{L}^* \bar{v}(x)\, dx, \end{aligned} \tag{3.15}$$

which is recognized as a classical stabilization term depending on the parameter [12, 13]

$$\tau_1 \equiv \tau_{1,(x_{i-1},x_i)} = \frac{\int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} g'(x,y)\, dxdy}{x_i - x_{i-1}} = \frac{h}{2\beta} \left(\coth(\alpha) - \frac{1}{\alpha} \right), \tag{3.16}$$

where $\alpha = (h\beta)/(2\kappa)$ is the mesh Peclét number and $h = x_i - x_{i-1}$ is the local mesh-size. We show plots of g' on the shifted element domain $(0, h) \times (0, h)$ in the diffusive and in the advective regime in Figure 3.2. A plot of τ_1 is presented in Figure 3.5.

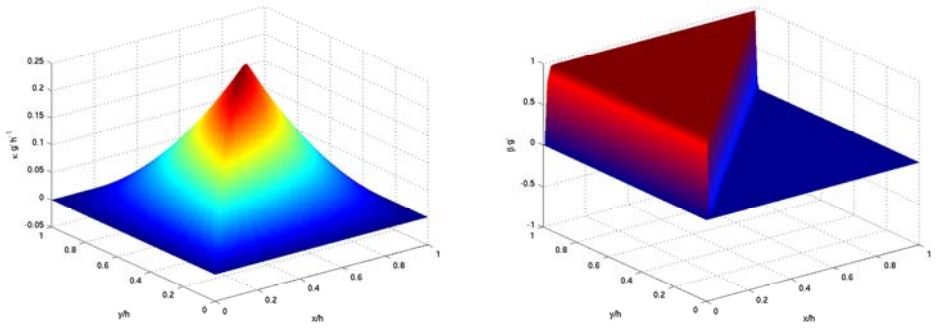


FIG. 3.2. *Fine-scale Green's functions g' at the element level $(0,h)\times(0,h)$ for the one-dimensional problem and linear elements. In the diffusive regime $\alpha = 10^{-2}$ (left), and in the advective regime $\alpha = 10^2$ (right). $\mathcal{P} = \mathcal{P}_{H_0^1}$.*

3.2. Higher-order elements and H_0^1-optimality in one dimension. We consider now higher-order piecewise-polynomial coarse scales on the grid $0 = x_0 < x_1 < \cdots < x_{n_{el}-1} < x_{n_{el}} = L$; i.e., we set

$$\bar{V} = \left\{\bar{v} \in H_0^1(0,L) \text{ such that } \bar{v}_{|(x_{i-1},x_i)} \in \mathbb{P}_k,\ 1 \le i \le n_{el}\right\},$$

where $\mathbb{P}_k$ is the space of polynomials of degree at most k. We still deal with $\mathcal{P} = \mathcal{P}_{H_0^1}$. The case of higher-order elements ($k \ge 2$) has not been studied in the literature of VMS methods, as far as we know. There are indeed additional difficulties with respect to the case of linear elements: V' is still a space of bubbles, but unlike the case $k = 1$, $\bar{V}$ also contains some (polynomial) bubbles, which are therefore missing in V'. This means that V' is a strict subset of bubbles

$$V' \subsetneq \bigoplus_{i=1,\ldots,n_{el}} H_0^1(x_{i-1},x_i), \tag{3.17}$$

or equivalently, V' is a space of bubbles with additional constraints. As a result, the fine-scale variational equation (2.6) can still be split element by element into

$$\int_{x_{i-1}}^{x_i} \mathcal{L}u'(x)v'(x)\,dx = \int_{x_{i-1}}^{x_i} (f(x) - \mathcal{L}\bar{u}(x))v'(x)\,dx \qquad \forall v' \in V', \quad i = 1,\ldots,n_{el}; \tag{3.18}$$

however, (3.18) is no longer equivalent to the strong form (3.14).

We can use the theory of section 2 for dealing with (3.18). Taking advantage of (3.17), we restrict from the beginning to a single element (x_{i-1},x_i) and to the bubbles supported on it. Then, we take as the fine-scale space $V_i' = V'_{|(x_{i-1},x_i)}$. The space of the bubbles which are polynomials of degree at most k plays the role of a coarse space on (x_{i-1},x_i); we set $\bar{V}_i = \bar{V}_{|(x_{i-1},x_i)} \cap H_0^1(x_{i-1},x_i)$. The space of unconstrained bubbles is $V_i = H_0^1(x_{i-1},x_i) = \bar{V}_i \oplus V_i'$. Precisely, $w \in V_i$ belongs to V_i' if and only if (integrating by parts)

$$0 = \int_{x_{i-1}}^{x_i} \frac{d}{dx}w(x)\frac{d}{dx}\bar{v}(x)\,dx = -\int_{x_{i-1}}^{x_i} w(x)\frac{d^2}{dx^2}\bar{v}(x)\,dx \qquad \forall \bar{v} \in \bar{V}_i. \tag{3.19}$$

The second-order derivatives of $\bar{V}_i$ functions are the polynomials of degree at most $k-2$. We need, as $\{\mu_j\}_{j=1,\ldots,N}$ (where $N = k-1$, now), a basis of $\mathbb{P}_{k-2}$. For example,

we can set $\mu_j(x) = (x - x_{i-1})^{j-1}$ for $1 \le j \le N$. The constraint is expressed, as in (2.17), by N scalar equations: $v \in V_i$ belongs to V_i' if and only if

$$\int_{x_{i-1}}^{x_i} \mu_j(x) v(x)\, dx = 0, \qquad 1 \le j \le N. \tag{3.20}$$

The Green's function of the unconstrained bubble problem is the element Green's function g^{el}. Then, we can use the formula (2.20) and derive an expression for g' in terms of g^{el}: on $(0,h)\times(0,h)$ we have

$$\begin{aligned} g'(x,y) = g^{el}(x,y) &- \begin{bmatrix} \int_0^h g^{el}(\tilde{x},y)d\tilde{x} & \cdots & \int_0^h \tilde{x}^{k-2} g^{el}(\tilde{x},y)d\tilde{x} \end{bmatrix} \\ &\times \begin{bmatrix} \int_0^h\int_0^h g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} & \cdots & \int_0^h\int_0^h \tilde{x}^{k-2} g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} \\ \vdots & \ddots & \vdots \\ \int_0^h\int_0^h \tilde{y}^{k-2} g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} & \cdots & \int_0^h\int_0^h \tilde{x}^{k-2}\tilde{y}^{k-2} g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} \end{bmatrix}^{-1} \\ &\times \begin{bmatrix} \int_0^h g^{el}(x,\tilde{y})d\tilde{y} \\ \vdots \\ \int_0^h \tilde{y}^{k-2} g^{el}(x,\tilde{y})d\tilde{y} \end{bmatrix}. \end{aligned} \tag{3.21}$$

We recall that $g'(x,y) = 0$ if x and y belong to different elements, while g' on each $(x_{i-1},x_i)\times(x_{i-1},x_i)$ can be obtained from (3.21) straightforwardly.

We discuss now in more detail the case of quadratic ($k = 2$) and cubic ($k = 3$) coarse-scale elements. If $k = 2$, then (3.21) yields

$$g'(x,y) = g^{el}(x,y) - \frac{\int_0^h g^{el}(\tilde{x},y)d\tilde{x}\int_0^h g^{el}(x,\tilde{y})d\tilde{y}}{\int_0^h\int_0^h g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y}}. \tag{3.22}$$

For $k = 3$, (3.21) gives

$$\begin{aligned} g'(x,y) = g^{el}(x,y) &- \begin{bmatrix} \int_0^h g^{el}(\tilde{x},y)d\tilde{x} & \int_0^h \tilde{x} g^{el}(\tilde{x},y)d\tilde{x} \end{bmatrix} \\ &\times \begin{bmatrix} \int_0^h\int_0^h g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} & \int_0^h\int_0^h \tilde{x} g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} \\ \int_0^h\int_0^h \tilde{y} g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} & \int_0^h\int_0^h \tilde{x}\tilde{y} g^{el}(\tilde{x},\tilde{y})d\tilde{x}d\tilde{y} \end{bmatrix}^{-1} \\ &\times \begin{bmatrix} \int_0^h g^{el}(x,\tilde{y})d\tilde{y} \\ \int_0^h \tilde{y} g^{el}(x,\tilde{y})d\tilde{y} \end{bmatrix}. \end{aligned} \tag{3.23}$$

Plots of g' on $(0,h)\times(0,h)$ are shown in Figures 3.3 and 3.4. (See [17] for explicit formulas.)

Observe that, from (2.18)–(2.19), g' is L^2-orthogonal to $\mathbb{P}_{k-2}$ in both variables x and y on each $(x_{i-1},x_i)\times(x_{i-1},x_i)$. Still assuming that the coefficients κ and β are piecewise-constant and the source term f is a piecewise-polynomial of degree at most $k-1$, then on (x_{i-1},x_i) we have

$$r(x) = \frac{d^{k-1}r}{dx^{k-1}}\frac{x^{k-1}}{(k-1)!} + \text{ “polynomial of degree } \le k-2\text{”}$$

and

$$\mathcal{L}^*\bar{v}(y) = \frac{d^{k-1}\mathcal{L}^*\bar{v}}{dx^{k-1}}\frac{y^{k-1}}{(k-1)!} + \text{ “polynomial of degree } \le k-2\text{”}.$$

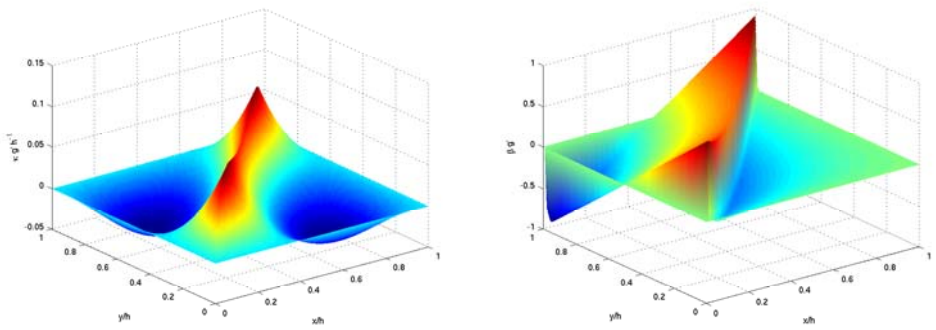


FIG. 3.3. *Fine-scale Green's functions g' at the element level $(0,h)\times(0,h)$ for the one-dimensional problem and quadratic elements. In the diffusive regime $\alpha = 10^{-2}$ (left), and in the advective regime $\alpha = 10^{2}$ (right). $\mathcal{P} = \mathcal{P}_{H_0^1}$.*

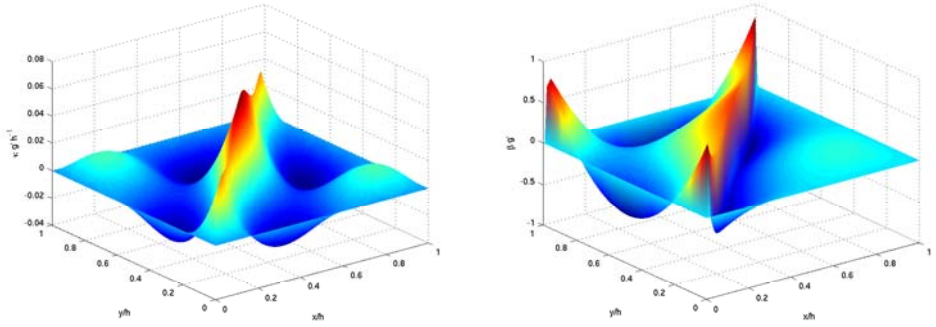


FIG. 3.4. *Fine-scale Green's functions g' at the element level $(0,h)\times(0,h)$ for the one-dimensional problem and cubic elements. In the diffusive regime $\alpha = 10^{-2}$ (left), and in the advective regime $\alpha = 10^{2}$ (right). $\mathcal{P} = \mathcal{P}_{H_0^1}$.*

Therefore, exploiting both the locality and the orthogonality of g' with respect to polynomials of degree $k-2$, the fine-scale effect on the coarse-scale equation can be written as

$$
\begin{aligned}
&\int_0^L \int_0^L \mathcal{L}^*\bar{v}(y) g'(x,y) r(x)\, dxdy \\
&\qquad = \sum_{i=1}^{n_{el}} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} \mathcal{L}^*\bar{v}(y) g'(x,y) r(x)\, dxdy \\
&\qquad = \frac{1}{((k-1)!)^2} \sum_{i=1}^{n_{el}} \int_{x_{i-1}}^{x_i} \int_{x_{i-1}}^{x_i} y^{k-1} \frac{d^{k-1}\mathcal{L}^*\bar{v}}{dx^{k-1}} g'(x,y) x^{k-1} \frac{d^{k-1} r}{dx^{k-1}}\, dxdy \\
&\qquad = \sum_{i=1}^{n_{el}} \frac{\int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_i} y^{k-1} g'(x,y) x^{k-1} dxdy}{((k-1)!)^2 (x_i - x_{i-1})} \int_{x_{i-1}}^{x_i} \frac{d^{k-1} r}{dx^{k-1}} \frac{d^{k-1}\mathcal{L}^*\bar{v}}{dx^{k-1}}\, dx \\
&\qquad = \sum_{i=1}^{n_{el}} \tau_{k,(x_{i-1},x_i)} \int_{x_{i-1}}^{x_i} \frac{d^{k-1} r}{dx^{k-1}} \frac{d^{k-1}\mathcal{L}^*\bar{v}}{dx^{k-1}}\, dx.
\end{aligned}
\tag{3.24}
$$

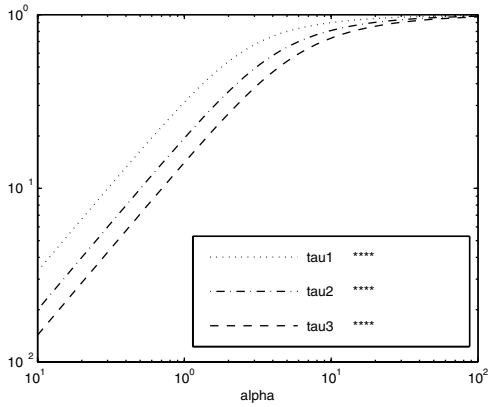


FIG. 3.5. *Stabilization parameters τ_1, τ_2, and τ_3 versus α.*

The stabilization term acts only locally and depends only on the derivative of degree $k-1$ of the residual; its effect is modulated by the parameter

$$\tau_k \equiv \tau_{k,(x_{i-1},x_i)} = \frac{\int_{x_{i-1}}^{x_i}\int_{x_{i-1}}^{x_i} y^{k-1}g'(x,y)x^{k-1}dxdy}{\left((k-1)!\right)^2(x_i - x_{i-1})}.$$

In the case of quadratic elements, from the previous formulas one can derive

$$\tau_2 = \frac{h^3}{72\,\beta}\frac{-3e^{2\alpha}\alpha^{-1} + e^{2\alpha} + 3e^{2\alpha}\alpha^{-2} - 3\alpha^{-2} - 1 - \alpha^{-1}}{e^{2\alpha} - e^{2\alpha}\alpha^{-1} + 1 + \alpha^{-1}},$$

while for cubic elements,

$$\tau_3 = \frac{h^5}{7200\,\beta}\,\frac{15e^{2\alpha}\alpha^{-2} - 6e^{2\alpha}\alpha^{-1} - 15e^{2\alpha}\alpha^{-3} + e^{2\alpha} + 15\alpha^{-3} + 6\alpha^{-1} + 15 + 1}{e^{2\alpha} - 3e^{2\alpha}\alpha^{-1} + 3e^{2\alpha}\alpha^{-2} - 1 - 3\alpha^{-2} - 3\alpha^{-1}}.$$

Plots of τ_1, τ_2, and τ_3 are compared in Figure 3.5.

Remark 1. From Figure 3.5 we see that the τ_k are positive and of order h^{2k-1}/β and $\alpha h^{2k-1}/\beta = h^{2k}/\kappa$ in the advective and in the diffusive regimes, respectively.

Remark 2. For linear elements, in one dimension, the H_0^1-optimal $\bar{u}$ is the nodal interpolant of u, which is a monotonicity preserving approximant. For higher-order elements, the H_0^1-optimal $\bar{u}$ is still nodally exact at the endpoints of each element, but we lose monotonicity inside the elements.

Remark 3. The format of (3.24) is reminiscent of the gradient least-squares stabilized method proposed by Franca and Dutra do Carmo [10].

3.3. L^2-optimality in one dimension and the localization of g'. We have shown that, for the one-dimensional problem and for the H_0^1-projector based VMS formulation (i.e., with $\mathcal{P} = \mathcal{P}_{H_0^1}$), the fine-scale Green's function is supported on the union of the $(x_{i-1}, x_i) \times (x_{i-1}, x_i)$ for $1 \le i \le n_{el}$. In this case, the g' is *fully localized* within each element, in that there is no coupling between elements. This allows a convenient evaluation of the fine-scale effect in the VMS formulation (see (3.15) and (3.24)). This feature, though, is not guaranteed for any projector $\mathcal{P}$. Take, e.g., the L^2-projector $\mathcal{P} = \mathcal{P}_{L^2}$, with piecewise-linear elements. We can still compute g' from (3.5), where now, in order to have (3.4), $\{\mu_i\}_{i=1,\dots,N}$ is a basis for $\bar{V}$ itself. For

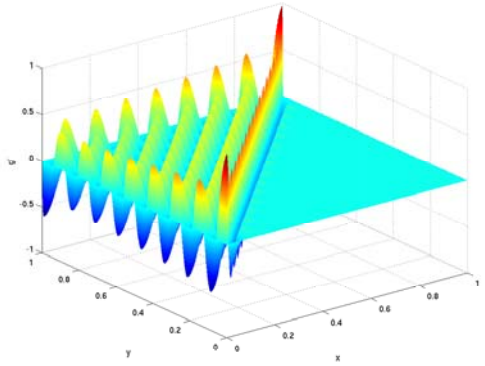


FIG. 3.6. *Fine-scale Green's function g' for the one-dimensional problem and linear elements, with $\mathcal{P} = \mathcal{P}_{L^2}$, $\kappa = 10^{-3}$, $\beta = 1$, $L = 1$, and a uniform grid of $n_{el} = 16$ elements. Note that in the case of $\mathcal{P} = \mathcal{P}_{L^2}$, g' is global and unattenuated when advection dominates.*

$\kappa = 10^{-3}$, $\beta = 1$, $L = 1$, and $n_{el} = 16$ elements, a plot of g' is presented in Figure 3.6; we see that the support of g' includes the entire upwind region $x \leq y$.

Remark 4. In practical applications, g' needs to be approximated, leading to classical stabilized methods. It is obviously more convenient and easy to approximate a highly localized g' than one that is global. This strongly suggests that the selection of the projector is crucial in the development of a multiscale method.

3.4. Linear elements in two dimensions. Turning to problems in two dimensions (as well as in multiple dimensions), we face two important differences.

First, it is technical and more difficult to obtain the analytical expression of the Green's function g and therefore of the fine-scale Green's function g' through (3.5). To overcome this difficulty, in this section we use the standard Galerkin method to numerically compute g and g' on a fine mesh of $524,288$ elements, which is able to resolve the fine scales of the problem under consideration.[1] We take here $\Omega = (0,1)^2$, the diffusivity is $\kappa = 10^{-3}$, and the unit advection velocity is $\beta = [1/2 \quad 1]/\sqrt{1.25}$. Since both g and g' are defined on $(0,1)^2 \times (0,1)^2$, for the purposes of a graphical representation we fix $y = y^* = [39/64 \quad 51/64] \approx [0.6 \quad 0.8]$ (see Figure 3.7), and we plot the Green's function g, and the fine-scale Green's function g' versus the argument x. The plot of $x \mapsto g(x, y^*)$ is presented in Figure 3.8. As is known, $x \mapsto g(x, y^*)$ is singular when $x = y^*$, and indeed the left graph in Figure 3.8 has been truncated at $g = 50$. Roughly speaking, g is supported around the upwind characteristic passing through y^*.

The second major difference, compared to the one-dimensional case, is that if the coarse scales are piecewise-polynomial, then the fine scales are not localized within each element (i.e., they are not bubble functions), and this happens for any choice of projector $\mathcal{P}$, including the H_0^1-projector. Indeed, since the coarse scales are polynomials on the edges of the elements of the triangulation, while the exact solution is arbitrary, the fine scales do not vanish there. Our aims here are the calculation of g'

[1] Let $\{\phi_i\}$ be a basis for piecewise-linear functions on the fine mesh, and think of $V \approx \text{span}\{\phi_i\}$, roughly speaking. Let $\mathbf{L}$ be the matrix representation of $\mathcal{L}$ (i.e., $L_{ij} = {}_{V^*}\langle \mathcal{L}\phi_j, \phi_i \rangle_V$), then the approximation of the Green's operator $\mathcal{G}$ is associated with $\mathbf{G} = \mathbf{L}^{-1}$. The approximation of $\mathcal{G}'$ in matrix form is obtained by $\mathbf{G}' = \mathbf{G} - \mathbf{G} \times \mathbf{P}^T \times (\mathbf{P} \times \mathbf{G} \times \mathbf{P}^T)^{-1} \times \mathbf{P} \times \mathbf{G}$, analogous to (2.11), where $\mathbf{P}$ is the matrix associated with the projector $\mathcal{P}$. The Green's function and fine-scale Green's function can be approximated as $g(x,y) \approx \sum_{i,j} G_{ij}\phi_j(x)\phi_i(y)$ and $g'(x,y) \approx \sum_{i,j} G'_{ij}\phi_j(x)\phi_i(y)$.

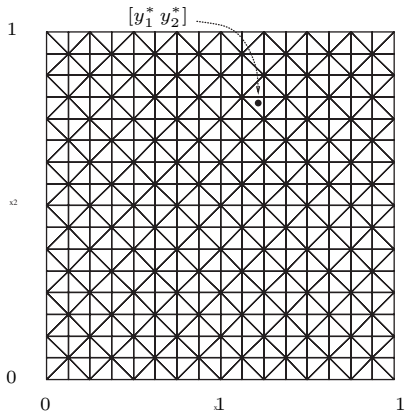


FIG. 3.7. *Mesh of the coarse-scale space $\bar{V}$ used for the calculation of the Green's functions in the two-dimensional case and the location of $y^* = [y_1^* \quad y_2^*]$.*

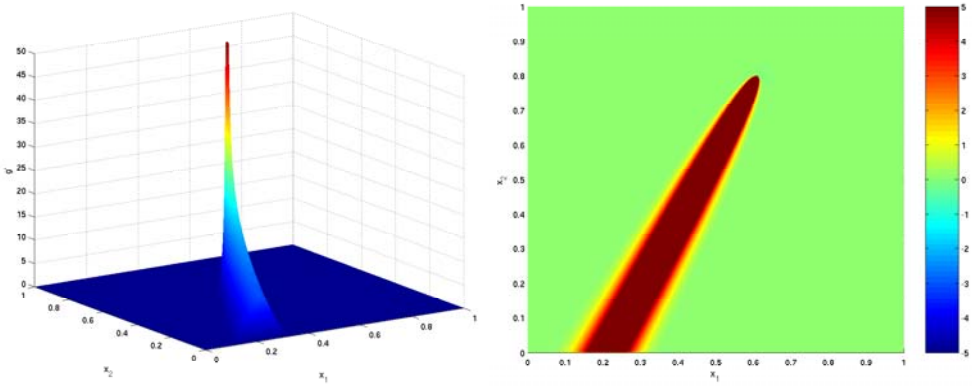


FIG. 3.8. *Plot (left) and contour plot (right) of $x \mapsto g(x, y^*)$.*

and the assessment of its attenuation compared with g and its locality for different choices of $\mathcal{P}$. We test both $\mathcal{P} = \mathcal{P}_{H_0^1}$ and $\mathcal{P} = \mathcal{P}_{L^2}$, which gave, in the one-dimensional case, a fully localized and a globally supported g', respectively. We take the space $\bar{V}$ formed by linear elements on the uniform triangulation shown in Figure 3.7. The plots of $x \mapsto g'(x, y^*)$ for $\mathcal{P} = \mathcal{P}_{H_0^1}$ and $\mathcal{P} = \mathcal{P}_{L^2}$ are presented in Figures 3.9 and 3.10, respectively. The singularity at $x = y^*$ is truncated at $g' = 50$ in the left-hand plots and at $g' = 5$ in the right-hand plots. Observe that in the case $\mathcal{P} = \mathcal{P}_{H_0^1}$, the fine-scale Green's function is more localized around y^*, compared with the case $\mathcal{P} = \mathcal{P}_{L^2}$, for which oscillations are spread over the entire domain. In addition, the g' for the case $\mathcal{P} = \mathcal{P}_{H_0^1}$ seems to be negligible outside a layer of a few elements around y^*. This is better seen in the (right-hand) contour plots of Figure 3.9 and 3.10, where the coarse mesh is overlaid.

Changing the position of y^* inside Ω and taking y^* on an edge or a vertex of the coarse triangulation produces similar results (not shown).

Remark 5. The upwind tail of g is global in the advection-dominated case, whereas it is highly attenuated for g' when $\mathcal{P} = \mathcal{P}_{H_0^1}$ (cf. Figure 3.8 with Figure 3.9). This has important implications for multiscale analysis of dominantly hyperbolic phenomena. In addition, the g' for $\mathcal{P} = \mathcal{P}_{H_0^1}$ is much more localized than that for

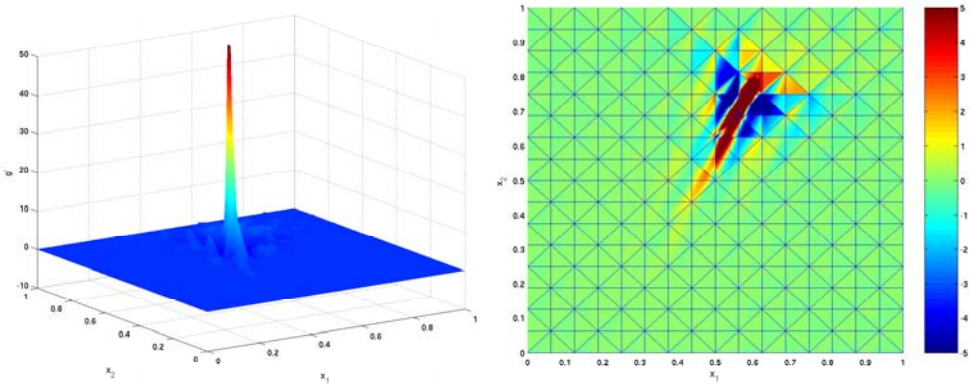


FIG. 3.9. *Plot (left) and contour plot (right) of* $x \mapsto g'(x, y^*)$ *for* $\mathcal{P} = \mathcal{P}_{H_0^1}$.

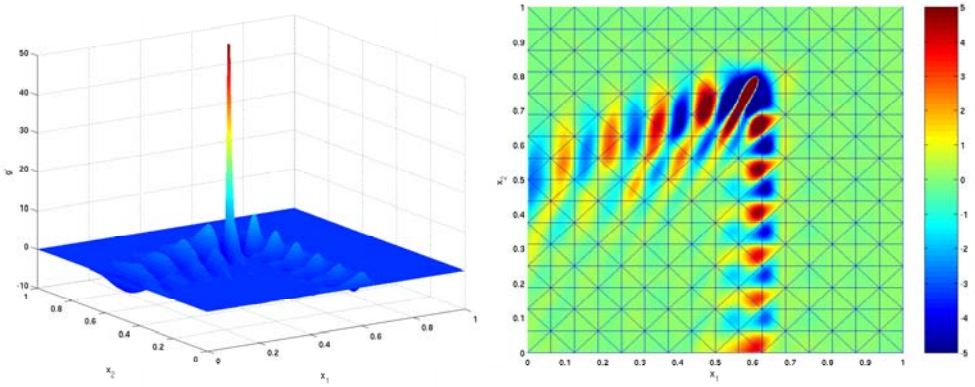


FIG. 3.10. *Plot (left) and contour plot (right) of* $x \mapsto g'(x, y^*)$ *for* $\mathcal{P} = \mathcal{P}_{L^2}$.

$\mathcal{P} = \mathcal{P}_{L^2}$. These results are consistent with the one-dimensional case and suggest that local approximations of g' for $\mathcal{P} = \mathcal{P}_{H_0^1}$ may achieve near H_0^1-optimality in multidimensional, advection-dominated cases.

Let us return to the model problem (3.1) with κ and β defined as above. We now consider a right-hand side $f = +1$ if $x_2 \geq 2x_1$, $f = -1$ if $x_2 < 2x_1$. The exact solution has an internal layer along $x_2 = 2x_1$, due to the discontinuity of f, and boundary layers at $x_1 = 1$ and $x_2 = 1$.

We consider three different meshes: The first two are shown in Figure 3.11, and the third is the same as the one depicted in Figure 3.7. The three meshes are quite coarse for the problem considered. The coarse-scale approximations $\bar{u}$ are given in Figures 3.12–3.14 for $\mathcal{P}_{H_0^1}$ and $\mathcal{P}_{L^2}$. In Figure 3.12 it is very clear that the solution for $\mathcal{P}_{H_0^1}$ is much better than that for $\mathcal{P}_{L^2}$. In Figure 3.13, the solution for $\mathcal{P}_{H_0^1}$ is better than that for $\mathcal{P}_{L^2}$ but not by as wide margin as in Figure 3.12. The trend continues in Figure 3.14, but the solution for $\mathcal{P}_{H_0^1}$ is only slightly better than that for $\mathcal{P}_{L^2}$. We have tested other meshes, obtaining results (not shown) similar to the ones of Figures 3.12–3.14. The superiority of $\mathcal{P}_{H_0^1}$ seems to be a general fact, though it is more apparent for finer meshes than coarser meshes. One might conclude that H_0^1-optimality is not as strong a condition as it is often thought to be and may not be enough in many practical cases for which monotonicity is deemed essential.

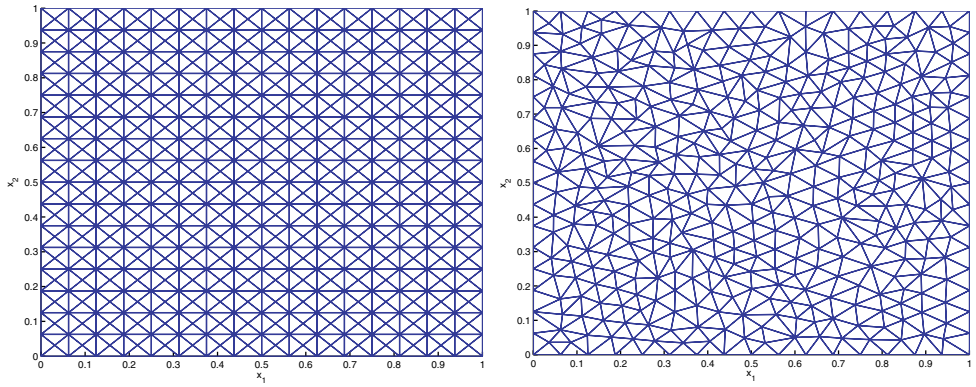


FIG. 3.11. *First two meshes associated with the coarse-scale spaces $\bar{V}$, used for the calculations of the coarse-scale components $\bar{u}$ of the model problem.*

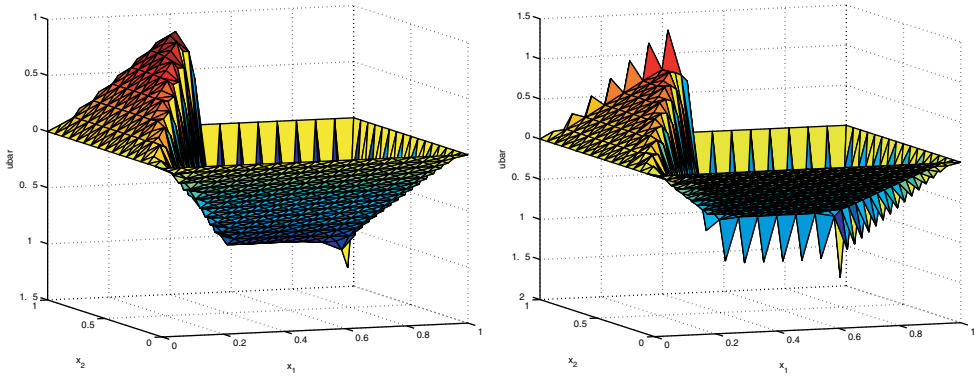


FIG. 3.12. *Coarse-scale component $\bar{u}$ for the model problem. $\mathcal{P} = \mathcal{P}_{H_0^1}$ (left) and $\mathcal{P} = \mathcal{P}_{L^2}$ (right). The coarse-scale space $\bar{V}$ is based on the left-hand mesh in Figure* 3.11.

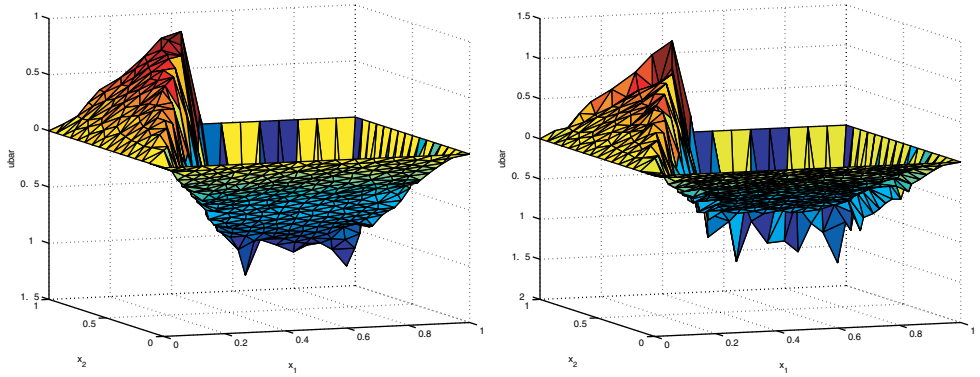


FIG. 3.13. *Coarse-scale component $\bar{u}$ for the model problem. $\mathcal{P} = \mathcal{P}_{H_0^1}$ (left) and $\mathcal{P} = \mathcal{P}_{L^2}$ (right). The coarse-scale space $\bar{V}$ is based on the right-hand mesh in Figure* 3.11.

Remark 6. In the case $\mathcal{P} = \mathcal{P}_{H_0^1}$, because of (3.6) we have, in the sense of distributions, $\int_\Omega g'(x,y)\Delta\bar{v}(y)\,dy = 0$ for all $\bar{v} \in \bar{V}$. Therefore, the fine-scale effect on

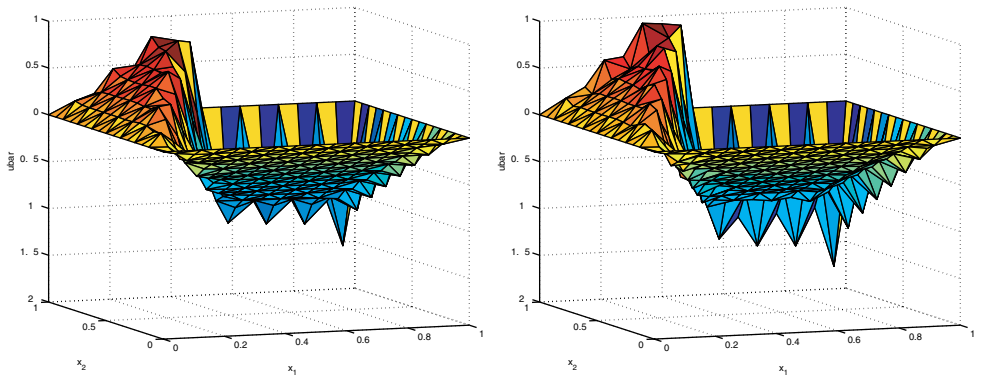


FIG. 3.14. *Coarse-scale component $\bar{u}$ for the model problem. $\mathcal{P} = \mathcal{P}_{H_0^1}$ (left) and $\mathcal{P} = \mathcal{P}_{L^2}$ (right). The coarse-scale space $\bar{V}$ is based on the mesh in Figure* 3.7.

the coarse-scale equation (3.7) becomes

$$\begin{aligned}\int_\Omega \int_\Omega \left(f(x) - \mathcal{L}\bar{u}(x)\right) g'(x,y) \mathcal{L}^* \bar{v}(y)\, dxdy \\ = -\int_\Omega \int_\Omega \left(f(x) - \mathcal{L}\bar{u}(x)\right) g'(x,y) \beta \cdot \nabla \bar{v}(y)\, dxdy.\end{aligned} \tag{3.25}$$

In one dimension, where g' is fully localized, the right-hand side of (3.25) is precisely the classical SUPG stabilization (see [9]); i.e., the residual is weighted only by the advective part of the operator, and the g' gives rise to the elementwise optimal τ, as described in sections 3.1 and 3.2. (Recall that f was assumed to be a piecewise-polynomial of degree at most $k-1$.) Note also that the diffusion operator in the residual in (3.25) may also be eliminated due to the aforementioned orthogonality property. These observations hold in higher dimensions as well, except g' is not fully localized within individual elements. In the classical multidimensional SUPG method [9], instead of (3.25) we have $-\sum_{e=1}^{n_{el}} \int_{\Omega_e} \left(f(x) - \mathcal{L}\bar{u}(x)\right) \tau(x)\, \beta \cdot \nabla \bar{v}(x)\, dx$, where Ω_e, $e = 1, \ldots, n_{el}$, are the elements of the mesh on Ω. The primary difference between SUPG and (3.25) is that g' is replaced by the elementwise constant τ. This approximation may be justified in light of the localized nature of g'. Indeed, SUPG has been shown to converge at optimal rates in higher dimensions (see, e.g., [19]), although in advection-dominated cases, the "stability" norm is not as strong as the H_0^1-seminorm in that it only contains the streamline derivative.

Remark 7. The residual-free bubble approach [4, 5, 6, 7, 8, 20, 21] has been shown in [3] to be equivalent to a multiscale method in which the fine-scale Green's function is approximated by a local element Green's function [12, 13]. Use of a local Green's function, in light of the framework described herein, can be rigorously justified only in the one-dimensional case in which the H_0^1-projector is employed. However, this amounts to a very convenient approximation in practice and one that is known to generate effective stabilized methods [1, 5, 6, 20]. With a better knowledge of g' in the multidimensional case, we would anticipate that improved stabilization schemes could be devised.

4. Conclusions. In this paper we have derived an expression for the fine-scale Green's function arising in VMS analysis. The specification of a projector, defining the direct-sum decomposition into coarse-scale and fine-scale components, renders the

problem for the fine-scale Green's function well-posed. Different projectors give rise to different fine-scale Green's functions, and their properties can vary considerably. It is felt to be beneficial if the fine-scale Green's function is more attenuated than the classical Green's function, and its support is dominantly local. It is found that the projector induced by the H_0^1-seminorm enjoys these properties whereas the projector induced by the L^2-norm does not.

The primary practical result of these studies is in the development of a framework for approximate multiscale methods. Indeed, in general it is not possible to exactly calculate the fine-scale Green's function. Despite its complexity, its orthogonality properties suggests simplified constructs in the form of stabilized methods. This is instantiated precisely in one dimension for the H_0^1-projector, and its possibility in higher dimensions is suggested as well. In fact, it is shown that the H_0^1-optimal method and SUPG have features in common.

The results obtained clarify the relationship between the fine-scale Green's function and the properties of the coarse-scale solution. However, we considered projectors associated only with inner products, and in particular, we studied only the H_0^1- and L^2-projectors. The coarse-scale solution achieves optimality in terms of the corresponding norm. One could conceive of requiring the coarse-scale solution to achieve optimality in other measures giving rise to nonlinear structure. This is an intriguing possibility in that one could, e.g., require monotonicity, or other desirable behavior. In the past, ad hoc procedures have been used to instill such properties in numerical methods, but the present ideas seem to have the potential for studying these issues in a more fundamental way.

Presently, most numerical methods are given as recipes, and they are evaluated ex post facto by the way they satisfy desired objectives. The present developments suggest a different approach: designing numerical methods to satisfy desired objectives ab initio. We are a long way from making this a practical reality, but we believe some small steps have been taken in this direction.

Acknowledgments. G. Sangalli thanks the Institute for Computational Engineering and Sciences (University of Texas at Austin) for kind hospitality. We thank Rich Lehoucq for insightful remarks and inspiration.

REFERENCES

[1] M. I. Asensio, A. Russo, and G. Sangalli, *The residual-free bubble numerical method with quadratic elements*, Math. Models Methods Appl. Sci., 14 (2004), pp. 641–661.

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

[3] F. Brezzi, L. P. Franca, T. J. R. Hughes, and A. Russo, $b = \int g$, Comput. Methods Appl. Mech. Engrg., 145 (1997), pp. 329–339.

[4] F. Brezzi, L. P. Franca, and A. Russo, *Further considerations on residual-free bubbles for advective-diffusive equations*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 25–33.

[5] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli, *A priori error analysis of residual-free bubbles for advection-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 1933–1948.

[6] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.

[7] F. Brezzi and L. D. Marini, *Augmented spaces, two-level methods, and stabilizing subgrids*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 31–46.

[8] F. Brezzi and A. Russo, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.

[9] A. N. Brooks and T. J. R. Hughes, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[10] L. P. Franca and E. G. Dutra do Carmo, *The Galerkin gradient least-squares method*, Comput. Methods Appl. Mech. Engrg., 74 (1989), pp. 41–54.

[11] J. Holmen, T. J. R. Hughes, A. A. Oberai, and G. N. Wells, *Sensitivity of the scale partition for variational multiscale large-eddy simulation of channel flow*, Phys. Fluids, 16 (2004), pp. 824–827.

[12] T. J. R. Hughes, *Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 387–401.

[13] T. J. R. Hughes, G. R. Feijóo, L. Mazzei, and J.-B. Quincy, *The variational multiscale method—A paradigm for computational mechanics*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 3–24.

[14] T. J. R. Hughes, L. Mazzei, and K. E. Jansen, *Large eddy simulation and the variational multiscale method*, Comput. Vis. Sci., 3 (2000), pp. 47–59.

[15] T. J. R. Hughes, L. Mazzei, A. A. Oberai, and A. A. Wray, *The multiscale formulation of large eddy simulation: Decay of homogeneous isotropic turbulence*, Phys. Fluids, 13 (2001), pp. 505–512.

[16] T. J. R. Hughes, A. A. Oberai, and L. Mazzei, *Large eddy simulation of turbulent channel flows by the variational multiscale method*, Phys. Fluids, 13 (2001), pp. 1784–1799.

[17] T. J. R. Hughes and G. Sangalli, *Variational multiscale analysis: The fine-scale Green's function, projection, optimization, localization, and stabilized methods*, Technical report 05-46, The Institute for Computational Engineering and Sciences, Austin, TX, 2005.

[18] T. J. R. Hughes, G. N. Wells, and A. A. Wray, *Energy transfers and spectral eddy viscosity in large-eddy simulations of homogeneous isotropic turbulence: Comparison of dynamic Smagorinsky and multiscale models over a range of discretizations*, Phys. Fluids, 16 (2004), pp. 4044–4052.

[19] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[20] G. Sangalli, *Global and local error analysis for the residual-free bubbles method applied to advection-dominated problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1496–1522.

[21] G. Sangalli, *Capturing small scales in elliptic problems using a residual-free bubbles finite element method*, Multiscale Model. Simul., 1 (2003), pp. 485–503.

[22] I. Stakgold, *Green's Functions and Boundary Value Problems*, Pure Appl. Math. (NY), 2, Wiley-Interscience, New York, 1998.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 558–571

COMPUTING THE GAMMA FUNCTION USING CONTOUR INTEGRALS AND RATIONAL APPROXIMATIONS*

THOMAS SCHMELZER† AND LLOYD N. TREFETHEN†

Abstract. Some of the best methods for computing the gamma function are based on numerical evaluation of Hankel's contour integral. For example, Temme evaluates this integral based on steepest descent contours by the trapezoid rule. Here we investigate a different approach to the integral: the application of the trapezoid rule on Talbot-type contours using optimal parameters recently derived by Weideman for computing inverse Laplace transforms. Relatedly, we also investigate quadrature formulas derived from best approximations to $\exp(z)$ on the negative real axis, following Cody, Meinardus, and Varga. The two methods are closely related, and both converge geometrically. We find that the new methods are competitive with existing ones, even though they are based on generic tools rather than on specific analysis of the gamma function.

Key words. gamma function, Hankel contour, numerical quadrature

AMS subject classifications. 65D20, 33F05

DOI. 10.1137/050646342

1. The gamma function. In his childhood Gauss rediscovered that the sum of the first n positive integers is given by

$$\sum_{k=1}^{n} k = \frac{n\,(n+1)}{2},$$

a formula which can be considered as an interpolation valid even for nonintegers. Starting in 1729 Euler discussed in a series of three letters to Goldbach, well known for the Goldbach conjecture, the problem of the product of the first n integers, which is today known as the factorial of n, $n!$. Davis [6] gives details about the history of the gamma function. We start here with the standard definition

$$\Gamma\,(z) = \int_0^\infty t^{z-1} e^{-t} dt, \qquad \text{Re } z > 0, \tag{1.1}$$

where

$$t^{z-1} = e^{(z-1)\log t} \text{ and } \log t \in \mathbb{R}.$$

The gamma function is analytic in the open right half-plane. Partial integration yields

$$\Gamma\,(z+1) = z\Gamma\,(z)\,, \tag{1.2}$$

and since $\Gamma\,(1) = 1$, we have

$$\Gamma\,(n+1) = n!.$$

*Received by the editors November 29, 2005; accepted for publication (in revised form) November 13, 2006; published electronically March 19, 2007.
http://www.siam.org/journals/sinum/45-2/64634.html
†Computing Laboratory, Oxford University, Oxford OX1 3QD, United Kingdom (thoms@comlab.ox.ac.uk, lnt@comlab.ox.ac.uk). The first author was supported by a Rhodes scholarship.

Any confusion caused by this identity dates back to Legendre. It is possible to continue the gamma function analytically into the left half-plane. This is often done by a representation of the reciprocal gamma function as an infinite product [1, eq. 6.1.2]:

$$\frac{1}{\Gamma(z)} = \lim_{n\to\infty} \frac{n^{-z}}{n!} z(z+1)\dots(z+n)$$

valid for all z. This representation shows that $\Gamma(z)$ has poles for $z = 0, -1, -2, \dots$. Of more practical use is the reflection formula [1, eq. 6.1.17]:

$$\Gamma(z)\Gamma(1-z) = \frac{\pi}{\sin \pi z}, \qquad z \notin \mathbb{Z}. \tag{1.3}$$

This identity implies $\Gamma(1/2) = \sqrt{\pi}$. It is standard to approximate the gamma function only for $\operatorname{Re} z \geq 1/2$ and to exploit (1.3) for $\operatorname{Re} z < 1/2$.

2. Hankel's representation. An alternative representation for the reciprocal gamma function, which is an entire function, is due to Hankel [11]. Substituting $t = su$ in (1.1) yields

$$F(s) := \frac{\Gamma(z)}{s^z} = \int_0^\infty u^{z-1} e^{-su} du,$$

which can be regarded as the Laplace transform of u^{z-1} for fixed complex z. Hence u^{z-1} can be interpreted as an inverse Laplace transform:

$$u^{z-1} = \mathcal{L}^{-1}\{F(s)\} = \frac{1}{2\pi i}\int_{\mathcal{C}} e^{ku} F(k) dk \ = \frac{1}{2\pi i}\int_{\mathcal{C}} e^{ku} \frac{\Gamma(z)}{k^z} dk.$$

The path $\mathcal{C}$ is any deformed Bromwich contour such that $\mathcal{C}$ winds around the negative real axis in the anticlockwise sense (see Figure 1). Now we substitute $s = ku$, which yields

$$u^{z-1} = \frac{1}{2\pi i}\int_{\mathcal{C}} e^s \frac{\Gamma(z) u^z}{s^z u} ds$$

and hence

$$\frac{1}{\Gamma(z)} = \frac{1}{2\pi i}\int_{\mathcal{C}} s^{-z} e^s ds. \tag{2.1}$$

The numerical evaluation of integrals of the form

$$I = \frac{1}{2\pi i}\int_{\mathcal{C}} e^s f(s) ds \tag{2.2}$$

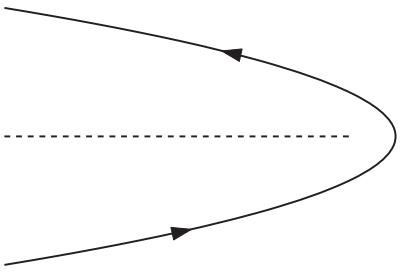


FIG. 1. *A typical Hankel contour, winding around the negative real axis (dashed) in the anticlockwise sense.*

has been discussed by Trefethen, Weideman, and Schmelzer [22]. The function s^{-z} has a branch cut on $\mathbb{R}^- = (-\infty, 0]$ but is analytic everywhere else. Hence (2.2) is independent of $\mathcal{C}$ under mild assumptions. The freedom to choose the path for inverse Laplace transforms has aroused a good deal of research interest. Recently Weideman and co-workers [22, 25, 24] have optimized parameters for the cotangent contours introduced by Talbot [19] as well as for other contours in the form of parabolas and hyperbolas. Here we focus on different numerical methods for which (2.1) is the common basis. In particular we shall compare

1. steepest descent contours,
2. Talbot-type contours,
3. rational approximation of e^s on $(-\infty, 0]$.

The first of these methods is an existing one, and the other two are new. Methods we do not compare are those of Spouge, Lanczos, and Stirling. Comments on these and on what is done in practice can be found in section 7.

In addition we mention in section 6 a generalization of (2.1) for matrices and introduce an idea for solving linear systems of the form $\Gamma(A)x = c$ without computing $\Gamma(A)$.

3. Saddle point method. Saddle point methods in general are extensively discussed in the book by Bender and Orszag [3, sect. 6.6]. The reciprocal gamma function is a standard example for this technique presented in this and many other textbooks. We keep the details to a minimum and follow an approach of Temme [20], who advocates the numerical evaluation of the integral along a steepest descent contour. A zero of the first derivative of an analytic function f indicates a saddle point of $\left|e^f\right|$. Through this point runs a path $\mathcal{C}$ where f has a constant imaginary part and a decreasing real part. This is a very desirable property for asymptotic analysis and numerical quadrature schemes. In order to apply these ideas here we fix the movable saddle by a change of variable $s = zt$. We get

$$\frac{1}{\Gamma(z)} = \frac{e^z z^{1-z}}{2\pi i} \int_{\mathcal{C}} e^{z\phi(t)} dt, \tag{3.1}$$

where $\phi(t) = t - 1 - \ln t$. If z is real and positive, then the integrand in (3.1) decreases exponentially as t moves away from 1 along the steepest descent contour. For complex z, on the other hand, the decrease becomes oscillatory, and in the limit of pure imaginary z, there is no decrease at all. Thus let us assume that z is a positive real number. Let $t = \rho e^{i\theta}$ be the steepest descent path parameterized by the radius ρ and the argument θ. The vanishing imaginary part at $t = 1$ induces the equation

$$0 = \operatorname{Im} \phi(t) = \rho \sin\theta - \theta.$$

Hence the path is given by $\rho = \theta / \sin\theta$. Temme [20] gives the reparameterization

$$\frac{1}{\Gamma(z)} = \frac{e^z z^{1-z}}{2\pi} \int_{-\pi}^{\pi} e^{-z\Phi(\theta)} d\theta,$$

where

$$\Phi(\theta) = 1 - \theta \cot\theta + \ln \frac{\theta}{\sin\theta}$$

with $\Phi(0) = 0$. Note that the real part of $dt/d\theta = \left(\cot\theta - \theta \csc^2\theta\right) + i$ is an odd function of θ.

The integral can be approximated by the midpoint rule, which is exponentially accurate. See [21] for a review of this phenomenon of high accuracy. The approximated integral is

$$I_N(z) = \frac{e^z z^{1-z}}{N} \sum_{k=1}^{N} e^{-z\Phi(\theta_k)}, \tag{3.2}$$

where the nodes are

$$\theta_k = -\pi + \left(k - \frac{1}{2}\right)\frac{2\pi}{N}, \qquad 1 \le k \le N.$$

This set of nodes is exponentially accurate, but it is not optimal for large z, for the nodes closer to $-\pi$ and π contribute negligibly because of the fast decay along the path. We could delete some of these points to make the method even more efficient, truncating the interval to $[-\tau, \tau]$ instead of $[-\pi, \pi]$.

4. Direct contour integration. Instead of working with saddle points, another approach is to apply the trapezoidal rule directly to (2.1). This makes it easy to evaluate $\Gamma(z)$ for complex as well as real arguments. Let $\phi(\theta)$ be an analytic function that maps the real line $\mathbb{R}$ onto the contour $\mathcal{C}$. Then (2.1) can be written as

$$I = \frac{1}{2\pi i} \int_{-\infty}^{\infty} \phi(\theta)^{-z} e^{\phi(\theta)} \phi'(\theta)\, d\theta. \tag{4.1}$$

Because of the term $e^{\phi(\theta)}$, the integrand decreases exponentially as $|\theta| \to \infty$ so that one commits an exponentially small error by truncating $\mathbb{R}$ to a finite interval. For simplicity we shall arbitrarily fix this interval as $[-\pi, \pi]$. In $[-\pi, \pi]$ we take N points θ_k spaced regularly at a distance $2\pi/N$, and our trapezoid approximation to (2.1) becomes

$$I_N = -iN^{-1} \sum_{k=1}^{N} e^{s_k} s_k^{-z} w_k, \tag{4.2}$$

where $s_k = \phi(\theta_k)$ and $w_k = \phi'(\theta_k)$. MATLAB codes are given in Figure 2.

Note that there is still the freedom left to choose a particular path. In Program 31 of the textbook [23], a closed circle with center $c = -11$ and radius $r = 16$ is used with 70 equidistant nodes on it. Although this contour crosses the branch cut, it does so sufficiently far down the real axis that the error introduced thereby is less than 10^{-11}.

A more systematic approach has been pursued by Weideman and co-workers [22, 25, 24], who have proposed, in particular, parameters for parabolic, hyperbolic, and cotangent contours:

1. Parabolic contour

$$s(\theta) = N\left[0.1309 - 0.1194\theta^2 + 0.2500i\theta\right], \tag{4.3}$$

2. Hyperbolic contour

$$s(\theta) = 2.246N\left[1 - \sin(1.1721 - 0.3443i\theta)\right], \tag{4.4}$$

3. Cotangent contour

$$s(\theta) = N\left[0.5017\theta\cot(0.6407\theta) - 0.6122 + 0.2645i\theta\right]. \tag{4.5}$$

```
function I = ContourIntegral(z,contour,N,f)
  [s,w] = feval(contour,N);                    % contour is a function
  I = zeros(size(z));                          % the different sums
  for k = 1:N                                  % quadrature via
    I = I+w(k)*exp(s(k)).*feval(f,s(k),z);     % evaluating f at the nodes
  end

function [s,w] = contourCot(N)
  t = (-N+1:2:N-1)*pi/N;                       % angles theta
  a = 0.5017; b = 0.2645i; ct = 0.6407*t; d = 0.6122;
  s = N*(a*t.*cot(ct)-d+b*t).';                % poles
  w = -i*(a*cot(ct)-a*ct./sin(ct).^2+b).';     % weights

function f = IntGamma(s,z)
  % for the reciprocal gamma function
  f = s.^(-z);
```

FIG. 2. *MATLAB codes to evaluate* (2.2) *by* (4.2). *The function* $f(s) = s^{-z}$ *and the contour* $\mathcal{C}$ *are defined in separate M-files and addressed as handles.*

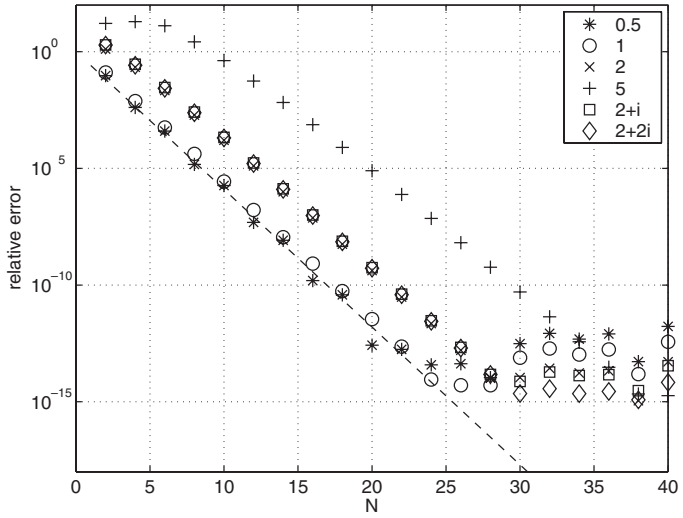
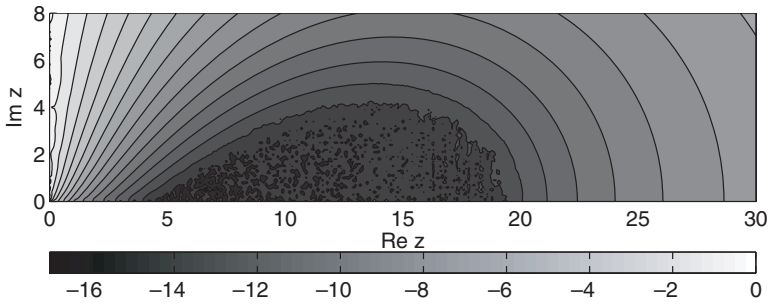


FIG. 3. *Convergence of* I_N *to* $1/\Gamma(z)$ *for the cotangent contour* (4.2), (4.5), *for six different values of* z. *The dashed line shows* 3.89^{-N}, *confirming Weideman's analysis.*

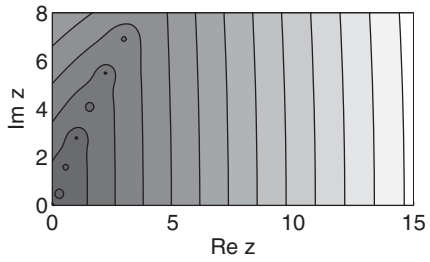
Using equidistant nodes with respect to θ, all of these contours show geometric convergence at rates approximately $O(3^{-N})$. Figure 3 illustrates this behavior by showing convergence as $N \to \infty$ for six values of z. According to Weideman the convergence rate for the cotangent contour is $O(3.89^{-N})$, which is shown as a dashed line in the figure.

In Figure 4, this behavior is compared in a region of the z-plane to the convergence for the parabolic and hyperbolic contours, the steepest descent contours, and the method of rational approximation to be introduced in the next section. All the methods are geometrically convergent (except steepest descents near the imaginary axis), and the cotangent contours and rational approximations are the best.

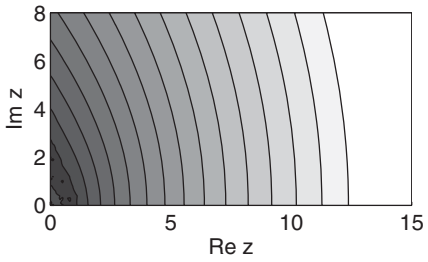
For all of these Talbot-type contours we encounter the same nonoptimality effect as for the saddle point method: The decay of the integrand is so fast that the left-



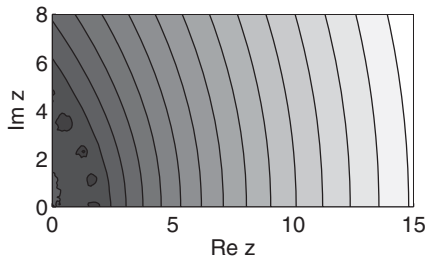
(a) *Saddle point method* (3.2), $N = 32$.



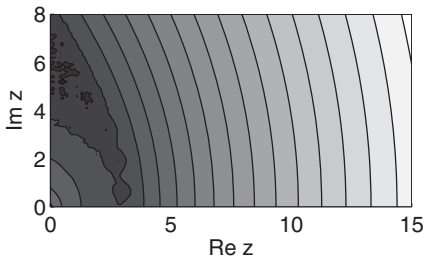
(b) *Circular contour from* [23], $N = 70$.



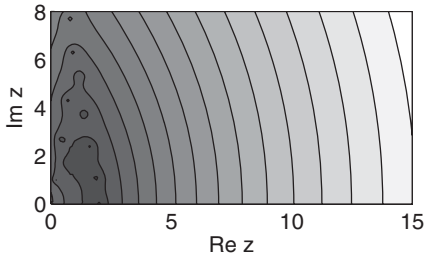
(c) *Parabolic contour* (4.3), $N = 32$.



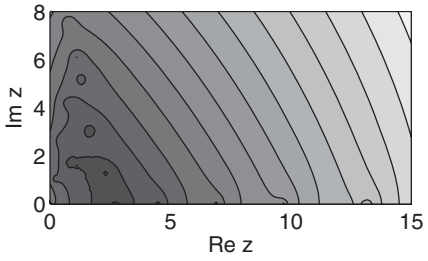
(d) *Hyperbolic contour* (4.4), $N = 32$.



(e) *Cotangent contour* (4.5), $N = 32$.



(f) *CMV approximation* (5.1) *with no shift*, $N = 16$.



(g) *CMV approximation* (5.1) *with shift* $b = 1$, $N = 16$.

FIG. 4. *Relative error in evaluating* $\Gamma(z)$ *in various points of the z-plane. The color bar in* (a) *indicates the scale for all seven plots (logs base* 10*). In practice, one would improve accuracy by reducing values of z to a fundamental strip, as shown in Figures* 5 *and* 8.

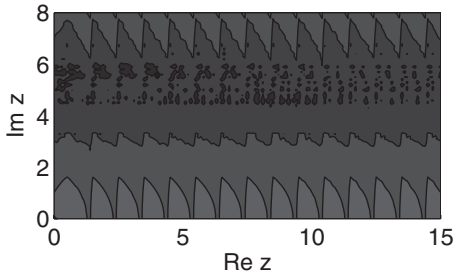


FIG. 5. *Relative error in evaluating* $\Gamma(z)$ *using a cotangent contour* (4.5), $N = 32$ *in* $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$ *and applying* (1.2) *and* (1.3) *for other points of the z-plane. The shading is the same as in Figure* 4.

```
% gammatalbot - Thomas Schmelzer & Nick Trefethen November 2005
%
% For real arguments this is around 20 times slower than Matlab's
% gamma, a factor roughly equal to the product of:
%   5 since this is an M-file rather than a .mex file
%   2 since it uses Talbot quadrature rather than best approximation
%   2 since the real symmetry is not exploited in the sum

function g = gammatalbot(z)                % complex Gamma function
  r = find(real(z)<0.5);                   % reflect to real(z)>=0.5
  z(r) = 1-z(r);
  shift = floor(real(z)-0.5);              % shift to fundamental strip
  zz = z-shift;
  g = 1./ContourIntegral(zz,@contourCot,32,@IntGamma);

  while any(shift)>0
    f = find(shift>0);
    g(f) = g(f).*zz(f);
    shift(f) = shift(f)-1;
    zz(f) = zz(f)+1;
  end
  g(r) = -pi./(g(r).*sin(pi*(z(r)-1)));    % reflect back
  j = find(imag(z)==0); g(j) = real(g(j)); % real inputs -> real outputs
```

FIG. 6. *A MATLAB routine for computing the gamma function. The fundamental identities* (1.2) *and* (1.3) *are used to reduce all arguments to the strip* $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$. *The code makes use of the functions listed in Figure* 2.

most nodes make a negligible contribution. The source of this phenomenon is the fact that Weideman's analysis considers only the factor e^s in (2.1), treating the factor s^{-z} as of order 1, whereas in fact, when z has a large real part, s^{-z} is very small. This effect is ubiquitous when computing with a fixed path and fixed nodes for all $z \in \mathbb{C}$. We could take advantage of it by fine-tuning Weideman's parameters in a manner specific to the gamma function, but we shall not do that here since our interest is in the application of generic methods for integrals of the form (2.2). Also, it is simpler and just as effective to use the fundamental identities (1.2) and (1.3) to reduce all arguments to the strip $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$. The effect of such reductions is illustrated for the cotangent contour in Figure 5. A MATLAB routine implementing this strategy is given in Figure 6.

5. Rational approximation. In a recent paper we, along with Weideman, interpreted the trapezoidal rule on a Hankel contour as a rational approximation of $\exp(z)$ on the negative real axis [22]. The analysis of best Chebyshev approximations of this kind is a problem made famous by Cody, Meinardus, and Varga (CMV) [5]; the errors are known to decrease asymptotically at the rate $O(H^N)$, where $H = 1/9.28903\ldots$ is known as *Halphen's constant* [10]. As shown in [22], these approximations can be used directly to evaluate integrals (2.2), bypassing the consideration of Talbot contours and the trapezoid rule. Given N, we define the best type (N, N) approximation to $\exp(s)$ to be the unique real rational function r_N^* of type (N, N) such that

$$\sup_{s\in\mathbb{R}^-} |r_N^*(s) - \exp(s)| = \inf_{r\in R_N} \sup_{s\in\mathbb{R}^-} |r(s) - \exp(s)|\,,$$

where R_N denotes the set of all rational functions of type (N, N). The coefficients of the polynomials in the numerator and denominator of r_N^* are given to very high accuracy in a paper by Carpenter, Ruttan, and Varga [4]. A practical way of determining these approximants on the fly is the Carathéodory–Fejér (CF) method. (In principle, the CF approximation is not best but near-best, but its difference from the true best approximation is negligible for $N \geq 2$ [22].) The function r_N^* can be represented in a partial fraction representation, i.e., by N poles $p_1, \ldots, p_N$ and residues $c_1, \ldots, c_N$ such that

$$r_N^*(s) = \sum_{k=1}^{N} \frac{c_k}{s - p_k} + c_0.$$

We define $\tilde{r}_N(s)$ to be the portion of this expression in the sum, i.e., $\tilde{r}_N(s) = r_N^*(s) - r_N^*(\infty)$, a rational function of type $(N-1, N)$ whose deviation from $\exp(s)$ on $\mathbb{R}^-$ decreases at the same asymptotic rate as that of r_N^* as $N \to \infty$.

These rational approximants can be used as the basis of another method for evaluating $1/\Gamma(z)$. We simply replace e^s in (2.1) by $\tilde{r}_N$ to obtain, with the aid of residue calculus,

$$I_N = \frac{1}{2\pi i} \int_{\mathcal{C}} \tilde{r}_N(s) s^{-z} ds = -\sum_{k=1}^{N} c_k p_k^{-z}, \tag{5.1}$$

which converges for $\mathrm{Re}\ z > 0$ as the decay of the integrand at infinity is fast enough. For $\mathrm{Re}\ z > 1$ we also have

$$I_N = \frac{1}{2\pi i} \int_{\mathcal{C}} r_N^*(s) s^{-z} ds. \tag{5.2}$$

For even N the poles come in conjugate pairs and (5.1) simplifies for real z to

$$I_N = -\sum_{k=1}^{N/2} 2\mathrm{Re}\left(c_k p_k^{-z}\right)$$

provided the first $N/2$ poles are all in the upper half-plane or all in the lower half-plane.

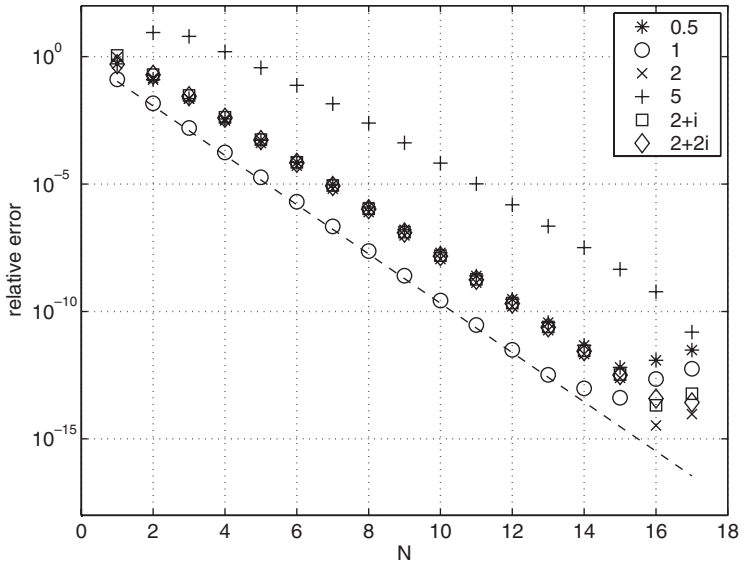


FIG. 7. *Convergence for the near-best rational approximation* (5.1) *of type* $(N-1,N)$ *with no shift. The convergence is about twice as fast as in Figure* 3*, with fifteen integrand evaluations sufficing to produce near machine precision. The dashed line shows* 9.28903^{-N}*, confirming Theorem* 5.2.

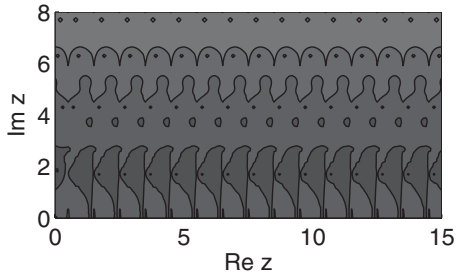


FIG. 8. *Relative error in evaluating* $\Gamma(z)$ *using a CMV approximation,* $N = 16$ *with no shift solely in* $\frac{1}{2} \le \mathrm{Re}\, z < \frac{3}{2}$*, and applying* (1.2) *and* (1.3) *for other points of the z-plane. The shading is the same as in Figure* 4.

For each z satisfying $\mathrm{Re}\, z > 0$ or $\mathrm{Re}\, z > 1$ as appropriate, I_N appears to converge to $1/\Gamma(z)$ at a geometric rate controlled by the same constant $H = 1/9.28903\ldots$ as indicated in Figures 7 and 8. A proof of this claim would follow from the following result, which we believe is true but have not yet proved.

CONJECTURE 5.1. *Let* $\{r_N^*\}$ *be the best approximations over* $\mathbb{R}^-$ *as defined above, let* K *be a compact set in* $\mathbb{C}$*, and let* $\|\cdot\|_K$ *denote the supremum norm over* K*. Then*

$$\limsup_{N\to\infty} \|\exp(s) - r_N^*(s)\|_K^{1/N} \le H = \frac{1}{9.28903\ldots}.$$

Here is the result that follows from the conjecture.

THEOREM 5.2. *Let* $\{\tilde{r}_N\}$ *and* $\{r_N^*\}$ *be the rational approximations defined above and let* z *be fixed with* $\mathrm{Re}\, z > 0$*. Then the approximations* (5.1) *and* (5.2) *(provided*

Re $z > 1$) *satisfy*

$$\limsup_{N\to\infty} \left| \frac{1}{\Gamma(z)} - I_N(z) \right|^{1/N} \leq H = \frac{1}{9.28903\ldots}.$$

Partial proof, assuming the validity of Conjecture 5.1. We introduce a special Hankel contour $\mathcal{C}_\rho$. It consists of a circle of radius ρ enclosing the origin and two rays joining $\rho e^{-i\pi}$ and $\rho e^{+i\pi}$ with the point $-\infty$. An upper bound for the error is deduced on $\mathcal{C}_\rho$. For the case of r_N^*, for example, we get by using (2.1) and (5.2)

$$\left| \frac{1}{\Gamma(z)} - I_N(z) \right| \leq \frac{1}{2\pi} \left\| r_N^*(s) - \exp(s) \right\|_{\mathcal{C}_\rho} \int_{\mathcal{C}_\rho} \left| s^{-z} \right| |ds|,$$

and we note that for any s, $|s^{-z}| \leq |s|^{-a} e^{\pi|b|}$ for $z = a + bi$ with $a > 1$. From here we readily obtain

$$\int_{\mathcal{C}_\rho} \left| s^{-z} \right| |ds| \leq \left(2\pi + \frac{2}{a-1} \right) e^{|b|\pi} \rho^{1-a}.$$

The convergence of $r_N^*(s)$ to $\exp(s)$ on the circle of radius ρ can be estimated by Conjecture 5.1, and therefore

$$\limsup_{N\to\infty} \left| \frac{1}{\Gamma(z)} - I_N(z) \right|^{1/N} \leq H.$$

It remains to show that the result just proved for r_N^* and Re $z > 1$ also holds for $\tilde{r}_N$ and Re $z > 0$. To do this split up the integral to obtain the estimate

$$\left| \frac{1}{\Gamma(z)} - I_N(z) \right| \leq \frac{1}{2\pi} \left\| s\left(\tilde{r}_N(s) - \exp(s)\right) \right\|_{\mathcal{C}_\rho} \int_{\mathcal{C}_\rho} \left| s^{-z-1} \right| |ds|.$$

The function $s\left(\tilde{r}_N - \exp(s)\right)$ in the left-hand term of this estimate approaches a constant as $s \to -\infty$ for each N, since $\tilde{r}_N - \exp(s)$ decreases at the rate $O(s^{-1})$. The essential point in showing that these Nth roots approach H as required is to make sure that the leftmost extremum of $\tilde{r}_N(s) - \exp(s)$ does not occur at a value of s that is exponentially large, in which case the Nth root of this value of s might fail to converge to 1. In fact, the results of Aptekarev [2] and Magnus [13] appear to confirm numerical evidence that the location of this extremum grows just algebraically, but we will not attempt a rigorous proof here. □

The fundamental property $\exp(a+b) = \exp(a)\exp(b)$ for any two complex arguments can be exploited in our algorithm. Given a positive parameter b, the function $\tilde{r}_N^b(s) = \exp(b)\tilde{r}_N(s-b)$ can be regarded as an approximation of $\exp(s)$ in the interval $(-\infty, b]$. In particular, (5.1) is the special case of this approximation for $b = 0$:

$$I_N^b = \frac{1}{2\pi i} \int_{\mathcal{C}} \tilde{r}_N^b(s) s^{-z} ds = -\sum_{k=1}^{N} e^b c_k (p_k + b)^{-z}. \tag{5.3}$$

It is easily proved that the *shifted* rational approximation $\tilde{r}_N^b(s)$ still converges with the same asymptotic rate H^N. In experiments we have observed that a shift of $O(1)$ gives better results especially for real arguments, as illustrated in Figures 7 and 9 and 4(f) and 4(g), where we used a shift of $b = 1$.

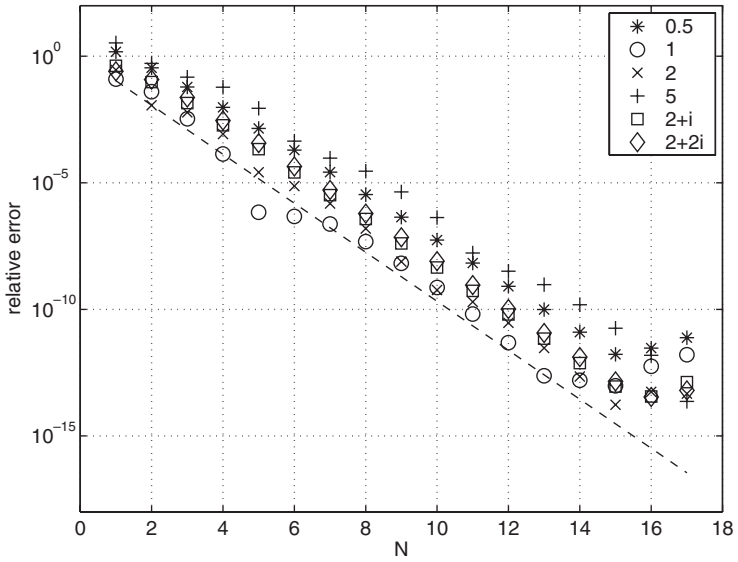


FIG. 9. *Convergence for the near-best rational approximation* (5.1) *of type* $(N-1, N)$ *with shift* $b = 1$. *Though the asymptotic behavior is the same, the constants are better than in Figure* 7, *and the use of such a shift might be a good idea in practice.*

6. Matrix arguments. Hankel's contour integral (2.1) can be generalized to square matrices A, and one can apply the methods introduced here to compute $\Gamma(A)^{-1}$ or to compute the solution vector x in a linear system $\Gamma(A)x = c$ without computing $\Gamma(A)$. We have confirmed this by numerical experiments not reported here. A drawback of such methods is that it is expensive to compute $s_k^{-A}c$ for every node; methods based on the algorithms of Spouge [18] and Lanczos [12] might be more efficient. We are currently not aware of applications where $\Gamma(A)$ is used for matrix arguments.

7. Other methods and existing software. There are a variety of existing methods for computing the gamma function. Are our methods competitive with these? As far as we can tell, the answer seems to be yes; they are "in the ballpark" in the sense of coming within a factor of 1–10 of the best methods, notably

- the method of Lanczos [12],
- the method of Spouge [18],
- the asymptotic Stirling series [1, eq. 6.1.37].

We emphasize that these methods are specialized algorithms designed for computing the gamma function and its close relatives, whereas our ideas are applicable in a much larger framework.

7.1. The method of Spouge. The method of Spouge [18] is attractive because of its simplicity and precise error estimates. Spouge introduced the approximation

$$\Gamma(z+1) \approx (z+\gamma)^{z+1/2} e^{-(z+\gamma)} \sqrt{2\pi} \left[c_0 + \sum_{k=1}^{N} \frac{c_k(\gamma)}{z+k} \right],$$

which is valid for $\mathrm{Re}\,(z+\gamma) > 0$ and dependent on a positive real parameter γ with $N = \lceil \gamma \rceil - 1$, which converges to an equality as $\gamma \to \infty$. Here $c_0 = 1$, and the other

coefficients are given by

$$c_k(\gamma) = \frac{1}{\sqrt{2\pi}} \frac{(-1)^{k-1}}{(k-1)!} (-k+\gamma)^{k-1/2} e^{-k+\gamma}, \qquad 1 \le k \le N.$$

The absolute error for this approximation can be bounded [18, Thm. 1.3.1] by

$$E_N(z) \le \left| \gamma(z) \frac{1}{\sqrt{N+1}(2\pi)^{N+3/2}} \right|.$$

Note that the relative error does not depend on z, making Spouge's method especially attractive for uniform approximations in the right half-plane. The above inequality implies that the method converges at least as fast as (6.28^{-N}), a rate lying midway between (3.89^{-N}) for Talbot contours and (9.29^{-N}) for best rational approximations. Actually, experiments suggest a better convergence rate, closer to $O(10^{-N})$.

7.2. The method of Lanczos. The method of Lanczos [12] is closely related to that of Spouge. Lanczos's method is based on the fast evaluation of the integral

$$F_\gamma(z) = \int_0^e \left[v(1-\log v)\right]^z v^\gamma dv,$$

where γ is a positive free parameter. The integral is approximated by a rational function

$$F_{N,\gamma}(z) = a_0 + \sum_{k=1}^{N} a_k/(z+k).$$

A variety of methods for computing the coefficients are discussed in a recent thesis by Pugh [16]. Their rate of decay depends strongly on a good choice for γ. However, it is unclear if it makes sense to ask about the asymptotic behavior for $N \to \infty$. Little is known about the decay of the error $|F_\gamma(z) - F_{N,\gamma}(z)|$ [16, Chap. 11]. Lanczos claimed that the higher γ becomes, the smaller is the value of the coefficients at which the convergence begins to slow down. At the same time, however, we have to wait longer before the asymptotic stage is reached. Pugh [16] calls this behavior the *Lanczos shelf* and is interested in finding good pairs of γ and N in order to guarantee a certain precision in the right half-plane. Godfrey [9] gives a 15-term expansion that provides an accuracy of about 15 significant digits along the real axis and about 13 digits in the rest of the complex plane. Because of the simple form of $F_{N,\gamma}(z)$, Lanczos's method is particularly suitable for matrix arguments.

7.3. Stirling's method. The asymptotic series that generalizes Stirling's formula[1] is still a standard and powerful method for evaluating the gamma function. There is a great deal of literature discussing efficient strategies and error estimates for these series (see the references in [14]). The goal here is to minimize the number of terms used to achieve the desired accuracy. This can be done in two ways, by either shifting the argument to the right or enforcing a faster asymptotic decay of the relative error using more terms in the series. (For fixed z and $N \to \infty$ the series does not converge.) The method is especially attractive for arguments with a large real part working in an arbitrary precision environment. Using an asymptotic series for $\log\Gamma(z)$, the error is bounded for $\operatorname{Re} z \ge 0$ by $|B_{2N}/(2N-1)|\,|z|^{1-2N}$, where B_{2N} denotes a Bernoulli number. This simple error estimate is due to Spira [17].

[1] Stirling was a student at the same Oxford college we both belong to, Balliol.

7.4. Software. Software libraries and programming environments for scientific computing all have routines to compute the gamma function, although quite a few do not deal with complex arguments. For our small survey we explored online documentation for various products, and yet it often remains unclear exactly which methods are used. For real arguments, a popular trick is to work with a rational Chebyshev approximation on the interval $[1, 2]$ and map this interval by the recurrence relation (1.2) to larger regions of the real line. The routine in the NAG library seems to map this interval to the whole real line, whereas MATLAB[2] uses a Stirling approximation for arguments larger than 12. On the fundamental interval, MATLAB uses a rational Chebyshev approximation of type $(8, 8)$. As the MATLAB routine was originally designed for Fortran, we imagine that many Fortran libraries use essentially the same method.

None of the above products provides a function for complex arguments. For Fortran the IMSL library has a routine of this kind. As there are no references to the work of Lanczos and Spouge in the IMSL documentation, we presume that it is based on asymptotic series. The Gnu Scientific Library provides a C function `gsl_sf_lngamma_complex_e` that evaluates $\log \Gamma(z)$ via the complex Lanczos method.

Mathematica uses the asymptotic Binet formula, which is another name for the Stirling series. We presume Maple uses the same method since the Maple documentation gives a reference to the classic book on special functions [7], which appeared before the methods of Lanczos and Spouge were introduced. Somewhat more interesting are the comments in [15]:

> There are a variety of methods in use for calculating the function $\Gamma(z)$ numerically, but none is quite as neat as the approximation derived by Lanczos. This scheme is entirely specific to the gamma function, seemingly plucked from thin air.

8. Conclusions. We have shown that $\Gamma(z)$ can be evaluated with geometric accuracy by two types of generic related methods:

- applying the trapezoidal rule on Talbot contours.
- using best rational approximations on the negative real axis.

Typically the second method is about twice as fast as the first. However, the first is simpler to implement as the construction of the best rational approximation is not trivial.

Amongst the Talbot contours, the cotangent contour has the best results. Using a shift from $(-\infty, 0]$ to $(-\infty, 1]$, one can improve the results for the best rational approximation a bit. For smaller values of z in the right half-plane, the approximations are excellent, and using the fundamental recurrence relation for the gamma function, one can extend the region of accuracy.

Even though the methods we have introduced are based on generic tools rather than on specific analysis of the gamma function, they are competitive with existing ones. The gamma function is just one of many special functions that have integral representations which can be evaluated efficiently by Talbot-type contours and rational approximations (see [8] for further examples). We believe that these methods can be useful in many areas of scientific computing.

Acknowledgments. We are grateful to André Weideman for discussions throughout this project, to Nico Temme for expert advice on the gamma function, and to

[2]In MATLAB 7.0 the command `type gamma` gives the source code of the corresponding `mex`-file. Previous versions do not offer this possibility.

Alphonse Magnus and Alexander Aptekarev for advice about Conjecture 5.1 and the challenges involved in proving it.

REFERENCES

[1] M.. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 1965.

[2] A. I. Aptekarev, *Sharp constants for rational approximations of analytic functions*, Mat. Sb., 193 (2002), pp. 1–72.

[3] C. M. Bender and S. A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, McGraw–Hill, New York, 1978.

[4] A. J. Carpenter, A. Ruttan, and R. S. Varga, *Extended numerical computations on the "1/9" conjecture in rational approximation theory*, in Rational Approximation and Interpolation, Lecture Notes in Math. 1105, P. R. Graves-Morris, E. B. Saff, and R. S. Varga, eds., Springer, Berlin, 1984, pp. 383–411.

[5] W. J. Cody, G. Meinardus, and R. S. Varga, *Chebyshev rational approximations to e^{-x} in $[0, +\infty)$ and applications to heat-conduction problems*, J. Approx. Theory, 2 (1969), pp. 50–65.

[6] P. J. Davis, *Leonhard Euler's integral: A historical profile of the gamma function*, Amer. Math. Monthly, 66 (1959), pp. 849–869.

[7] A. Erdélyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, *Higher Transcendental Functions Volumes* I and II, McGraw–Hill, New York, 1953.

[8] A. Gil, J. Segura, and N. M. Temme, *Computing special functions by using quadrature rules*, Numer. Algorithms, 33 (2003), pp. 265–275.

[9] P. Godfrey, *A Note on the Computation of the Convergent Lanczos Complex Gamma Approximation*, http://my.fit.edu/~gabdo/gamma.txt, 2001.

[10] A. A. Gonchar and E. A. Rakhmanov, *Equilibrium distributions and degree of rational approximation of analytic functions*, Sb. Math., 62 (1989), pp. 305–348.

[11] H. Hankel, *Die Euler'schen Integrale bei unbeschränkter Variabilität des Arguments*, Z. Math. Phys., 9 (1864), pp. 1–21.

[12] C. Lanczos, *A precision approximation of the gamma function*, J. Soc. Indust. Appl. Math. B, 1 (1964), pp. 86–96.

[13] A. P. Magnus, *Asymptotics and super asymptotics of best rational approximation error norms for the exponential function (the '1/9' problem) by the Carathéodory–Fejér method*, in Nonlinear Methods and Rational Approximation II, A. Cuyt et al., eds., Kluwer, Dordrecht, 1994, pp. 173–185.

[14] E. W. Ng, *A comparison of computational methods and algorithms for the complex gamma function*, ACM Trans. Math. Software, 1 (1975), pp. 56–70.

[15] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical recipes in C*, 2nd ed., Cambridge University Press, Cambridge, 1992.

[16] G. R. Pugh, *An Analysis of the Lanczos Gamma Approximation*, Ph.D. thesis, University of British Columbia, Vancouver, BC, Canada, 2004.

[17] R. Spira, *Calculation of the gamma function by Stirling's formula*, Math. Comp., 25 (1971), pp. 317–322.

[18] J. L. Spouge, *Computation of the gamma, digamma, and trigamma functions*, SIAM J. Numer. Anal., 31 (1994), pp. 931–944.

[19] A. Talbot, *The accurate numerical inversion of Laplace transforms*, J. Inst. Math. Appl., 23 (1979), pp. 97–120.

[20] N. M. Temme, *Special Functions: An Introduction to the Classical Functions of Mathematical Physics*, Wiley, New York, 1996.

[21] L. N. Trefethen and J. A. C. Weideman, *The Fast Trapezoid Rule in Scientific Computing*, manuscript, 2006.

[22] L. N. Trefethen, J. A. C. Weideman, and T. Schmelzer, *Talbot quadratures and rational approximations*, BIT, 46 (2006), pp. 653–670.

[23] L. N. Trefethen, *Spectral Methods in MATLAB*, SIAM, Philadelphia, 2000.

[24] J. A. C. Weideman and L. N. Trefethen, *Parabolic and hyperbolic contours for computing the Bromwich integral*, Math. Comp., to appear.

[25] J. A. C. Weideman, *Optimizing Talbot's contours for the inversion of the Laplace transform*, SIAM J. Numer. Anal., 44 (2006), pp. 2342–2362.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 572–591

NUMERICAL APPROXIMATION OF A TIME DEPENDENT, NONLINEAR, SPACE-FRACTIONAL DIFFUSION EQUATION*

VINCENT J. ERVIN†, NORBERT HEUER‡, AND JOHN PAUL ROOP§

Abstract. In this article we analyze a fully discrete numerical approximation to a time dependent fractional order diffusion equation which contains a nonlocal quadratic nonlinearity. The analysis is performed for a general fractional order diffusion operator. The nonlinear term studied is a product of the unknown function and a convolution operator of order 0. Convergence of the approximation and a priori error estimates are given. Numerical computations are included, which confirm the theoretical predictions.

Key words. anomalous diffusion, nonlinear parabolic equation, finite element approximation

AMS subject classification. 65N30

DOI. 10.1137/050642757

1. Introduction.

In this paper we study the numerical approximation to time dependent fractional order diffusion equations containing a nonlocal quadratic nonlinearity. Specifically, we consider equations of the form

$$u_t + \mathcal{D}^{2\alpha} u - \nabla \cdot (u\, B(u)) = f(x), \quad x \in \Omega\ ,\ t \in (0,T], \tag{1.1}$$

$$u(x,t) = 0, \quad x \in \partial\Omega\ ,\ t \in (0,T], \tag{1.2}$$

$$u(x,0) = u^0(x), \quad x \in \Omega, \tag{1.3}$$

which arise from models in statistical mechanics. In such a setting, u can be thought of as describing the density of particles filling up a domain $\Omega \subset \mathbb{R}^d$. In (1.1), $\mathcal{D}^{2\alpha}$ denotes a general fractional order diffusion operator of order 2α, $1/2 < \alpha \leq 1$. The term $\nabla \cdot (u\, B(u))$ models particle interactions.

For the classical diffusion case ($\alpha = 1$) the diffusion operator models a Brownian diffusion process. For fractional diffusion ($1/2 < \alpha < 1$) the $\mathcal{D}^{2\alpha}$ operator is commonly referred to as *anomalous diffusion*, where the underlying stochastic process is a Lévy α-stable flight. A key difference between fractional diffusion operators and the usual diffusion operator is that fractional diffusion operators are *nonlocal* operators. Equations containing fractional diffusion have also been investigated in modeling turbulent flow [6, 18], chaotic dynamics of classical conservative systems [19], and contaminant transport in groundwater flow [2, 12].

*Received by the editors November 4, 2005; accepted for publication (in revised form) September 8, 2006; published electronically April 3, 2007. Supported by FONDAP Program in Applied Mathematics and Fondecyt project 1040615 (Chile).

http://www.siam.org/journals/sinum/45-2/64275.html

†Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975 (vjervin@clemson.edu). This research was undertaken while this author was a visitor in the Departamento de Ingeniería Matemática, Universidad de Concepcíon. Research partially supported by the National Science Foundation under award DMS-0410792.

‡BICOM, Department of Mathematical Sciences, Brunel University, Uxbridge, Middlesex UB8 3PH, UK (norbert.heuer@brunel.ac.uk). This research was undertaken at the Departamento de Ingeniería Matemática, Universidad de Concepcíon.

§Department of Mathematics, North Carolina A & T University, Greensboro, NC 27411 (jproop@ncat.edu). This research was undertaken while this author was a post-doc in the Department of Mathematics at Virginia Tech. Research partially supported by the Air Force Office of Scientific Research under award F49620-1-00-0299.

There are a number of definitions for fractional derivatives in $\mathbb{R}^1$, for example the Riemann–Liouville, Grüwald–Letnikov, and Caputo fractional derivatives [13], the Riemann–Liouville being the most commonly used. In higher dimensions there are also several definitions for the fractional diffusion operator. In this article (see section 2.2) we describe two: the fractional Laplacian operator, and the weighted directional, fractional diffusion operator. For problems posed on bounded domains, a disadvantage of the fractional Laplacian operator is that it is formally defined in terms of the Fourier transform. From practical considerations we have focused our attention on the weighted directional, fractional diffusion operator.

From the fact that $\mathcal{D}^{2\alpha}$ is a strongly elliptic operator, the existence and uniqueness for the space-fractional diffusion equation parallels that of the usual diffusion equation, the difference being in the function spaces, i.e., $H_0^\alpha(\Omega)$ instead of $H_0^1(\Omega)$. (See [7, 8, 14].) The existence of solutions to (1.1)–(1.3) has been studied by Biler and Woyczyński in [4]. For $B(\cdot)$ given by (2.8), (2.9) they showed the existence of a local in time weak solution. In this paper we do not investigate the existence of u satisfying (1.1)–(1.3) but, assuming a sufficiently regular solution u exists, the existence and convergence properties of its approximation u_h. The results presented in this paper extend the work developed in [7], [8] (see also [15]) for a steady-state linear fractional advection dispersion equation.

A finite element approximation scheme is described and shown to be computable in section 3. A priori error estimates for the approximation are presented in section 4. Hölder-type inequalities for Sobolev spaces, used in the analysis, are derived in section 2.4. Our analysis does not rely on the particular form of the fractional diffusion operator, requiring only that it satisfy properties of *continuity* and *coercivity* (see (2.2), (2.3)). Several examples of nonlocal operators $B(\cdot)$ are given in section 2.3. Again, our analysis does not assume a particular form for $B(\cdot)$, only that it is linear and an *operator of order* 0 (see (2.4)). Finally, in section 5 we present some numerical experiments which support the theoretical estimates.

2. Mathematical preliminaries.

2.1. Mathematical notation. In this section we summarize the mathematical notation used and state our assumptions regarding properties satisfied by the fractional diffusion operator $\mathcal{D}^{2\alpha}$ and the operator $B(\cdot)$.

The following notation is used. The $L^2(\Omega)$ inner product is denoted by $(\cdot,\cdot)$, and the $L^p(\Omega)$ norm by $\|\cdot\|_{L^p}$, with the special cases of $L^2(\Omega)$ and $L^\infty(\Omega)$ norms being written as $\|\cdot\|$ and $\|\cdot\|_\infty$, respectively. For $k \in \mathbb{N}$, we denote the norm associated with the Sobolev space $W^{k,p}(\Omega)$ by $\|\cdot\|_{W^{k,p}}$, with the special case $W^{k,2}(\Omega)$ being written as $H^k(\Omega)$ with norm $\|\cdot\|_k$ and seminorm $|\cdot|_k$. For the definition of fractional order Sobolev spaces $W^{s,p}(\Omega)$, $s \in \mathbb{R}^+\backslash\mathbb{N}$, we use the real method of interpolation between two Banach spaces [3, 16].

When $v(\mathbf{x},t)$ is defined on the entire time interval $(0,T)$, we define

$$\|v\|_{\infty,k} := \sup_{0<t<T} \|v(\cdot,t)\|_k\,, \quad \|v\|_{0,k} := \left(\int_0^T \|v(\cdot,t)\|_k^2\,dt\right)^{1/2},$$
$$\|v\|(t) := \|v(\cdot,t)\|.$$

For convenience we let X denote the space

$$X := H_0^\alpha(\Omega) := \text{ closure of } C_0^\infty(\Omega) \text{ in } H^\alpha(\Omega). \tag{2.1}$$

We use $H^{-\alpha}(\Omega)$ to denote the dual space of $H_0^\alpha(\Omega)$, with norm denoted $\|\cdot\|_{-\alpha}$.

Throughout the paper we use C to denote a *generic constant* whose actual value may change from line to line.

We make the following general assumptions regarding the diffusion operator. There exist constants $C_c, C_t > 0$ such that for $v, w \in X$

$$\langle \mathcal{D}^{2\alpha} v, w\rangle \le C_t \|v\|_\alpha \|w\|_\alpha \quad \text{(continuity on } X \times X\text{)}, \tag{2.2}$$

$$\langle \mathcal{D}^{2\alpha} v, v\rangle \ge C_c \|v\|_\alpha^2 \quad \text{(coercivity on } X\text{)}, \tag{2.3}$$

where $\langle\cdot,\cdot\rangle$ denotes the duality pairing of $H^{-\alpha}(\Omega)$ and $H_0^\alpha(\Omega)$.

For the nonlocal operator $B(\cdot)$ we assume the following:

(i) $B(\cdot)$ is linear,

(ii) $B(\cdot)$ is an *operator of order 0*; i.e., for $\beta \ge 0$, $u \in H^\beta(\Omega)$,

$$\|B(u)\|_\beta \le C_B \|u\|_\beta. \tag{2.4}$$

2.2. Examples of fractional diffusion operators satisfying (2.2), (2.3).

1. *Fractional Laplacian operator.* In the context of pseudo differential operators [17], a fractional diffusion operator may be defined in terms of the negative Laplacian operator, $-\Delta$ [4].

We have that for ω the Fourier transform variable,

$$\mathcal{F}(-\Delta u(x)) = |\omega|^2 \hat{u}(\omega).$$

The fractional Laplacian operator is then defined via the inverse Fourier transform as

$$(-\Delta)^{\gamma/2} u(x) := \mathcal{F}^{-1}\left(|\omega|^\gamma \hat{u}(\omega)\right). \tag{2.5}$$

Associated with (2.5), a fractional differential operator of order 2α may be formally defined as

$$\mathcal{D}^{2\alpha} u(x) := (-\Delta)^\alpha u(x). \tag{2.6}$$

For $\mathcal{D}^{2\alpha}$ defined by (2.6), we have for $v, w \in H_0^\alpha(\Omega)$, $\alpha > 1/2$,

$$\begin{aligned}\langle \mathcal{D}^{2\alpha} v, w\rangle &= \left((-\Delta)^{\alpha/2} v,\, (-\Delta)^{\alpha/2} w\right)\\ &\le C_1 \|v\|_\alpha \|w\|_\alpha.\end{aligned}$$

Also,

$$\begin{aligned}\langle \mathcal{D}^{2\alpha} v, v\rangle &= \left((-\Delta)^{\alpha/2} v,\, (-\Delta)^{\alpha/2} v\right)\\ &\ge C_2 \|v\|_\alpha^2.\end{aligned}$$

2. *Weighted directional, fractional diffusion operator.* In [8, 14] the following fractional diffusion operator was introduced and analyzed:

$$D_M^{2\alpha} u(x) := -\int_{\|\nu\|=1} D_\nu^{2\alpha} u(x)\, M(d\nu), \tag{2.7}$$

where $M(d\nu)$ denotes a general probability measure on the unit sphere in $\mathbb{R}^{\acute{d}}$; for $\sigma > 0$,

$$D_\nu^{-\sigma} v(x) := \frac{1}{\Gamma(\sigma)} \int_0^\infty \xi^{\sigma-1} v(x - \xi\nu)\, d\xi \ ,$$

and for $n-1<\beta\le n$, $\sigma=n-\beta$,

$$D_\nu^\beta v(x) \;:=\; (\nu\cdot\nabla)^n\, D_\nu^{-\sigma} v(x) \quad .$$

Properties (2.2), (2.3) were established in [8, 14] for $D_M^{2\alpha}$.

2.3. Examples of operators B(·) satisfying (2.4). Typically $B(\cdot)$ is of the form

$$B(u)(x) \;=\; \int b(x,y)\, u(y)\, dy \;. \tag{2.8}$$

For the ordinary diffusion equation the following operators have been considered. The choice

$$b(x,y) \;=\; c\,(x-y)\,|x-y|^{-\acute{d}} \tag{2.9}$$

has been used in a model for Brownian diffusion of charge carriers interacting via Coulomb forces. For $c>0$, (2.8) has been used to model the mutual gravitational attraction of particles in a cloud [4]. For $\acute{d}=2$ and

$$b(x,y) \;=\; (x_2-y_2,\,-(x_1-y_1))\,|x-y|^{-2} \tag{2.10}$$

the ordinary diffusion equation becomes the vorticity equation for the Navier–Stokes equations.

A general *potential kernel* for $B(\cdot)$ has the form

$$b(x,y) \;=\; c\,(x-y)\,|x-y|^{-\acute{d}+\beta-1} \qquad \text{for } 0<\beta\le \acute{d}-1. \tag{2.11}$$

To determine the order of the operators $B(\cdot)$ defined above we have the following from [9]:

1. For $P(x_1,\dots,x_{\acute{d}})$ a polynomial in $\acute{d}$ variables,

$$\mathcal{F}(P(x_1,\dots,x_{\acute{d}}))(\boldsymbol{\omega}) \;=\; (2\pi)^{\acute{d}}\, P\left(\frac{-i\partial}{\partial\omega_1},\dots,\frac{-i\partial}{\partial\omega_{\acute{d}}}\right)\delta(\boldsymbol{\omega}) \;. \tag{2.12}$$

2. For $r=|x|$, $\rho=|\boldsymbol{\omega}|$, $m\in\{0,1,2,\dots\}$, c_{-1},c_0 constants dependent on $\acute{d}+2m$,

$$\mathcal{F}(r^{-\lambda}) = 2^{\acute{d}-\lambda}\pi^{\acute{d}/2}\,\frac{\Gamma((\acute{d}-\lambda)/2)\,\rho^{\lambda-\acute{d}}}{\Gamma(\lambda/2)}, \qquad \lambda\ne\acute{d}+2m, \tag{2.13}$$

$$\mathcal{F}(r^{-\acute{d}-2m}) = \frac{1}{2}\,\Gamma\left(\frac{\acute{d}}{2}\right)\,\pi^{-\acute{d}/2}\,\left(c_{-1}\rho^{2m}\ln\rho \;+\; c_0\rho^{2m}\right). \tag{2.14}$$

For the kth component of $x/|x|^\lambda = x_k/|x|^\lambda$, $\lambda\ne\acute{d}+2m$, combining (2.12) and (2.13), and using the Fourier transform property of the convolution operator $\star$,

$$\begin{aligned}
\mathcal{F}\left(\frac{x_k}{|x|^\lambda}\right) &= (2\pi)^{\acute{d}}\,\left(\frac{-i\,\partial}{\partial\omega_k}\right)\delta(\boldsymbol{\omega}) \;\star\; 2^{\acute{d}-\lambda}\pi^{\acute{d}/2}\,\frac{\Gamma((\acute{d}-\lambda)/2)\,\rho^{\lambda-\acute{d}}}{\Gamma(\lambda/2)} \\
&= C\int\left(\frac{-i\,\partial}{\partial\sigma_k}\right)\delta(\boldsymbol{\sigma})\,|\boldsymbol{\omega}-\boldsymbol{\sigma}|^{\lambda-\acute{d}}\,d\boldsymbol{\sigma} \\
&= C\,i\int\delta(\boldsymbol{\sigma})\,(\sigma_k-\omega_k)|\boldsymbol{\omega}-\boldsymbol{\sigma}|^{\lambda-\acute{d}-2}\,d\boldsymbol{\sigma} \\
&= C\,i\,\omega_k\,|\boldsymbol{\omega}|^{\lambda-\acute{d}-2} \;. \qquad (2.15)
\end{aligned}$$

The zero extension of $f \in H_0^\gamma(\Omega)$, $\tilde{f}$, satisfies $\tilde{f} \in H^\gamma(\mathbb{R}^{\acute{d}})$. Thus, $f \in H_0^\gamma(\Omega)$ implies $|\boldsymbol{\omega}|^j \mathcal{F}(\tilde{f}) \in L^2(\mathbb{R}^{\acute{d}})$ for $0 \le j \le \gamma$.

For the kth component of $B(u)(x)$ defined by (2.8), (2.11) with $\beta \neq 1$, we have, as $B(\cdot)$ is a convolution operator,

$$\begin{aligned}\int_{\mathbb{R}^{\acute{d}}} |\boldsymbol{\omega}|^{2j} \left|\mathcal{F}(B(u)_k)\right|^2 d\boldsymbol{\omega} &= \int_{\mathbb{R}^{\acute{d}}} |\boldsymbol{\omega}|^{2j} \left|C\,\omega_k\,|\boldsymbol{\omega}|^{-\beta-1}\hat{u}(\boldsymbol{\omega})\right|^2 d\boldsymbol{\omega} \\ &\le C \int_{\mathbb{R}^{\acute{d}}} |\boldsymbol{\omega}|^{2(j-\beta)} \left|\hat{u}\right|^2 (\boldsymbol{\omega})\, d\boldsymbol{\omega}\ .\end{aligned}$$

Thus, if $u \in H_0^\gamma(\Omega)$, then $(B(u))_k \in H^{\beta+\gamma}(\Omega)$ for $k = 1, \dots \acute{d}$. Hence $B(u) \in H^{\beta+\gamma}(\Omega)$. Consequently, $B(\cdot)$ is an operator of order $-\beta$. (Also, then an operator of order 0.) For $\beta = 1$, using (2.14), $B(\cdot)$ is an operator of order -1.

2.4. Hölder-type inequalities for Sobolev spaces. In this section we present a number of estimates which are useful in handling the nonlinear term in the error analysis.

LEMMA 1. *Let $\Omega \subset \mathbb{R}^{\acute{d}}$ be bounded, $\partial\Omega \in C^1$. Then for u and v such that the given norms are finite we have*

$$\|uv\| \le C \begin{cases} \|u\|_s \, \|v\|_{\acute{d}/2-s}, & 0 < s < \frac{\acute{d}}{2}, \\ \|u\|_\infty \, \|v\| & \\ \|u\|_s \, \|v\|, & s > \acute{d}/2. \end{cases} \tag{2.16}$$

Proof. For $z \in W^{j,p}(\Omega) \cap W^{m,r}(\Omega)$ we have the following embedding properties for Sobolev spaces [1, p. 218]. For $1 < r \le p < \infty$,

$$\|z\|_{W^{j,p}} \ \le \ C\,\|z\|_{W^{m,r}}, \tag{2.17}$$

$$\text{where } \frac{1}{p} = \frac{j}{\acute{d}} + \frac{1}{r} - \frac{m}{\acute{d}} \quad \text{and} \quad \begin{cases} j \ge 0, \text{ if } r < p, & \text{or} \\ j > 0, \ \ j \text{ not an integer}, & \text{or} \\ j \ge 0, \ \ 1 < r \le 2. \end{cases}$$

Note that the above inequality (2.17) holds for $m \in \mathbb{R}$, $m > 0$. Using Hölder's inequality, with $p, \tilde{p} > 1$, satisfying $1/p + 1/\tilde{p} = 1$, and the embedding theorem

$$\begin{aligned}\|uv\| &\le \|u\|_{L^{2p}} \, \|v\|_{L^{2\tilde{p}}} \\ &= \|u\|_{W^{0,2p}} \, \|v\|_{W^{0,2\tilde{p}}} \\ &\le C\|u\|_{W^{\acute{d}(p-1)/(2p),2}} \, \|v\|_{W^{\acute{d}/(2p),2}} \\ &= C\|u\|_{\acute{d}(p-1)/(2p)} \, \|v\|_{\acute{d}/(2p)}. \end{aligned} \tag{2.18}$$

The first inequality in (2.16) follows from (2.18) with the choice $s = \acute{d}(p-1)/(2p)$. The second and third inequalities are straightforward to establish. ∎

Remark. The boundary regularity assumption on Ω in Lemma 1, $\partial\Omega \in C^1$, can be relaxed. The Sobolev embedding theorems on bounded domains require sufficient regularity of the domain to enable functions defined in Ω to be appropriately extended to $\mathbb{R}^{\acute{d}}$. In particular, for the analysis presented in sections 3 and 4 it suffices for Ω to be a Lipschitz domain.

The following results are Hölder-type inequalities for Sobolev spaces.

THEOREM 1. *Let $\Omega \subset \mathbb{R}^{\acute{d}}$ be bounded, $\partial\Omega \in C^1$. Then for $0 \le \alpha, \beta \le 1$, $\tilde{\epsilon} > 0$, $p > 1$, $0 < s \le 1/2$, u, and v such that the given norms are finite we have*

(2.19)

$$\|uv\|_\alpha \le C \begin{cases} \|u\|_1 \, \|v\|_{\alpha+\tilde{\epsilon}}, & \acute{d} = 2, \\ \|u\|_{3/2\,-\,s} \, \|v\|_{\alpha+s+\tilde{\epsilon}}, & \acute{d} = 3, \quad 0 < s \le \frac{1}{2}, \end{cases}$$

(2.20)

$$\|uv\|_{\alpha\beta} \le C\|u\|_{\beta\,+\,\acute{d}(p-1)(1-\beta)/2p} \, \|v\|_{\acute{d}(1-\alpha+\alpha p)/(2p)\,+\,\tilde{\epsilon}} \quad \textit{for} \begin{cases} \acute{d} = 2, & 1 < p \\ \acute{d} = 3, & 1 < p \le 3. \end{cases}$$

Proof. We have that

$$\|uv\|_1 \le \|uv\| \, + \, |uv|_1 \tag{2.21}$$

and

$$|uv|_1 \le \|u \, \nabla v\| \, + \, \|\nabla u \, v\|.$$

Proceeding as in the proof of Lemma 1, with $q, \tilde{q} > 1$, $1/q + 1/\tilde{q} = 1$,

$$\begin{aligned} \|u \, \nabla v\| \le \|u\|_{L^{2q}} \, \|\nabla v\|_{L^{2\tilde{q}}} \; &\le \; \|u\|_{W^{0,2q}} \, \|v\|_{W^{1,2\tilde{q}}} \\ &\le C\|u\|_{\acute{d}(q-1)/(2q)} \, \|v\|_{(\acute{d}+2q)/(2q)}. \end{aligned} \tag{2.22}$$

Also, for $\epsilon > 0$

$$\|u \, \nabla v\| \; \le \; C\|u\|_{\acute{d}/2\,+\,\epsilon} \, \|v\|_1. \tag{2.23}$$

Similarly, with $r > 1$

$$\|\nabla u \, v\| \; \le \; C\|u\|_{(\acute{d}+2r)/(2r)} \, \|v\|_{\acute{d}(r-1)/(2r)} \quad \text{and} \quad \|\nabla u \, v\| \; \le \; C\|u\|_1 \, \|v\|_{\acute{d}/2\,+\,\epsilon}. \tag{2.24}$$

Combining (2.21), (2.22), (2.24), (2.18), for $s > 1$, we have

$$\begin{aligned} \|uv\|_1 \; \le \; C \Big(&\|u\|_{\acute{d}(s-1)/(2s)} \, \|v\|_{\acute{d}/(2s)} + \|u\|_{\acute{d}(q-1)/(2q)} \, \|v\|_{(\acute{d}+2q)/(2q)} \\ &+ \|u\|_{(\acute{d}+2r)/(2r)} \, \|v\|_{\acute{d}(r-1)/(2r)} \Big). \end{aligned} \tag{2.25}$$

From (2.22), (2.24)(b), and (2.25) it follows that

$$\|uv\|_1 \; \le \; C \, \|u\|_1 \, \|v\|_{\acute{d}(1+\epsilon)/2}, \quad \acute{d} = 2, 3. \tag{2.26}$$

Also, equating the norms for u in the last two terms of (2.25), we have, for s appropriately chosen,

$$\|uv\|_1 \; \le \; C \, \|u\|_{\acute{d}(q-1)/(2q)} \, \|v\|_{(\acute{d}+2q)/(2q)} \quad \text{for } q > 3, \; \acute{d} = 3. \tag{2.27}$$

Next we interpolate between spaces to obtain the stated estimates.

For u fixed, let operators T_0, T_1 be dependent on v defined by

$$T_0(v) \; = \; T_1(v) \; = \; uv.$$

Using (2.26), we consider T_1 as a bounded linear operator between $H^{\acute{d}(1+\epsilon)/2}$ and H^1, with norm $\leq C\|u\|_1$. Also, using (2.18), we consider T_0 as a bounded linear operator between $H^{\acute{d}/(2p)}$ and L^2, with norm $\leq C\|u\|_{\acute{d}(p-1)/(2p)}$. By interpolation [3, 16] we obtain

$$\begin{aligned}\|uv\|_\alpha &= \|uv\|_{[L^2,H^1]_{\alpha,2}}\\ &\leq \|T_1\|^\alpha \|T_0\|^{1-\alpha}\|v\|_{[H^{\acute{d}/(2p)},H^{\acute{d}(1+\epsilon)/2}]_{\alpha,2}}\\ &\leq C\|u\|_1^\alpha \,\|u\|^{1-\alpha}_{\acute{d}(p-1)/(2p)}\,\|v\|_{\acute{d}(1-\alpha+\alpha p)/(2p)+\tilde{\epsilon}}\\ &\leq C\|u\|_1\,\|v\|_{\acute{d}(1-\alpha+\alpha p)/(2p)+\tilde{\epsilon}} \qquad (2.28)\end{aligned}$$

for $1 < p \leq 3$ in $\mathbb{R}^3$ (i.e., $\acute{d} = 3$) and no restriction in $\mathbb{R}^2$ (i.e., $\acute{d} = 2$). Note that $(\acute{d}(1-\alpha+\alpha p)/(2p))$ is a decreasing function of p. Minimizing with respect to p, we obtain (2.19)(a) and (2.19)(b) for $s = 1/2$.

Next we interpolate with v held fixed. Let $S_1 : H^1 \to H^\alpha$ be the operator defined by $S_1(u) = u\,v$. Using (2.28), we have that S_1 is a bounded linear operator with norm $\leq C\|v\|_{\acute{d}(1-\alpha+\alpha p)/(2p)+\tilde{\epsilon}}$. From (2.18), for $S_0 : H^{\acute{d}(p-1)/(2p)} \to L^2$ given by $S_0(u) = u\,v$, S_0 is a bounded linear operator with norm $\leq C\|v\|_{\acute{d}/(2p)}$. By interpolation we obtain

$$\begin{aligned}\|uv\|_{\alpha\beta} &= \|uv\|_{[L^2,H^\alpha]_{\beta,2}}\\ &\leq \|S_1\|^\beta\|S_0\|^{1-\beta}\|u\|_{[H^{\acute{d}(p-1)/(2p)},H^1]_{\beta,2}}\\ &\leq C\|v\|^\beta_{\acute{d}(1-\alpha+\alpha p)/(2p)+\tilde{\epsilon}}\,\|v\|^{1-\beta}_{\acute{d}/(2p)}\,\|u\|_{\beta+\acute{d}(p-1)(1-\beta)/(2p)}\\ &\leq C\|u\|_{\beta+\acute{d}(p-1)(1-\beta)/(2p)}\,\|v\|_{\acute{d}(1-\alpha+\alpha p)/(2p)+\tilde{\epsilon}}. \qquad (2.29)\end{aligned}$$

For the case (2.19)(b), in view of (2.18), for the choice $q = p$, for $p > 3$, from (2.27),

$$\|uv\|_1 \;\leq\; C\|u\|_{\acute{d}(p-1)/(2p)}\,\|v\|_{\acute{d}/(2p)+1}. \qquad (2.30)$$

Interpolating, as above, with u fixed, we obtain

$$\begin{aligned}\|uv\|_\alpha &\leq C\|u\|_{\acute{d}(p-1)/(2p)}\,\|v\|_{\alpha(\acute{d}/(2p)+1)+(1-\alpha)(\acute{d}/2p)}\\ &= C\|u\|_{\acute{d}(p-1)/(2p)}\,\|v\|_{\acute{d}/(2p)+\alpha}. \qquad (2.31)\end{aligned}$$

Letting $s = \acute{d}/(2p)$ in (2.31), the stated result follows. □

Also used below is the following lemma.

LEMMA 2. *For $\Omega \subset \mathbb{R}^{\acute{d}}$, $\alpha > \acute{d}/4$, $v, w \in X$, $\epsilon > 0$, there exists $C > 0$ such that*

$$(v\,B(w),\,\nabla v) \leq C\frac{(q\,\epsilon)^{-p/q}}{p}\|\nabla\cdot B(w)\|^p\|v\|^2 \,+\, \epsilon\,\|v\|^2_\alpha, \qquad (2.32)$$

where $p = 4\alpha/(4\alpha - \acute{d})$, $q = 4\alpha/\acute{d}$.

Proof. We begin by rewriting the inner product as

$$\begin{aligned}(v\,B(w),\,\nabla v) &= \frac{1}{2}\left(B(w),\,\nabla v^2\right) \;=\; -\frac{1}{2}\left(\nabla\cdot B(w),\,v^2\right)\\ &\leq \frac{1}{2}\,\|v\|^2_{L^4}\,\|\nabla\cdot B(w)\|.\end{aligned}$$

For $\Omega \subset \mathbb{R}^{\acute{d}}$, $H^{\acute{d}/4}(\Omega)$ is continuously imbedded in $L^4(\Omega)$ (cf. (2.17)), and as an interpolation space

$$H^{\acute{d}/4}(\Omega) \;=\; \left[L^2,\, H^{\alpha}\right]_{\frac{\acute{d}}{4\alpha},\,2}\;.$$

Hence,

$$\begin{aligned}\|v\|_{L^4}^2 \le C\,\|v\|_{H^{\acute{d}/4}}^2 \;&\le\; C\left(\|v\|^{1-\acute{d}/4\alpha}\,\|v\|_{\alpha}^{\acute{d}/4\alpha}\right)^2\\ &\le C\|v\|^{2-\acute{d}/2\alpha}\,\|v\|_{\alpha}^{\acute{d}/2\alpha}.\end{aligned}$$

Thus,

$$\|v\|_{L^4}^2\,\|\nabla\cdot B(w)\| \le C\,\|v\|^{2-\acute{d}/2\alpha}\,\|\nabla\cdot B(w)\|\,\|v\|_{\alpha}^{\acute{d}/2\alpha}. \tag{2.33}$$

Applying Young's inequality, $ab \le |a|^p/p + |b|^q/q$ for $1/p + 1/q = 1$, with the choice $p = 4\alpha/(4\alpha - \acute{d})$, $q = 4\alpha/\acute{d}$, the stated results (2.32) follow. ∎

3. Finite element approximation. In this section we formulate a fully discrete finite element method for (1.1)–(1.3). We begin by describing the finite element approximation framework and listing the approximating properties and inverse estimates used in the analysis.

Let $\Omega \subset \mathbb{R}^{\acute{d}}$ be a polygonal domain, and let T_h be a triangulation of Ω made of triangles (in $\mathbb{R}^2$) or tetrahedrons (in $\mathbb{R}^3$). Thus, the computational domain is defined by

$$\Omega = \bigcup K; \qquad K \in T_h.$$

We assume that there exist constants c_1, c_2 such that

$$c_1 h \le h_K \le c_2 \rho_K,$$

where h_K is the diameter of triangle (tetrahedron) K, ρ_K is the diameter of the greatest ball (sphere) included in K, and $h = \max_{K\in T_h} h_K$. For $k \in \mathbb{N}$, let $P_k(A)$ denote the space of polynomials on A of degree no greater than k. Then we define the finite element space X_h as follows:

$$X_h := \left\{v \in X \cap C(\bar{\Omega}) : v|_K \in P_k(K)\;\; \forall K \in T_h\right\}. \tag{3.1}$$

We summarize several properties of finite element spaces and Sobolev spaces which we will use in our subsequent analysis. For $w \in H^{k+1}(\Omega)$ we have (see [10]) that there exists $\mathcal{W} \in X_h$ such that

$$\|w - \mathcal{W}\| \;+\; h\|\nabla(w - \mathcal{W})\| \;\le\; C_I h^{k+1}\|w\|_{k+1}. \tag{3.2}$$

LEMMA 3 (see [5]). *Let $\{T_h\}$, $0 < h \le 1$, denote a quasi-uniform family of subdivisions of a polyhedral domain $\Omega \subset \mathbb{R}^{\acute{d}}$. Let $(\hat{K}, P, N)$ be a reference finite element such that $P \subset W^{l,p}(\hat{K}) \cap W^{m,q}(\hat{K})$ is a finite-dimensional space of functions on $\hat{K}$, N is a basis for P', where $1 \le p \le \infty$, $1 \le q \le \infty$, and $0 \le m \le l$. For $K \in T_h$, let (K, P_K, N_K) be the affine equivalent element, and $V_h = \{v : v$ is measurable and $v|_K \in P_K\ \forall K \in T_h\}$. Then there exists $C = C(l, p, q)$ such that*

$$\left[\sum_{K\in T_h} \|v\|_{W^{l,p}(K)}^p\right]^{1/p} \;\le\; C\,h^{m-l+\min(0,\,\frac{\acute{d}}{p}-\frac{\acute{d}}{q})}\left[\sum_{K\in T_h} \|v\|_{W^{m,q}(K)}^q\right]^{1/q} \tag{3.3}$$

for all $v \in V_h$. ∎

Let Δt denote the step size for t so that $t_n = n\Delta t$, $n = 0, 1, 2, \ldots, N$. For notational convenience, we denote $v^n := v(\cdot, t_n)$ and

$$d_t f^n := \frac{f(t_n) - f(t_{n-1})}{\Delta t}. \tag{3.4}$$

The following norms are also used in the analysis:

$$|||v|||_{\infty,k} := \max_{1 \le n \le N} \|v^n\|_k,$$

$$|||v|||_{0,k} := \left[\sum_{n=1}^{N} \|v^n\|_k^2 \Delta t \right]^{1/2}.$$

Approximating system. *For* $n = 1, 2, \ldots, N$, *find* $u_h^n \in X_h$ *such that*

$$(d_t u_h^n, v) + \langle \mathcal{D}^{2\alpha} u_h^n, v \rangle + (u_h^n B(u_h^{n-1}), \nabla v) = (f^n, v) \quad \forall\, v \in X_h. \tag{3.5}$$

For notational convenience we define $A(w; u, v)$ as

$$A(w; u, v) := \langle \mathcal{D}^{2\alpha} u, v \rangle + (uB(w), \nabla v). \tag{3.6}$$

Then, the linear system of equations (3.5) can be written equivalently as

$$(d_t u_h^n, v) + A(u_h^{n-1}; u_h^n, v) = (f^n, v) \quad \forall\, v \in X_h. \tag{3.7}$$

To ensure computability of the algorithm, we begin by showing that (3.5) is uniquely solvable for u_h^n at each time step n. We use the following induction hypothesis, which simply states that the computed iterates u_h^n are bounded independent of h and n:

$$\text{(IH1)} \qquad \|u_h^j\|_1 \le \mathcal{K}, \quad j = 0, \ldots, n-1. \tag{3.8}$$

Lemma 4. *Assume that* (IH1) *holds; i.e.,* $\|u_h^j\|_1 \le \mathcal{K}$ *for* $j = 0, 1, \ldots, n-1$. *For a sufficiently small step size* Δt, *there exists a unique solution* $u_h^n \in X_h$ *satisfying* (3.5).

Proof. As (3.5) represents a finite system of linear equations, the positivity of $(u_h^n, u_h^n)/\Delta t + A(u_h^{n-1}; u_h^n, u_h^n)$ is a sufficient condition for the existence and uniqueness of u_h^n.

We have, using (2.3) and (2.32),

$$\begin{aligned}
&\frac{(u_h^n, u_h^n)}{\Delta t} + A(u_h^{n-1}; u_h^n, u_h^n) \\
&\quad = \frac{1}{\Delta t}\|u_h^n\|^2 + \langle \mathcal{D}^{2\alpha} u_h^n, u_h^n \rangle + (u_h^n B(u_h^{n-1}), \nabla u_h^n) \\
&\quad \ge \frac{1}{\Delta t}\|u_h^n\|^2 + C_c \|u_h^n\|_\alpha^2 - C_1 \epsilon_2^{-C_2} \|\nabla \cdot B(u_h^{n-1})\|^{C_3} \|u_h^n\|^2 - \epsilon_2 \|u_h^n\|_\alpha^2 \\
&\quad = \left(\frac{1}{\Delta t} - C_1 \epsilon_2^{-C_2} \|\nabla \cdot B(u_h^{n-1})\|^{C_3} \right) \|u_h^n\|^2 + (C_c - \epsilon_2) \|u_h^n\|_\alpha^2 \qquad (3.9) \\
&\quad \ge \left(\frac{1}{\Delta t} - \tilde{C}_1 \epsilon_2^{-C_2} C_B^{C_3} \|u_h^{n-1}\|_1^{C_3} \right) \|u_h^n\|^2 + (C_c - \epsilon_2) \|u_h^n\|_\alpha^2 \qquad (3.10) \\
&\quad \ge \left(\frac{1}{\Delta t} - \tilde{C}_1 \epsilon_2^{-C_2} C_B^{C_3} \mathcal{K}^{C_3} \right) \|u_h^n\|^2 + (C_c - \epsilon_2) \|u_h^n\|_\alpha^2. \qquad (3.11)
\end{aligned}$$

Hence, for Δt chosen sufficiently small we have that (3.5) is uniquely solvable for u_h^n. ∎

The discrete Gronwall's lemma plays an important role in the following analysis.

LEMMA 5 (discrete Gronwall's lemma [11]). *Let Δt, H, and a_n, b_n, c_n, γ_n (for integers $n \geq 0$) be nonnegative numbers such that*

$$a_l \;+\; \Delta t \sum_{n=0}^{l} b_n \;\leq\; \Delta t \sum_{n=0}^{l} \gamma_n\, a_n \;+\; \Delta t \sum_{n=0}^{l} c_n \;+\; H \quad \text{for } l \geq 0.$$

Suppose that $\Delta t\, \gamma_n < 1$ for all n, and set $\sigma_n = (1 - \Delta t\, \gamma_n)^{-1}$. Then,

$$a_l \;+\; \Delta t \sum_{n=0}^{l} b_n \;\leq\; \exp\left(\Delta t \sum_{n=0}^{l} \sigma_n\, \gamma_n\right) \left\{ \Delta t \sum_{n=0}^{l} c_n \;+\; H \right\} \quad \text{for } l \geq 0. \tag{3.12}$$

4. A priori error estimate. In this section we analyze the error between the finite element approximation given by (3.5) and the true solution. A priori error estimates for the approximation are given in Theorem 2.

THEOREM 2. *Assume that for $\acute{d}/4 < \alpha < 1$, (1.1)–(1.3) has a solution u satisfying $u_t \in L^2(0,T;H^{k+1}(\Omega))$, $u_{tt} \in L^2(0,T;L^2(\Omega))$, with $u^0 \in H^{k+1}(\Omega)$. In addition, assume that $\Delta t \leq c\, h$. Then, the finite element approximation (3.5) is convergent to the solution of (1.1)–(1.3) on the interval $(0,T)$ as $\Delta t, h \to 0$. The approximation u_h satisfies the following error estimates:*

(4.1)

$$|||u - u_h|||_{0,\alpha} \leq C\, \left(h^{k+1} \|u_t\|_{0,k+1} \;+\; h^{(k+1-\alpha)}\, |||u|||_{0,k+1} \;+\; \Delta t\, \|u_t\|_{0,1} \;+\; \Delta t \|u_{tt}\|_{0,0} \right),$$

$$|||u - u_h|||_{\infty,0} \leq C\, \Big(h^{k+1} \|u_t\|_{0,k+1} \;+\; h^{(k+1-\alpha)}\, |||u|||_{0,k+1} \;+\; \Delta t\, \|u_t\|_{0,1} \;+\; \Delta t \|u_{tt}\|_{0,0}$$

$$+\, h^{k+1}\, |||u|||_{\infty,k+1} \Big). \tag{4.2}$$

Remarks. 1. $u_t \in L^2(0,T;H^{k+1}(\Omega))$, $u_0 \in H^{k+1}(\Omega)$ implies $u \in L^2(0,T;H^{k+1}(\Omega)) \cap L^\infty(0,T;H^{k+1}(\Omega))$.

2. As previously defined in (3.1), k is the polynomial order of the approximation functions u_h^n.

In order to establish the estimates (4.1), (4.2), we begin by introducing the following notation. Let $u^n = u(t_n)$ represent the solution of (1.1)–(1.3), and u_h^n denote the solution of (3.5).

For $\mathcal{U}^n \in X_h$, define $\Lambda^n, E^n, \epsilon_u$, as

$$\Lambda^n = u^n - \mathcal{U}^n, \qquad E^n = \mathcal{U}^n - u_h^n, \qquad \epsilon_u = u^n - u_h^n.$$

The proof of Theorem 2 is established in three steps:

1. Prove a lemma, assuming the induction hypothesis.
2. Show that the induction hypothesis is true.
3. Prove the error estimates given in (4.1), (4.2).

Step 1. We prove the following lemma.

LEMMA 6. *Under the induction hypothesis $\|u_h^j\|_1 \leq \mathcal{K}$ for $j = 0, 1, \ldots, l-1$, we have that*

$$\|E^l\|^2 \leq G(\Delta t, h), \tag{4.3}$$

where

$$G(\Delta t,h) \;=\; C\Big(h^{2(k+1)}\|u_t\|^2_{0,k+1} \;+\; h^{2(k+1-\alpha)}\,|||u|||^2_{0,k+1} \\ +\;(\Delta t)^2\,\|u_t\|^2_{0,1} + \;(\Delta t)^2\|u_{tt}\|^2_{0,0}\Big).$$

Proof of Lemma 6. We present the proof for $\Omega \subset \mathbb{R}^2$; the case for $\Omega \subset \mathbb{R}^3$ follows analogously.

From (1.1), (1.2) we have that the true solution u satisfies

$$(d_t u^n, v) \;+\; \langle \mathcal{D}^{2\alpha} u^n,\, v\rangle \;+\; (u^n B(u_h^{n-1}),\, \nabla v) = (f^n, v) \;-\; (u_t - d_t u^n, v) \\ -\,(u^n B(u^n - u_h^{n-1}),\, \nabla v), \quad v \in X_h. \tag{4.4}$$

Subtracting (3.5) from (4.4), we obtain the following equation for ϵ_u:

$$(d_t \epsilon_u, v) \;+\; \langle \mathcal{D}^{2\alpha} \epsilon_u,\, v\rangle \;+\; (\epsilon_u B(u_h^{n-1}),\, \nabla v) = -\,(u_t - d_t u^n, v) \\ -\,(u^n B(u^n - u_h^{n-1}),\, \nabla v), \quad v \in X_h. \tag{4.5}$$

Substituting $\epsilon_u = E^n + \Lambda^n$, $v = E^n$ into (4.5), we obtain

$$(d_t E^n, E^n) \;+\; \langle \mathcal{D}^{2\alpha} E^n,\, E^n\rangle \;+\; (E^n B(u_h^{n-1}),\, \nabla E^n) \;=\; F(E^n), \tag{4.6}$$

where

$$F(E^n) := -(d_t \Lambda^n, E^n) \;-\; \langle \mathcal{D}^{2\alpha} \Lambda^n,\, E^n\rangle \;-\; (\Lambda^n B(u_h^{n-1}),\, \nabla E^n) \\ -\,(u_t - d_t u^n, E^n) \;-\; (u^n B(u^n - u_h^{n-1}),\, \nabla E^n).$$

Note that

$$(d_t E^n, E^n) = \frac{1}{\Delta t}\left((E^n, E^n) - (E^{n-1}, E^n)\right) \\ \geq \frac{1}{2\Delta t}\left(\|E^n\|^2 - \|E^{n-1}\|^2\right),$$

and from (2.32)

$$(E^n B(u_h^{n-1}),\, \nabla E^n) \leq \epsilon_2\,\|E^n\|^2_\alpha \;+\; C_1 \epsilon_2^{-C_2} \|\nabla \cdot B(u_h^{n-1})\|^{C_3} \|E^n\|^2. \tag{4.7}$$

Multiplying (4.6) by $2\,\Delta t$, summing from $n = 1$ to l, and using (2.3), we have

$$\left(\|E^l\|^2 - \|E^0\|^2\right) + 2\,(C_c - \epsilon_2) \sum_{n=1}^{l} \Delta t \|E^n\|^2_\alpha \\ \leq\; 2\Delta t \sum_{n=1}^{l} C_1 \epsilon_2^{-C_2} \|\nabla \cdot B(u_h^{n-1})\|^{C_3} \|E^n\|^2 \;+\; 2\Delta t \sum_{n=1}^{l} F(E^n). \tag{4.8}$$

We now estimate each term in $\mathrm{F}(E^n)$:

$$\begin{aligned}(d_t\Lambda^n, E^n) &\le \|E^n\|\,\|d_t\Lambda^n\| \\ &\le \frac{1}{2}\|E^n\|^2 + \frac{1}{2}\|d_t\Lambda^n\|^2. \end{aligned} \tag{4.9}$$

Using (2.2),

$$\begin{aligned}\langle \mathcal{D}^{2\alpha}\Lambda^n,\, E^n\rangle &\le C_t\|E^n\|_\alpha\,\|\Lambda^n\|_\alpha \\ &\le \epsilon_4\|E^n\|_\alpha^2 + \frac{C_t^2}{4\epsilon_4}\|\Lambda^n\|_\alpha^2. \end{aligned} \tag{4.10}$$

Using duality with respect to the L^2 inner product,

$$\begin{aligned}(\Lambda^n B(u_h^{n-1}),\, \nabla E^n) &\le \|\nabla E^n\|_{-(1-\alpha)}\,\|\Lambda^n B(u_h^{n-1})\|_{(1-\alpha)} \\ &\le \epsilon_5\|E^n\|_\alpha^2 + \frac{C_4}{4\epsilon_5}\|\Lambda^n B(u_h^{n-1})\|_{(1-\alpha)}^2. \end{aligned} \tag{4.11}$$

For the next term in $F(E^n)$ we use

$$(u_t - d_t u^n, E^n) \;\le\; \|E^n\|\,\|u_t - d_t u^n\| \;\le\; \frac{1}{2}\|E^n\|^2 + \frac{1}{2}\|u_t - d_t u^n\|^2. \tag{4.12}$$

The remaining term is rewritten as the sum of three terms.

$$\begin{aligned}(u^n B(u^n - u_h^{n-1}),\, \nabla E^n) &= (u^n B(u^n - u^{n-1}),\, \nabla E^n) + (u^n B(u^{n-1} - u_h^{n-1}),\, \nabla E^n) \\ &= (u^n B(u^n - u^{n-1}),\, \nabla E^n) + (u^n B(\Lambda^{n-1}),\, \nabla E^n) \\ &\quad + (u^n B(E^{n-1}),\, \nabla E^n)\ . \end{aligned}$$

Each of these terms is rewritten in a similar fashion as in (4.11):

$$(u^n B(u^n - u^{n-1}),\, \nabla E^n) \le \epsilon_7\|E^n\|_\alpha^2 + \frac{C_4}{4\epsilon_7}\,\|u^n B(u^n - u^{n-1})\|_{(1-\alpha)}^2, \tag{4.13}$$

$$(u^n B(\Lambda^{n-1}),\, \nabla E^n) \le \epsilon_8\|E^n\|_\alpha^2 + \frac{C_4}{4\epsilon_8}\,\|u^n B(\Lambda^{n-1})\|_{(1-\alpha)}^2, \tag{4.14}$$

$$(u^n B(E^{n-1}),\, \nabla E^n) \le \epsilon_9\|E^n\|_\alpha^2 + \frac{C_4}{4\epsilon_9}\,\|u^n B(E^{n-1})\|_{(1-\alpha)}^2. \tag{4.15}$$

Combining (4.8)–(4.15), for $\epsilon_1, \ldots, \epsilon_9$ appropriately chosen, there exist constants

$C_j > 0$ such that

$$\begin{aligned}
\left(\|E^l\|^2 - \|E^0\|^2\right) + C_{12} \sum_{n=1}^{l} \Delta t \|E^n\|_\alpha^2 & \\
&\leq \Delta t \sum_{n=1}^{l} \left(C_{13}\|\nabla \cdot B(u_h^{n-1})\|^{C_3} + C_{14}\right) \|E^n\|^2 \\
&+ \Delta t \sum_{n=1}^{l} C_{15} \|u^n B(E^{n-1})\|_{(1-\alpha)}^2 \\
&+ \Delta t \sum_{n=1}^{l} C_{16} \|d_t \Lambda^n\|^2 + \Delta t \sum_{n=1}^{l} C_{17} \|\Lambda^n\|_\alpha^2 \\
&+ \Delta t \sum_{n=1}^{l} C_{18} \|\Lambda^n B(u_h^{n-1})\|_{(1-\alpha)}^2 \\
&+ \Delta t \sum_{n=1}^{l} C_{19} \|u^n B(u^n - u^{n-1})\|_{(1-\alpha)}^2 \\
&+ \Delta t \sum_{n=1}^{l} C_{20} \|u^n B(\Lambda^{n-1})\|_{(1-\alpha)}^2 \\
&+ \Delta t \sum_{n=1}^{l} C_{21} \|u_t - d_t u^n\|^2. \qquad (4.16)
\end{aligned}$$

We now apply the interpolation property of the approximation space to estimate the terms on the right-hand side (RHS) of (4.16).

$$\begin{aligned}
\sum_{n=1}^{l} \Delta t \|d_t \Lambda^n\|^2 &= \sum_{n=1}^{l} \Delta t \left\| \frac{1}{\Delta t} \int_{t_{n-1}}^{t_n} 1 \, \frac{\partial \Lambda}{\partial t} \, dt \right\|^2 \\
&\leq \sum_{n=1}^{l} \Delta t \left(\frac{1}{\Delta t}\right)^2 \int_\Omega \left(\int_{t_{n-1}}^{t_n} 1 \, dt \right) \left(\int_{t_{n-1}}^{t_n} \left(\frac{\partial \Lambda}{\partial t}\right)^2 dt \right) dx \\
&\leq C h^{2k+2} \|u_t\|_{0,k+1}^2. \qquad (4.17)
\end{aligned}$$

Note that $(d_t u^n - u_t^n)$ may be expressed as

$$d_t u^n - u_t^n = \frac{1}{\Delta t} \int_{t_{n-1}}^{t_n} u_{tt}(\cdot, t)(t_{n-1} - t) \, dt.$$

Also,

$$\begin{aligned}
\left(\frac{1}{\Delta t} \int_{t_{n-1}}^{t_n} u_{tt}(\cdot, t)(t_{n-1} - t) \, dt \right)^2 &\leq \frac{1}{(\Delta t)^2} \int_{t_{n-1}}^{t_n} u_{tt}^2(\cdot, t) \, dt \int_{t_{n-1}}^{t_n} (t_{n-1} - t)^2 \, dt \\
&= \frac{1}{3} \Delta t \int_{t_{n-1}}^{t_n} u_{tt}^2(\cdot, t) \, dt.
\end{aligned}$$

Therefore it follows that

$$\sum_{n=1}^{l} \Delta t \, \|u_t - d_t u^n\|^2 \leq \sum_{n=1}^{l} \Delta t \int_\Omega \frac{1}{3} \Delta t \int_{t_{n-1}}^{t_n} u_{tt}^2(\cdot, t)\, dt\, dx$$
$$= \frac{1}{3} (\Delta t)^2 \|u_{tt}\|_{0,0}^2. \tag{4.18}$$

Next we estimate the terms in (4.16) involving $B(\cdot)$ using (2.4), (2.19), and (3.8). These estimates for $B(\cdot)$ are dimension-specific. We present the case for $\acute{d} = 2$.

For the first term on the RHS,

$$\|\nabla \cdot B(u_h^{n-1})\| \;\leq\; C\|B(u_h^{n-1})\|_1 \;\leq\; C\|u_h^{n-1}\|_1 \;\leq\; C\mathcal{K}\,. \tag{4.19}$$

Using interpolation between L^2 and H^α and Young's inequality, we obtain, for $\delta \in (0, 2\alpha - 1)$,

$$\|u^n B(E^{n-1})\|_{(1-\alpha)}^2 \leq C\, \|B(E^{n-1})\|_{(1-\alpha+\delta)}^2\, \|u^n\|_1^2$$
$$\leq C\|E^{n-1}\|^{2(2\alpha-1-\delta)/\alpha}\, \|E^{n-1}\|_\alpha^{2(1-\alpha+\delta)/\alpha}\, \|u^n\|_1^2$$
$$\leq \epsilon_{10}\|E^{n-1}\|_\alpha^2 \;+\; C\, \|u^n\|_1^{2(\alpha/(2\alpha-1-\delta))}\|E^{n-1}\|^2. \tag{4.20}$$

With the interpolation error bound $\|\Lambda^n\|_{(1-\alpha+\delta)} \leq Ch^{k+\alpha-\delta}\, \|u^n\|_{k+1}$,

$$\|\Lambda^n B(u_h^{n-1})\|_{(1-\alpha)} \leq C\, \|\Lambda^n\|_{(1-\alpha+\delta)} \|B(u_h^{n-1})\|_1$$
$$\leq C\mathcal{K} h^{k+\alpha-\delta}\, \|u^n\|_{k+1}. \tag{4.21}$$

To estimate $\|u^n B(\Lambda^{n-1})\|_{(1-\alpha)}$ we proceed similarly.

$$\|u^n B(\Lambda^{n-1})\|_{(1-\alpha)} \leq C\, \|u^n\|_1\, \|B(\Lambda^{n-1})\|_{(1-\alpha+\delta)}$$
$$\leq Ch^{k+\alpha-\delta}\, \|u^{n-1}\|_{k+1}\, \|u^n\|_1. \tag{4.22}$$

Using

$$\|u^n - u^{n-1}\|_1^2 \;\leq\; \Delta t \int_{t_{n-1}}^{t_n} \|u_t\|_1^2\, dt,$$

we have that

$$\|u^n B(u^n - u^{n-1})\|_{(1-\alpha)}^2 \leq C\, \|u^n\|_{(1-\alpha+\delta)}^2\, \|B(u^n - u^{n-1})\|_1^2$$
$$\leq C\|u^n\|_{(1-\alpha+\delta)}^2\, \|u^n - u^{n-1}\|_1^2$$
$$\leq C\Delta t\, \|u^n\|_{(1-\alpha+\delta)}^2 \int_{t_{n-1}}^{t_n} \|u_t\|_1^2\, dt. \tag{4.23}$$

From (4.17)–(4.23) and $\|E^0\| = 0$, estimate (4.16) becomes (using that $\|u\|_1$ is bounded for $t \in [0, T]$)

$$\|E^l\|^2 \;+\; C_{12} \sum_{n=1}^{l} \Delta t \|E^n\|_\alpha^2 \leq \Delta t \sum_{n=1}^{l} C_{22}\|E^n\|^2$$
$$+ Ch^{2(k+1)}\|u_t\|_{0,k+1}^2 \;+\; Ch^{2(k+1-\alpha)}\, |||u|||_{0,k+1}^2$$
$$+ C(\Delta t)^2\, \|u_t\|_{0,1}^2$$
$$+ C\, h^{2(k+\alpha-\delta)}\, |||u|||_{0,k+1}^2$$
$$+ C\, (\Delta t)^2 \|u_{tt}\|_{0,0}^2. \tag{4.24}$$

Finally, as $\alpha > 0.5$, with $\Delta t < 1/C_{22}$ and the associations $a_l = \|E^l\|^2$, $b_n = C_{12}\|E^n\|_\alpha^2$, $\gamma_n = C_{22}$, $c_n = 0$, $H = \tilde{C}\left(h^{2(k+1)}\|u_t\|_{0,k+1}^2 + h^{2(k+1-\alpha)}\|\!|u|\!\|_{0,k+1}^2 + (\Delta t)^2\|u_t\|_{0,1}^2 + (\Delta t)^2\|u_{tt}\|_{0,0}^2\right)$, applying Gronwall's lemma, we obtain the bound given in (4.3), where $C = \tilde{C}\exp(TC_{22}/(1-\Delta t C_{22}))$. ∎

Step 2. We show that the induction hypothesis (IH1) is true.

Assume that $\|u_h^j\|_1 \le \mathcal{K}$ for $j = 0, 1, \ldots, l-1$. Using the interpolation property, inverse estimate (3.3), and (4.3), we have that

$$\begin{aligned}\|\nabla u_h^l\| &\le \|\nabla(u_h^l - u^l)\| + \|\nabla u^l\| \\ &\le \|\nabla E^l\| + \|\nabla \Lambda^l\| + \|\nabla u^l\| \\ &\le C\left(h^{-1}\|E^l\| + \|\nabla u^l\|\right) \\ &\le Ch^{-1}(h^{k+1-\alpha} + \Delta t) + C\|\nabla u^l\|. \end{aligned} \tag{4.25}$$

Thus as C is independent of l, $u \in L^\infty(0,T;H^1(\Omega))$, for $\Delta t \le ch$, we have that $\|\nabla u_h^l\|$ is bounded.

An analogous argument shows that $\|u_h^l\|$ is also bounded independent of h and l. ∎

Step 3. We derive the error estimates in (4.1) and (4.2).

Proof of Theorem 2. To establish (4.1), from (4.24) and (4.3), and using $T = N\Delta t$,

$$\|\!|E|\!\|_{0,\alpha}^2 \;=\; \sum_{n=1}^{N} \Delta t\,\|E^n\|_\alpha^2 \;\le\; C\,(T+1)G(\Delta t, h).$$

Hence, using the interpolation property and that

$$\|\!|u - u_h|\!\|_{0,\alpha} \;\le\; \|\!|E|\!\|_{0,\alpha} + \|\!|\Lambda|\!\|_{0,\alpha},$$

estimate (4.1) then follows.

Using estimates (4.3) and approximation properties, we have

$$\begin{aligned}\|\!|u - u_h|\!\|_{\infty,0}^2 &\le \|\!|E|\!\|_{\infty,0}^2 + \|\!|\Lambda|\!\|_{\infty,0}^2 \\ &\le G(\Delta t, h) + h^{2k+2}\,\|\!|u|\!\|_{\infty,k+1}^2,\end{aligned}$$

which yields estimate (4.2). ∎

5. Numerical results. In this section we illustrate the predicted convergence results given in Theorem 2 with numerical computations for $\Omega \subset \mathbb{R}^2$. For points $x, y \in \mathbb{R}^2$ we use $x = (x_1, x_2)$ and $y = (y_1, y_2)$. For ease of notation, for $\theta \in [0, 2\pi)$ we let $D_\theta^{-\sigma} u := D_\nu^{-\sigma} u$, where $\nu = [\cos\theta, \sin\theta]^T$.

It is noteworthy to again remark that fractional derivatives/diffusion operators are *nonlocal* operators. Consequently, a sparse coefficient matrix, characteristic of using a finite element basis for the test and trial space, does not occur when approximating space-fractional diffusion equations. For the fractional diffusion operator given by (2.7) it is computationally convenient to use the bilinear form (see (3.5))

$$\mathcal{B}(u_h^n, v) \;:=\; \frac{1}{\Delta t}(u_h^n, v) \;-\; \int_0^{2\pi} a\,(D_\theta^{2\alpha-1} u_h^n,\, D_{\theta+\pi} v)\, M(d\theta) \;+\; (u_h^n\, B(u_h^{n-1}),\, \nabla v).$$

Note that $\mathrm{supp}(D_{\theta+\pi} v) \subseteq \mathrm{supp}(v)$, whereas, in general $\mathrm{supp}(D_{\theta+\pi}^{\alpha} v) \supset \mathrm{supp}(v)$. For a discussion on the implementation of the finite element method approximation for the fractional diffusion operator (5.1) in $\mathbb{R}^2$ see [15].

TABLE 5.1
Experimental error results for Example 1 for the fractional diffusion operator and no B term.

h	$\|\|\|u-u_h\|\|\|_{\infty,0}$	Cvge. rate	$\|\|\|u-u_h\|\|\|_{0,0}$	Cvge. rate
1/4	$7.44283 \cdot 10^{-3}$		$5.735602 \cdot 10^{-3}$	
1/8	$2.991413 \cdot 10^{-3}$	1.32	$2.281621 \cdot 10^{-3}$	1.33
1/12	$1.784701 \cdot 10^{-3}$	1.27	$1.365371 \cdot 10^{-3}$	1.27
1/16	$1.232144 \cdot 10^{-3}$	1.29	$9.449112 \cdot 10^{-4}$	1.28
1/20	$9.411762 \cdot 10^{-4}$	1.21	$7.232338 \cdot 10^{-4}$	1.20
1/24	$7.470209 \cdot 10^{-4}$	1.27	$5.748151 \cdot 10^{-4}$	1.26

TABLE 5.2
Experimental error results for Example 1 for the fractional diffusion operator and $b(x,y) = (x-y)$.

h	$\|\|\|u-u_h\|\|\|_{\infty,0}$	Cvge. rate	$\|\|\|u-u_h\|\|\|_{0,0}$	Cvge. rate
1/4	$7.160147 \cdot 10^{-3}$		$5.603630 \cdot 10^{-3}$	
1/8	$2.907621 \cdot 10^{-3}$	1.30	$2.242255 \cdot 10^{-3}$	1.32
1/12	$1.729328 \cdot 10^{-3}$	1.28	$1.336439 \cdot 10^{-3}$	1.28
1/16	$1.185318 \cdot 10^{-3}$	1.31	$9.186409 \cdot 10^{-4}$	1.30
1/20	$8.977671 \cdot 10^{-4}$	1.25	$6.978723 \cdot 10^{-4}$	1.23
1/24	$7.053820 \cdot 10^{-4}$	1.32	$5.498338 \cdot 10^{-4}$	1.31

TABLE 5.3
Experimental error results for Example 1 for the fractional diffusion operator and $b(x,y) = (x-y)/|x-y|^2$.

h	$\|\|\|u-u_h\|\|\|_{\infty,0}$	Cvge. rate	$\|\|\|u-u_h\|\|\|_{0,0}$	Cvge. rate
1/4	$6.940074 \cdot 10^{-3}$		$5.532543 \cdot 10^{-3}$	
1/8	$2.735276 \cdot 10^{-3}$	1.34	$2.160750 \cdot 10^{-3}$	1.36
1/12	$1.564182 \cdot 10^{-3}$	1.38	$1.248073 \cdot 10^{-3}$	1.35
1/16	$1.041509 \cdot 10^{-3}$	1.41	$8.274240 \cdot 10^{-4}$	1.43
1/20	$7.742380 \cdot 10^{-4}$	1.33	$6.060513 \cdot 10^{-4}$	1.40
1/24	$5.960361 \cdot 10^{-4}$	1.50	$4.586714 \cdot 10^{-4}$	1.53

The fractional differential operator used in our computations was (see (2.7))

$$D_M^{2\alpha} u(x) \;:=\; -\frac{1}{\pi} \int_{\theta=0}^{2\pi} D_\theta^{2\alpha} u(x)\, d\theta, \tag{5.1}$$

which we approximate as

$$D_M^{2\alpha} u(x) \;\approx\; -\;\frac{1}{2} D_0^{2\alpha} u(x) \;-\; \frac{1}{2} D_{\pi/2}^{2\alpha} u(x) \;-\; \frac{1}{2} D_{\pi}^{2\alpha} u(x) \;-\; \frac{1}{2} D_{3\pi/2}^{2\alpha} u(x)\;. \tag{5.2}$$

The value of α used was $\alpha = 0.75$.

The approximation space X_h was taken to be the space of continuous piecewise linear functions, i.e., $k = 1$.

From Theorem 2, (4.1), and (4.2), we have the predicted rates of convergence for $\Delta t = Ch^{k+1-\alpha}$ $(= Ch^{1.25}$ for $k = 1$, $\alpha = 0.75)$ of

$$\|\|\|u-u_h\|\|\|_{0,\alpha} \;\sim O(h^{1.25}), \qquad \|\|\|u-u_h\|\|\|_{\infty,0} \;\sim O(h^{1.25}). \tag{5.3}$$

In Tables 5.1–5.6 we give the results for $\|\|\|u-u_h\|\|\|_{0,0}$, which from (4.1) and (5.3) are predicted to satisfy

$$\|\|\|u-u_h\|\|\|_{0,0} \;\sim O(h^{1.25}).$$

TABLE 5.4
Experimental error results for Example 1 *for the usual diffusion operator and no B term.*

h	$\|\|\|u-u_h\|\|\|_{\infty,0}$	Cvge. rate	$\|\|\|u-u_h\|\|\|_{0,0}$	Cvge. rate
1/4	$1.478687 \cdot 10^{-3}$		$1.228611 \cdot 10^{-3}$	
1/8	$8.125659 \cdot 10^{-4}$	0.86	$6.938199 \cdot 10^{-4}$	0.82
1/12	$5.538746 \cdot 10^{-4}$	0.95	$4.874202 \cdot 10^{-4}$	0.87
1/16	$4.180683 \cdot 10^{-4}$	0.98	$3.751933 \cdot 10^{-4}$	0.91
1/20	$3.353536 \cdot 10^{-4}$	0.99	$3.048365 \cdot 10^{-4}$	0.93
1/24	$2.798554 \cdot 10^{-4}$	0.99	$2.566538 \cdot 10^{-4}$	0.94

TABLE 5.5
Experimental error results for Example 1 *for the usual diffusion operator and* $b(x,y) = (x-y)$.

h	$\|\|\|u-u_h\|\|\|_{\infty,0}$	Cvge. rate	$\|\|\|u-u_h\|\|\|_{0,0}$	Cvge. rate
1/4	$1.470130 \cdot 10^{-3}$		$1.223586 \cdot 10^{-3}$	
1/8	$8.119423 \cdot 10^{-4}$	0.87	$6.874692 \cdot 10^{-4}$	0.83
1/12	$5.527071 \cdot 10^{-4}$	0.95	$4.812465 \cdot 10^{-4}$	0.88
1/16	$4.172209 \cdot 10^{-4}$	0.98	$3.690746 \cdot 10^{-4}$	0.92
1/20	$3.346173 \cdot 10^{-4}$	0.99	$2.987263 \cdot 10^{-4}$	0.95
1/24	$2.791593 \cdot 10^{-4}$	0.99	$2.505369 \cdot 10^{-4}$	0.96

TABLE 5.6
Experimental error results for Example 1 *for the usual diffusion operator and* $b(x,y) = (x-y)/|x-y|^2$.

h	$\|\|\|u-u_h\|\|\|_{\infty,0}$	Cvge. rate	$\|\|\|u-u_h\|\|\|_{0,0}$	Cvge. rate
1/4	$1.481271 \cdot 10^{-3}$		$1.203806 \cdot 10^{-3}$	
1/8	$8.093125 \cdot 10^{-4}$	0.87	$6.612167 \cdot 10^{-4}$	0.86
1/12	$5.477165 \cdot 10^{-4}$	0.96	$4.556000 \cdot 10^{-4}$	0.92
1/16	$4.135933 \cdot 10^{-4}$	0.98	$3.436790 \cdot 10^{-4}$	0.98
1/20	$3.314616 \cdot 10^{-4}$	0.99	$2.734522 \cdot 10^{-4}$	1.02
1/24	$2.761739 \cdot 10^{-4}$	1.00	$2.253575 \cdot 10^{-4}$	1.06

For comparison, computations were also performed with the usual diffusion operator in place of $D_M^{2\alpha}u$, namely on the equation

$$u_t \;-\; \Delta u \;-\; \nabla \cdot (u\,B(u)) \;=\; f(x). \tag{5.4}$$

For the usual diffusion operator, Δt was chosen as $\Delta t = Ch$. From Theorem 2, the predicted rate of convergence is then

$$\|\|u-u_h\|\|_{0,0} \;\sim O(h), \qquad \|\|u-u_h\|\|_{\infty,0} \;\sim O(h). \tag{5.5}$$

Example 1. For the problem described in (1.1)–(1.3) we take $\Omega = (0,1)\times(0,1)$, and a known solution $u(x_1,x_2,t) = (4t^2-4t+1)(x_1-x_1^2)(x_2-x_2^2)$, with $u^0(x_1,x_2) = u(x_1,x_2,0)$. The RHS of (1.1) was computed using the true solution, and the approximation to $D_M^{2\alpha}u(x)$ given in (5.2).

Computations were performed for $B(u)$ given by (2.8) with

(a) $b(x,y) = 0$, i.e., $B(u) = 0$ (see Tables 5.1, 5.4),

(b) $b(x,y) = x-y$, i.e., a smooth operator B (see Tables 5.2, 5.5),

(c) $b(x,y) = (x-y)/|x-y|^2$ (see Tables 5.3, 5.6).

The results presented in Tables 5.1–5.6 are consistent with those predicted by Theorem 2, given in (5.3). ("Cvge. rate" stands for convergence rate.)

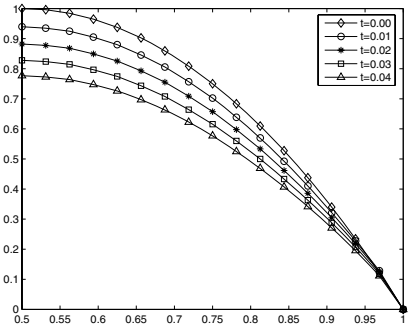


FIG. 5.1. *Time evolution of* (1.1) *for* $c = 0$.

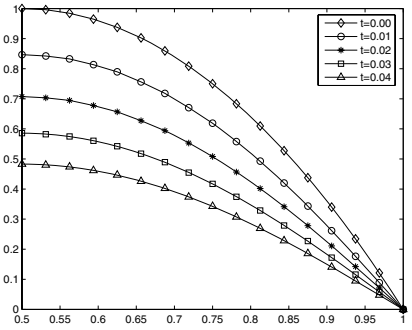


FIG. 5.2. *Time evolution of* (5.4) *for* $c = 0$.

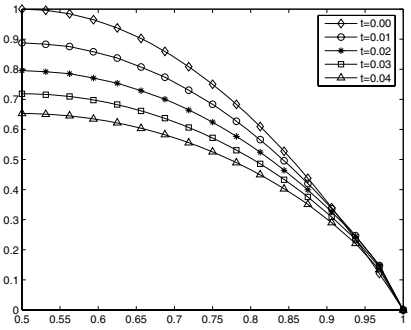


FIG. 5.3. *Time evolution of* (1.1) *for* $c = -1$.

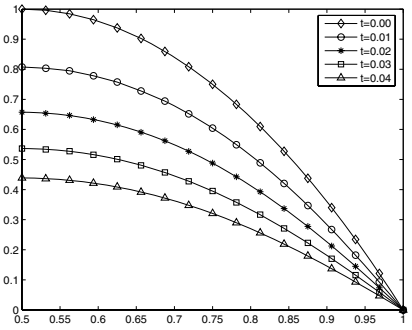


FIG. 5.4. *Time evolution of* (5.4) *for* $c = -1$.

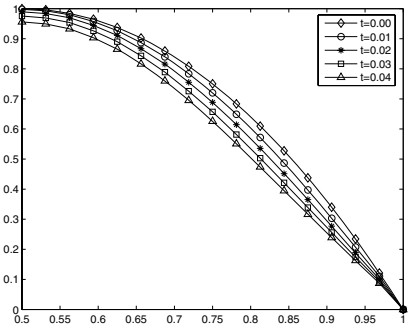


FIG. 5.5. *Time evolution of* (1.1) *for* $c = 1$.

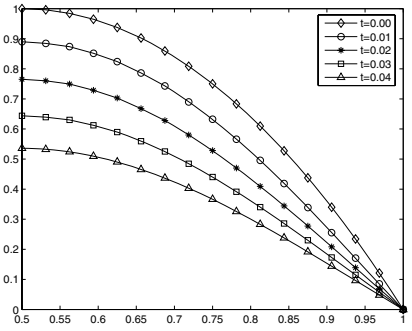


FIG. 5.6. *Time evolution of* (5.4) *for* $c = 1$.

Example 2. In order to demonstrate the influence of the nonlocal quadratic nonlinearity $\nabla \cdot (uB(u))$, we present in Figures 5.1–5.10 the plots of the time evolution of the approximation u_h for the initial value $u^0(x_1, x_2) = 16(x_1 - x_1^2)(x_2 - x_2^2)$. Plots for the fractional diffusion equation are displayed on the left, the usual diffusion equation on the right. Note that u^0 has height 1 at $(1/2, 1/2)$ and is symmetric with respect

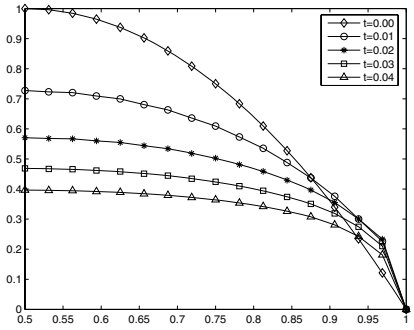


FIG. 5.7. *Time evolution of* (1.1) *for* $c = -5$.

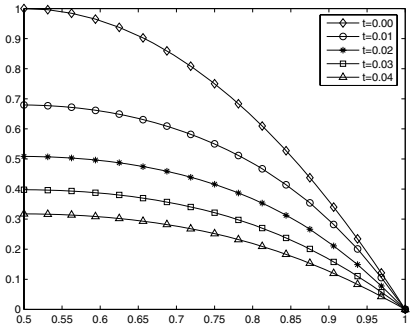


FIG. 5.8. *Time evolution of* (5.4) *for* $c = -5$.

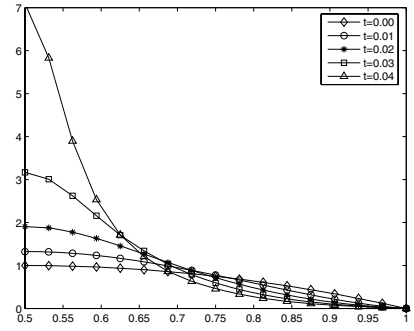


FIG. 5.9. *Time evolution of* (1.1) *for* $c = 5$.

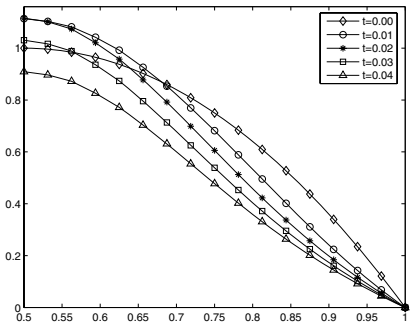


FIG. 5.10. *Time evolution of* (5.4) *for* $c = 5$.

to x_1 and x_2. The profiles given are along the line segment $[x_1, 1/2]$, $1/2 \leq x_1 \leq 1$. The operator $B(u)$ was chosen as in (2.8) with $b(x, y)$ given by (2.9). Values for $c = 0$ (Figures 5.1, 5.2), $c = \pm 1$ (Figures 5.3–5.6), and $c = \pm 5$ (Figures 5.7–5.10) were used.

For the negative values of c the diffusion of u away from the maximum at $(1/2, 1/2)$ is enhanced. For positive values of c the $\nabla \cdot (uB(u))$ term acts "against the diffusion operator" to try and concentrate u at $(1/2, 1/2)$. This behavior is consistent with the case $c < 0$ modeling Brownian diffusion and $c > 0$ being used to model mutual gravitational attraction of particles in clouds (see [4]).

REFERENCES

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

[2] D. A. BENSON, S. W. WHEATCRAFT, AND M. M. MEERSCHAEERT, *The fractional order governing equations of Lévy motion*, Water Resour. Res., 36 (2000), pp. 1413–1423.

[3] J. BERGH AND J. LÖFSTRÖM, *Interpolation Spaces, An Introduction*, Springer-Verlag, New York, Berlin, 1976.

[4] P. BILER AND W. A. WOYCZYŃSKI, *Global and exploding solutions for nonlocal quadratic evolution problems*, SIAM J. Appl. Math., 59 (1998), pp. 845–869.

[5] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, Berlin, 1994.

[6] B. A. CARRERAS, V. E. LYNCH, AND G. M. ZASLAVSKY, *Anomalous diffusion and exit time distribution of particle tracers in plasma turbulence models*, Phys. Plasmas, 8 (2001), pp. 5096–5103.

[7] V. J. Ervin and J. P. Roop, *Variational formulation for the stationary fractional advection dispersion equation*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 558–576.

[8] V. J. Ervin and J. P. Roop, *Variational solution of fractional advection dispersion equation on bounded domains in* $\mathbb{R}^{\acute{d}}$, Numer. Methods for Partial Differential Equations, to appear.

[9] I. M. Gelfand and G. E. Shilov, *Generalized Functions*, Vol. I, Academic Press, New York, 1964.

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

[11] J. G. Heywood and R. Rannacher, *Finite-element approximation of the nonstationary Navier–Stokes problem. Part* IV: *Error analysis for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[12] M. M. Meerschaert, D. A. Benson, and B. Baeumer, *Multidimensional advection and fractional dispersion*, Phys. Rev. E, 59 (1999), pp. 5026–5028.

[13] K. S. Miller and B. Ross, *An Introduction to the Fractional Calculus and Fractional Differential Equations*, John Wiley & Sons, New York, 1993.

[14] J. P. Roop, *Variational Solution of the Fractional Advection Dispersion Equation*, Ph.D. thesis, Department of Mathematical Sciences, Clemson University, Clemson, SC, 2004.

[15] J. P. Roop, *Computational aspects of FEM approximations of fractional advection dispersion equations on bounded domains in* $\mathbb{R}^2$, J. Comput. Appl. Math., 193 (2006), pp. 243–268.

[16] Ch. Schwab, *p− and hp− Finite Element Methods*, Oxford University Press, London, 1998.

[17] R. Seeley, *Topics in pseudo-differential operators*, in Pseudo-Differential Operators, L. Nirenberg, ed., C.I.M.E., Cremonese, Roma, 1968, pp. 168–305.

[18] M. F. Shlesinger, B. J. West, and J. Klafter, *Lévy dynamics of enhanced diffusion: Application to turbulence*, Phys. Rev. Lett., 58 (1987), pp. 1100–1103.

[19] G. M. Zaslavsky, D. Stevens, and H. Weitzner, *Self-similar transport in incomplete chaos*, Phys. Rev. E, 48 (1993), pp. 1683–1694.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 592–609

ALMOST SURE AND MOMENT EXPONENTIAL STABILITY IN THE NUMERICAL SIMULATION OF STOCHASTIC DIFFERENTIAL EQUATIONS*

DESMOND J. HIGHAM†, XUERONG MAO‡, AND CHENGGUI YUAN§

Abstract. Relatively little is known about the ability of numerical methods for stochastic differential equations (SDEs) to reproduce almost sure and small-moment stability. Here, we focus on these stability properties in the limit as the timestep tends to zero. Our analysis is motivated by an example of an exponentially almost surely stable nonlinear SDE for which the Euler–Maruyama (EM) method fails to reproduce this behavior for any nonzero timestep. We begin by showing that EM correctly reproduces almost sure and small-moment exponential stability for sufficiently small timesteps on scalar linear SDEs. We then generalize our results to multidimensional nonlinear SDEs. We show that when the SDE obeys a linear growth condition, EM recovers almost surely exponential stability very well. Under the less restrictive condition that the drift coefficient of the SDE obeys a one-sided Lipschitz condition, where EM may break down, we show that the backward Euler method maintains almost surely exponential stability.

Key words. backward Euler, Euler–Maruyama, implicit, one-sided Lipschitz condition, linear growth condition, Lyapunov exponent, stochastic theta method

AMS subject classifications. 65C30, 60H10

DOI. 10.1137/060658138

1. Introduction.

Stability theory for numerical simulations of stochastic differential equations (SDEs) typically deals with mean-square behavior. Asymptotic, or almost sure, stability is at least as relevant in typical applications, but does not benefit from a well-developed theory. Our general aim here is to address this imbalance.

We begin with a brief overview of relevant work.

A characterization of asymptotic linear stability for a wide class of SDE methods was given in [10, Lemma 5.1], but this turns out to be of limited use in proving analytical results. Some properties for weak methods were derived in [10, section 6], and results for the related T-stability concept can be found in [19]. The issue of whether the asymptotic linear stability region is bounded was analyzed in [5]. Other authors [7, 8, 17] have tested asymptotic stability via numerical experiments.

The related concept of pth moment stability for $0 < p \leq 2$ is interesting in its own right, and in the linear scalar SDE case it is known that as $p \to 0$ this property is equivalent to asymptotic stability; see Theorem 4.1. We note that some analysis in [1] on stochastic difference equations is relevant to the application of a weak Euler–Maruyama (EM) method to a scalar SDE, and further strengthens the connection between pth moment and asymptotic stability. Similarly, the results in [2] are relevant to EM on a scalar SDE; in this case the emphasis is on polynomial, rather

*Received by the editors April 25, 2006; accepted for publication (in revised form) October 2, 2006; published electronically April 3, 2007.
http://www.siam.org/journals/sinum/45-2/65813.html
†Department of Mathematics, University of Strathclyde, Glasgow, G1 1XH, UK (djh@maths.strath.ac.uk).
‡Department of Statistics and Modelling Science, University of Strathclyde, Glasgow G1 1XH, UK (xuerong@stams.strath.ac.uk).
§Department of Mathematics, University of Wales Swansea, Swansea SA2 8PP, UK (c.yuan@swansea.ac.uk).

than the generic exponential, rates of convergence. In [4], moment stability for SDEs with delay is studied in the presence of a suitable Lyapunov function.

Unlike in the mean-square case [10], we are not aware of any numerical methods that, on a reasonable class of SDEs, have been proved to possess an asymptotic stability analogue of deterministic A-stability [9]; "problem stable implies numerical method stable for all stepsizes."

In this work, we focus on a more fundamental property of the form "problem stable implies numerical method stable for sufficiently small stepsizes," where stability is meant in the exponential asymptotic sense and independently of the size of initial data. To show that this is a nontrivial issue, we give a nonlinear example in section 3 where, for arbitrarily small timesteps, the basic EM method may fail to preserve stability. This motivates the subsequent analysis. We find conditions under which EM does preserve exponential asymptotic stability for small timesteps, and we show that introducing implicitness, in the form of the backward Euler method, produces good results on a class of SDEs that includes our motivating example.

More precisely, we prove positive results for scalar-noise SDEs that are linear (section 4) or satisfy linear growth conditions (section 5). Then in section 6 we show that backward Euler is successful under a one-sided Lipschitz condition on the drift. Section 7 shows how the results generalize to multidimensional noise.

2. Notation. Throughout this paper, we let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathbb{P})$ be a complete probability space with a filtration $\{\mathcal{F}_t\}_{t\geq 0}$ that is increasing and right continuous, with $\mathcal{F}_0$ containing all $\mathbb{P}$-null sets. Let $B(t)$ be a scalar Brownian motion defined on the probability space. Let $|\cdot|$ denote both the Euclidean norm in $\mathbb{R}^n$ and the trace (or Frobenius) norm in $\mathbb{R}^{n\times m}$. The inner product of x, y in $\mathbb{R}^n$ is denoted by $\langle x, y\rangle$. We use $a \vee b$ to denote $\max(a, b)$, $a \wedge b$ to denote $\min(a, b)$, and a.s. to mean almost surely.

We are concerned with the n-dimensional nonlinear Itô SDE

$$dx(t) = f(x(t))dt + g(x(t))dB(t), \quad t \geq 0, \qquad \text{given } 0 \neq x(0) \in \mathbb{R}^n. \tag{2.1}$$

As a standing hypothesis, we assume that $f, g : \mathbb{R}^n \to \mathbb{R}^n$ are smooth enough for the SDE (2.1) to have a unique global solution $x(t)$ on $[0, \infty)$ (see, for example, [15], for sufficient conditions). We make two remarks.

- Scalar Brownian motion $B(t)$ is used to make the analysis in sections 5 and 6 more accessible. In section 7 we state how our results extend to the case of multidimensional noise.
- The restriction to a deterministic initial condition is convenient and does not lose any generality when asymptotic stability is studied; see, for example, [15, section 4.2].

The EM method applied to (2.1) produces approximations $X_k \approx x(k\Delta t)$, where $X_0 = x(0)$ and

$$X_{k+1} = X_k + \Delta t f(X_k) + g(X_k)\Delta B_k. \tag{2.2}$$

Here $\Delta t > 0$ is the timestep and $\Delta B_k := B((k+1)\Delta t) - B(k\Delta t)$ is the Brownian increment. We will also consider the more general stochastic theta (ST) method which takes the form

$$X_{k+1} = X_k + \Delta t\left((1-\theta)f(X_k) + \theta f(X_{k+1})\right) + g(X_k)\Delta B_k, \tag{2.3}$$

where $\theta \in [0, 1]$ is a fixed parameter. For $\theta = 0$, ST reduces to EM. For $\theta \neq 0$ (2.3) defines X_{k+1} implicitly. We will refer to the $\theta = 1$ case as backward Euler (BE).

3. Motivating example. For the scalar cubic SDE

$$dx(t) = \left(x(t) - x(t)^3\right) dt + 2x(t)dB(t) \tag{3.1}$$

it follows from Theorem 6.1 in section 6 below that

$$\limsup_{t\to\infty} \frac{1}{t} \log|x(t)| \le -1 \quad \text{a.s.} \tag{3.2}$$

The EM method (2.2) applied to (3.1) produces

$$X_{k+1} = X_k \left(1 + \Delta t - \Delta t X_k^2 + 2\Delta B_k\right). \tag{3.3}$$

LEMMA 3.1. *Suppose* $0 < \Delta t < 1$. *If* $|X_1| \ge 2^4/\sqrt{\Delta t}$ *in* (3.3), *then*

$$\mathbb{P}\left(|X_k| \ge \frac{2^{k+3}}{\sqrt{\Delta t}} \ \forall\, k \ge 1\right) \ge \exp\left(-4e^{-2/\sqrt{\Delta t}}\right).$$

Proof. First, we show that

$$|X_k| \ge \frac{2^{k+3}}{\sqrt{\Delta t}} \text{ and } |\Delta B_k| \le 2^k \quad \Rightarrow \quad |X_{k+1}| \ge \frac{2^{k+4}}{\sqrt{\Delta t}}. \tag{3.4}$$

To see this, suppose $|X_k| \ge 2^{k+3}/\sqrt{\Delta t}$. Then

$$\begin{aligned} |X_{k+1}| &\ge |X_k| \left|\Delta t |X_k^2| - 1 - \Delta t - 2|\Delta B_k|\right| \\ &\ge \frac{2^{k+3}}{\sqrt{\Delta t}} \left|2^{2k+6} - 1 - \Delta t - 2|\Delta B_k|\right|. \end{aligned}$$

Hence, $|X_{k+1}| \ge 2^{k+4}/\sqrt{\Delta t}$ if

$$2^{2k+6} - 1 - \Delta t - 2|\Delta B_k| \ge 2;$$

that is,

$$2|\Delta B_k| \le 2^{2k+6} - 3 - \Delta t.$$

Since $2^{2k+6} - 3 - \Delta t \ge 2^{2k+6} - 4 \ge 2^{k+1}\ \forall\, k \ge 0$, the implication (3.4) follows.

From (3.4), given that $|X_1| \ge 2^4/\sqrt{\Delta t}$, the event that $\{|X_k| \ge 2^{k+3}/\sqrt{\Delta t}, \forall\, 1 \le k \le K\}$ contains the event that $\{|\Delta B_k| \le 2^k\ \forall\, 1 \le k \le K\}$. So, because the $\{\Delta B_k\}$ are independent,

$$\mathbb{P}\left(|X_k| \ge \frac{2^{k+3}}{\sqrt{\Delta t}} \ \forall\, 1 \le k \le K\right) \ge \prod_{k=1}^{K} \mathbb{P}\left(|\Delta B_k| \le 2^k\right). \tag{3.5}$$

Now, because $\Delta B_k \sim N(0, \Delta t)$, we have

$$\begin{aligned} \mathbb{P}\left(|\Delta B_k| \ge 2^k\right) &= \mathbb{P}\left(\frac{|\Delta B_k|}{\sqrt{\Delta t}} \ge \frac{2^k}{\sqrt{\Delta t}}\right) \\ &= \frac{2}{\sqrt{2\pi}} \int_{2^k/\sqrt{\Delta t}}^{\infty} e^{-x^2/2}\, dx \\ &\le \frac{2}{\sqrt{2\pi}} \int_{2^k/\sqrt{\Delta t}}^{\infty} e^{-x}\, dx \\ &= \frac{2}{\sqrt{2\pi}} \exp\left(-\frac{2^k}{\sqrt{\Delta t}}\right). \end{aligned}$$

Hence, in (3.5)

$$\mathbb{P}\left(|X_k| \geq \frac{2^{k+3}}{\sqrt{\Delta t}} \ \forall\, 1 \leq k \leq K\right) \geq \prod_{k=1}^{K} \left(1 - \exp\left(-\frac{2^k}{\sqrt{\Delta t}}\right)\right).$$

Since

$$\log(1-u) \geq -2u \qquad \text{for } 0 < u < \frac{1}{2},$$

we then have

$$\begin{aligned} \log\left(\mathbb{P}\left(|X_k| \geq \frac{2^{k+3}}{\sqrt{\Delta t}} \ \forall\, 1 \leq k \leq K\right)\right) &\geq \sum_{k=1}^{K} \log\left(1 - \exp\left(-\frac{2^k}{\sqrt{\Delta t}}\right)\right) \\ &\geq -2\sum_{k=1}^{K} \exp\left(-\frac{2^k}{\sqrt{\Delta t}}\right). \end{aligned} \tag{3.6}$$

Next, using $2^k \geq 2k$,

$$\sum_{k=1}^{K} \exp\left(\frac{-2^k}{\sqrt{\Delta t}}\right) \leq \sum_{k=1}^{K} \exp\left(-\frac{2k}{\sqrt{\Delta t}}\right).$$

The right-hand side is a geometric series that converges monotonically from below to $e^{-2/\sqrt{\Delta t}}/(1 - e^{-2/\sqrt{\Delta t}}) \leq 2e^{-2/\sqrt{\Delta t}}$. Hence, in (3.6),

$$\log\left(\mathbb{P}\left(|X_k| \geq \frac{2^{k+3}}{\sqrt{\Delta t}} \ \forall\, 1 \leq k \leq K\right)\right) \geq -4e^{-2/\sqrt{\Delta t}},$$

and the result follows. □

To interpret Lemma 3.1, we note that given any $x(0) \neq 0$ and any $\Delta t > 0$, there is a nonzero probability that the first Brownian increment, ΔB_1, will cause $|X_1| \geq 2^4/\sqrt{\Delta t}$. Hence, there is a nonzero probability that EM will produce a numerical solution that blows up at a geometric rate. This contrasts with the initial-data-independent exponential stability of the underlying SDE, shown by (3.2).

In sections 4 and 5 we show that this poor behavior cannot happen when EM is applied to linear scalar problems or an appropriate class of nonlinear SDEs. In section 6 we study a class of SDEs that includes (3.1) and show that the correct stability can be retained by moving to an implicit method. We note that in all results, when we state the existence of a suitable upper limit, $\Delta t^\star$, on the stepsize, we implicitly mean that $\Delta t^\star$ does not depend on the initial data.

4. Linear scalar SDEs. In this section we focus on the linear scalar SDE

$$dx(t) = \alpha x(t)dt + \sigma x(t)dB(t), \qquad \text{with } 0 \neq x(0) \in \mathbb{R}, \tag{4.1}$$

where α and σ are real numbers. The following result is classical; see, e.g., [3, 13, 14].

THEOREM 4.1. *The sample Lyapunov exponent of the solution to the SDE* (4.1) *is*

$$\lim_{t\to\infty} \frac{1}{t} \log|x(t)| = \alpha - \frac{1}{2}\sigma^2 \quad \text{a.s.},$$

and the pth moment Lyapunov exponent is

$$\lim_{t\to\infty}\frac{1}{t}\log\mathbb{E}\left(|x(t)|^p\right)=p\alpha+\frac{1}{2}p(p-1)\sigma^2,$$

for any $p>0$. Hence, the zero solution of the SDE (4.1) is a.s. exponentially stable if and only if $\alpha-\frac{1}{2}\sigma^2<0$, while it is pth moment exponentially stable if and only if $\alpha+\frac{1}{2}(p-1)\sigma^2<0$.

We hence observe that the zero solution of the SDE (4.1) is a.s. exponentially stable if and only if it is pth moment exponentially stable for some sufficiently small positive p.

In the following three subsections we show that for small Δt, EM and ST recover almost sure and pth moment exponential stability of (4.1).

4.1. Almost sure exponential stability of Euler–Maruyama.

THEOREM 4.2. *If $\alpha-\frac{1}{2}\sigma^2<0$ in (4.1), then for any $\varepsilon\in(0,1)$ there is a $\Delta t^\star\in(0,1)$ such that for any $\Delta t<\Delta t^\star$, the EM approximation has the property that*

$$\lim_{k\to\infty}\frac{1}{k\Delta t}\log|X_k|\le(1-\varepsilon)\left(\alpha-\frac{1}{2}\sigma^2\right)<0\quad a.s. \tag{4.2}$$

Proof. The EM method (2.2) applied to (4.1) has the form

$$X_{k+1}=X_k(1+\alpha\Delta t+\sigma\Delta B_k). \tag{4.3}$$

It follows that $X_k=x_0\prod_{j=0}^{k-1}(1+\alpha\Delta t+\sigma\Delta B_j)$, and thus

$$\log|X_k|=\log|x_0|+\sum_{j=0}^{k-1}\log|1+\alpha\Delta t+\sigma\Delta B_j|.$$

Dividing both sides by k, letting $k\to\infty$, and then applying the classical strong law of large numbers we obtain

$$\lim_{k\to\infty}\frac{1}{k\Delta t}\log|X_k|=\frac{1}{\Delta t}\mathbb{E}\log|1+\alpha\Delta t+\sigma\sqrt{\Delta t}Z|\quad\text{a.s.,}\quad\text{where } Z\sim N(0,1). \tag{4.4}$$

Writing

$$\begin{aligned}\log|1+\alpha\Delta t+\sigma\sqrt{\Delta t}Z|&=\tfrac{1}{2}\log([1+\alpha\Delta t+\sigma\sqrt{\Delta t}Z]^2)\\&=\tfrac{1}{2}\log(1+2[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]+[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]^2),\end{aligned}$$

and recalling the fundamental inequality

$$\log(1+u)\le u-\frac{1}{2}u^2+\frac{1}{3}u^3,\qquad u\ge-1,$$

we have

$$\begin{aligned}\log|1+\alpha\Delta t+\sigma\sqrt{\Delta t}Z|\le\frac{1}{2}\Big(&2[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]+[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]^2\\&-\frac{1}{2}\big(2[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]+[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]^2\big)^2\\&+\frac{1}{3}\big(2[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]+[\alpha\Delta t+\sigma\sqrt{\Delta t}Z]^2\big)^3\Big).\end{aligned}$$

Making use of the properties $\mathbb{E}(Z^{2n}) = (2n-1)!!$ and $\mathbb{E}(Z^{2n-1}) = 0$, for $n = 1, 2, 3, \ldots$, we can compute

$$(4.5)\quad \begin{cases} \mathbb{E}[\alpha\Delta t + \sigma\sqrt{\Delta t}Z] &= \alpha\Delta t, \\ \mathbb{E}([\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^2) &= \alpha^2\Delta t^2 + \sigma^2\Delta t, \\ \mathbb{E}([\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^3) &= \alpha^3\Delta t^3 + 3\alpha\sigma^2\Delta t^2, \\ \mathbb{E}([\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^4) &= \alpha^4\Delta t^4 + 6\alpha^2\sigma^2\Delta t^3 + 3\sigma^4\Delta t^2, \\ \mathbb{E}([\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^5) &= \alpha^5\Delta t^5 + 10\alpha^3\sigma^2\Delta t^4 + 15\alpha\sigma^4\Delta t^3, \\ \mathbb{E}([\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^6) &= \alpha^6\Delta t^6 + 15\alpha^4\sigma^2\Delta t^5 + 45\alpha^2\sigma^4\Delta t^4 + 15\sigma^6\Delta t^3, \end{cases}$$

and hence obtain

$$(4.6)\qquad \mathbb{E}\log|1 + \alpha\Delta t + \sigma\sqrt{\Delta t}Z| \le (\alpha - \tfrac{1}{2}\sigma^2)\Delta t + C_1\Delta t^2,$$

where $C_1 = C_1(\alpha, \sigma) > 0$ is a constant independent of Δt. Now, choose $\Delta t^\star \in (0, 1)$ so small that $C_1\Delta t^\star \le \varepsilon(\frac{1}{2}\sigma^2 - \alpha)$. Then for any $\Delta t < \Delta t^\star$ we can substitute (4.6) into (4.4) to obtain (4.2). □

4.2. Moment exponential stability of Euler–Maruyama.

THEOREM 4.3. *Let $p \in (0, 2]$. If $\alpha + \frac{1}{2}(p-1)\sigma^2 < 0$ in (4.1), then for any $\varepsilon \in (0, 1)$ there is a $\Delta t^\star \in (0, 1)$ such that for any $\Delta t < \Delta t^\star$, the EM approximation has the property that*

$$(4.7)\qquad \lim_{k\to\infty} \frac{1}{k\Delta t}\log\mathbb{E}(|X_k|^p) \le (1-\varepsilon)p\left(\alpha + \frac{1}{2}(p-1)\sigma^2\right) < 0.$$

Proof. It follows from (4.3) that $\mathbb{E}(|X_k|^p) = |x_0|^p\prod_{j=0}^{k-1}\mathbb{E}(|1 + \alpha\Delta t + \sigma\Delta B_j|^p)$, and hence

$$\mathbb{E}(|X_k|^p) = |x_0|^p(\mathbb{E}|1 + \alpha\Delta t + \sigma\sqrt{\Delta}Z|^p)^k, \qquad \text{where } Z \sim N(0, 1).$$

This implies

$$(4.8)\qquad \lim_{k\to\infty} \frac{1}{k\Delta t}\log\mathbb{E}(|X_k|^p) = \frac{1}{\Delta t}\log\mathbb{E}(|1 + \alpha\Delta t + \sigma\sqrt{\Delta}Z|^p).$$

Writing

$$\begin{aligned} |1 + \alpha\Delta t + \sigma\sqrt{\Delta t}Z|^p &= ([1 + \alpha\Delta t + \sigma\sqrt{\Delta t}Z]^2)^{p/2} \\ &= (1 + 2[\alpha\Delta t + \sigma\sqrt{\Delta t}Z] + [\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^2)^{p/2}, \end{aligned}$$

and recalling the fundamental inequality

$$(4.9)\qquad (1+u)^{p/2} \le 1 + \frac{p}{2}u + \frac{p(p-2)}{8}u^2 + \frac{p(p-2)(p-4)}{2^3\times 3!}u^3, \qquad u \ge -1,$$

we have

$$\begin{aligned} |1 + \alpha\Delta t + \sigma\sqrt{\Delta t}Z|^p \le 1 &+ \frac{p}{2}\left(2[\alpha\Delta t + \sigma\sqrt{\Delta t}Z] + [\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^2\right) \\ &+ \frac{p(p-2)}{8}\left(2[\alpha\Delta t + \sigma\sqrt{\Delta t}Z] + [\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^2\right)^2 \\ &+ \frac{p(p-2)(p-4)}{48}\left(2[\alpha\Delta t + \sigma\sqrt{\Delta t}Z] + [\alpha\Delta t + \sigma\sqrt{\Delta t}Z]^2\right)^3. \end{aligned}$$

Making use of (4.5), we obtain

$$\mathbb{E}(|1+\alpha\Delta t+\sigma\sqrt{\Delta t}Z|^p) \le 1+p[\alpha+\tfrac{1}{2}(p-1)\sigma^2]\Delta t + C_3\Delta t^2,$$

where $C_3 = C_3(\alpha,\sigma,p) > 0$ is a constant independent of Δt. Now, choose $\Delta t^\star \in (0,1)$ so small that for all $\Delta t < \Delta t^\star$

$$C_3\Delta t \le \varepsilon p|\alpha+\tfrac{1}{2}(p-1)\sigma^2| \qquad \text{and} \qquad -1 < (1-\varepsilon)p[\alpha+\tfrac{1}{2}(p-1)\sigma^2]\Delta t < 0.$$

Then

$$\mathbb{E}(|1+\alpha\Delta t+\sigma\sqrt{\Delta t}Z|^p) \le 1+(1-\varepsilon)p[\alpha+\tfrac{1}{2}(p-1)\sigma^2]\Delta t.$$

Substituting this into (4.8) we obtain

$$\lim_{k\to\infty}\frac{1}{k\Delta t}\log\mathbb{E}(|X_k|^p) \le \frac{1}{\Delta t}\log\left(1+(1-\varepsilon)p\left[\alpha+\frac{1}{2}(p-1)\sigma^2\right]\Delta t\right).$$

But $\log(1+u) \le u$ for $-1 < u < 0$, and thus (4.7) follows. □

4.3. Exponential stability of the stochastic theta method. If we assume that Δt is chosen so small that $\Delta t\alpha\theta < 1$, then the ST method (2.3) applied to the linear SDE (4.1) may be written in the form

$$X_{k+1} = X_k\left(1+\frac{\alpha}{1-\theta\alpha\Delta t}\Delta t+\frac{\sigma}{1-\theta\alpha\Delta t}\Delta B_k\right).$$

This approximation coincides with the EM method applied to the modified linear SDE

$$dy(t) = \frac{\alpha}{1-\theta\alpha\Delta t}y(t)dt+\frac{\sigma}{1-\theta\alpha\Delta t}y(t)dB(t).$$

Using this observation, it follows almost immediately that the statements of Theorems 4.2 and 4.3 also apply to the ST method.

5. Generalization to multidimensional nonlinear SDEs. To analyze the n-dimensional nonlinear SDE (2.1), we begin by imposing the linear growth assumption

$$|f(x)| \vee |g(x)| \le K|x| \qquad \forall x \in \mathbb{R}^n. \tag{5.1}$$

This implies

$$f(0) = 0, \quad g(0) = 0, \tag{5.2}$$

and we will be concerned with pathwise convergence of the solution $x(t)$ of (2.1) to the zero solution, as $t \to \infty$, and the preservation of this property under discretization. We also note that condition (5.1) ensures that, with probability one, the solution will never reach the origin; see, for example, [15, Lemma 3.2].

We begin by giving sufficient conditions for almost sure exponential stability of the SDE.

THEOREM 5.1. *Let* (5.1) *hold. If*

$$-\lambda := \sup_{x\in\mathbb{R}^n, x\neq 0}\left(\frac{\langle x, f(x)\rangle+\frac{1}{2}|g(x)|^2}{|x|^2}-\frac{\langle x, g(x)\rangle^2}{|x|^4}\right) < 0, \tag{5.3}$$

then the solution of (2.1) *obeys*

$$\limsup_{t\to\infty} \frac{1}{t}\log|x(t)| \le -\lambda \quad \text{a.s.}, \tag{5.4}$$

and given any $\varepsilon \in (0,\lambda)$ *there exists a* $p^\star \in (0,1)$ *such that for all* $0 < p < p^\star$

$$\limsup_{t\to\infty} \frac{1}{t}\log \mathbb{E}(|x(t)|^p) \le -p(\lambda-\varepsilon). \tag{5.5}$$

Proof. See Appendix A. □

Next, we analyze the EM discretization (2.2).

THEOREM 5.2. *Let* (5.1) *and* (5.3) *hold. Then for any* $\varepsilon \in (0,\lambda)$ *there is a constant* $\Delta t^\star \in (0,1)$ *such that for any* $0 < \Delta t < \Delta t^\star$ *the EM approximation* (2.2) *satisfies*

$$\limsup_{k\to\infty} \frac{1}{k\Delta t}\log|X_k| \le -(\lambda-\varepsilon) \quad \text{a.s.} \tag{5.6}$$

Further, for any $\varepsilon \in (0,\lambda)$ *and any sufficiently small* $p > 0$, *there is a constant* $\Delta t^\star \in (0,1)$ *such that for any* $0 < \Delta t < \Delta t^\star$ *the EM approximation* (2.2) *satisfies*

$$\limsup_{k\to\infty} \frac{1}{k\Delta t}\log \mathbb{E}(|X_k|^p) \le -p(\lambda-\varepsilon). \tag{5.7}$$

Proof. By condition (5.2), we compute from (2.2) that

$$\begin{aligned}|X_{k+1}|^2 &= |X_k|^2 + 2\langle X_k, f(X_k)\Delta t + g(X_k)\Delta B_k\rangle + |f(X_k)\Delta t + g(X_k)\Delta B_k|^2\\ &= |X_k|^2(1+\xi_k),\end{aligned}$$

where

$$\xi_k = \frac{1}{|X_k|^2}\left[2\langle X_k, f(X_k)\Delta t + g(X_k)\Delta B_k\rangle + |f(X_k)\Delta t + g(X_k)\Delta B_k|^2\right]$$

if $X_k \neq 0$, otherwise it is set to -1. Clearly, $\xi_k \ge -1$. For any $p \in (0,1)$, by inequality (4.9) we have

$$\begin{aligned}|X_{k+1}|^p &= |X_k|^p(1+\xi_k)^{p/2}\\ &\le |X_k|^p\left(1 + \frac{p}{2}\xi_k + \frac{p(p-2)}{8}\xi_k^2 + \frac{p(p-2)(p-4)}{2^3\times 3!}\xi_k^3\right).\end{aligned}$$

Hence the conditional expectation

$$\begin{aligned}\mathbb{E}(|X_{k+1}|^p\Big|\mathcal{F}_{k\Delta t}) &\le |X_k|^p\ \mathbb{E}\left(1 + \frac{p}{2}\xi_k + \frac{p(p-2)}{8}\xi_k^2 + \frac{p(p-2)(p-4)}{2^3\times 3!}\xi_k^3\Big|\mathcal{F}_{k\Delta t}\right)\\ &= |X_k|^p\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}\left(1 + \frac{p}{2}\xi_k + \frac{p(p-2)}{8}\xi_k^2 + \frac{p(p-2)(p-4)}{2^3\times 3!}\xi_k^3\Big|\mathcal{F}_{k\Delta t}\right),\end{aligned} \tag{5.8}$$

where $\mathbf{1}_A$ denotes the indicator function for A. Now,

$$\begin{aligned}&\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\xi_k|\mathcal{F}_{k\Delta t})\\ &= \mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}\left(\frac{1}{|X_k|^2}\left[2\langle X_k, f(X_k)\Delta t + g(X_k)\Delta B_k\rangle + |f(X_k)\Delta t + g(X_k)\Delta B_k|^2\right]\Big|\mathcal{F}_{k\Delta t}\right).\end{aligned}$$

Since ΔB_k is independent of $\mathcal{F}_{k\Delta t}$, we have $\mathbb{E}(\Delta B_k|\mathcal{F}_{k\Delta t}) = \mathbb{E}(\Delta B_k) = 0$ and $\mathbb{E}((\Delta B_k)^2|\mathcal{F}_{k\Delta t}) = \mathbb{E}((\Delta B_k)^2) = \Delta t$. It is then easy to obtain that

$$\begin{aligned}\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\xi_k|\mathcal{F}_{k\Delta t}) &= \mathbf{1}_{\{X_k\neq 0\}}\left(\frac{1}{|X_k|^2}\left[2\langle X_k, f(X_k)\Delta t\rangle + |f(X_k)|^2\Delta t^2 + |g(X_k)|^2\Delta t\right]\right)\\ &\le \mathbf{1}_{\{X_k\neq 0\}}\left(\frac{1}{|X_k|^2}\left[2\langle X_k, f(X_k)\Delta t\rangle + |g(X_k)|^2\right]\Delta t + K^2\Delta t^2\right), \qquad (5.9)\end{aligned}$$

where (5.1) has been used. Similarly, we can show that

$$\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\xi_k^2|\mathcal{F}_{k\Delta t}) \ge \frac{4}{|X_k|^4}\langle X_k, g(X_k)\rangle^2\Delta t - c_K\Delta t^2 \qquad (5.10)$$

and

$$\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\xi_k^3|\mathcal{F}_{k\Delta t}) \le c_K\Delta t^2, \qquad (5.11)$$

where $c_K > 0$ is a constant dependent only on K. Substituting (5.9), (5.10), and (5.11) into (5.8) and then using (5.3) and (5.1) we derive that

$$\begin{aligned}\mathbb{E}(|X_{k+1}|^p\big|\mathcal{F}_{k\Delta t}) &\le |X_k|^p\mathbf{1}_{\{X_k\neq 0\}}\left(1 + \frac{p}{2|X_k|^2}\left[2\langle X_k, f(X_k)\Delta t\rangle + |g(X_k)|^2\right]\Delta t\right.\\ &\qquad \left. + \frac{p(p-2)}{2|X_k|^4}\langle X_k, g(X_k)\rangle^2\Delta t + C\Delta t^2\right)\\ &= |X_k|^p\mathbf{1}_{\{X_k\neq 0\}}\left\{1 + p\Delta t\left(\frac{\langle X_k, f(X_k)\rangle + \frac{1}{2}|g(X_k)|^2}{|X_k|^2} - \frac{\langle X_k, g(X_k)\rangle^2}{|X_k|^4}\right)\right.\\ &\qquad \left. + \frac{p^2\Delta t\langle X_k, g(X_k)\rangle^2}{2|X_k|^4} + C\Delta t^2\right\}\\ &\le |X_k|^p\left(1 - p\lambda\Delta t + \frac{p^2\Delta t K^2}{2} + C\Delta t^2\right), \qquad (5.12)\end{aligned}$$

where $C = C(K,p) > 0$ is a constant independent of Δt. Now, for any given $\varepsilon \in (0,\lambda)$ and $p \in (0,1)$ sufficiently small for $pK^2 < \varepsilon$, choose $\Delta t^\star \in (0,1)$ sufficiently small for $p\lambda\Delta t^\star < 1$ and $C\Delta t^\star < \frac{1}{2}p\varepsilon$. It then follows from (5.12) that for any $\Delta t < \Delta t^\star$

$$\mathbb{E}(|X_{k+1}|^p\big|\mathcal{F}_{k\Delta t}) \le |X_k|^p(1 - p(\lambda-\varepsilon)\Delta t).$$

Taking expectations on both sides yields

$$\mathbb{E}(|X_{k+1}|^p) \le \mathbb{E}(|X_k|^p)(1 - p(\lambda-\varepsilon)\Delta t).$$

Since this holds for any $k \ge 0$, we have

$$\mathbb{E}(|X_k|^p) \le |x(0)|^p(1 - p(\lambda-\varepsilon)\Delta t)^k \le |x(0)|^p e^{-pk(\lambda-\varepsilon)\Delta t} \quad \forall k \ge 1. \qquad (5.13)$$

This implies (5.7). Moreover, we have

$$\mathbb{P}\{|X_k|^p > k^2 e^{-pk(\lambda-\varepsilon)\Delta t}\} \le \frac{|x(0)|^p}{k^2} \quad \forall k \ge 1.$$

By the Borel–Cantelli lemma, we see that for almost all $\omega \in \Omega$

$$|X_k|^p \le k^2 e^{-pk(\lambda-\varepsilon)\Delta t} \qquad (5.14)$$

holds for all but finitely many k. Hence, there exists a $k_0(\omega)$, for all $\omega \in \Omega$ excluding a $\mathbb{P}$-null set, for which (5.14) holds whenever $k \geq k_0$. Consequently, for almost all $\omega \in \Omega$,

$$\frac{1}{k\Delta t} \log |X_k| \leq -(\lambda - \varepsilon) + \frac{2\log(k)}{pk\Delta t}$$

whenever $k \geq k_0$. Letting $k \to \infty$ we obtain (5.6). □

Let us now apply Theorem 5.2 to the linear SDE system

$$dx(t) = Ax(t)dt + Gx(t)dB(t), \quad t \geq 0, \qquad \text{given } 0 \neq x(0) \in \mathbb{R}^n, \tag{5.15}$$

where $A, G \in \mathbb{R}^{n\times n}$. This corresponds to $f(x) = Ax$ and $g(x) = Gx$ in (2.1). Note

$$\tfrac{1}{2}\lambda_{\min}(A + A^T)|x|^2 \leq \langle x, Ax\rangle = \tfrac{1}{2}\langle x, (A + A^T)x\rangle \leq \tfrac{1}{2}\lambda_{\max}(A + A^T)|x|^2$$

and

$$\lambda_{\min}(G^T G)|x|^2 \leq \langle x, G^T Gx\rangle = |Gx|^2 \leq \|G\|^2|x|^2,$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ denote the maximum and minimum eigenvalues of a symmetric matrix, respectively. Moreover,

$$0 \leq \langle x, Gx\rangle^2 = \tfrac{1}{4}\langle x, (G + G^T)x\rangle^2 \leq \tfrac{1}{4}\lambda^2_{\max}(G + G^T)|x|^2,$$

while if $G + G^T$ is either nonpositive definite or nonnegative definite,

$$\langle x, Gx\rangle^2 \geq \tfrac{1}{4}\left[|\lambda_{\max}(G + G^T)| \wedge |\lambda_{\min}(G + G^T)|\right]^2 |x|^4.$$

We hence observe that

$$\frac{\langle x, Ax\rangle + \frac{1}{2}|Gx|^2}{|x|^2} - \frac{\langle x, Gx\rangle^2}{|x|^4} \geq \tfrac{1}{2}\lambda_{\min}(A + A^T) + \tfrac{1}{2}\lambda_{\min}(G^T G) - \tfrac{1}{4}\lambda^2_{\max}(G + G^T),$$

while if $G + G^T$ is either nonpositive definite or nonnegative definite,

$$\frac{\langle x, Ax\rangle + \frac{1}{2}|Gx|^2}{|x|^2} - \frac{\langle x, Gx\rangle^2}{|x|^4} \leq \tfrac{1}{2}\lambda_{\max}(A + A^T) + \tfrac{1}{2}\|G\|^2 - \tfrac{1}{4}\left[|\lambda_{\max}(G + G^T)| \wedge |\lambda_{\min}(G + G^T)|\right]^2.$$

By Theorem 5.2 we reach the following conclusion.

COROLLARY 5.3. *If $G + G^T$ is either nonpositive definite or nonnegative definite and*

$$-\lambda := \tfrac{1}{2}\lambda_{\max}(A + A^T) + \tfrac{1}{2}\|G\|^2 - \tfrac{1}{4}\left[|\lambda_{\max}(G + G^T)| \wedge |\lambda_{\min}(G + G^T)|\right]^2 < 0,$$

then for any $\varepsilon \in (0, \lambda)$ there is a pair of constants $p \in (0, 1)$ and $\Delta t^\star \in (0, 1)$ such that for any $\Delta t < \Delta t^\star$ the EM approximation of the linear SDE (5.15) *has the properties* (5.7) *and* (5.6).

6. Backward Euler. So far, we have proved positive results about EM for sufficiently small Δt. However, we saw in section 3 that this behavior does not extend to the cubic example (3.1). This SDE does not satisfy the linear growth condition (5.1); thus, of course, the theorems in section 5 do not apply. However, (3.1) does satisfy (5.3), since

$$\sup_{x\in\mathbb{R},x\neq 0}\left(\frac{\langle x,f(x)\rangle+\frac{1}{2}|g(x)|^2}{|x|^2}-\frac{\langle x,g(x)\rangle^2}{|x|^4}\right)=\sup_{x\in\mathbb{R},x\neq 0}\left(\frac{x^2-x^4+2x^2}{x^2}-\frac{4x^4}{x^4}\right)\leq -1,$$

and we note that the proof of Theorem 5.1 did not use the condition $|f(x)|\leq K|x|$ explicitly, though $|g(x)|\leq K|x|$ was used. Of course, the linear growth condition (5.1) was used implicitly to guarantee that the solution stays away from the origin with probability one. However, for this property we need only a weaker condition (see [15, Lemma 3.2 on p. 120]). Let us form this improved result as a new theorem.

THEOREM 6.1. *The conclusions of Theorem* 5.1 *still hold if condition* (5.1) *is replaced by the following: for each integer* $i\geq 1$ *there is a* $K_i>0$ *such that*

$$|f(x)|\leq K_i|x| \quad \forall x\in\mathbb{R}^n \text{ with } |x|\leq i, \tag{6.1}$$

while there is a $K>0$ *such that*

$$|g(x)|\leq K|x| \quad \forall x\in\mathbb{R}^n. \tag{6.2}$$

An application of this theorem to the SDE (3.1) shows that its solution obeys (3.2), as claimed in section 3. We also saw from Lemma 3.1 that EM does not preserve this almost sure asymptotic stability for any $\Delta t>0$. Hence, it is not possible to extend Theorem 5.2 to the case where (5.1) is replaced by (6.1) and (6.2).

An interesting open question is whether any other numerical methods preserve exponential asymptotic stability for small Δt under (6.1) and (6.2).

In this section we pursue a different approach. We consider a structural constraint that is known to allow positive results to be proved for the BE method in other contexts. More precisely, we assume that there is a constant $\mu\in\mathbb{R}$ such that

$$\langle x-y,f(x)-f(y)\rangle\leq\mu|x-y|^2 \quad \forall x,y\in\mathbb{R}^n. \tag{6.3}$$

This *one-sided Lipschitz condition* has been applied in the deterministic and stochastic literature [9, 11, 12, 16, 20] to establish results about long-term behavior and boundedness in a manner that is connected with the use of Lyapunov functions [6, 18]. In particular, we note that under (6.3) the condition $\mu\Delta t<1$ ensures that (2.3) with $\theta=1$ can be solved uniquely for X_{k+1}.

The next theorem concerns the exponential stability of BE under conditions (6.3) and (6.2). Although (6.2) implies $g(0)=0$, (6.3) may not force $f(0)=0$, and thus we still need to assume it for the purpose of stability analysis.

THEOREM 6.2. *Let* (6.2) *and* (6.3) *hold and* $f(0)=0$. *Assume also that* $\mu+\frac{1}{2}\rho<0$, *where*

$$\rho:=\sup_{x\in\mathbb{R}^n,x\neq 0}\Big(\frac{|g(x)|^2}{|x|^2}-\frac{2\langle x,g(x)\rangle^2}{|x|^4}\Big). \tag{6.4}$$

Then (5.4) *holds with* $-\lambda=\mu+\frac{1}{2}\rho$, *and for any* $\varepsilon\in(0,|\mu+\frac{1}{2}\rho|)$ *there is a pair of constants* $p\in(0,1)$ *and* $\Delta t^\star\in(0,1)$ *with* $\mu\Delta t^\star<1$ *such that for any* $\Delta t<\Delta t^\star$, *the*

BE method (that is, (2.3) *with* $\theta = 1$*) has the properties that*

$$\limsup_{k\to\infty} \frac{1}{k\Delta t} \log \mathbb{E}(|X_k|^p) \le p\left(\mu + \frac{1}{2}\rho + \varepsilon\right) < 0 \tag{6.5}$$

and

$$\limsup_{k\to\infty} \frac{1}{k\Delta t} \log |X_k| \le \mu + \frac{1}{2}\rho + \varepsilon < 0 \quad \text{a.s.} \tag{6.6}$$

Proof. It is straightforward to adapt the proof of Theorem 5.1 in order to establish (5.4) under (6.2) and (6.3). From (2.3) with $\theta = 1$, we have

$$|X_{k+1}|^2 = \langle X_{k+1}, X_k + g(X_k)\Delta B_k\rangle + \langle X_{k+1}, f(X_{k+1})\Delta t\rangle.$$

By (6.3) and $f(0) = 0$, we have

$$\langle X_{k+1}, f(X_{k+1})\Delta t\rangle \le \mu\Delta t|X_{k+1}|^2.$$

But,

$$\langle X_{k+1}, X_k + g(X_k)\Delta B_k\rangle \le \tfrac{1}{2}|X_{k+1}|^2 + \tfrac{1}{2}|X_k + g(X_k)\Delta B_k|^2.$$

We hence obtain

$$\begin{aligned}
|X_{k+1}|^2 &\le \frac{1}{1-2\mu\Delta t}|X_k + g(X_k)\Delta B_k|^2 \\
&\le \frac{1}{1-2\mu\Delta t}\left(|X_k|^2 + 2\langle X_k, g(X_k)\rangle\Delta B_k + |g(X_k)|^2\Delta B_k^2\right) \\
&= \frac{|X_k|^2}{1-2\mu\Delta t}(1+\zeta_k),
\end{aligned}$$

where

$$\zeta_k = \frac{1}{|X_k|^2}\left(2\langle X_k, g(X_k)\rangle\Delta B_k + |g(X_k)|^2\Delta B_k^2\right)$$

if $X_k \neq 0$, otherwise it is set to -1. Clearly, $\zeta_k \ge -1$. For any $p \in (0,1)$, by inequality (4.9) we can then show that

$$\begin{aligned}
\mathbb{E}(|X_{k+1}|^p\big|\mathcal{F}_{k\Delta t}) \le \frac{|X_k|^p}{(1-2\mu\Delta t)^{p/2}}\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}\Bigg(1 + \frac{p}{2}\zeta_k \\
+ \frac{p(p-2)}{8}\zeta_k^2 + \frac{p(p-2)(p-4)}{2^3\times 3!}\zeta_k^3\Big|\mathcal{F}_{k\Delta t}\Bigg).
\end{aligned} \tag{6.7}$$

In the same way as in the proof of Theorem 5.2 we can show that

$$\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\zeta_k|\mathcal{F}_{k\Delta t}) = \mathbf{1}_{\{X_k\neq 0\}}\frac{|g(X_k)|^2}{|X_k|^2}\Delta t,$$

$$\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\zeta_k^2|\mathcal{F}_{k\Delta t}) \ge \frac{4\langle X_k, g(X_k)\rangle^2}{|X_k|^4}\Delta t - K^4\Delta t^2,$$

and

$$\mathbf{1}_{\{X_k\neq 0\}}\mathbb{E}(\zeta_k^3|\mathcal{F}_{k\Delta t}) \le c_K\Delta t^2,$$

where $c_K > 0$ is a constant dependent only on K. Substituting the three inequalities above into (6.7) and then using (6.4) and (6.2) we derive that

$$\begin{aligned}\mathbb{E}(|X_{k+1}|^p\big|\mathcal{F}_{k\Delta t}) &\le \frac{|X_k|^p}{(1-2\mu\Delta t)^{p/2}}\mathbf{1}_{\{X_k\neq 0\}}\left(1+\frac{p}{2}\frac{|g(X_k)|^2}{|X_k|^2}\Delta t\right.\\ &\quad\left.+\frac{p(p-2)}{8}\left[\frac{4\langle X_k,g(X_k)\rangle^2}{|X_k|^4}\Delta t - K^4\Delta t^2\right]+\frac{p(p-2)(p-4)}{2^3\times 3!}c_K\Delta t^2\right)\\ &\le \frac{|X_k|^p}{(1-2\mu\Delta t)^{p/2}}\left(1+\frac{1}{2}p\rho\Delta t+\frac{1}{2}p^2K^2\Delta t+C\Delta t^2\right),\end{aligned}$$

where $C = C(p,K)$ is a positive constant. Taking expectations on both sides, we arrive at

$$\mathbb{E}(|X_{k+1}|^p) \le \frac{1+\frac{1}{2}p\rho\Delta t+\frac{1}{2}p^2K^2\Delta t+C\Delta t^2}{(1-2\mu\Delta t)^{p/2}}\,\mathbb{E}(|X_k|^p). \tag{6.8}$$

Now, for any $\varepsilon\in(0,|\mu+\frac{1}{2}\rho|)$, we may choose p sufficiently small for $pK^2\le\frac{1}{4}\varepsilon$. Then we have

$$(1-2\mu\Delta t)^{p/2} \ge 1-p\mu\Delta t-\widehat{C}\Delta t^2 > 0, \tag{6.9}$$

for sufficiently small Δt, where $\widehat{C}=\widehat{C}(p,\mu)$ is a positive constant. By further reducing Δt, if necessary, we may ensure that

$$C\Delta t<\tfrac{1}{8}p\varepsilon,\qquad \widehat{C}\Delta t<\tfrac{1}{4}\varepsilon,\qquad |p(\mu+\tfrac{1}{4}\varepsilon)\Delta t|\le\tfrac{1}{2}. \tag{6.10}$$

Using (6.9) and (6.10) in (6.8) gives

$$\mathbb{E}(|X_{k+1}|^p) \le \frac{1+\frac{1}{2}p(\rho+\frac{1}{2}\varepsilon)\Delta t}{1-p(\mu+\frac{1}{4}\varepsilon)\Delta t}\,\mathbb{E}|X_k|^p. \tag{6.11}$$

Note that for any $u\in[-\frac{1}{2},\frac{1}{2}]$

$$\frac{1}{1-u} = 1+u+u^2\sum_{i=0}^{\infty}u^i \le 1+u+u^2\sum_{i=0}^{\infty}(\tfrac{1}{2})^i = 1+u+2u^2.$$

By further reducing Δt, if necessary, so that

$$4p(\mu+\tfrac{1}{4}\varepsilon)^2\Delta t+(\rho+\tfrac{1}{2}\varepsilon)\left(p(\mu+\tfrac{1}{4}\varepsilon)\Delta t+2[p(\mu+\tfrac{1}{4}\varepsilon)\Delta t]^2\right)\le\varepsilon,$$

and using (6.11), we compute that

$$\begin{aligned}\mathbb{E}(|X_{k+1}|^p) &\le \left(1+\tfrac{1}{2}p(\rho+\tfrac{1}{2}\varepsilon)\Delta t\right)\left(1+p(\mu+\tfrac{1}{4}\varepsilon)\Delta t+2[p(\mu+\tfrac{1}{4}\varepsilon)\Delta t]^2\right)\mathbb{E}(|X_k|^p)\\ &\le [1+p(\mu+\tfrac{1}{2}\rho+\varepsilon)\Delta t]\mathbb{E}(|X_k|^p).\end{aligned}$$

From this we can show the assertions (6.5) and (6.6) in the same way as in the proof of Theorem 5.2. □

Let us return to the scalar SDE (3.1), where $f(x) = x - x^3$ and $g(x) = 2x$. In this case, we have $\langle x-y,\, f(x)-f(y)\rangle \le |x-y|^2$, so we may take $\mu = 1$ in (6.3), while

$$\rho := \sup_{x\in\mathbb{R}, x\neq 0} \left(\frac{|g(x)|^2}{|x|^2} - \frac{2\langle x, g(x)\rangle^2}{|x|^4} \right) = -4,$$

whence $\mu + \frac{1}{2}\rho = -1$, which gives another confirmation of (3.2).

By Theorem 6.2, for any given $\varepsilon \in (0,1)$, there is a $\Delta t^\star > 0$ sufficiently small so that if $\Delta t < \Delta t^\star$, the BE approximate solution of the SDE (3.1) obeys

$$\limsup_{k\to\infty} \frac{1}{k\Delta t} \log |X_k| \le -1 + \varepsilon \qquad \text{a.s.},$$

which recovers property (3.2) very well indeed.

It is also interesting to observe that in the scalar case (that is, $n = 1$),

$$\rho = \sup_{x\in\mathbb{R}, x\neq 0} \left(-\frac{|g(x)|^2}{|x|^2} \right) \le 0.$$

In this case, if (6.3) also holds with $\mu < 0$, then the BE method is a.s. exponentially stable as long as the stepsize is sufficiently small. For example, the BE approximate solution to the scalar SDE

$$dx(t) = (\mu x - x^3)dt + g(x)dB(t)$$

is always a.s. exponentially stable as long as the stepsize is sufficiently small, $\mu < 0$, and g obeys the linear growth condition (6.2). However, in the case $\mu \ge 0$, we will need that

$$|g(x)|^2 \ge \rho x^2, \quad x \in \mathbb{R},$$

holds for some $\rho > 2\mu$ in order to conclude that the BE method is a.s. exponentially stable.

7. Multidimensional noise. So far, in order to streamline the presentation, we have only considered scalar noise. In this section we state, without proof, how the nonlinear results generalize to the multinoise case, as follows:

$$dx(t) = f(x(t))dt + \sum_{j=1}^{d} g_j(x(t))dB_j(t), \quad t \ge 0, \qquad \text{given } 0 \neq x(0) \in \mathbb{R}^n. \tag{7.1}$$

Here $(B_1(t), \ldots, B_d(t))$ is a d-dimensional Brownian motion. As before, we assume, as a standing hypothesis, that $f, g_1, \ldots, g_d : \mathbb{R}^n \to \mathbb{R}^n$ are smooth enough for the SDE (7.1) to have a unique global solution $x(t)$ on $[0,\infty)$.

The following generalization of Theorem 6.1 gives a criterion for the almost sure and moment exponential stability of the SDE.

THEOREM 7.1. *Assume that for each integer $i \ge 1$ there is a $K_i > 0$ such that*

$$|f(x)| \le K_i|x| \quad \forall x \in \mathbb{R}^n \text{ with } |x| \le i, \tag{7.2}$$

while there is a $K > 0$ such that

$$|g_j(x)| \le K|x| \quad \forall x \in \mathbb{R}^n \text{ and } 1 \le j \le d. \tag{7.3}$$

If

$$-\lambda := \sup_{x\in\mathbb{R}^n, x\neq 0}\left(\frac{\langle x, f(x)\rangle + \frac{1}{2}\sum_{j=1}^d |g_j(x)|^2}{|x|^2} - \frac{\sum_{j=1}^d \langle x, g_j(x)\rangle^2}{|x|^4}\right) < 0,$$

then the solution of (7.1) *obeys*

$$\limsup_{t\to\infty} \frac{1}{t}\log|x(t)| \leq -\lambda \quad a.s., \tag{7.4}$$

and given any $\varepsilon \in (0,\lambda)$ *there exists a* $p^\star \in (0,1)$ *such that for all* $0 < p < p^\star$

$$\limsup_{t\to\infty} \frac{1}{t}\log \mathbb{E}(|x(t)|^p) \leq -p(\lambda - \varepsilon).$$

This theorem can be proved in a similar way that Theorem 5.1 is proved in the appendix.

The EM method applied to (7.1) produces approximations $X_k \approx x(k\Delta t)$ with $X_0 = x(0)$ and

$$X_{k+1} = X_k + \Delta t f(X_k) + \sum_{j=1}^d g(X_k)\Delta B_{jk}, \tag{7.5}$$

where $\Delta B_{jk} := B_j((k+1)\Delta t) - B_j(k\Delta t)$. Recalling the motivating example in section 3, we will replace the local linear growth condition (7.2) by a global one.

THEOREM 7.2. *Assume that all the conditions of Theorem* 7.1 *hold with condition* (7.2) *replaced by*

$$|f(x)| \leq K|x| \quad \forall x \in \mathbb{R}^n.$$

Then for any $\varepsilon \in (0,\lambda)$ *there is a constant* $\Delta t^\star \in (0,1)$ *such that for any* $0 < \Delta t < \Delta t^\star$ *the EM approximation* (7.5) *satisfies*

$$\limsup_{k\to\infty} \frac{1}{k\Delta t}\log|X_k| \leq -(\lambda - \varepsilon) \quad a.s.$$

Further, for any $\varepsilon \in (0,\lambda)$ *and any sufficiently small* $p > 0$, *there is a constant* $\Delta t^\star \in (0,1)$ *such that for any* $0 < \Delta t < \Delta t^\star$ *the EM approximation* (7.5) *satisfies*

$$\limsup_{k\to\infty} \frac{1}{k\Delta t}\log \mathbb{E}(|X_k|^p) \leq -p(\lambda - \varepsilon).$$

The BE method applied to (7.1) produces approximations $X_k \approx x(k\Delta t)$ with $X_0 = x(0)$ and

$$X_{k+1} = X_k + \Delta t f(X_{k+1}) + \sum_{j=1}^d g_j(X_k)\Delta B_{jk}. \tag{7.6}$$

THEOREM 7.3. *Let* (7.3) *and* (6.3) *hold and* $f(0) = 0$. *Assume also that* $\mu + \frac{1}{2}\rho < 0$, *where*

$$\rho := \sup_{x\in\mathbb{R}^n, x\neq 0}\left(\frac{\sum_{j=1}^d |g_j(x)|^2}{|x|^2} - \frac{2\sum_{j=1}^d \langle x, g_j(x)\rangle^2}{|x|^4}\right).$$

Then (7.4) *holds with* $-\lambda = \mu + \frac{1}{2}\rho$, *and for any* $\varepsilon \in (0, |\mu + \frac{1}{2}\rho|)$ *there is a pair of constants* $p \in (0,1)$ *and* $\Delta t^\star \in (0,1)$ *with* $\mu \Delta t^\star < 1$ *such that for any* $\Delta t < \Delta t^\star$, *the BE method* (7.6) *has the properties that*

$$\limsup_{k\to\infty} \frac{1}{k\Delta t} \log \mathbb{E}(|X_k|^p) \le p\left(\mu + \frac{1}{2}\rho + \varepsilon\right) < 0$$

and

$$\limsup_{k\to\infty} \frac{1}{k\Delta t} \log |X_k| \le \mu + \frac{1}{2}\rho + \varepsilon < 0 \quad \text{a.s.}$$

Theorems 7.2 and 7.3 can be proved in the same way as the scalar noise versions, Theorems 5.2 and 6.2.

Appendix A. Proof of Theorem 5.1.

Proof. The result (5.4) may be proved by generalizing the analysis in [15, pp. 121–123], so we give only an outline. By the Itô formula, we can show that

$$\begin{aligned}\log(|x(t)|^2) &= \log(|x(0)|^2) + M(t) \\ &\quad + \int_0^t 2\left(\frac{\langle x(s), f(x(s))\rangle + \frac{1}{2}|g(x(s))|^2}{|x(s)|^2} - \frac{\langle x(s), g(x(s))\rangle^2}{|x(s)|^4}\right) ds,\end{aligned}$$

where

$$M(t) = \int_0^t \frac{2\langle x(s), g(x(s))\rangle}{|x(s)|^2} dB(s).$$

From the condition $|g(x)| \le K|x|$, it is straightforward to show that

$$\lim_{t\to\infty} \frac{M(t)}{t} = 0 \quad \text{a.s.}$$

Now, if (5.3) holds, then

$$\log(|x(t)|^2) \le \log(|x_0|^2) + M(t) - 2\lambda t.$$

Dividing both sides by $2t$ and then letting $t \to \infty$ we obtain (5.4).

Now we show (5.5). For $0 < p < 1$ we have, from the Itô formula,

$$\begin{aligned} d\left(|x(t)|^p\right) &= d\left(\left(|x(t)|^2\right)^{\frac{1}{2}p}\right) \\ &= \frac{p}{2}\left(|x(t)|^2\right)^{\frac{1}{2}p-1} dx(t) \\ &\quad + \frac{1}{2}\frac{p}{2}\left(\frac{p}{2} - 1\right)\left(|x(t)|^2\right)^{\frac{1}{2}p-2} 4\langle x(t), g(x(t))\rangle^2 \, dt \\ &= p|x(t)|^p \left[\frac{\langle x(t), f(x(t))\rangle + \frac{1}{2}|g(x(t))|^2}{|x(t)|^2} - \frac{\langle x(t), g(x(t))\rangle^2}{|x(t)|^4}\right. \\ &\quad \left. + \frac{p}{2}\frac{\langle x(t), g(x(t))\rangle^2}{|x(t)|^4}\right] dt \\ &\quad + p|x(t)|^p \frac{\langle x(t), g(x(t))\rangle}{|x(t)|^2} \, dB(t). \end{aligned} \tag{A.1}$$

Under (5.1) and (5.3) this implies

$$d\left(|x(t)|^p\right) \le p|x(t)|^p \left(-\lambda + \frac{p}{2}K^2\right) dt + p|x(t)|^p \frac{\langle x(t), g(x(t))\rangle}{|x(t)|^2}\, dB(t).$$

Given $\varepsilon \in (0, \lambda)$ we may choose $p \in (0, 1)$ so small that $pK^2/2 < \varepsilon$, whence

$$\begin{aligned} d\left(e^{(\lambda-\varepsilon)pt}|x(t)|^p\right) &\le e^{(\lambda-\varepsilon)pt}|x(t)|^p \left[(\lambda-\varepsilon)p + p\left(-\lambda + \frac{1}{2}pK^2\right)\right] dt \\ &\quad + e^{(\lambda-\varepsilon)pt} p|x(t)|^p \frac{\langle x(t), g(x(t))\rangle}{|x(t)|^2}\, dB(t) \\ &\le e^{(\lambda-\varepsilon)pt} p|x(t)|^p \frac{\langle x(t), g(x(t))\rangle}{|x(t)|^2}\, dB(t). \end{aligned}$$

We deduce that

$$e^{(\lambda-\varepsilon)pt}\mathbb{E}|x(t)|^p \le \mathbb{E}|x(0)|^p,$$

and (5.5) follows. □

REFERENCES

[1] J. A. D. Appleby, G. Berkolaiko, and A. Rodkina, *On the Asymptotic Behavior of the Moments of Solutions of Stochastic Difference Equations*, Technical report MS-05-21, School of Mathematical Sciences, Dublin City University, Dublin, Ireland, 2005.

[2] J. A. D. Appleby, D. Mackey, and A. Rodkina, *Almost Sure Polynomial Asymptotic Stability of Stochastic Difference Equations*, Technical report MS-05-20, School of Mathematical Sciences, Dublin City University, Dublin, Ireland, 2005.

[3] L. Arnold, *Stochastic Differential Equations: Theory and Applications*, Wiley, Chichester, UK, 1972.

[4] C. T. H. Baker and E. Buckwar, *Exponential stability in p-th mean of solutions, and of convergent Euler-type solutions, of stochastic delay differential equations*, J. Comput. Appl. Math., 184 (2005), pp. 404–427.

[5] A. Bryden and D. J. Higham, *On the boundedness of asymptotic stability regions for the stochastic theta method*, BIT, 43 (2003), pp. 1–6.

[6] E. Buckwar, R. Horvath-Bokor, and R. Winkler, *Asymptotic mean-square stability of two-step methods for stochastic ordinary differential equations*, BIT, 46 (2006), pp. 261–282.

[7] K. Burrage and T. Tian, *A note on the stability properties of the Euler methods for solving stochastic differential equations*, New Zealand J. Math., 29 (2000), pp. 115–127.

[8] P. M. Burrage, *Runge–Kutta Methods for Stochastic Differential Equations*, Ph.D. thesis, University of Queensland, Brisbane, Australia, 1999.

[9] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II, *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, Berlin, 1996.

[10] D. J. Higham, *Mean-square and asymptotic stability of the stochastic theta method*, SIAM J. Numer. Anal., 38 (2000), pp. 753–769.

[11] D. J. Higham and P. E. Kloeden, *Numerical methods for nonlinear stochastic differential equations with jumps*, Numer. Math., 101 (2005), pp. 101–119.

[12] D. J. Higham, X. Mao, and A. M. Stuart, *Strong convergence of Euler-like methods for nonlinear stochastic differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1041–1063.

[13] X. Mao, *Stability of Stochastic Differential Equations with Respect to Semimartingales*, Pitman Res. Notes Math. Ser. 251, Longman Scientific and Technical, Harlow, UK, 1991.

[14] X. Mao, *Exponential Stability of Stochastic Differential Equations*, Marcel Dekker, New York, 1994.

[15] X. Mao, *Stochastic Differential Equations and Applications*, Horwood, Chichester, UK, 1997.

[16] J. Mattingly, A. M. Stuart, and D. J. Higham, *Ergodicity for SDEs and approximations: Locally Lipschitz vector fields and degenerate noise*, Stochastic Process. Appl., 101 (2002), pp. 185–232.

[17] G. N. Milstein, E. Platen, and H. Schurz, *Balanced implicit methods for stiff stochastic systems*, SIAM J. Numer. Anal., 35 (1998), pp. 1010–1019.

[18] G. N. Milstein and M. V. Tretyakov, *Numerical integration of stochastic differential equations with nonglobally Lipschitz coefficients*, SIAM J. Numer. Anal., 43 (2005), pp. 1139–1154.

[19] Y. Saito and T. Mitsui, *T-stability of numerical scheme for stochastic differential equations*, in Contributions in Numerical Mathematics, World Sci. Ser. Appl. Anal. 2, World Scientific, River Edge, NJ, 1993, pp. 333–344.

[20] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, UK, 1996.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 610–640

A GALERKIN BOUNDARY ELEMENT METHOD FOR HIGH FREQUENCY SCATTERING BY CONVEX POLYGONS*

S. N. CHANDLER-WILDE† AND S. LANGDON†

Abstract. In this paper we consider the problem of time-harmonic acoustic scattering in two dimensions by convex polygons. Standard boundary or finite element methods for acoustic scattering problems have a computational cost that grows at least linearly as a function of the frequency of the incident wave. Here we present a novel Galerkin boundary element method, which uses an approximation space consisting of the products of plane waves with piecewise polynomials supported on a graded mesh, with smaller elements closer to the corners of the polygon. We prove that the best approximation from the approximation space requires a number of degrees of freedom to achieve a prescribed level of accuracy that grows only logarithmically as a function of the frequency. Numerical results demonstrate the same logarithmic dependence on the frequency for the Galerkin method solution. Our boundary element method is a discretization of a well-known second kind combined-layer-potential integral equation. We provide a proof that this equation and its adjoint are well-posed and equivalent to the boundary value problem in a Sobolev space setting for general Lipschitz domains.

Key words. Galerkin boundary element method, high frequency scattering, convex polygons, Helmholtz equation, large wave number, Lipschitz domains

AMS subject classifications. 35J05, 65R20

DOI. 10.1137/06065595X

1. Introduction.

The scattering of time-harmonic acoustic waves by bounded obstacles is a classical problem that has received much attention in the literature over the years. Much effort has been put into the development of efficient numerical schemes, but an outstanding question yet to be fully resolved is how to achieve an accurate approximation to the scattered wave with a reasonable computational cost in the case that the scattering obstacle is large compared to the wavelength of the incident field.

The standard boundary or finite element method approach is to seek an approximation to the scattered field from a space of piecewise polynomial functions. However, due to the oscillatory nature of the solution, such an approach suffers from the limitation that a fixed number of degrees of freedom K are required per wavelength in order to achieve a good level of accuracy, with the accepted guideline in the engineering literature being to take $K = 10$ (see, e.g., [53] and the references therein). A further difficulty, at least for the finite element method, is the presence of "pollution errors," phase errors in wave propagation across the domain, which can lead to even more severe restrictions on the value of K when the wavelength is short [9, 39].

Let L be a linear dimension of the scattering obstacle, and set $k = 2\pi/\lambda$, where λ is the wavelength of the incident wave, so that k is the wave number, proportional to the frequency of the incident wave. Then a consequence of fixing K is that the number of degrees of freedom will be proportional to $(kL)^d$, where $d = N$ in the case of the finite element method, $d = N - 1$ in the case of the boundary element method,

*Received by the editors March 31, 2006; accepted for publication (in revised form) October 20, 2006; published electronically April 3, 2007.
http://www.siam.org/journals/sinum/45-2/65595.html
†Department of Mathematics, University of Reading, Whiteknights, PO Box 220, Berkshire RG6 6AX, UK (s.n.chandler-wilde@reading.ac.uk, s.langdon@reading.ac.uk). The work of the second author was supported by a Leverhulme Trust Early Career Fellowship.

and $N = 2$ or 3 is the number of space dimensions of the problem. Thus, as either the frequency of the incident wave or the size of the obstacle grows, so does the number of degrees of freedom, and hence the computational cost of the numerical scheme. As a result, the numerical solution of many realistic physical problems is intractable using current technologies. In fact, for some of the most powerful recent algorithms for three-dimensional (3D) scattering problems (e.g., [13, 21]), the largest obstacles for which numerical results have been reported have diameter not more than a few hundred times the wavelength.

For boundary element methods, the cost of setting up and solving the large linear systems which arise can be reduced substantially through a combination of preconditioned iterative methods [4, 22, 36] combined with fast matrix-vector multiply methods based on the fast multipole method [5, 26, 21] or the FFT [13]. However, this does nothing to reduce the growth in the number of degrees of freedom as kL increases (linear with respect to kL in two dimensions, quadratic in three dimensions). Thus computations become infeasible as $kL \to \infty$.

1.1. Reducing the number of degrees of freedom for kL large. To achieve a dependence of the number of degrees of freedom on kL which is lower than $(kL)^d$, it seems essential to use an approximation space better able to replicate the behavior of the scattered field at high frequencies than piecewise polynomials. To that end, much attention in the recent literature has focused on enriching the approximation space with oscillatory functions, specifically plane waves or Bessel functions.

A common approach (see, e.g., [8, 16, 27, 37, 53]) is to form an approximation space consisting of standard finite element basis functions multiplied by plane waves travelling in a large number of directions, approximately uniformly distributed on the unit circle (in two dimensions) or sphere (in three dimensions). Theoretical analysis (e.g., [8]) and computational results (e.g., [53]) suggest that these methods converge rapidly as the number of plane wave directions increases, with a significant reduction in the number of degrees of freedom required per wavelength, compared to standard finite and boundary element methods. But the number of degrees of freedom is still proportional to $(kL)^d$, and serious conditioning problems occur when the number of plane wave directions is large.

A related idea is to attempt to identify the important wave propagation directions at high frequencies, and to incorporate the oscillatory part of this high frequency asymptotic behavior into the approximation space. This is the idea behind the finite element method of [34] and the boundary element methods of [25, 19, 12, 33, 45]. This idea has been investigated most thoroughly in the case that the scattering obstacle is smooth and strictly convex. In this case the leading order oscillatory behavior is particularly simple on the boundary of the scattering obstacle, so that this approach is perhaps particularly well adapted for boundary element methods. If a direct integral equation formulation is used, in which the solution to be determined is the trace of the total field or its normal derivative on the boundary, the most important wave direction to include is that of the incident wave (see, e.g., [1, 25, 12, 28]). This approach is equivalent, in the case of a sound hard scatterer, to approximating the ratio of the total field to the incident field, with physical optics predicting that this ratio is approximately constant on the illuminated side and approximately zero on the shadow side of the obstacle at high frequencies.

In [1], Abboud, Nédélec, and Zhou consider the two-dimensional (2D) problem of scattering by a smooth, strictly convex obstacle. They suggest that the ratio of the scattered field to the incident field can be approximated with error of order

$N^{-\nu} + ((kL)^{1/3}/N)^{\nu+1}$ using a uniform mesh of piecewise polynomials of degree ν, so that the total number of degrees of freedom N need be proportional only to $(kL)^{1/3}$ in order to maintain a fixed level of accuracy. In fact, this paper appears to be the first in which the dependence of the error estimates on the wave number k is indicated, and the requirement that the number of degrees of freedom is proportional to $(kL)^{1/3}$ is a big improvement over the usual requirement for proportionality to kL. This approach is coupled with a fast multipole method in [25], where impressive numerical results are reported for large scale 3D problems.

The same approach is combined with a mesh refinement concentrating degrees of freedom near the shadow boundary in [12]. The numerical results in [12] for scattering by a circle suggest that, with this mesh refinement, both the number of degrees of freedom and the total computational cost required to maintain a fixed level of accuracy remain constant as $kL \to \infty$. The method of [12] has recently been applied to deal with each of the multiple scatters which occur when a wave is incident on two, separated, smooth convex 2D obstacles [33]. Numerical experiments have also recently been presented in [29], where the convergence of this iterative approach to the multiple scattering problem is analyzed.

In [28] a numerical method in the spirit of [12] is proposed, namely a p-version boundary element method with a k-dependent mesh refinement in a transition region around the shadow boundary. A rigorous error analysis, which combines estimates using high frequency asymptotics of derivatives of the solution on the surface with careful numerical analysis, demonstrates that the approximation space is able to represent the oscillatory solution to any desired accuracy provided the number of degrees of freedom increases approximately in proportion to $(kL)^{1/9}$ as kL increases. And, in fact, numerical experiments in [28], using this approximation space as the basis of a Galerkin method, suggest that a prescribed accuracy can be achieved by keeping the number of degrees of freedom fixed as the wave number increases.

The boundary element method and its analysis that we will present in this paper for the problem of scattering by a convex polygon are most closely related to our own recent work [19, 45] on the specific problem of 2D acoustic scattering by an inhomogeneous, piecewise constant impedance plane. In [19, 45] a Galerkin boundary element method for this problem is proposed, in which the leading order high frequency behavior as $k \to \infty$, consisting of the incident and reflected ray contributions, is first subtracted off. The remaining scattered wave, consisting of rays diffracted by discontinuities in impedance, is expressed as a sum of products of oscillatory and nonoscillatory functions, with the nonoscillatory functions being approximated by piecewise polynomials supported on a graded mesh, with larger elements away from discontinuities in impedance. For the method in [19] it was shown in that paper that the number of degrees of freedom needed to maintain accuracy as $k \to \infty$ grows only logarithmically with k. This result was improved in [45], where it was shown, via sharper regularity results and a modified mesh, that for a fixed number of degrees of freedom the error is bounded independently of k.

1.2. The oscillatory integral problem. In the above paragraphs we have reviewed methods for reducing the dependence on k of the number of degrees of freedom necessary to achieve a required accuracy. Indeed some of the methods we have described above [12, 45, 33, 28] appear, in numerical experiments, to require only a number of degrees of freedom $M = O(1)$ as $k \to \infty$. Further, for one specific scattering problem [45] this has been shown by a rigorous numerical analysis. However, it should be emphasised strongly that this is not the end of the story; $M = O(1)$ as $k \to \infty$

does not imply a computational cost which is $O(1)$ as $k \to \infty$. The reason is that, while M fixed implies a fixed size of the approximating linear system, the matrix entries become increasingly difficult to evaluate, at least by conventional quadrature methods, as $k \to \infty$. This observation is perhaps particularly true for boundary integral equation based methods where the difficulty arises from the high frequency behavior of both the oscillatory basis functions (necessary to keep M fixed as $k \to \infty$) and the oscillatory kernels of the integral operators. As a consequence, each matrix entry is a highly oscillatory integral when k is large. We discuss only briefly in this paper the effective evaluation of the matrix entries in the Galerkin method we will propose, referring the reader to [44] for most of the details. And the methods we describe in [44] are $O(1)$ in computational cost as $k \to \infty$ for many but not all of the matrix entries, so that further work is required to make the algorithm we will propose fully effective at high frequency. But we note that, of the papers cited above, only the methods of Bruno et al. [12], Geuzaine, Bruno, and Reitich [33], and Langdon and Chandler-Wilde [45] appear to achieve an $O(1)$ computational cost as $k \to \infty$.

The issue in evaluating the matrix entries is one of numerical evaluation of oscillatory integrals. In Bruno et al. this is achieved by a "localized integration" strategy described in [12]. This strategy might be termed a "numerical method of stationary phase," in which the integrals are approximated by localized integrals over small, wave number-dependent neighborhoods of the stationary points of the oscillatory integrand. A similar strategy for integrals of the same type arising in high frequency boundary integral methods for 3D problems is developed in [32]. Promising alternative approaches are two older methods for oscillatory integrals due to Filon [31] (recently reanalyzed by Iserles [40, 41]; see [6] for a discussion of its application to the matrix entries in a high frequency collocation boundary element method) and Levin [47], and methods based on deformation of paths of integration into the complex plane to steepest descent paths [38]. We note that, in contrast to [12, 33, 6], where Nyström/collocation methods are used and the oscillatory integrals are one-dimensional, the matrix entries in our Galerkin methods are, of course, 2D oscillatory integrals, so that development of a robust method for their evaluation is a harder problem.

1.3. The main results of the paper. In this paper, we consider specifically the problem of scattering by convex polygons. This is, in at least one respect, a more challenging problem than the smooth convex obstacle since the corners of the polygon give rise to strong diffracted rays which illuminate the shadow side of the obstacle much more strongly than the rays that creep into the shadow zone of a smooth convex obstacle. These creeping rays decay exponentially, so that it is enough to remove the oscillation of the incident field to obtain a sufficiently simple field to approximate by piecewise polynomials, though a wave number-dependent, carefully graded mesh (cf. [12, 28]) must be used to resolve the transition zone between illuminated and shadow regions.

This approach, of removing the oscillation of the incident field and then approximating by a piecewise polynomial, does not suffice for a scatterer with corners. In brief, our algorithm for the convex polygon is as follows, inspired by our previously developed algorithm for scattering by a piecewise constant impedance plane [19], discussed in the last paragraph of section 1.1. From the geometrical theory of diffraction, one expects, on the sides of the polygon, incident, reflected, and diffracted ray contributions. On each illuminated side, the leading order behavior as $k \to \infty$ consists of the incident wave and a known reflected wave. The first stage in our algorithm is to separate this part of the solution explicitly. (On sides in shadow this step is omitted.)

The remaining field on the boundary consists of waves which have been diffracted at the corners and which travel along the polygon sides. We approximate this remaining field by taking linear combinations of products of piecewise polynomials with plane waves, the plane waves travelling parallel to the polygon sides. A key ingredient in our algorithm is to design a graded mesh to go on each side of the polygon for the piecewise polynomial approximation. This mesh has larger elements away from the corners and a mesh grading near the corners depending on the internal angles, in such a way as to equidistribute the approximation error over the subintervals of the mesh, based on a careful study of the oscillatory behavior of the solution.

The major results of the paper are as follows. We begin in section 2 by introducing the exterior Dirichlet scattering problem that we will solve numerically via a second kind boundary integral equation formulation. Our boundary integral equation is well known (e.g., [23]), obtained from Green's representation theorem. The boundary integral operator is a linear combination of a single-layer potential and its normal derivative, so that the integral equation is precisely the adjoint of the equation proposed independently for the exterior Dirichlet problem by Brakhage and Werner [11], Leis [46], and Panič [52]. However, it seems (see, e.g., the introduction to [14]) not to be widely appreciated that these formulations are well-posed for Lipschitz as well as smooth domains in a range of boundary Sobolev spaces; indeed there exists only a brief and partial account of these standard formulations for the Lipschitz domain case in the literature [50] (the treatment in [23] is for domains of class C^2). We remedy this gap in the literature in section 2, showing that our operator is a bijection on the boundary Sobolev space $H^{s-1/2}(\Gamma)$ and the adjoint operator of [11] is a bijection on $H^{s+1/2}(\Gamma)$, both for $|s| \leq 1/2$. Our starting points are known results on the (Laplace) double-layer potential operator on Lipschitz domains [57, 30] coupled with mapping properties of the single-layer potential operator [49]. (We note that this obvious approach of deducing results for the Helmholtz equation as a perturbation from the Laplace case has previously been employed for second kind boundary integral equations in Lipschitz domains in [56, 50, 48].) Of course the results we obtain apply in particular to a polygonal domain in two dimensions.

The design of our numerical algorithm depends on a careful analysis of the oscillatory behavior of the solution of the integral equation (which is the normal derivative of the total field on the boundary Γ). This is the content of section 3 of the paper. In contrast, e.g., to [28], where this information is obtained by difficult high frequency asymptotics, we adapt a technique from [19, 45], where explicit representations of the solution in a half-plane are obtained from Green's representation theorem. In the estimates we obtain of high order derivatives, we take care to obtain as precise information as possible, with a view to the future design of alternative numerical schemes, perhaps based on a *p*- or *hp*-boundary element method.

Section 4 of the paper contains, arguably, the most significant theoretical and practical results. In this section we design an approximation space for the normal derivative of the total field on Γ. As outlined above, on each side we approximate this unknown as the sum of the leading order asymptotics (known explicitly, and zero on a side in shadow) plus an expression of the form $\exp(\mathrm{i}ks)V_+(s) + \exp(-\mathrm{i}ks)V_-(s)$, where s is arc-length distance along the side and $V_\pm(s)$ are piecewise polynomials. We show, as a main result of the paper, that the approximation space based on this representation has the property that the error in best approximation of the normal derivative of the total field is bounded by $C_\nu(n[1+\log(kL)])^{\nu+3/2}M_N^{-\nu-1}$, where M_N is the total number of degrees of freedom, L is the length of the perimeter, n is the number of sides of the polygon, ν is the polynomial degree, and the constant C_ν

depends only on ν and the corner angles of the polygon. This is a strong result, showing that the number of degrees of freedom need only increase like $\log^{3/2}(kL)$ as $kL \to \infty$ to maintain accuracy.

In section 5 we analyze a Galerkin method, based on the approximation space of section 4. We show that the same bound holds for our Galerkin method approximation to the solution of the integral equation, except that an additional stability constant is introduced. We do not attempt the (difficult) task of ascertaining the dependence of this stability constant on k. In section 6 we present some numerical results which fully support our theoretical estimates, and we discuss, briefly, some numerical implementation issues, including conditioning and evaluation of the integrals, that arise. We finish the paper with some concluding remarks and open problems.

We note that the Galerkin method is, of course, not the only way to select a numerical solution from a given approximation space. In [6] we present some results for a collocation method, based on the approximation space results in section 4. The attraction of the Galerkin method we present in section 5 is that we are able to establish stability, at least in the asymptotic limit of sufficient mesh refinement, which we do not know how to do for the collocation method.

2. The boundary value problem and integral equation formulation. Consider scattering of a time-harmonic acoustic plane wave u^i by a sound-soft convex polygon Υ, with boundary $\Gamma := \bigcup_{j=1}^n \Gamma_j$, where Γ_j, $j = 1, \ldots, n$, are the n sides of the polygon with j increasing counterclockwise, as shown in Figure 2.1. We denote by $P_j := (p_j, q_j)$, $j = 1, \ldots, n$, the vertices of the polygon, and we set $P_{n+1} = P_1$ so that, for $j = 1, \ldots, n$, Γ_j is the line joining P_j with P_{j+1}. We denote the length of Γ_j by $L_j := |P_{j+1} - P_j|$, the external angle at each vertex P_j by $\Omega_j \in (\pi, 2\pi)$, the unit normal perpendicular to Γ_j and pointing out of Υ by $\mathbf{n}_j := (n_{j1}, n_{j2})$, and the angle of incidence of the plane wave, as measured counterclockwise from the downward vertical, by $\theta \in [0, 2\pi)$. Writing $\mathbf{x} = (x_1, x_2)$ and $\mathbf{d} := (\sin\theta, -\cos\theta)$, we then have

$$u^i(\mathbf{x}) = \mathrm{e}^{\mathrm{i}k(x_1 \sin\theta - x_2 \cos\theta)} = \mathrm{e}^{\mathrm{i}k\mathbf{x}\cdot\mathbf{d}}.$$

We will say that Γ_j is in shadow if $\mathbf{n}_j \cdot \mathbf{d} \geq 0$ and is illuminated if $\mathbf{n}_j \cdot \mathbf{d} < 0$. If n_s is the number of sides in shadow and it is convenient to choose the numbering so that sides $1, \ldots, n_s$ are in shadow and sides $n_s + 1, \ldots, n$ are illuminated.

We will formulate the boundary value problem we wish to solve for the total acoustic field u in a standard Sobolev space setting. For an open set $G \subset \mathbb{R}^N$, let $H^1(G) := \{v \in L^2(G) : \nabla v \in L^2(G)\}$ (∇v denoting here the weak gradient of v). We recall [49] that if G is a Lipschitz domain, then there is a well-defined trace operator, the unique bounded linear operator $\gamma : H^1(G) \to H^{1/2}(\partial G)$ which satisfies $\gamma v = v|_{\partial G}$ in the case when $v \in C^\infty(\bar{G}) := \{w|_{\bar{G}} : w \in C^\infty(\mathbb{R}^N)\}$. Let $H^1(G; \Delta) := \{v \in H^1(G) : \Delta v \in L^2(G)\}$ (Δ the Laplacian in a weak sense), a Hilbert space with the norm $\|v\|_{H^1(G;\Delta)} := \{\int_G [|v|^2 + |\nabla v|^2 + |\Delta v|^2] \mathrm{d}x\}^{1/2}$. If G is Lipschitz, then [49] there is also a well-defined normal derivative operator, the unique bounded linear operator $\partial_{\mathbf{n}} : H^1(G; \Delta) \to H^{-1/2}(\partial G)$ which satisfies

$$\partial_{\mathbf{n}} v = \frac{\partial v}{\partial \mathbf{n}} := \mathbf{n} \cdot \nabla v,$$

almost everywhere on Γ, when $v \in C^\infty(\bar{G})$. $H^1_{\mathrm{loc}}(G)$ denotes the set of measurable $v : G \to \mathbb{C}$ for which $\chi v \in H^1(G)$ for every compactly supported $\chi \in C^\infty(\bar{G})$.

The polygonal domain Υ is Lipschitz as is its exterior $D := \mathbb{R}^2 \setminus \bar{\Upsilon}$. Let $\gamma_+ : H^1(D) \to H^{1/2}(\Gamma)$ and $\gamma_- : H^1(\Upsilon) \to H^{1/2}(\Gamma)$ denote the exterior and interior trace

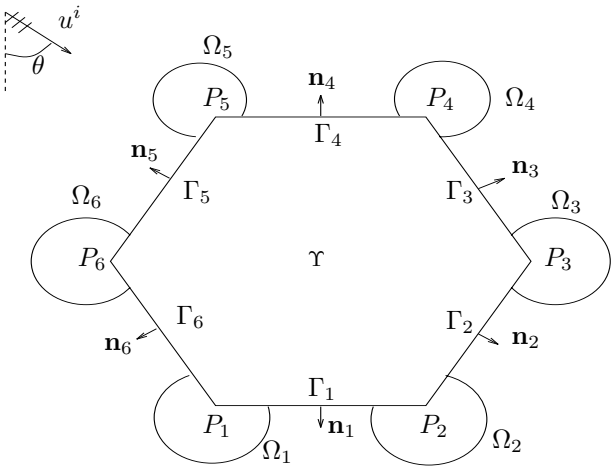


FIG. 2.1. *Our notation for the polygon.*

operators, respectively, and let $\partial_{\mathbf{n}}^+ : H^1(D;\Delta) \to H^{-1/2}(\Gamma)$ and $\partial_{\mathbf{n}}^- : H^1(\Upsilon;\Delta) \to H^{-1/2}(\Gamma)$ denote the exterior and interior normal derivative operators, respectively, the unit normal vector $\mathbf{n}$ directed out of Υ. Then the boundary value problem we seek to solve is the following: given $k > 0$ (the wave number), find $u \in C^2(D) \cap H^1_{\rm loc}(D)$ such that

$$\Delta u + k^2 u = 0 \quad \text{in } D, \tag{2.1}$$

$$\gamma_+ u = 0 \quad \text{on } \Gamma, \tag{2.2}$$

and the scattered field, $u^s := u - u^i$, satisfies the Sommerfeld radiation condition

$$\lim_{r\to\infty} r^{1/2} \left(\frac{\partial u^s}{\partial r}(\mathbf{x}) - \mathrm{i}k u^s(\mathbf{x}) \right) = 0, \tag{2.3}$$

where $r = |\mathbf{x}|$ and the limit holds uniformly in all directions $\mathbf{x}/|\mathbf{x}|$.

THEOREM 2.1 (see, e.g., [49, Theorem 9.11]). *The boundary value problem* (2.1)–(2.3) *has exactly one solution.*

Remark 2.2. While for compatibility with most of the boundary element literature we formulate the above boundary value problem in a standard Sobolev space setting, where one looks for a solution in the energy space $H^1_{\rm loc}(D)$, we note that other alternatives are available. In particular, we might seek the solution in classical function spaces as $u \in C^2(D) \cap C(\overline{D})$; this is commonly done when the boundary is sufficiently smooth [23, 24], but is also reasonable when D is Lipschitz, as it follows from standard elliptic regularity estimates up to the boundary (e.g., [42]) that if D is Lipschitz, then every solution to the Sobolev space formulation is continuous up to the boundary. A weaker requirement than $u \in C^2(D) \cap C(\overline{D})$ is usual in the harmonic analysis literature, namely to seek $u \in C^2(D)$ which satisfies the boundary condition (2.2) in the sense of almost everywhere tangential convergence, and to require that the nontangential maximal function of u is in $L^p(\Gamma)$ for some $p \in (1,\infty)$ (most commonly $p = 2$). For details of this latter formulation for the sound-soft scattering problem for the Helmholtz equation, and proofs of its well-posedness (for $2 - \epsilon < p < \infty$ and some $\epsilon > 0$) via second kind integral equation formulations, see Torres and Welland [56] for the case $\mathrm{Im}\, k > 0$, and Liu [48] and Mitrea [50] for the case $k > 0$.

Suppose that $u \in C^2(D) \cap H^1_{\mathrm{loc}}(D)$ satisfies (2.1)–(2.3). Then, by standard elliptic regularity estimates [35, section 8.11], $u \in C^\infty(\bar{D} \setminus \Gamma_C)$, where $\Gamma_C := \{P_1, \ldots, P_n\}$ is the set of corners of Γ. It is, moreover, possible to derive an explicit representation for u near the corners. For $j = 1, \ldots, n$, let $R_j := \min(L_{j-1}, L_j)$ (with $L_{-1} := L_N$). Let (r, θ) be polar coordinates local to a corner P_j, chosen so that $r = 0$ corresponds to the point P_j, the side Γ_{j-1} lies on the line $\theta = 0$, the side Γ_j lies on the line $\theta = \Omega_j$, and the part of $\bar{D}$ within distance R_j of P_j is the set of points with polar coordinates $\{(r, \theta) : 0 \le r < R_j,\ 0 \le \theta \le \Omega_j\}$. Choose R so that $R \le R_j$ and $\rho := kR < \pi/2$, and let G denote the set of points with polar coordinates $\{(r, \theta) : 0 \le r < R, 0 \le \theta \le \Omega_j\}$ (see Figure 2.2). The following result, in which J_ν denotes the Bessel function of the first kind of order ν, follows by standard separation of variables arguments.

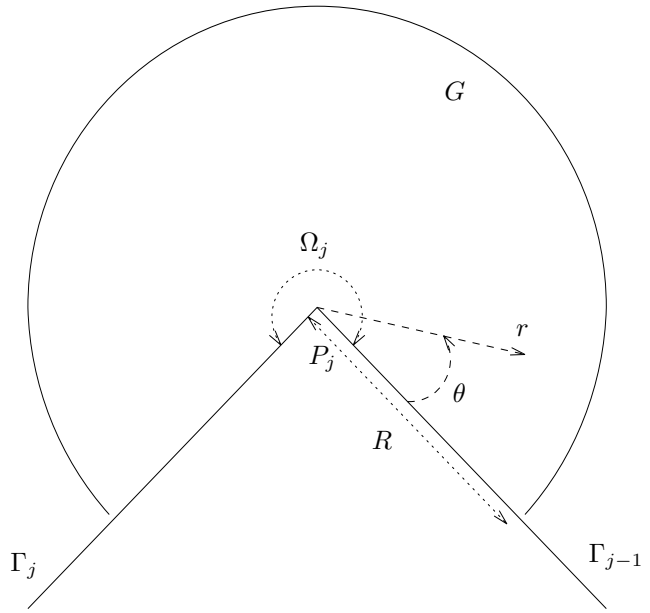


FIG. 2.2. *Neighborhood of a corner.*

THEOREM 2.3 (representation near corners). *Let $g(\theta)$ denote the value of u at the point with polar coordinates (R, θ). Then, where (r, θ) denotes the polar coordinates of $\mathbf{x}$, it holds that*

$$u(\mathbf{x}) = \sum_{n=1}^{\infty} a_n J_{n\pi/\Omega_j}(kr) \sin\left(\frac{n\theta\pi}{\Omega_j}\right), \quad \mathbf{x} \in G, \tag{2.4}$$

where

$$a_n := \frac{2}{\Omega_j J_{n\pi/\Omega_j}(kR)} \int_0^{\Omega_j} g(\theta) \sin\left(\frac{n\theta\pi}{\Omega_j}\right) \mathrm{d}\theta, \quad n \in \mathbb{N}. \tag{2.5}$$

Remark 2.4. The condition $\rho = kR < \pi/2$ ensures that $J_{n\pi/\Omega_j}(kR) \neq 0$, $n \in \mathbb{N}$, in fact (see (3.14)), that $|a_n J_{n\pi/\Omega_j}(kr)| \le C(r/R)^{n\pi/\Omega_j}$, where the constant C is independent of n and $\mathbf{x}$, so that the series (2.4) converges absolutely and uniformly in G. Thus $u \in C(\bar{D})$. Moreover, from this representation and the behavior of the

Bessel function J_ν (cf. Theorem 3.3) it follows that near the corner P_j, $\nabla u(\mathbf{x})$ has the standard singular behavior that

$$|\nabla u(\mathbf{x})| = \mathcal{O}\left(r^{\pi/\Omega_j - 1}\right) \text{ as } r \to 0. \tag{2.6}$$

From [24, Theorem 3.12] and [49, Theorems 7.15 and 9.6] we see that if u satisfies the boundary value problem (2.1)–(2.3), then a form of Green's representation theorem holds, namely

$$u(\mathbf{x}) = u^i(\mathbf{x}) - \int_\Gamma \Phi(\mathbf{x},\mathbf{y})\partial_{\mathbf{n}}^+ u(\mathbf{y})\,\mathrm{d}s(\mathbf{y}), \quad \mathbf{x} \in D, \tag{2.7}$$

where $\mathbf{n}$ is the normal direction directed out of Υ and $\Phi(\mathbf{x},\mathbf{y}) := (\mathrm{i}/4)H_0^{(1)}(k|\mathbf{x}-\mathbf{y}|)$ is the standard fundamental solution for the Helmholtz equation, with $H_0^{(1)}$ the Hankel function of the first kind of order zero. Note that, since $u \in C^\infty(\bar{D} \setminus \Gamma_C)$ and the bound (2.6) holds, we have in fact that $\partial_{\mathbf{n}}^+ u = \partial u/\partial \mathbf{n} \in L^2(\Gamma) \cap C^\infty(\Gamma \setminus \Gamma_C)$.

Starting from the representation (2.7) for u, we will obtain the boundary integral equation for $\partial u/\partial \mathbf{n}$ which we will solve numerically later in the paper. This integral equation formulation is expressed in terms of the standard single-layer potential operator ($\mathcal{S}$) and the adjoint of the double-layer potential operator ($\mathcal{T}$), defined, for $v \in L^2(\Gamma)$, by

$$\mathcal{S}v(\mathbf{x}) := 2\int_\Gamma \Phi(\mathbf{x},\mathbf{y})v(\mathbf{y})\,\mathrm{d}s(\mathbf{y}), \quad \mathcal{T}v(\mathbf{x}) := 2\int_\Gamma \frac{\partial \Phi(\mathbf{x},\mathbf{y})}{\partial \mathbf{n}(\mathbf{x})} v(\mathbf{y})\,\mathrm{d}s(\mathbf{y}), \ \mathbf{x} \in \Gamma \setminus \Gamma_C. \tag{2.8}$$

We note that both $\mathcal{S}$ and $\mathcal{T}$ are bounded operators on $L^2(\Gamma)$. In fact, more generally ([56, Lemma 6.1] or see [49]), $\mathcal{S} : H^{s-1/2}(\Gamma) \to H^{s+1/2}(\Gamma)$ and $\mathcal{T} : H^{s-1/2}(\Gamma) \to H^{s-1/2}(\Gamma)$ for $|s| \le 1/2$, and these mappings are bounded. We state the integral equation we will solve in the next theorem. Our proof of this theorem is based on the proof in [23] for domains of class C^2, modified to use more recent results on layer potentials on Lipschitz domains.

THEOREM 2.5. *If $u \in C^2(D) \cap H^1_{\mathrm{loc}}(D)$ satisfies the boundary value problem (2.1)–(2.3), then, for every $\eta \in \mathbb{R}$, $\partial_{\mathbf{n}}^+ u = \frac{\partial u}{\partial \mathbf{n}} \in L^2(\Gamma)$ satisfies the integral equation*

$$(\mathcal{I} + \mathcal{K})\partial_{\mathbf{n}}^+ u = f \quad \text{on } \Gamma, \tag{2.9}$$

where $\mathcal{I}$ is the identity operator, $\mathcal{K} := \mathcal{T} + \mathrm{i}\eta\mathcal{S}$, and

$$f(\mathbf{x}) := 2\frac{\partial u^i}{\partial \mathbf{n}}(\mathbf{x}) + 2\mathrm{i}\eta u^i(\mathbf{x}), \quad \mathbf{x} \in \Gamma \setminus \Gamma_C.$$

Conversely, if $v \in H^{-1/2}(\Gamma)$ satisfies $(\mathcal{I} + \mathcal{K})v = f$ for some $\eta \in \mathbb{R} \setminus \{0\}$, and u is defined in D by (2.7), with $\partial_{\mathbf{n}}^+ u$ replaced by v, then $u \in C^2(D) \cap H^1_{\mathrm{loc}}(D)$ and satisfies the boundary value problem (2.1)–(2.3). Moreover, $\partial_{\mathbf{n}}^+ u = v$.

Proof. Suppose first that $v \in H^{-1/2}(\Gamma)$ satisfies $(\mathcal{I} + \mathcal{K})v = f$ and define u by $u := u^i - Sv$, where

$$Sv(\mathbf{x}) := \int_\Gamma \Phi(\mathbf{x},\mathbf{y})v(\mathbf{y})\,\mathrm{d}s(\mathbf{y}), \quad \mathbf{x} \in \mathbb{R}^2 \setminus \Gamma.$$

Then [49, Chapter 9, Theorem 6.11] $u \in C^2(\mathbb{R}^2 \setminus \Gamma) \cap H^1_{\mathrm{loc}}(\mathbb{R}^2)$ and satisfies (2.1) in $\mathbb{R}^2 \setminus \Gamma$ and (2.3). Thus u satisfies the boundary value problem as long as $\gamma_+ u = 0$.

Now standard results on boundary traces of the single-layer potential on Lipschitz domains [49] give us that

$$2\gamma_{\pm}Sv = \mathcal{S}v, \quad 2\partial_{\mathbf{n}}^{\pm}(Sv) = (\mp\mathcal{I} + \mathcal{T})v. \tag{2.10}$$

On the other hand, we have that $(\mathcal{I} + \mathcal{T} + \mathrm{i}\eta\mathcal{S})v = f$. Thus

$$2\partial_{\mathbf{n}}^{-}u = 2\frac{\partial u^i}{\partial \mathbf{n}} - (\mathcal{I} + \mathcal{T})v = \mathrm{i}\eta\mathcal{S}v - 2\mathrm{i}\eta\gamma_+ u^i = -2\mathrm{i}\eta\gamma_- u.$$

Applying Green's first identity [49, Theorem 4.4] to $u \in H^1(\Upsilon;\Delta)$, we deduce that

$$-\eta\int_\Gamma |\gamma_- u|^2\,\mathrm{d}s = \mathrm{Im}\int_\Gamma \partial_{\mathbf{n}}^{-}u\,\gamma_-\bar{u}\,\mathrm{d}s = 0.$$

Thus $\gamma_+ u = \gamma_- u = 0$, so that u satisfies the boundary value problem (2.1)–(2.3). Further, $\partial_{\mathbf{n}}^{-}u = 0$ and $\partial_{\mathbf{n}}^{+}u = v + \partial_{\mathbf{n}}^{-}u = v$.

Conversely, if u satisfies the boundary value problem, in which case $\partial_{\mathbf{n}}^{+}u = \frac{\partial u}{\partial \mathbf{n}} \in L^2(\Gamma) \subset H^{-1/2}(\Gamma)$ and (2.7) holds, then, applying the trace results (2.10), we deduce

$$2\gamma_+ u^i = \mathcal{S}\partial_{\mathbf{n}}^{+}u, \quad 2\frac{\partial u^i}{\partial \mathbf{n}} = (\mathcal{I} + \mathcal{T})\partial_{\mathbf{n}}^{+}u.$$

Hence (2.9) holds. ∎

The above theorem, together with Theorem 2.1, implies that the integral equation (2.9) has exactly one solution in $H^{-1/2}(\Gamma)$, provided that we choose $\eta \neq 0$.

Remark 2.6. The idea of taking a linear combination of first and second kind integral equations to obtain a uniquely solvable boundary integral equation equivalent to an exterior scattering problem for the Helmholtz equation dates back to Brakhage and Werner [11], Leis [46], and Panič [52] for the exterior Dirichlet problem and Burton and Miller [15] for the Neumann problem. In fact, the integral equation in [11, 46, 52] is precisely the adjoint of (2.9) (see the discussion and Corollary 2.8 and Remark 2.9 below). The above proof is based on that in [23]. But, while Colton and Kress [23] restrict attention to the case when Γ is sufficiently smooth (of class C^2), the proof of Theorem 2.5 is valid for arbitrary Lipschitz Γ, and in an arbitrary number of dimensions. (Note, however, that, for general Lipschitz Γ, $\mathcal{T}v$, for $v \in H^{-1/2}(\Gamma)$, must be understood as the sum of the normal derivatives of Sv on the two sides of Γ [49, Chapter 7]. This definition of $\mathcal{T}v$ is equivalent to that in (2.8) when $v \in L^2(\Gamma)$ [56, section 4],[50, section 7].)

The following theorem, which shows that the operator $\mathcal{I} + \mathcal{K}$ is bijective on a range of Sobolev spaces, holds for a general Lipschitz boundary Γ (with $\mathcal{T}$ defined as in Remark 2.6 in the general case) in any number of space dimensions ≥ 2.

THEOREM 2.7. *Let $\mathcal{A} := \mathcal{I} + \mathcal{K}$ and suppose that $\eta \in \mathbb{R} \setminus \{0\}$. Then, for $|s| \leq 1/2$, the bounded linear operator $\mathcal{A} : H^{s-1/2}(\Gamma) \to H^{s-1/2}(\Gamma)$ is bijective with bounded inverse $\mathcal{A}^{-1}$.*

Proof. It is enough to show this result for $s = \pm 1/2$; it then follows for all s by interpolation [49]. We note first that, since $H^1(\Gamma)$ is compactly embedded in $L^2(\Gamma)$ so that $L^2(\Gamma)$ is compactly embedded in $H^{-1}(\Gamma)$, and since $\mathcal{S}$ is a bounded operator from $H^{-1}(\Gamma)$ to $L^2(\Gamma)$, it follows that $\mathcal{S}$ is a compact operator on $H^{-1}(\Gamma)$ and $L^2(\Gamma)$. Let $\mathcal{T}_0$ denote the operator corresponding to $\mathcal{T}$ in the case $k = 0$; explicitly, in the case when Γ is a 2D polygon, $\mathcal{T}_0 v$, for $v \in L^2(\Gamma)$, is defined by (2.8) with $\Phi(\mathbf{x},\mathbf{y})$ replaced by $\Phi_0(\mathbf{x},\mathbf{y}) := -(2\pi)^{-1}\log|\mathbf{x}-\mathbf{y}|$. Then $\mathcal{T}_0 - \mathcal{T}$ is a bounded operator

from $H^{-1}(\Gamma)$ to $L^2(\Gamma)$ and thus a compact operator on $H^{-1}(\Gamma)$ and $L^2(\Gamma)$. (To see the boundedness of $\mathcal{T}_0 - \mathcal{T}$ it is perhaps easiest to show that the adjoint operator, $\mathcal{T}_0' - \mathcal{T}'$, is a bounded operator from $L^2(\Gamma)$ to $H^1(\Gamma)$, which follows since $D(\mathcal{T}_0' - \mathcal{T}')$ is a bounded operator on $L^2(\Gamma)$. Here D is the surface gradient operator, $\mathcal{T}'$ and $\mathcal{T}_0'$ are standard double-layer potential operators [49, Theorem 6.17], in particular

$$T'v(\mathbf{x}) := \int_\Gamma \frac{\partial \Phi(\mathbf{x},\mathbf{y})}{\partial \mathbf{n}(\mathbf{y})}\, v(\mathbf{y}) \mathbf{d}s(\mathbf{y}), \quad \mathbf{x} \in \Gamma,$$

and the boundedness of the integral operator $D(\mathcal{T}_0' - \mathcal{T}')$ follows since its kernel is continuous or weakly singular.) Thus $\mathcal{A}$, as an operator on $H^{s-1/2}(\Gamma)$, $s = \pm 1/2$, is a compact perturbation of $\mathcal{I} + \mathcal{T}_0$. But it is known that $\mathcal{I} + \mathcal{T}_0'$ is Fredholm of index zero on $H^{s+1/2}(\Gamma)$ for $|s| \leq 1/2$ (see [57, 30]), from which it follows from [49, Theorem 6.17] that the adjoint operator $\mathcal{I} + \mathcal{T}_0'$ is Fredholm of index zero on $H^{s-1/2}(\Gamma)$ for $|s| \leq 1/2$. Thus $\mathcal{A}$ is Fredholm of index zero on $H^{s-1/2}(\Gamma)$, $s = \pm 1/2$. Since $\mathcal{A}$ is Fredholm with the same index on $H^{-1}(\Gamma)$ and $L^2(\Gamma)$, and $L^2(\Gamma)$ is dense in $H^{-1}(\Gamma)$, it follows from a standard result on Fredholm operators (see, e.g., [54, section 1]) that the null-space of $\mathcal{A}$, as an operator on $H^{-1}(\Gamma)$, is a subset of $L^2(\Gamma)$. But it follows from Theorems 2.1 and 2.5 that $\mathcal{A}v = 0$ has no nontrivial solution in $H^{-1/2}(\Gamma) \supset L^2(\Gamma)$. Thus $\mathcal{A} : H^{s-1/2}(\Gamma) \to H^{s+1/2}(\Gamma)$ is invertible for $s = \pm 1/2$. □

We have observed in Remark 2.6 that an alternative integral equation formulation for the exterior Dirichlet problem was introduced in [11, 46, 52]. In this formulation one seeks a solution to the exterior Dirichlet problem in the form of a combined single- and double-layer potential with some unknown density $\tilde{\phi}$ and arrives at the boundary integral equation $\mathcal{A}'\tilde{\phi} = 2\gamma_+ u^i$, where

$$\mathcal{A}' = \mathcal{I} + \mathcal{T}' + i\eta S$$

is the adjoint of $\mathcal{A}$ in the sense that the duality relation holds that $\langle A\phi, \psi\rangle_\Gamma = \langle \phi, A'\psi\rangle_\Gamma$ for $\phi \in H^{-1/2}(\Gamma)$, $\psi \in H^{1/2}(\Gamma)$, where $\langle \phi, \psi\rangle_\Gamma := \int_\Gamma \phi(y)\psi(y)\mathrm{d}s(y)$ [49, Theorems 6.15 and 6.17]. It is known that $\mathcal{A}'$ maps $H^{s+1/2}(\Gamma)$ to $H^{s+1/2}(\Gamma)$ and that this mapping is bounded for $|s| \leq 1/2$ [56, 49]. This, the duality relation, and Theorem 2.7 imply the invertibility of $\mathcal{A}'$. Precisely, we have the following result.

COROLLARY 2.8. *For $|s| \leq 1/2$ and $\eta \in \mathbb{R} \setminus \{0\}$, the mapping $\mathcal{A}' : H^{s+1/2}(\Gamma) \to H^{s+1/2}(\Gamma)$ is bijective with bounded inverse ${\mathcal{A}'}^{-1}$.*

Remark 2.9. We note that brief details of a proof that the related operator $\tilde{\mathcal{A}}' := \mathcal{I} + \mathcal{T}' + i\eta S S_0^2$, where S_0 denotes S in the case $k = 0$, is invertible as an operator on $L^2(\Gamma)$ if $\eta \in \mathbb{R} \setminus \{0\}$ are given in Mitrea [50]. Moreover, the argument outlined in [50], which follows the same pattern that we have used to prove Theorem 2.7, namely to show that $\tilde{\mathcal{A}}'$ is Fredholm of index zero by perturbation from the Laplace case, and then to establish uniqueness by mirroring the usual uniqueness argument for smooth domains [23] (though the details of this are omitted in [50]), could be applied equally to show that $\mathcal{A}'$ is invertible on $L^2(\Gamma)$ for $\eta \in \mathbb{R} \setminus \{0\}$. Then, arguing by duality in the same way in which we deduce Corollary 2.8, we could deduce that A is invertible on $L^2(\Gamma)$. Thus the argument outlined in [50] offers an alternative route to that written out above for establishing that $\mathcal{A}$ and $\mathcal{A}'$ are invertible as operators on $L^2(\Gamma)$ for $\eta \in \mathbb{R} \setminus \{0\}$.

We also note that for the case $\eta = 0$ when $\mathcal{A}' = \mathcal{I} + \mathcal{T}'$, it is shown that $\mathcal{A}'$ is invertible as an operator on $L^2(\Gamma)$ if $\mathrm{Im}k > 0$ in [56]. This result is sharpened in [50], where it is shown that $\mathcal{A}'$ is also invertible as an operator on $L^2(\Gamma)$ if $k > 0$ is not

an eigenvalue of an appropriately stated interior Neumann problem in Υ. See [50] (and Liu [48]) for further discussion of the case when $k > 0$ is an interior Neumann eigenvalue when $\mathcal{A}'$ has a finite-dimensional kernel.

In the remainder of the paper we will focus on the properties of $\mathcal{A}$ as an operator on $L^2(\Gamma)$. We remark that the result that $\mathcal{I} + \mathcal{T}_0'$ is Fredholm of index zero on $L^2(\Gamma)$ dates back to [58] in the case when Γ is a 2D polygon. Letting $\|\cdot\|_2$ denote the norm on $L^2(\Gamma)$, the technique in [58] (or see [17]) is to show that $\mathcal{T}_0' = \mathcal{T}_1' + \mathcal{T}_2'$, where $\|\mathcal{T}_1'\|_2 < 1$. Since taking adjoints preserves norms and compactness, and since $\mathcal{S}$ and $\mathcal{T} - \mathcal{T}_0$ are compact operators on $L^2(\Gamma)$, it holds in the case of a 2D polygon that $\mathcal{A} = \mathcal{I} + \mathcal{K} = \mathcal{I} + \mathcal{K}_1 + \mathcal{K}_2$, where $\|\mathcal{K}_1\|_2 < 1$ and $\mathcal{K}_2$ is a compact operator on $L^2(\Gamma)$.

Throughout the remainder of the paper we suppose that $\eta \in \mathbb{R}$ with $\eta \neq 0$, so that $\mathcal{A}$ is invertible, and let

$$C_S := \|\mathcal{A}^{-1}\|_2 = \|(\mathcal{I} + \mathcal{K})^{-1}\|_2. \tag{2.11}$$

We note that the value of C_S depends on k, η, and the geometry of Γ. But recently an upper bound has been obtained for C_S as a function of k, η, and the geometry of Γ in the case when Γ is (in two dimensions or three dimensions) the boundary of a piecewise smooth, starlike Lipschitz domain [20, Theorem 4.3], by using Rellich-type identities. In particular, for the commonly recommended choice $|\eta| = k$ (see, e.g., [28]), this bound implies for the convex polygon that

$$C_S \leq \frac{1}{2}\left(1 + 9\theta + 4\theta^2\right) \tag{2.12}$$

for $kR_0 \geq 1$. Here it is assumed that the coordinate system is chosen so that the origin lies inside Γ, and we define $R_0 := \max_{x\in\Gamma} |x|$, $\theta := R_0/\delta_-$, and δ_- to be the perpendicular distance from the origin to the nearest side of the polygon. For example, in the case of a square (for which we carry out computations in section 6, choosing $\eta = k$), taking the origin at the center of the square gives $\theta = \sqrt{2}$ and so $C_S \leq \frac{9}{2}(1 + \sqrt{2}) < 11$ for $kR_0 \geq 1$.

3. Regularity results. In this section we aim to understand the behavior of $\partial u/\partial \mathbf{n}$, the normal derivative of the total field on Γ, which is the unknown function in the integral equation (2.9). Precisely, we will obtain bounds on the surface tangential derivatives of $\partial u/\partial \mathbf{n}$ in which the dependence on the wave number is completely explicit. This will enable us in section 4 to design a family of approximation spaces well adapted to approximating $\partial u/\partial \mathbf{n}$.

To understand the behavior of $\partial u/\partial \mathbf{n}$ near the corners P_j, our technique will be to use the explicit representation (2.4). To understand the behavior away from the corners, we will need another representation for $\partial u/\partial \mathbf{n}$ which we now derive.

Our starting point is the observation that if $U = \{\mathbf{x} = (x_1, x_2), x_1 \in \mathbb{R}, x_2 > 0\}$ is the upper half-plane and $v \in C^2(U) \cap C(\bar{U})$ satisfies the Helmholtz equation in U and the Sommerfeld radiation condition, then [18, Theorem 3.1]

$$v(\mathbf{x}) = 2\int_{\partial U} \frac{\partial \Phi(\mathbf{x}, \mathbf{y})}{\partial y_2} v(\mathbf{y})\, \mathrm{d}s(\mathbf{y}), \quad \mathbf{x} \in U. \tag{3.1}$$

The same formula holds [18] if v is a horizontally or upwards propagating plane wave, i.e., if $v(\mathbf{x}) = \mathrm{e}^{\mathrm{i}k\mathbf{x}.\mathbf{d}}$ with $\mathbf{d} = (d_1, d_2)$, $|\mathbf{d}| = 1$, and $d_2 \geq 0$.

To make use of this observation, we make the following construction. Extend the line Γ_j to infinity in both directions; the resulting infinite line comprises Γ_j and the

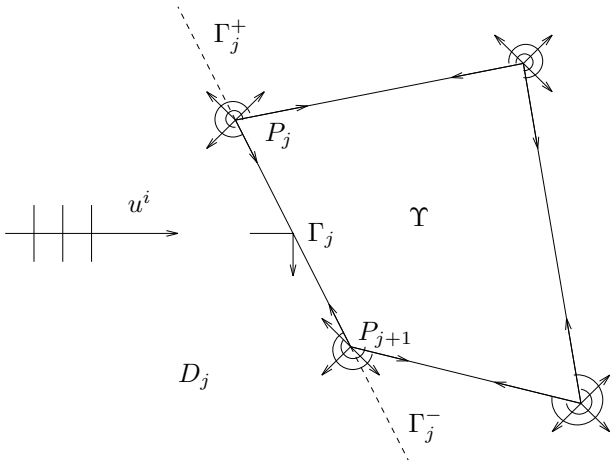


FIG. 3.1. *Extension of Γ_j, for derivation of regularity estimates.*

half-lines Γ_j^+ and Γ_j^-, above P_j and below P_{j+1}, respectively; see Figure 3.1. Let $D_j \subset D$ denote the half-plane on the opposite side of this line to Υ.

Now consider first the case when Γ_j is in shadow, by which we mean that $\mathbf{n}_j.\mathbf{d} \geq 0$. Then it follows from (3.1) that

$$u^s(\mathbf{x}) = 2\int_{\Gamma_j^+\cup\Gamma_j\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u^s(\mathbf{y})\,\mathrm{d}s(y), \quad \mathbf{x}\in D_j, \tag{3.2}$$

and also that

$$u^i(\mathbf{x}) = 2\int_{\Gamma_j^+\cup\Gamma_j\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u^i(\mathbf{y})\,\mathrm{d}s(y), \quad \mathbf{x}\in D_j. \tag{3.3}$$

Since $u = u^i + u^s$ and $u = 0$ on Γ, we deduce that

$$u(\mathbf{x}) = 2\int_{\Gamma_j^+\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u(\mathbf{y})\,\mathrm{d}s(y), \quad \mathbf{x}\in D_j.$$

In the case when Γ_j is illuminated ($\mathbf{n}_j.\mathbf{d} < 0$), (3.2) holds, but (3.3) is replaced by

$$u^i(\mathbf{x}) = -2\int_{\Gamma_j^+\cup\Gamma_j\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u^i(\mathbf{y})\,\mathrm{d}s(y), \quad \mathbf{x}\in \mathbb{R}^2\backslash\bar{D}_j. \tag{3.4}$$

Now let $u^r(\mathbf{x}) := -u^i(\mathbf{x}')$ for $\mathbf{x}\in D_j$, where $\mathbf{x}'$ is the reflection of $\mathbf{x}$ in the line $\Gamma_j^+\cup\Gamma_j\cup\Gamma_j^-$. (The physical interpretation of u^r is that it is the plane wave that would be reflected if Γ_j were infinitely long.) From (3.4), for $\mathbf{x}\in D_j$,

$$u^r(\mathbf{x}) = 2\int_{\Gamma_j^+\cup\Gamma_j\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x}',\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u^i(\mathbf{y})\,\mathrm{d}s(y) = -2\int_{\Gamma_j^+\cup\Gamma_j\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u^i(\mathbf{y})\,\mathrm{d}s(y),$$

and adding this to (3.2) we find that

$$u(\mathbf{x}) = u^i(\mathbf{x}) + u^r(\mathbf{x}) + 2\int_{\Gamma_j^+\cup\Gamma_j^-} \frac{\partial\Phi(\mathbf{x},\mathbf{y})}{\partial\mathbf{n}(\mathbf{y})} u(\mathbf{y})\,\mathrm{d}s(y), \quad \mathbf{x}\in D_j.$$

Thus on an illuminated side it holds that

$$\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = 2\frac{\partial u^i}{\partial \mathbf{n}}(\mathbf{x}) + 2\int_{\Gamma_j^+\cup\Gamma_j^-} \frac{\partial^2\Phi(\mathbf{x},\mathbf{y})}{\partial \mathbf{n}(\mathbf{x})\partial \mathbf{n}(\mathbf{y})} u(\mathbf{y})\,\mathrm{d}s(\mathbf{y}), \quad \mathbf{x}\in\Gamma_j. \tag{3.5}$$

The same expression, but without the term $2\frac{\partial u^i}{\partial \mathbf{n}}(\mathbf{x})$, holds when Γ_j is in shadow. The high frequency Kirchhoff or physical optics approximation to $\partial u/\partial \mathbf{n}$ is just $\partial u/\partial \mathbf{n} = 2\partial u^i/\partial \mathbf{n}$ on the illuminated sides and zero on the sides in shadow. Thus the integral in (3.5) is an explicit expression for the correction to the physical optics approximation.

The representation (3.5) is very useful in understanding the oscillatory nature of the solution on a typical side Γ_j. In particular we note that, in physical terms, the integral over Γ_j^+ can be interpreted as the normal derivative on Γ_j of the field due to dipoles distributed along Γ_j^+. The point is that the field due to each dipole has the same oscillatory behavior $\mathrm{e}^{\mathrm{i}ks}$ on Γ_j. To exhibit this explicitly, we calculate, using standard properties of Bessel functions [2], that for $\mathbf{x}\in\Gamma_j$, $\mathbf{y}\in\Gamma_j^{\pm}$, with $\mathbf{x}\neq\mathbf{y}$,

$$\frac{\partial^2\Phi(\mathbf{x},\mathbf{y})}{\partial \mathbf{n}(\mathbf{x})\partial \mathbf{n}(\mathbf{y})} = \frac{\mathrm{i}kH_1^{(1)}(k|\mathbf{x}-\mathbf{y}|)}{4|\mathbf{x}-\mathbf{y}|} = \frac{\mathrm{i}k^2}{4}\mathrm{e}^{\mathrm{i}k|\mathbf{x}-\mathbf{y}|}\mu(k|\mathbf{x}-\mathbf{y}|), \tag{3.6}$$

where $\mu(z) := \mathrm{e}^{-\mathrm{i}z}H_1^{(1)}(z)/z$ for $z>0$. The function $\mu(z)$ is singular at $z=0$ but increasingly smooth as $z\to\infty$, as quantified in the next theorem (cf. [19, Lemma 2.5]).

THEOREM 3.1. *For every $\epsilon>0$,*

$$|\mu^{(m)}(z)| \le C_\epsilon (m+1)!\, z^{-3/2-m}$$

for $z\ge\epsilon$ and $m=0,1,\ldots,$ where

$$C_\epsilon = \frac{2\sqrt[4]{5}(1+\epsilon^{-1/2})}{\pi}. \tag{3.7}$$

Proof. From [51, equation (12.31)], $\mu(z) = (-2\mathrm{i}/\pi)\int_0^\infty (t^2-2\mathrm{i}t)^{1/2}\mathrm{e}^{-zt}\,\mathrm{d}t$ for $\mathrm{Re}z>0$, where the branch of $(t^2-2\mathrm{i}t)^{1/2}$ is chosen so that $\mathrm{Re}(t^2-2\mathrm{i}t)^{1/2}\ge 0$. Thus

$$\mu^{(m)}(z) = (-1)^{m+1}\frac{2\mathrm{i}}{\pi}\int_0^\infty t^{m+1/2}(t-2\mathrm{i})^{1/2}\mathrm{e}^{-zt}\,\mathrm{d}t$$

and hence

$$|\mu^{(m)}(z)| \le \frac{2}{\pi}\int_0^\infty t^{m+1/2}(t^2+4)^{1/4}\mathrm{e}^{-zt}\,\mathrm{d}t.$$

Now for $t\in[0,1]$, $(t^2+4)^{1/4}\le 5^{1/4}$ and for $t\in[1,\infty)$, $(t^2+4)^{1/4}\le 5^{1/4}t^{1/2}$. So

$$\begin{aligned}\frac{\pi}{2\sqrt[4]{5}}\,|\mu^{(m)}(z)| &\le \int_0^\infty t^{m+1/2}\mathrm{e}^{-zt}\,\mathrm{d}t + \int_0^\infty t^{m+1}\mathrm{e}^{-zt}\,\mathrm{d}t \\ &= \Gamma(m+3/2)z^{-3/2-m} + \Gamma(m+2)z^{-2-m} \le (1+\epsilon^{-1/2})\Gamma(m+2)z^{-3/2-m}\end{aligned}$$

for $z\ge\epsilon$. ∎

To make use of the above result, let $\mathbf{x}(s)$ denote the point on Γ whose arc-length distance measured counterclockwise from P_1 is s. Explicitly,

$$\mathbf{x}(s) = P_j + \left(s-\tilde{L}_{j-1}\right)\left(\frac{P_{j+1}-P_j}{L_j}\right) \quad \text{for } s\in[\tilde{L}_{j-1},\tilde{L}_j], \quad j=1,\ldots,n,$$

where $\tilde{L}_0 := 0$ and for $j = 1, \ldots, n$, $\tilde{L}_j := \sum_{m=1}^{j} L_m$ is the arc-length distance from P_1 to P_{j+1}. Define

$$\phi(s) := \frac{1}{k}\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}(s)) \quad \text{for } s \in [0, L], \tag{3.8}$$

where $L := \tilde{L}_n$ so that $\phi(s)$ is the unknown function of arc-length whose behavior we seek to determine. Let

$$\Psi(s) := \begin{cases} \frac{2}{k}\frac{\partial u^i}{\partial \mathbf{n}}(\mathbf{x}(s)) & \text{if } s \in (\tilde{L}_{n_s}, L), \\ 0 & \text{if } s \in (0, \tilde{L}_{n_s}), \end{cases}$$

so that $\Psi(s)$ is the physical optics approximation to $\phi(s)$, and set $\psi_j(s) := u(\tilde{\mathbf{x}}_j(s))$, $s \in \mathbb{R}$, where $\tilde{\mathbf{x}}_j(s) \in \Gamma_j^+ \cup \Gamma_j \cup \Gamma_j^-$ is the point

$$\tilde{\mathbf{x}}_j(s) := P_j + \left(s - \tilde{L}_{j-1}\right)\left(\frac{P_{j+1} - P_j}{L_j}\right), \quad -\infty < s < \infty.$$

From (3.5) and (3.6) we have the explicit representation for ϕ on the side Γ_j that

$$\phi(s) = \Psi(s) + \frac{\mathrm{i}}{2}[\mathrm{e}^{\mathrm{i}ks} v_j^+(s) + \mathrm{e}^{-\mathrm{i}ks} v_j^-(s)], \quad s \in [\tilde{L}_{j-1}, \tilde{L}_j], \quad j = 1, \ldots, n, \tag{3.9}$$

where

$$v_j^+(s) := k \int_{-\infty}^{\tilde{L}_{j-1}} \mu(k|s-t|)\mathrm{e}^{-\mathrm{i}kt}\psi_j(t)\,\mathrm{d}t, \quad s \in [\tilde{L}_{j-1}, \tilde{L}_j], \quad j = 1, \ldots, n,$$
$$v_j^-(s) := k \int_{\tilde{L}_j}^{\infty} \mu(k|s-t|)\mathrm{e}^{\mathrm{i}kt}\psi_j(t)\,\mathrm{d}t, \quad s \in [\tilde{L}_{j-1}, \tilde{L}_j], \quad j = 1, \ldots, n.$$

The terms $\mathrm{e}^{\mathrm{i}ks} v_j^+(s)$ and $\mathrm{e}^{-\mathrm{i}ks} v_j^-(s)$ in (3.9) are the integrals over Γ_j^+ and Γ_j^-, respectively, in (3.5) and can be thought of as the contributions to $\partial u/\partial \mathbf{n}$ on Γ_j due to the diffracted rays travelling from P_j to P_{j+1} and from P_{j+1} to P_j, respectively, including all multiply diffracted ray components.

Thus the equation we wish to solve is (2.9), and we have the explicit representation (3.9) for its solution. At first glance this may not appear to help us, since the unknown solution u appears (as ψ_j) on the right-hand side of (3.9). However, (3.9) is extremely helpful in understanding how ϕ behaves since it explicitly separates out the oscillatory part of the solution. The functions $v_j^{\pm}$ are not oscillatory away from the corners, as the following theorem quantifies. In this theorem and hereafter we let

$$u_M := \sup_{\mathbf{x} \in D} |u(\mathbf{x})| < \infty \tag{3.10}$$

and note that $\|\psi_j\|_\infty \le u_M$, $j = 1, \ldots, n$.

THEOREM 3.2 (solution behavior away from corners). *For $\epsilon > 0$, $j = 1, \ldots, n$, and $m = 0, 1, \ldots$, it holds for $s \in [\tilde{L}_{j-1}, \tilde{L}_j]$ that*

$$|v_j^{+(m)}(s)| \le 2C_\epsilon m! u_M k^m (k(s - \tilde{L}_{j-1}))^{-1/2-m}, \quad k(s - \tilde{L}_{j-1}) \ge \epsilon,$$
$$|v_j^{-(m)}(s)| \le 2C_\epsilon m! u_M k^m (k(\tilde{L}_j - s))^{-1/2-m}, \quad k(\tilde{L}_j - s) \ge \epsilon,$$

where C_ϵ is given by (3.7).

Proof. From Theorem 3.1, for $s \in [\tilde{L}_{j-1} + \epsilon/k, \tilde{L}_j]$

$$\begin{aligned}
|v_j^{+(m)}(s)| &= k^{m+1} \left| \int_{-\infty}^{\tilde{L}_{j-1}} \mu^{(m)}(k|s-t|) \mathrm{e}^{-\mathrm{i}kt} \psi_j(t)\, \mathrm{d}t \right| \\
&\le C_\epsilon (m+1)! k^{m+1} \|\psi_j\|_\infty \int_{-\infty}^{\tilde{L}_{j-1}} (k|s-t|)^{-3/2-m}\, \mathrm{d}t \\
&= C_\epsilon \frac{(m+1)!}{(m+1/2)} k^{-1/2} \|\psi_j\|_\infty (s - \tilde{L}_{j-1})^{-1/2-m} \\
&\le 2C_\epsilon m! u_M k^m (k(s-\tilde{L}_{j-1}))^{-1/2-m}.
\end{aligned}$$

The bound on $v_j^{-(m)}(s)$ is obtained similarly. ∎

The above theorem quantifies precisely the behavior of $\partial u/\partial \mathbf{n}$ away from the corners. Complementing this bound, using Theorem 2.3 we can study the behavior of $\partial u/\partial \mathbf{n}$ near the corners. To state this result it is convenient to extend the definition of ϕ from $[0, L]$ to $\mathbb{R}$ by the periodicity condition $\phi(s + L) = \phi(s)$, $s \in \mathbb{R}$.

THEOREM 3.3 (solution behavior near corners). *If $kR_j = \min(kL_{j-1}, kL_j) \ge \pi/4$ for $j = 1, \ldots, n$, then for $j = 1, \ldots, n$ and $0 < k|s - \tilde{L}_{j-1}| \le \pi/12$, it holds that*

$$\left| \phi^{(m)}(s) \right| \le C u_M \sqrt{m + \frac{1}{2}}\, m! k^m (k|s - \tilde{L}_{j-1}|)^{-\alpha_j - m}, \quad m = 0, 1, \ldots,$$

where

$$\alpha_j := 1 - \frac{\pi}{\Omega_j} \in (0, 1/2) \tag{3.11}$$

and $C = 72\sqrt{2}\, \pi^{-1}\, \mathrm{e}^{1/\mathrm{e} + \pi/6}$.

Proof. To analyze the behavior of u using (2.4) we will use the representation for the Bessel function of order ν [2, equation (9.1.20)],

$$J_\nu(z) = \frac{2(z/2)^\nu}{\pi^{1/2}\Gamma(\nu + 1/2)} \int_0^1 (1 - t^2)^{\nu - 1/2} \cos(zt)\, \mathrm{d}t \text{ for } \mathrm{Re} z > 0,\ \nu > -1/2,$$

where the branch of $(z/2)^\nu$ is chosen so that $(z/2)^\nu > 0$ for $z > 0$ and $(z/2)^\nu$ is analytic in $\mathrm{Re} z > 0$. This representation implies that

$$\cos z \le \frac{J_\nu(z) \pi^{1/2} \Gamma(\nu + 1/2)}{2(z/2)^\nu \int_0^1 (1 - t^2)^{\nu - 1/2}\, \mathrm{d}t} \le 1, \quad 0 \le z \le \pi/2. \tag{3.12}$$

Recalling the definitions of R and G before Theorem 2.3 and the Definition (2.5) of the coefficient a_n, we have that $\rho := kR < \pi/2$ and

$$|a_n| \le \frac{2u_M}{J_{n\pi/\Omega_j}(\rho)}. \tag{3.13}$$

Thus, for $0 < r < R$,

$$\left| a_n J_{n\pi/\Omega_j}(kr) \right| \le \frac{2u_M}{\cos \rho} \left(\frac{r}{R} \right)^{n\pi/\Omega_j}, \tag{3.14}$$

confirming that the series (2.4) converges for $0 \le r < R$. Further, the bound (3.14) justifies differentiating (2.4) term by term to get that for $\mathbf{x} \in \Gamma_{j-1} \cap G$, $\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = kF(kr)$, where

$$F(z) := \frac{\pi}{\Omega_j z} \sum_{n=1}^{\infty} n a_n J_{n\pi/\Omega_j}(z), \quad \mathrm{Re} z > 0, \quad |z| < \rho. \tag{3.15}$$

Since $|\cos z| \le \mathrm{e}^{|\mathrm{Im} z|}$, $z \in \mathbb{C}$, so that $|\cos zt| \le \mathrm{e}^{|\mathrm{Im} z|}$ for $z \in \mathbb{C}$, $0 \le t \le 1$, we see from (3.13) that for $\mathrm{Re} z > 0$,

$$\left| n a_n J_{n\pi/\Omega_j}(z) \right| \le \frac{2u_M n}{\cos \rho} \mathrm{e}^{|\mathrm{Im} z|} \left(\frac{|z|}{\rho} \right)^{n\pi/\Omega_j}. \tag{3.16}$$

Thus the series (3.15) is absolutely and uniformly convergent in $\mathrm{Re} z > 0$, $|z| < \rho_0$, for every $\rho_0 < \rho$, and F is analytic in $\mathrm{Re} z > 0$, $|z| < \rho$. Further, from (3.16) and since for $0 \le \alpha < 1$, $\sum_{n=1}^{\infty} n\alpha^n = \alpha \frac{\mathrm{d}}{\mathrm{d}\alpha} \sum_{n=1}^{\infty} \alpha^n = \frac{\alpha}{(1-\alpha)^2}$, we see that for $\mathrm{Re} z > 0$, $|z| < \rho$,

$$|F(z)| \le \frac{\pi}{\Omega_j |z|} \frac{2u_M}{\cos \rho} \frac{\mathrm{e}^{|\mathrm{Im} z|}}{(1 - |z/\rho|^{\pi/\Omega_j})^2} \left(\frac{|z|}{\rho} \right)^{\pi/\Omega_j}.$$

We can use this bound to obtain bounds on derivatives of F, and hence bounds on derivatives of $\partial u/\partial \mathbf{n}$. For $0 < t \le \rho/3$, $0 < \varepsilon < t$, from Cauchy's integral formula we have that

$$|F^{(m)}(t)| = \frac{m!}{2\pi} \left| \int_{\Gamma_\varepsilon} \frac{F(z)}{(z-t)^{m+1}} \, \mathrm{d}z \right|,$$

where Γ_ε is the circle of radius ε centered on t, which lies in $\mathrm{Re} z > 0$, $|z| < \rho$. Since

$$|F(z)| \le \frac{2\pi u_M \mathrm{e}^{|\mathrm{Im} z|} (t-\varepsilon)^{\pi/\Omega_j - 1}}{\Omega_j \rho^{\pi/\Omega_j} \cos \rho (1 - (2/3)^{\pi/\Omega_j})^2}$$

for $z \in \Gamma_\varepsilon$, we see that

$$|F^{(m)}(t)| \le \frac{2\pi u_M \mathrm{e}^t (t-\varepsilon)^{\pi/\Omega_j - 1} \varepsilon^{-m} m!}{\Omega_j \rho^{\pi/\Omega_j} \cos \rho (1 - (2/3)^{\pi/\Omega_j})^2}. \tag{3.17}$$

Now, for $\alpha > 0$, $\beta > 0$, $(t-\varepsilon)^{-\alpha} \varepsilon^{-\beta}$ is minimized on $(0,t)$ by the choice $\varepsilon = \beta t/(\alpha+\beta)$. Setting $\varepsilon = mt/(m+1-\pi/\Omega_j)$ in (3.17), we see that

$$|F^{(m)}(t)| \le \frac{2\pi u_M \mathrm{e}^t m! (m+1-\pi/\Omega_j)^{m+1-\pi/\Omega_j} t^{\pi/\Omega_j - 1 - m}}{\Omega_j \rho^{\pi/\Omega_j} \cos \rho (1 - (2/3)^{\pi/\Omega_j})^2 m^m (1-\pi/\Omega_j)^{1-\pi/\Omega_j}}.$$

Now

$$\frac{(m+1-\pi/\Omega_j)^{m+1-\pi/\Omega_j}}{m^m} \le \frac{(m+1/2)^{m+1/2}}{m^m} = \left(1 + \frac{1}{2m}\right)^m \sqrt{m+\frac{1}{2}} \le \mathrm{e}^{1/2} \sqrt{m+\frac{1}{2}},$$

$$\frac{2\pi}{\Omega_j (1-\pi/\Omega_j)^{1-\pi/\Omega_j} (1 - (2/3)^{\pi/\Omega_j})^2} \le \frac{18}{(1-\pi/\Omega_j)^{1-\pi/\Omega_j}} \le 18\mathrm{e}^{1/\mathrm{e}},$$

and hence

$$|F^{(m)}(t)| \le \frac{18 \mathrm{e}^{1/\mathrm{e}+1/2+t} \sqrt{m+1/2}\, m! u_M}{\rho^{\pi/\Omega_j} \cos \rho} t^{\pi/\Omega_j - 1 - m}, \quad 0 < t \le \rho/3. \tag{3.18}$$

Since $\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}) = kF(kr)$, this implies that

$$\left|\frac{\partial^{(m)}}{\partial r^m}\left[\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x})\right]\right| \leq \tilde{C} u_M\, k^{m+1}(kr)^{\pi/\Omega_j - 1 - m}, \quad 0 < r \leq R/3 < \frac{\pi}{6k},$$

where $\tilde{C} = (18\mathrm{e}^{1/\mathrm{e}+1/2+\pi/6}\sqrt{m+1/2}\, m!)/(\rho^{\pi/\Omega_j}\cos\rho)$. Choosing $\rho = \pi/4$, the result follows. ∎

From Theorems 3.2 and 3.3, and (3.9), which gives that

$$v_j^{\pm}(s) = -2\mathrm{i}\mathrm{e}^{\mp \mathrm{i}ks}(\phi(s) - \Psi(s)) - \mathrm{e}^{\pm 2\mathrm{i}ks} v_j^{\mp}(s),$$

we deduce the following corollary, in which $\alpha_{n+1} := \alpha_1$.

COROLLARY 3.4. *Suppose that $kR_j = \min(kL_{j-1}, kL_j) \geq \pi/4$ for $j = 1, \ldots, n$. Then, for $m = 0, 1, \ldots$, there exists $C_m > 0$, dependent only on m, such that if $j \in \{1, \ldots, n\}$, then*

$$|v_j^{+(m)}(s)| \leq C_m u_M k^m (k(s - \tilde{L}_{j-1}))^{-\alpha_j - m}, \quad 0 < k(s - \tilde{L}_{j-1}) \leq \pi/12,$$
$$|v_j^{-(m)}(s)| \leq C_m u_M k^m (k(\tilde{L}_j - s))^{-\alpha_{j+1} - m}, \quad 0 < k(\tilde{L}_j - s) \leq \pi/12.$$

The following limiting case suggests that the bounds in Theorem 3.2 and Corollary 3.4 are optimal in their dependence on k, $s - \tilde{L}_{j-1}$, and $\tilde{L}_j - s$, in the sense that no sharper bound holds uniformly in the angle of incidence. Suppose that Υ lies in the right-hand half-plane with P_1 located at the origin and $\mathbf{d} \cdot \mathbf{n}_1 = 0$, and consider the limit $\min(kL_0, kL_1) \to \infty$ and $\Omega_1 \to 2\pi$. In this limit $\alpha_1 \to 1/2$, and it is plausible that $u(\mathbf{x}) \to u_{\text{k.e.}}(\mathbf{x})$, where $u_{\text{k.e.}}$ is the solution to the following "knife edge" diffraction problem: where $\Gamma_{\text{k.e.}} := \{(x_1, 0) : x_1 \geq 0\}$, given the incident plane wave u^i, find the total field $u_{\text{k.e.}} \in C^2(\mathbb{R}^2 \setminus \Gamma_{\text{k.e.}}) \cap C(\mathbb{R}^2)$ such that $\Delta u_{\text{k.e.}} + k^2 u_{\text{k.e.}} = 0$ in $\mathbb{R}^2 \setminus \Gamma_{\text{k.e.}}$, $u_{\text{k.e.}} = 0$ on $\Gamma_{\text{k.e.}}$, and $u_{\text{k.e.}} - u^i$ has the correct radiating behavior. The solution to this problem which satisfies the physically correct radiation condition is given by [10, equation (8.24)]. This solution implies that $\varphi(s) := \frac{1}{k}\frac{\partial u_{\text{k.e.}}}{\partial \mathbf{n}}((s, 0)) = \pm \mathrm{e}^{\mathrm{i}ks} v(s)$, where the $+/-$ sign is taken on the upper/lower surface of the knife edge and $v(s) := \hat{c}(ks)^{-1/2}$, where $\hat{c} = \mathrm{e}^{-\mathrm{i}\pi/4}\sqrt{2/\pi}$. The function $v(s)$ and its derivatives satisfy the bounds on v_1^+ in Theorem 3.2 and Corollary 3.4 (with $\alpha_j = 1/2$), but do not satisfy any sharper bounds in terms of dependence on k or $s - \tilde{L}_{j-1}$.

4. The approximation space. Our aim now is to use the regularity results of section 3 to design an optimal approximation space for the numerical solution of (2.9). We begin by rewriting (2.9) in parametric form. Defining, for $j = 1, \ldots, n$,

$$a_j := \frac{p_{j+1} - p_j}{L_j}, \quad b_j := \frac{q_{j+1} - q_j}{L_j}, \quad c_j := p_j - a_j \tilde{L}_{j-1}, \quad d_j := q_j - b_j \tilde{L}_{j-1},$$

and noting that $n_{j1} = b_j$, $n_{j2} = -a_j$, we can rewrite (2.9) as

$$\phi(s) + \int_0^L \kappa(s, t)\phi(t)\,\mathrm{d}t = f(s), \quad s \in [0, L], \tag{4.1}$$

where, for $\mathbf{x}(s) \in \Gamma_l$, $\mathbf{y}(t) \in \Gamma_j$, i.e., for $s \in (\tilde{L}_{l-1}, \tilde{L}_l)$, $t \in (\tilde{L}_{j-1}, \tilde{L}_j)$,

$$\kappa(s, t) := -\frac{1}{2}\left[\eta H_0^{(1)}(kR) + \mathrm{i}k\left[(a_l b_j - b_l a_j)t + b_l(c_l - c_j) - a_l(d_l - d_j)\right]\frac{H_1^{(1)}(kR)}{R}\right],$$

with $R = R(s,t) := \sqrt{(a_l s - a_j t + c_l - c_j)^2 + (b_l s - b_j t + d_l - d_j)^2}$ and $f \in L^2(0,L)$ defined by

$$f(s) := 2\mathrm{i}[b_l \sin\theta + a_l \cos\theta + (\eta/k)]\mathrm{e}^{\mathrm{i}k((a_l s + c_l)\sin\theta - (b_l s + d_l)\cos\theta)}.$$

The first step in our numerical method is to separate off the explicitly known leading order behavior, the physical optics approximation $\Psi(s)$. Thus we introduce a new unknown,

$$\varphi := \phi - \Psi \in L^2(0,L). \tag{4.2}$$

Substituting into (4.1) we have

$$\varphi + K\varphi = F, \tag{4.3}$$

where the integral operator $K : L^2(0,L) \to L^2(0,L)$ and $F \in L^2(0,L)$ are defined by

$$K\psi(s) := \int_0^L \kappa(s,t)\psi(t)\,\mathrm{d}t,\ 0 \le s \le L, \quad F := f - \Psi - K\Psi.$$

Equation (4.3) is the integral equation we will solve numerically. By Theorem 2.7, (4.3) has a unique solution in $L^2(0,L)$ and $\|(I+K)^{-1}\|_2 = C_S$, where C_S is defined in (2.11) and I is the identity operator on $L^2(0,L)$.

We will design an approximation space to represent φ based on (3.9). The novelty of the scheme we propose is that on each side Γ_j, $j = 1,\dots,n$, of the polygon, we approximate $v_j^{\pm}$ by conventional piecewise polynomials, rather than approximating φ itself. This makes sense since, as quantified by Theorem 3.2, the functions $v_j^{\pm}$ are smooth (their higher order derivatives are small) away from the corners P_j and P_{j+1}. To approximate $v_j^{\pm}$ we use piecewise polynomials of a fixed degree $\nu \ge 0$ on a graded mesh, the mesh grading adapted in an optimal way to the bounds of Theorems 3.2 and 3.3. In [19] the 2D problem of scattering of a plane wave by a straight boundary of piecewise constant surface impedance was considered. We will construct a similar mesh on each side of the polygon as was used on each interval of constant impedance in [19], except that we use a different grading near the corners, with the grading near each corner dependent on the angle at that corner.

To construct this mesh we choose a constant $c^* > 0$ (we take $c^* = 2\pi$ in the numerical examples in section 6) and set $\lambda^* := c^*/k$. Next, for every $A > \lambda^*$, we define a composite graded mesh on $[0,A]$, with a polynomial grading on $[0,\lambda^*]$ and a geometric grading on $[\lambda^*, A]$ (note that the mesh on $[0,\lambda^*]$ is similar to that classically used near corners (e.g., [17, 7]) for solving Laplace's equation on polygonal domains).

DEFINITION 4.1. *For $A > \lambda^*$, $N = 2,3,\dots$, $\Lambda_{N,A,q} := \{y_0,\dots,y_{N+N_{A,q}}\}$ is the mesh consisting of the points*

$$y_i = \lambda^*\left(\frac{i}{N}\right)^q,\ i = 0,\dots,N, \quad \text{and} \quad y_{N+j} := \lambda^*\left(\frac{A}{\lambda^*}\right)^{j/N_{A,q}},\ j = 1,\dots,N_{A,q}, \tag{4.4}$$

where $N_{A,q} := \lceil N^* \rceil$, i.e., $N_{A,q}$ is the smallest integer greater than or equal to N^*, and

$$N^* := \frac{-\log(A/\lambda^*)}{q\log(1-1/N)}.$$

Let us explain the rationale behind this definition. Having the bounds of Theorems 3.2 and 3.3 in mind, the mesh on $[0, \lambda^*]$ is chosen to be approximately optimal if q is chosen appropriately (see Theorem 4.2 below), in terms of equidistributing the error between the subintervals of the mesh when $s^{-\alpha}$, with $0 < \alpha < 1/2$, is approximated on $[0, \lambda^*]$ in the L^2 norm. That the mesh we propose on $[0, \lambda^*]$ has this property and the appropriate choice of q as a function of α is well known and dates back to Rice [55]. Similarly, the mesh on $[\lambda^*, A]$ is chosen to be approximately optimal, in terms of equidistributing the error between the subintervals of the mesh, when $s^{-1/2}$ is approximated on $[\lambda^*, A]$ in the L^2 norm. Finally, the choice of N^* ensures a smooth transition between the two parts of the mesh, and thus approximately the same L^2 error in the two adjacent subintervals on either side of λ^*. In particular, in the case that $N_{A,q} = N^*$, it holds that $y_{N+1}/y_N = y_N/y_{N-1}$, so that y_{N-1} and y_N are points in both the polynomial and the geometric parts of the mesh. Note that by the mean value theorem, $-\log(1 - 1/N) = 1/(\xi N)$ for some $\xi \in (1 - 1/N, 1)$, and hence

$$N_{A,q} < \frac{N \log(kA/c^*)}{q} + 1. \tag{4.5}$$

For $a < b$ let $\|\cdot\|_{2,(a,b)}$ denote the norm on $L^2(a, b)$, $\|f\|_{2,(a,b)} := \{\int_a^b |f(s)|^2 \mathrm{d}s\}^{1/2}$. Similarly, for $f \in C[a, b]$, let $\|f\|_{\infty,(a,b)} := \sup_{a<s<b} |f(s)|$. For $A > \lambda^*$, $\nu \in \mathbb{N} \cup \{0\}$, $q \geq 1$, let $\Pi_{N,\nu} \subset L^2(0, A)$ denote the set of piecewise polynomials

$$\Pi_{N,\nu} := \{\sigma : \sigma|_{(y_{j-1},y_j)} \text{ is a polynomial of degree} \leq \nu \text{ for } j = 1, \ldots, N + N_{A,q}\},$$

and let P_N^* be the orthogonal projection operator from $L^2(0, A)$ to $\Pi_{N,\nu}$, so that setting $p = P_N^* f$ minimizes $\|f - p\|_{2,(0,A)}$ over all $p \in \Pi_{N,\nu}$.

THEOREM 4.2. *Suppose that $f \in C^\infty(0, \infty)$, $kA > c^*$, and $\alpha \in (0, 1/2)$, and that for $m = 0, 1, 2, \ldots$, there exist constants $c_m > 0$ such that*

$$|f^{(m)}(s)| \leq \begin{cases} c_m k^m (ks)^{-\alpha-m}, & ks \leq 1, \\ c_m k^m (ks)^{-1/2-m}, & ks \geq 1. \end{cases} \tag{4.6}$$

Then, with the choice $q := (2\nu + 3)/(1 - 2\alpha)$, there exists a constant C_ν, dependent only on c^*, ν, and α, such that for $N = 2, 3, \ldots$,

$$\|f - P_N^* f\|_{2,(0,A)} \leq \frac{C_\nu \tilde{c}_\nu (1 + \log(kA/c^*))^{1/2}}{k^{1/2} N^{\nu+1}},$$

where $\tilde{c}_\nu := \max(c_0, c_{\nu+1})$.

Proof. Throughout the proof let C_ν denote a positive constant whose value depends on ν, c^*, and α, not necessarily the same at each occurrence. For $0 \leq a < b \leq A$, let $p_{a,b,\nu}$ denote the polynomial of degree $\leq \nu$ which is the best approximation to f in the L^2 norm on (a, b). Then it follows from Taylor's theorem that

$$\|f - p_{a,b,\nu}\|_{2,(a,b)} \leq C_\nu (b - a)^{\nu+3/2} \|f^{(\nu+1)}\|_{\infty,(a,b)}. \tag{4.7}$$

Now

$$\|f - P_N^* f\|_{2,(0,A)}^2 = \sum_{j=1}^{N+N_{A,q}} \int_{y_{j-1}}^{y_j} |f - P_N^* f|^2 \, \mathrm{d}s = \sum_{j=1}^{N+N_{A,q}} e_j, \tag{4.8}$$

where $e_j := \|f - p_{y_{j-1},y_j,\nu}\|^2_{2,(y_{j-1},y_j)}$. From the definition (4.4) we see that

$$e_1 \le \int_0^{y_1} |f(s)|^2 \,\mathrm{d}s \le c_0^2 k^{-2\alpha} \int_0^{\lambda^*/N^q} s^{-2\alpha} \,\mathrm{d}s \le \frac{C_\nu c_0^2}{kN^{2\nu+3}}. \tag{4.9}$$

Using (4.7) we have, for $j = 2, 3, \ldots, N + N_{A,q}$,

$$e_j \le C_\nu (y_j - y_{j-1})^{2\nu+3} \|f^{(\nu+1)}\|^2_{\infty,(y_{j-1},y_j)}. \tag{4.10}$$

Further, for $j = 2, \ldots, N$,

$$y_j - y_{j-1} = \frac{c^*}{kN^q}\,[j^q - (j-1)^q] \le \frac{c^* q j^{q-1}}{kN^q}, \tag{4.11}$$

and, using (4.6) and since $N/(j-1) \le 2N/j$,

$$\|f^{(\nu+1)}\|_{\infty,(y_{j-1},y_j)} \le c_{\nu+1} k^{-\alpha} y_{j-1}^{-\alpha-\nu-1} \le c_{\nu+1} k^{\nu+1} \left(\frac{2N}{j}\right)^{q(\alpha+\nu+1)}. \tag{4.12}$$

Combining (4.10)–(4.12) we see that for $j = 2, \ldots, N$,

$$e_j \le \frac{C_\nu c_{\nu+1}^2}{kN^{2\nu+3}}. \tag{4.13}$$

For $j = N+1, \ldots, N_{A,q}$, recalling (4.4) and the choice of N^* and then using (4.11),

$$y_j - y_{j-1} = y_{j-1}\left(\frac{y_j - y_{j-1}}{y_{j-1}}\right) \le y_{j-1}\left(\frac{y_N - y_{N-1}}{y_{N-1}}\right) \le y_{j-1}\frac{q}{N-1} \le 2y_{j-1}\frac{q}{N}.$$

Also, from (4.6),

$$\|f^{(\nu+1)}\|_{\infty,(y_{j-1},y_j)} \le c_{\nu+1} k^{-1/2} y_{j-1}^{-\nu-3/2}.$$

Using these bounds in (4.10), we see that the bound (4.13) holds also for $j = N + 1, \ldots, N + N_{A,q}$. Combining (4.8), (4.9), and (4.13),

$$\|f - P_N^* f\|^2_{2,(0,A)} \le \frac{C_\nu \tilde{c}_\nu^2 (N + N_{A,q})}{kN^{2\nu+3}} \le \frac{C_\nu \tilde{c}_\nu^2 (1 + \log(kA/c^*))}{kN^{2\nu+2}},$$

using (4.5). Hence the result follows. ∎

We assume through the remainder of the paper that $c^* > 0$ is chosen so that

$$kL_j \ge c^*, \quad j = 1, \ldots, n. \tag{4.14}$$

For $j = 1, \ldots, n$, recalling (3.11), we define $q_j := (2\nu+3)/(1-2\alpha_j)$ and the two meshes

$$\Gamma_j^+ := \tilde{L}_{j-1} + \Lambda_{N,L_j,q_j}, \quad \Gamma_j^- := \tilde{L}_j - \Lambda_{N,L_j,q_{j+1}}.$$

Letting $\mathrm{e}_\pm(s) := \mathrm{e}^{\pm \mathrm{i}ks}$, $s \in [0, L]$, we then define

$$V_{\Gamma_j^+,\nu} := \{\sigma \mathrm{e}_+ : \sigma \in \Pi_{\Gamma_j^+,\nu}\}, \quad V_{\Gamma_j^-,\nu} := \{\sigma \mathrm{e}_- : \sigma \in \Pi_{\Gamma_j^-,\nu}\}$$

for $j = 1, \ldots, n$, where

$$\Pi_{\Gamma_j^+,\nu} := \{\sigma \in L^2(0,L) : \sigma|_{(\tilde{L}_{j-1}+y_{m-1},\tilde{L}_{j-1}+y_m)} \text{ is a polynomial of degree } \le \nu \text{ for } m = 1, \ldots, N + N_{L_j,q_j}, \text{ and } \sigma|_{(0,\tilde{L}_{j-1})\cup(\tilde{L}_j,L)} = 0\},$$

$$\Pi_{\Gamma_j^-,\nu} := \{\sigma \in L^2(0,L) : \sigma|_{(\tilde{L}_j-\tilde{y}_m,\tilde{L}_j-\tilde{y}_{m-1})} \text{ is a polynomial of degree } \le \nu \text{ for } m = 1, \ldots, N + N_{L_j,q_{j+1}}, \text{ and } \sigma|_{(0,\tilde{L}_{j-1})\cup(\tilde{L}_j,L)} = 0\},$$

with $0 = y_0 < y_1 < \cdots < y_{N+N_{L_j,q_j}} = L_j$ the points of the mesh Λ_{N,L_j,q_j} and $0 = \tilde{y}_0 < \tilde{y}_1 < \cdots < \tilde{y}_{N+N_{L_j,q_{j+1}}} = L_j$ the points of the mesh $\Lambda_{N,L_j,q_{j+1}}$. We define P_N^+ and P_N^- to be the orthogonal projection operators from $L^2(0,L)$ onto $\Pi_{\Gamma_+,\nu}$ and $\Pi_{\Gamma_-,\nu}$, respectively, where $\Pi_{\Gamma_\pm,\nu}$ denotes the linear span of $\bigcup_{j=1,\ldots,n} \Pi_{\Gamma_j^\pm,\nu}$. We also define the functions $v_\pm \in L^2(0,L)$ by

$$v_+(s) := v_j^+(s), \quad v_-(s) := v_j^-(s), \quad \tilde{L}_{j-1} < s < \tilde{L}_j, \ j = 1, \ldots, n.$$

We then have the following error estimate, in which u_M is as defined in (3.10) and we abbreviate $\|\cdot\|_{2,(0,L)}$ by $\|\cdot\|_2$.

THEOREM 4.3. *There exists a constant $C_\nu > 0$, dependent only on c^*, ν, and Ω_1, $\Omega_2, \ldots, \Omega_n$, such that*

$$\|v_+ - P_N^+ v_+\|_2 \le C_\nu u_M \frac{n^{1/2}(1 + \log(k\bar{L}/c^*))^{1/2}}{k^{1/2} N^{\nu+1}},$$

where $\bar{L} := (L_1 \ldots L_n)^{1/n}$, with an identical bound holding on $\|v_- - P_N^- v_-\|_2$.

Proof. From Theorem 3.2, Corollary 3.4, and Theorem 4.2,

$$\|v_+ - P_N^+ v_+\|_2^2 = \sum_{j=1}^n \|v_j^+ - P_N^+ v_j^+\|_{2,(\tilde{L}_{j-1},\tilde{L}_j)}^2 \le n \frac{C_\nu^2 u_M^2 (1 + \log(k\bar{L}))}{k N^{2\nu+2}},$$

and the result follows. ∎

Our approximation space $V_{\Gamma,\nu}$ is the linear span of

$$\bigcup_{j=1,\ldots,n} \{V_{\Gamma_j^+,\nu} \cup V_{\Gamma_j^-,\nu}\}.$$

The dimension of this approximation space, i.e., the number of degrees of freedom, is

$$M_N = 2(\nu+1) \sum_{j=1}^n (N + N_{L_j,q_j}) < 2(\nu+1) n N (1 + N^{-1} + \log(k\bar{L}/c^*)) \tag{4.15}$$

by (4.5). We define P_N to be the operator of orthogonal projection from $L^2(0,L)$ onto $V_{\Gamma,\nu}$. It remains to prove a bound on $\|\varphi - P_N\varphi\|_2$, showing that our mesh and approximation space are well adapted to approximating φ.

To use Theorem 4.3 we note from (3.9) and (4.2) that $\varphi = \frac{\mathrm{i}}{2}(\mathrm{e}_+ v_+ + \mathrm{e}_- v_-)$. But $\mathrm{e}_+ P_N^+ v_+ + \mathrm{e}_- P_N^- v_- \in V_{\Gamma,\nu}$ and $P_N\varphi$ is the best approximation to φ in $V_{\Gamma,\nu}$. Applying

Theorem 4.3 we thus have that

$$\begin{aligned}\|\varphi - P_N\varphi\|_2 &\le \left\|\varphi - \frac{\mathrm{i}}{2}(e_+P_N^+v_+ + e_-P_N^-v_-)\right\|_2 \\ &= \frac{1}{2}\|e_+(v_+ - P_N^+v_+) + e_-(v_- - P_N^-)\|_2 \\ &\le \|e_+\|_\infty\|v_+ - P_N^+v_+\|_2 + \|e_-\|_\infty\|v_- - P_N^-v_-\|_2 \\ &\le C_\nu u_M \frac{n^{1/2}(1+\log^{1/2}(k\bar{L}))}{k^{1/2}N^{\nu+1}}.\end{aligned}$$

Combining this bound with (4.15), we obtain the following main result of the paper. We remind the reader that we are assuming throughout that (4.14) holds.

THEOREM 4.4. *There exist positive constants C_ν and C'_ν, depending only on c^*, ν, and Ω_1, $\Omega_2, \ldots, \Omega_n$, such that*

$$k^{1/2}\|\varphi - P_N\varphi\|_2 \le C_\nu u_M \frac{n^{1/2}(1+\log(k\bar{L}/c^*))^{1/2}}{N^{\nu+1}} \le C'_\nu u_M \frac{(n[1+\log(k\bar{L}/c^*)])^{\nu+3/2}}{M_N^{\nu+1}}.$$

A comment on the factor $k^{1/2}$ on the left-hand side is probably helpful. Reflecting that the solution of the physical problem must be independent of the unit of length measurement and that we are designing our numerical scheme to preserve this property, it is easy to see that the values of both $k^{1/2}\|\varphi\|_2$ and $k^{1/2}\|\varphi - P_N\varphi\|_2$ remain fixed as k changes if we keep kL_j fixed for $j = 1, \ldots, n$ (and also, of course, keep Ω_j, $j = 1, \ldots, n$, c^*, and ν fixed). Thus inclusion of the factor $k^{1/2}$ ensures that the value of $k^{1/2}\|\varphi - P_N\varphi\|_2$ is independent of the unit of length measurement as are the bounds on the right-hand side.

5. Galerkin method. Theorem 4.4 has shown that it is possible to approximate accurately the solution of the integral equation (4.3) with a number of degrees of freedom that grows only very modestly as the wave number increases. To select an approximation, φ_N, from the approximation space $V_{\Gamma,\nu}$ we use the Galerkin method. Let $(\cdot,\cdot)$ denote the usual inner product on $L^2(0,L)$, defined by $(\chi_1,\chi_2) := \int_0^L \chi_1(s)\bar{\chi}_2(s)\,\mathrm{d}s$, so that $\|\chi\|_2 = (\chi,\chi)^{1/2}$. Then our Galerkin method approximation $\varphi_N \in V_{\Gamma,\nu}$ is defined by

$$(\varphi_N,\rho) + (K\varphi_N,\rho) = (F,\rho) \quad \text{for all } \rho \in V_{\Gamma,\nu}; \tag{5.1}$$

equivalently

$$\varphi_N + P_NK\varphi_N = P_NF. \tag{5.2}$$

Our goal now is to show that (5.2) has a unique solution φ_N, to establish a bound on the error $\|\varphi - \varphi_N\|_2$ in this numerical method, and to relate this error to the best approximation error $\|\varphi - P_N\varphi\|_2$. We begin by establishing that $I + P_NK$ is invertible if N is large enough. We remind the reader (see the end of section 2) that we are assuming that $\eta \in \mathbb{R}$, the coupling parameter in the integral equation, is chosen with $\eta \neq 0$, which ensures that $I + K$ is invertible.

THEOREM 5.1. *For all $v \in L^2(0,L)$, $\|P_Nv - v\|_2 \to 0$ as $N \to \infty$.*

Proof. Since $\|P_N\|_2 = 1$, it is enough to show that $P_Nv \to v$ in $L^2(0,L)$ for all $v \in C^\infty[0,L]$, a dense subset of $L^2(0,L)$. But this follows from Theorem 4.2 and the definition of P_N. ∎

THEOREM 5.2. *There exists a constant $N^* \geq 2$, dependent only on Γ, k, and η, such that, for $N \geq N^*$, the operator $I + P_N K : L^2(0, L) \to L^2(0, L)$ is bijective with*

$$C_s := \sup_{N \geq N^*} \|(I + P_N K)^{-1}\|_2 < \infty, \tag{5.3}$$

so that (5.2) *has exactly one solution for $N \geq N^*$.*

Proof. Recalling the discussion at the end of section 2, we note that it holds that $K = K_1 + K_2$, where $\|K_1\|_2 < 1$ and K_2 is a compact operator on $L^2(0, L)$. Since $\|P_N K_1\|_2 \leq \|K_1\|_2 < 1$, $I + P_N K_1$ is invertible and $\|(I + P_N K_1)^{-1}\|_2 \leq (1 - \|K_1\|_2)^{-1}$. Since K_2 is compact and $I + K$ is injective, it follows from Theorem 5.1 and standard perturbation arguments for projection methods (e.g., [7, Theorem 8.2.1], [17]) that $(I + P_N K)^{-1}$ exists and is uniformly bounded for all N sufficiently large. ∎

From (4.3) and (5.2) it follows that $\varphi - \varphi_N = (I + P_N K)^{-1}(\varphi - P_N \varphi)$, and hence

$$\|\varphi - \varphi_N\|_2 \leq \|(I + P_N K)^{-1}\|_2 \|\varphi - P_N \varphi\|_2. \tag{5.4}$$

Combining (5.3) and (5.4) with Theorem 4.4, we obtain our final error estimate.

THEOREM 5.3. *There exist positive constants C_ν and C'_ν, depending only on c^*, ν, and Ω_1, $\Omega_2, \ldots, \Omega_n$, such that*

$$\begin{aligned} k^{1/2}\|\varphi - \varphi_N\|_2 &\leq C_s C_\nu u_M \frac{n^{1/2}(1 + \log(k\bar{L}/c^*))^{1/2}}{N^{\nu+1}} \\ &\leq C_s C'_\nu u_M \frac{(n[1 + \log(k\bar{L}/c^*)])^{\nu+3/2}}{M_N^{\nu+1}} \end{aligned} \tag{5.5}$$

for $N \geq N^*$, where N^* and C_s are as defined in Theorem 5.2.

Note that we will take $c^* = 2\pi$ and $\eta = k$ in all our numerical calculations in the next section. Note also that, while the constants C_ν and C'_ν, from the best approximation theorem, Theorem 4.4, depend only on c^*, ν, and the corner angles of Γ, the numbers N^* and C_s depend additionally on k, L_1, $L_2, \ldots, L_n$, and η. We do not attempt the difficult task of elucidating this dependence in this paper. We note only that, very recently, for the boundary integral equation formulation (2.9) applied to scattering by a circle, Domínguez, Graham, and Smyshlyaev [28] have shown that $\mathcal{I} + \mathcal{K}$ is elliptic if $\eta = \pm k$ and k is sufficiently large, so that every Galerkin method is automatically stable; specifically, (5.3) holds for every N^* if P_N is the orthogonal projection from $L^2(0, L)$ onto the Galerkin approximation space. Further it follows from results in [28] that, at worst, $C_s = O(k^{1/3})$ as $k \to \infty$ in the circle case. Our numerical results in section 6 will suggest the stronger result that for our particular scheme and geometry, the bound of Theorem 5.3 holds with a constant C_s independent of k. We recall from section 2 (2.12) that it has been shown that the corresponding continuous continuity constant $C_S = O(1)$ as $k \to \infty$ if the choice $\eta = k$ is made.

Of course our aim in approximating φ by φ_N is to approximate $\partial_{\mathbf{n}}^+ u$ and hence, via (2.7), the solution u of the scattering problem. Clearly, from (3.8) and (4.2), an approximation to $\partial u / \partial \mathbf{n}$ is

$$\frac{\partial u}{\partial \mathbf{n}}(\mathbf{x}(s)) \approx k(\Psi(s) + \varphi_N(s)), \quad 0 \leq s \leq L.$$

Using this approximation in (2.7), we conclude that

$$u(\mathbf{x}) \approx u_N(\mathbf{x}) := u^i(\mathbf{x}) - k \int_0^L \Phi(\mathbf{x}, \mathbf{x}(s))[\Psi(s) + \varphi_N(s)]\, \mathrm{d}s, \quad \mathbf{x} \in D. \tag{5.6}$$

Theorem 5.3 implies the following error estimate.

THEOREM 5.4. *There exist positive constants C_ν and C'_ν, depending only on c^*, ν, and Ω_1, $\Omega_2, \ldots, \Omega_n$, such that*

$$\frac{\sup_{\mathbf{x}\in D}|u(\mathbf{x})-u_N(\mathbf{x})|}{\sup_{\mathbf{x}\in D}|u(\mathbf{x})|} \le C_s C_\nu \frac{n(1+\log(k\bar{L}/c^*))}{N^{\nu+1}} \le C_s C'_\nu \frac{(n[1+\log(k\bar{L}/c^*)])^{\nu+2}}{M_N^{\nu+1}}$$

for $N \ge N^*$, where N^* and C_s are as defined in Theorem 5.2.

Proof. From (2.7) and (5.6),

$$\begin{aligned}
|u(\mathbf{x})-u_N(\mathbf{x})| &= k\left|\int_0^L \Phi(\mathbf{x},\mathbf{x}(s))\,[\varphi(s)-\varphi_N(s)]\,\mathrm{d}s\right| \\
&\le \frac{k}{4}\left\{\int_0^L |H_0^{(1)}(k|\mathbf{x}-\mathbf{x}(s)|)|^2\,\mathrm{d}s\right\}^{1/2} \|\varphi-\varphi_N\|_2 \\
&\le \frac{k}{4}\left\{2\sum_{j=1}^{n}\int_0^{L_j/2} |H_0^{(1)}(kt)|^2\,\mathrm{d}t\right\}^{1/2} \|\varphi-\varphi_N\|_2 \\
&\le C_\nu k^{1/2} n^{1/2}(1+\log(k\bar{L}/c^*))^{1/2}\|\varphi-\varphi_N\|_2,
\end{aligned}$$

where we have used that $|H_0^{(1)}(t)|$ is a monotonic decreasing function of t on $(0,\infty)$ and that $|H_0^{(1)}(t)| = O(t^{-1/2})$ as $t \to \infty$ (see e.g., [2]). The result now follows from Theorem 5.3. ∎

6. Numerical results. There has been much work on the optimal choice of the parameter η in (2.9) (see, e.g., [3, 43]). Here we choose $\eta = k$ as in [28]. We also set $c^* = 2\pi$ and restrict attention to the case $\nu = 0$. For higher values of ν the implementation of the scheme is similar. Note that, with $c^* = 2\pi$ and $\nu = 0$, there are approximately N degrees of freedom used to represent the solution in the first wavelength on each side adjacent to a corner.

The equation we wish to solve is (5.1) with $\nu = 0$. Writing φ_N as a linear combination of the basis functions of $V_{\Gamma,0}$, we have

$$\varphi_N(s) = \sum_{j=1}^{M_N} v_j \rho_j(s), \quad 0 \le s \le L,$$

where ρ_j is the jth basis function and M_N is the dimension of $V_{\Gamma,0}$. For $p = 1, \ldots, n$, where n is the number of sides of the polygon, we define $n_p^\pm$ to be the number of points in the mesh $\Gamma_p^\pm$, so that $n_p^+ = N + N_{L_p,q_p}$, $n_p^- = N + N_{L_p,q_{p+1}}$, and we denote the points of the mesh $\Gamma_p^\pm$ by $s_{p,l}^\pm$ for $l = 1, \ldots, n_p^\pm$, with $s_{p,1}^\pm < \cdots < s_{p,n_p^\pm}^\pm$. Setting $n_1 := 0$, $n_p := \sum_{j=1}^{p-1}(n_j^+ + n_j^-)$ for $p = 2, \ldots, n-1$, we define, for $p = 1, \ldots, n$,

$$\rho_{n_p+j}(s) := \begin{cases} \mathrm{e}^{\mathrm{i}ks}\chi_{(s^+_{p,j-1},s^+_{p,j})}(s)/\sqrt{s^+_{p,j}-s^+_{p,j-1}}, & j = 1, \ldots, n_p^+, \\ \mathrm{e}^{-\mathrm{i}ks}\chi_{(s^-_{p,j-1},s^-_{p,j})}(s)/\sqrt{s^-_{p,j}-s^-_{p,j-1}}, & j = n_p^+ + 1, \ldots, n_p^+ + n_p^-, \end{cases}$$

where $\chi_{(y_1,y_2)}$ denotes the characteristic function of the interval (y_1, y_2). From (4.15), $M_N = \sum_{j=1}^{n}(n_j^+ + n_j^-) < 2nN(3/2 + \log(k\bar{L}/c^*))$.

Equation (5.1) with $\nu = 0$ is equivalent to the linear system

$$\sum_{j=1}^{M_N} v_j((\rho_j, \rho_m) + (K\rho_j, \rho_m)) = (F, \rho_m), \quad m = 1, \ldots, M_N. \tag{6.1}$$

In order to set up this linear system we need to determine the integrals (ρ_j, ρ_m), $(K\rho_j, \rho_m)$, and (F, ρ_m). We note that many of the integrals $(K\rho_j, \rho_m)$ and (F, ρ_m) are highly oscillatory; in particular, all these integrals become highly oscillatory in the limit as $k \to \infty$ with N fixed. The efficient calculation of these integrals is an aspect of the numerical scheme which requires further research, as discussed in section 1.2. But note that explicit formulae for the analytic evaluation of some of these integrals, and a consideration of the quadrature techniques required to evaluate the rest of them numerically, are presented in [44].

Another important issue is the conditioning of the linear system. Standard analysis of the Galerkin method for second kind equations [7] implies that, where $M := [(\rho_j, \rho_m)]$ is the mass matrix (which is necessarily Hermitian and positive definite) and $A = [(\rho_j, \rho_m) + (K\rho_j, \rho_m)]$ is the whole matrix, it holds that $\mathrm{cond}_2 A \le C_s \mathrm{cond}_2 M$, where C_s is defined by (5.3). Thus Theorem 5.2 implies that $\mathrm{cond}_2 A$ is bounded as $N \to \infty$ if the mass matrix is well conditioned. Unfortunately, it appears that, as $N \to \infty$ with k fixed, M must ultimately become badly conditioned. However, the results below will show only moderate condition numbers of A even for large values of N (see Table 6.1). More positively, in the limit as $k \to \infty$ with N fixed, $\mathrm{cond}_2 M \to 1$. To see this we observe that if (ρ_j, ρ_m) is a nonzero off-diagonal element of the mass matrix (in which case the supports of ρ_j and ρ_m are overlapping subintervals of the meshes Γ_p^+ and Γ_p^- for some side p), it holds that $|(\rho_j, \rho_m)| = |\sin(ko)|\sqrt{o/(kS_1S_2)}$, where S_1 and S_2 are the lengths of the two subintervals, and o is the length of the overlap.

As a numerical example, we consider the problem of scattering by a square with sides of length 2π. In this case $n = 4$ and $\Omega_j = 3\pi/2$, $j = 1, 2, 3, 4$. The corners of the square are $P_1 := (0, 0)$, $P_2 := (2\pi, 0)$, $P_3 := (2\pi, 2\pi)$, $P_4 := (0, 2\pi)$, and the incident angle is $\theta = \pi/4$, so the incident field is directed towards P_4, with P_2 in shadow (as shown in Figure 6.1, where the total acoustic field is plotted for $k = 10$).

In Figure 6.2 we plot $|\varphi_N(s)|$ against s for $k = 10$ and $N = 4, 16, 64, 256$. As we expect, $|\varphi_N(s)|$ is highly peaked at the corners of the polygon, $s = 0$, 2π, 4π, 6π and 8π (which is the same corner as $s = 0$), where $\varphi(s)$ is infinite. Except at these corners, $|\varphi_N(s)|$ appears to be converging pointwise as N increases. (We do not plot $\varphi_N(s)$ itself, which is highly oscillatory.)

In order to test the convergence of our scheme, we take the "exact" solution to be that computed with a large number of degrees of freedom, namely with $N = 256$. For $k = 5$ and $k = 320$ the relative L^2 errors $\|\varphi_N - \varphi_{256}\|_2 / \|\varphi_{256}\|_2$ are shown in Table 6.1 (all L^2 norms are computed by approximating by discrete L^2 norms, sampling at 100000 evenly spaced points around the boundary of the square). For this example, Theorem 5.3 predicts that for $N \ge N^*$, $\|\varphi - \varphi_N\|_2 \le CN^{-1}$, where C is a constant. Thus Theorem 5.3 predicts that for $N > N^*$, the average rate of convergence is

$$EOC := \frac{\log(\|\varphi - \varphi_N\|_2 / \|\varphi - \varphi_{N^*}\|_2)}{\log(N/N^*)} \ge 1 - \frac{\hat{C}}{\log(N/N^*)} \sim 1$$

as $N \to \infty$, where $\hat{C} := \log(\|\varphi - \varphi_N\|_2 / C)$. This behavior is clearly seen in the EOC

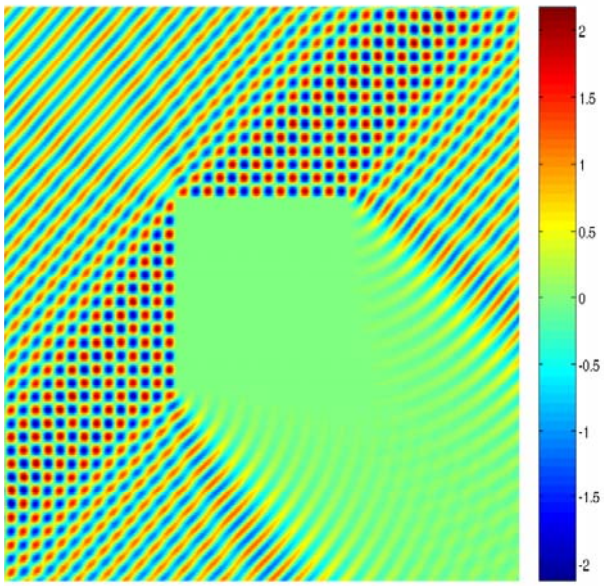


FIG. 6.1. *Total acoustic field, scattering by a square, $k = 10$. Incident field is directed from the top left corner towards the bottom right corner.*

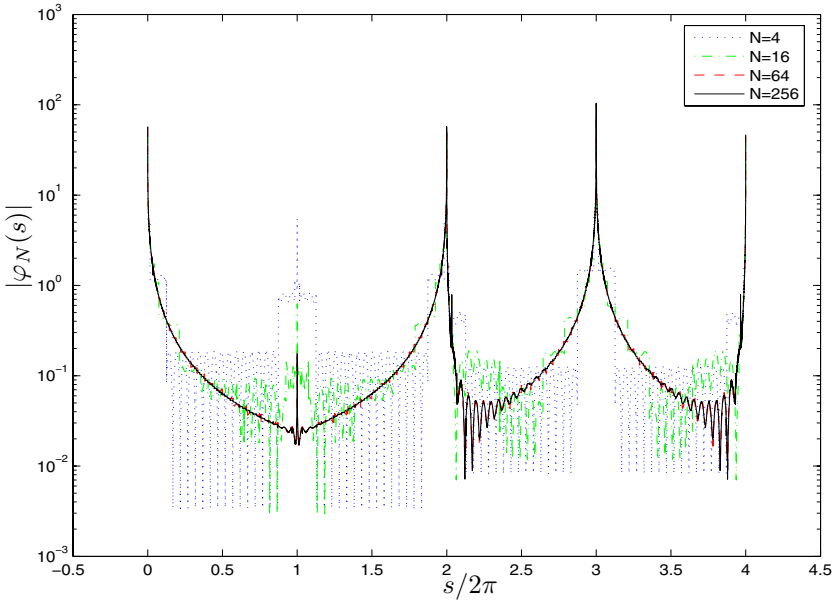


FIG. 6.2. *$|\varphi_N(s)|$ plotted against s, various N, for scattering by a square of side length ten wavelengths.*

values (defined with $N^* = 8$) in Table 6.1 for both values of k. We also show in Table 6.1 the 2 norm condition number, $\mathrm{cond}_2 A$, of the matrix $A = [(\rho_j, \rho_m) + (K\rho_j, \rho_m)]$ for each example. Unlike methods where the approximation space is formed by multiplying standard finite element basis functions by many plane waves travelling in a large number of directions [27, 53, 37], the condition number does not grow signifi-

TABLE 6.1
Errors and relative L^2 errors, various N, $k = 5$, and $k = 320$.

k	N	M_N	$k^{1/2}\|\varphi_N - \varphi_{256}\|_2$	$\|\varphi_N - \varphi_{256}\|_2/\|\varphi_{256}\|_2$	*EOC*	$\text{cond}_2 A$
5	8	88	5.7339×10^{-1}	2.4426×10^{-1}		9.5×10^{0}
	16	176	3.7454×10^{-1}	1.5955×10^{-1}	0.6	4.6×10^{1}
	32	360	1.6176×10^{-1}	6.8909×10^{-2}	0.9	2.6×10^{1}
	64	712	7.7267×10^{-2}	3.2916×10^{-2}	1.0	2.4×10^{2}
	128	1416	3.3541×10^{-2}	1.4289×10^{-2}	1.0	1.5×10^{3}
320	8	120	7.0765×10^{-1}	3.6736×10^{-1}		2.4×10^{2}
	16	240	5.9792×10^{-1}	3.1040×10^{-1}	0.2	6.9×10^{2}
	32	472	1.9668×10^{-1}	1.0211×10^{-1}	0.9	8.1×10^{2}
	64	944	7.5808×10^{-2}	3.9354×10^{-2}	1.1	1.1×10^{3}
	128	1888	4.8814×10^{-2}	2.5341×10^{-2}	1.0	3.8×10^{3}

TABLE 6.2
Errors and relative L^2 errors, various k, $N = 64$.

k	M_N	$k^{1/2}\|\varphi_{64} - \varphi_{256}\|_2$	$\|\varphi_{64} - \varphi_{256}\|_2/\|\varphi_{256}\|_2$	$\text{cond}_2 A$
5	712	7.7267×10^{-2}	3.2916×10^{-2}	2.4×10^{2}
10	752	6.6373×10^{-2}	2.8702×10^{-2}	8.4×10^{1}
20	792	3.8309×10^{-1}	1.6914×10^{-1}	5.1×10^{3}
40	824	1.3162×10^{-1}	5.9856×10^{-2}	1.2×10^{3}
80	864	7.4315×10^{-2}	3.4801×10^{-2}	2.7×10^{3}
160	904	7.0884×10^{-2}	3.4570×10^{-2}	1.4×10^{3}
320	944	7.5808×10^{-2}	3.9354×10^{-2}	1.1×10^{3}
640	984	6.4280×10^{-2}	3.5693×10^{-2}	1.5×10^{3}

TABLE 6.3
Relative errors, $|u_N(\mathbf{x}) - u_{256}(\mathbf{x})|/|u_{256}(\mathbf{x})|$, as a function of N, at three points $\mathbf{x}$.

k	N	$\mathbf{x} = (-\pi, 3\pi)$	$\mathbf{x} = (3\pi, 3\pi)$	$\mathbf{x} = (3\pi, -\pi)$
5	4	1.9587×10^{-2}	1.0071×10^{-3}	1.5885×10^{-2}
	8	4.2629×10^{-3}	2.8031×10^{-3}	2.3215×10^{-3}
	16	3.6284×10^{-4}	3.1410×10^{-4}	1.3513×10^{-3}
	32	6.7523×10^{-5}	2.9803×10^{-5}	1.7939×10^{-5}
	64	1.2675×10^{-5}	5.9626×10^{-6}	4.6158×10^{-6}
320	4	2.2938×10^{-3}	2.9350×10^{-3}	2.0897×10^{-2}
	8	4.3176×10^{-3}	1.5157×10^{-3}	1.1652×10^{-2}
	16	3.3908×10^{-3}	9.6409×10^{-4}	9.3922×10^{-3}
	32	3.3898×10^{-4}	1.6984×10^{-4}	9.0526×10^{-4}
	64	1.0022×10^{-4}	9.6493×10^{-5}	2.6204×10^{-4}

cantly as the number of degrees of freedom increases.

In Table 6.2 we fix $N = 64$ and show $\|\varphi_{64}-\varphi_{256}\|_2/\|\varphi_{256}\|_2$ and $k^{1/2}\|\varphi_{64}-\varphi_{256}\|_2$ for increasing values of k. Both measures of errors remain approximately constant in magnitude as k increases. Recall that, keeping N fixed as k increases corresponds to keeping the number of degrees of freedom per wavelength fixed near each corner and increasing the total number of degrees of freedom, M_N, approximately in proportion to $\log(k\bar{L})$. Thus these results are consistent with the approximation error estimate of Theorem 4.2 which suggests that increasing M_N proportional to $\log^{3/2}(k\bar{L})$ is enough to keep the error bounded; indeed these results are suggestive that the bound (5.5) in the Galerkin error estimate, Theorem 5.3, holds with a constant C_s which is independent of k. Note that the condition number of the coefficient matrix A only increases modestly as k increases, and is approximately constant for $k \geq 40$.

In Table 6.3 we show numerical convergence of the total field $u_N(\mathbf{x})$ at the three

points $\mathbf{x} = (-\pi, 3\pi)$ (illuminated), $\mathbf{x} = (3\pi, 3\pi)$, and $\mathbf{x} = (3\pi, -\pi)$ (shadow), for $k = 5$ and $k = 320$. The errors are consistent with the estimate of Theorem 5.4. As might be expected for the computation of linear functionals of φ_N, the relative errors in Table 6.3 are a lot smaller and converge to zero more rapidly than the relative errors in the computation of the boundary data in Tables 6.1 and 6.2.

7. Conclusions. In this paper we have described a novel Galerkin boundary integral equation method for solving problems of high frequency scattering by convex polygons. In section 2, building on previous results for Lipschitz domains [56, 48, 50, 49], we have shown that the standard second kind boundary integral equations for the exterior Dirichlet problem for the Helmholtz equation are well-posed for general Lipschitz domains in a scale of Sobolev spaces. We have understood very completely in section 3 the oscillatory behavior of the normal derivative of the field on the boundary of the polygon. We have then used this understanding to design an optimal graded mesh for approximation of the diffracted field by products of piecewise polynomials and plane waves. Our error analysis demonstrates that the number of degrees of freedom required to achieve a prescribed level of accuracy using the best approximation to the solution from the approximation space grows only logarithmically with respect to the wave number k as $k \to \infty$. Numerical experiments indicate that the same statement holds for the Galerkin approximation from the same approximation space. However, while we have established that the error in the Galerkin approximation space is bounded by the stability constant C_s times the best approximation error, our Theorem 5.3 holds only for a sufficiently refined mesh and we have not established a bound on C_s which is independent of k, to mirror the recently established bound (2.12) on the corresponding continuous stability constant.

There are very many open problems in extending the results of this paper to more general scatterers. In this extension we expect that our mesh design and parts of our analysis will have relevance for representing certain components of the total field. For example, in the case of 2D convex curvilinear polygons, something close to the mesh grading we use may be appropriate on each side of the polygon, especially at higher frequencies when the waves diffracted by the corners become more localized near the corners. In the case of three-dimensional scattering by convex polyhedra, it seems to us that the mesh we propose may be useful in representing the variation of edge scattered waves in the direction perpendicular to the edge.

Acknowledgments. The authors gratefully acknowledge helpful discussions with Markus Melenk (Vienna) and Johannes Elschner (WIAS, Berlin) and the helpful comments of the anonymous referees.

REFERENCES

[1] T. Abboud, J.-C. Nédélec, and B. Zhou, *Méthode des équations intégrales pour les hautes fréquences*, C. R. Acad. Sci. Paris Sér. I Math., 318 (1994), pp. 165–170.

[2] M. Abramowitz and I. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1972.

[3] S. Amini, *On the choice of the coupling parameter in boundary integral formulations of the exterior acoustic problem*, Appl. Anal., 35 (1990), pp. 75–92.

[4] S. Amini and N. D. Maines, *Preconditioned Krylov subspace methods for boundary element solution of the Helmholtz equation*, Internat. J. Numer. Methods Engrg., 41 (1998), pp. 875–898.

[5] S. Amini and A. T. J. Profit, *Multi-level fast multipole Galerkin method for the boundary integral solution of the exterior Helmholtz equation*, in Current Trends in Scientific Computing (Xi'an, 2002), Contemp. Math. 329, AMS, Providence, RI, 2003, pp. 13–19.

[6] S. Arden, S. N. Chandler-Wilde, and S. Langdon, *A collocation method for high frequency scattering by convex polygons*, J. Comput. Appl. Math., to appear.
[7] K. E. Atkinson, *The Numerical Solution of Integral Equations of the Second Kind*, Cambridge University Press, Cambridge, UK, 1997.
[8] I. Babuška and J. M. Melenk, *The partition of unity method*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 727–758.
[9] I. M. Babuška and S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM J. Numer. Anal., 34 (1997), pp. 2392–2423. Reprinted in SIAM Rev., 42 (2000), pp. 451–484.
[10] J. J. Bowman, T. B. A. Senior, and P. L. E. Uslenghi, *Electromagnetic and Acoustic Scattering by Simple Shapes*, North-Holland, Amsterdam, 1969.
[11] H. Brakhage and P. Werner, *Über das Dirichletsche Aussenraumproblem für die Helmholtzsche Schwingungsgleichung*, Arch. Math., 16 (1965), pp. 325–329.
[12] O. P. Bruno, C. A. Geuzaine, J. A. Monro, Jr., and F. Reitich, *Prescribed error tolerances within fixed computational times for scattering problems of arbitrarily high frequency: The convex case*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 629–645.
[13] O. P. Bruno and L. A. Kunyansky, *A fast, high-order algorithm for the solution of surface scattering problems: Basic implementation, tests, and applications*, J. Comput. Phys., 169 (2001), pp. 80–110.
[14] A. Buffa and S. Sauter, *On the acoustic single layer potential: Stabilization and Fourier analysis*, SIAM J. Sci. Comput., 28 (2006), pp. 1974–1999.
[15] A. J. Burton and G. F. Miller, *The application of integral equation methods to the numerical solution of some exterior boundary-value problems*, Proc. Roy. Soc. Lond. Ser. A, 323 (1971), pp. 201–210.
[16] O. Cessenat and B. Després, *Using plane waves as base functions for solving time harmonic equations with the ultra weak variational formulation*, J. Comput. Acoust., 11 (2003), pp. 227–238.
[17] G. Chandler, *Galerkin's method for boundary integral equations on polygonal domains*, J. Aust. Math. Soc. Ser. B, 26 (1984), pp. 1–13.
[18] S. N. Chandler-Wilde, *Boundary value problems for the Helmholtz equation in a half-plane*, in Proceedings of the Third International Conference on Mathematical and Numerical Aspects of Wave Propagation, G. Cohen, ed., SIAM, Philadelphia, 1995, pp. 188–197.
[19] S. N. Chandler-Wilde, S. Langdon, and L. Ritter, *A high-wavenumber boundary-element method for an acoustic scattering problem*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 647–671.
[20] S. N. Chandler-Wilde and P. Monk, *Wave-number-explicit bounds in time-harmonic scattering*, SIAM J. Math. Anal., submitted.
[21] W. C. Chew, J. M. Song, T. J. Cui, S. Velamparambil, M. L. Hastriter, and B. Hu, *Review of large scale computing in electromagnetics with fast integral equation solvers*, CMES Comput. Model. Eng. Sci., 5 (2004), pp. 361–372.
[22] S. H. Christiansen and J. C. Nédélec, *Preconditioners for the numerical solution of boundary integral equations from electromagnetism*, C. R. Acad. Sci. Paris Sér. I Math., 331 (2000), pp. 733–738 (in French).
[23] D. L. Colton and R. Kress, *Integral Equation Methods in Scattering Theory*, Wiley, New York, 1983.
[24] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, Springer-Verlag, Berlin, 1992.
[25] E. Darrigrand, *Coupling of fast multipole method and microlocal discretization for the* 3*-D Helmholtz equation*, J. Comput. Phys., 181 (2002), pp. 126–154.
[26] E. Darve and P. Havé, *A fast multipole method for Maxwell equations stable at all frequencies*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 603–628.
[27] A. de La Bourdonnaye and M. Tolentino, *Reducing the condition number for microlocal discretization problems*, Philos. Trans. R. Soc. Lond. A, 362 (2004), pp. 541–559.
[28] V. Domínguez, I. G. Graham, and V. P. Smyshlyaev, *A Hybrid Numerical-Asymptotic Boundary Integral Method for High-Frequency Acoustic Scattering*, Preprint 1/2006, Bath Institute for Complex Systems, University of Bath, Bath, UK, 2006.
[29] F. Ecevit and F. Reitich, *A high-frequency integral equation method for electromagnetic and acoustic scattering simulations: Rate of convergence of multiple scattering iterations*, in Proceedings of the Seventh International Conference on Mathematical and Numerical Aspects of Wave Propagation, Providence, RI, 2005, pp. 145–147.
[30] J. Elschner, *The double layer potential operator over polyhedral domains* I: *Solvability in weighted Sobolev spaces*, Appl. Anal., 45 (1992), pp. 117–134.

[31] L. N. G. Filon, *On a quadrature formula for trigonometric integrals*, Proc. Roy. Soc. Edinburgh, 49 (1928), pp. 38–47.
[32] M. Ganesh, S. Langdon, and I. Sloan, *Efficient evaluation of highly oscillatory acoustic scattering integrals*, J. Comput. Appl. Math., to appear.
[33] C. Geuzaine, O. Bruno, and F. Reitich, *On the $O(1)$ solution of multiple-scattering problems*, IEEE Trans. Magnetics, 41 (2005), pp. 1488–1491.
[34] E. Giladi and J. B. Keller, *A hybrid numerical asymptotic method for scattering problems*, J. Comput. Phys., 174 (2001), pp. 226–247.
[35] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, Berlin, 1983.
[36] P. J. Harris and K. Chen, *On efficient preconditioners for iterative solution of a Galerkin boundary element equation for the three-dimensional exterior Helmholtz problem*, J. Comput. Appl. Math., 156 (2003), pp. 303–318.
[37] T. Huttunen, P. Monk, F. Collino, and J. P. Kaipio, *The ultra-weak variational formulation for elastic wave problems*, SIAM J. Sci. Comput., 25 (2004), pp. 1717–1742.
[38] D. Huybrechs and S. Vandewalle, *On the evaluation of highly oscillatory integrals by analytic continuation*, SIAM J. Numer. Anal., 44 (2006), pp. 1026–1048.
[39] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer-Verlag, New York, 1998.
[40] A. Iserles, *On the numerical quadrature of highly-oscillating integrals.* I. *Fourier transforms*, IMA J. Numer. Anal., 24 (2004), pp. 365–391.
[41] A. Iserles, *On the numerical quadrature of highly-oscillating integrals.* II. *Irregular oscillations*, IMA J. Numer. Anal., 25 (2005), pp. 25–44.
[42] C. E. Kenig, *Harmonic Analysis Techniques for Second Order Elliptic Boundary Value Problems*, AMS, Providence, RI, 1994.
[43] R. Kress, *Minimizing the condition number of boundary integral operators in acoustic and electromagnetic scattering*, Quart. J. Mech. Appl. Math., 38 (1985), pp. 323–341.
[44] S. Langdon and S. N. Chandler-Wilde, *Implementation of a boundary element method for high frequency scattering by convex polygons*, in Advances in Boundary Integral Methods, Proceedings of the 5th UK Conference on Boundary Integral Methods, K. Chen, ed., Liverpool University Press, Liverpool, UK, 2005, pp. 2–11.
[45] S. Langdon and S. N. Chandler-Wilde, *A wavenumber independent boundary element method for an acoustic scattering problem*, SIAM J. Numer. Anal., 43 (2006), pp. 2450–2477.
[46] R. Leis, *Zur Dirichletschen Randwertaufgabe des Außenraumes der Schwingungsgleichung*, Math. Z., 90 (1965), pp. 205–211.
[47] D. Levin, *Analysis of a collocation method for integrating rapidly oscillatory functions*, J. Comput. Appl. Math., 78 (1997), pp. 131–138.
[48] C. Liu, *The Helmholtz Equation on Lipschitz Domains*, Preprint 1356, IMA, University of Minnesota, Minneapolis, MN, 1995.
[49] W. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[50] M. Mitrea, *Boundary value problems and Hardy spaces associated to the Helmholtz equation in Lipschitz domains*, J. Math. Anal. Appl., 202 (1996), pp. 819–842.
[51] F. Oberhettinger and L. Badii, *Tables of Laplace Transforms*, Springer-Verlag, New York, Heidelberg, 1973.
[52] O. J. Panič, *On the solubility of exterior boundary-value problems for the wave equation and for a system of Maxwell's equations*, Uspekhi Mat. Nauk, 20 (1965), pp. 221–226 (in Russian).
[53] E. Perrey-Debain, O. Lagrouche, P. Bettess, and J. Trevelyan, *Plane-wave basis finite elements and boundary elements for three-dimensional wave scattering*, Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 362 (2004), pp. 561–577.
[54] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*, Birkhäuser Verlag, Basel, Switzerland, 1991.
[55] J. R. Rice, *On the degree of convergence of nonlinear spline approximation*, in Approximations with Special Emphasis on Spline Functions, I. J. Schoenburg, ed., Academic Press, New York, 1969.
[56] R. H. Torres and G. V. Welland, *The Helmholtz equation and transmission problems with Lipschitz interfaces*, Indiana Univ. Math. J., 42 (1993), pp. 1457–1485.
[57] G. Verchota, *Layer potentials and boundary value problems for Laplace's equation in Lipschitz domains*, J. Funct. Anal., 59 (1984), pp. 572–611.
[58] V. J. Šelepov, *The index of an integral operator of potential type in the space L_p*, Soviet Math. Dokl., 10 (1969), pp. 754–757.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 641–665

CONVERGENCE OF ADAPTIVE DISCONTINUOUS GALERKIN APPROXIMATIONS OF SECOND-ORDER ELLIPTIC PROBLEMS*

OHANNES A. KARAKASHIAN† AND FREDERIC PASCAL‡

Abstract. A residual-type a posteriori error estimator is introduced and analyzed for a discontinuous Galerkin formulation of a model second-order elliptic problem with Dirichlet–Neumann-type boundary conditions. An adaptive algorithm using this estimator together with specific marking and refinement strategies is constructed and shown to achieve any specified error level in the energy norm in a finite number of cycles. The convergence rate is in effect linear with a guaranteed error reduction at every cycle. Results of numerical experiments are presented.

Key words. discontinuous Galerkin methods, a posteriori estimates, convergence of adaptive methods

AMS subject classifications. 65N55, 65F10

DOI. 10.1137/05063979X

1. Introduction.

Let $\Omega \subset \mathbf{R}^d$, $d = 2, 3$, be a bounded open polyhedral domain. We consider the following boundary value problem:

$$-\Delta u = f \quad \text{in } \Omega, \tag{1.1}$$

$$u = g_D \quad \text{on } \Gamma_D, \tag{1.2}$$

$$\nabla u \cdot n = g_N \quad \text{on } \Gamma_N, \tag{1.3}$$

where $\partial\Omega := \Gamma = \Gamma_D \cup \Gamma_N$ and n is the unit normal vector exterior to Ω. We assume that Γ_D has positive measure, $f \in L^2(\Omega)$, $g_N \in L^2(\Gamma_N)$. Assumptions on f, g_D, and g_N are given later.

Recently there has been a flurry of activity concerning a posteriori error estimates for the discontinuous Galerkin (DG) method for elliptic as well as other problems. In [7], Becker, Hansbo, and Larson use a Helmholtz-type decomposition of the error to derive estimates in the energy norm. Bustinza, Gatica, and Cockburn [11] use a similar technique to derive estimates for linear and nonlinear elliptic problems for the local discontinuous Galerkin (LDG) method. (See also the article by Castillo [12] in the same issue.) Creusé and Nicaise [13] consider the interesting issue of anisotropic elements, i.e., those with large aspect ratio, in the context of the stationary Stokes problem. Houston, Schötzau, and Wihler [15] derive energy norm a posteriori estimates for the hp-version of the DG method for elliptic problems. The fact that the penalty parameter γ appears with a different exponent in their a posteriori estimates provides an interesting alternative to ours. We also mention [21] and [16] for L^2-norm or functional error estimation for the DG method.

In [17] we presented residual-type a posteriori estimates in the energy norm for DG approximations of a special case of the boundary value problem (1.1)–(1.3) corresponding to $\Gamma_D = \Gamma$ and $g_D = 0$. In the present work, we extend these estimates to encompass the more general mixed boundary conditions (1.2), (1.3), continuing

*Received by the editors September 7, 2005; accepted for publication (in revised form) August 18, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sinum/45-2/63979.html
†Department of Mathematics, University of Tennessee, Knoxville, TN 37996 (ohannes@math.utk.edu). The research of this author was supported partially by NSF grant DMS-0411448.
‡CMLA, ENS de Cachan et CNRS, 61 Av. du President Wilson, 94235 Cachan Cedex, France (Frederic.Pascal@cmla.ens-cachan.fr).

work already begun in [18]. These estimates are used to provide a mesh modification strategy which is then shown to be convergent in the energy norm induced by the bilinear form defining the DG method.

The principal goal of an adaptive algorithm is to achieve a user specified error level in a finite number of cycles. While it is typical for the error to be measured in the energy norm ($\|\nabla e\|$ for the standard Galerkin method for second-order elliptic problems, and an appropriate energy norm for the DG method), other interesting and useful measures of the error, "quantities of interest" (QOI), are emerging; see, e.g., [8]. To date, however, no convergence results are known except with respect to the energy norm, given that the proofs make use in an essential way of an orthogonality relation and such relations are not known for other QOI's.

A typical cycle consists of the following basic steps: (1) Given a mesh $\mathcal{T}_H$, calculate the approximation on this mesh. (2) Estimate the error of the approximation at hand using an error estimator. (3) Refine/coarsen $\mathcal{T}_H$ using the information to obtain a new mesh $\mathcal{T}_h$.

The rigorous treatment of the convergence of adaptive algorithms for elliptic problems can be said to have started with the paper of Babuşka and Vogelius [4] where a detailed treatment of the one-dimensional case was given. In 1996 Dörfler [14] gave a convergence proof for the two-dimensional case for the standard Galerkin method using linear elements while outlining an extension to quadratic elements. One of the highlights of this work is that bounds on the convergence rate were provided, which was not the case for [4]. On the other hand, the initial mesh had to be fine enough to essentially resolve the solution. The latter issue provided the starting point for the work of Morin, Nochetto, and Siebert [19], [20], who introduced the concept of *data oscillation* $osc(f, \mathcal{T}_H) = (\sum_{K\in\mathcal{T}_H} \|H(f - f_K)\|^2)^{1/2}$ to circumvent this requirement. The nagging issue of calculating this quantity accurately on a coarse mesh is not resolved and should be treated within the larger and important framework of accounting for the quadrature errors arising from the implementation of the finite element formulation as well as from the calculation of certain terms in the a posteriori estimators. More recently, Binev, Dahmen, and DeVore [9] have proposed a modification of the algorithm in [20] that incorporates coarsening to prove optimal work estimates. More specifically, they have shown that if the solution u can be approximated by a piecewise linear function to an accuracy of $O(n^{-s})$ on a triangulation with n cells, then the algorithm constructs an approximation with the same asymptotic accuracy at a cost of $O(n)$ arithmetic operations.

In this paper we take up the issue of convergence of an adaptive algorithm in the context of a DG formulation for the problem (1.1)–(1.3). The specific DG method used is of an interior penalty type that can be traced to the work of [5] and [2]. We refer to [3] for a survey and unified view of DG methods. Our main result can be summarized as follows: With $a_h^\gamma(\cdot,\cdot)$, $h > 0$, denoting the bilinear form associated with the DG formulation of the problem, we have $a_h^\gamma(e_h, e_h) \le \rho a_H^\gamma(e_H, e_H)$, $\rho < 1$. The principal assumptions which enable this result are the following:

(i) The data of the problem f, g_D, g_N belong to the same polynomial spaces which contain the numerical solution.
(ii) The mesh $\mathcal{T}_h$ is not too fine with respect to the mesh $\mathcal{T}_H$.
(iii) The penalty parameter is not too small; specifically, it must be larger than a constant depending only on the minimum angle of the triangles and the degree of the polynomials in the discontinuous finite element spaces.
(iv) While the marking strategy used is the one used by Dörfler, the refinement strategy is designed to accommodate the DG approach.

While the assumption on the data of the problem may seem to be restrictive, we should note that for one there are practically important cases satisfying these assumptions, e.g., piecewise constant data. Another mitigating argument is that the numerical integration rules cannot distinguish between the data functions and their Lagrange interpolants. Therefore, these assumptions can be relaxed in tandem with an effort to take into account the effect of numerical integration.

The paper is organized as follows. Section 2 is devoted to preliminaries. In addition to establishing notation, we quote a result whose details can be found in [17] and [18] concerning the approximation of discontinuous piecewise polynomial functions by continuous functions of the same type. This result has so far played a key role in the a posteriori estimates as well as in the convergence proof. Section 3 is devoted to the derivation of the residual-type a posteriori estimates extending the results of [17] to include the more general mixed boundary conditions (1.2), (1.3). A novel contribution is estimate (3.12) (Theorem 3.2(iv)). It completes the a posteriori error estimates of [17] by providing lower bounds for the gradients of the error. That the jump terms in (3.12) are multiplied by γ^2 is significant in that γ appears with exponent one in the bilinear form. Interestingly, this fact plays an important role in the proof of convergence of the adaptive algorithm. In section 4 we outline the marking and refinement strategies and prove convergence of the adaptive scheme. It is worth noting that the analysis of the DG formulation presents some complications not present in the standard method. One is due to the fact that the energy norm is mesh dependent. Another more basic one is due to the fact that the bilinear form which defines the method is not coercive on the energy space. Both issues are successfully resolved. Let us also note that while the marking and refinement strategies are couched in two dimensions, we believe that appropriate modifications can be introduced to obtain convergence in three dimensions as well. In particular, only the refinement strategy, Lemma 4.1, and Corollary 4.1 need be extended. Finally, in section 5 we present the results of some numerical experiments.

2. Preliminaries.

2.1. Notation. For a domain $D \subseteq \mathbf{R}^d$ and integer $m \geq 0$, $H^m(D)$ will denote the (Hilbert) Sobolev space with inner product $(u,v)_{m,D} = \sum_{|\alpha|\leq m} \int_D D^\alpha u\, D^\alpha v\, dx$ and norm $\|u\|_{m,D} = (u,u)_{m,D}^{1/2}$ (cf. [1]). To simplify the notation, we shall drop m when its value is zero. Also, we shall often encounter functions that vanish on Γ_D, and thus we let $H^1_{0,\Gamma_D} = \{v \in H^1(\Omega),\ v = 0 \text{ on } \Gamma_D\}$.

Extensive use will be made of edge/surface integrals. Therefore, for a $(d-1)$-dimensional subset e of $\mathbf{R}^d$, we set $\langle u, v\rangle_e = \int_e u\, v\, ds$ and $|u|_e = \langle u, u\rangle_e^{1/2}$.

2.2. Triangulations. Let $\mathcal{T}_h = \{K_i : i = 1, 2, \ldots, m_h\}$ be a family of star-like partitions (triangulations) of the domain Ω parametrized by $0 < h \leq 1$. We assume the following:

(i) The elements of $\mathcal{T}_h$ satisfy the minimal angle condition.

(ii) $\mathcal{T}_h$ is locally quasi-uniform; that is, if two elements K_j and K_ℓ are adjacent in the sense that $\mu_{d-1}(\partial K_j \cap \partial K_\ell) > 0$, then $\mathrm{diam}(\mathrm{K_j}) \approx \mathrm{diam}(\mathrm{K}_\ell)$.

We define $\mathcal{E}_h^I$ and $\mathcal{E}_h^B$ to be the set of all interior and boundary edges (faces in the case $d = 3$), respectively, as follows:

$$\mathcal{E}_h^I = \{e = \partial K_j \cap \partial K_\ell, \quad \mu_{d-1}(\partial K_j \cap \partial K_\ell) > 0\},$$
$$\mathcal{E}_h^B = \{e = \partial K \cap \partial\Omega, \quad \mu_{d-1}(\partial K \cap \partial\Omega) > 0\}, \qquad \mathcal{E}_h = \mathcal{E}_h^I \cup \mathcal{E}_h^B.$$

For each $e \in \mathcal{E}_h^I$, we denote the two triangles that "share" it by K^+ and K^-, respectively. Which of the two is K^+ is completely arbitrary but not irrelevant! If $e \in \mathcal{E}_h^B$, then $e = \partial K^+ \cap \partial\Omega \equiv \partial K \cap \partial\Omega$.

We assume that for each $e \in \mathcal{E}_h^B$, either $e \subset \Gamma_D$ or $e \subset \Gamma_N$. We then set $\mathcal{E}_h^B = \mathcal{E}_h^D \cup \mathcal{E}_h^N$, where $\mathcal{E}_h^D$ and $\mathcal{E}_h^N$ are, respectively, the set of boundary edges on Γ_D and on Γ_N. From the previous assumption, we have $\mathcal{E}_h^D \cap \mathcal{E}_h^N = \emptyset$.

Given a partition or mesh $\mathcal{T}_h$ of Ω, we find it convenient to use the spaces $H^m(\mathcal{T}_h) = \Pi_{K\in\mathcal{T}_h} H^m(K)$. In this context we consider K to be open so that elements of $H^m(\mathcal{T}_h)$ are single-valued. In particular, the "energy space" for the DG method for this problem will be $E_h = H^2(\mathcal{T}_h)$.

We shall also use the discontinuous finite element spaces $V_h^r = \Pi_{K\in\mathcal{T}_h} P_{r-1}(K)$, $r \geq 2$, where $P_k(K)$ is the space of polynomials of total degree k defined on K.

It is essential to be able to define values of functions in $H^m(\mathcal{T}_h)$ and V_h^r on the edges e. Thus, for $v \in H^m(\mathcal{T}_h)$, $m \geq 1$, and $e \in \mathcal{E}_h^I \cup \mathcal{E}_h^B$, v_e^+ will denote the trace on e of the restriction v^+ of v to K^+. Similarly we define v_e^- for $e \in \mathcal{E}_h^I$.

We also define *jumps* and *averages* of such traces as follows:

$$[v] = v_e^+ - v_e^-,\ e \in \mathcal{E}_h^I, \qquad [v] = v_e^+,\ e \in \mathcal{E}_h^B,$$
$$\{v\} = \tfrac{1}{2}\big(v_e^+ + v_e^-\big),\ e \in \mathcal{E}_h^I, \qquad \{v\} = v_e^+,\ e \in \mathcal{E}_h^B.$$

Finally, for $v \in H^2(\mathcal{T}_h)$ we let $\{\partial_n v\} = \frac{1}{2}\big(\nabla v^+ + \nabla v^-\big)\cdot \mathbf{n}^+$, $e \in \mathcal{E}_h^I$, where $\mathbf{n}^+$ is the unit outward normal to K^+ and $[\partial_n v] = \nabla v^+ \cdot \mathbf{n}^+ - \nabla v^- \cdot \mathbf{n}^+$, $e \in \mathcal{E}_h^I$.

2.3. Some useful results. We shall make frequent use of the following trace and inverse inequalities (cf. [10], [17]):

$$|v|_{\partial D}^2 \leq c\left(h_D^{-1}\|v\|_D^2 + h_D\|\nabla v\|_D^2\right) \quad \forall v \in H^1(D), \tag{2.1}$$

where $h_D = \operatorname{diam}(D)$;

$$|v|_{j,D} \leq c h_D^{i-j}|v|_{i,D} \quad \forall v \in P_k(D),\ 0 \leq i \leq j \leq 2. \tag{2.2}$$

We shall also make essential use of the fact that an element of V_h^r can be approximated by *continuous* piecewise polynomial functions, specifically by elements of $V_h^r \cap H^1(\Omega)$; the degree of approximation being controlled, not surprisingly, by the jumps of the discontinuous function. Here we extend the result established in [17] to allow approximation by functions that also satisfy Dirichlet-type conditions on the boundary. We also include a significant observation that the approximation result holds in the L^2-norm as well. We omit the proof since its essential points were provided in [17] and [18].

THEOREM 2.1. *Let $\mathcal{T}_h$ be a conforming or nonconforming mesh consisting of triangles when $d = 2$, and tetrahedra when $d = 3$. Then for any $v_h \in V_h^r$ and multi-index α with $|\alpha| = 0, 1$ the following approximation results hold:*

(i) *Let g be the restriction to Γ of a function in $V_h^r \cap H^1(\Omega)$. Then there exists $\chi \in V_h^r \cap H^1(\Omega)$ satisfying $\chi\big|_\Gamma = g$ and*

$$\sum_{K\in\mathcal{T}_h} \|D^\alpha(v_h - \chi)\|_K^2 \leq c\left(\sum_{e\in\mathcal{E}_h^I} h_e^{1-2|\alpha|}|[v_h]|_e^2 + \sum_{e\in\mathcal{E}_h^B} h_e^{1-2|\alpha|}|v_h - g|_e^2\right). \tag{2.3}$$

(ii) *Let g be the restriction to Γ_D of a function in $V_h^r \cap H^1(\Omega)$. Then there exists $\chi \in V_h^r \cap H^1(\Omega)$ satisfying $\chi\big|_{\Gamma_D} = g$ and*

(2.4)
$$\sum_{K\in\mathcal{T}_h} \|D^\alpha(v_h-\chi)\|_K^2 \le c\left(\sum_{e\in\mathcal{E}_h^I} h_e^{1-2|\alpha|}|[v_h]|_e^2 + \sum_{e\in\mathcal{E}_h^D} h_e^{1-2|\alpha|}|v_h-g|_e^2\right).$$

(iii) *There exists $\chi \in V_h^r \cap H^1(\Omega)$ satisfying*

$$\sum_{K\in\mathcal{T}_h} \|D^\alpha(v_h-\chi)\|_K^2 \le c\sum_{e\in\mathcal{E}_h^I} h_e^{1-2|\alpha|}|[v_h]|_e^2 \tag{2.5}$$

for some constant C independent of h and v_h but which may depend on r and the minimal angle θ_0 of the triangles in $\mathcal{T}_h$.

REMARK 2.1. *The proof of this result is constructive and is based on an averaging process. It should also hold for more general partitions of Ω such as quadrilaterals and parallelepipeds.*

3. A posteriori error estimates.

3.1. The discrete problem. In order to construct a weak formulation for the problem (1.1)–(1.3), we introduce the bilinear form $a_h^\gamma : E_h \times E_h \to \mathbf{R}$:

$$a_h^\gamma(u,v) = \sum_{K\in\mathcal{T}_h} (\nabla u, \nabla v)_K - \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \Big(\langle\{\partial_n u\},[v]\rangle_e + \langle\{\partial_n v\},[u]\rangle_e - \gamma h_e^{-1}\langle[u],[v]\rangle_e\Big),$$

where $h_e = \operatorname{diam}(e)$ and γ is the interior penalty parameter. We point out that we have adopted the averaged value $\{\partial_n v\}_e$ of the normal derivatives attributed to Arnold [2]. The results of this paper also apply to the so-called Baker formulation for which $\{\partial_n v\} = \nabla v^+ \cdot \mathbf{n}^+$.

The bilinear form a_h^γ is consistent with the BVP (1.1)–(1.3) in the following sense: If $u \in H^2(\Omega)$ satisfies (1.1)–(1.3), then

$$a_h^\gamma(u,v) = F(v) := (-\Delta u, v)_\Omega - \sum_{e\in\mathcal{E}_h^D} \left\langle u, \partial_n v - \gamma h_e^{-1} v\right\rangle_e + \sum_{e\in\mathcal{E}_h^N} \langle \partial_n u, v\rangle_e \quad \forall v \in E_h.$$

Thus, we define the DG approximation u_h^γ of the solution u of the BVP (1.1)–(1.3) as the element of V_h^r that satisfies

$$a_h^\gamma(u_h^\gamma, v) = F(v) \quad \forall v \in V_h^r. \tag{3.1}$$

We thus have the orthogonality relation

$$a_h^\gamma(u-u_h^\gamma, v) = 0 \quad \forall v \in V_h^r, \tag{3.2}$$

which will play an important role in the derivation of the a posteriori estimates as well as the proof of the convergence of the adaptive scheme.

Concerning the continuity and coercivity of the form a_h^γ, we can prove the following result.

LEMMA 3.1. (i)

$$|a_h^\gamma(u,v)| \le 2\|u\|_{1,h}\|v\|_{1,h} \qquad \forall u, v \in E_h.$$

(ii) *There exist positive constants* γ_0 *and* c_a *such that for all* $\gamma \geq \gamma_0$

$$a_h^\gamma(v,v) \geq c_a \|v\|_{1,h}^2 \qquad \forall v \in V_h^r,$$

where

$$\|v\|_{1,h} = \left(\sum_{K\in\mathcal{T}_h} \|\nabla v\|_K^2 + \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \left(h_e \left|\{\partial_n v\}\right|_e^2 + \gamma h_e^{-1} |[v]|_e^2 \right) \right)^{1/2}.$$

Let us mention here that γ_0 depends only on r and θ_0. Also, the proof of (i) is merely an application of the Cauchy–Schwarz inequality. To prove (ii), we have to use the trace and inverse inequalities (2.1), (2.2).

3.2. A residual-type a posteriori estimate. This section is devoted to the generalization of the residual-type a posteriori estimates given in [17]. The estimators as well as the exposition follow the lines found in Verfürth [23], with the exception of the technical details stemming from the discontinuous nature of V_h^r.

THEOREM 3.1. *Suppose that* g_D *in* (1.2) *is the restriction to* Γ_D *of a function in* $V_h^r \cap H^1(\Omega)$. *Then with* $e = u - u_h^\gamma$ *there holds*

$$\sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2 \leq c \Bigg\{ \sum_{K\in\mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \sum_{e\in\mathcal{E}_h^I} h_e \left|\left[\partial_n u_h^\gamma\right]\right|_e^2 + \sum_{e\in\mathcal{E}_h^N} h_e \left|g_N - \partial_n u_h^\gamma\right|_e^2$$

$$+ \gamma^2 \sum_{e\in\mathcal{E}_h^I} h_e^{-1} \left|[u_h^\gamma]\right|_e^2 + \gamma^2 \sum_{e\in\mathcal{E}_h^D} h_e^{-1} \left|g_D - u_h^\gamma\right|_e^2 \Bigg\}. \tag{3.3}$$

In particular, the constant c *is independent of the meshsize and* γ.

Proof. Integrating by parts, we obtain

$$\sum_{K\in\mathcal{T}_h} (\nabla e, \nabla\eta)_K = \sum_{K\in\mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K + \sum_{e\in\mathcal{E}_h^I} \left(\langle \{\partial_n e\}, [\eta] \rangle_e + \langle \{\eta\}, [\partial_n e] \rangle_e \right)$$

$$+ \sum_{e\in\mathcal{E}_h^D} \langle \partial_n e, \eta \rangle_e + \sum_{e\in\mathcal{E}_h^N} \langle \partial_n e, \eta \rangle_e \quad \forall \eta \in H^1(\mathcal{T}_h). \tag{3.4}$$

Letting $\eta = e - v_h$, where v_h is piecewise constant on $\mathcal{T}_h$, we have

$$\sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2 = \sum_{K\in\mathcal{T}_h} (f + \Delta u_h^\gamma, \eta)_K - \sum_{e\in\mathcal{E}_h^I} \left(\langle \{\partial_n e\}, [u_h^\gamma + v_h] \rangle_e + \langle \{\eta\}, [\partial_n u_h^\gamma] \rangle_e \right)$$

$$- \sum_{e\in\mathcal{E}_h^D} \langle \partial_n e, u_h^\gamma + v_h - g_D \rangle_e + \sum_{e\in\mathcal{E}_h^N} \langle g_N - \partial_n u_h^\gamma, \eta \rangle_e. \tag{3.5}$$

Now, from the orthogonality relation (3.2), for all χ in $V_h^r \cap H^1(\Omega)$ with $\chi\big|_{\Gamma_D} = g_D$, we have

$$\begin{aligned}
0 &= a_h^\gamma(e, u_h^\gamma + v_h - \chi) \\
&= \sum_{K\in\mathcal{T}_h} (\nabla e, \nabla(u_h^\gamma - \chi))_K - \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \langle \{\partial_n (u_h^\gamma - \chi)\}, [e] \rangle_e \\
&\quad - \sum_{e\in\mathcal{E}_h^I} \langle \{\partial_n e\}, [u_h^\gamma + v_h] \rangle_e - \sum_{e\in\mathcal{E}_h^D} \langle \partial_n e, u_h^\gamma + v_h - g_D \rangle_e \\
&\quad + \sum_{e\in\mathcal{E}_h^I} \gamma h_e^{-1} \langle [e], [u_h^\gamma + v_h] \rangle_e + \sum_{e\in\mathcal{E}_h^D} \gamma h_e^{-1} \langle g_D - u_h^\gamma, u_h^\gamma + v_h - g_D \rangle_e.
\end{aligned}$$

Using this relation in (3.5) to eliminate the terms containing $\partial_n e$, we obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2 = &\sum_{K\in\mathcal{T}_h}(f+\Delta u_h^\gamma,\eta)_K - \sum_{K\in\mathcal{T}_h}(\nabla e,\nabla(u_h^\gamma-\chi))_K\\ &-\sum_{e\in\mathcal{E}_h^I}\Big(\langle\{\partial_n(u_h^\gamma-\chi)\},[u_h^\gamma]\rangle_e+\gamma h_e^{-1}\langle[u_h^\gamma],[\eta]\rangle_e+\langle\{\eta\},[\partial_n u_h^\gamma]\rangle_e\Big)\\ &+\sum_{e\in\mathcal{E}_h^D}\Big(\langle\partial_n(u_h^\gamma-\chi),g_D-u_h^\gamma\rangle_e+\gamma h_e^{-1}\langle g_D-u_h^\gamma,\eta\rangle_e\Big)\\ &+\sum_{e\in\mathcal{E}_h^N}\langle g_N-\partial_n u_h^\gamma,\eta\rangle_e\,. \end{aligned}\tag{3.6}$$

We now obtain bounds for the terms on the right-hand side of (3.6). Those that contain η are bounded by $\frac{1}{2}$ times,

$$\begin{aligned}&\frac{1}{\epsilon_1}\sum_{K\in\mathcal{T}_h}h_K^2\|f+\Delta u_h^\gamma\|_K^2+\frac{1}{\epsilon_2}\sum_{e\in\mathcal{E}_h^I}h_e\big|\big[\partial_n u_h^\gamma\big]\big|_e^2+\frac{1}{\epsilon_3}\gamma\sum_{e\in\mathcal{E}_h^I}h_e^{-1}|[u_h^\gamma]|_e^2\\ &+\frac{1}{\epsilon_4}\gamma\sum_{e\in\mathcal{E}_h^D}h_e^{-1}|g_D-u_h^\gamma|_e^2+\frac{1}{\epsilon_5}\sum_{e\in\mathcal{E}_h^N}h_e|g_N-\partial_n u_h^\gamma|_e^2\\ &+\epsilon_1\sum_{K\in\mathcal{T}_h}h_K^{-2}\|\eta\|_K^2+\epsilon_2\sum_{e\in\mathcal{E}_h^I}h_e^{-1}|\{\partial_n\eta\}|_e^2+\epsilon_3\gamma\sum_{e\in\mathcal{E}_h^I}h_e^{-1}|[\eta]|_e^2\\ &+\epsilon_4\gamma\sum_{e\in\mathcal{E}_h^D}h_e^{-1}|\eta|_e^2+\epsilon_5\sum_{e\in\mathcal{E}_h^N}h_e^{-1}|\eta|_e^2,\end{aligned}\tag{3.7}$$

for any $\epsilon_i>0$, $i=1,\dots,5$. To estimate the "η" terms in (3.7) we choose as v_h the best piecewise constant approximation of e that gives, using an approximation result of [6], the estimate

$$\|\eta\|_K\le ch_K\|\nabla e\|_K,\ K\in\mathcal{T}_h.$$

Since the mesh is locally quasi-uniform, using this approximation result and the trace inequality (2.1), we obtain

$$\begin{aligned}\sum_{K\in\mathcal{T}_h}h_K^{-2}\|\eta\|_K^2&\le c\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2,\\ \sum_{e\in\mathcal{E}_h^I}h_e^{-1}\left(|\{\partial_n\eta\}|_e^2+|[\eta]|_e^2\right)&\le c\sum_{e\in\mathcal{E}_h^I}\sum_{K=K^+,K^-}h_e^{-1}(h_K^{-1}\|\eta\|_K^2+h_K\|\nabla\eta\|_K^2)\\ &\le c\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2,\\ \sum_{e\in\mathcal{E}_h^B}h_e^{-1}|\eta|_e^2&\le c\sum_{e\in\mathcal{E}_h^B}\sum_{K=K^+}h_e^{-1}(h_K^{-1}\|\eta\|_K^2+h_K\|\nabla\eta\|_K^2)\\ &\le c\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2.\end{aligned}$$

We can now hide the "η" terms in the left-hand side of (3.6) by taking the ϵ's sufficiently small. In particular, we must take $\epsilon_3\approx 1/\gamma$ and $\epsilon_4\approx 1/\gamma$.

To obtain (3.3), we also need to estimate the terms containing $u_h^\gamma - \chi$. Indeed, these are bounded by

$$\epsilon \sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2 + \frac{1}{\epsilon}\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-\chi)\|_K^2 + \sum_{e\in\mathcal{E}_h^I} h_e\big|\{\partial_n(u_h^\gamma-\chi)\}\big|_e^2 + \sum_{e\in\mathcal{E}_h^I} h_e^{-1}|[u_h^\gamma]|_e^2 + \sum_{e\in\mathcal{E}_h^D} h_e\big|\partial_n(u_h^\gamma-\chi)\big|_e^2 + \sum_{e\in\mathcal{E}_h^D} h_e^{-1}|g_D-u_h^\gamma|_e^2. \tag{3.8}$$

Using the trace and inverse inequalities, we see that the two terms in (3.8) that contain $\partial_n(u_h^\gamma-\chi)$ are bounded by $\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-\chi)\|_K^2$. In view of Theorem 2.1(ii), the latter is bounded by $\sum_{e\in\mathcal{E}_h^I} h_e^{-1}\,|[u_h^\gamma]|_e^2 + \sum_{e\in\mathcal{E}_h^D} h_e^{-1}|g_D-u_h^\gamma|_e^2$. ∎

THEOREM 3.2. *The following estimates hold:*

(i) *Suppose that f is a piecewise polynomial on $\mathcal{T}_h$. Then for each $K\in\mathcal{T}_h$*

$$h_K^2\|f+\Delta u_h^\gamma\|_K^2 \le c\,\|\nabla e\|_K^2. \tag{3.9}$$

(ii) *For $e=\partial K^+\cap\partial K^-\in\mathcal{E}_h^I$,*

$$h_e\big|\big[\partial_n u_h^\gamma\big]\big|_e^2 \le c(\|\nabla e\|_{K^+}^2+\|\nabla e\|_{K^-}^2). \tag{3.10}$$

(iii) *Suppose that g_N is a piecewise polynomial on $\mathcal{E}_h^N$. Then for $e=\partial K^+\cap\partial\Omega\in\mathcal{E}_h^N$*

$$h_e\big|g_N-\partial_n u_h^\gamma\big|_e^2 \le c\|\nabla e\|_{K^+}^2. \tag{3.11}$$

(iv) *Suppose that g_D is the restriction to Γ_D of a function in $V_h^r\cap H^1(\Omega)$. Then there exists γ_1 depending only on r and θ_0 such that for $\gamma\ge\gamma_1$*

$$\gamma^2\sum_{e\in\mathcal{E}_h^I} h_e^{-1}\big|\big[u_h^\gamma\big]\big|_e^2 + \gamma^2\sum_{e\in\mathcal{E}_h^D} h_e^{-1}\big|g_D-u_h^\gamma\big|_e^2 \le c\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2. \tag{3.12}$$

The constants c in (3.9)–(3.12) depend on r, θ_0 the degrees of f and g_N, but are independent of the meshsize and γ.

Proof. Proofs of assertions (i), (ii), (iii) are similar to the proof of Theorem 3.2 in [17].

Concerning assertion (iv), let $u_h^G\in V_h^r\cap H^1(\Omega)$ with $u_h^G\big|_{\Gamma_D}=g_D$ denote the "standard" continuous Galerkin approximation of u given as the solution of

$$(\nabla u_h^G,\nabla\chi) = (f,\chi)+\sum_{e\in\mathcal{E}_h^N}\langle\chi,g_N\rangle_e \quad \forall\chi\in V_h^r\cap H^1_{0,\Gamma_D}. \tag{3.13}$$

It is easily seen that u_h^G satisfies

$$a_h^\gamma(u_h^G,\chi) = (f,\chi)-\sum_{e\in\mathcal{E}_h^D}\langle g_D,\partial_n\chi\rangle_e+\sum_{e\in\mathcal{E}_h^N}\langle\chi,g_N\rangle_e \quad \forall\chi\in V_h^r\cap H^1_{0,\Gamma_D}, \tag{3.14}$$

and then one gets the following orthogonality relation for u_h^G:

$$a_h^\gamma(u_h^G-u,\chi)=0 \quad \forall\chi\in V_h^r\cap H^1_{0,\Gamma_D}. \tag{3.15}$$

Now for all $\chi \in V_h^r \cap H^1_{0,\Gamma_D}$,

$$\begin{aligned}
a_h^\gamma(u_h^\gamma - u_h^G, u_h^\gamma - u_h^G) &= a_h^\gamma(u - u_h^G, u_h^\gamma - u_h^G) \\
&= a_h^\gamma(u - u_h^G, u_h^\gamma - u_h^G - \chi) \\
&= \sum_{K\in\mathcal{T}_h} (\nabla(u - u_h^G), \nabla(u_h^\gamma - u_h^G - \chi))_K \\
&\quad - \sum_{e\in\mathcal{E}_h^I} \left\langle \left\{\partial_n(u - u_h^G)\right\}, [u_h^\gamma] \right\rangle_e - \sum_{e\in\mathcal{E}_h^D} \left\langle \partial_n(u - u_h^G), u_h^\gamma - g_D \right\rangle_e \\
&= \sum_{K\in\mathcal{T}_h} (\nabla e + \nabla(u_h^\gamma - u_h^G), \nabla(u_h^\gamma - u_h^G - \chi))_K \\
&\quad - \sum_{e\in\mathcal{E}_h^I} \left\langle \{\partial_n e\} + \left\{\partial_n(u_h^\gamma - u_h^G)\right\}, [u_h^\gamma] \right\rangle_e \\
&\quad - \sum_{e\in\mathcal{E}_h^D} \left\langle \partial_n e + \partial_n(u_h^\gamma - u_h^G), u_h^\gamma - g_D \right\rangle_e .
\end{aligned}$$

Then integration by parts of $(\nabla e, \nabla(u_h^\gamma - u_h^G - \chi))_K$ gives

$$\begin{aligned}
&a_h^\gamma(u_h^\gamma - u_h^G, u_h^\gamma - u_h^G) \\
&\quad = \sum_{K\in\mathcal{T}_h} \Big(\big(f + \Delta u_h^\gamma, u_h^\gamma - u_h^G - \chi\big)_K + \big(\nabla(u_h^\gamma - u_h^G), \nabla(u_h^\gamma - u_h^G - \chi)\big)_K \Big) \\
&\qquad - \sum_{e\in\mathcal{E}_h^I} \Big(\big\langle [\partial_n u_h^\gamma], \{u_h^\gamma - u_h^G - \chi\} \big\rangle_e + \big\langle \left\{\partial_n(u_h^\gamma - u_h^G)\right\}, [u_h^\gamma] \big\rangle_e \Big) \\
&\qquad - \sum_{e\in\mathcal{E}_h^D} \left\langle \partial_n(u_h^\gamma - u_h^G), u_h^\gamma - g_D \right\rangle_e - \sum_{e\in\mathcal{E}_h^N} \left\langle \partial_n u_h^\gamma - g_N, u_h^\gamma - u_h^G - \chi \right\rangle_e .
\end{aligned}$$

In view of the coercivity of a_h^γ on V_h^r, using the arithmetic-geometric mean inequality, we obtain

$$\begin{aligned}
c_a \sum_{K\in\mathcal{T}_h} \|\nabla(u_h^\gamma - u_h^G)\|_K^2 &+ (\gamma - \gamma_0) \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} h_e^{-1} |[u_h^\gamma - u_h^G]|_e^2 \\
&\le \frac{\epsilon_1}{2} \sum_{K\in\mathcal{T}_h} \|\nabla(u_h^\gamma - u_h^G)\|_K^2 + \frac{1}{2\epsilon_1} \sum_{K\in\mathcal{T}_h} \|\nabla(u_h^\gamma - u_h^G - \chi)\|_K^2 \\
&\quad + \frac{1}{2\epsilon_2\gamma} \sum_{K\in\mathcal{T}_h} h_K^2 \|f + \Delta u_h^\gamma\|_K^2 + \frac{\epsilon_2\gamma}{2} \sum_{K\in\mathcal{T}_h} h_K^{-2} \|u_h^\gamma - u_h^G - \chi\|_K^2 \\
&\quad + \frac{1}{2\epsilon_2\gamma} \sum_{e\in\mathcal{E}_h^I} h_e |[\partial_n u_h^\gamma]|_e^2 + \frac{\epsilon_2\gamma}{2} \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^N} h_e^{-1} |\{u_h^\gamma - u_h^G - \chi\}|_e^2 \\
&\quad + \frac{\epsilon_1}{2} \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} h_e |\left\{\partial_n(u_h^\gamma - u_h^G)\right\}|_e^2 + \frac{1}{2\epsilon_1} \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} h_e^{-1} |[u_h^\gamma - u_h^G]|_e^2 \\
&\quad + \frac{1}{2\epsilon_2\gamma} \sum_{e\in\mathcal{E}_h^N} h_e |\partial_n u_h^\gamma - g_N|_e^2, \qquad (3.16)
\end{aligned}$$

where c_a and γ_0 are as in Lemma 3.1. Using the trace and inverse inequalities, it

follows that

$$\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} h_e|\left\{\partial_n(u_h^\gamma-u_h^G)\right\}|_e^2 \le c_1\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-u_h^G)\|_K^2, \tag{3.17}$$

$$\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^N} h_e^{-1}|\{u_h^\gamma-u_h^G-\chi\}|_e^2 \le c_2\sum_{K\in\mathcal{T}_h} h_K^{-2}\|u_h^\gamma-u_h^G-\chi\|_K^2, \tag{3.18}$$

where c_1 and c_2 depend only on r and θ_0. Now choose $\chi\in V_h^r\cap H^1_{0,\Gamma_D}$ to approximate $u_h^\gamma-u_h^G$ as in Theorem 2.1(ii). Using (3.18) and (3.17) in (3.16), we have

$$\begin{aligned}
&c_a\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-u_h^G)\|_K^2+(\gamma-\gamma_0)\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D}h_e^{-1}|[u_h^\gamma-u_h^G]|_e^2\\
&\qquad\le\frac{\epsilon_1}{2}(1+c_1)\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-u_h^G)\|_K^2+\frac{1}{2\epsilon_2\gamma}\sum_{K\in\mathcal{T}_h}h_K^2\|f+\Delta u_h^\gamma\|_K^2\\
&\qquad+\frac{1}{2\epsilon_2\gamma}\sum_{e\in\mathcal{E}_h^I}h_e|[\partial_n u_h^\gamma]|_e^2+\frac{1}{2\epsilon_2\gamma}\sum_{e\in\mathcal{E}_h^N}h_e|\partial_n u_h^\gamma-g_N|_e^2\\
&\qquad+\left(\frac{1+c_3}{2\epsilon_1}+\frac{\epsilon_2(1+c_2)c_3\gamma}{2}\right)\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D}h_e^{-1}|[u_h^\gamma-u_h^G]|_e^2,
\end{aligned}\tag{3.19}$$

where c_3 is the constant in (2.4). Now choose ϵ_1 small so that $\epsilon_1(1+c_1)\le c_a$ and choose ϵ_2 small so that $\epsilon_2(1+c_2)c_3\le 1$. Then for $\gamma\ge\gamma_1:=4\big(\gamma_0+\frac{1+c_3}{2\epsilon_1}\big)$ we will have $\gamma-\gamma_0-\frac{1+c_3}{2\epsilon_1}-\frac{\gamma}{2}\ge\frac{1}{4}\gamma$. Note that c_a,γ_0,c_1,c_2,c_3 and consequently $\epsilon_1,\epsilon_2,\gamma_1$ depend only on r and θ_0. Thus, from (3.19) we obtain

$$\begin{aligned}
&\frac{c_a}{2}\sum_{K\in\mathcal{T}_h}\|\nabla(u_h^\gamma-u_h^G)\|_K^2+\frac{1}{4}\gamma\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D}h_e^{-1}|[u_h^\gamma-u_h^G]|_e^2\\
&\le\frac{1}{2\epsilon_2\gamma}\left(\sum_{K\in\mathcal{T}_h}h_K^2\|f+\Delta u_h^\gamma\|_K^2+\sum_{e\in\mathcal{E}_h^I}h_e|[\partial_n u_h^\gamma]|_e^2+\sum_{e\in\mathcal{E}_h^N}h_e|\partial_n u_h^\gamma-g_N|_e^2\right).
\end{aligned}\tag{3.20}$$

Using assertions (3.9)–(3.11) in (3.20), we obtain (3.12). This concludes the proof. ∎

Remark 3.1.

(i) *In view of* (3.20)*, one may obtain lower and upper bounds for* $\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2$ *with only the three terms on the left sides of* (3.9)*,* (3.10)*, and* (3.11)*.*

(ii) *Inequality* (3.12) *is important in that it confirms the right side of* (3.3) *as both an upper and a lower bound for the error* $\sum_{K\in\mathcal{T}_h}\|\nabla e\|_K^2$ *and thus completes Theorem* 3.2 *of* [17].

(iii) *In* [15] *the upper bound analogous to* (3.3) *contains only* γ *and thus constitutes a stronger result than ours in this respect. On the other hand, the fact that the lower bound* (3.12) *contains* γ^2 *is significant and plays an important role in the convergence proof of the adaptive algorithm.*

4. Convergence of the adaptive scheme. In this section we will describe in detail our adaptive algorithm and prove its convergence under appropriate assumptions. The algorithm, which is iterative in nature, consists of constructing a sequence of meshes and corresponding approximations whereby each cycle consists of the following four steps:

1. Given a mesh $\mathcal{T}_H$, a DG approximation u_H^γ is constructed by solving (3.1) exactly (to machine precision). In practice, however, only an approximate solution is found by a fast iterative method, e.g., Multigrid. In that case, the additional errors caused must be taken into account.
2. An (a posteriori) estimation of the error e_H is obtained by calculating, e.g., the right side of (3.3) without the terms containing γ, their exclusion being motivated by (3.20).
3. Based on the information supplied by the a posteriori error estimate certain triangles and edges of $\mathcal{T}_H$ are marked for refinement. This is the *marking* strategy. It is patterned after that of [14].
4. The triangles and edges marked for refinement in step 3 lead to a set of triangles to be refined in a specific way. This is the *refinement* strategy and defines the new mesh $\mathcal{T}_h$.

Our convergence result can be summarized as follows: Let $\mathcal{T}_H$ be a mesh with V_H^r denoting the corresponding discontinuous finite element space, and let $\mathcal{T}_h$ denote a refinement of $\mathcal{T}_H$ obtained by following the above steps. Let u_h^γ and u_H^γ denote the DG solutions in V_h^r and V_H^r, respectively, and e_h and e_H the corresponding errors. Then, under certain assumptions on the data of the BVP (1.1)–(1.3) and for γ sufficiently large, there holds

$$a_h^\gamma(e_h, e_h) \le \rho\, a_H^\gamma(e_H, e_H), \qquad 0 < \rho < 1. \tag{4.1}$$

Let us note that such convergence results are based in an essential manner on an orthogonality relation which in this context is written as

$$a_h^\gamma(e_H, e_H) = a_h^\gamma(e_h, e_h) + a_h^\gamma(u_h^\gamma - u_H^\gamma, u_h^\gamma - u_H^\gamma). \tag{4.2}$$

The convergence of the algorithm hinges on obtaining a fixed reduction in the error, and this depends in a crucial manner on the nonnegative quantity $a_h^\gamma(u_h^\gamma - u_H^\gamma, u_h^\gamma - u_H^\gamma)$ being sufficiently large with respect to the other two terms in (4.2). However, examples of problems can be constructed, in particular when the solution u is oscillatory, whereby $a_h^\gamma(u_h^\gamma - u_H^\gamma,\ u_h^\gamma - u_H^\gamma) = 0$ on an arbitrarily long sequence of meshes, each one obtained from the previous one by full refinement. See, e.g., [19], [20], [14]. It turns out that our assumptions on the data preclude such occurrences, resulting in the linear convergence rate (4.1).

Before engaging in the proof of the theorem, we immediately notice a difficulty presented by the fact that we have $a_h^\gamma(e_H, e_H)$ on the left-hand side of (4.2) instead of $a_H^\gamma(e_H, e_H)$. Another basic problem is that $a_h^\gamma(\cdot,\cdot)$ is not coercive on the energy space. We will show below that $a_h^\gamma(e_h, e_h)$ behaves like a norm, thus giving a meaning to the convergence result $\lim_{h\to 0} a_h^\gamma(e_h, e_h) = 0$ implied by (4.1).

We deal with the first problem by showing that $a_h^\gamma(e_H, e_H)$ is bounded by $a_H^\gamma(e_H, e_H)$ plus a nonnegative quantity that can be absorbed in other terms.

PROPOSITION 4.1. *Suppose the mesh $\mathcal{T}_h$ is not too fine with respect to $\mathcal{T}_H$. Then*

$$\begin{aligned} a_h^\gamma(e_H, e_H) &\le a_H^\gamma(e_H, e_H) + c\gamma \sum_{e\in\mathcal{E}_H^I\cup\mathcal{E}_H^D} h_e^{-1} |\,[e_H]\,|_e^2 \\ &= a_H^\gamma(e_H, e_H) + c\gamma \sum_{e\in\mathcal{E}_H^I} h_e^{-1} |\,[u_H^\gamma]\,|_e^2 + c\gamma \sum_{e\in\mathcal{E}_H^D} h_e^{-1} |g_D - u_H^\gamma|_e^2. \end{aligned} \tag{4.3}$$

Proof. Indeed, we have

$$a_h^\gamma(e_H, e_H) = \sum_{K\in\mathcal{T}_h} \|\nabla e_H\|_K^2 - \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \Big(2\,\langle\{\partial_n e_H\}, [e_H]\rangle_e - \gamma h_e^{-1} |\,[e_H]\,|_e^2\Big). \tag{4.4}$$

Since u is smooth and u_H^γ is a polynomial on each $K \in \mathcal{T}_H$, we have $\sum_{K\in\mathcal{T}_h} \|\nabla e_H\|_K^2 = \sum_{K\in\mathcal{T}_H} \|\nabla e_H\|_K^2$. Now if $e \in \mathcal{E}_h^I$ is a "completely" new edge, i.e., is in the interior of some $K \in \mathcal{T}_H$, then $[e_H]\big|_e = 0$. Also, since Γ is polygonal, edges $e \in \mathcal{E}_h^D$ are parts of edges in $\mathcal{E}_H^D$. Thus

$$\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \langle\{\partial_n e_H\}, [e_H]\rangle_e = \sum_{e\in\mathcal{E}_H^I\cup\mathcal{E}_H^D} \langle\{\partial_n e_H\}, [e_H]\rangle_e .$$

As for the terms in (4.4) that contain γ, the problem is to contend with the weights h_e^{-1} of the new edges. Again, there are no contributions from the completely new edges. Thus

$$\gamma \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} h_e^{-1} |[e_H]|_e^2 \le \gamma \sum_{e\in\mathcal{E}_H^I\cup\mathcal{E}_H^D} \nu(e) h_e^{-1} |[e_H]|_e^2,$$

where for $e \in \mathcal{E}_H^I \cup \mathcal{E}_H^D$, $\nu(e) = \max\{\frac{h_e}{h_{e'}} \mid e' \in \mathcal{E}_h^I \cup \mathcal{E}_h^D,\ e' \in e\} \ge 2$ is a number that measures the fineness of $\mathcal{T}_h$ with respect to $\mathcal{T}_H$. Assuming that $\nu(e)$ is uniformly bounded, i.e., $\mathcal{T}_h$ is not too fine relative to $\mathcal{T}_H$, we finally obtain (4.3). ∎

We next tackle the lack of coercivity of $a_h^\gamma(.,.)$ on the energy space E_h by showing that, nevertheless, as far as e_h is concerned, $a_h^\gamma(.,.)$ behaves like a norm!

PROPOSITION 4.2. *There exists a constant γ_2 depending only on r and θ_0 such that if $\gamma \ge \gamma_2$, then for some constant $C_1 > 0$ depending only on r and θ_0 there holds*

(4.5)
$$a_h^\gamma(e_h, e_h) \ge \frac{1}{2} \sum_{K\in\mathcal{T}_h} \|\nabla e_h\|_K^2 + C_1\gamma^2 \sum_{e\in\mathcal{E}_h^I} h_e^{-1} |[u_h^\gamma]|_e^2 + C_1\gamma^2 \sum_{e\in\mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2.$$

Proof. We have

$$a_h^\gamma(e_h, e_h) = \sum_{K\in\mathcal{T}_h} \|\nabla e_h\|_K^2 - \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \Big(2\langle\{\partial_n e_h\}, [e_h]\rangle_e - \gamma h_e^{-1} |[e_h]|_e^2\Big). \tag{4.6}$$

Moreover, for all $\chi \in V_h^r \cap H^1(\Omega)$ satisfying $\chi\big|_{\Gamma_D} = g_D$, we have

$$\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \langle\{\partial_n e_h\}, [e_h]\rangle_e = \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \langle\{\partial_n e_h\}, [\chi - u_h^\gamma]\rangle_e .$$

On the other hand, by virtue of the orthogonality identity (3.2),

$$0 = a_h^\gamma(e_h, \chi - u_h^\gamma) = \sum_{K\in\mathcal{T}_h} (\nabla e_h, \nabla(\chi - u_h^\gamma))_K - \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \Big(\langle\{\partial_n(\chi - u_h^\gamma)\}, [e_h]\rangle_e + \langle\{\partial_n e_h\}, [\chi - u_h^\gamma]\rangle_e - \gamma h_e^{-1} \langle[e_h], [\chi - u_h^\gamma]\rangle_e\Big).$$

Thus

$$\begin{aligned}\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \langle\{\partial_n e_h\}, [e_h]\rangle_e = {} & \sum_{K\in\mathcal{T}_h} (\nabla e_h, \nabla(\chi - u_h^\gamma))_K \\ & + \sum_{e\in\mathcal{E}_h^I} \Big(\langle\{\partial_n(\chi - u_h^\gamma)\}, [u_h^\gamma]\rangle_e + \gamma h_e^{-1} |[u_h^\gamma]|_e^2\Big) \\ & - \sum_{e\in\mathcal{E}_h^D} \Big(\langle\{\partial_n(\chi - u_h^\gamma)\}, g_D - u_h^\gamma\rangle_e - \gamma h_e^{-1} |g_D - u_h^\gamma|_e^2\Big).\end{aligned} \tag{4.7}$$

We now choose χ in (4.7) as in Theorem 2.1(ii). Also, using the trace and inverse inequalities, for any $\epsilon > 0$ we obtain

$$\left| \sum_{e \in \mathcal{E}_h^I \cup \mathcal{E}_h^D} \langle \{\partial_n e_h\}, [e_h] \rangle_e \right| \leq c\epsilon \sum_{K \in \mathcal{T}_h} \|\nabla e_h\|_K^2 + \left(\gamma + \frac{c}{\epsilon}\right) \sum_{e \in \mathcal{E}_h^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \left(\gamma + \frac{c}{\epsilon}\right) \sum_{e \in \mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2. \tag{4.8}$$

Using this in (4.6), we obtain

$$a_h^\gamma(e_h, e_h) \geq (1 - 2c\epsilon) \sum_{K \in \mathcal{T}_h} \|\nabla e_h\|_K^2 - \left(\gamma + \frac{2c}{\epsilon}\right) \left(\sum_{e \in \mathcal{E}_h^I} h_e^{-1} |[u_h^\gamma]|_e^2 + \sum_{e \in \mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2 \right). \tag{4.9}$$

Now note that the last two sums in (4.9) are dominated by $\frac{c}{\gamma^2} \sum_{K \in \mathcal{T}_h} \|\nabla e_h\|_K^2$ as shown by the a posteriori estimate (3.12). Hence, choosing $\epsilon = \frac{1}{8c}$ and then using (3.12), for γ sufficiently large we obtain the desired result. ∎

REMARK 4.1. *The proof of Proposition* 4.2 *also yields*

$$a_h^\gamma(e_h, e_h) \leq \left(1 + \frac{1}{\gamma}\right) \sum_{K \in \mathcal{T}_h} \|\nabla e_h\|_K^2 + C_2 \gamma \sum_{e \in \mathcal{E}_h^I} h_e^{-1} |[u_h^\gamma]|_e^2 + C_2 \gamma \sum_{e \in \mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2. \tag{4.10}$$

We now begin the proof of (4.1). Let $\mathcal{T}_h$ be a refinement of $\mathcal{T}_H$. Since $e_H \in E_H \subseteq E_h$, we integrate $\sum_{K \in \mathcal{T}_h} (\nabla e_H, \nabla v)_K$ by parts to obtain

$$\sum_{K \in \mathcal{T}_h} (\nabla e_H, \nabla v)_K = \sum_{K \in \mathcal{T}_h} (f + \Delta u_H^\gamma, v)_K + \sum_{e \in \mathcal{E}_h^I} \left(\langle \{\partial_n e_H\}, [v] \rangle_e + \langle \{v\}, [\partial_n e_H] \rangle_e \right) + \sum_{e \in \mathcal{E}_h^D} \langle \partial_n e_H, v \rangle_e + \sum_{e \in \mathcal{E}_h^N} \langle g_N - \partial_n u_H^\gamma, v \rangle_e \qquad \forall v \in E_h. \tag{4.11}$$

It then follows from (4.11) and the definition of $a_h^\gamma(\cdot, \cdot)$ that

$$\begin{aligned} \sum_{K \in \mathcal{T}_h} (f + \Delta u_H^\gamma, v)_K - \sum_{e \in \mathcal{E}_h^I} \langle [\partial_n u_H^\gamma], \{v\} \rangle_e + \sum_{e \in \mathcal{E}_h^N} \langle g_N - \partial_n u_H^\gamma, v \rangle_e &= a_h^\gamma(e_H, v) \\ &+ \sum_{e \in \mathcal{E}_h^I} \left(\langle \{\partial_n v\}, [e_H] \rangle_e - \gamma h_e^{-1} \langle [e_H], [v] \rangle_e \right) \\ &+ \sum_{e \in \mathcal{E}_h^D} \left(\langle \partial_n v, e_H \rangle_e - \gamma h_e^{-1} \langle e_H, v \rangle_e \right) \quad \forall v \in E_h. \end{aligned} \tag{4.12}$$

At this point we write $a_h^\gamma(e_H, v) = a_h^\gamma(u_h^\gamma - u_H^\gamma, v) + a_h^\gamma(e_h, v)$ and note that $a_h^\gamma(e_h, v) = 0 \; \forall v \in V_h^r$. Also, it turns out that it is crucial to eliminate the troublesome terms containing γ. These considerations lead us to use test functions from the subspaces $V_h^r \cap H^1_{0,\Gamma_D}$ of E_h encountered in the proof of Theorem 3.2. We then have

the key identity

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h}(f+\Delta u_H^\gamma, v)_K - \sum_{e\in\mathcal{E}_h^I}\langle[\partial_n u_H^\gamma],\{v\}\rangle_e + \sum_{e\in\mathcal{E}_h^N}\langle g_N-\partial_n u_H^\gamma, v\rangle_e \\
&\quad = a_h^\gamma(u_h^\gamma-u_H^\gamma,v) - \sum_{e\in\mathcal{E}_h^I}\langle\{\partial_n v\},[u_H^\gamma]\rangle_e + \sum_{e\in\mathcal{E}_h^D}\langle\partial_n v, g_D-u_H^\gamma\rangle_e \qquad (4.13)\\
&\quad = \sum_{K\in\mathcal{T}_h}(\nabla(u_h^\gamma-u_H^\gamma),\nabla v)_K - \sum_{e\in\mathcal{E}_h^I}\langle\{\partial_n v\},[u_h^\gamma]\rangle_e + \sum_{e\in\mathcal{E}_h^D}\langle\partial_n v, g_D-u_h^\gamma\rangle_e
\end{aligned}$$

$\forall v\in V_h^r\cap H^1_{0,\Gamma_D}$. The principal thrust of the proof of convergence is to use (4.13) to bound the terms $h_K^2\|f+\Delta u_H^\gamma\|_K^2$, $h_e|\,[\partial_n u_H^\gamma]\,|_e^2$, and $h_e|g_N-\partial_n u_H^\gamma|_e^2$ by an appropriate functional of $u_h^\gamma-u_H^\gamma$. This estimation is accomplished by, on the one hand, marking certain triangles and edges of $\mathcal{T}_H$ for refinement (marking strategy) and, on the other hand, ensuring that the test function space $V_h^r\cap H^1_{0,\Gamma_D}$ is large enough to yield the desired estimates. Consequently, the refinement must be done according to some specific rules (refinement strategy).

We next describe our marking strategy, which is modeled after the one in Dörfler [14].

Marking Strategy.

For some number $\theta\in(0,1)$, let $\mathcal{R}_H^K$, $\mathcal{R}_H^I$, and $\mathcal{R}_H^N$ be any subsets of $\mathcal{T}_H$, $\mathcal{E}_H^I$, and $\mathcal{E}_H^N$, respectively, such that

$$\begin{aligned}
\sum_{K\in\mathcal{R}_H^K} h_K^2\|f+\Delta u_H^\gamma\|_K^2 &\ge \theta\sum_{K\in\mathcal{T}_H} h_K^2\|f+\Delta u_H^\gamma\|_K^2,\\
\sum_{e\in\mathcal{R}_H^I} h_e|\,[\partial_n u_H^\gamma]\,|_e^2 &\ge \theta\sum_{e\in\mathcal{E}_H^I} h_e|\,[\partial_n u_H^\gamma]\,|_e^2,\\
\sum_{e\in\mathcal{R}_H^N} h_e|g_N-\partial_n u_H^\gamma|_e^2 &\ge \theta\sum_{e\in\mathcal{E}_H^N} h_e|g_N-\partial_n u_H^\gamma|_e^2.
\end{aligned}$$

With $E_{\mathcal{R}}$ and E denoting the sums on the left and right sides, respectively, we have

$$E_{\mathcal{R}}\ge\theta E. \qquad (4.14)$$

Refinement Strategy.

(I) A marked triangle $K\in\mathcal{R}_H^K$ will be cut into a number of equivalent triangles. This number depends on r, as shown in Figure 4.1.

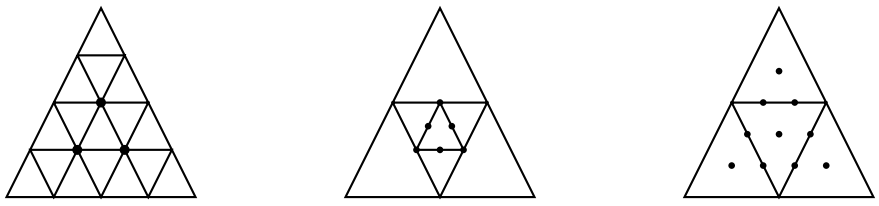


Fig. 4.1.

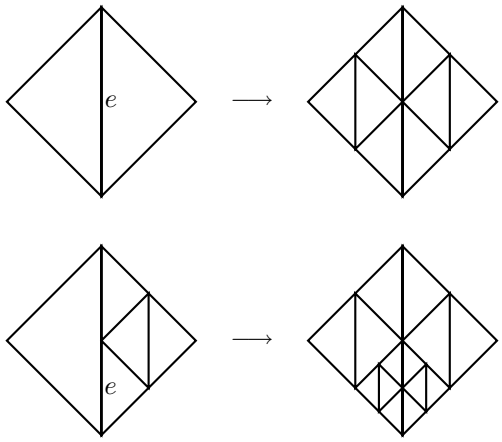


FIG. 4.2.

(II) Let $e = \partial K^+ \cap \partial K^- \in \mathcal{R}_H^I$ be a marked interior edge. Then one or both of K^+ and K^- will be cut in a manner depending on whether e is a full edge of both K^+ and K^- (see Figure 4.2).

(III) Let $e = \partial K \cap \Gamma_N$ be a marked edge in $\mathcal{R}_H^N$. Then K will be cut into four equivalent triangles.

REMARK 4.2. (i) *There may be some overlap between requirements* (I), (II), *and* (III).

(ii) *Additional requirements may also be imposed. For instance, one may wish to curtail the number of hanging nodes after refinement. Indeed, to simplify the programming we impose a maximum of one hanging node per interior edge. The combination of* (I), (II), *and such rules may lead to a finer mesh. This is acceptable since larger spaces* V_h^r *and* $V_h^r \cap H^1_{0,\Gamma_D}$ *will correspond to a finer mesh.*

Estimation of $h_K^2\|f + \Delta u_H^\gamma\|_K^2$. For $K \in \mathcal{R}_H^K$ consider the partition $\mathcal{T}_K$ shown in Figure 4.1, corresponding to a given r with the understanding that the eventual refinement of K may be finer than $\mathcal{T}_K$. We introduce the finite-dimensional spaces S_K given by

$$S_K = \{v \in C^0(K),\ v\big|_{K'} \in P_{r-1}(K')\ \forall K' \in \mathcal{T}_K,\ v = 0 \text{ on } \partial K\}.$$

It is clear that S_K is a subspace of $V_h^r \cap H^1_{0,\Gamma_D}$. Also, it is easily seen that a function in S_K is uniquely determined by its values at the nodes shown in Figure 4.1. Thus $\dim(S_K) \le d := r(r+1)/2 = \dim(P_{r-1}(K))$. Furthermore, for each r, a basis $\{\phi_i\}_{i=1}^d$ for S_K can be constructed by "gluing" together Lagrangian-type functions corresponding to the individual triangles in the partition $\mathcal{T}_K$. Indeed, it is not hard to show that the functions $\{\phi_i\}_{i=1}^d$ are linearly independent.

Now letting $\{\psi_i\}_{i=1}^d$ be the usual Lagrangian basis for $P_{r-1}(K)$ corresponding to the nodes shown in Figure 4.3, we form the "Gramian" matrix G given by $G_{ij} = (\phi_j, \psi_i)_K$, $i, j = 1, \ldots, d$. We have the following lemma.

LEMMA 4.1. *G is nonsingular.*

Proof. We will consider only the case $r = 2$; the remaining cases may be handled in a similar manner or verified by direct (and tedious) calculation. With ν_1, ν_2, ν_3 denoting the three nodes shown in Figure 4.1, let $\mathbf{v^2}, \mathbf{v^3}$ be the vectors emanating from

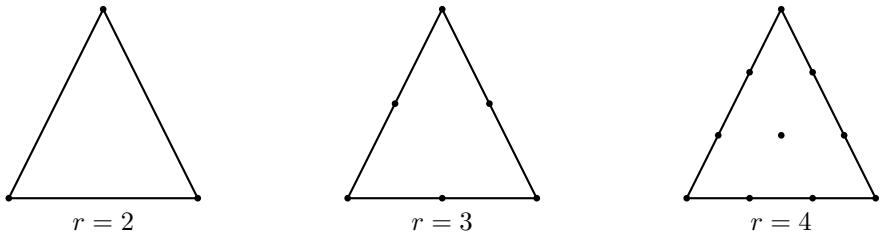


FIG. 4.3.

ν_1 and terminating at ν_2 and ν_3, respectively. Let also ϕ_1, ϕ_2, ϕ_3 be the (pyramidal) basis functions corresponding to the nodes ν_1, ν_2, ν_3 and denote their supports by S^1, S^2, S^3. Clearly,

$$\phi_2(x,y) = \phi_1(x - v_1^2, y - v_2^2) \quad \forall (x,y) \in S^2 \quad \text{and}$$
$$\phi_3(x,y) = \phi_1(x - v_1^3, y - v_2^3) \quad \forall (x,y) \in S^3.$$

Suppose there exists $\psi = ax + by + c \in P_1(K)$ such that $(\phi_j, \psi) = 0$, $j = 1,2,3$. We will show that $a = b = c = 0$, thus implying the linear independence of the rows of G.

$$\begin{aligned} 0 = (\psi, \phi_2)_K &= \int_{S^2} \psi(x,y)\phi_2(x,y)dxdy = \int_{S^2} \psi(x,y)\phi_1(x - v_1^2, y - v_2^2)dxdy \\ &= \int_{S^1} \psi(x + v_1^2, y + v_2^2)\phi_1(x,y)dxdy \\ &= \int_{S^1} \psi(x,y)\phi_1(x,y)dxdy + (av_1^2 + bv_2^2)\int_{S^1} \phi_1(x,y)dxdy. \end{aligned}$$

Now $\int_{S^1} \psi(x,y)\phi_1(x,y)dxdy = (\psi, \phi_1)_K = 0$. On the other hand, ϕ_1 is nonnegative and nonzero; thus we conclude from the above that $av_1^2 + bv_2^2 = 0$. In a similar way, we obtain $av_1^3 + bv_2^3 = 0$. Since the vectors $\mathbf{v^2}, \mathbf{v^3}$ are linearly independent, it follows that $a = b = 0$. Now that this has been shown, the fact that $c = 0$ readily follows from $(\psi, \phi_1)_K = 0$. ∎

COROLLARY 4.1. *Let $P : P_{r-1}(K) \to S_K$ denote the operator given by $(Pv, \chi)_K = (v, \chi)_K$ $\forall \chi \in S_K$. Then $\|P \cdot\|_K$ is a norm equivalent to $\|\cdot\|_K$ on $P_{r-1}(K)$ with constants that are independent of h_K.*

Proof. We only need to check the positivity of $\|P \cdot\|_K$ to see that it is a norm. Indeed, suppose $Pv = 0$ for some $v \in P_{r-1}(K)$. It then follows that $(v, \phi)_K = 0$ $\forall \phi \in S_K$. Since G is nonsingular, it follows that $v = 0$. The equivalence of the norms is a consequence of finite dimensionality. The fact that the constants involved are $O(1)$ follows from a scaling argument. ∎

To estimate $f + \Delta u_H^\gamma$ we take $v = P(f + \Delta u_H^\gamma)$ in (4.13). We get

$$\begin{aligned} \|P(f + \Delta u_H^\gamma)\|_K^2 &= (f + \Delta u_H^\gamma, P(f + \Delta u_H^\gamma))_K \\ &= \sum_{K' \in \mathcal{T}_{h,K}} (\nabla(u_h^\gamma - u_H^\gamma), \nabla P(f + \Delta u_H^\gamma))_{K'} \\ &\quad - \sum_{e \in \mathcal{E}_{h,K}^I} \langle \{\partial_n P(f + \Delta u_H^\gamma)\}, [u_h^\gamma] \rangle_e \\ &\quad + \sum_{e \in \mathcal{E}_h^D \cap \partial K} \langle \partial_n P(f + \Delta u_H^\gamma), g_D - u_h^\gamma \rangle_e, \end{aligned} \tag{4.15}$$

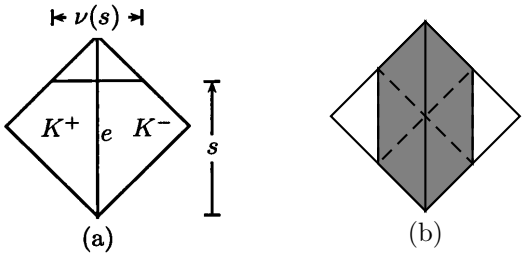


FIG. 4.4.

where $\mathcal{T}_{h,K} = \{K' \in \mathcal{T}_h,\ K' \subseteq K\}$ and $\mathcal{E}^I_{h,K} = \{e \in \mathcal{E}^I_h,\ e \subseteq K\}$.

Now using the trace and inverse inequalities for any $\epsilon > 0$, we have

$$\sum_{K' \in \mathcal{T}_{h,K}} (\nabla(u_h^\gamma - u_H^\gamma), \nabla P(f + \Delta u_H^\gamma))_{K'} \le c\epsilon \|f + \Delta u_H^\gamma\|_K^2 + \frac{c}{\epsilon} \sum_{K' \in \mathcal{T}_{h,K}} h_{K'}^{-2} \|\nabla(u_h^\gamma - u_H^\gamma)\|_{K'}^2. \tag{4.16}$$

Moreover,

$$\sum_{e \in \mathcal{E}^I_{h,K}} \langle \{\partial_n P(f + \Delta u_H^\gamma)\}, [u_h^\gamma] \rangle_e \le c\epsilon \|f + \Delta u_H^\gamma\|_K^2 + \frac{c}{\epsilon} \sum_{e \in \mathcal{E}^I_{h,K}} h_e^{-3} |[u_h^\gamma]|_e^2, \tag{4.17}$$

and

$$\sum_{e \in \mathcal{E}^D_h \cap \partial K} |\langle \partial_n P(f + \Delta u_H^\gamma), g_D - u_h^\gamma \rangle_e| \le c\epsilon \|f + \Delta u_H^\gamma\|_K^2 + \frac{c}{\epsilon} \sum_{e \in \mathcal{E}^D_h \cap \partial K} h_e^{-3} |g_D - u_h^\gamma|_e^2. \tag{4.18}$$

Now using (4.16), (4.17), and (4.18) with a small ϵ in (4.15), it follows from Corollary 4.1 that

$$h_K^2 \|f + \Delta u_H^\gamma\|_K^2 \le c \sum_{K' \in \mathcal{T}_{h,K}} \|\nabla(u_h^\gamma - u_H^\gamma)\|_{K'}^2 + c \sum_{e \in \mathcal{E}^I_{h,K}} h_e^{-1} |[u_h^\gamma]|_e^2 + \sum_{e \in \mathcal{E}^D_h \cap \partial K} h_e^{-1} |g_D - u_h^\gamma|_e^2. \tag{4.19}$$

Estimation of $h_e |[\partial_n u_H^\gamma]|_e^2$. Let $e \in \mathcal{R}^I_H$ be a marked edge. It follows from the refinement strategy (see Figure 4.2) that e is a full edge of both K^+ and K^-, where K^+, K^- may belong to $\mathcal{T}_H$ or one of them at most may have been formed after refinement. We construct a test function $v \in V_h^r \cap H^1_{0,\Gamma_D}$ as follows:

(i) Let $\tilde{v}$ be the extension of $[\partial_n u_H^\gamma]|_e$ to $\tilde{K} := K^+ \cup K^-$ by constants along lines normal to e.

(ii) Let ℓ be the continuous piecewise linear function whose support is the shaded region in Figure 4.4(b) and which assumes the value 1 at the midpoint of e.

We take $v = \tilde{v}\ell$. Note that v belongs to $V_h^r \cap H^1_{0,\Gamma_D}$. Using this v in (4.13), we obtain

$$\begin{aligned}\langle [\partial_n u_H^\gamma], \{v\}\rangle_e = &\sum_{K'\in\mathcal{T}_{h,\tilde{K}}} \Big((f+\Delta u_H^\gamma, v)_{K'} - (\nabla(u_h^\gamma - u_H^\gamma), \nabla v)_{K'}\Big) \\ &+ \sum_{e\in\mathcal{E}^I_{h,\tilde{K}}} \langle \{\partial_n v\}, [u_h^\gamma]\rangle_e - \sum_{e\in\mathcal{E}_h^D\cap\partial\tilde{K}} \langle \partial_n v, g_D - u_h^\gamma\rangle_e,\end{aligned} \tag{4.20}$$

where $\mathcal{T}_{h,\tilde{K}} = \{K' \in \mathcal{T}_h,\ K' \subseteq \tilde{K}\}$ and $\mathcal{E}^I_{h,\tilde{K}} = \{e \in \mathcal{E}_h^I,\ e \subseteq \tilde{K}\}$. Now note that

$$\langle [\partial_n u_H^\gamma], \{v\}\rangle_e = \int_e |\,[\partial_n u_H^\gamma]\,|^2 \ell(s) ds. \tag{4.21}$$

With ℓ acting as a weight function, we have

$$h_e \int_e |\,[\partial_n u_H^\gamma]\,|^2 \ell(s) ds \geq c h_e |\,[\partial_n u_H^\gamma]\,|_e^2, \tag{4.22}$$

where c is independent of h_e. Moreover, since $0 \leq \ell \leq 1$ and $\tilde{v}$ is constant along lines normal to e,

$$\|v\|^2_{\tilde{K}} \leq \|\tilde{v}\|^2_{\tilde{K}} = \int_e |\,[\partial_n u_H^\gamma]\,|^2 \nu(s) ds \leq c h_e |\,[\partial_n u_H^\gamma]\,|_e^2, \tag{4.23}$$

where ν is as in Figure 4.4(a). Now using the trace and inverse inequalities in (4.20), for any $\epsilon > 0$ we obtain

$$\begin{aligned}h_e |\,[\partial_n u_H^\gamma]\,|_e^2 \leq c\epsilon \|v\|^2_{\tilde{K}} + \frac{c}{\epsilon}\Bigg(&\sum_{K'\in\mathcal{T}_{h,\tilde{K}}} \Big(h^2_{K'}\|f+\Delta u_H^\gamma\|^2_{K'} + \|\nabla(u_h^\gamma - u_H^\gamma)\|^2_{K'}\Big) \\ &+ \sum_{e\in\mathcal{E}^I_{h,\tilde{K}}} h_e^{-1} |\,[u_h^\gamma]\,|_e^2 + \sum_{e\in\mathcal{E}_h^D\cap\partial\tilde{K}} h_e^{-1} |g_D - u_h^\gamma|_e^2\Bigg).\end{aligned} \tag{4.24}$$

Hence using (4.21)–(4.23) in (4.24), and choosing ϵ sufficiently small, we arrive at

$$\begin{aligned}h_e |\,[\partial_n u_H^\gamma]\,|_e^2 \leq c &\sum_{K'\in\mathcal{T}_{h,\tilde{K}}} \Big(h^2_{K'}\|f+\Delta u_H^\gamma\|^2_{K'} + \|\nabla(u_h^\gamma - u_H^\gamma)\|^2_{K'}\Big) \\ &+ c \sum_{e\in\mathcal{E}^I_{h,\tilde{K}}} h_e^{-1} |\,[u_h^\gamma]\,|_e^2 + c \sum_{e\in\mathcal{E}_h^D\cap\partial\tilde{K}} h_e^{-1} |g_D - u_h^\gamma|_e^2.\end{aligned} \tag{4.25}$$

Estimation of $h_e|g_N - \partial_n u_H^\gamma|_e^2$. We define a test function $v \in V_h^r \cap H^1_{0,\Gamma_D}$ as follows: Let $\tilde{v}$ be the extension of $g_N - \partial_n u_H^\gamma$ by constants along lines normal to e, and let ℓ be the continuous piecewise linear function that vanishes outside of K and that assumes the value 1 at the midpoint of e. We let $v = \tilde{v}\ell$ and note that since $g_N - \partial_n u_H^\gamma$ is a polynomial of degree less than or equal to $r-2$, $v \in V_h^r \cap H^1_{0,\Gamma_D}$. Using this v in (4.13), we obtain

$$\begin{aligned}\langle g_N - \partial_n u_H^\gamma, v\rangle_e = &-(f+\Delta u_H^\gamma, v)_K + \sum_{K'\in\mathcal{T}_{h,K}} (\nabla(u_h^\gamma - u_H^\gamma), \nabla v)_{K'} \\ &- \sum_{e\in\mathcal{E}^I_{h,K}} \langle \{\partial_n v\}, [u_h^\gamma]\rangle_e.\end{aligned} \tag{4.26}$$

Using the trace and inverse inequalities, we obtain

$$h_e \langle g_N - \partial_n u_H^\gamma, v\rangle_e \le c\epsilon \|v\|_K^2 + \frac{c}{\epsilon}\left(h_K^2 \|f + \Delta u_H^\gamma\|_K^2 + \sum_{K' \in \mathcal{T}_{h,K}} \|\nabla(u_h^\gamma - u_H^\gamma)\|_{K'}^2 + \sum_{e \in \mathcal{E}_{h,K}^I} h_e^{-1} |\,[u_h^\gamma]\,|_e^2 \right). \tag{4.27}$$

As in steps (4.21), (4.22), and (4.23), we have

$$h_e |g_N - \partial_n u_H^\gamma|_e^2 \le c h_e \langle g_N - \partial_n u_H^\gamma, v\rangle_e \quad \text{and} \quad \|v\|_K^2 \le c h_e |g_N - \partial_n u_H^\gamma|_e^2. \tag{4.28}$$

Thus, from (4.27) it follows that

$$h_e |g_N - \partial_n u_H^\gamma|_e^2 \le c h_K^2 \|f + \Delta u_H^\gamma\|_K^2 + c \sum_{K' \in \mathcal{T}_{h,K}} \|\nabla(u_h^\gamma - u_H^\gamma)\|_{K'}^2 + c \sum_{e \in \mathcal{E}_{h,K}^I} h_e^{-1} |\,[u_h^\gamma]\,|_e^2. \tag{4.29}$$

We are now ready to state and prove the main result of this paper.

THEOREM 4.1. *Let u_h^γ and u_H^γ denote the DG solutions in V_h^r and V_H^r, respectively, and e_H and e_h the corresponding errors. Assume that*

(i) *The data of the BVP (1.1)–(1.3) is such that $f \in P_{r-1}(\Omega)$, $g_D \in P_{r-1}(\Gamma_D)$, and $g_N \in P_{r-2}(\Gamma_N)$.*
(ii) *$\mathcal{T}_h$ is not too fine with respect to $\mathcal{T}_H$.*
(iii) *For some $\theta \in (0,1)$ the marking of triangles and edges of the mesh $\mathcal{T}_H$ and their refinement is done according to the rules specified above.*

Then, there exists γ_3 depending only on r, θ_0, and θ such that for all $\gamma \ge \gamma_3$, (4.1) holds with ρ given by (4.36).

Proof. First, using the trace and inverse inequalities, we have for $\gamma \ge \gamma_4(r, \theta_0)$

$$a_h^\gamma(u_h^\gamma - u_H^\gamma, u_h^\gamma - u_H^\gamma) \ge \frac{1}{2} \sum_{K \in \mathcal{T}_h} \|\nabla(u_h^\gamma - u_H^\gamma)\|_K^2 + \frac{1}{2} \sum_{e \in \mathcal{E}_h^I \cup \mathcal{E}_h^D} h_e^{-1} |\,[u_h^\gamma - u_H^\gamma]\,|_e^2. \tag{4.30}$$

On the other hand, from (4.19), (4.25), and (4.29), it follows that for some constant $C_3 > 0$ depending only on r and θ_0

$$E_{\mathcal{R}} \le C_3 \left(\sum_{K \in \mathcal{T}_h} \|\nabla(u_h^\gamma - u_H^\gamma)\|_K^2 + \sum_{e \in \mathcal{E}_h^I} h_e^{-1} |\,[u_h^\gamma]\,|_e^2 + \sum_{e \in \mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2 \right). \tag{4.31}$$

Next, using (4.2), (4.3), (4.30), and (4.31), we obtain

$$\begin{aligned} a_H^\gamma(e_H, e_H) + c\gamma \sum_{e \in \mathcal{E}_H^I} h_e^{-1} |\,[u_H^\gamma]\,|_e^2 + c\gamma \sum_{e \in \mathcal{E}_H^D} h_e^{-1} |g_D - u_H^\gamma|_e^2 &\ge a_h^\gamma(e_H, e_H) \\ &= a_h^\gamma(e_h, e_h) + a_h^\gamma(u_h^\gamma - u_H^\gamma, u_h^\gamma - u_H^\gamma) \\ &\ge a_h^\gamma(e_h, e_h) + \frac{1}{2} \sum_{K \in \mathcal{T}_h} \|\nabla(u_h^\gamma - u_H^\gamma)\|_K^2 \\ &\ge a_h^\gamma(e_h, e_h) + \frac{E_{\mathcal{R}}}{2C_3} - \frac{1}{2} \left(\sum_{e \in \mathcal{E}_h^I} h_e^{-1} |\,[u_h^\gamma]\,|_e^2 + \sum_{e \in \mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2 \right). \end{aligned} \tag{4.32}$$

Now from (3.20) it follows for $\gamma \geq \gamma_5(r, \theta_0, \theta)$ that

$$c\gamma \left(\sum_{e \in \mathcal{E}_H^I} h_e^{-1} |\, [u_H^\gamma]\, |_e^2 + \sum_{e \in \mathcal{E}_H^D} h_e^{-1} |g_D - u_H^\gamma|_e^2 \right) \leq \frac{cE}{\gamma} \leq \frac{\theta E}{4C_3}. \tag{4.33}$$

Also, from (4.5) we have

$$\sum_{e \in \mathcal{E}_h^I} h_e^{-1} |\, [u_h^\gamma]\, |_e^2 + \sum_{e \in \mathcal{E}_h^D} h_e^{-1} |g_D - u_h^\gamma|_e^2 \leq \frac{1}{C_1 \gamma^2} a_h^\gamma(e_h, e_h). \tag{4.34}$$

Thus, using (4.33) and (4.34) in (4.32), we obtain

$$a_H^\gamma(e_H, e_H) \geq \left(1 - \frac{1}{2C_1\gamma^2}\right) a_h^\gamma(e_h, e_h) + \frac{\theta E}{4c_3}. \tag{4.35}$$

We choose γ large so that $1 - \frac{1}{2C_1\gamma^2} > 0$. On the other hand, using (4.10) with H instead of h (recall that this result holds for a generic mesh), it follows from (3.20) and (3.3) that for some constant $C_4 > 0$ depending only on r and θ_0 one has

$$E \geq C_4 a_H^\gamma(e_H, e_H).$$

Using this in (4.35), it follows that

$$\left(1 - \frac{\theta C_4}{4C_3}\right) a_H^\gamma(e_H, e_H) \geq \left(1 - \frac{1}{2C_1\gamma^2}\right) a_h^\gamma(e_h, e_h).$$

If $\frac{\theta C_4}{4C_3} \geq 1$, then this means that $a_h^\gamma(e_h, e_h) = 0$. If, on the other hand, $0 < \frac{\theta C_4}{4C_3} < 1$, then we obtain (4.1) with ρ given by

$$\rho = \frac{1 - \frac{\theta C_4}{4C_3}}{1 - \frac{1}{2C_1\gamma^2}}. \tag{4.36}$$

The conclusion of the theorem now follows for γ sufficiently large. ∎

REMARK 4.3. *The conditions on the data of the BVP* (1.1)–(1.3) *are restrictive and are the price paid to simplify the proofs. We believe that they can be relaxed or dispensed with by introducing appropriate projections of the data functions. See, e.g.,* [15] *and* [20]. *The generalization of our results including an accounting for the effects of quadrature errors is being pursued.*

5. Numerical experiments. In this section we present the results of some numerical experiments to exhibit the performance of the adaptive strategy outlined in section 4. We used the Baker version of the method since the forms $\{\partial_n v\} = \nabla v^+ \cdot \mathbf{n}^+$ are easier to implement.

As a representative of a problem with a smooth solution we chose

$$-\Delta u = 2\pi^2 \sin \pi x \sin \pi y \quad \text{in } \Omega = [0,1]^2, \quad u = 0 \quad \text{on } \partial\Omega, \tag{P1}$$

with $u = \sin \pi x \sin \pi y$. The next problem has the smooth but oscillatory solution $u = \sin 8\pi x \sin 8\pi y$:

$$-\Delta u = 128\pi^2 \sin 8\pi x \sin 8\pi y \quad \text{in } \Omega = [0,1]^2, \quad u = 0 \quad \text{on } \partial\Omega. \tag{P2}$$

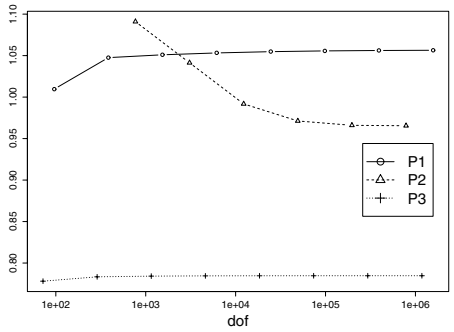


FIG. 5.1. *Effectivity indices;* $r = 3$, $\gamma = 6.3$.

Finally, as an example of a solution with a singularity we considered the problem

$$-\Delta u = 0 \quad \text{in } \Omega, \quad u = g_D \quad \text{on } \Gamma_D = \partial\Omega, \tag{P3}$$

where Ω is the polygon with vertices $(0,0)$, $(-1,-1)$, $(1,-1)$, $(1,1)$, $(-1,1)$, $(0,0)$ and has a reentrant corner at $(0,0)$. The datum g_D is adjusted so that the solution is $u = r^{2/3}\sin\frac{2\theta}{3}$ in polar coordinates.

We generated an adaptive code written in the C language and ran the experiments on a workstation with an Intel Pentium 4 chip rated at 3.06 GHz. The linear systems were solved by Multigrid with point Gauss–Seidel smoothing as a preconditioner for the conjugate gradient method. To assess the performance of the estimator and the adaptive algorithm, we monitored the following three quantities:

$$a_h^\gamma(e,e) = \sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2 - 2\sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \langle\{\partial_n e\},[e]\rangle_e + \gamma(r-1)^2 \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} h_e^{-1}|[e]|_e^2,$$

$$\|e\|_{1,h} = \left(\sum_{K\in\mathcal{T}_h} \|\nabla e\|_K^2 + \sum_{e\in\mathcal{E}_h^I\cup\mathcal{E}_h^D} \left(h_e\,|\{\partial_n e\}|_e^2 + \gamma(r-1)^2 h_e^{-1}|[e]|_e^2\right)\right)^{1/2},$$

$$\eta = \left(\sum_{K\in\mathcal{T}_h} h_K^2\|f+\Delta u_h^\gamma\|_K^2 + \sum_{e\in\mathcal{E}_h^I} h_e\left|\left[\partial_n u_h^\gamma\right]\right|_e^2 + \sum_{e\in\mathcal{E}_h^N} h_e\left|g_N - \partial_n u_h^\gamma\right|_e^2 \right.$$
$$\left. + \gamma^2(r-1)^4 \sum_{e\in\mathcal{E}_h^I} h_e^{-1}\left|[u_h^\gamma]\right|_e^2 + \gamma^2(r-1)^4 \sum_{e\in\mathcal{E}_h^D} h_e^{-1}\left|g_D - u_h^\gamma\right|_e^2\right)^{1/2}.$$

These quantities are modified versions of the bilinear form, the energy norm, and the residual error estimator. Since the coercivity threshold is known to increase quadratically as a function of the degree $r-1$, we replaced γ by $\gamma(r-1)^2$. This way, the calculations could be performed without the need for adjusting γ with r. Following the same reasoning, we attached $\gamma^2(r-1)^4$ to the jump terms of the residual estimator since γ^2 accompanied these terms both in the upper and lower bounds.

The first set of experiments concerned a study of the effectivity index $\eta/\|e\|_{1,h}$ as a function of the degrees of freedom (dof's). Figure 5.1 shows the effectivity indices for all three test problems. Starting with an initial mesh of 16 triangles (96 dof's), the mesh was refined uniformly until a maximum of about 10^6. In particular, the indices behaved rather well with values close to 1 (more so for (P1) and (P2) than

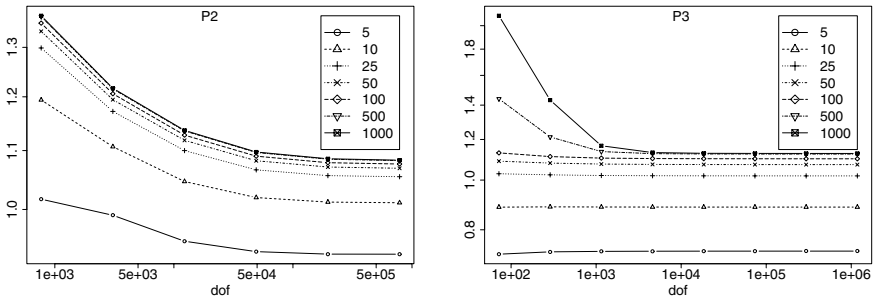


FIG. 5.2. *Dependence of effectivity indices on γ; $r = 3$.*

TABLE 5.1
Total CPU time (sec).

θ	P1	P2	P3
0.3	65	95	35
0.5	24	34	17
0.9	11	13	11

TABLE 5.2
Adaptive iterations.

θ	P1	P2	P3
0.3	108	75	89
0.5	40	26	39
0.9	9	6	14

TABLE 5.3
Total triangles in final mesh.

θ	P1	P2	P3
0.3	12853	16943	4458
0.5	12823	17780	5064
0.9	20170	22394	8208

for (P3)), and the index for (P2) took longer to stabilize given the oscillatory nature of the solution. Similar behavior was observed for $r = 2, 4, 5$.

We also wanted to study the effect of γ on the effectivity indices. Figure 5.2 shows the results of experiments concerning test problems (P2) and (P3), $r = 3$, and values of γ from 5 to 1000. While such an effect does indeed exist, it is nevertheless quite mild, as evidenced by the narrow range of the changes in the effectivity indices. We also note that the effectivity indices seem to be convergent as γ increases. Similar results were obtained for (P1) and $r = 2, 4, 5$.

The remaining experiments were devoted to the validation of the convergence characteristics of the adaptive algorithm. We ran several experiments with $r = 2, 3, 4$, all three test problems, and several values of θ and γ. In all cases all three quantities $a_h^\gamma(e, e)$, $\|e\|_{1,h}$, and η decreased monotonically. We should also mention that in order to simplify the program, we cut the marked triangles into four triangles only, in variance with the patterns shown in Figure 4.1. The plots in Figure 5.3 show the excellent agreement between the error $\|e\|_{1,h}$ and the estimator η. On the other hand, the bilinear form $a_h^\gamma(e, e)$ seems to follow a very similar but parallel trajectory, evidence of its equivalence to the other two. The two plots of Figure 5.4 show the corresponding final meshes for (P2) and (P3), respectively.

Next, we wanted to study the effect of the choice of θ on the performance of the adaptive algorithm. Indeed, the experiments indicate that while convergence is not in doubt, the patterns of refinement are strongly influenced by this choice, as evidenced by the final triangle count and, more importantly, the CPU time. Postponing a detailed study of this important issue to a future work, we nevertheless maintain that if we accept the criterion that the most efficient algorithm is the one with the least execution time, then larger values of θ should be preferred. Tables 5.1–5.3 show, respectively, the CPU time, the number of iterations to convergence, and the triangle count in the final mesh for the three test problems and $\theta = 0.3, 0.5, 0.9$. While smaller values of θ lead to a smaller number of triangles, they are up to six times costlier in CPU time. This is due to the fact that at every cycle, relatively few triangles and

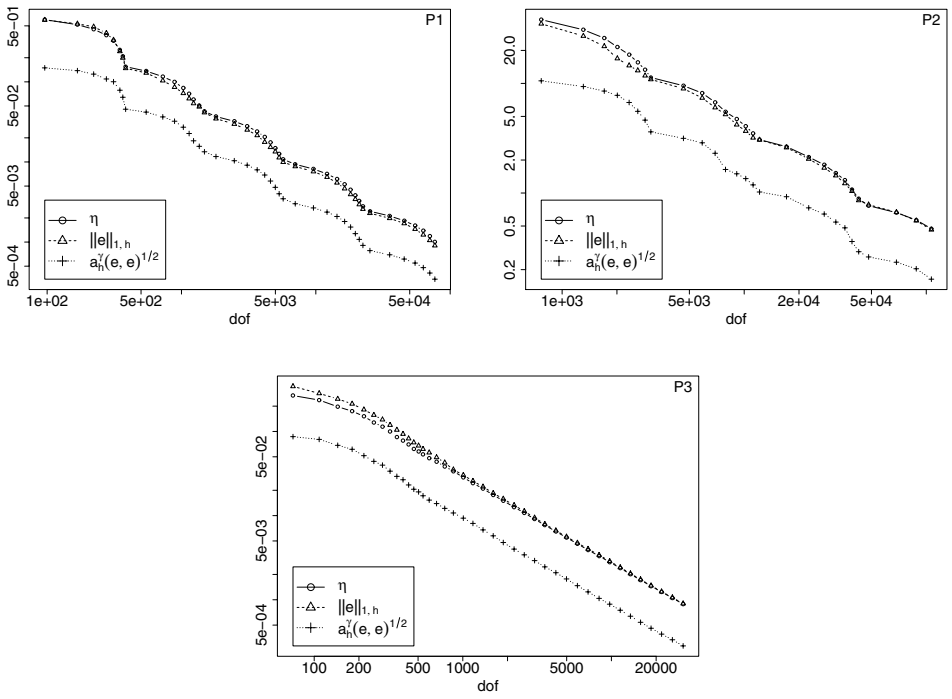


FIG. 5.3. $r = 3$, $\gamma = 6.3$, $\theta = 0.5$.

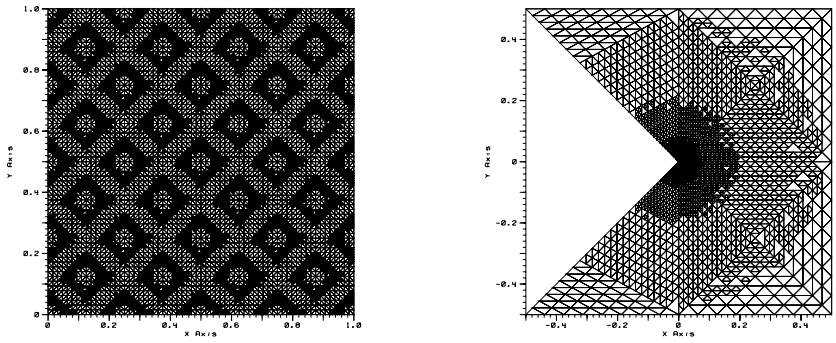


FIG. 5.4. *Final meshes for* (P2), (P3)*;* $r = 3$, $\gamma = 6.3$, $\theta = 0.5$.

edges are refined, resulting in a large, one could say unacceptable, number of cycles.

Tables 5.4, 5.5, and 5.6 encapsulate the results of an attempt to study the effect of θ on both the number of refinement levels and the distribution of triangles over the levels; in a sense they provide a spectral analysis of the mesh hierarchy. For test problem (P1), the value $\theta = 0.9$ caused a shift of the refinement to a higher level with a substantial number of triangles on level 6 (Table 5.4). On the other hand, Table 5.6 shows the opposite behavior for test problem (P3), whereby the smaller values of $\theta = 0.3, 0.5$ created five additional levels, albeit with a relatively small number of additional triangles.

These experiments, although limited in scope, provide a validation of the theo-

TABLE 5.4
(P1) *($|\mathcal{T}_0| = 16$) Level leaf triangle distribution.*

Level / θ	0.3	0.5	0.9
0–3	0	0	0
4	1177	1187	36
5	11676	11636	14942
6	-	-	5192

TABLE 5.5
(P2) *($|\mathcal{T}_0| = 128$) Level leaf triangle distribution.*

Level / θ	0.3	0.5	0.9
0–2	0	0	0
3	5275	4996	3458
4	11668	12784	18936

TABLE 5.6
(P3) *($|\mathcal{T}_0| = 12$) Level leaf triangle distribution.*

Level / θ	0.3	0.5	0.9
0–2	0	0	0
3	374	314	57
4	1311	1496	2256
5	873	1065	1955
6	625	715	1316
7	409	483	910
8	275	325	587
9	192	208	416
10	120	148	281
11	78	92	186
12	60	66	110
13	42	50	94
14	18	18	40
15	18	18	-
16	18	18	-
17	18	18	-
18	23	22	-
19	4	8	-

retical results of the paper. They also point to the importance of further exploration of the mechanisms of marking and refinement. In particular, a static choice of θ is far from being satisfactory and must be replaced by a more dynamic (adaptive!) mechanism.

Acknowledgments. The authors thank Mr. Mike Saum for help in the development of the code and the generation of tables and figures. The initial meshes were generated by the program "Triangle" developed by J. R. Shewchuk [22].

REFERENCES

[1] R. A. ADAMS, *Sobolev Spaces*, Academic Press, New York, 1975.
[2] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.
[3] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[4] I. Babuşka and M. Vogelius, *Feedback and adaptive finite element solution of one dimensional boundary value problems*, Numer. Math., 44 (1984), pp. 75–102.
[5] G. A. Baker, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.
[6] G. A. Baker, W. N. Jureidini, and O. A. Karakashian, *Piecewise solenoidal vector fields and the Stokes problem*, SIAM J. Numer. Anal., 27 (1990), pp. 1466–1485.
[7] R. Becker, P. Hansbo, and M. Larson, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 723–733.
[8] R. Becker and R. Rannacher, *A feedback approach to error control in finite element methods: Basic analysis and examples*, East-West J. Numer. Math., 4 (1996), pp. 237–264.
[9] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.
[10] S. Brenner and R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[11] R. Bustinza, G. Gatica, and B. Cockburn, *An a posteriori error estimate for the local discontinuous Galerkin method applied to linear and nonlinear diffusion problems*, J. Sci. Comput., 22/23 (2005), pp. 147–185.
[12] P. Castillo, *An a posteriori error estimate for the local discontinuous Galerkin method*, J. Sci. Comput., 22/23 (2005), pp. 187–204.
[13] E. Creusé and S. Nicaise, *Anisotropic a posteriori error estimation for the mixed discontinuous Galerkin approximation of the Stokes problem*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 449–483.
[14] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
[15] P. Houston, D. Schötzau, and T. Wihler, *Energy Norm A Posteriori Error Estimation of hp-Adaptive Discontinuous Galerkin Methods for Elliptic Problems*, IMA Preprint Series 1985, 2004.
[16] G. Kanschat and R. Rannacher, *Local error analysis of the interior penalty discontinuous Galerkin method for second order elliptic problems*, J. Numer. Math., 10 (2002), pp. 249–274.
[17] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.
[18] O. Karakashian and F. Pascal, *Adaptive discontinuous Galerkin approximations of second-order elliptic problems*, in Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2004), P. Neittaanmäki, T. Rossi, S. Korotov, E. Oñate, J. Périaux, and D. Knörzer, eds., Jyväskylä, Finland, 2004.
[19] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
[20] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.
[21] B. Rivière and M. Wheeler, *A posteriori error estimates for a discontinuous Galerkin method applied to elliptic problems*, Comput. Math. Appl., 46 (2003), pp. 141–163.
[22] J. R. Shewchuk, *Triangle: Engineering a 2D quality mesh generator and Delauney triangulator*, in Applied Computational Geometry: Towards Geometric Engineering, Lecture Notes in Comput. Sci. 1148, M. C. Lin and D. Manocha, eds., Springer-Verlag, Berlin, 1996, pp. 203–222.
[23] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Wiley-Teubner, New York, 1996.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 666–697

OPTIMIZED SCHWARZ WAVEFORM RELAXATION METHODS FOR ADVECTION REACTION DIFFUSION PROBLEMS*

M. J. GANDER† AND L. HALPERN‡

Abstract. We study in this paper a new class of waveform relaxation algorithms for large systems of ordinary differential equations arising from discretizations of partial differential equations of advection reaction diffusion type. We show that the transmission conditions between the subsystems have a tremendous influence on the convergence speed of the waveform relaxation algorithms, and we identify transmission conditions with optimal performance. Since these optimal transmission conditions are expensive to use, we introduce a class of local transmission conditions of Robin type, which approximate the optimal ones and can be used at the same cost as the classical transmission conditions. We determine the transmission conditions in this class with the best performance of the associated waveform relaxation algorithm. We show that the new algorithm is well posed and converges much faster than the classical one. We illustrate our analysis with numerical experiments.

Key words. domain decomposition, waveform relaxation, Schwarz methods, time parallelism

AMS subject classification. 65M12

DOI. 10.1137/050642137

1. Introduction.

Waveform relaxation algorithms have been invented to solve extremely large systems of ordinary differential equations arising in circuit simulation [26]. They use a partition of the original problem into subproblems, which are then solved separately, and an iteration with information exchange between subproblems leads to the solution of the original problem. Since the solution of the subproblems can be done in parallel, these algorithms are very well suited for implementations on parallel computers. The main drawback of waveform relaxation algorithms is in general their slow convergence, for a review, see [1].

There are two main classical approaches in the literature to solve parabolic problems in parallel. The first approach consists of discretizing the equations uniformly in time with an implicit scheme and then applying a domain decomposition technique to the elliptic problems obtained at each time step, see, for example, [2, 32, 3] and references therein. This approach has the disadvantage that a uniform time discretization needs to be enforced across different subdomains, and one thus loses one of the main features of domain decomposition algorithms, namely, to be able to treat different subdomains numerically differently with an adapted procedure for each subdomain. A second disadvantage is that small amounts of information need to be exchanged at every time level, which can be costly in a parallel environment where the startup cost to send information is significant. In addition, the iteration in time cannot proceed before all the subdomains have converged. An interesting variant, which avoids iterating by making explicit predictions at the interfaces, can be found in [36].

The second classical approach consists of discretizing the equations in space first and then applying a waveform relaxation algorithm to the large system of ordinary

*Received by the editors October 7, 2005; accepted for publication (in revised form) October 6, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sinum/45-2/64213.html
†Section de Mathématiques, Université de Genève, 2-4 rue du Lièvre, CP 64, CH-1211 Genève, Switzerland (martin.gander@unige.ch).
‡LAGA, Institut Galilée, Université Paris XIII, 93430 Villetaneuse, France (halpern@math.univ-paris13.fr).

differential equations obtained from the spatial discretization. A formulation using discretized subdomains can be found in [25]. The main disadvantage of this approach is that spatial information of the connectivity of the subsystems in the large system of ordinary differential equations is lost, and parameters like physical overlap and information exchange are difficult to interpret in this context. Using a different approach and abandoning the idea of subsystems, efficient waveform relaxation algorithms of multigrid type, see [29, 38, 21, 22], and also convolution waveform relaxation algorithms, see [20, 23], have been introduced and analyzed.

To avoid the inherent problems of the classical decomposition approaches, waveform relaxation algorithms for problems in space-time were formulated in [16, 14, 13] and independently in [18] directly at the continous level without discretization. The spatial domain is decomposed into subdomains, and time dependent problems are solved iteratively on subdomains, exchanging information at the interfaces between subdomains. This approach permits a different numerical treatment in both space and time of the subdomain problems, and information is exchanged less often between subdomains. The iteration is defined as in the classical Schwarz case, but as in waveform relaxation, time dependent subproblems are solved, which explains the names of these methods. Unfortunately these algorithms, although robust with respect to refinement, if the overlap is held constant, are still converging only slowly.

We show in this paper for a model problem of advection reaction diffusion type why the convergence of the Schwarz waveform relaxation algorithm is slow. By analyzing the convergence behavior of the classical overlapping Schwarz waveform relaxation algorithm applied to the model problem, we show that the classical algorithm is using ineffective transmission conditions. The classical transmission conditions inhibit the information exchange between subdomains and therefore slow down the convergence of the algorithm. Using ideas introduced in [10], we derive optimal transmission conditions for the Schwarz waveform relaxation algorithm. These transmission conditions coincide with the transparent boundary conditions used to truncate computational domains, which were first studied in [7] for hyperbolic problems and in [19] for advection diffusion problems. Transparent transmission conditions lead to Schwarz waveform relaxation algorithms which converge in a finite number of steps, equal to the number of subdomains, see [11] for the wave equation case. In general, however, the transparent boundary conditions are expensive to compute since they involve nonlocal operators. Similar to the approach for stationary problems in [34, 24, 8, 17], we approximate the transparent transmission conditions locally at the interfaces between subdomains, see [10, 12, 31]. This leads to algorithms which converge even without overlap and are well suited to couple different numerical methods, like finite element and finite differences methods, see [4]. We then optimize the convergence rate, including an overlap in the optimization if desired.

Since the algorithms are eventually discretized to be used on a parallel computer, we analyze the performance of the optimized algorithms asymptotically as the discretization parameter goes to zero. This analysis reveals an interesting relationship between the space-time discretization (implicit-explicit) and the convergence of the optimized algorithm. Numerical experiments show that the convergence rates are improved by orders of magnitude over the rate of the classical overlapping Schwarz waveform relaxation methods.

This paper is organized as follows: In section 2, we present the model problem for which we study the overlapping Schwarz waveform relaxation algorithm in what follows. We include fundamental existence results for the solution, which are later used to prove well posedness and convergence of the algorithm. In section 3, we intro-

duce the classical overlapping Schwarz waveform relaxation algorithm, show that it is well posed when applied to the advection reaction diffusion equation, and analyze its convergence. In section 4, we introduce the optimal Schwarz waveform relaxation algorithm, an algorithm that uses, instead of Dirichlet transmission conditions, transparent ones. Because such transmission conditions can be expensive to use, we also introduce local approximations of these transmission conditions. The core of this paper is contained in section 5, where we analyze the optimized Schwarz waveform relaxation algorithm with Robin transmission conditions. We show that the algorithm is well posed and convergent. We also derive the optimal parameters in the Robin transmission conditions and their dependence on the problem parameters, and we study the asymptotic dependence of the discretized algorithm on the mesh parameters. In section 6, we show numerical results for the classical and optimized Schwarz waveform relaxation algorithms, which show how drastically the convergence behavior is improved using optimized transmission conditions. We present our conclusions in section 7. All our analysis is performed for the simple case of a two subdomain decomposition, since we improve the algorithm locally between subdomains. We show, however, numerical experiments for more than two subdomains, which indicate that the results of our analysis are valid in that case as well.

2. Model problem and function spaces. Our guiding example is the one dimensional advection reaction diffusion equation

$$\mathcal{L}u := \partial_t u - \nu \partial_{xx} u + a\partial_x u + bu = f \quad \text{in } \Omega \times (0,T), \tag{2.1}$$

where $\Omega = \mathbb{R}$, $\nu > 0$, and a and b are constants which do not both vanish simultaneously. The case of the heat equation needs special treatment and can be found in [12]. Without loss of generality, we can assume that the advection coefficient a is nonnegative since $a < 0$ amounts to changing x into $-x$. We can also assume that the reaction coefficient b is nonnegative. If not, a change of variables $v = ue^{-\sigma t}$ with $\sigma + b > 0$ will lead to (2.1) with a positive reaction coefficient.

A weak solution of (2.1) for $\Omega = \mathbb{R}$ is defined to be a $u \in \mathcal{C}(0,T;L^2(\Omega)) \cap L^2(0,T;H^1(\Omega))$ such that for any v in $H^1(\Omega)$

$$\frac{d}{dt}(u,v) + \nu(\partial_x u, \partial_x v) + \frac{a}{2}((\partial_x u, v) - (\partial_x v, u)) + b(u,v) = (f,v) \text{ in } \mathcal{D}'(0,T), \tag{2.2}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$. Problem (2.1) is completed by the initial condition

$$u(x,0) = u_0(x) \quad \text{in } \Omega. \tag{2.3}$$

We have a first existence and uniqueness result, proved in [27].

THEOREM 2.1. *For $\Omega = \mathbb{R}$, if the initial value u_0 is in $L^2(\Omega)$ and the right-hand side f is in $L^2(0,T;L^2(\Omega))$, then there exists a unique weak solution u of* (2.1), (2.3).

With the transmission conditions we will introduce later, we will need more regularity in our analysis, in the anisotropic Sobolev spaces $H^{r,s}(\Omega \times (0,T)) = L^2(0,T;H^r(\Omega)) \cap H^s(0,T;L^2(\Omega))$ defined in [27].

THEOREM 2.2. *For $\Omega = \mathbb{R}$, if the initial value u_0 is in $H^1(\Omega)$ and the right-hand side f is in $L^2(0,T;L^2(\Omega))$, then the weak solution u of* (2.1), (2.3) *is in $H^{2,1}(\Omega \times (0,T))$. If u_0 is in $H^2(\Omega)$ and f is in $H^{1,1/2}(\Omega \times (0,T))$, then u is in $H^{3,3/2}(\Omega \times (0,T))$.*

For the proof, and the trace theorems in $H^{r,s}$, we refer to [27].

3. Classical Schwarz waveform relaxation. We decompose the spatial domain $\Omega = \mathbb{R}$ into two overlapping subdomains $\Omega_1 = (-\infty, L)$ and $\Omega_2 = (0, \infty)$, $L > 0$. The overlapping Schwarz waveform relaxation algorithm consists then of solving iteratively subproblems on $\Omega_1 \times (0,T)$ and $\Omega_2 \times (0,T)$ using as a boundary condition at the interfaces $x = 0$ and $x = L$ the values obtained from the previous iteration. The algorithm is thus for iteration index $k = 1, 2, \ldots$, given by

$$
\begin{array}{llll}
\mathcal{L}u_1^k = f & \text{in } \Omega_1 \times (0,T), & \mathcal{L}u_2^k = f & \text{in } \Omega_2 \times (0,T), \\
u_1^k(\cdot,0) = u_0 & \text{in } \Omega_1, & u_2^k(\cdot,0) = u_0 & \text{in } \Omega_2, \\
u_1^k(L,\cdot) = u_2^{k-1}(L,\cdot) & \text{in } (0,T), & u_2^k(0,\cdot) = u_1^{k-1}(0,\cdot) & \text{in } (0,T),
\end{array} \tag{3.1}
$$

where an initial guess $u_1^0(0,t)$ and $u_2^0(L,t)$, $t \in (0,T)$, needs to be provided. We first study the well posedness of algorithm (3.1) and then analyze its convergence properties. While algorithm (3.1) is also well defined without overlap, $L = 0$, it is not convergent, since no information is exchanged in that case. This will be different for the optimized algorithms proposed in section 5.

3.1. Well posedness of the algorithm. We first need to show the well posedness of each subdomain problem. Without loss of generality, we consider the subdomain problem on Ω_1 only:

$$
\begin{array}{rcll}
\mathcal{L}v & = & f & \text{in } \Omega_1 \times (0,T), \\
v(\cdot,0) & = & u_0 & \text{in } \Omega_1, \\
v(L,\cdot) & = & g & \text{in } (0,T).
\end{array} \tag{3.2}
$$

THEOREM 3.1. *If $f \in L^2(0,T;L^2(\Omega_1))$, $u_0 \in H^1(\Omega_1)$, and $g \in H^{\frac{3}{4}}(0,T)$ and if the compatibility condition*

$$u_0(L) = g(0) \tag{3.3}$$

is satisfied, then problem (3.2) *has a unique solution v in $H^{2,1}(\Omega_1 \times (0,T))$. Moreover $v(0,\cdot)$ is in $H^{\frac{3}{4}}(0,T)$, and the following compatibility property holds:*

$$\lim_{t \to 0+} v(0,t) = u_0(0). \tag{3.4}$$

Proof. The proof of existence and uniqueness in $H^{2,1}(\Omega_1 \times 0,T))$ can be found in [27]. The compatibility relation follows from the trace theorem in [27]. ∎

By the Sobolev embedding theorem, u_0 is continuous on $\bar{\Omega}_1$ and g is continuous on $[0,T]$, which gives a classical meaning to the compatibility condition (3.3). The preceding result ensures that the subdomain problems are well posed in the classical algorithm, provided the initial and boundary conditions satisfy the compatibility condition (3.3) for each iteration step.

To show that this is indeed the case, let $u_2^0(L,\cdot)$ and $u_1^0(0,\cdot)$ be given in $H^{\frac{3}{4}}(0,T)$ such that $u_2^0(L,\cdot) = u_0(L)$ and $u_1^0(0,\cdot) = u_0(0)$. Then, by Theorem 3.1, the first iteration of the overlapping Schwarz waveform relaxation algorithm (3.1) defines a unique first iterate (u_1^1, u_2^1) in $H^{2,1}(\Omega_1 \times (0,T)) \times H^{2,1}(\Omega_2 \times (0,T))$. Furthermore, $u_1^1(0,\cdot)$ and $u_2^1(L,\cdot)$ are in $H^{\frac{3}{4}}(0,T)$, $\lim_{t\to 0+} u_1^1(0,t) = u_0(0)$, and $\lim_{t\to 0+} u_2^1(L,t) = u_0(L)$. Hence by induction, the algorithm is well posed.

3.2. Convergence of the algorithm. By linearity, the error between the solution u and the iterates u_j^k, $j = 1, 2$, of the overlapping Schwarz waveform relaxation algorithm (3.1) satisfies a homogeneous advection reaction diffusion equation with a

homogeneous initial condition. We therefore study in what follows the homogeneous problem with data on the interfaces only. Let h_L and h_0 be given in $H^{\frac{3}{4}}(0,T)$ with $h_L(0)=0$ and $h_0(0)=0$, to satisfy the compatibility conditions, and let (e_1,e_2) be the solution in $H^{2,1}(\Omega_1\times(0,T))\times H^{2,1}(\Omega_2\times(0,T))$ of the equations

$$\begin{array}{rclll} \mathcal{L}e_1 & = & 0 & \text{in } \Omega_1\times(0,T), & \mathcal{L}e_2 = 0 \quad \text{in } \Omega_2\times(0,T),\\ e_1(\cdot,0) & = & 0 & \text{in } \Omega_1, & e_2(\cdot,0) = 0 \quad \text{in } \Omega_2,\\ e_1(L,\cdot) & = & h_L & \text{in } (0,T), & e_2(0,\cdot) = h_0 \quad \text{in } (0,T). \end{array} \tag{3.5}$$

Our analysis is based on the Fourier transform, which we denote for any function $h\in L^2(\mathbb{R})$ by $\hat{h}:=\mathcal{F}h$. We define the one-sided space ${}_0H^{\frac{3}{4}}(0,T)=\{\varphi\in H^{\frac{3}{4}}(0,T),\varphi(0)=0\}$, equipped with the norm $\|\varphi\|_{{}_0H^{\frac{3}{4}}(0,T)}:=\inf\{\|\Phi\|_{H^{\frac{3}{4}}(\mathbb{R})},\Phi=\varphi \text{ a.e. in } (0,T),\Phi=0 \text{ a.e. in } (-\infty,0)\}$.

LEMMA 3.2. *Let $L>0$. If $a>0$ or $a=0$ and $b>0$, then the map $\mathcal{G}_D$ associated with equations* (3.5)*,*

$$\mathcal{G}_D:(h_L,h_0)\mapsto(e_2(L,\cdot),e_1(0,\cdot)), \tag{3.6}$$

is defined from $({}_0H^{\frac{3}{4}}(0,T))^2$ into itself, and $\mathcal{G}_D^2$ is a strict contraction on $({}_0H^{\frac{3}{4}}(0,T))^2$.

Proof. Since h_L and h_0 are in ${}_0H^{\frac{3}{4}}(0,T)$, we can extend them in $H^{\frac{3}{4}}(\mathbb{R})$ to obtain $\tilde{h}_L$ and $\tilde{h}_0$ vanishing on $(-\infty,0)$. We then extend equations (3.5) in time to $\mathbb{R}$, and their solution coincides with (e_1,e_2) on $(0,T)$. Therefore, we still call it (e_1,e_2). By Fourier transform in time, we find in each subdomain the same ordinary differential equation

$$i\omega\hat{e}_j-\nu\partial_{xx}\hat{e}_j+a\partial_x\hat{e}_j+b\hat{e}_j=0,\quad j=1,2, \tag{3.7}$$

with the characteristic roots

$$r^+=\frac{a+\sqrt{d}}{2\nu},\quad r^-=\frac{a-\sqrt{d}}{2\nu},\quad d=a^2+4\nu(b+i\omega), \tag{3.8}$$

where $\sqrt{d}$ is the complex square root with positive real part. Therefore, $\Re(r^+)>0$ and $\Re(r^-)<0$, and we find, using that e_j is in $L^2(\Omega_j)$,

$$\hat{e}_1(x,\omega)=\mathcal{F}\tilde{h}_L(\omega)e^{r^+(x-L)},\qquad \hat{e}_2(x,\omega)=\mathcal{F}\tilde{h}_0(\omega)e^{r^-x}. \tag{3.9}$$

On the interfaces of the subdomains, we therefore have

$$\mathcal{F}(\mathcal{G}_D(\tilde{h}_L,\tilde{h}_0))(\omega)=(\mathcal{F}\tilde{h}_0(\omega)e^{r^-L},\mathcal{F}\tilde{h}_L(\omega)e^{-r^+L}).$$

Since $\tilde{h}_0$ and $\tilde{h}_L$ vanish in $\mathbb{R}_-$, their Fourier transforms are analytic in the half-plane $\Im\tau<0$, and by (3.9) and the Paley–Wiener theorem [37], $e_1(0,\cdot)$ and $e_2(L,\cdot)$ vanish in $\mathbb{R}_-$. Since they are in $H^{\frac{3}{4}}(\mathbb{R})$, they are continuous, and therefore $e_2(L,0)=0$ and $e_1(0,0)=0$: the map $\mathcal{G}_D$ maps $({}_0H^{\frac{3}{4}}(0,T))^2$ into itself. We have furthermore

$$\mathcal{F}(\mathcal{G}_D^2(\tilde{h}_L,\tilde{h}_0))(\omega)=e^{(r^--r^+)L}(\mathcal{F}\tilde{h}_0(\omega),\mathcal{F}\tilde{h}_L(\omega)). \tag{3.10}$$

Denoting by

$$C_D:=\sup_{\omega\in\mathbb{R}}e^{(r^--r^+)L}=e^{-\frac{L}{\nu}(\sqrt{a^2+4\nu b})}, \tag{3.11}$$

we get for any extension $(\tilde{h}_0, \tilde{h}_L)$ of (h_0, h_L)

$$\|\mathcal{G}_D^2(h_L, h_0)\|_{({}_0H^{\frac{3}{4}}(0,T))^2} \le \|\mathcal{G}_D^2(\tilde{h}_L, \tilde{h}_0)\|_{(H^{\frac{3}{4}}(\mathbb{R}))^2} \le C_D \|(\tilde{h}_L, \tilde{h}_0)\|_{(H^{\frac{3}{4}}(\mathbb{R}))^2}.$$

Taking the infimum on all the extensions on the right-hand side, we get

$$\|\mathcal{G}_D^2(h_L, h_0)\|_{({}_0H^{\frac{3}{4}}(0,T))^2} \le C_D \|(h_L, h_0)\|_{({}_0H^{\frac{3}{4}}(0,T))^2},$$

and since C_D is positive and strictly less than 1, we have proved that $\mathcal{G}_D^2$ is a contraction. ∎

We now prove convergence of the overlapping Schwarz waveform relaxation algorithm.

THEOREM 3.3. *Let $L > 0$. For $a > 0$ or $a = 0$ and $b > 0$, the iterates (u_1^k, u_2^k) of algorithm (3.1) converge to the solution of (2.1), (2.3) for any initial guess g_0 and g_L in $H^{\frac{3}{4}}(0,T)$ such that $g_0(0) = u_0(0)$ and $g_L(0) = u_0(L)$.*

Proof. The errors $e_j^k = u_j^k - u$, $j = 1, 2$, satisfy for $k \ge 1$ (3.1) with $f = 0$ and $u_0 = 0$. For positive k, we introduce the interface functions $h_L^k = e_2^k(L, \cdot)$ and $h_0^k = e_1^k(0, \cdot)$ and denote by $h_0^0 = h_0$ and $h_L^0 = h_L$. Using the map $\mathcal{G}_D$, we obtain by induction

$$(h_L^{2k}, h_0^{2k}) = \mathcal{G}_D^{2k}(h_L^0, h_0^0),$$

and thus by Lemma 3.2

$$\|(h_L^{2k}, h_0^{2k})\|_{({}_0H^{\frac{3}{4}}(0,T))^2} \le C_D^k \|(h_L^0, h_0^0)\|_{({}_0H^{\frac{3}{4}}(0,T))^2},$$

with C_D given in (3.11). Solving (3.5) and using (3.9), we obtain for e_1

$$\|e_1\|^2_{L^2(0,T;H^2(\Omega_1))} \le \int_{-\infty}^{\infty} |r^+|^4 \int_{\Omega_1} e^{2\Re(r^+)(x-L)} |\mathcal{F}\tilde{h}_L|^2 dx\, d\omega = \int_{-\infty}^{\infty} \frac{|r^+|^4}{2\Re(r^+)} |\mathcal{F}\tilde{h}_L|^2 d\omega.$$

For $a > 0$ or $a = 0$ and $b > 0$, the denominator in the factor in front of $|\mathcal{F}\tilde{h}_L|^2$ is bounded from below, and for $|\omega|$ large, the factor behaves like $|\omega|^{3/2}$. Therefore

$$\|e_1\|_{L^2(0,T;H^2(\Omega_1))} \le C \|\tilde{h}_L\|_{H^{\frac{3}{4}}(\mathbb{R})},$$

and the same result also holds for $\|e_1\|_{H^1(0,T;L^2(\Omega_1))}$. Hence

$$\|e_1\|_{H^{2,1}((\Omega_1)\times(0,T))} \le M_D \|h_L\|_{{}_0H^{\frac{3}{4}}(0,T)}, \tag{3.12}$$

and similarly for e_2. Now we apply (3.12) to the errors e_1^{2k+1} and e_2^{2k+1} in the iteration and obtain

$$\begin{aligned}&\|(e_1^{2k+1}, e_2^{2k+1})\|_{H^{2,1}(\Omega_1\times(0,T))\times H^{2,1}(\Omega_2\times(0,T))} \\ &\quad \le M_D \|(h_L^{2k}, h_0^{2k})\|_{({}_0H^{\frac{3}{4}}(0,T))^2} \le M_D C_D^k \|(g_L - u(L,\cdot), g_0 - u(0,\cdot))\|_{({}_0H^{\frac{3}{4}}(0,T))^2},\end{aligned}$$

which together with Lemma 3.2 completes the proof. A similar argument also holds for even iteration numbers. ∎

Theorem 3.3 shows that the overlapping Schwarz waveform relaxation algorithm converges and that the convergence rate is at least linear and is independent of the length of the time interval. It does however depend on the problem parameters ν,

a, and b and the overlap L. Using also the preceding Lemma 3.2, the error in the overlapping Schwarz waveform relaxation algorithm satisfies on the interfaces over a double iteration step in Fourier space relation (3.10) or equivalently

$$\hat{e}_1^{k+1}(L,\omega) = \rho_D \hat{e}_1^{k-1}(L,\omega), \quad \hat{e}_2^{k+1}(0,\omega) = \rho_D \hat{e}_2^{k-1}(0,\omega), \tag{3.13}$$

where the convergence factor $\rho_D = \rho_D(\omega, L, \nu, a, b)$ is given by

$$\rho_D(\omega, L, \nu, a, b) := e^{(r^- - r^+)L} = e^{-\frac{\sqrt{a^2+4\nu(b+i\omega)}}{\nu}L}. \tag{3.14}$$

Note that the convergence factor ρ_D is uniformly bounded in modulus for all ω by a quantity strictly less than 1,

(3.15)

$$R_D(\omega, L, \nu, a, b) := |\rho_D(\omega, L, \nu, a, b)| \le \bar{R}_D(L, \nu, a, b) := R_D(0, L, \nu, a, b) = e^{-\frac{\sqrt{a^2+4\nu b}}{\nu}L},$$

and for L small, we have

$$\bar{R}_D = 1 - \frac{\sqrt{a^2+4\nu b}}{\nu}L + O(L^2). \tag{3.16}$$

Using the convergence factor ρ_D from Fourier analysis allows us to obtain a sharper convergence result for bounded time intervals.

THEOREM 3.4 (superlinear convergence). *For the advection reaction diffusion equation on a bounded time interval $(0,T)$, the asymptotic convergence rate of the overlapping Schwarz waveform relaxation algorithm* (3.1) *is superlinear:*

$$||e_j^{2k}(0,\cdot)||_{L^\infty(0,T)} \le \operatorname{erfc}\left(\frac{kL}{\sqrt{\nu T}}\right) ||e_j^0(0,\cdot)||_{L^\infty(0,T)}, \quad j = 1,2,$$

where the error function complement is defined by $\operatorname{erfc}(x) := \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-s^2} ds$.

Proof. By induction on the relations (3.13), we obtain

$$\hat{e}_1^{2k}(0,\omega) = \rho_D^k \hat{e}_1^0(0,\omega), \qquad \hat{e}_2^{2k}(L,\omega) = \rho_D^k \hat{e}_2^0(L,\omega). \tag{3.17}$$

Using the inverse Fourier transform and the convolution theorem, we find

$$e_1^{2k}(0,t) = (\mathcal{F}^{-1}\rho_D^k) * e_1^0(0,t), \qquad e_2^{2k}(L,t) = (\mathcal{F}^{-1}\rho_D^k) * e_2^0(L,t). \tag{3.18}$$

Now the inverse Fourier transform of ρ_D^k is

$$\mathcal{F}^{-1}\rho_D^k = \frac{kL}{\sqrt{\nu\pi}t^{3/2}} e^{-\frac{(kL)^2}{\nu t} - \left(\frac{a^2}{4\nu}+b\right)t},$$

and we can therefore estimate for $j = 1, 2$

$$\|e_j^{2k}(0,\cdot)\|_{L^\infty(0,T)} \le \|\mathcal{F}^{-1}\rho_D^k\|_{L^1(0,T)} \|e_j^0(0,\cdot)\|_{L^\infty(0,T)} \le \operatorname{erfc}\left(\frac{kL}{\sqrt{\nu T}}\right) \|e_j^0(0,\cdot)\|_{L^\infty(0,T)},$$

where the last inequality follows from estimating the term $e^{-(\frac{a^2}{4\nu}+b)t}$ by 1. By a similar argument for the second subdomain, the result follows. ∎

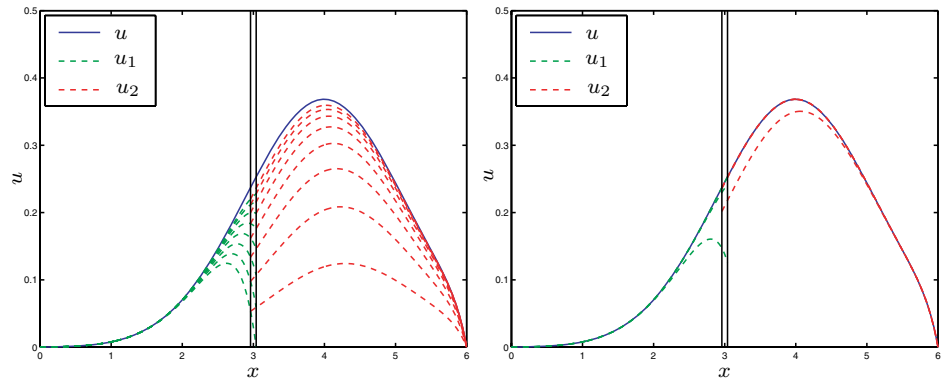


FIG. 3.1. *On the left, the first few iterates of the classical Schwarz waveform relaxation algorithm (dashed) at the end of the time interval $t = T$, together with the exact solution (solid), and on the right the first iterates of the new optimized Schwarz waveform relaxation algorithm.*

This result was first proved for bounded domains in [18], see also [15]. It also holds in higher dimensions and for general decompositions, for the heat equation, see [14], and for advection diffusion, see [5]. The result differs significantly from the classical linear convergence result of the overlapping Schwarz method for elliptic problems and also from the classical superlinear convergence results for waveform relaxation, which is slower, see [35]. Furthermore, one can show that the convergence rate depends only on the number of subdomains in higher order terms, see [15], and hence coarse grid preconditioners are not necessary for evolution problems of this type.

The Dirichlet transmission conditions at the interfaces are however responsible for slow convergence in the classical Schwarz waveform relaxation algorithm: in Figure 3.1 on the left, we show the first few iterations at the end of the time interval for a model problem. On the right, we show the first few iterations of the new, much faster algorithm we will develop in what follows.

4. Optimal Schwarz waveform relaxation. We now introduce transmission conditions which are more effective for the information exchange between subdomains. The new algorithm is

(4.1)
$$\begin{aligned} \mathcal{L}u_1^k &= f \quad \text{in } \Omega_1 \times (0,T), & \mathcal{L}u_2^k &= f \quad \text{in } \Omega_2 \times (0,T), \\ u_1^k(\cdot,0) &= u_0, & u_2^k(\cdot,0) &= u_0, \\ (\partial_x + \mathcal{S}_1)u_1^k(L,\cdot) &= (\partial_x + \mathcal{S}_1)u_2^{k-1}(L,\cdot), & (\partial_x + \mathcal{S}_2)u_2^k(0,\cdot) &= (\partial_x + \mathcal{S}_2)u_1^{k-1}(0,\cdot), \end{aligned}$$

where $\mathcal{S}_1$ and $\mathcal{S}_2$ are linear operators in time, possibly pseudodifferential.

4.1. Optimal transmission conditions. The following theorem gives the optimal choice for $\mathcal{S}_1$ and $\mathcal{S}_2$.

THEOREM 4.1. *For $a > 0$ or $a = 0$ and $b > 0$, algorithm (4.1) converges to the solution u of (2.1) in two iterations for all initial guesses $u_1^0 \in H^{2,1}(\Omega_1 \times (0,T))$ and $u_2^0 \in H^{2,1}(\Omega_2 \times (0,T))$, independently of the size of the overlap $L \geq 0$, if and only if the operators $\mathcal{S}_1$ and $\mathcal{S}_2$ have the corresponding symbols*

(4.2)
$$\sigma_1 = -r^-, \quad \sigma_2 = -r^+,$$

where $r^\pm$ are defined in (3.8).

Proof. Using the Fourier transform with parameter ω as in Lemma 3.2, we find for the error

$$\hat{e}_1^k(x,\omega) = \alpha^k(\omega)e^{r^+(x-L)}, \quad \hat{e}_2^k(x,\omega) = \beta^k(\omega)e^{r^-x}, \quad k \geq 1, \tag{4.3}$$

where α^k and β^k are constants, which, using the new transmission conditions, satisfy for $k \geq 1$ the recurrence relation

$$\begin{aligned} \alpha^{k+1}(r^+ + \sigma_1) &= \beta^k(r^- + \sigma_1)e^{r^-L}, \\ \beta^{k+1}(r^- + \sigma_2) &= \alpha^k(r^+ + \sigma_2)e^{-r^+L}. \end{aligned} \tag{4.4}$$

Now for an arbitrary initial guess u_1^0 and u_2^0, the coefficients α^1 and β^1 will in general not vanish. Since $r^- + \sigma_1 = r^+ + \sigma_2 = 0$ implies $r^+ + \sigma_1 \neq 0$ and $r^- + \sigma_2 \neq 0$, we obtain from (4.4) that α^2 and β^2 are identically zero if and only if $r^- + \sigma_1 = r^+ + \sigma_2 = 0$. ∎

Note that the symbols σ_1, σ_2 given in (4.2) are not polynomials in $i\omega$, and hence the optimal corresponding transmission operators $\mathcal{S}_1$, $\mathcal{S}_2$ are nonlocal operators in time; they correspond to integral transfer operators in time along the interfaces between subdomains. Even though such operators can be efficiently implemented, see, for example, [30], they are more costly than local transfer operators and the latter are in general preferred. It is therefore of interest to approximate the nonlocal operators by local ones, whose symbols are polynomials in $i\omega$. Using each equation in (4.4) at iteration k in the other one at iteration $k+1$, we find

$$\alpha^{k+1} = \rho\alpha^{k-1}, \quad \beta^{k+1} = \rho\beta^{k-1}$$

with the new convergence factor

$$\rho = \frac{r^- + \sigma_1}{r^+ + \sigma_1} \cdot \frac{r^+ + \sigma_2}{r^- + \sigma_2} e^{(r^- - r^+)L}. \tag{4.5}$$

4.2. Approximations of the optimal transmission conditions. We approximate the symbols σ_1 and σ_2 from (4.2) corresponding to the optimal transmission operators by constants, which leads to Robin transmission conditions in the Schwarz waveform relaxation algorithm (4.1), i.e.,

$$\mathcal{S}_1 := \frac{-a+p}{2\nu}, \qquad \mathcal{S}_2 := \frac{-a-p}{2\nu}. \tag{4.6}$$

The choice of the parameter p is restricted by the requirement that the subdomain problems need to be well posed, and a good choice should lead to a rapidly converging algorithm; both issues we will analyze in detail in the following section.

Notice that using the knowledge of the symbols (4.2) of the optimal transmission conditions, we have chosen a particular form for the low order approximation, leading to (4.6). In general one is not required to do so; in particular, we could have chosen, for example, two different parameters p in (4.6), which would have given more freedom in the optimization process we study later, or even chosen a different p at each iteration, as it was done for a steady problem in [9]. One could also choose higher order transmission conditions, i.e., approximations by polynomials in $i\omega$. Having one parameter only however greatly simplifies the optimization process, so we leave the more general cases for future studies.

5. Optimized Schwarz waveform relaxation. We now study the Schwarz waveform relaxation algorithm with Robin transmission conditions. We start with the overlapping case, $L > 0$. We first show under what conditions on the free parameter p the algorithm is well posed and then prove convergence of the algorithm for a general choice of p satisfying these conditions. We also study the influence of p on the performance of the algorithm and propose two choices for p: one choice motivated by a low frequency approximation and a second choice which optimizes the performance of the algorithm. We then prove that the algorithm converges also without overlap, and we again study the influence of p on the performance of the nonoverlapping algorithm.

5.1. Well posedness of the algorithm. As in the case of the classical Schwarz waveform relaxation algorithm studied in section 3, we first need to analyze under which conditions the subdomain problems of the algorithm with Robin transmission conditions is well posed. Without loss of generality, we study only the well posedness of the subdomain problem on Ω_1:

$$\begin{aligned} \mathcal{L}v &= f && \text{in } \Omega_1 \times (0,T), \\ v(\cdot,0) &= u_0 && \text{in } \Omega_1, \\ (\partial_x v + \mathcal{S}_1 v)(L,\cdot) &= g_L && \text{in } (0,T). \end{aligned} \tag{5.1}$$

We first show an extension result, which allows us to reduce the study of the well posedness to the case with homogeneous initial and boundary conditions.

LEMMA 5.1. *If u_0 is in $H^1(\Omega_1)$ and g_L is in $H^{\frac{1}{4}}(0,T)$, then there exists an extension w in $H^{2,1}(\Omega_1 \times (0,T))$ such that $w(\cdot,0) = u_0$ in Ω_1 and $(\partial_x w + \mathcal{S}_1 w)(L,\cdot) = g_L$ on $(0,T)$.*

Proof. Let $\tilde{g}_L$ be in $H^{\frac{3}{4}}(0,T)$ such that $\tilde{g}_L(0) = u_0(L)$. By the continuous extension theorem, there exists a w_1 in $H^{2,1}(\Omega \times (0,T))$ such that

$$w_1(\cdot,0) = u_0, \quad w_1(L,\cdot) = \tilde{g}_L, \quad \partial_x w_1(L,\cdot) = 0$$

and a w_2 in $H^{2,1}(\Omega \times (0,T))$ such that

$$w_2(\cdot,0) = 0, \quad w_2(L,\cdot) = 0, \quad \partial_x w_2(L,\cdot) = g_L - \mathcal{S}_1 \tilde{g}_L.$$

Now the sum $w := w_1 + w_2$ is the desired extension in $H^{2,1}(\Omega \times (0,T))$. ∎

Thus it suffices to analyze the well posedness of the problem with homogeneous initial and boundary conditions:

$$\begin{aligned} \mathcal{L}\tilde{v} &= F && \text{in } \Omega_1 \times (0,T), \\ \tilde{v}(\cdot,0) &= 0 && \text{in } \Omega_1, \\ (\partial_x \tilde{v} + \mathcal{S}_1 \tilde{v})(L,\cdot) &= 0 && \text{in } (0,T), \end{aligned} \tag{5.2}$$

where $\tilde{v} = v - w$ and the right-hand side function $F = f - \mathcal{L}w$ is in $L^2(0,T;L^2(\Omega_1))$ if f is in $L^2(0,T;L^2(\Omega_1))$. We start with the weak formulation: for any φ in $H^1(\Omega_1)$, we multiply the equation by φ, integrate, and use Green's formula and the boundary condition to obtain in $\mathcal{D}'(0,T)$

$$\frac{d}{dt}(\tilde{v},\varphi) + \nu(\partial_x \tilde{v}, \partial_x \varphi) + \frac{a}{2}((\partial_x \tilde{v},\varphi) - (\partial_x \varphi,\tilde{v})) + b(\tilde{v},\varphi) + \frac{p}{2}\tilde{v}(L)\varphi(L) = (F,\varphi). \tag{5.3}$$

The following Theorem gives existence, uniqueness, and regularity of the weak solution.

THEOREM 5.2. *Suppose F is in $L^2(0,T;L^2(\Omega_1))$. Then, for any p, problem* (5.2) *has a unique weak solution $\tilde{v}$ in $H^{2,1}(\Omega_1 \times (0,T))$.*

Proof. The proof is based on a priori estimates.

1. Multiplying equation (5.2) by $\tilde{v}$, integrating in space, and using the boundary condition, we obtain

$$\frac{1}{2}\frac{d}{dt}\|\tilde{v}\|^2 + \nu\|\partial_x\tilde{v}\|^2 + b\|\tilde{v}\|^2 + \frac{p}{2}\tilde{v}^2(L) = (F,\tilde{v}). \tag{5.4}$$

(a) Suppose first $p \geq 0$.

- If $b > 0$, by the Cauchy–Schwarz inequality, and using the inequality

$$\alpha\beta \leq \frac{\eta}{2}\,\alpha^2 + \frac{1}{2\eta}\,\beta^2 \quad \text{for all } \alpha,\beta \in \mathbb{R} \text{ and } \eta > 0 \tag{5.5}$$

in the form $||F||\,||\tilde{v}|| \leq \frac{1}{2b}||F||^2 + \frac{b}{2}||\tilde{v}||^2$, we obtain

$$\frac{1}{2}\frac{d}{dt}\|\tilde{v}\|^2 + \nu\|\partial_x\tilde{v}\|^2 + \frac{b}{2}\|\tilde{v}\|^2 \leq \frac{1}{2b}\|F\|^2,$$

which gives, after integration on any time interval $(0,t)$,

$$\frac{1}{2}||\tilde{v}||^2(t) + \nu\int_0^t ||\partial_x\tilde{v}||^2 + \frac{b}{2}\int_0^t ||\tilde{v}||^2 \leq \frac{1}{2b}\int_0^t ||F||^2. \tag{5.6}$$

- If $b = 0$, we use (5.5) with $\eta = 1$ and get through integration on $(0,t)$

$$\frac{1}{2}\|\tilde{v}\|^2(t) + \nu\int_0^t \|\partial_x\tilde{v}\|^2 \leq \frac{1}{2}\|F\|^2_{L^2(0,T;L^2(\Omega_1))} + \frac{1}{2}\int_0^t \|\tilde{v}\|^2.$$

We then apply the Gronwall lemma and obtain

$$\|\tilde{v}\|^2(t) + 2\nu\int_0^t \|\partial_x\tilde{v}\|^2 \leq e^T\|F\|^2_{L^2(0,T;L^2(\Omega_1)}.$$

(b) Suppose now $p < 0$. We move the boundary term in (5.4) to the right-hand side; using the Sobolev inequality in $H^1(\Omega_1)$,

$$\|\tilde{v}\|^2_{L^\infty(\overline{\Omega_1})} \leq 2\|\partial_x\tilde{v}\|\,\|\tilde{v}\|, \tag{5.7}$$

we bound the boundary term, applying again (5.5),

$$-\frac{p}{2}\tilde{v}^2(L) \leq \frac{\nu}{2}\|\partial_x\tilde{v}\|^2 + \frac{p^2}{2\nu}\|\tilde{v}\|^2;$$

and we conclude using the Gronwall lemma as before.

Thus, in both cases, we have a bound for $\tilde{v}$ in $L^\infty(0,T;L^2(\Omega_1)) \cap L^2(0,T;H^1(\Omega_1))$,

$$\|\tilde{v}\|_{L^\infty(0,T;L^2(\Omega_1))}, \|\tilde{v}\|_{L^2(0,T;H^1(\Omega_1))} \leq C(T)\|F\|_{L^2(0,T;L^2(\Omega_1))}. \tag{5.8}$$

2. To obtain the higher regularity result in the theorem, we need to show that $\partial_x^2\tilde{v}$ and $\partial_t\tilde{v}$ are also in $L^2(0,T;L^2(\Omega_1))$. Multiplying the equation by $-\partial_x^2\tilde{v}$ and integrating in space, we get

$$\frac{1}{2}\frac{d}{dt}||\partial_x\tilde{v}||^2+\nu||\partial_x^2\tilde{v}||^2+b||\partial_x\tilde{v}||^2 - \left((\partial_t\tilde{v})\partial_x\tilde{v}+\frac{a}{2}(\partial_x\tilde{v})^2+b(\partial_x\tilde{v})\tilde{v}\right)(L)=-\int_{-\infty}^{L}F\partial_x^2\tilde{v}.$$

Now using the boundary condition to replace $\partial_x\tilde{v}$, we obtain

$$\frac{d}{dt}\left(\frac{1}{2}\|\partial_x\tilde{v}\|^2+\frac{p-a}{4\nu}\tilde{v}^2(L)\right)+\nu\|\partial_x^2\tilde{v}\|^2+b\|\partial_x\tilde{v}\|^2 +\frac{p-a}{2\nu}\left(b-\frac{a}{2}\frac{p-a}{2\nu}\right)\tilde{v}^2(L)=-\int_{-\infty}^{L}F\partial_x^2\tilde{v}.$$

Again using the Cauchy–Schwarz inequality and (5.5) on the right, we find after integrating in time

$$\left(\frac{1}{2}\|\partial_x\tilde{v}\|^2+\frac{p-a}{4\nu}\tilde{v}^2(L)\right)(t)+\frac{\nu}{2}\int_0^t\|\partial_x^2\tilde{v}\|^2+b\int_0^t\|\partial_x\tilde{v}\|^2 +\frac{p-a}{2\nu}\left(b-\frac{a}{2}\frac{p-a}{2\nu}\right)\int_0^t\tilde{v}^2(L)\le\frac{1}{2\nu}\int_0^t||F||^2. \tag{5.9}$$

First the term $\frac{p-a}{2\nu}\left(b-\frac{a}{2}\frac{p-a}{2\nu}\right)\int_0^t\tilde{v}^2(L)$ is handled as in 1, using (5.7) and (5.8). Then, if $p\ge a$, we obtain

$$\|\partial_x\tilde{v}\|^2+\nu\|\partial_x^2\tilde{v}\|_{L^2(0,T;L^2(\Omega_1))}\le C(T)\|F\|_{L^2(0,T;L^2(\Omega_1))}.$$

If $p<a$, then we pass the term containing $\tilde{v}(L)$ to the right-hand side, and using (5.7), we obtain

$$\frac{1}{2}\|\partial_x\tilde{v}\|^2+\frac{\nu}{2}\int_0^t\|\partial_x^2\tilde{v}\|^2\le\frac{a-p}{4\nu}\left(\alpha||\partial_x\tilde{v}||^2+\frac{1}{\alpha}||\tilde{v}||^2\right)+C(T)\int_0^t||F||^2.$$

Now choosing $\alpha=\nu/(a-p)$ and using (5.8), we obtain

$$\|\partial_x\tilde{v}\|_{L^\infty(0,T;L^2(\Omega_1))},\|\partial_x^2\tilde{v}\|_{L^2(0,T;L^2(\Omega_1))}\le C(T)\|F\|_{L^2(0,T;L^2(\Omega_1))},$$

where we omit the dependence of the constant C on a, p, b, and ν.

Now using (5.2), we have

$$\partial_t\tilde{v}=\nu\partial_x^2\tilde{v}-a\partial_x\tilde{v}-b\tilde{v}+F,$$

and since all the terms on the right-hand side are in $L^2(0,T;L^2(\Omega_1))$ by the previous estimates, it follows that $\partial_t\tilde{v}$ is in $L^2(0,T;L^2(\Omega_1))$, which concludes the a priori estimates in $H^{2,1}(\Omega_1\times(0,T))$. Existence and uniqueness can now be shown using a Galerkin method [27]. □

Using Lemma 5.1 and Theorem 5.2, we obtain now the well posedness of the subdomain problems.

THEOREM 5.3. *If f is in $L^2(0,T;L^2(\Omega_1))$, u_0 is in $H^1(\Omega_1)$, and g_L is in $H^{\frac{1}{4}}(0,T)$, then, for any p, problem* (5.1) *has a unique solution v in $H^{2,1}(\Omega_1\times(0,T))$.*

The same result also holds on subdomain Ω_2, the only difference being that $-a$ becomes $+a$ in the estimate (5.9).

THEOREM 5.4. *Let g_L and g_0 be given in $H^{\frac{1}{4}}(0,T)$. If algorithm* (4.1) *with $\mathcal{S}_j$ defined in* (4.6) *is initialized by $(\partial_x u_1^1 + \mathcal{S}_1 u_1^1)(L,\cdot) = g_L$ and $(\partial_x u_2^1 + \mathcal{S}_2 u_2^1)(0,\cdot) = g_0$, then, for any p,* (4.1) and (4.6) *define a sequence of iterates (u_1^k, u_2^k) in $H^{2,1}(\Omega_1 \times (0,T)) \times H^{2,1}(\Omega_2 \times (0,T))$.*

Proof. The proof is done by induction: for $k = 1$, (4.1) defines a unique first iterate (u_1^1, u_2^1) in $H^{2,1}(\Omega_1 \times (0,T)) \times H^{2,1}(\Omega_2 \times (0,T))$ by Theorem 5.3. Assuming now that (u_1^k, u_2^k) is in $H^{2,1}(\Omega_1 \times (0,T)) \times H^{2,1}(\Omega_2 \times (0,T))$, by the trace theorem, we have that $(\partial_x u_2^k + \mathcal{S}_1 u_2^k)(L,\cdot)$ and $(\partial_x u_1^k + \mathcal{S}_2 u_1^k)(0,\cdot)$ are in $H^{\frac{1}{4}}(0,T)$, and thus by Theorem 5.3, (u_1^{k+1}, u_2^{k+1}) must be in $H^{2,1}(\Omega_1 \times (0,T)) \times H^{2,1}(\Omega_2 \times (0,T))$, which concludes the proof. ∎

For the proof of convergence in the overlapping case, we need however more regularity.

THEOREM 5.5. *For $a > 0$ or $a = 0$ and $b > 0$, let $p \geq 0$ and f be in $H^{1,\frac{1}{2}}(\Omega_1 \times (0,T))$, u_0 be in $H^2(\Omega)$, and g_L be in $H^{\frac{3}{4}}(0,T)$, with the compatibility conditions*

$$g_L(0) = \partial_x u_0(L) + \mathcal{S}_1 u_0(L). \tag{5.10}$$

Then the solution v of the subdomain problem (5.1) *is in $H^{3,\frac{3}{2}}(\Omega_1 \times (0,T))$. Furthermore the following compatibility property at $x = 0$ is satisfied:*

$$\lim_{t \to 0+} (\partial_x v + \mathcal{S}_2 v)(0,t) = \partial_x u_0(0) + \mathcal{S}_2 u_0(0). \tag{5.11}$$

Proof. In this more regular situation, the solution u of (2.1) is in $H^{3,\frac{3}{2}}(\Omega \times (0,T))$ by Theorem 2.2. Furthermore $\tilde{g}_L = (\partial_x u + \mathcal{S}_1 u)(L,\cdot)$ is in $H^{\frac{3}{4}}(0,T)$. Subtracting u from v, the difference e is in $H^{3,\frac{3}{2}}(\Omega_1 \times (0,T))$, the solution of (5.1) with data $(0,0,h_L = g_L - \tilde{g}_L)$. By Fourier transform, the same calculation as in Lemma 3.2 gives with $\tilde{h}_L$ being an extension of h_L on $\mathbb{R}$ vanishing in $\mathbb{R}_-$

$$\hat{e} = \frac{2\nu}{\sqrt{d} + p} \mathcal{F}\tilde{h}_L(\omega) e^{r^+(x-L)}. \tag{5.12}$$

The norm of $\partial_x^3 e$ is therefore given by

$$\|\partial_x^3 e\|^2_{L^2(\Omega_1 \times \mathbb{R})} = \int_{\mathbb{R}} \frac{4\nu^2 |r^+|^6}{2\Re r^+ |\sqrt{d} + p|^2} |\mathcal{F}\tilde{h}_L(\omega)|^2 d\omega,$$

and the norm of e in $H^{\frac{3}{2}}(\mathbb{R}, L^2(\Omega_1))$ is

$$\|e\|^2_{H^{3/2}(\mathbb{R}, L^2(\Omega_1))} = \int_{\mathbb{R}} \frac{4\nu^2 (1+\omega^2)^{3/2}}{2\Re r^+ |\sqrt{d} + p|^2} |\mathcal{F}\tilde{h}_L(\omega)|^2 d\omega.$$

In both cases, for $a > 0$ or $a = 0$ and $b > 0$, and $p \geq 0$, the denominator in the factor in front of $|\mathcal{F}\tilde{h}_L(\omega)|^2$ is bounded from below, and it is easy to see that for large $|\omega|$, it is equivalent to a constant times $|\omega|^{3/2}$. Therefore we have the bound

$$\|e\|_{H^{3,\frac{3}{2}}(\Omega_1 \times (0,T))} \leq C \|\tilde{h}_L\|_{H^{\frac{3}{4}}(\mathbb{R})}. \tag{5.13}$$

For the compatibility condition, since $\tilde{h}_L$ vanishes in $\mathbb{R}_-$, its Fourier transform is analytic in the half-plane $\Im\tau < 0$, and by (5.12) and the Paley–Wiener theorem [37],

$e(0,\cdot)$ and $\partial_x e(0,\cdot)$ vanish in $\mathbb{R}_-$. Since they are in $H^{\frac{5}{4}}(\mathbb{R})$ and $H^{\frac{3}{4}}(\mathbb{R})$, respectively, they are continuous, and therefore $\lim_{t\to 0+}(\partial_x e + \mathcal{S}_2 e)(0,t) = 0$, which gives the compatibility property for v. ∎

This regularity result shows the well posedness of the algorithm in $H^{3,\frac{3}{2}}(\Omega_1 \times (0,T))$.

THEOREM 5.6. *For $a > 0$ or $a = 0$ and $b > 0$, and $p \geq 0$, let f be in $H^{1,\frac{1}{2}}(\Omega_1 \times (0,T))$, u_0 be in $H^2(\Omega)$, and g_L and g_0 be in $H^{\frac{3}{4}}(0,T)$, with the compatibility conditions*

$$g_L(0) = \partial_x u_0(L) + \mathcal{S}_1 u_0(L), \qquad g_0(L) = \partial_x u_0(0) + \mathcal{S}_2 u_0(0). \tag{5.14}$$

Then, algorithm (4.1) *with transmission operators* (4.6) *defines a sequence of iterates (u_1^k, u_2^k) in $H^{3,\frac{3}{2}}(\Omega_1 \times (0,T)) \times H^{3,\frac{3}{2}}(\Omega_2 \times (0,T))$.*

5.2. Convergence of the overlapping algorithm. Let h_L and h_0 be given in ${}_0H^{\frac{3}{4}}(0,T)$. Let (e_1, e_2) be the solution in $H^{3,\frac{3}{2}}(\Omega_1 \times (0,T)) \times H^{3,\frac{3}{2}}(\Omega_2 \times (0,T))$ of the problem

$$\begin{aligned} \mathcal{L}e_1 &= 0 \quad \text{in } \Omega_1 \times (0,T), & \mathcal{L}e_2 &= 0 \quad \text{in } \Omega_2 \times (0,T), \\ e_1(\cdot,0) &= 0 \quad \text{in } \Omega_1, & e_2(\cdot,0) &= 0 \quad \text{in } \Omega_2, \\ (\partial_x e_1 + \mathcal{S}_1 e_1)(L,\cdot) &= h_L \text{ in } (0,T), & (\partial_x e_2 + \mathcal{S}_2 e_2)(0,\cdot) &= h_0 \text{ in } (0,T). \end{aligned} \tag{5.15}$$

LEMMA 5.7. *For $a > 0$ or $a = 0$ and $b > 0$, if $p \geq 0$ and $L > 0$, the map $\mathcal{G}_0$ associated with* (5.15),

$$\mathcal{G}_0 : (h_L, h_0) \mapsto ((\partial_x e_2 + \mathcal{S}_1 e_2)(L,\cdot), (\partial_x e_1 + \mathcal{S}_2 e_1)(0,\cdot)), \tag{5.16}$$

is defined from $({}_0H^{\frac{3}{4}}(0,T))^2$ into itself, and $\mathcal{G}_0^2$ is strictly contracting.

Proof. The proof is analogous to the proof of Lemma 3.2 using Fourier analysis. Defining $\tilde{h}_L$ and $\tilde{h}_0$ as any extensions of h_L and h_0 in $H^{\frac{3}{4}}(\mathbb{R})$, vanishing in $\mathbb{R}^-$, we obtain after a short calculation

$$\mathcal{F}(\mathcal{G}_0^2(\tilde{h}_L, \tilde{h}_0))(\omega) = \left(\frac{\sqrt{d} - p}{\sqrt{d} + p}\right)^2 (\mathcal{F}\tilde{h}_0(\omega) e^{(r^- - r^+)L}, \mathcal{F}\tilde{h}_L(\omega) e^{(r^- - r^+)L}),$$

where d and $r^\pm$ are defined in (3.8). Since $p \geq 0$, we have $\left|\frac{\sqrt{d}-p}{\sqrt{d}+p}\right| \leq 1$ and thus

$$\|\mathcal{G}_0^2(h_L, h_0)\|_{({}_0H^{\frac{3}{4}}(0,T))^2} \leq C_D \|(h_L, h_0)\|_{({}_0H^{\frac{3}{4}}(0,T))^2},$$

and since C_D defined in (3.11) satisfies $C_D < 1$, the result follows. ∎

From the proof of this Lemma, we can see that the contraction of the overlapping Schwarz waveform relaxation map with Robin transmission conditions, $\mathcal{G}_0$ given in (5.16), is at least as good as the contraction of the classical map with Dirichlet transmission conditions, $\mathcal{G}_D$ given in (3.6), no matter what one chooses for the parameter $p \geq 0$ in the Robin transmission conditions. Before doing a more thorough comparison, we use the contraction property from Lemma 5.7 to prove convergence of the new algorithm.

THEOREM 5.8. *Let f be in $H^{1,\frac{1}{2}}(\Omega_1 \times (0,T))$ and u_0 be in $H^2(\Omega)$. For $a > 0$ or $a = 0$ and $b > 0$, if $p \geq 0$ and $L > 0$, then the solution (u_1^k, u_2^k) of algorithm* (4.1), (4.6) *converges to the solution u of* (2.1) *for any initial guess $(g_0, g_L) \in (H^{\frac{3}{4}}(0,T))^2$ with the compatibility conditions* (5.14).

Proof. The errors $e_j^k = u_j^k - u$, $j = 1, 2$, satisfy for $k \geq 1$ (4.1) with $f = 0$ and $u_0 = 0$. Introducing the interface functions $h_L^k = (\partial_x e_2^k + \mathcal{S}_1 e_2^k)(L, \cdot)$, $h_0^k = (\partial_x e_1^k + \mathcal{S}_2 e_1^k)(0, \cdot)$ and using the map $\mathcal{G}_0$, we obtain by induction $(h_L^{2k}, h_0^{2k}) = \mathcal{G}_0^{2k}(h_L^0, h_0^0)$, and thus by Lemma 5.7

$$\|(h_L^{2k}, h_0^{2k})\|_{(H^{\frac{3}{4}}(0,T))^2} \leq C_D^k \|(h_L^0, h_0^0)\|_{(H^{\frac{3}{4}}(0,T))^2}.$$

We have by (5.13)

$$\begin{aligned}\|(e_1^{2k+1}, e_2^{2k+1})\|_{H^{3,\frac{3}{2}}(\Omega_1\times(0,T))\times H^{3,\frac{3}{2}}(\Omega_2\times(0,T))} &\leq C\|(h_L^{2k}, h_0^{2k})\|_{(H^{\frac{3}{4}}(0,T))^2} \\ &\leq CC_D^k \|(h_L^0, h_0^0)\|_{(H^{\frac{3}{4}}(0,T))^2},\end{aligned}$$

which together with Lemma 5.7 completes the proof. ∎

Having proved convergence, we now compare the performance of the classical Schwarz waveform relaxation algorithm and the new one with Robin transmission conditions. We do this first at the continuous level, which motivates the optimization procedure we introduce in subsections 5.4 and 5.7 for the discretized case. Using Theorem 5.8 and Lemma 5.7, the error in the overlapping Schwarz waveform relaxation algorithm with Robin transmission conditions satisfies on the interfaces over a double iteration step in Fourier the relation

$$\mathcal{F}(\mathcal{G}_0^2(\tilde{h}_L, \tilde{h}_0))(\omega) = \left(\frac{\sqrt{d} - p}{\sqrt{d} + p}\right)^2 e^{(r^- - r^+)L}(\mathcal{F}\tilde{h}_L(\omega), \mathcal{F}\tilde{h}_0(\omega)),$$

where d, r^-, and r^+ are defined in (3.8). Equivalently, we have

$$\hat{e}_1^{k+1}(L, \omega) = \rho_0 \hat{e}_1^{k-1}(L, \omega), \quad \hat{e}_2^{k+1}(0, \omega) = \rho_0 \hat{e}_2^{k-1}(0, \omega), \tag{5.17}$$

where the convergence factor $\rho_0 = \rho_0(\omega, p, L, \nu, a, b)$ of the new algorithm with Robin transmission conditions is given by

$$\rho_0(\omega, p, L, \nu, a, b) := \left(\frac{\sqrt{a^2 + 4\nu(b + i\omega)} - p}{\sqrt{a^2 + 4\nu(b + i\omega)} + p}\right)^2 e^{-\frac{\sqrt{a^2+4\nu(b+i\omega)}}{\nu} L}. \tag{5.18}$$

For any frequency ω, we can therefore directly compare the performance of the classical Schwarz waveform relaxation algorithm with the new one with Robin transmission conditions: we have $\rho_0 = (\frac{\sqrt{d}-p}{\sqrt{d}+p})^2 \rho_D$, where ρ_D is the classical convergence factor defined in (3.14). This shows that for each ω we have $|\rho_0| < |\rho_D|$ for $p > 0$. Furthermore, for any $\epsilon > 0$ there exists an ω_ϵ such that

$$\int_{|\omega|>\omega_\epsilon} (1 + \omega^2)^{3/4} |\hat{e}_1^0(L, \omega)|^2 d\omega \leq \epsilon$$

because we assume that $e_1^0(L, \cdot)$ is in $H^{\frac{3}{4}}$. Since $|\rho_0| < 1$, we obtain

$$\|\hat{e}_1^{2k}(L, \cdot)\|^2_{H^{\frac{3}{4}}(0,T)} \leq \epsilon + \int_{|\omega|\leq\omega_\epsilon} (1 + \omega^2)^{\frac{3}{4}} |\rho_0(\omega, p, L, \nu, a, b)|^{2k} |\hat{e}_1^0(L, \omega)|^2 d\omega,$$

and taking the supremum of the convergence factor out of the integral, we have

$$\|\hat{e}_1^{2k}(L, \cdot)\|^2_{H^{\frac{3}{4}}(0,T)} \leq \epsilon + \sup_{|\omega|\leq\omega_\epsilon} |\rho_0(\omega, p, L, \nu, a, b)|^{2k} \|\hat{e}_1^0(L, \cdot)\|^2_{H^{\frac{3}{4}}(0,T)}.$$

A similar estimate for the classical algorithm gives

$$\|\hat{e}_1^{2k}(L,\cdot)\|^2_{H^{\frac{3}{4}}(0,T)} \leq \epsilon + \sup_{|\omega|\leq\omega_\epsilon} |\rho_D(\omega,L,\nu,a,b)|^{2k} \|\hat{e}_1^0(L,\cdot)\|^2_{H^{\frac{3}{4}}(0,T)},$$

which shows that improving the convergence factor on a sufficiently large bounded frequency range improves the overall convergence of the algorithm. The choice of a bounded frequency range is further motivated by the fact that computations are performed on a discretized problem, whose grid cannot carry arbitrarily high frequencies. We carefully analyze how to chose the free parameter p for optimal performance of the algorithm in the next subsections.

5.3. Low frequency approximation for the algorithm with overlap. We have seen that the convergence factor of the new algorithm with Robin transmission conditions is given by (5.18), and any choice of the free parameter $p \geq 0$ is admissible to obtain a well posed algorithm. But how should p be chosen, apart from $p \geq 0$? A simple choice is to use a low frequency approximation of the symbols σ_j, $j = 1, 2$, of the optimal transmission operators given in (4.2), based on a Taylor expansion about $\omega = 0$. This is motivated by the fact that with overlap, $L > 0$, the exponential term in the convergence factor (5.18) is exponentially small for ω large, and hence p should be used to make the transmission conditions effective for ω small. Using a Taylor expansion of the square root $\sqrt{a^2 + 4\nu(b + i\omega)}$ in (4.2) about $\omega = 0$, we find

$$\sqrt{a^2 + 4\nu(b + i\omega)} = \sqrt{a^2 + 4\nu b} + \frac{2\nu}{\sqrt{a^2 + 4\nu b}} i\omega + O(\omega^2), \tag{5.19}$$

and hence the low frequency approximation choice for p in the Robin transmission condition is

$$p = p_T := \sqrt{a^2 + 4\nu b}. \tag{5.20}$$

With this choice, the convergence factor vanishes for $\omega = 0$ and also when ω goes to infinity, since $L > 0$. To further analyze the convergence factor, we introduce a change of variables based on the real part of the square root in the convergence factor (5.18),

$$x := \Re(\sqrt{a^2 + 4\nu(b + i\omega)}). \tag{5.21}$$

In this new variable, the convergence factor (5.18) in modulus becomes

$$R_0(x, p, x_0, L) := |\rho_0| = \frac{(x-p)^2 + x^2 - x_0^2}{(x+p)^2 + x^2 - x_0^2} e^{-\frac{Lx}{\nu}}, \tag{5.22}$$

where $x_0^2 := a^2 + 4\nu b$. Note that $R_0 \geq 0$ by definition, which can also be seen from $x^2 \geq a^2 + 4\nu b = x_0^2$ from the change of variables (5.21). Using now the parameter p_T from the Taylor expansion, we find for the Taylor–Robin method (T0 for Taylor of order 0) the convergence factor in modulus to be

$$R_{T0}(x, x_0, L) := R_0(x, p_T, x_0, L) = \frac{x - x_0}{x + x_0} e^{-\frac{Lx}{\nu}} \geq 0, \qquad x \geq x_0. \tag{5.23}$$

THEOREM 5.9 (T0 performance with overlap). *Let $L > 0$ and $x_0 := \sqrt{a^2 + 4\nu b}$. The convergence factor R_{T0} in* (5.23) *of the overlapping Schwarz waveform relaxation*

algorithm with Robin transmission conditions (4.1),(4.6) *and* $p = p_T$ *from the Taylor low frequency approximation* (5.20) *is for* $x_0 \le x < \infty$ *uniformly bounded by*

(5.24)
$$R_{T0}(x, x_0, L) \le \bar{R}_{T0}(x_0, L) := R_{T0}(\bar{x}, x_0, L) = \frac{\bar{x} - x_0}{\bar{x} + x_0} e^{-\frac{L\bar{x}}{\nu}}, \quad \bar{x} = \sqrt{x_0^2 + \frac{2\nu x_0}{L}}.$$

For L *small, we have* $\bar{R}_{T0}(x_0, L) = 1 - 2\sqrt{\frac{2x_0}{\nu}}\sqrt{L} + O(L)$.

Proof. Taking a derivative of R_{T0} with respect to x shows that there is a unique maximum of R_{T0} for $x_0 \le x < \infty$ at $\bar{x}$, which leads to the bound given in (5.24). An expansion for L small leads then to the asymptotic result of the theorem. ∎

Now in a numerical calculation, two additional issues come into play: First the frequency parameter ω cannot be arbitrarily high, there is a maximum frequency that can be represented on a grid with spacing Δt, and an estimate for this maximum frequency is $\omega_{\max} = \frac{\pi}{\Delta t}$, the signal that oscillates between ± 1 from grid point to grid point. Second, the overlap L is in general not a fixed quantity, one can afford only a few grid cells overlap, and often $L = \Delta x$. The question therefore arises, if for a particular discretization, which might have to satisfy a stability constraint, the bound on the contraction factor in (5.24) is really relevant, or if the highest frequencies represented on the grid of the particular discretization stay below $\bar{\omega}$ where the maximum of R_{T0} is attained, which corresponds to $\bar{x}$ given in (5.24) in the transformed problem. To answer this question, we need to study for which cases the maximum numerical frequency $\omega_{\max}$ stays below $\bar{\omega}$ or, in the transformed problem, under which conditions

$$x_{\max} = \sqrt{\frac{\sqrt{x_0^4 + 16\nu^2\omega_{\max}^2} + x_0^2}{2}} \tag{5.25}$$

stays below $\bar{x}$ given in (5.24). A direct comparison shows that for

$$L > L_0 := \frac{4\nu x_0}{\sqrt{x_0^4 + 16\nu^2\omega_{\max}^2} - x_0^2} \tag{5.26}$$

the maximum numerical frequency $\omega_{\max} > \bar{\omega}$, and hence the bound given in Theorem 5.9 determines the convergence rate of the algorithm. If however $L \le L_0$, then the maximum on the numerically relevant convergence factor is attained at $\omega_{\max}$. Numerically, we therefore have

(5.27)
$$R_{T0}(x, x_0, L) \le \tilde{R}_{T0}(x_0, L) := \begin{cases} R_{T0}(\bar{x}, x_0, L) = \dfrac{\bar{x} - x_0}{\bar{x} + x_0} e^{-\frac{L\bar{x}}{\nu}} & \text{if } L > L_0, \\ R_{T0}(x_{\max}, x_0, L) = \dfrac{x_{\max} - x_0}{x_{\max} + x_0} e^{-\frac{Lx_{\max}}{\nu}} & \text{if } L \le L_0. \end{cases}$$

To obtain a concrete asymptotic result for the case where the overlap L is linked to the space discretization Δx, $L = C_1\Delta x$, and the space discretization Δx is linked to the time discretization Δt by a stability or accuracy constraint, $\Delta t = C_2\Delta x^\beta$, $\beta > 0$, we insert these relations into L_0 and expand to find

$$L_0 = \frac{x_0}{\pi}\Delta t + O(\Delta t^2), \tag{5.28}$$

which leads to the following asymptotic results.

THEOREM 5.10 (T0 discrete convergence estimate with overlap). *Let* $x_0 := \sqrt{a^2 + 4\nu b}$. *If* $L = C_1\Delta x$ *and* $\Delta t = C_2\Delta x^\beta$, *then the bound* $\tilde{R}_{T0}$ *in* (5.27) *on the convergence factor estimate of the discretized overlapping Schwarz waveform relaxation algorithm with Robin transmission conditions* (4.1),(4.6) *and* $p = p_T$ *from the Taylor low frequency approximation* (5.20) *is for* Δx *small given by*

(5.29)
$$\tilde{R}_{T0} = \begin{cases} 1 - 2\sqrt{\dfrac{2C_1x_0}{\nu}}\sqrt{\Delta x} + O(\Delta x) & \text{if } \beta > 1 \text{ or } \beta = 1 \text{ and } \frac{C_1}{C_2} > \frac{x_0}{\pi}, \\ 1 - \dfrac{\sqrt{2}(C_2x_0 + C_1\pi)}{\sqrt{C_2\pi\nu}}\sqrt{\Delta x} + O(\Delta x) & \text{if } \beta = 1 \text{ and } \frac{C_1}{C_2} \le \frac{x_0}{\pi}, \\ 1 - x_0\sqrt{\dfrac{2C_2}{\pi\nu}}\Delta x^{\frac{\beta}{2}} + o(\Delta x^{\frac{\beta}{2}}) & \text{if } \beta < 1. \end{cases}$$

Proof. Comparing $L = C_1\Delta x$ with the expansion of L_0 given in (5.28) and using that $\Delta t = C_2\Delta x^\beta$, we see that for Δx small the first case in (5.27) corresponds to the first case given in (5.29). The asymptotic bound on the convergence factor then follows by simply expanding for $L = C_1\Delta x$ and Δx small. For the second case, one can set $\beta = 1$ and directly expand the second case of (5.27) to find the result given. For the last case, we first notice that $x_{\max}$ satisfies

$$x_{\max} = \sqrt{2\pi\nu}\frac{1}{\sqrt{\Delta t}} + O(\sqrt{\Delta t}) = \sqrt{\frac{2\pi\nu}{C_2}}\Delta x^{-\frac{\beta}{2}} + O(\Delta x^{\frac{\beta}{2}}), \tag{5.30}$$

which together with $L = C_1\Delta x$ gives for the exponential the expansion

$$e^{-\frac{Lx_{\max}}{\nu}} = 1 - C_1\sqrt{\frac{2\pi}{C_2\nu}}\Delta x^{1-\frac{\beta}{2}} + O(\Delta x^{2-\beta}). \tag{5.31}$$

Multiplying this with the expansion for the coefficient in front of the exponential in (5.27), whose expansion is

$$\frac{x_{\max} - x_0}{x_{\max} + x_0} = 1 - x_0\sqrt{\frac{2C_2}{\pi\nu}}\Delta x^{\frac{\beta}{2}} + O(\Delta x^\beta), \tag{5.32}$$

the result follows. ∎

The preceding theorem shows that for explicit discretizations, which have a stability constraint of the type $\Delta t = C_2\Delta x^2$ for this problem and for which the present algorithm would still be of interest for nonmatching time grids, the optimized Schwarz waveform relaxation algorithm with Robin transmission conditions based on a low frequency approximation and an overlap of the order of the spatial discretization Δx will have an asymptotic convergence factor $1 - O(\sqrt{\Delta x})$, as one could expect from the continuous analysis in Theorem 5.9. This is still true for implicit discretizations, as long as Δt is of the same order as Δx. Once Δt becomes much larger than Δx, however, one can expect the algorithm to converge faster asymptotically because of the last relation in (5.29).

5.4. Optimization of the algorithm with overlap. We investigate now if there exists a better choice for p such that the overall convergence factor is smaller than with the parameter from the low frequency approximation. We will use the label O0 for these methods, which stands for optimized of order 0. We place ourselves first

again in the continuous context, where $\omega \in \mathbb{R}$, and thus $\omega_{\max} = \infty$. Later, we will also investigate the discretized case where $\omega_{\max} < \infty$. We introduce a change of variables, which will greatly simplify the analysis of the optimal parameter p. Setting $x := \frac{y\nu}{L}$, $p := \frac{\tilde{p}\nu}{L}$, and $x_0 := \frac{y_0\nu}{L}$ in the convergence factor (5.22), we obtain

$$R_0(y, \tilde{p}, y_0) = \frac{(y-\tilde{p})^2 + y^2 - y_0^2}{(y+\tilde{p})^2 + y^2 - y_0^2} e^{-y}, \tag{5.33}$$

which is now an expression independent of the overlap parameter L and the viscosity parameter ν. The best choice for the parameter $\tilde{p}$ is the one that makes R_0 as small as possible uniformly for all $y \geq y_0$ and is hence the solution of the min-max problem

$$\min_{\tilde{p}} \left(\max_{y \geq y_0} R_0(y, \tilde{p}, y_0) \right) = \min_{\tilde{p} \geq 0} \left(\max_{y \geq y_0} \frac{(y-\tilde{p})^2 + y^2 - y_0^2}{(y+\tilde{p})^2 + y^2 - y_0^2} e^{-y} \right), \tag{5.34}$$

where minimizing over nonnegative $\tilde{p}$ is equivalent to minimizing over all $\tilde{p}$, as one can see from (5.33). Note that $\tilde{p}$ nonnegative is also a requirement for the convergence proof of the algorithm in Theorem 5.8. To analyze the min-max problem (5.34), we need two lemmas.

LEMMA 5.11. *The function $y \mapsto R_0(y, \tilde{p}, y_0)$ defined in* (5.33) *has a unique local maximum at*

$$\bar{y}(y_0, \tilde{p}) = \sqrt{\frac{y_0^2 + 2\tilde{p} + \sqrt{d(y_0, \tilde{p})}}{2}}, \quad d(y_0, \tilde{p}) = \tilde{p}(-\tilde{p}^3 - 4\tilde{p}^2 + (4 + 2y_0^2)\tilde{p} + 8y_0^2), \tag{5.35}$$

if $0 \leq \tilde{p} < \tilde{p}_1(y_0)$, where $\tilde{p}_1(y_0)$ is the unique positive root of $d(y_0, \tilde{p})$ for $y_0 > 0$. If $\tilde{p} \geq \tilde{p}_1(y_0)$, then $R_0(y, \tilde{p}, y_0)$ is a monotonically decreasing function of y.

Proof. A partial derivative of $R_0(y, \tilde{p}, y_0)$ with respect to y gives

$$\frac{\partial R_0}{\partial y} = -\frac{e^{-y}(4y^4 - 4(2\tilde{p} + y_0^2)y^2 + (\tilde{p}^2 - y_0^2)(y_0^2 - 4\tilde{p} - \tilde{p}^2))}{((\tilde{p}+y)^2 + y^2 - y_0^2)^2},$$

and therefore $R_0(y, \tilde{p}, y_0)$ can have at most two extrema, $\bar{y} = \sqrt{(y_0^2 + 2\tilde{p} + \sqrt{d(y_0, \tilde{p})})/2}$ and $\underline{y} = \sqrt{(y_0^2 + 2\tilde{p} - \sqrt{d(y_0, \tilde{p})})/2}$, with the discriminant $d(y_0, \tilde{p})$ given in (5.35). The larger of the two, $\bar{y}$, must be a maximum, since $R_0 \geq 0$ and R_0 goes to 0 as y goes to ∞. Since the discriminant is positive for small positive $\tilde{p}$ and is negative for large positive $\tilde{p}$, it must have by continuity at least one real positive root $\tilde{p}_1(y_0) > 0$, $d(y_0, \tilde{p}_1) = 0$. To prove that this root is unique, we use the derivative of $d(y_0, \tilde{p})/\tilde{p}$ with respect to $\tilde{p}$, which shows that there are two extrema, one at $r_1 = -\frac{1}{3}(4 - \sqrt{28 + 6y_0^2})$ and one at $r_2 = -\frac{1}{3}(4 + \sqrt{28 + 6y_0^2})$. The larger one, r_1, must be a maximum, since the discriminant goes to $-\infty$ as $\tilde{p}$ goes to ∞, and thus r_2 is a minimum. Since r_2 is negative, $\tilde{p}_1$ is the only positive root of the discriminant, since this latter is still positive for arbitrary small $\tilde{p}$. For $\tilde{p} \geq \tilde{p}_1$, R_0 has no extrema in y and hence decreases monotonically to 0 as y goes to infinity. ∎

LEMMA 5.12. *For fixed $y > y_0$ and $\tilde{p} > 0$, we have $\frac{\partial R_0(y, \tilde{p}, y_0)}{\partial \tilde{p}}(\tilde{p} - \tilde{p}_2(y)) \geq 0$, where $\tilde{p}_2(y) := \sqrt{2y^2 - y_0^2}$.*

Proof. A partial derivative of $R_0(y, \tilde{p}, y_0)$ with respect to $\tilde{p}$ gives

$$\frac{\partial R_0}{\partial \tilde{p}} = -\frac{4e^{-y}y(-\tilde{p}^2 + 2y^2 - y_0^2)}{((\tilde{p}+y)^2 + y^2 - y_0^2)^2},$$

which has only one root in $\tilde{p}$, $\tilde{p}_2(y) = \sqrt{2y^2 - y_0^2}$, which is positive. For $\tilde{p} < \tilde{p}_2$, $\frac{\partial R_0}{\partial \tilde{p}}$ is negative and hence $R_0(y, \tilde{p}, y_0)$ decreases when $\tilde{p}$ increases, whereas for $\tilde{p} > \tilde{p}_2$, $R_0(y, \tilde{p}, y_0)$ increases when $\tilde{p}$ increases. ∎

THEOREM 5.13 (O0 performance with overlap). *Let $L > 0$ and $x_0 := \sqrt{a^2 + 4\nu b}$. The best performance of the optimized overlapping Schwarz waveform relaxation algorithm at the continuous level with Robin transmission conditions* (4.1),(4.6) *is obtained for $p = p^* := \frac{\tilde{p}^* \nu}{L}$, where $\tilde{p}^*$, the solution of the min-max problem* (5.34)*, is for $y_0 := \frac{x_0 L}{\nu} < y_c$ given by the unique solution $\tilde{p}^* \geq y_0$ of the nonlinear equation*

$$R_0(y_0, \tilde{p}^*, y_0) = R_0(\bar{y}(y_0, \tilde{p}^*), \tilde{p}^*, y_0), \tag{5.36}$$

where $R_0(y, \tilde{p}, y_0)$ is given in (5.33) *and $\bar{y}(y_0, \tilde{p})$ is given in* (5.35)*. For $y_0 \geq y_c$, $\tilde{p}^*$ is given by the unique solution of*

$$y_0 = \tilde{p}^* \sqrt{\frac{\tilde{p}^*}{(4 + \tilde{p}^*)}}. \tag{5.37}$$

The constant y_c is universal, $y_c = 1.618386576\ldots$, and the convergence factor with the optimal p^* is uniformly bounded by

$$R_0(y, \tilde{p}^*, y_0) \leq \bar{R}_{O0}(y_0, \tilde{p}^*) := R_0(\bar{y}(y_0, \tilde{p}^*), \tilde{p}^*, y_0). \tag{5.38}$$

For L small, we have the asymptotic result

$$p^* = \frac{\tilde{p}^* \nu}{L} \approx (2x_0^2 \nu)^{\frac{1}{3}} L^{-\frac{1}{3}}, \quad \bar{R}_{O0} \approx 1 - \left(\frac{2^5 x_0}{\nu}\right)^{\frac{1}{3}} L^{\frac{1}{3}}. \tag{5.39}$$

Proof. By Lemma 5.12, the optimal $\tilde{p}^* \geq y_0$ since for $\tilde{p} < \tilde{p}_2(y_0) = y_0$, increasing $\tilde{p}$ decreases $R_0(y, \tilde{p}, y_0)$ for all $y > y_0$. Now Lemma 5.11 implies that for $y_0 \leq \tilde{p} \leq \tilde{p}_1(y_0)$, the maximum of R_0 in the min-max problem can be attained at $y = y_0$ or at the interior maximum at $\bar{y}$ given in (5.35). For $\tilde{p} = y_0$, we have $R_0(y_0, \tilde{p}, y_0) = R_0(y_0, y_0, y_0) = 0$ and $d(y_0, \tilde{p}) = d(y_0, y_0) = y_0^2(2 + y_0)^2 \geq 0$, and hence $R_0(y, \tilde{p}, y_0)$ has for $y \geq y_0$ a unique maximum at $\bar{y}(y_0, y_0) = \sqrt{y_0(2 + y_0)} > y_0$. Increasing $\tilde{p}$ from y_0 increases $R_0(y_0, \tilde{p}, y_0)$ by Lemma 5.12 monotonically for all $\tilde{p} > \tilde{p}_2(y_0) = y_0$. Increasing $\tilde{p}$ from y_0 also decreases $R_0(\bar{y}(y_0, \tilde{p}), \tilde{p}, y_0)$ by Lemma 5.12, as long as it exists, $\tilde{p} < \tilde{p}_1(y_0)$ according to Lemma 5.11, and $\tilde{p} < \tilde{p}_2(\bar{y}(y_0, \tilde{p})) = \sqrt{2\bar{y}^2 - y_0^2}$, after which $R_0(\bar{y}, \tilde{p}, y_0)$ will increase again according to Lemma 5.12. By continuity, the maximum of R_0 is minimized either for $\tilde{p}_1^*$ satisfying

$$R_0(y_0, \tilde{p}_1^*, y_0) = R_0(\tilde{y}, \tilde{p}_1^*, y_0), \tag{5.40}$$

provided that $\tilde{p}_1^* \leq \tilde{p}_2(\bar{y}(y_0, \tilde{p}_1^*)) = \sqrt{2(\bar{y}(y_0, \tilde{p}_1^*))^2 - y_0^2}$, or for $\tilde{p}_2^*$ given by

$$\tilde{p}_2^* = \tilde{p}_2(\bar{y}(y_0, \tilde{p}_2^*)) = \sqrt{2(\bar{y}(y_0, \tilde{p}_2^*))^2 - y_0^2}. \tag{5.41}$$

It depends on the only parameter left, y_0, which of these two cases is the solution. Imposing $\tilde{p}_1^* = \tilde{p}_2^*$ and both (5.40) and (5.41), we can solve for the value of y_0 where both are equally optimal. We find

$$y_0 = y_c = 1.618386576\ldots, \quad \tilde{p}_1^* = \tilde{p}_2^* = 2.583490822\ldots.$$

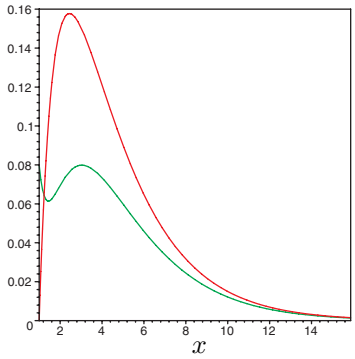


FIG. 5.1. *The top curve is the convergence factor R_{T0} from the Taylor low frequency approximation, and the curve below is the optimized convergence factor R_{O0}, for an example from the numerical section.*

TABLE 5.1
Comparison of the optimal $\tilde{p}^*$ from Theorem 5.13 *and its asymptotic approximation.*

y_0	0.1	0.01	0.001	0.0001	0.00001
$\tilde{p}^*$	0.2936	0.05952	0.01265	0.002717	0.0005849
Asymptotic $\tilde{p}^*$	0.2714	0.05848	0.01260	0.002714	0.0005848

Hence for $y_0 < y_c$ and for $y_0 \geq y_c$ (5.40) and (5.41), respectively, give the solution. (5.41) can be simplified by solving it for y_0, which gives (5.37) stated in the theorem, and a derivative with respect to $\tilde{p}^*$ shows that there is a unique positive root $\tilde{p}^*$ for $y_0 > 0$.

The uniform bound given in (5.38) is a direct consequence of (5.36) and (5.37), since in both cases the maximum is attained at $\bar{y}$.

To show the asymptotic result (5.39), we note that for L small and the other problem parameters a, b, and ν fixed, we have y_0 small, since from the variable transform we have $y_0 = \frac{x_0}{\nu} L = \frac{\sqrt{a^2+4\nu b}}{\nu} L$ and therefore the first result (5.36) applies asymptotically, $y_0 < y_c$. To solve (5.36) asymptotically, we insert the ansatz $\tilde{p}^* = C_p y_0^\alpha$ into (5.36) and expand both sides for y_0 small. Using that $\tilde{p}^* \geq y_0$, we find from its definition that asymptotically $\bar{y} \approx \sqrt{2C_p} y_0^{\frac{\alpha}{2}}$. Using this in equation (5.36), we find for the leading order terms

$$1 - 2^{\frac{5}{3}} y_0^{1-\alpha} - y_0 + 2^{\frac{5}{3}} y_0^{1-\alpha+1} y_0 + \cdots = 1 - 2^{\frac{3}{2}} \sqrt{C_p} y_0^{\frac{\alpha}{2}} + 4 C_p y_0^\alpha + \cdots ,$$

which implies $1 - \alpha = \frac{\alpha}{2}$ and thus $\alpha = \frac{2}{3}$ and $C_p = 2^{\frac{1}{3}}$, which leads to (5.39). ∎

In Figure 5.1 we show the convergence factors R_{T0} and R_{O0} for an example with $x_0 = 1$, $L = 0.08$, and $\nu = 0.2$ from the numerical section. One can see the better performance of the optimized Robin transmission conditions over the Taylor transmission conditions and also the equioscillation of the optimal choice.

Table 5.1 gives a comparison of the optimal $\tilde{p}^*$ from (5.36) with the asymptotic approximation (5.39). One can see that the asymptotic approximation is very close to the optimal $\tilde{p}^*$ already for moderately small values of y_0, which corresponds to a small overlap L. For larger values of y_0, the asymptotic approximation can be a valuable initial guess for the nonlinear equation solver to find the optimal $\tilde{p}^*$ from (5.36).

In Figure 3.1 on the right, we show the first few iterations, at the end of the time interval, of the optimized Schwarz waveform relaxation algorithm with Robin

transmission conditions for the same model problem for which the iterates of the classical Schwarz waveform relaxation algorithm are shown on the left. One can clearly see that the new algorithm with Robin transmission conditions converges much faster than the algorithm with Dirichlet transmission conditions.

Theorem 5.13 gives the optimal choice for the parameter in the Robin transmission conditions of the optimized Schwarz waveform relaxation algorithm at the continuous level. In a numerical setting, however, not all the frequencies are present, as we have seen, and we have to address the question again if the maximum of the convergence factor attained at $\bar{y}$ is relevant in a computation. Letting $L = C_1\Delta x$ and $\Delta t = C_2\Delta x^\beta$ as before, the maximum numerical frequency we can expect on the time discretization grid is $\omega_{\max} = \frac{\pi}{\Delta t} = \frac{\pi}{C_2\Delta x^\beta}$, which corresponds with the variable transform to

(5.42)
$$y_{\max} = \frac{Lx_{\max}}{\nu} = C_1\Delta x\sqrt{\frac{\sqrt{x_0^4 + \left(\frac{4\nu\pi}{C_2\Delta x^\beta}\right)^2} + 2x_0^2}{2}} = C_1\sqrt{\frac{2\pi}{\nu C_2}}\Delta x^{1-\frac{\beta}{2}} + O(\Delta x^{1+\frac{\beta}{2}}),$$

whereas $\bar{y}$ from the optimization in (5.36) has the expansion

$$\bar{y} = 2^{\frac{2}{3}}\left(\frac{x_0C_1}{\nu}\right)^{\frac{1}{3}}\Delta x^{\frac{1}{3}} + O(\Delta x). \tag{5.43}$$

Hence, if $1 - \frac{\beta}{2} = \frac{1}{3}$ or $\beta = \frac{4}{3}$ and $C_1 = \sqrt{x_0}\left(\frac{2\nu C_2^3}{\pi^3}\right)^{\frac{1}{4}} =: C_c$, the numerical $y_{\max}$ and $\bar{y}$ from the optimization are asymptotically at the same location, which represents the boundary between the usefulness of the continuous optimization result (5.36) and a different optimization for the discretized algorithm, which we show in the following theorem.

THEOREM 5.14 (O0 discrete convergence estimate with overlap). *Let* $x_0 := \sqrt{a^2 + 4\nu b}$. *If* $L = C_1\Delta x$ *and* $\Delta t = C_2\Delta x^\beta$, *then the convergence factor* $R_0(y, \tilde{p}, y_0)$ *of the discretized overlapping Schwarz waveform relaxation algorithm with Robin transmission conditions* (4.1),(4.6) *is for* Δx *small bounded for all* $y \in [y_0, y_{\max}]$ *by* $\tilde{R}_{O0}$, *where* $y_{\max}$ *is given in* (5.42), $\tilde{R}_{O0}$ *and* $\tilde{p}^*$ *satisfy*

(5.44)
$$\tilde{R}_{O0} \approx 1 - \left(\frac{2^5C_1x_0}{\nu}\right)^{\frac{1}{3}}\Delta x^{\frac{1}{3}},\ p^* \approx \left(\frac{2x_0^2\nu}{C_1}\right)^{\frac{1}{3}}\Delta x^{-\frac{1}{3}} \quad \text{if } \beta > \tfrac{4}{3} \text{ or } \beta = \tfrac{4}{3} \text{ and } C_1 > C_c,$$
$$\tilde{R}_{O0} \approx 1 - \tfrac{8C_1x_0}{C_p\nu}\Delta x^{\frac{1}{3}}, \qquad p^* \approx C_p\Delta x^{-\frac{1}{3}} \qquad \text{if } \beta = \tfrac{4}{3} \text{ and } C_1 \le C_c,$$
$$\tilde{R}_{O0} \approx 1 - 2\left(\frac{2C_2x_0^2}{\nu\pi}\right)^{\frac{1}{4}}\Delta x^{\frac{\beta}{4}},\ p^* \approx \left(\frac{2^3x_0^2\nu\pi}{C_2}\right)^{\frac{1}{4}}\Delta x^{-\frac{\beta}{4}} \text{ if } \beta < \tfrac{4}{3},$$

and the constants are given by $C_p = \frac{1}{2C_2}(\sqrt{\pi^2C_1^2 + 8\sqrt{2\nu\pi}C_2^{\frac{3}{2}}x_0} - \pi C_1)$, $C_c = \sqrt{x_0}(\frac{2\nu C_2^3}{\pi^3})^{\frac{1}{4}}$.

Proof. The first case is a direct consequence of Theorem 5.13, which applies in this case, since the maximum $\bar{y}$ is relevant for the numerical discretization if $\beta > \frac{4}{3}$ or $\beta = \frac{4}{3}$ and $C_1 > C_c$, as can be see from (5.42) and (5.43). For case two and three, the local maximum at $\bar{y}$ lies outside of the numerical frequencies, $\bar{y} > y_{\max}$, and hence the min-max problem needs to be adapted to this situation; the maximum needs now

be minimized only for $y \in [y_0, y_{\max}]$. For a small overlap, which corresponds to y_0 small, the solution is achieved according to Theorem 5.13 when

$$R_0(y_0, \tilde{p}^*, y_0) = R_0(y_{\max}, \tilde{p}^*, y_0). \tag{5.45}$$

Expanding both sides asymptotically for small Δx, we find, using the ansatz $\tilde{p}^* = \tilde{C}_p \Delta x^\alpha$, that the leading order terms of (5.45) are

$$1 - \frac{4x_0 C_1}{\tilde{C}_p \nu} \Delta x^{1-\alpha} + \cdots = 1 - \frac{2\tilde{C}_p}{C_1} \sqrt{\frac{\nu C_2}{2\pi}} \Delta x^{\alpha - 1 + \frac{\beta}{2}} - C_1 \sqrt{\frac{2\pi}{\nu C_2}} \Delta x^{1-\frac{\beta}{2}} + \cdots .$$

Hence in the limiting case, where $\beta = \frac{4}{3}$, we have $\alpha = \frac{2}{3}$, and both terms on the right have the same exponent. This leads to the constant C_p given in the theorem in case two, after having used the back transform $p^* = \frac{\tilde{p}^* \nu}{C_1 \Delta x}$. If however $\beta < \frac{4}{3}$, then the last term on the right-hand side is of lower order. Balancing the remaining two, we find for the exponents $1 - \alpha = \alpha - 1 + \frac{\beta}{2}$ or $\alpha = 1 - \frac{\beta}{4}$ and the constant $\tilde{C}_p = C_1 (\frac{2^3 x_0^2 \pi}{\nu^3 C_2})^{\frac{1}{4}}$, which leads after the back transform to the last case stated in the theorem. ∎

5.5. Convergence for the nonoverlapping algorithm. We now assume that the overlap is zero, $L = 0$. We first analyze convergence of the algorithm in the appropriate Sobolev spaces. The convergence analysis for the nonoverlapping case is based on energy estimates and follows an idea from [28], which has also been used in [6] and [33] for Schwarz algorithms applied to steady problems and in [11] for a nonoverlapping Schwarz waveform relaxation algorithm for hyperbolic evolution equations.

THEOREM 5.15. *Without overlap, $L = 0$, the Schwarz waveform relaxation algorithm* (4.1),(4.6) *converges for $p > 0$ in $(L^\infty(0,T;L^2(\Omega_1)) \cap L^2(0,T;H^1(\Omega_1))) \times (L^\infty(0,T;L^2(\Omega_2)) \cap L^2(0,T;H^1(\Omega_2)))$ to the solution u of* (2.1) *for any initial guess $g_0 \in H^{\frac{1}{4}}(0,T)$ and $g_L \in H^{\frac{1}{4}}(0,T)$.*

Proof. As in the proof of Theorem 5.2 we obtain the energy estimates

$$\frac{1}{2}\frac{d}{dt}\|e_1^k\|^2 + \nu\|\partial_x e_1^k\|^2 + b\|e_1^k\|^2 - \left(\nu\partial_x e_1^k - \frac{a}{2}e_1^k\right)(0)e_1^k(0) = 0, \tag{5.46}$$

$$\frac{1}{2}\frac{d}{dt}\|e_2^k\|^2 + \nu\|\partial_x e_2^k\|^2 + b\|e_2^k\|^2 + \left(\nu\partial_x e_2^k - \frac{a}{2}e_2^k\right)(0)e_2^k(0) = 0. \tag{5.47}$$

Introducing the boundary operators $\mathcal{B}^+ = \partial_x + \mathcal{S}_1$, $\mathcal{B}^- = \partial_x + \mathcal{S}_2$ and rewriting the terms on the interface in the form $(\nu\partial_x e - \frac{a}{2}e)e = \frac{\nu^2}{2p}((\mathcal{B}^+ e)^2 - (\mathcal{B}^- e)^2)$, we obtain the new energy estimates

$$\frac{1}{2}\frac{d}{dt}\|e_1^k\|^2 + \nu\|\partial_x e_1^k\|^2 + b\|e_1^k\|^2 + \frac{\nu^2}{2p}(\mathcal{B}^- e_1^k)^2(0) = \frac{\nu^2}{2p}(\mathcal{B}^+ e_1^k)^2(0), \tag{5.48}$$

$$\frac{1}{2}\frac{d}{dt}\|e_2^k\|^2 + \nu\|\partial_x e_2^k\|^2 + b\|e_2^k\|^2 + \frac{\nu^2}{2p}(\mathcal{B}^+ e_2^k)^2(0) = \frac{\nu^2}{2p}(\mathcal{B}^- e_2^k)^2(0). \tag{5.49}$$

Now note that the transmission conditions can be expressed with the operators $\mathcal{B}^\pm$,

$$\mathcal{B}^+ e_1^k = \mathcal{B}^+ e_2^{k-1}, \quad \mathcal{B}^- e_2^k = \mathcal{B}^- e_1^{k-1} \text{ on } \{0\} \times (0,T).$$

Replacing the corresponding terms in the two equations (5.48) and (5.49), adding the resulting equations, and summing in k, we get a telescopic sum on the interfaces and

therefore

$$\begin{aligned}\sum_{k=1}^{K}\left[\tfrac{1}{2}\tfrac{d}{dt}(\|e_1^k\|^2+\|e_2^k\|^2)+\nu(\|\partial_x e_1^k\|^2+\|\partial_x e_2^k\|^2)+b(\|e_1^k\|^2+\|e_2^k\|^2)\right]+\\ +\tfrac{\nu^2}{2p}((\mathcal{B}^- e_1^K)^2+(\mathcal{B}^+ e_2^K)^2)(0)=\tfrac{\nu^2}{2p}((\mathcal{B}^- e_1^1)^2+(\mathcal{B}^+ e_2^1)^2)(0).\end{aligned} \tag{5.50}$$

We can now integrate in time, and since the initial values of the error vanish, the sum of the energies over all the iterates remains bounded. Hence the energy in the iterates needs to go to zero and the algorithm converges. ∎

5.6. Low frequency approximation for the algorithm without overlap. One can choose the free parameter p in the Robin transmission conditions based on a low frequency Taylor approximation, as given in (5.20). But now there is no overlap to be effective on the high frequencies, the convergence factor (5.23) becomes

$$R_{T0}(x,x_0)=\frac{x-x_0}{x+x_0}, \tag{5.51}$$

where $x \geq x_0$ is given by the variable transform (5.21). Clearly R_{T0} is a monotonically increasing function of x for $x \geq x_0$ and tends to one as x tends to infinity. There is therefore no uniform bound on R_{T0} which is strictly less than one for all $x \geq x_0$ in the case without overlap. But we have already seen that in a numerical calculation the frequency parameter ω cannot be arbitrarily high. It suffices therefore for the numerical case to find a bound for R_{T0} for $x_0 \leq x \leq x_{\max}$, where $x_{\max}$ is given in (5.25).

THEOREM 5.16 (T0 discrete convergence estimate without overlap). *Let* $x_0 := \sqrt{a^2+4\nu b}$ *and* $L=0$. *Then the convergence factor estimate* R_{T0} *of the discretized nonoverlapping Schwarz waveform relaxation algorithm with Robin transmission conditions* (4.1),(4.6) *and* $p=p_T$ *from the Taylor low frequency approximation* (5.20) *is for* $x_0 \leq x \leq x_{\max}$, *where* $x_{\max}$ *is defined in* (5.25), *bounded by*

$$R_{T0}(x,x_0) \leq \tilde{R}_{T0}(x_0) := R_{T0}(x_{\max},x_0) = \frac{\sqrt{\sqrt{\Delta t^2 x_0^4+16\nu^2\pi^2}+x_0^2\Delta t}-\sqrt{2\Delta t}x_0}{\sqrt{\sqrt{\Delta t^2 x_0^4+16\nu^2\pi^2}+x_0^2\Delta t}+\sqrt{2\Delta t}x_0}. \tag{5.52}$$

For Δt *small, we have* $\tilde{R}_{T0}(x_0)=1-x_0\sqrt{\frac{2}{\nu\pi}}\sqrt{\Delta t}+O(\Delta t)$.

Proof. By the monotonicity of R_{T0} in x, the bound for $x_0 \leq x \leq x_{\max}$ on R_{T0} is attained at $x=x_{\max}$, which leads, using the variable transform (5.21) and $\omega_{\max}=\frac{\pi}{\Delta t}$, to the bound given in (5.52). ∎

Now we can compare the asymptotic performance of the algorithm without overlap to the performance of the algorithm with overlap. If in the discretization the time step Δt is linked to the spatial discretization step Δx by the relation $\Delta t = C_2 \Delta x^\beta$, then we see by comparing the results of Theorem 5.10 with the results of Theorem 5.16 that for $\beta \geq 1$ adding an overlap of size Δx does improve the asymptotic performance of the algorithm, whereas for $\beta < 1$ adding an overlap of the order of Δx does not improve the asymptotic performance. In particular this shows that with the Taylor transmission conditions and using an explicit time discretization with the stability constraint $\Delta t = C_1 \Delta x^2$, an overlap is helpful. Note that if an explicit scheme is used with the same time steps in both subdomains, there is no need to iterate, since one can explicitly advance the algorithm on the interface as well. A subdomain iteration would still be of interest if one uses nonmatching time grids, however, see, for example, [11].

5.7. Optimization of the algorithm without overlap. As in the case with overlap, there is a better choice for p than the low frequency approximation based on a Taylor expansion of the optimal symbol. We can again try to choose p such that the convergence factor

$$R_0(x,p,x_0) = \frac{(x-p)^2+x^2-x_0^2}{(x+p)^2+x^2-x_0^2} \tag{5.53}$$

is minimized over all $x_0 \le x \le x_{\max}$. Hence the optimal choice for p for the discretized algorithm is the solution of the min-max problem

$$\min_p \left(\max_{x_0 \le x \le x_{\max}} R_0(x,p,x_0) \right) = \min_{p \ge 0} \left(\max_{x_0 \le x \le x_{\max}} \frac{(x-p)^2+x^2-x_0^2}{(x+p)^2+x^2-x_0^2} \right), \tag{5.54}$$

where minimizing over nonnegative p is equivalent to minimizing over all p, as one can see from (5.53). The following theorem can be proved as in the case with overlap.

THEOREM 5.17 (O0 performance without overlap). *Let $L = 0$, $x_0 := \sqrt{a^2+4\nu b}$, and $x_{\max} < \infty$ be given. Then the best performance of the optimized nonoverlapping Schwarz waveform relaxation algorithm with Robin transmission conditions* (4.1),(4.6) *is obtained for $p = p^*$, where p^*, the solution of the min-max problem* (5.54)*, is for $x_{\max} \ge \frac{1+\sqrt{5}}{2} x_0$ given by*

(5.55)

$$p^* = \sqrt{x_0(2x_{\max}+x_0)}, \qquad R_0(x,p^*,x_0) \le \tilde{R}_{O0} = \frac{x_{\max}+x_0-\sqrt{2x_{\max}x_0+x_0^2}}{x_{\max}+x_0+\sqrt{2x_{\max}x_0+x_0^2}},$$

and for $x_{\max} < \frac{1+\sqrt{5}}{2} x_0$ we have

$$p^* = \sqrt{2x_{\max}^2 - x_0^2}, \qquad R_0(x,p^*,x_0) \le \tilde{R}_{O0} = \frac{\sqrt{2x_{\max}^2-x_0^2}-x_{\max}}{\sqrt{2x_{\max}^2-x_0^2}+x_{\max}}. \tag{5.56}$$

THEOREM 5.18 (O0 discrete convergence estimate without overlap). *Let $L = 0$ and $x_0 := \sqrt{a^2+4\nu b}$. If the nonoverlapping Schwarz waveform relaxation algorithm with optimized Robin transmission conditions is discretized in time with time step Δt, then for Δt small we have*

$$p^* = (2^3 x_0^2 \pi\nu)^{\frac{1}{4}} \Delta t^{-\frac{1}{4}} + O(\Delta t^{\frac{1}{4}}), \quad \tilde{R}_{O0} = 1 - 2\left(\frac{2x_0^2}{\pi\nu}\right)^{\frac{1}{4}} \Delta t^{\frac{1}{4}} + O(\sqrt{\Delta t}). \tag{5.57}$$

Proof. Using the variable transform (5.21), $x_{\max}$ behaves like

$$x_{\max} = \sqrt{2\pi\nu}\Delta t^{-\frac{1}{2}} + O(\sqrt{\Delta t}),$$

and thus the first result of (5.55) in Theorem 5.17 applies. Expanding p^* and $\tilde{R}_{O0}$ from (5.55) leads to (5.57). ∎

To summarize the results of this section, we show in Table 5.2 an overview of the performance one can obtain with the various choices of the parameter p in the transmission conditions. It is interesting to note that, for optimized Schwarz waveform relaxation methods, without overlap does not necessarily mean less performance than with overlap: in the T0 case, if $\beta \le 1$, the performance of the overlapping and nonoverapping algorithms is the same, and the same holds in the O0 case if $\beta \le \frac{4}{3}$.

TABLE 5.2
Summary of the asymptotic convergence factors for the various parameter choices in the Robin transmission conditions for $\Delta t = \Delta x^\beta$.

Method	Convergence factor	Parameter p
T0 overlap Δx	$\begin{cases} 1 - O(\sqrt{\Delta x}) & \text{if } \beta \geq 1 \\ 1 - O(\Delta x^{\frac{\beta}{2}}) & \text{if } \beta < 1 \end{cases}$	$\sqrt{a^2 + 4\nu b}$
O0 overlap Δx	$\begin{cases} 1 - O(\Delta x^{\frac{1}{3}}) & \text{if } \beta \geq \frac{4}{3} \\ 1 - O(\Delta x^{\frac{\beta}{4}}) & \text{if } \beta < \frac{4}{3} \end{cases}$	$(2\nu(a^2 + 4\nu b))^{\frac{1}{3}} \Delta x^{-\frac{1}{3}}$ $(8\nu\pi(a^2 + 4\nu b))^{\frac{1}{4}} \Delta x^{-\frac{\beta}{4}}$
T0 no overlap	$1 - O(\sqrt{\Delta t})$	$\sqrt{a^2 + 4\nu b}$
O0 no overlap	$1 - O(\Delta t^{\frac{1}{4}})$	$(8\nu\pi(a^2 + 4\nu b))^{\frac{1}{4}} \Delta t^{-\frac{1}{4}}$

6. Numerical results. We perform in this section numerical experiments to measure the convergence factors of the numerical implementation of the Schwarz waveform relaxation algorithms analyzed at the continuous level in this paper. We use the parabolic model problem (2.1) with $\Omega = (0, 6)$. We impose homogeneous boundary conditions, $u(0, t) = 0$ and $u(6, t) = 0$, and use various initial conditions $u(x, 0)$, $x \in \Omega$. We first use a decomposition of the domain Ω into the two subdomains $\Omega_1 = (0, L_2)$ and $\Omega_2 = (L_1, 6)$, $L_1 \leq L_2$, and hence $L = L_2 - L_1$. We refer with the term iteration to a double iteration of the respective algorithms, since for two subdomains one can perform all the iterations in an alternating fashion and thus obtain the even iterates on one subdomain and the odd ones on the other without having to compute the remaining ones. We show results of numerical experiments for only the algorithm with overlap since with overlap we can compare the results to the classical Schwarz waveform relaxation algorithm with Dirichlet transmission conditions, which does not converge without overlap.

6.1. Dirichlet transmission conditions. In this first set of experiments, we use the classical Schwarz waveform relaxation algorithm with Dirichlet transmission conditions analyzed in section 3. We chose for the problem parameters $\nu = 0.2$, $a = 1$, and $b = 0$. We discretize (2.1) using an upwind finite difference discretization in space with mesh parameter $\Delta x = 0.02$ and a backward Euler discretization in time, with time step $\Delta t = 0.005$. We chose $L_1 = 2.96$ and $L_2 = 3.04$, which means the overlap is $L = 0.08$, and we compute the numerical solution in the time interval $[0, T]$. Using as initial condition

$$u(x, 0) = e^{-3(1.2-x)^2},$$

we have already shown in Figure 3.1 for this example the first few iterations at the end of the time interval $T = 2.5$, where we started the algorithm with a zero initial guess. We show in Figure 6.1 the convergence behavior of the classical Schwarz waveform relaxation algorithm for this example for three different lengths of the time interval, $T = 1$, $T = 2.5$, and $T = 10$, together with the linear bound on the convergence rate from Theorem 3.3 and the superlinear convergence bound from Theorem 3.4. The dashed curve shows the error measured in the L_2 norm between the converged solution and the iterates at the interface L_2. One can clearly see that the behavior of the algorithm depends on the length of the time interval T, as predicted by Theorem 3.4. For short time intervals, the superlinear bound on the convergence rate is sharper, and hence the algorithm must converge superlinearly, as shown in Figure 6.1 on the left. If the time interval becomes longer, as in the middle graph of Figure 6.1, the linear bound of Theorem 3.3 is sharper than the superlinear one early in the iteration, and hence the algorithm converges linearly. But later the superlinear bound becomes

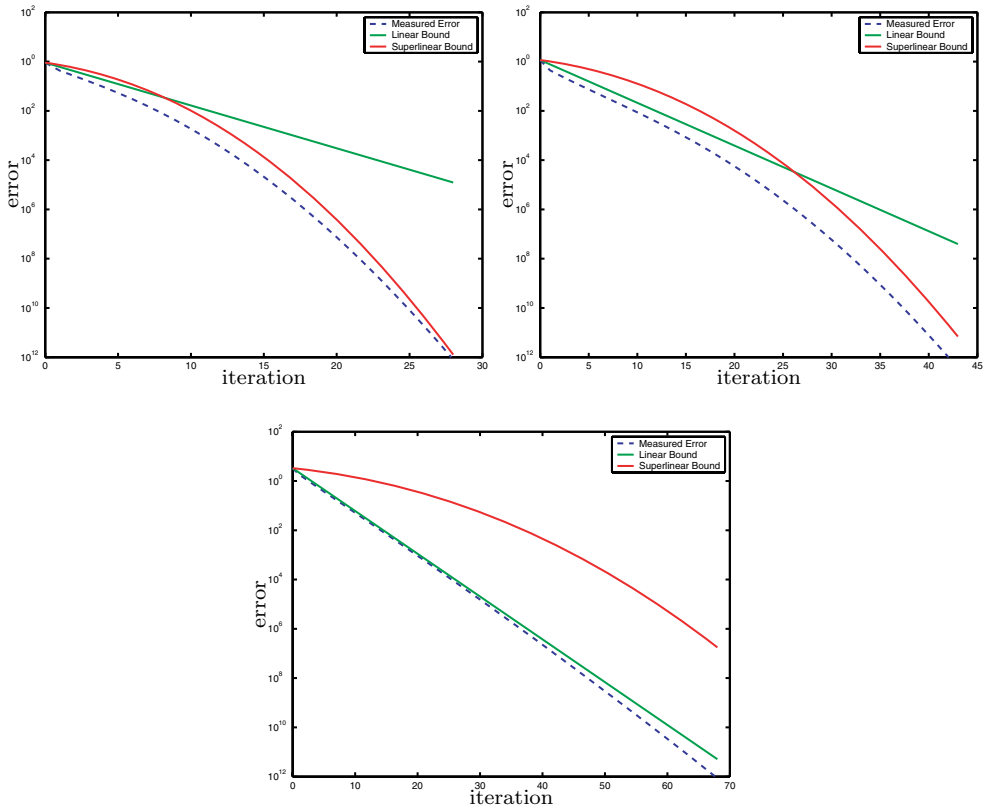


FIG. 6.1. *Convergence rate of the classical Schwarz waveform relaxation algorithm with Dirichlet transmission conditions together with the theoretical linear and superlinear bounds on the convergence rates: on the left for* $T = 1$*, in the middle for* $T = 2.5$*, and on the right for* $T = 10$*.*

sharper and hence a transition to the superlinear convergence regime occurs. For long time intervals, the initial linear convergence regime also prevails for more iterations, as one can see in Figure 6.1 on the right.

6.2. Robin transmission conditions. We now change the transmission conditions in the Schwarz waveform relaxation algorithm to Robin transmission conditions. Using the same numerical configuration as in the previous subsection, we obtain for the parameter p in the transmission conditions using a Taylor expansion $p = p_T = 1$, and using the optimization from Theorem 5.13, we obtain $p = p^* = 2.054275607$. In Figure 3.1 on the right, we have already seen the first few iterations at the end of the time interval $T = 2.5$ for this example with the optimal parameter p^*, starting the iteration with the zero initial guess. In Figure 6.2 one can see how much faster the algorithm converges with Robin transmission conditions compared to the classical algorithm. One can also see that the optimized parameter p^* leads to an even better performance than the parameter p_T from the Taylor transmission conditions. Note that for all the results comparing the performance of the algorithms, we started the iteration with a random initial guess. This is important to obtain a relevant comparison since, for smooth solutions starting with a smooth initial guess, high frequencies would not be present on the mesh and thus a much coarser mesh would have been sufficient for the computation. The random initial guess has the effect that the mesh resolution is indeed needed to resolve the iteration and thus corresponds to the

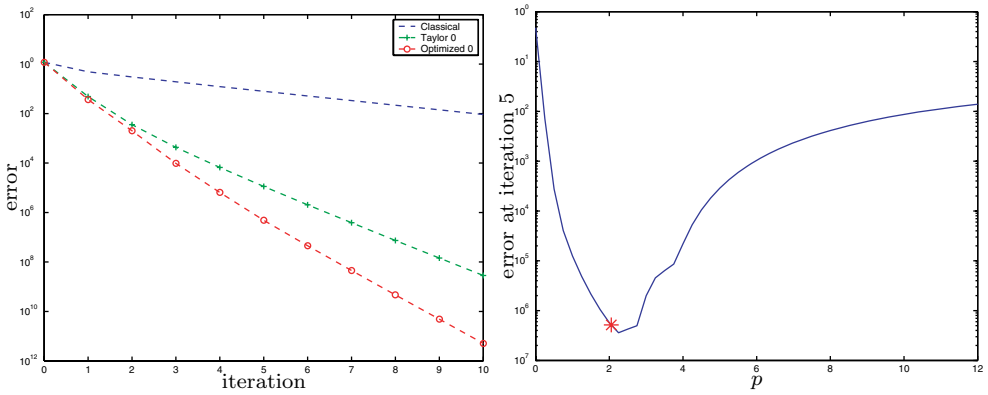


FIG. 6.2. *Left: Convergence rates of the classical Schwarz waveform relaxation algorithm with Dirichlet transmission conditions compared to the same algorithm with the new Robin transmission conditions, with the low frequency Taylor approximation or optimized. Right: The error obtained running the algorithm with Robin transmission conditions for* 5 *steps and various choices of the free parameter* p*, and indicated by a star the choice* p^* *predicted by the theory.*

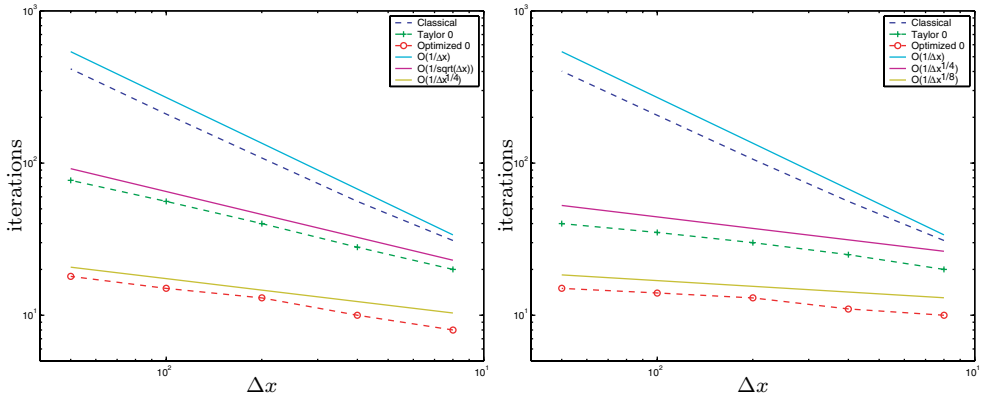


FIG. 6.3. *Asymptotic behavior as the mesh is refined with an overlap* $L = \Delta x$*: on the left the case where* $\Delta t = O(\Delta x)$ *and on the right the case where* $\Delta t = O(\sqrt{\Delta x})$*, together with the predicted rates from the analysis, both for the classical and the optimized Schwarz waveform relaxation algorithms with Taylor and optimized Robin transmission conditions.*

relevant case in practice.

Next, we verify if the optimal choice for the parameter $p = p^*$ derived using the continuous Fourier analysis in Theorem 5.13 really corresponds to the best choice one can make in the fully discretized algorithm. In Figure 6.2 on the right we show the error obtained after running the Schwarz waveform relaxation algorithm with Robin transmission conditions for five steps using various values for the free parameter p in the transmission conditions. The optimal choice p^* from Theorem 5.13 is indicated by a star. Clearly the continuous analysis predicts the optimal choice of the parameter p very well.

Finally, we illustrate the asymptotic analysis by performing two sets of experiments according to Theorems 5.10 and 5.14. We choose the same problem parameters as before but start now with a coarser mesh both in space and time, $\Delta x = 0.08$ and $\Delta t = 0.02$, and we fix the overlap to be $L = \Delta x$. We then run the classical and optimized Schwarz waveform relaxation algorithms with Taylor–Robin transmission

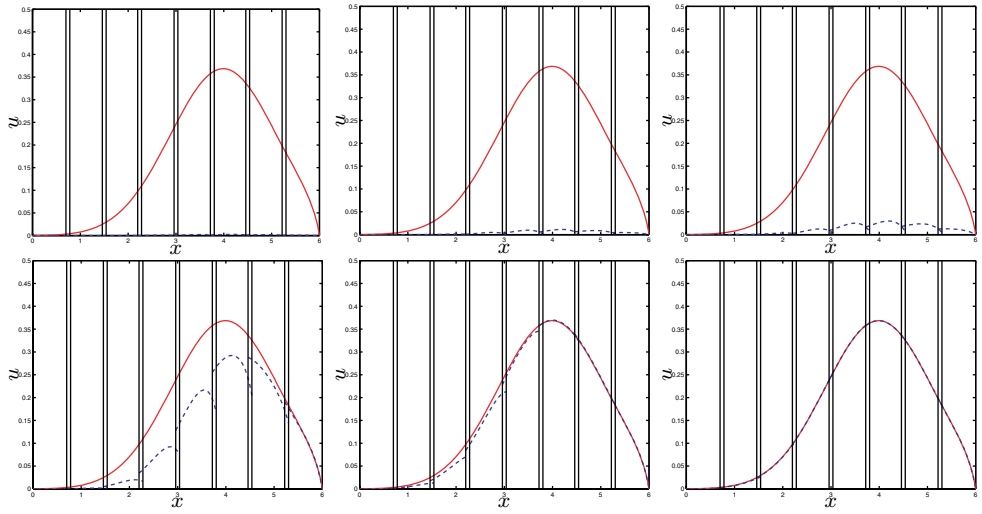


FIG. 6.4. *From left to right, the first iterates $u_j^k(x,T)$, $j=1,\ldots,8$, (dashed) at the end of the time interval $t=T$ together with the exact solution (solid) for the same model problem as before: top row the classical algorithm and bottom row the optimized algorithm.*

conditions until the error becomes smaller than 10^{-6} and count the number of iterations. We repeat this experiment dividing Δx and Δt by 2 several times, which implies $\Delta t = O(\Delta x)$. This corresponds to (3.16) for the classical algorithm, where the convergence factor should behave like $1 - O(\Delta x)$. For the algorithm with Taylor transmission conditions it corresponds to the case in Theorem 5.10, where the convergence factor should behave like $1 - O(\sqrt{\Delta x})$, and for the optimized algorithm it corresponds to the case in Theorem 5.14, where the convergence factor should behave like $1 - O(\Delta x^{\frac{1}{4}})$. Figure 6.3 shows on the left the results obtained from these experiments. One can see that the asymptotic analysis predicts very well the numerical behavior of the algorithms. Next, we perform a similar experiment, starting with the same values for Δx and Δt, but now we divide Δx by 2 each time and Δt only by $\sqrt{2}$ (such a refinement is admissible since our scheme is implicit), which implies

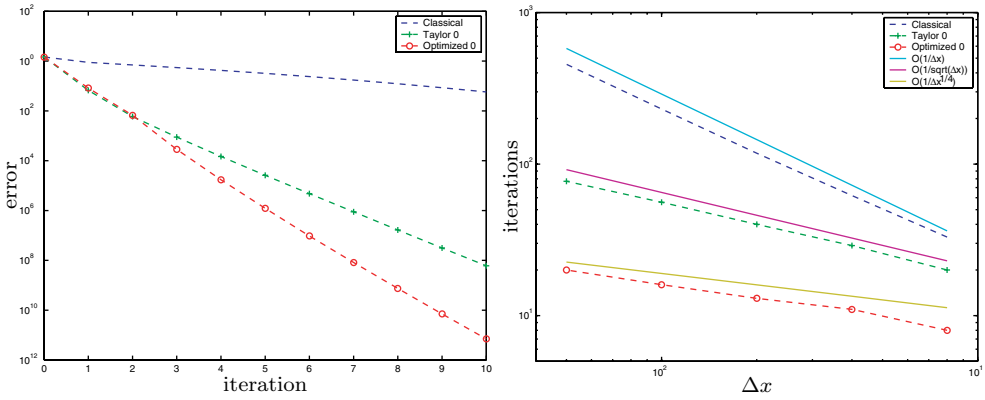


FIG. 6.5. *Left: convergence rate comparison for the eight subdomain case. Right: Asymptotic behavior as the mesh is refined with an overlap $L=\Delta x$ for the eight subdomain case, with $\Delta t = O(\Delta x)$, together with the predicted rates from the two subdomain analysis.*

$\Delta t = O(\sqrt{\Delta x})$. While this does not change anything for the classical algorithm, which still has the same bad convergence factor $1 - O(\Delta x)$, for the algorithm with Taylor–Robin transmission conditions now case 3 of Theorem 5.10 applies, and the algorithm should show the much better convergence factor $1 - O(\Delta x^{\frac{1}{4}})$. The optimized algorithm has according to Theorem 5.14 now the even better convergence factor $1 - O(\Delta x^{\frac{1}{8}})$, virtually independent of Δx. In Figure 6.3 on the right, one can clearly see that this is the case. The algorithm has different asymptotic convergence factors with the same overlap, depending on the discretization in time, as predicted.

6.3. Experiments with many subdomains. We now show experiments which indicate that the results we obtained for two subdomains are also relevant for many subdomains. Using the same model problem as before, we now decompose the domain into eight subdomains. In Figure 6.4, we show in the top row the first three iterations of the classical Schwarz waveform relaxation algorithm, and below we show the same iterations for the algorithm with optimized Robin transmission conditions. This clearly shows how important the transmission conditions are in the many subdomain case. We show the corresponding convergence rates in Figure 6.5 on the left, and on the right we perform the same asymptotic experiments as in Figure 6.3 on the left but now with eight subdomains, which indicates that the results of Theorems 5.10 and 5.14 also hold for more than two subdomains.

7. Conclusions. We have analyzed Schwarz waveform relaxation algorithms for advection reaction diffusion equations. We have shown that these methods, using the classical Dirichlet transmission conditions, are well defined and have a convergence rate which is bounded both by a linear and a superlinear rate. Both rates can be sharp, depending on the length of the time interval of the simulation. We then showed that there exist much better transmission conditions than the classical Dirichlet conditions. Optimal transmission conditions are transparent conditions, but they are in general nonlocal and thus less convenient to use. We introduced instead Robin transmission conditions in the Schwarz waveform relaxation algorithm, showed that the new algorithm is well posed and convergent, even if there is no overlap, and analyzed how to chose the free parameter in the new transmission conditions. We also gave asymptotic results when the overlap or the mesh parameters become small. We finally illustrated our findings with numerical experiments which document the relevance of our continuous analysis.

REFERENCES

[1] K. Burrage, *Parallel and Sequential Methods for Ordinary Differential Equations*, Oxford University Press, New York, 1995.

[2] X.-C. Cai, *Additive Schwarz algorithms for parabolic convection-diffusion equations*, Numer. Math., 60 (1991), pp. 41–61.

[3] X.-C. Cai, *Multiplicative Schwarz methods for parabolic problems*, SIAM J. Sci. Comput., 15 (1994), pp. 587–603.

[4] P. D'Anfray, L. Halpern, and J. Ryan, *New trends in coupled simulations featuring domain decomposition and metacomputing*, M2AN Math. Model. Numer. Anal., 36 (2002), pp. 953–970.

[5] D. S. Daoud and M. J. Gander, *Overlapping Schwarz waveform relaxation for convection reaction diffusion problems*, in Proceedings of the 13th International Conference on Domain Decomposition, Theory and Engineering Applications of Computational Methods, 2001, pp. 253–260.

[6] B. Després, *Méthodes de décomposition de domaine pour les problèmes de propagation d'ondes en régimes harmoniques*, Ph.D. thesis, Paris IX, 1991.

[7] B. Engquist and A. Majda, *Absorbing boundary conditions for the numerical simulation of waves*, Math. Comp., 31 (1977), pp. 629–651.
[8] B. Engquist and H.-K. Zhao, *Absorbing boundary conditions for domain decomposition*, Appl. Numer. Math., 27 (1998), pp. 341–365.
[9] M. J. Gander and G. H. Golub, *A non-overlapping optimized Schwarz method which converges with an arbitrarily weak dependence on h*, in Proceedings of the Fourteenth International Conference on Domain Decomposition Methods, National Autonomous University of Mexico, 2002.
[10] M. J. Gander, L. Halpern, and F. Nataf, *Optimal convergence for overlapping and non-overlapping Schwarz waveform relaxation*, in Proceedings of the Eleventh International Conference of Domain Decomposition Methods, Domain Decomposition Press, 1999.
[11] M. J. Gander, L. Halpern, and F. Nataf, *Optimal Schwarz waveform relaxation for the one dimensional wave equation*, SIAM J. Numer. Anal., 41 (2003), pp. 1643–1681.
[12] M. J. Gander and L. Halpern, *Méthodes de relaxation d'ondes pour l'équation de la chaleur en dimension* 1, C. R. Acad. Sci. Paris Ser. I, 336 (2003), pp. 519–524.
[13] M. J. Gander and A. M. Stuart, *Space-time continuous analysis of waveform relaxation for the heat equation*, SIAM J. Sci. Comput., 19 (1998), pp. 2014–2031.
[14] M. J. Gander and H. Zhao, *Overlapping Schwarz waveform relaxation for parabolic problems in higher dimension*, in Proceedings of Algoritmy 14, Slovak Technical University, 1997, pp. 42–51.
[15] M. J. Gander, *Analysis of Parallel Algorithms for Time Dependent Partial Differential Equations*, Ph.D. thesis, Stanford University, Stanford, CA, 1997.
[16] M. J. Gander, *Overlapping Schwarz for parabolic problems*, in Proceedings of the Ninth International Conference on Domain Decomposition Methods, Domain Decomposition Press, 1997, pp. 97–104.
[17] M. J. Gander, *Optimized Schwarz Methods*, Siam J. Numer. Anal., 44 (2006), pp. 699–731
[18] E. Giladi and H. B. Keller, *Space time domain decomposition for parabolic problems*, Numer. Math., 93 (2002), pp. 279–313.
[19] L. Halpern, *Artificial boundary conditions for the advection-diffusion equations*, Math. Comp., 174 (1986), pp. 425–438.
[20] M. Hu, K. Jackson, J. Janssen, and S. Vandewalle, *Remarks on the optimal convolution kernel for CSOR waveform relaxation*, Adv. Comput. Math., 7 (1997), pp. 135–156.
[21] J. Janssen and S. Vandewalle, *Multigrid waveform relaxation on spatial finite element meshes: The continuous-time case*, SIAM J. Numer. Anal., 33 (1996), pp. 456–474.
[22] J. Janssen and S. Vandewalle, *Multigrid waveform relaxation on spatial finite element meshes: The discrete-time case*, SIAM J. Sci. Comput., 17 (1996), pp. 133–155.
[23] J. Janssen and S. Vandewalle, *On SOR waveform relaxation methods*, SIAM J. Numer. Anal., 34 (1997), pp. 2456–2481.
[24] C. Japhet, *Optimized Krylov-Ventcell method. Application to convection-diffusion problems*, in Proceedings of the 9th International Conference on Domain Decomposition Methods, Domain Decomposition Press, 1998, pp. 382–389.
[25] R. Jeltsch and B. Pohl, *Waveform relaxation with overlapping splittings*, SIAM J. Sci. Comput., 16 (1995), pp. 40–49.
[26] E. Lelarasmee, A. E. Ruehli, and A. L. Sangiovanni-Vincentelli, *The waveform relaxation method for time-domain analysis of large scale integrated circuits*, IEEE Trans. CAD IC Syst., 1 (1982), pp. 131–145.
[27] J.-L. Lions and E. Magenes, *Problèmes aux limites non homogènes et applications*, Trav. Recherches Math. 17–18, Dunod, Paris, 1968.
[28] P.-L. Lions, *On the Schwarz alternating method.* III: *A variant for nonoverlapping subdomains*, in Proceedings of the Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, Houston, TX, SIAM, Philadelphia, PA, 1990.
[29] C. Lubich and A. Ostermann, *Multi-grid dynamic iteration for parabolic equations*, BIT, 27 (1987), pp. 216–234.
[30] C. Lubich and A. Schädle, *Fast convolution for non-reflecting boundary conditions*, SIAM J. Sci. Comput., 24 (2002), pp. 161–182.
[31] V. Martin, *An optimized Schwarz waveform relaxation method for unsteady convection diffusion equation*, Appl. Numer. Math., 52 (2005), pp. 401–428.
[32] G. A. Meurant, *Numerical experiments with a domain decomposition method for parabolic problems on parallel computers*, in Proceedings of the Fourth International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, PA, 1991.
[33] F. Nataf, F. Rogier, and E. de Sturler, *Optimal Interface Conditions for Domain Decom-*

position Methods, Technical report 301, CMAP (Ecole Polytechnique), Palaiseau, France, 1994.

[34] F. Nataf and F. Rogier, *Factorization of the convection-diffusion operator and the Schwarz algorithm*, M^3AS, 5 (1995), pp. 67–93.

[35] O. Nevanlinna, *Remarks on Picard-Lindelöf iterations part i*, BIT, 29 (1989), pp. 328–346.

[36] R. Rannacher and G. Zhou, *Analysis of a domain-splitting method for nonstationary convection-diffusion problems*, East-West J. Numer. Math., 2 (1994), pp. 151–174.

[37] W. Rudin, *Real and Complex Analysis*, McGraw–Hill, New York, 1966.

[38] S. Vandewalle and G. Horton, *Fourier mode analysis of the multigrid waveform relaxation and time-parallel multigrid methods*, Computing, 54 (1995), pp. 317–330.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 698–730

THE OPTIMAL CONVERGENCE OF THE *h-p* VERSION OF THE FINITE ELEMENT METHOD WITH QUASI-UNIFORM MESHES*

BENQI GUO† AND WEIWEI SUN‡

Dedicated to Professor R. S. Wong on the occasion of his sixtieth birthday

Abstract. In the framework of the Jacobi-weighted Besov spaces, we analyze the convergence of the *h-p* version of finite element solutions on quasi-uniform meshes and the lower and upper bounds of errors for elliptic problems on polygons. Both lower and upper bounds are proved to be optimal in h and p, which leads to the optimal convergence of the *h-p* version of the finite element method with quasi-uniform meshes for elliptic problems on polygons. The results proved for the *h-p* version include the *h*-version with quasi-uniform meshes and the *p*-version with quasi-uniform degrees as two special cases.

Key words. the *h-p* version, finite element method, quasi-uniform meshes, singularity, Jacobi-weighted Besov and Sobolev spaces, optimal rate of convergence

AMS subject classifications. 65N30, 65N25, 35D10

DOI. 10.1137/05063756X

1. Introduction.

In this paper we investigate the convergence of the *h-p* version of the finite element method (FEM) with quasi-uniform meshes in two dimensions in the framework of the Jacobi-weighted Besov spaces. In particular, we prove asymptotically exact upper and lower bounds for the approximation error in finite element solutions of the *h-p* version for elliptic problems on polygonal domains whose solutions exhibit typical corner singularities. Our analysis is done within the framework of the Jacobi-weighted Besov spaces, which already has been proved an appropriate tool for obtaining optimal estimates for the *p*-version of the FEM for this type of problem; see [5, 6, 7]. Here we incorporate the mesh dependence into the analysis for the *p*-version and provide optimal estimates for any combination of mesh size and polynomial degree for the case of quasi-uniform meshes and uniform polynomial degrees.

The *p*-version of FEM uses a fixed mesh and improves the approximation of the solution by increasing degrees of piecewise polynomials. The *h*-version is based on mesh refinement and piecewise polynomials of low and fixed degrees. The *h-p* version combines mesh refinement with an increase of degrees. Let us recall the main theoretical achievements for the *h-p* version since its beginning. An analysis of the *h-p* version of the FEM in one dimension was given by Gui and Babuška [16]. They considered the approximation of typical singularities x^γ and proved optimal upper and lower bounds of error in the p and *h-p* finite element solutions in H^1 and L^2 norms. The approximation of singularities $x^\gamma \log^\nu x$ in one dimension was addressed, and

*Received by the editors August 4, 2005; accepted for publication (in revised form) October 20, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sinum/45-2/63756.html

†Department of Mathematics, Shanghai Normal University, Shanghai, China, and Department of Mathematics, University of Manitoba, Winnipeg, MB R3T 2N2, Canada (guo@cc.umanitoba.ca). The work of this author was partially supported by the NSERC of Canada under grant OGP0046726, and partially supported by E-institutes of Shanghai Municipal Education Commission under project E03004 while this author visited Shanghai Normal University in 2006.

‡Department of Mathematics, City University of Hong Kong, Kowloon Tong, Hong Kong, China (maweiw@cityu.edu.hk). The work of this author was supported by the Research Grant Council of the Hong Kong Special Administrative Region, China, under project CityU 102103.

asymptotic analysis on the approximation error was given in [7]. The h-p version of the FEM with quasi-uniform meshes in two dimensions was analyzed by Babuška and Suri [10] after improving the approximation results of the p-version of the FEM [9]. They gave an upper bound of error in finite element approximation for elliptic problems with singularities $r^\gamma \log^\nu r$ ($r = |x|$ is the distance to the origin), which is actually of order $O(h^{-\gamma} p^{-2\gamma} \log^\nu (p/h))$. This upper bound is sharp for noninteger γ, and it can be sharper for integer γ and $\nu > 0$. Since then, the h-p FEM has been widely used for various scientific and engineering computations such as solid and fluid mechanics and electromagnetics [2, 3, 4, 13, 15, 29, 30]. The h-p FEM has been generalized to the h-p versions of the boundary element method (BEM) [19, 20, 28] and the discontinuous Galerkin method (DG FEM)[24, 25, 27, 26, 31] to solve linear and nonlinear elliptic, parabolic, and hyperbolic problems as well as multiscale problems. Commercial and research software packages based on the h-p FEM such as MSC/PROBE (MacNeal Schwendler, California), Poly FEM (IBM, Massachusetts), PHLEX (Computational Mechanics, Texas), STRESSCHECK (Engineering Software Research & Development, Missouri), and STRIPE (Aeronautical Research Institute of Sweden) are now widely used in engineering computation. While great progress in theory, algorithms, and applications of the h-p FEM were made in the last two decades, there has been no breakthrough in the optimal convergence of the h-p FEM for elliptic problems with singularities of $r^\gamma \log^\nu r$-type on polygonal domains. The lower bound and the sharpest estimation of the upper bound of the error in the h-p FEM solutions for elliptic problems with singular solutions of $r^\gamma \log^\nu r$-type have not been addressed until now. Therefore, the optimal convergence of the h-p version of FEM has not been well established yet.

The h-p version with quasi-uniform meshes is, from methodology and approximation theory, the p-version on scaled meshes. The approach of the p-version gives the p-dependence in the approximation errors, and a proper scaling argument will fully reveal the information of the h-dependence. Hence, the analysis for the best approximation of the h-p version with quasi-uniform meshes is not feasible unless the optimal convergence of the p-version is rigorously proved. Fortunately, the best a priori error estimation for the p-version has been recently established, and we are now ready to pursue the best a priori error estimation for the h-p version. In the last few years, with a series of papers by one of the authors and his collaborators [5, 6, 7, 19], a new analysis of the p-version has started in the framework of the Jacobi-weighted Besov and Sobolev spaces. The approximation theory of the FEM and BEM in two dimensions in this new mathematical framework is systematically developed in these papers, which demonstrates that Jacobi-weighted Besov and Sobolev spaces are the most appropriate tools for obtaining optimal upper and lower bounds of approximation errors when dealing with singular solutions on polygons. This framework has been generalized to the p-version of the FEM in three dimensions [17] and the h-p version of the BEM [20]. The Jacobi projection and interpolation have been developed recently in the spectral methods as well; see, e.g., [12, 21, 22, 23] and references therein. In this paper we will further develop the approximation theory of the h-p version of the FEM with quasi-uniform meshes. Based on the analysis of the upper and lower bounds of approximation error in the FEM solutions in terms of h and p, we will establish the optimal convergence of the h-p version of the FEM for elliptic problems on polygonal domains.

The rest of the paper is organized as follows. In section 2 we analyze approximation in the Jacobi-weighted Besov and Sobolev spaces on a standard domain

$Q = (-1,1)^2$. We introduce the Jacobi-weighted Besov and Sobolev spaces on a scaled domain $Q_h = (-h,h)^2$ in section 3 and analyze the approximation properties for smooth and singular functions of $r^\gamma \log^\nu r$-type in the framework of these spaces. These approximation results on scaled square domains are applied in section 4 to prove the convergence of the *h-p* version FEM with quasi-uniform meshes for elliptic problems with smooth solutions, and to derive the sharpest estimate of lower and upper bounds of error in FEM solutions of the *h-p* version for elliptic problems with singular solutions on polygonal domains, which leads to the optimal convergence of the *h-p* version of the FEM with quasi-uniform meshes. In the last section, we will make some concluding remarks.

2. Approximation in Jacobi-weighted spaces on a square domain.

2.1. Weighted Besov and Sobolev spaces.

Let $Q = I^2 = (-1,1)^2$, and let

$$w_{\alpha,\beta}(x) = \prod_{i=1}^{2} (1-x_i^2)^{\alpha_i+\beta_i} \tag{2.1}$$

be a weight function with integer $\alpha_i \geq 0$ and real number $\beta_i > -1$, which is referred to as the Jacobi weight. The weighted Sobolev space $H^{k,\beta}(Q)$ is defined as a closure of C^∞ functions in the norm with the Jacobi weight

$$\|u\|^2_{H^{k,\beta}(Q)} = \sum_{|\alpha|=0}^{k} \int_Q |D^\alpha u|^2 \, w_{\alpha,\beta}(x)\, dx, \tag{2.2}$$

where $D^\alpha u = u_{x_1^{\alpha_1},x_2^{\alpha_2}}$, $\alpha = (\alpha_1,\alpha_2), |\alpha| = \alpha_1+\alpha_2$, and $\beta = (\beta_1,\beta_2)$. By $|u|_{H^{k,\beta}(Q)}$ we denote the seminorm,

$$|u|_{H^{k,\beta}(Q)} = \sum_{|\alpha|=k} \int_Q |D^\alpha u|^2 \, w_{\alpha,\beta}(x)\, dx. \tag{2.3}$$

The Jacobi-weighted Sobolev space $H^{s,\beta}(Q)$ and Besov space $B^{s,\beta}(Q)$ are defined as interpolation spaces $\mathcal{B}^{s,\beta}_{2,q}(Q) = (H^{\ell,\beta}(Q), H^{k,\beta}(Q))_{\theta,q}$ with $q = 2$ and $q = \infty$, respectively, by the K-method, where $0 < \theta < 1$, $1 \leq q \leq \infty$, $s = (1-\theta)\ell + \theta k$, and ℓ and k are integers, $\ell < k$, furnished with norms

$$\|u\|_{\mathcal{B}^{s,\beta}_{2,2}(Q)} = \left(\int_0^\infty t^{-2\theta} \, |K(t,u)|^2 \frac{dt}{t} \right)^{1/q}, \quad \|u\|_{\mathcal{B}^{s,\beta}_{2,\infty}(Q)} = \sup_{t>0} t^{-\theta} \, K(t,u) \tag{2.4}$$

with

$$K(t,u) = \inf_{u=v+w} \Big(\|v\|_{H^{\ell,\beta}(Q)} + t\|w\|_{H^{k,\beta}(Q)} \Big).$$

The space $H^{s,\beta}(Q)$ is called the Jacobi-weighted Sobolev space with fractional order if s is not an integer, and the space $B^{s,\beta}(Q)$ is referred as to the Jacobi-weighted Besov space.

The modified weighted Besov space $B^{s,\beta}_\nu(Q)$ with $\nu \geq 0$ is defined as an interpolation space by a modified K-method,

$$B^{s,\beta}_\nu(Q) = \Big(H^{\ell,\beta}(Q), H^{k,\beta}(Q) \Big)_{\theta,\infty,\nu},$$

with a modified norm,

$$\|u\|_{B^{s,\beta}_{\nu}(Q)} = \sup_{t>0} K(t,u)\ \frac{t^{-\theta}}{(1+|\log t|)^{\nu}}. \tag{2.5}$$

Remark 2.1. The spaces $H^{s,\beta}(Q)$ and $B^{s,\beta}(Q) = B^{s,\beta}_0(Q)$ are exact of θ-exponent, but $B^{s,\beta}_{\nu}(Q)$ with $\nu > 0$ is not. It was proved in [5] that $B^{s,\beta}_{\nu}(Q)$ is weakly exact of θ-exponent. Suppose that E realizes a linear operator: $H_l \to H^{m_l,\beta}(Q)$, $l = 1, 2$, with norms denoted by $\|E\|_l$, where $H_l, l = 1, 2$, are Banach spaces. Then E is a linear operator: $(H_1, H_2)_{\theta,q} \to (H^{m_1,\beta}(Q), H^{m_2,\beta}(Q))_{\theta,q,\nu}$ such that for $\nu = 0$

$$\|E\|_{(H_1,H_2)_{\theta,q}\to(H^{m_1,\beta}(Q),H^{m_2,\beta}(Q))_{\theta,q,0}} \le \|E\|_1^{1-\theta}\|E\|_2^{\theta} \tag{2.6}$$

and for $\nu > 0$

$$\|E\|_{(H_1,H_2)_{\theta,q}\to(H^{m_1,\beta}(Q),H^{m_2,\beta}(Q))_{\theta,\infty,\nu}} \le \|E\|_1^{1-\theta}\|E\|_2^{\theta}\left(1+\log\frac{\|E\|_2}{\|E\|_1}\right)^{\nu}. \tag{2.7}$$

For the exact interpolation spaces of θ-exponent we refer to [11], and for the weak exactness of θ-exponent for the space $B^{s,\beta}_{\nu}(Q), \nu > 0$, we refer to [5].

We shall introduce the Jacobi projection and analyze its properties. Let $J^{\alpha,\beta}_n(x)$ be the Jacobi polynomial of degree n,

$$J^{\alpha,\beta}_n(x) = \frac{(1-x)^{-\alpha}(1+x)^{-\beta}}{2^n\, n!}\,\frac{d^n\,(1-x)^{\alpha+n}(1+x)^{\beta+n}}{dx^n} \tag{2.8}$$

with $\alpha, \beta > -1$. In particular, for $\alpha = \beta$, we write $J^{\beta,\beta}_n(x) = J^{\beta}_n(x)$. The Jacobi polynomials $J^{\beta}_n(x)$ are orthogonal with the Jacobi weight $w_\beta(x) = (1-x^2)^{\beta}$:

$$\int_I J^{\beta}_m(x)\, J^{\beta}_n(x)\, w_\beta(x)\, dx = \begin{cases} \gamma^{\beta}_n, & m = n, \\ 0, & m \ne n \end{cases}$$

with

$$\gamma^{\beta}_n = \frac{2^{2\beta+1}\Gamma^2(n+\beta+1)}{(2n+2\beta+1)\Gamma(n+1)\Gamma(n+2\beta+1)}. \tag{2.9}$$

Let $J^{\beta}_{n,k}(x) = \frac{d^k}{dx^k} J^{\beta}_n(x)$. Then for $0 \le k \le n$

$$J^{\beta}_{n,k}(x) = 2^{-k}\frac{\Gamma(n+2\beta+k+1)}{\Gamma(n+2\beta+1)} J^{\beta+k}_{n-k}(x),$$

which are orthogonal with the Jacobi weight $w_{\beta+k}(x) = (1-x^2)^{\beta+k}$:

$$\int_I J^{\beta}_{m,k}(x) J^{\beta}_{n,k}(x)\, w_{\beta+k}(x)\, dx = \begin{cases} \gamma^{\beta}_{n,k}, & m = n \ge k, \\ 0 & \text{otherwise} \end{cases}$$

with

$$\gamma^{\beta}_{n,k} = \frac{2^{2\beta+1}\,\Gamma(n+2\beta+k+1)\,\Gamma^2(n+\beta+1)}{(2n+2\beta+1)\,\Gamma(n+1-k)\,\Gamma^2(n+2\beta+1)}. \tag{2.10}$$

By the Stirling formula [14], there hold asymptotically

(2.11)
$$\gamma_n^\beta = \frac{2^{2\beta+1}}{(2n+2\beta+1)}(1+O((n+1)^{-1/5})),\ \ \gamma_{n,k}^\beta = \frac{2^{2\beta+1}n^{2k}}{(2n+2\beta+1)}(1+O((n+1)^{-1/5})).$$

It is known [1, 23] that for $x \in [-1,1]$ there holds

$$|J_n^\beta(x)| \le C(n+1)^{\max\{\beta,-1/2\}} \tag{2.12}$$

and for $x = \pm 1$, we have the more precise estimation

$$|J_n^\beta(1)| = |J_n^\beta(-1)| \le C(n+1)^\beta. \tag{2.13}$$

For $u \in H^{k,\beta}(Q), k \ge 0$, with $\beta = (\beta_1, \beta_2)$, there is the Jacobi–Fourier expansion

$$u = \sum_{i,j=0}^{\infty} c_{i,j} J_i^{\beta_1}(x_1) J_j^{\beta_2}(x_2)$$

with

$$c_{i,j} = \frac{1}{\gamma_i^{\beta_1}\gamma_j^{\beta_2}} \int_Q u(x)\, J_i^{\beta_1}(x_1) J_j^{\beta_2}(x_2)\, w_\beta(x) dx,$$

where $J_i^{\beta_1}(x_1)$ and $J_j^{\beta_2}(x_2)$ are the Jacobi polynomials and $w_\beta(x) = (1-x_1^2)^{\beta_1}(1-x_1^2)^{\beta_1}$. Then

$$\|u\|_{L_\beta^2}^2(Q) = \sum_{i,j=0}^{\infty} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} \tag{2.14}$$

and

(2.15)
$$|u|_{H^{k,\beta}(Q)}^2 = \sum_{|\alpha|=k} \sum_{i\ge\alpha_1, j\ge\alpha_2} |c_{i,j}|^2 \gamma_{i,\,\alpha_1}^{\beta_1}\gamma_{j,\,\alpha_2}^{\beta_2} \cong \sum_{|\alpha|=k} \sum_{i\ge\alpha_1, j\ge\alpha_2} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2\alpha_1} j^{2\alpha_2}.$$

Hereafter $A \cong B$ means $c_1 B \le A \le c_2 B$ with constants c_1 and c_2 independent of A and B.

Let $\mathcal{P}_p(Q)$ with $p \ge 0$ be the set of all polynomials of (separate) degree $\le p$ on Q. Then

$$u_p = \prod_p^\beta u = \sum_{i,j=0}^{p} c_{i,j} J_i^{\beta_1}(x_1) J_j^{\beta_2}(x_2)$$

is the Jacobi projection of $u \in H^{k,\beta}(Q), k \ge 0$, on $\mathcal{P}_p(Q)$.

2.2. Approximation properties of the Jacobi projection on $Q = (-1,1)^2$.

THEOREM 2.1. *Let $u \in H^{k,\beta}(Q)$ with integer $k \ge 0$, $\beta = (\beta_1, \beta_2)$, $\beta_i > -1$, $1 \le i \le 2$, and let $u_p = \Pi_p^\beta u$ be its Jacobi projection on $\mathcal{P}_p(Q)$ with $p \ge 0$. Then we have for integer $\ell \le k$*

$$\|u - u_p\|_{H^{\ell,\beta}(Q)} \le C\,(p+1)^{-(k-l)}\ \|u\|_{H^{k,\beta}(Q)}. \tag{2.16}$$

Furthermore, if $u \in H^{k,\beta}(Q)$ with $k > 1$, $\beta_i \le -1/2$, $1 \le i \le 2$, then for $x \in \bar{Q}$

$$\|u - u_p\|_{C^0(\bar{Q})} \le C(p+1)^{-(k-l)} \|u\|_{H^{k,\beta}(Q)}, \tag{2.17}$$

$$|(u - u_p)(\pm 1, x_2)| \le C(p+1)^{-(k-3/2-\beta_1)} \|u\|_{H^{k,\beta}(Q)}, \tag{2.18}$$

$$|(u - u_p)(x_1, \pm 1)| \le C(p+1)^{-(k-3/2-\beta_2)} \|u\|_{H^{k,\beta}(Q)}, \tag{2.19}$$

$$|(u - u_p)(\pm 1, \pm 1)| \le C(p+1)^{-(k-2-\beta_1-\beta_2)} \|u\|_{H^{k,\beta}(Q)}. \tag{2.20}$$

If $p \ge k-1$, $k \ge 1$, $l \le k$, we have estimations in seminorms

$$|u - u_p|_{H^{\ell,\beta}(Q)} \le C\,(p+1)^{-(k-\ell)}\, |u|_{H^{k,\beta}(Q)}, \tag{2.21}$$

and if, in addition, $k > 1$, $\beta_i \le -1/2$, $1 \le i \le 2$, there hold for $x \in \bar{Q}$

$$|(u - u_p)(x)| \le C(p+1)^{-(k-1)} |u|_{H^{k,\beta}(Q)}, \tag{2.22}$$

$$|(u - u_p)(\pm 1, x_2)| \le C(p+1)^{-(k-3/2-\beta_1)} |u|_{H^{k,\beta}(Q)}, \tag{2.23}$$

$$|(u - u_p)(x_1, \pm 1)| \le C(p+1)^{-(k-3/2-\beta_2)} |u|_{H^{k,\beta}(Q)}, \tag{2.24}$$

$$|(u - u_p)(\pm 1, \pm 1)| \le C(p+1)^{-(k-2-\beta_1-\beta_2)} |u|_{H^{k,\beta}(Q)}. \tag{2.25}$$

The constants C in the above inequalities are independent of p and u.

Proof. By (2.15) we have for $\alpha = (\alpha_1, \alpha_2)$ with $|\alpha| = l$

$$\begin{aligned} \int_Q |D^\alpha (u - u_p)|^2 w_{\alpha,\beta}(x)dx &= \sum_{i \ge \max\{p+1,\alpha_1\}, j \ge \max\{p+1,\alpha_2\}} |c_{i,j}|^2 \gamma^{\beta_1}_{i,\alpha_1} \gamma^{\beta_2}_{j,\alpha_2} \\ &+ \left(\sum_{\alpha_1 \le i \le p, j \ge \max\{p+1,\alpha_2\}} + \sum_{i \ge \max\{p+1,\alpha_1\}, \alpha_2 \le j \le p} \right) |c_{i,j}|^2 \gamma^{\beta_1}_{i,\alpha_1} \gamma^{\beta_2}_{j,\alpha_2}. \end{aligned} \tag{2.26}$$

Due to (2.11) there hold

$$\begin{aligned} &\sum_{i \ge \max\{p+1,\alpha_1\}, j \ge \max\{p+1,\alpha_2\}} |c_{i,j}|^2 \gamma^{\beta_1}_{i,\alpha_1} \gamma^{\beta_2}_{j,\alpha_2} \\ &\le C \sum_{i \ge \max\{p+1,\alpha_1\}, j \ge \max\{p+1,\alpha_2\}} |c_{i,j}|^2 \gamma^{\beta_1}_{i} \gamma^{\beta_2}_{j} i^{2\alpha_1} j^{2\alpha_2} \\ &\le C(p+1)^{-2(k-l)} \sum_{i \ge \max\{p+1,\alpha_1\}, j \ge \max\{p+1,\alpha_2\}} |c_{i,j}|^2 \gamma^{\beta_1}_{i} \gamma^{\beta_2}_{j} i^{2\alpha_1} j^{2(k-\alpha_1)}. \end{aligned} \tag{2.27}$$

Similarly, we have

$$\begin{aligned} &\sum_{\alpha_1 \le i \le p, j \ge \max\{p+1,\alpha_2\}} |c_{i,j}|^2 \gamma^{\beta_1}_{i,\alpha_1} \gamma^{\beta_2}_{j,\alpha_2} \\ &\le C(p+1)^{-2(k-l)} \sum_{\alpha_1 \le i \le p, j \ge \max\{p+1,\alpha_2\}} |c_{i,j}|^2 \gamma^{\beta_1}_{i} \gamma^{\beta_2}_{j} \gamma^{\beta_2}_{j} i^{2\alpha_1} j^{2(k-\alpha_1)} \end{aligned} \tag{2.28}$$

and

$$\begin{aligned} &\sum_{i \ge \max\{p+1,\alpha_1\}, \alpha_2 \le j \le p} |c_{i,j}|^2 \gamma^{\beta_3,\beta_1}_{i,\alpha_1} \gamma^{\beta_4,\beta_2}_{j,\alpha_2} \\ &\le C(p+1)^{-2(k-l)} \sum_{i \ge \max\{p+1,\alpha_1\}, \alpha_2 \le j \le p} |c_{i,j}|^2 \gamma^{\beta_1}_{i} \gamma^{\beta_2}_{j} \gamma^{\beta_2}_{j} i^{2(k-\alpha_2)} j^{2\alpha_2}. \end{aligned} \tag{2.29}$$

A combination of (2.26)–(2.29) leads to (2.17).

We shall next prove (2.17). Since $k > 1, \beta_1, \beta_2 \le -1/2$, $u \in C^0(\bar{Q})$ [18], and due to (2.11) and (2.12), there holds

$$|(u-u_p)(x)| \le \left(\sum_{i,j\ge p+1} + \sum_{i\ge p+1, 0\le j\le p} + \sum_{0\le i\le p, j\ge p+1} \right) |c_{i,j}| \sqrt{\gamma_i^{\beta_1}\gamma_j^{\beta_2}}. \tag{2.30}$$

By the Schwartz inequality there holds

(2.31)

$$\begin{aligned} \sum_{i\ge p+1,0\le j\le p} |c_{i,j}| \sqrt{\gamma_i^{\beta_1}\gamma_j^{\beta_2}} &\le \left(\sum_{i\ge p+1,0\le j<p} i^{-2k} \right)^{\frac{1}{2}} \left(\sum_{i\ge p+1,0\le j\le p} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2k} \right)^{\frac{1}{2}} \\ &\le C(p+1)^{-(k-1)} \left(\sum_{i\ge p+1,0\le j\le p} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2k} \right)^{\frac{1}{2}}. \end{aligned}$$

Similarly, we can prove that

$$\sum_{0\le i\le p, j\ge p+1} |c_{i,j}| \sqrt{\gamma_i^{\beta_1}\gamma_j^{\beta_2}} \le C(p+1)^{-(k-1)} \left(\sum_{0\le i\le p, j\ge p+1} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} j^{2k} \right)^{\frac{1}{2}} \tag{2.32}$$

and

$$\sum_{i,j\ge p+1} |c_{i,j}| \sqrt{\gamma_i^{\beta_1}\gamma_j^{\beta_2}} \le C(p+1)^{-(k-1)} \left(\sum_{p+1\le i,j<\infty} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2k} \right)^{1/2}. \tag{2.33}$$

Combining (2.30)–(2.33), we obtain (2.17). The above arguments can be carried out for (2.18)–(2.20) except that we use (2.13) instead of (2.12).

For $p \ge k-1 \ge 0$, there holds

$$\begin{aligned} &\int_Q |D^\alpha (u-u_p)|^2 w_{\alpha,\beta}(x)dx \\ &\qquad = \left(\sum_{i,j\ge p+1} + \sum_{\alpha_1\le i\le p, j\ge p+1} + \sum_{i\ge p+1,\alpha_2\le j\le p} \right) |c_{i,j}|^2 \gamma_{i,\alpha_1}^{\beta_3,\beta_1} \gamma_{j,\alpha_2}^{\beta_4,\beta_2}. \end{aligned}$$

Note that

$$\sum_{i,j\ge p+1} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2\alpha_1} j^{2(k-\alpha_1)} \le \sum_{i,j\ge k} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2\alpha_1} j^{2(k-\alpha_1)} \le |u|^2_{H^{k,\beta}(Q)},$$

$$\begin{aligned} \sum_{\alpha_1\le i\le p, j\ge p+1} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2\alpha_1} j^{2(k-\alpha_1)} &\le \sum_{\alpha_1\le i\le p, j\ge k} |c_{i,j}|^2 \gamma_i^{\beta_1}\gamma_j^{\beta_2} i^{2\alpha_1} j^{2(k-\alpha_1)} \\ &\le |u|^2_{H^{k,\beta}(Q)}, \end{aligned}$$

$$\sum_{i\ge p+1,\alpha_2\le j\le p}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}i^{2(k-\alpha_2)}j^{2\alpha_2}\le\sum_{i\ge k,\alpha_2\le j\le p}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}i^{2(k-\alpha_2)}j^{2\alpha_2}\le|u|^2_{H^{k,\beta}(Q)},$$

$$\sum_{i\ge p+1,0\le j\le p}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}i^{2k}\le\sum_{i\ge k,0\le j\le p}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}i^{2k}\le|u|^2_{H^{k,\beta}(Q)},$$

$$\sum_{0\le i\le p,j\ge p+1}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}j^{2k}\le\sum_{0\le i\le p,j\ge k}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}j^{2k}\le|u|^2_{H^{k,\beta}(Q)},$$

$$\sum_{i,j\ge p+1}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}i^{2k}\le\sum_{i,j\ge k}|c_{i,j}|^2\gamma_i^{\beta_1}\gamma_j^{\beta_2}i^{2k}\le|u|^2_{H^{k,\beta}(Q)},$$

which lead to (2.21)–(2.25) immediately. ∎

Remark 2.2. The estimates in seminorms are derived in [5] for $p\ge k-1$. The estimates in Jacobi-weighted Sobolev norms and C^0 norms for $p\ge 0$ are available in this paper for the first time.

The argument of interpolation spaces yields the approximation results in the spaces $H^{s,\beta}(Q)$ and $B^{s,\beta}_\nu(Q)$, $\nu\ge 0$. We refer to [11] for the details of the proof for the spaces $H^{s,\beta}(Q)$ and $B^{s,\beta}(Q)$, and to [5] for the spaces $B^{s,\beta}_\nu(Q)$, $\nu>0$.

THEOREM 2.2. *Let $u\in H^{s,\beta}(Q)$ (resp., $B^{s,\beta}_\nu(Q)$), $s>0$, $\beta_i=(\beta_1,\beta_2)$, $\beta_i>-1$, $1\le i\le 2$, integer $\nu\ge 0$, and let u_p be the Jacobi projection of u on $\mathcal{P}_p(Q)$ with $p\ge 1$. Then for any integer $\ell<s$ there holds*

(2.34)
$$\|u-u_p\|_{H^{\ell,\beta}(Q)}\le C\,(p+1)^{-(s-\ell)}\,\|u\|_{H^{s,\beta}(Q)}\left(resp.,\ \tfrac{(1+\log(p+1))^\nu}{(p+1)^{s-\ell}}\|u\|_{B^{s,\beta}_\nu(Q)}\right).$$

Furthermore, if $u\in H^{s,\beta}(Q)$ (resp., $B^{s,\beta}_\nu(Q)$) with $s>1,\beta_i\le -1/2, 1\le i\le 2$, then for $x\in\bar Q$

(2.35) $$\|u-u_p\|_{C^0(\bar Q)}\le C(p+1)^{-(s-1)}\,\|u\|_{H^{s,\beta}(Q)}\left(resp.,\ \tfrac{(1+\log(p+1))^\nu}{(p+1)^{s-1}}\|u\|_{B^{s,\beta}_\nu(Q)}\right),$$

(2.36)
$$|(u-u_p)(\pm1,x_2)|\le C(p+1)^{-(s-3/2-\beta_1)}\,\|u\|_{H^{s,\beta}(Q)}\left(resp.,\ \tfrac{(1+\log(p+1))^\nu}{(p+1)^{s-3/2-\beta_1}}\|u\|_{B^{s,\beta}_\nu(Q)}\right),$$

(2.37)
$$|(u-u_p)(x_1,\pm1)|\le C(p+1)^{-(s-3/2-\beta_2)}\,\|u\|_{H^{s,\beta}(Q)}\left(resp.,\ \tfrac{(1+\log(p+1))^\nu}{(p+1)^{s-3/2-\beta_2}}\|u\|_{B^{s,\beta}_\nu(Q)}\right),$$

(2.38)
$$|(u-u_p)(\pm1,\pm1)|\le C(p+1)^{-(s-2-\beta_1-\beta_2)}\,\|u\|_{H^{s,\beta}(Q)}\left(resp.,\ \tfrac{(1+\log(p+1))^\nu}{(p+1)^{s-2-\beta_1-\beta_2}}\|u\|_{B^{s,\beta}_\nu(Q)}\right).$$

The constants C in the above inequalities are independent of p and u.

Remark 2.3. By the usual argument of interpolation spaces defined by the real method, e.g., the K-method, Theorems 2.1–2.2 stand for noninteger ℓ.

2.3. Approximability of singular functions on $Q = (-1,1)^2$. Let (r,θ) be the polar coordinates with respect to the vertex $(-1,-1)$, where $r = \{(x_1+1)^2 + (x_2+1)^2\}^{1/2}$, $\theta = \arctan(\frac{x_2+1}{x_1+1})$. For $\gamma > 0$ and integer $\nu \geq 0$

$$u(x) = r^\gamma \log^\nu r\, \chi(r)\, \Phi(\theta) \tag{2.39}$$

is a singular function defined on $Q = (-1,1)^2$, where $\chi(r)$ and $\Phi(\theta)$ are C^∞ functions such that for $0 < r_0 < 2$

$$\chi(r) = \begin{cases} 1 & \text{for} \quad 0 < r \leq \frac{r_0}{2}, \\ 0 & \text{for} \quad r \geq r_0. \end{cases} \tag{2.40}$$

Let $R_0 = R_{r_0,\theta_0}$ be a subregion of Q with $\theta_0 \in (0, \pi/4)$ and $r_0 \in (0,2)$:

$$R_{r_0,\theta_0} = \left\{ x \in Q \;\middle|\; r < r_0, \quad \theta_0 < \theta < \pi/2 - \theta_0 \right\}, \tag{2.41}$$

which is shown in Figure 2.1. For $x \in R_0$ we have $0 < 2 - r_0 < (1 - x_i) < 2$, and

$$\kappa_0 = \tan\theta_0 \leq \frac{1+x_2}{1+x_1} \leq \frac{1}{\kappa_0}.$$

Now we characterize the singularity of $u(x)$ in terms of the weighted Besov spaces $B^{s,\beta}_\nu(Q)$.

THEOREM 2.3 (see [5, Theorem 3.10]). *Let $u = r^\gamma \log^\nu r\, \chi(r)\, \Phi(\theta)$ be as given in (2.39) with $\gamma > 0$ and integer $\nu \geq 0$. Then $u \in B^{s,\beta}_{\nu^*}(Q)$ with $s = 2 + 2\gamma + \beta_1 + \beta_2$, $\beta_i > -1$, $i = 1,2$, and*

$$\nu^* = \begin{cases} \nu & \text{if } \gamma \text{ is not an integer or } \nu = 0, \\ \nu - 1 & \text{if } \gamma \text{ is an integer and } \nu \geq 1. \end{cases} \tag{2.42}$$

A combination of Theorems 2.3 and 2.1–2.2 leads to the approximabilities of the singular functions of $r^\gamma \log^\nu r$-type with $\gamma > 0$ and integer $\nu \geq 0$.

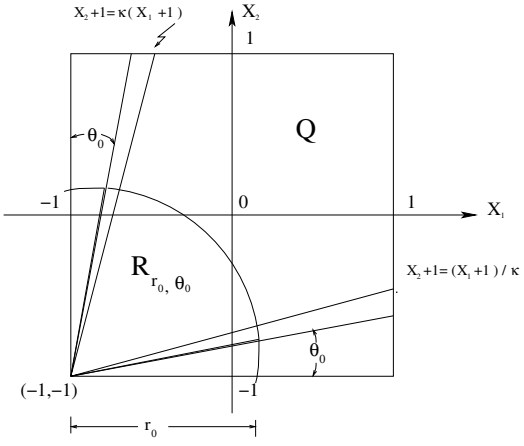


FIG. 2.1. *Square domain Q and subregion R_{ρ_0,θ_0}.*

THEOREM 2.4 (see [7, Theorem 3.8]). *Let $u(x)$ be as given in* (2.39) *with $\gamma > 0$ and integer $\nu \geq 0$, and let ψ and φ be the Jacobi projection of u on $\mathcal{P}_p(Q)$, $p \geq 1$, associated with $\beta = (0,0)$ and $\beta = (-1/2,-1/2)$, respectively. Then*

$$\|u-\psi\|_{L^2(Q)} \leq C\,(p+1)^{-2-2\gamma}(1+\log(1+p))^{\nu^*}\|v\|_{B^{2+2\gamma,\beta}_{\nu^*}(Q)} \tag{2.43}$$

with $\beta = (0,0)$, and

$$\|u-\phi\|_{H^1(R_0)} \leq C\,(p+1)^{-2\gamma}(1+\log(1+p))^{\nu^*}\|v\|_{B^{1+2\gamma,\beta}_{\nu^*}(Q)} \tag{2.44}$$

with $\beta = (-1/2,-1/2)$, where R_0 and ν^* are as given in (2.41) *and* (2.42).

THEOREM 2.5 (see [7, Theorems 2.9–2.11]). *Let $u(x)$ be as given in* (2.39) *with $\gamma > 0$ and $\nu \geq 1$.*

(i) *If $r^\gamma\,\Phi(\theta)$ is not a polynomial and $\nu = 0$, then*

$$\inf_{\varphi\in\mathcal{P}_p(Q)}\|u-\varphi\|_{H^1(R_0)} \geq C\,(p+1)^{-2\gamma}; \tag{2.45}$$

(ii) *If the integer $\nu \geq 1$, there holds*

$$\inf_{\varphi\in\mathcal{P}_p(Q)}\|v-\varphi\|_{H^1(R_0)} \geq C\,(p+1)^{-2\gamma}(1+\log(1+p))^{\nu^*}. \tag{2.46}$$

Here the constant C independent of p, R_0 and ν^* are as given in (2.41) *and* (2.42).

Remark 2.4. We exclude the trivial case that $\nu = 0$ and $r^\gamma\,\Phi(\theta)$ is a polynomial for which there is no approximation error. Throughout our paper, we assume that $r^\gamma\,\Phi(\theta)$ is not a polynomial while $\nu = 0$.

3. Approximation in Jacobi-weighted spaces on a scaled domain.

3.1. The Jacobi-weighted Besov and Sobolev spaces on $Q_h = (-h,h)^2$. For analyzing the approximation properties for smooth and singular functions on a scaled region we need to introduce the Jacobi-weighted Sobolev spaces $H^{k,\beta}(Q_h)$ and Besov spaces $B^{s,\beta}_\nu(Q_h)$ on a scaled square $Q_h = (-h,h)^2$.

Let $w^h_{\alpha,\beta}(x)$ be a weighted function on $Q_h = (-h,h)^2$:

$$w^h_{\alpha,\beta}(x) = \prod_{i=1}^{2}\left(\frac{h^2-x_i^2}{h^2}\right)^{\alpha_i+\beta_i} = \prod_{i=1}^{2}\left(1-\left(\frac{x_i}{h}\right)^2\right)^{\alpha_i+\beta_i},$$

where $\alpha = (\alpha_1,\alpha_2)$, $\alpha_i \geq 0$, integer, and $\beta = (\beta_1,\beta_2), \beta_i > -1$, real, $1 \leq i \leq 2$.

The Jacobi-weighted Sobolev space $H^{k,\beta}(Q_h)$, $k \geq 0$, is the closure of C^∞ functions furnished with the norm

$$\|u\|^2_{H^{k,\beta}(Q_h)} = \sum_{0\leq|\alpha|\leq k}\int_Q |D^\alpha u(x)|^2 w^h_{\alpha,\beta}(x)dx,$$

and $|u|_{H^{k,\beta}(Q_h)}$ denotes the seminorm involving only the kth derivatives.

The Jacobi-weighted Sobolev spaces $H^{s,\beta}(Q_h)$ and Besov spaces $B^{s,\beta}(Q_h)$ can be introduced as usual interpolation spaces by the K-method:

$$H^{s,\beta}(Q_h) = \Big(H^{\ell,\beta}(Q_h), H^{k,\beta}(Q_h)\Big)_{\theta,2}, \quad B^{s,\beta}(Q_h) = \Big(H^{\ell,\beta}(Q_h), H^{k,\beta}(Q)\Big)_{\theta,\infty}.$$

The space $B_\nu^{s,\beta}(Q_h)$ is an interpolation defined by the modified K-method:

$$B_\nu^{s,\beta}(Q_h) = \Big(H^{\ell,\beta}(Q_h), H^{k,\beta}(Q_h)\Big)_{\theta,\infty,\nu}.$$

Due to the definition of interpolation spaces and a simple scaling, the following proposition can be easily proved.

PROPOSITION 3.1. *Let $u(x)$ and $U(\xi) = u(h\xi)$ be functions defined on Q_h and Q, respectively.*

(i) *$u \in H^{k,\beta}(Q_h)$ with integer $k \geq 0$ if $U(\xi) = u(h\xi) \in H^{k,\beta}(Q)$, and vice versa. Furthermore, there holds for $l \leq k$*

$$|u|^2_{H^{\ell,\beta}(Q_h)} = h^{1-l}|U|_{H^{l,\beta}(Q)}. \tag{3.1}$$

(ii) *$u \in H^{s,\beta}(Q_h)$ with noninteger $s \geq 0$ if $U(\xi) \in H^{s,\beta}(Q)$, and vice versa.*

(iii) *$u \in B_\nu^{s,\beta}(Q_h)$ with real $s > 0$ and integer $\nu \geq 0$ if $U(\xi) \in B_\nu^{s,\beta}(Q)$, and vice versa.*

3.2. Approximation in Jacobi-weighted spaces on $Q_h = (-h, h)^2$. Let $\mathcal{P}_p(Q_h)$ be a set of polynomials of degree $\leq p$ on the scaled square Q_h, and let $\prod_{p,h}^\beta$ be the Jacobi projection operator on $\mathcal{P}_p(Q)$. Obviously, for $u \in H^{k,\beta}(Q_h)$ with $k \geq 0$, $u_{hp}(x) = \prod_{p,h}^\beta u$ is the Jacobi projection of $u \in H^{k,\beta}(Q_h)$ on $\mathcal{P}_p(Q_h)$ if and only if $U_p(\xi) = u_{hp}(h\xi)$ is the Jacobi projection of $U(\xi) = u(h\xi)$ on $\mathcal{P}_p(Q)$.

LEMMA 3.1. *Let $u \in H^{k,\beta}(Q_h), k \geq 0$. Then*

$$\|U - U_p\|_{H^{k,\beta}(Q)} \leq Ch^{\mu-1}\|u\|_{H^{k,\beta}(Q_h)}, \tag{3.2}$$

where $\mu = \min\{k, p+1\}$ and C is independent of p, h, k, and u.

Proof. For $k = 0$

$$\|U - U_p\|_{H^{0,\beta}(Q)} \leq \|U\|_{H^{0,\beta}(Q)} \leq h^{-1}\|u\|_{H^{0,\beta}(Q_h)}. \tag{3.3}$$

We now assume that the integer $k \geq 1$. Then we have by (2.21) of Theorem 2.1

$$\begin{aligned}\|U - U_p\|_{H^{k,\beta}(Q)} &\leq \|U - U_p\|_{H^{\mu,\beta}(Q)} + \sum_{m=\mu+1}^{k} (|U|_{H^{m,\beta}(Q)} + |U_p|_{H^{m,\beta}(Q)}) \\ &\leq C\left(|U|_{H^{\mu,\beta}(Q)} + \sum_{m=\mu+1}^{k} |U|_{H^{m,\beta}(Q)}\right).\end{aligned}$$

Here $\sum_{m=\mu+1}^{k} = 0$ if $\mu + 1 < k$. By the scaling argument (3.1), we obtain

$$\|U - U_p\|_{H^{k,\beta}(Q)} \leq C\sum_{m=\mu}^{k} h^{m-1}|u|_{H^{m,\beta}(Q_h)} \leq Ch^{\mu-1}\|u\|_{H^{k,\beta}(Q_h)},$$

which completes the proof. ∎

THEOREM 3.2. *Let $u \in H^{k,\beta}(Q_h)$ and let u_{ph} be its Jacobi projection on $\mathcal{P}_p(Q_h)$ with $p \geq 1$. Then, for $0 \leq l \leq k$,*

$$\|u - u_{hp}\|_{H^{l,\beta}(Q_h)} \leq C\frac{h^{\mu-l}}{(p+1)^{k-l}}\|u\|_{H^{k,\beta}(Q_h)} \tag{3.4}$$

with $\mu = \min\{k, p+1\}$. *Furthermore, if* $k > 1$, $\beta_\ell \le -1/2$, $1 \le \ell \le 2$, *then*

$$\|u - u_{hp}\|_{C^0(\bar{Q})} \le C \frac{h^{\mu-1}}{(p+1)^{k-1}} \|u\|_{H^{k,\beta}(Q_k)}. \tag{3.5}$$

The constants C *are independent of* $p, h, k,$ *and* u.

Proof. Let $\xi = \frac{x}{h}$ and $U(\xi) = u(h\xi)$. Then, due to Proposition 3.1, $U \in H^{k,\beta}(Q)$, and

$$\begin{aligned}\|U - U_p\|_{H^{l,\beta}(Q)} &= \|U - U_p - \textstyle\prod_p^\beta (U - U_p)\|_{H^{l,\beta}(Q)} \\ &\le C(p+1)^{-(k-l)} \|U - U_p\|_{H^{k,\beta}(Q)} \le C(p+1)^{-(k-l)} h^{\mu-1} \|u\|_{H^{k,\beta}(Q_h)},\end{aligned}$$

which together with the scaling argument (3.1) leads to

$$\|u - u_{hp}\|_{H^{l,\beta}(Q_h)} \le C h^{1-l} \|(U - U_p)\|_{H^{l,\beta}(Q)} \le C p^{-(k-l)} h^{\mu-l} \|u\|_{H^{k,\beta}(Q_h)}.$$

If $k > 1$ and $\beta_\ell \le -1/2, 1 \le \ell \le 2$, then there holds by Theorem 2.1 and Lemma 3.1

$$\begin{aligned}&|(u - u_{hp})(x)| = |U - U_p(\xi)| \le |(U - U_p - \textstyle\prod_p^\beta (U - U_p))(\xi)| \\ &\le C(p+1)^{-(k-1)} \|U - U_p\|_{H^{k,\beta}(Q)} \le C(p+1)^{-(k-1)} h^{\mu-1} \|u\|_{H^{k,\beta}(Q_h)}. \qquad \square\end{aligned}$$

By the argument of interpolation spaces, we have the approximation results in the spaces $H^{s,\beta}(Q_h)$ and $B_\nu^{s,\beta}(Q_h)$.

THEOREM 3.3. *Let* $u \in H^{s,\beta}(Q_h)$ $(resp., B_\nu^{s,\beta}(Q_h))$ *with* ≥ 0, *and let* u_{hp} *be the projection of* u *on* $\mathcal{P}_p(Q_h)$ *with* $p \ge 0$. *Then for* $0 \le l \le s$,

$$\begin{aligned}&\|u - u_{hp}\|_{H^{l,\beta}(Q_h)} \\ &\le C \frac{h^{\mu-l}}{(p+1)^{s-l}} \|u\|_{H^{s,\beta}(Q_h)} \left(resp., \frac{h^{\mu-1}}{(p+1)^{s-1}} \left(1 + \log \frac{p+1}{h}\right)^\nu B_\nu^{s,\beta}(Q_h) \right)\end{aligned} \tag{3.6}$$

with $\mu = \min\{s, p+1\}$. *Furthermore, if* $u \in H^{s,\beta}(Q_h)$ *with* $s > 1, \beta_\ell \le -1/2, 1 \le \ell \le 2$, *then for* $x \in \bar{Q}_h$

$$\begin{aligned}&|(u - u_{hp}(x)| \\ &\le C \frac{h^{\mu-l}}{(p+1)^{s-l}} \|u\|_{H^{s,\beta}(Q_k)} \left(resp., \frac{h^{\mu-1}}{(p+1)^{s-1}} \left(1 + \log \frac{p+1}{h}\right)^\nu B_\nu^{s,\beta}(Q_h) \right).\end{aligned} \tag{3.7}$$

The constants C *are independent of* $p, h,$ *and* u.

Proof. We will prove the theorem for $u \in B_\nu^{s,\beta}(Q_h)$. Let l and k be integers such that $0 \le l < s < k$ and $H^{s,\beta}(Q_h) = \left(H^{l,\beta}(Q_h), H^{k,\beta}(Q_h)\right)_{\theta,\infty,\nu}$ with $\theta = \frac{s-l}{k-l} \in (0,1)$. We have by Theorem 3.2

$$\|u - u_{hp}\|_{H^{l,\beta}(Q_h)} \le C h^{\mu_1 - l} \|u\|_{H^{l,\beta}(Q_h)} \tag{3.8}$$

with $\mu_1 = \min\{p+1, l\}$, and

$$\|u - u_{hp}\|_{H^{l,\beta}(Q_h)} \le C \frac{h^{\mu_2 - l}}{p^{k-l}} \|u\|_{H^{k,\beta}(Q_h)} \tag{3.9}$$

with $\mu_2 = \min\{p+1, k\}$. The weak exactness of θ-exponent (2.7) for the modified Jacobi-weighted Besov space $B^{s,\beta}_{\nu}(Q_h)$, together with (3.8) and (3.9), leads to

$$\begin{aligned}\|u-u_{hp}\|_{H^{l,\beta}(Q_h)} &\le C\frac{h^{(1-\theta)(\mu_1-l)+\theta(\mu_2-l)}}{(p+1)^{\theta(k-l)}}\left(1+\log\frac{(p+1)^{-(k-l)}h^{\mu_2-l}}{h^{\mu_1-l}}\right)^{\nu}\|u\|_{B^{s,\beta}_{\nu}(Q_h)}\\ &= C\frac{h^{\mu-l}}{(p+1)^{s-l}}\left(1+(k-l)\log(p+1)+(\mu_1-\mu_2)\log\frac{1}{h}\right)^{\nu}\|u\|_{B^{s,\beta}_{\nu}(Q_h)}\\ &\le C\frac{h^{\mu-l}}{(p+1)^{s-l}}\left(1+\log\frac{p+1}{h}\right)^{\nu}\|u\|_{B^{s,\beta}_{\nu}(Q_h)}.\end{aligned}$$

If $s>1, \beta_\ell \le -1/2, 1\le \ell\le 2$, select l and k such that $1<l<s<k$. Then by Theorem 3.2 there holds for $x\in\bar{Q}_h$

$$\|u-u_{hp}\|_{C^0(\bar{Q}_h)} \le Cp^{-(l-1)}h^{\mu_1-1}\|u\|_{H^{l,\beta}(Q_h)} \tag{3.10}$$

and

$$\|u-u_{hp}\|_{C^0(\bar{Q}_h)} \le Cp^{-(k-1)}h^{\mu_2-1}\|u\|_{H^{k,\beta}(Q_h)}. \tag{3.11}$$

The weak exactness of θ-exponent (2.7) together with (3.10)–(3.11) leads to

$$\begin{aligned}\|u-u_{hp}\|_{C^0(\bar{Q}_h)} &\le C\frac{h^{(1-\theta)(\mu_1-1)+\theta(\mu_2-1)}}{p^{\theta(k-1)+(1-\theta)(l-1)}}\left(1+\log\frac{(p+1)^{-(k-l)}h^{\mu_2-1}}{(p+1)^{-(l-1)}h^{\mu_1-1}}\right)^{\nu}\|u\|_{B^{s,\beta}_{\nu}(Q_h)}\\ &= C\frac{h^{\mu-l}}{(p+1)^{s-l}}\left(1+(k-l)\log(p+1)+(\mu_1-\mu_2)\log\frac{1}{h}\right)^{\nu}\|u\|_{B^{s,\beta}_{\nu}(Q_h)}\\ &\le C\frac{h^{\mu-l}}{(p+1)^{s-l}}\left(1+\log\frac{p+1}{h}\right)^{\nu}\|u\|_{B^{s,\beta}_{\nu}(Q_h)}.\end{aligned}$$

The proof for $u\in H^{s,\beta}(Q_h)$ is almost the same except that we apply (2.6) instead of (2.7). ∎

3.3. Approximation of singular functions on a scaled domain $Q_h = (-h,h)^2$. In this section we investigate the approximability of singular functions on a scaled square $Q_h=(-h,h)^2$:

$$u = r^{\gamma}\log^{\nu} r\chi_h(r)\,\Phi(\theta) \tag{3.12}$$

with $\gamma>0$ and integer $\nu\ge 0$, where $r=\sqrt{(x_1+h)^2+(x_2+h)^2}$, $\theta=\arctan\frac{h+x_2}{h+x_1}$, $\chi_h(r)=\chi(\frac{r}{h})$, and $\chi(\cdot)$ and $\Phi(\cdot)$ are C^∞ functions defined on the standard square as in the previous section.

Due to Proposition 3.1 and Theorem 2.3, a simple scaling leads us to the regularity of u in terms of the Jacobi-weighted Besov spaces.

THEOREM 3.4. *Let u be as given in* (3.12) *with $\gamma>0$ and $\nu\ge 0$. Then $u\in B^{s,\beta}_{\nu^*}(Q_h)$ with $s=2+2\gamma+\beta_1+\beta_2, \beta_l>-1, l=1,2$, and ν^* as given in* (2.42).

Proof. Let $\tilde{u}(\xi)=u(h\xi)$. Then for $\nu=0$

$$\tilde{u}(\xi)=u(h\xi)=h^{\gamma}\zeta^{\gamma}\chi(\zeta)\,\Phi(\theta)=h^{\gamma}w(\xi), \tag{3.13}$$

and for $\nu\ge 1$

$$\begin{aligned}\tilde{u}(\xi) &= h^{\gamma}\zeta^{\gamma}(\log h+\log\zeta)^{\nu}\chi(\zeta)\,\Phi(\theta)\\ &= h^{\gamma}\zeta^{\gamma}\,\chi(\zeta)\,\Phi(\theta)\sum_{m=0}^{\nu}\binom{\nu}{m}\log^{\nu-m}h\;\log^{m}\zeta = h^{\gamma}\sum_{m=0}^{\nu}\binom{\nu}{m}\tilde{v}_m(\xi)\log^{\nu-m}h,\end{aligned} \tag{3.14}$$

where $w(\xi) = \zeta^\gamma \chi(\zeta)\,\Phi(\theta), \tilde{v}_m(\xi) = \zeta^\gamma \chi(\zeta)\,\Phi(\theta)\log^m \zeta, \zeta = \sqrt{(\xi_1+1)^2+(\xi_2+1)^2}$. Due to Theorem 2.3, $w(\xi) \in B^{s,\beta}(Q)$ and $\tilde{v}_m(\xi) \in B^{s,\beta}_{m^*}(Q)$ with $s = 2+\gamma+\beta_1+\beta_2, \beta_\ell > -1, \ell = 1,2$, and

$$m^* = \begin{cases} m-1 & \text{if } \gamma \text{ is an integer and } m \geq 1, \\ m & \text{otherwise.} \end{cases} \tag{3.15}$$

The assertions of the theorem follow immediately from Theorem 2.3 and Proposition 3.1. ∎

By $R_0^h = R^h_{r_0,\theta_0}$ we denote a subregion of Q_h with $\theta_0 \in (0,\pi/4)$ and $r_0 \in (0,h)$,

$$R_0^h = R^h_{r_0,\theta_0} = \{x \in Q \mid r < r_0, \quad \theta_0 < \theta < \pi/2 - \theta_0\}. \tag{3.16}$$

A combination of Theorem 2.4 and a proper scaling gives a sharp estimation on the upper bound of approximation error in the Jacobi projections for the singular functions.

THEOREM 3.5. *Let $u(x)$ be as given in* (3.12). *Then there exist polynomials $\psi_{hp}(x)$ and $\varphi_{hp}(x)$ in $\mathcal{P}_p(Q)$, $p \geq 1$, such that*

$$\|u - \psi_{hp}\|_{L^2(Q_h)} \leq C\,\frac{h^{1+\gamma}}{(p+1)^{2(1+\gamma)}} F_\nu(p,h) \tag{3.17}$$

with $\beta = (0,0)$, and

$$\|u - \varphi_{hp}\|_{H^1(R_0^h)} \leq C\,\frac{h^{\gamma}}{(p+1)^{2\gamma}} F_\nu(p,h) \tag{3.18}$$

with $\beta = (-1/2,-1/2)$, where $F_\nu(p,h)$ is a log-polynomial,

$$F_\nu(p,h) = \begin{cases} (1+\log\frac{p+1}{h})^\nu & \text{for noninteger } \gamma, \\ (1+\log\frac{p+1}{h})^{\nu-1} & \text{for integer } \gamma \text{ and } r^\gamma\Phi(\theta) \in \mathcal{P}_\gamma(Q_h), \\ \max\Big\{ (1+\log\frac{p+1}{h})^{\nu-1},\ \log^\nu \frac{1}{h} \Big\} & \text{for integer } \gamma \text{ and } r^\gamma\Phi(\theta) \notin \mathcal{P}_\gamma(Q_h). \end{cases} \tag{3.19}$$

Furthermore, there holds

$$\|u - \varphi_{hp}\|_{C^0(\bar{Q}_h)} \leq C\,\frac{h^{\gamma}}{(p+1)^{2\gamma}} F_\nu(p,h). \tag{3.20}$$

The constants C in (3.17)–(3.20) *are independent of h and p.*

Proof. We shall concentrate on the proof of (3.18); the proofs for (3.17) and (3.20) are similar to what follows. By (3.13) for $\nu = 0$,

$$\tilde{u}(\xi) = u(h\xi) = h^\gamma \zeta^\gamma \chi(\zeta)\,\Phi(\theta) = h^\gamma w(\xi).$$

Then (3.17) and (3.18) with $\nu = 0$ follow from Theorem 2.4 immediately.

Due to (3.14) for $\nu \geq 1$,

$$\begin{aligned} \tilde{u}(\xi) &= h^\gamma \zeta^\gamma (\log h + \log\zeta)^\nu \chi(\zeta)\,\Phi(\theta) = h^\gamma \zeta^\gamma\,\chi(\zeta)\,\Phi(\theta) \sum_{m=0}^{\nu} \binom{\nu}{m} \log^{\nu-m} h\ \log^m \zeta \\ &= h^\gamma \sum_{m=0}^{\nu} \binom{\nu}{m} \tilde{v}_m(\xi) \log^{\nu-m} h. \end{aligned}$$

By Theorem 2.3, $\tilde{v}_m(\xi) \in B^{1+2\gamma,\beta}_{m^*}(Q)$ with $\beta = (-1/2,-1/2)$, and then due to Theorem 2.4, $\tilde{\varphi}_m(\xi) = \Pi^{\beta}_p \tilde{v}_m$ with $\beta = (-1/2,-1/2)$ satisfies

$$\|\tilde{v}_m(\xi) - \tilde{\varphi}_m(\xi)\|_{H^1(R_0)} \le C(p+1)^{-2\gamma}(1+\log(1+p))^{m^*}\|\tilde{v}_m\|_{B^{1+2\gamma,\beta}_{m^*}(Q)} \tag{3.21}$$

with m^* as given in (3.15).

If γ is not an integer, let $\tilde{\varphi}(\xi) = h^\gamma \sum_{m=0}^{\nu} \binom{\nu}{m} \tilde{\varphi}_m(\xi)\log^{\nu-m} h$, and let $\varphi(x) = \tilde{\varphi}(x/h) = \Pi^{\beta}_{p,h} u$ be associated with $\beta = (-1/2,-1/2)$. Then there hold

$$\begin{aligned}\|\tilde{u}(\xi) - \tilde{\varphi}(\xi)\|_{H^1(R_0)} &\le \frac{Ch^\gamma}{(p+1)^{2\gamma}} \sum_{m=0}^{\nu} \binom{\nu}{m} (1+\log(1+p))^m \log^{\nu-m}\frac{1}{h} \\ &\le C\frac{h^\gamma \left(1+\log\frac{p+1}{h}\right)^\nu}{(p+1)^{2\gamma}}\end{aligned}$$

and

$$\|u(x) - \varphi(x)\|_{H^1(R^h_0)} = \|\tilde{u}(\xi) - \tilde{\varphi}(\xi)\|_{H^1(R_0)} \le C\frac{h^\gamma}{(p+1)^{2\gamma}}\left(1+\log\frac{p+1}{h}\right)^\nu.$$

Thus (3.18) is proved for noninteger γ.

If γ is an integer, we have by (3.21)

$$\begin{aligned}\|\tilde{u}(\xi) - \tilde{\varphi}(\xi)\|_{H^1(R_0)} &\le \frac{Ch^\gamma}{(p+1)^{2\gamma}}\left(\log^\nu\frac{1}{h} + \sum_{m=1}^{\nu}\binom{\nu}{m}(1+\log(p+1))^{m-1}\log^{\nu-m}\frac{1}{h}\right) \\ &\le \frac{Ch^\gamma}{(p+1)^{2\gamma}}\max\left\{\left(1+\log\frac{p+1}{h}\right)^{\nu-1},\ \log^\nu\frac{1}{h}\right\},\end{aligned}$$

which implies (3.18) for integer γ.

If γ is an integer and $r^\gamma\Phi(\theta)$ is a polynomial of degree γ in Q_h, then $\zeta^\gamma\Phi(\theta)$ is a polynomial of degree γ in Q. We rewrite (3.14) as

$$\tilde{u}(\xi) = h^\gamma\left(\tilde{v}_0(\xi)\log^\nu h + \sum_{m=1}^{\nu}\binom{\nu}{m}\tilde{v}_m(\xi)\log^{\nu-m}h\right) = h^\gamma\Big(\tilde{v}_0(\xi)\log^\nu h + \tilde{w}(\xi)\Big).$$

By the arguments above, there exists polynomial $\tilde{\varphi}_w(\xi) \in \mathcal{P}_p(Q)$ such that

$$\begin{aligned}\|\tilde{w}(\xi) - \tilde{\varphi}_w(\xi)\|_{H^1(R_0)} &\le C\sum_{m=1}^{\nu}\binom{\nu}{m}\frac{(1+\log(p+1))^{(m-1)}}{(p+1)^{2\gamma}}\log^{\nu-m}\frac{1}{h} \\ &\le C\frac{\left(1+\log\frac{p+1}{h}\right)^{\nu-1}}{(p+1)^{2\gamma}}.\end{aligned}$$

Let $u_0(x) = \tilde{u}(\frac{x}{h}) = r^\gamma\chi_h(r)\Phi(\theta)\log^\nu h$ and $w(x) = h^\gamma\tilde{w}(\frac{x}{h})$. Then

$$u(x) = u_0(x) + w(x). \tag{3.22}$$

Since $u_0(x)$ is a C^∞ function, there exists a polynomial $\varphi_0(x) \in \mathcal{P}_p(Q_h)$ such that

$$\|u_0 - \varphi_0\|_{H^1(R^h_0)} \le C\left(\frac{h}{(p+1)^2}\right)^\gamma\left(1+\log\frac{p+1}{h}\right)^{\nu-1}. \tag{3.23}$$

Let $\varphi_{hp}(x) = \varphi_0(x) + \varphi_w(x)$ with $\varphi_w(x) = h^\gamma \tilde{\varphi}_w(\frac{x}{h})$. By (3.22)–(3.23), we have

$$\begin{aligned}\|v(x) - \varphi_{hp}(x)\|_{H^1(R_0^h)} &\le \|w(x) - \varphi_w(x)\|_{H^1(R_0^h)} + \|u_0 - \varphi_0\|_{H^1(R_0^h)} \\ &\le \frac{C\,h^\gamma}{p^{2\gamma}} \left(1 + \log\frac{p+1}{h}\right)^{\nu-1},\end{aligned}$$

which leads to the estimation (3.18) in the case that $r^\gamma \Phi(\theta)$ is a polynomial.

Arguments similar to the above can be carried out for (3.17) and (3.20), and we will not elaborate here. ∎

3.4. Asymptotic error analysis for Legendre projection for singular functions of $x^\gamma \log^\nu x$-type on a scaled interval. The estimation of the lower bounds of approximation error for the singular function v of $r^\gamma \log^\nu r$-type on Q_h is not trivial and has never been addressed in the literature because it is not a simple generalization of the approximation on standard square Q with a simple scaling. To this end we need asymptotic error analysis for singular functions of $x^\gamma \log^\nu x$-type on a scaled interval $(-h, h)$. The asymptotic error analysis of Legendre projection for singular functions of $x^\gamma \log^\nu x$-type on a standard interval $I = (-1, 1)$ was studied in [7]. We are here generalizing the result to a scaled interval $I_h = (-h, h)$ in terms of p and h.

Let

$$w_\nu(x) = (x + h)^\gamma \log^\nu(x + h), \quad x \in I_h = (-h, h), \tag{3.24}$$

with real $\gamma > 0$ and integer $\nu \ge 1$. Let $\mathcal{P}_p(I_h)$ be a set of polynomials of degree $\le p$ on I_h. We shall analyze the asymptotic of the approximation error of Legendre projection, which is essential to the sharpest estimates of the lower bounds of error in the finite element solutions of the h-p versions.

It was shown in [16] that on the standard interval $I = (-1, 1)$ the singular function $(1 + \xi)^\gamma$ has the Legendre expansion

$$(1 + \xi)^\gamma = \sum_{i=0}^{\infty} a_i(\gamma) L_i(\xi),$$

where $L_i(\xi)$ is the Legendre polynomial of degree i, and

$$a_i(\gamma) = (-1)^{i-1} \left(i + \frac{1}{2}\right) C_0(\gamma) \frac{\Gamma(i - \gamma)}{\Gamma(i + \gamma + 2)} \tag{3.25}$$

with

$$C_0(\gamma) = \frac{2^{1+\gamma}\Gamma^2(1 + \gamma) \sin \pi\gamma}{\pi}. \tag{3.26}$$

For the singular function $(1 + \xi)^\gamma \log^\nu(1 + \xi)$, it was proved in [7] that

$$(1 + \xi)^\gamma \log^\nu(1 + \xi) = \frac{d^\nu}{d\gamma^\nu}(1 + \xi)^\gamma = \sum_{i=0}^{\infty} b_i(\gamma) L_i(\xi),$$

where for $i > 0$

$$b_i(\gamma) = a_i^{(\nu)}(\gamma) = \frac{d^\nu}{d\gamma^\nu} a_i(\gamma) = \frac{(-1)^{i-1}}{i^{2\gamma+1}} \sum_{\ell=0}^{\nu} (-1)^{\nu-\ell} C_\ell(\gamma) \log^{\nu-\ell} i \ \left(1 + O\left(\frac{1}{i}\right)\right) \tag{3.27}$$

with $C_1(\gamma) = \nu C_0'(\gamma)$. By a scaling $x = h\xi$, we have

$$w_0(x) = (h+x)^\gamma = h^\gamma(1+\xi)^\gamma = \sum_{i=0}^{\infty} h^\gamma a_i(\gamma) L_i(\xi) = \tilde{w}_0(\xi)$$

and

$$w_\nu(x) = (h+x)^\gamma \log^\nu(h+x) = \frac{d^\nu}{d\gamma^\nu}(h+x)^\gamma = \sum_{i=0}^{\infty} b_i(\gamma,h) L_i(\xi) = \tilde{w}_\nu(\xi)$$

with

$$b_i(\gamma,h) = \frac{d^\nu}{d\gamma^\nu}\left(h^\gamma a_i(\gamma)\right) = h^\gamma \sum_{m=0}^{\nu} \left(\begin{array}{c} \nu \\ m \end{array} \right) a_i^{(m)}(\gamma)\, \log^{\nu-m} h. \tag{3.28}$$

Due to (3.27),

$$\begin{aligned} &a_i^{(m)}(\gamma) \\ &= (-1)^{i-1} \frac{(-1)^m C_0(\gamma) \ln^m i + (-1)^{m-1} C_1(\gamma) \ln^{m-1} i + \cdots + C_m(\gamma)}{i^{2\gamma+1}} \left(1 + O\left(\frac{1}{i}\right)\right) \end{aligned}$$

with $C_1(\gamma) = mC_0'(\gamma)$.

For noninteger $\gamma, C_0(\gamma) \neq 0$, and

$$\begin{aligned} b_i(\gamma,h) &= (-1)^{i-1+\nu} h^\gamma \sum_{m=0}^{\nu} \left(\begin{array}{c} \nu \\ m \end{array} \right) C_0(\gamma) \log^{\nu-m} \frac{1}{h}\ \log^m i \left(1 + O\left(\frac{1}{\log i}\right)\right) \\ &= \frac{(-1)^{i-1+\nu} h^\gamma C_0(\gamma)}{i^{2\gamma+1}} \log^\nu \frac{i}{h}. \end{aligned} \tag{3.29}$$

If γ is an integer, then $C_0(\gamma) = 0$ and $C_1(\gamma) = (-1)^\gamma(\gamma!)^2 \neq 0$. Hence

$$\begin{aligned} b_i(\gamma,h) &= \frac{(-1)^{i+\nu} h^\gamma}{i^{2\gamma+1}} \sum_{m=1}^{\nu} \left(\begin{array}{c} \nu \\ m \end{array} \right) (-1)^{\nu-m} C_1(\gamma) \log^{\nu-m} h \log^{m-1} i\ \left(1 + O\left(\frac{1}{\log i}\right)\right) \\ &= \frac{(-1)^{i+\nu} h^\gamma C_1(\gamma)}{i^{2\gamma+1}} \sum_{m'=0}^{\nu-1} \left(\begin{array}{c} \nu-1 \\ m' \end{array} \right) \frac{\nu}{m'+1} \log^{\nu-1-m'} \frac{1}{h} \log^{m'} i \left(1 + O\left(\frac{1}{\log i}\right)\right). \end{aligned}$$

Note that $1 \leq \frac{\nu}{m'+1} \leq \nu$ for $0 \leq m' \leq \nu - 1$; therefore there holds

$$\log^{\nu-1} \frac{i}{h} \leq \sum_{m'=0}^{\nu} \left(\begin{array}{c} \nu-1 \\ m' \end{array} \right) \frac{\nu}{m'+1} \log^{\nu-m} \frac{1}{h} \log^{m-1} i \leq \nu \log^{\nu-1} \frac{i}{h},$$

which implies that for integer γ

$$|C_1(\gamma)| \frac{h^\gamma}{i^{2\gamma+1}} \log^{\nu-1} \frac{i}{h} \leq |b_i(\gamma,h)| \leq \nu |C_1(\gamma)| \frac{h^\gamma}{i^{2\gamma+1}} \log^{\nu-1} \frac{i}{h}. \tag{3.30}$$

Let $\tilde{\varphi}_p^\nu(\xi) = \sum_{i=0}^p b_i(\gamma,h) L_i(\xi)$ and $\varphi_{hp}^\nu(x) = \sum_{i=0}^p b_i(\gamma,h) L_i(\frac{x}{h})$, which are the Legendre projection of $\tilde{w}_\nu(\xi)$ on $\mathcal{P}_p(I)$ and the Legendre projection of $w_\nu(x)$ on $\mathcal{P}_p(I_h)$, respectively. Then we have the following asymptotic error estimation.

THEOREM 3.6. *Let $w_\nu(x)$ be as given in* (3.24)*, and let $\varphi_{hp}^\nu(x)$ be its Legendre projection on $\mathcal{P}_p(I_h)$. There holds for noninteger $\gamma > 0$ and $\nu \geq 0$*

$$\|w_\nu(x) - \varphi_{hp}^\nu(x)\|_{L^2(I_h)} \cong \frac{h^{\gamma+1/2}}{(p+1)^{2\gamma+1}} \left(1 + \log \frac{p+1}{h}\right)^\nu, \tag{3.31}$$

and for integer $\gamma > 0$ and $\nu \geq 1$

$$\|w_\nu(x) - \varphi_{hp}^\nu(x)\|_{L^2(I_h)} \cong \frac{h^{\gamma+1/2}}{(p+1)^{2\gamma+1}} \left(1 + \log \frac{p+1}{h}\right)^{\nu-1}. \tag{3.32}$$

Proof. By a scaling argument we have

$$\|w_\nu(x) - \varphi_{hp}^\nu(x)\|_{L^2(I_h)} = h^{1/2} \|\tilde{w}_\nu(\xi) - \tilde{\varphi}_p^\nu(\xi)\|_{L^2(I)}.$$

It is sufficient to show that for noninteger γ

$$\|\tilde{w}_\nu(\xi) - \tilde{\varphi}_p^\nu(\xi)\|_{L^2(I)} \cong \frac{h^\gamma}{(p+1)^{2\gamma+1}} \log^\nu \frac{p+1}{h}, \tag{3.33}$$

and for integer $\gamma > 0$ and $\nu \geq 1$

$$\|\tilde{w}_\nu(\xi) - \tilde{\varphi}_p^\nu(\xi)\|_{L^2(I)} \cong \frac{h^\gamma}{(p+1)^{2\gamma+1}} \log^{\nu-1} \frac{p+1}{h}. \tag{3.34}$$

Because of the orthogonality of Legendre polynomials, there holds

$$\|\tilde{w}_\nu(\xi) - \tilde{\varphi}_p^\nu(\xi)\|_{L^2(I)}^2 = \sum_{i=p+1}^{\infty} |b_i(\gamma, h)|^2 \frac{2}{2i+1}.$$

Therefore it holds by (3.29) that for noninteger γ and integer $\nu \geq 1$,

$$\begin{aligned} \|\tilde{w}_\nu(\xi) - \tilde{\varphi}_p^\nu(\xi)\|_{L^2(I)}^2 &= |C_0(\gamma)|^2 h^{2\gamma} \sum_{i=p+1}^{\infty} i^{-(4\gamma+3)} \log^{2\nu} \frac{i}{h} \left(1 + O\left(\frac{1}{\log i}\right)\right) \\ &= \frac{|C_0(\gamma)|^2 h^{2\gamma}}{(4\gamma+2)p^{4\gamma+2}} \log^{2\nu} \frac{p+1}{h} \left(1 + O\left(\frac{1}{\log(p+1)}\right)\right). \end{aligned} \tag{3.35}$$

For noninteger γ and integer $\nu = 0$, $b_i(\gamma, h) = h^\gamma a_i(\gamma)$. By (3.25) we have

$$\begin{aligned} \|\tilde{w}_0(\xi) - \tilde{\varphi}_p^0(\xi)\|_{L^2(I)}^2 &= |C_0(\gamma)|^2 h^{2\gamma} \sum_{i=p+1}^{\infty} |a_i(\gamma)|^2 \frac{2}{2i+1} \\ &= \frac{|C_0(\gamma)|^2 h^{2\gamma}}{(4\gamma+2)(p+1)^{4\gamma+2}} \left(1 + O\left(\frac{1}{p+1}\right)\right), \end{aligned}$$

which together with (3.35) leads to (3.33) for noninteger γ and $\nu \geq 0$.

For integer γ and $\nu \geq 1$, (3.30) implies that

$$\begin{aligned} &\|\tilde{w}_\nu(\xi) - \tilde{\varphi}_p^\nu(\xi)\|_{L^2(I)}^2 \\ &\leq \nu |C_1(\gamma)|^2 h^{2\gamma} \sum_{i=p+1}^{\infty} i^{-(4\gamma+3)} \log^{2(\nu-1)} \frac{i}{h} \left(1 + O\left(\frac{1}{\log i}\right)\right) \\ &= \frac{\nu |C_1(\gamma)|^2 h^{2\gamma}}{(4\gamma+2)(p+1)^{4\gamma+2}} \log^{2(\nu-1)} \frac{p+1}{h} \left(1 + O\left(\frac{1}{\log(p+1)}\right)\right) \end{aligned} \tag{3.36}$$

and

(3.37)
$$\|\tilde{w}_\nu(\xi)-\tilde{\varphi}_p^\nu(\xi)\|^2_{L^2(I)}\geq\frac{|C_1(\gamma)|^2h^{2\gamma}}{(4\gamma+2)(p+1)^{4\gamma+2}}\log^{2(\nu-1)}\frac{p+1}{h}\ \left(1+O\left(\frac{1}{\log(p+1)}\right)\right).$$

Thus, (3.36)–(3.37) lead to (3.34) for integer γ and $\nu\geq 1$ and complete the proof of the theorem. ∎

3.5. Lower bound of approximation error for singular functions of $r^\gamma\log^\nu r$-type. We shall introduce the Jacobi-weighted Sobolev spaces $H^{k,\beta}(I_h)$ on scaled interval $I_h=(-h,h)$ and prove a lemma concerning error in the seminorm of $H^{k,\beta}(I_h)$, which will be used in deriving the lower bound of approximation error for singular functions of $r^\gamma\log^\nu r$-type.

The Jacobi-weighted Sobolev spaces $H^{k,\beta}(I_h), k\geq 0$, with $\beta>-1$ are furnished with weighted norm

$$\|u\|^2_{H^{k,\beta}(I_h)}=\sum_{\ell=0}^k\int_{I_h}|u^{(\ell)}(x)|^2\ w_{\ell,\beta}(x)\,dx,$$

where $w_{\ell,\beta}(x)=(1-\frac{x^2}{h^2})^{\ell+\beta}$. The seminorm $|u|_{H^{k,\beta}(I_h)}$ is involved in the kth derivative only. In the special case that $\beta=0$, the Jacobi weight $w_{\ell,0}(x)$ is called a Legendre weight, and the corresponding spaces are referred to as Legendre-weighted Sobolev spaces. For the Jacobi-weighted Sobolev spaces on standard interval $I=(-1,1)$ and the approximation theory in the framework of these spaces, including special cases $\beta=0$ (Legendre) and $\beta=-1/2$ (Chebyshev), we refer to [7, 19, 20].

LEMMA 3.7. *Let $w\in H^{1,\beta}(I_h)$ with $\beta=(0,0)$, and let w_p be its Legendre projection on $\mathcal{P}_p(I_h)$. Then there holds*

$$|w-w_p|_{H^{1,\beta}(I_h)}\geq\frac{p+1}{h}|w-w_p|_{L^2(I_h)}. \tag{3.38}$$

Proof. Let $x=h\xi$ for $\xi\in I=(-1,1)$, which introduces the functions $\tilde{w}(\xi)=w(h\xi)$ and $\tilde{w}_p(\xi)=w_p(h\xi)$ on I. By Lemma 2.1 of [7] and a scaling argument we have

$$\begin{aligned}\int_{I_h}|w'-w_p'|^2\left(1-\left(\frac{x}{h}\right)^2\right)dx&=\int_I|\tilde{w}'-\tilde{w}_p'|^2\frac{1-\xi^2}{h}d\xi\geq\frac{(p+1)^2}{h}\int_I|\tilde{w}-\tilde{w}_p|^2d\xi\\&=\frac{(p+1)^2}{h^2}\int_{I_h}|w-w_p|^2dx,\end{aligned}$$

which leads to (3.38). ∎

THEOREM 3.8. *Let $u(x)$ be as given in* (3.12) *with $\gamma>0$ and integer $\nu\geq 0$. Then*

$$\inf_{\phi\in\mathcal{P}_p(Q_h)}\|v-\phi\|_{H^1(R_0^h)}\geq C\frac{h^\gamma}{(p+1)^{2\gamma}}F_\nu(p,h) \tag{3.39}$$

with $F_\nu(p,h)$ as given in (3.19).

Proof. For $\nu=0$, $u(x)=h^\gamma w(\xi)$ with $w(\xi)=\zeta^\gamma\chi(\zeta)\,\Phi(\theta)$, where $\zeta=\{(\xi_1+1)^2+(\xi_2+1)^2\}^{1/2}$. Due to Theorem 2.5, there holds

$$\inf_{\phi\in\mathcal{P}_p(Q_h)}\|v-\phi\|_{H^1(R_0^h)}=h^\gamma\inf_{\tilde{\phi}\in\mathcal{P}_p(Q)}\|w-\tilde{\phi}\|_{H^1(R_0)}\geq Ch^\gamma(p+1)^{-2\gamma},$$

which proves (3.39) with $\nu = 0$. We assume that the integer $\nu \geq 1$ and further assume without loss of generality that $\Phi(\theta) \not\equiv 0$ and $\chi_h(r) \equiv 1$ for $0 \leq r \leq h$. There is an interval $[\theta_1, \theta_2]$ on which $|\Phi(\theta)| > \Phi_0 > 0$. For any $\varphi \in \mathcal{P}_p(Q_h)$

$$\begin{aligned}\int_{R_0^h} \left|\frac{\partial}{\partial r}(v-\varphi)\right|^2 dx &\geq \int_{\theta_1}^{\theta_2}\int_0^h \left|\frac{\partial}{\partial r}(r^\gamma \log^\nu r\Phi(\theta) - \varphi)\right|^2 r dr \\ &= \int_{\theta_1}^{\theta_2} |\Phi(\theta)|^2 \left(\int_0^h \left|\frac{\partial}{\partial r}(r^\gamma \log^\nu r - \Phi^{-1}(\theta)\varphi)\right|^2 r dr\right) d\theta.\end{aligned}$$

Since $\Phi^{-1}(\theta)\varphi(r,\theta)$ is a well-defined polynomial of degree p in variable r with θ as a parameter, by Lemma 3.7, we have that

$$\begin{aligned}&\int_0^h \left|\frac{\partial}{\partial r}(r^\gamma \log^\nu r - \Phi^{-1}(\theta)\varphi)\right|^2 r dr \geq h\int_0^h \left|\frac{\partial}{\partial r}(r^\gamma \log^\nu r - \Phi^{-1}(\theta)\varphi)\right|^2 \left(\frac{r}{h}\right)\left(\frac{h-r}{h}\right) dr \\ &\geq h\int_0^h \left|\frac{\partial}{\partial r}(r^\gamma \log^\nu r - \Phi^{-1}(\theta)\varphi))\right|^2 \left(\frac{r}{h}\right)\left(\frac{h-r}{h}\right) dr \geq \frac{(p+1)^2}{h}\|r^\gamma \log^\nu r - \psi_p^\nu\|^2_{L^2(I_h)},\end{aligned}$$

where ψ is the Legendre projection of $r^\gamma \log^\nu r$ on $\mathcal{P}_p(\tilde{I}_h), \tilde{I}_h = (0,h)$, which gives

$$\int_{Q_h} \left|\frac{\partial}{\partial r}(v-\varphi)\right|^2 dx \geq (\theta_2-\theta_1)\Phi_0\frac{(p+1)^2}{h}\|r^\gamma \log^\nu r - \psi\|^2_{L^2(\tilde{I}_h)}. \tag{3.40}$$

By Theorem 3.6, there holds

$$\|r^\gamma \log^\nu r - \psi\|^2_{L^2(\tilde{I}_h)} \geq C\frac{h^{\gamma+1/2}}{(p+1)^{2\gamma+1}}\left(1+\log\frac{p+1}{h}\right)^\nu \tag{3.41}$$

if γ is not an integer, and

$$\|r^\gamma \log^\nu r - \psi\|^2_{L^2(\tilde{I}_h)} \geq C\frac{h^{\gamma+1/2}}{(p+1)^{2\gamma+1}}\left(1+\log\frac{p+1}{h}\right)^{\nu-1} \tag{3.42}$$

if γ is an integer and $\nu \geq 1$. The combination of (3.40)–(3.42) leads to the third case of (3.39).

If $\log^\nu \frac{1}{h}$ is asymptotically larger than $\log^{\nu-1}\frac{p}{h}$, the lower bound is not sharp in the case that γ is an integer and $r^\gamma\Phi(\theta)$ is not a polynomial while comparing the upper bound. A sharper analysis is needed in that case. According to (3.22), $u(x)$ can be decomposed as

$$u(x) = u_0(x) + w(x),$$

where $u_0(x) = \tilde{v}_0(\frac{x}{h}) = r^\gamma\chi_h(r)\Phi(\theta)\log^\nu h$ and $w(x) = \tilde{w}(\frac{x}{h})$. It was shown that

$$\inf_{\varphi\in\mathcal{P}_p(Q_h)}\|w-\varphi\|_{H^1(R_0^h)} \geq C\left(\frac{h}{(p+1)^2}\right)^\gamma\left(1+\log\frac{p+1}{h}\right)^{\nu-1}.$$

Since we assume that $r^\gamma\Phi(\theta)$ is not a polynomial in variable x, it is a singular function. By Theorem 3.6,

$$\begin{aligned}\inf_{\varphi\in\mathcal{P}_p(Q_h)}\|u_0-\varphi\|_{H^1(R_0^h)} &= \log^\nu\frac{1}{h}\inf_{\varphi\in\mathcal{P}_p(Q_h)}\|r^\gamma\chi_h(r)\Phi(\theta)-\tilde{\varphi}\|_{H^1(R_0^h)} \\ &\geq C\frac{h^\gamma}{(p+1)^{2\gamma}}\log^\nu\frac{1}{h}.\end{aligned}$$

Therefore for any $\phi_1, \phi_2 \in \mathcal{P}_p(Q_h)$ and $\phi = \phi_1 + \phi_1$ there holds

$$\|u-\phi\|_{H^1(R_0^h)} \geq \|u_0-\phi_1\|_{H^1(R_0^h)} - \|w-\phi_2\|_{H^1(R_0^h)},$$

which implies

$$\begin{aligned}(3.43)\quad &\inf_{\varphi\in\mathcal{P}_p(Q_h)} \|u-\varphi\|_{H^1(R_0^h)} \\ &\geq \inf_{\varphi_1\in\mathcal{P}_p(Q_h)} \|u_0-\varphi_1\|_{H^1(R_0^h)} - \inf_{\varphi_2\in\mathcal{P}_p(Q_h)} \|w-\varphi_2\|_{H^1(R_0^h)} \\ &\geq C\frac{h^\gamma}{(p+1)^{2\gamma}}\log^\nu\frac{1}{h} - \tilde{C}\frac{h^\gamma}{(p+1)^{2\gamma}}\left(1+\log\frac{p+1}{h}\right)^{\nu-1} \geq C\frac{h^\gamma}{(p+1)^{2\gamma}}\log^\nu\frac{1}{h}.\end{aligned}$$

If $\log^{\nu-1}\frac{p}{h}$ is asymptotically larger than $\log^\nu\frac{1}{h}$, we can show similarly that

$$\begin{aligned}&\inf_{\varphi\in\mathcal{P}_p(Q_h)} \|u-\varphi\|_{H^1(R_0^h)} \geq \inf_{\varphi_2\in\mathcal{P}_p(Q_h)} \|w-\varphi_2\|_{H^1(R_0^h)} - \inf_{\varphi_1\in\mathcal{P}_p(Q_h)} \|u_0-\varphi_1\|_{H^1(R_0^h)} \\ &\geq C\frac{h^\gamma}{(p+1)^{2\gamma}}\left(1+\log\frac{p+1}{h}\right)^{\nu-1} - \tilde{C}\frac{h^\gamma}{(p+1)^{2\gamma}}\log^\nu\frac{1}{h} \\ &\geq C\frac{h^\gamma}{(p+1)^{2\gamma}}\left(1+\log\frac{p+1}{h}\right)^{\nu-1}.\end{aligned}$$

This with (3.43) gives the second case of (3.39) and completes the proof of the theorem. □

Remark 3.1. By the usual argument of interpolation spaces defined by the real method, e.g., the K-method, Theorems 3.2, 3.3, 3.5, and 3.8 stand for noninteger ℓ.

4. Optimal rate of convergence of the *h-p* version with quasi-uniform meshes. In this section we demonstrate how the optimal approximation results obtained in the previous section lead to optimal a priori upper and lower error estimates for the *h-p* version of the FEM with quasi-uniform meshes. We follow the notation and symbols introduced in [6] where we analyzed the performance of the *p*-version.

Let Ω be a polygon, shown in Figure 4.1, with vertices A_i, $1 \leq i \leq M$ ($A_{M+1} = A_1$), and (open) edges Γ_i connecting the vertices A_i and A_{i+1}. By ω_i we denote the internal angle between Γ_i and Γ_{i+1}. Let $\mathcal{D}$ be a subset of $\mathcal{B} = \{1, 2, \ldots, M\}$ and $\mathcal{N} = \mathcal{B} \setminus \mathcal{D}$. We refer to $\Gamma_{\mathrm{D}} = \cup_{i\in\mathcal{D}}\overline{\Gamma}_i$ as the Dirichlet boundary and to $\Gamma_{\mathrm{N}} = \cup_{i\in\mathcal{N}}\Gamma_i$ as the Neumann boundary. We also allow polygons with internal angle 2π, which is important in applications.

Consider the following boundary value problem:

$$\begin{aligned}&-\Delta u + u = f \qquad \text{in } \Omega, \\ &u|_{\Gamma_{\mathrm{D}}} = 0, \qquad \left.\frac{\partial u}{\partial n}\right|_{\Gamma_{\mathrm{N}}} = g.\end{aligned} \tag{4.1}$$

By this simple model problem we shall show how to derive the lower and upper bounds of the approximation error of the *h-p* version, which can be applied to general elliptic problems on nonsmooth domains. By $H^k(\Omega)$, $k \geq 0$, integer, we denote the usual Sobolev space and $H^1_{\mathrm{D}}(\Omega) = \{u \in H^1(\Omega) \mid u|_{\Gamma_{\mathrm{D}}} = 0\}$. The variational form of (4.1) is to seek $u(x) \in H^1_{\mathrm{D}}(\Omega)$ such that

$$B(u,v) = F(v) \qquad \forall\, v \in H^1_{\mathrm{D}}(\Omega), \tag{4.2}$$

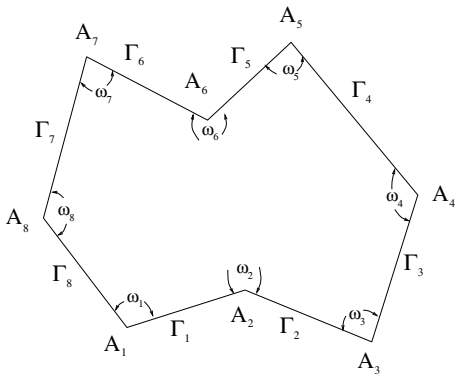


FIG. 4.1. *Polygonal domain* Ω.

where B is a bilinear form on $H^1_{\rm D}(\Omega)\times H^1_{\rm D}(\Omega)$ and F is a linear functional on $H^1_{\rm D}(\Omega)$, given by

$$B(u,v)=\int_\Omega (\nabla u\cdot\nabla v+u\,v)\,dx \tag{4.3}$$

and

$$F(v)=\int_\Omega f\,v\,dx+\int_{\Gamma_{\rm N}} g\,v\,ds. \tag{4.4}$$

4.1. The *h*-*p* version of the finite element method for problems with smooth solutions. Let $\Omega_h=\{\Omega_j,\ 1\le j\le J\}$ be a quasi-uniform mesh over the domain Ω. The elements Ω_i are shape-regular triangular and quadrilateral elements. We shall assume that $\overline{\Omega}_i\cap\overline{\Omega}_j$ is either the empty set, an entire side, or a vertex of Ω_i and Ω_j, and assume that all vertices of Ω are vertices of some Ω_i. By h_i we denote the size of element Ω_i; then there exists a constant C_q such that

$$1\le\frac{\max_i h_i}{\min_i h_i}\le C_q.$$

By $\mathcal{P}_p(\Omega)$ (or $\mathcal{P}_p(\Omega_i)$), we denote the space of all polynomials of degree $\le p$ defined on Ω (or Ω_i) and let $S^p(\Omega;\Delta_h)=\left\{u\mid u|_{\Omega_j}\in\mathcal{P}_p(\Omega_j),\ j=1,2,\ldots,J\right\}$ and $S^{p,1}_D(\Omega;\Delta_h)=S^p(\Omega;\Delta_h)\cap H^1_D(\Omega)$.

In practical applications, the finite element spaces $S^{p,1}_D(\Omega;\Delta_h;\mathcal{M})$ are used in computations. Here $\mathcal{M}=\{M_j,1\le j\le J\}$ denotes a mapping vector and M_j is an affine mapping of standard triangle T or square S onto Ω_j. Let $S^p(\Omega;\Delta_h;\mathcal{M})=\left\{\phi(x)\ |\phi\ |_{\Omega_j}=\tilde{\phi}_j\circ M_j^{-1},\tilde{\phi}_j\in\mathcal{P}_p(T)\text{ or }\mathcal{P}_p(S),j=1,2,\ldots,J\right\}$, where $\mathcal{P}_p(T)$ or $\mathcal{P}_p(S)$ is a set of polynomials of total or separate degree p on T or S, and let $S^{p,1}_D(\Omega;\Delta_h;\mathcal{M})=S^p(\Omega;\Delta_h;\mathcal{M})\cap H^1_D(\Omega)$ and $S^{p,1}(\Omega;\Delta_h;\mathcal{M})=S^p(\Omega;\Delta_h;\mathcal{M})\cap H^1(\Omega)$. The polynomials in the space $S^{p,1}_D(\Omega;\Delta_h;\mathcal{M})$ are called piecewise pull-back polynomials, and $\mathcal{P}_p(\Omega_i)$ denotes a set of pull-back polynomials of degree $\le p$ on Ω_i.

Obviously, if all elements are triangles or parallelograms, then $S^{p,1}_D(\Omega;\Delta_h)=S^{p,1}_D(\Omega;\Delta_h;\mathcal{M})$, and if all elements are triangles or quadrilaterals, then $S^{p,1}_D(\Omega;\Delta_h)\subset S^{p,1}_D(\Omega;\Delta_h;\mathcal{M})$. We shall establish the results in a general setting, i.e., in $S^{p,1}_D(\Omega;\Delta_h;\mathcal{M})$.

The h-p version finite element solution $u_{hp} \in S_D^{p,1}(\Omega;\Delta_h;\mathcal{M})$ is such that

$$B(u_{hp},v) = F(v) \qquad \forall\, v \in S_D^{p,1}(\Omega;\Delta_h;\mathcal{M}). \tag{4.5}$$

Due to the coercivity and continuity of the bilinear form (4.3), one can show that

$$\|u-u_{hp}\|_{H^1(\Omega)} \le C \inf_{w\in S_D^{p,1}(\Omega;\Delta;\mathcal{M})} \|u-w\|_{H^1(\Omega)}. \tag{4.6}$$

LEMMA 4.1. *Let γ_h be an edge of T_h which is a triangle or a parallelogram, and let ψ be a polynomial of degree p on γ_h vanishing at the end points of γ_h. Then there exists an extension $\Psi(x) \in \mathcal{P}_p(T_h)$ such that $\Psi(x)\,|_{\gamma_h} = \psi$ and vanishes on other edges of T_h, and*

$$\|\Psi\|_{H^1(T_h)} \le C\|\psi\|_{H_{00}^{1/2}(\gamma_h)} \tag{4.7}$$

with the constant C independent of h.

If T_h is a quadrilateral and M is a bilinear mapping of standard square $S = (-1,1)^2$ onto T_h, then the extension is a pull-back polynomial $\Psi = \tilde{\Psi}\circ M^{-1}$ with $\tilde{\Psi} \in \mathcal{P}_p(S)$ such that $\Psi\,|_\gamma = \psi$ and vanishes on other edges of T_h, and (4.7) holds. If T_h is a curved triangle or quadrilateral and $\tilde{\psi} = \psi\circ M$ is a polynomial of degree p on $\tilde{\gamma}_h = \gamma_h \circ M$, where M is a mapping of a standard triangle or square T onto T_h, then there exists an extension $\Psi = \tilde{\Psi}\circ M^{-1}$ with $\tilde{\Psi} \in \mathcal{P}_p(T)$ such that $\Psi\,|_\gamma = \psi$ and vanishes on other edges of T_h, and (4.7) holds.

Proof. Let M be the mapping of T onto T_h and $\gamma = (-1,1)$ onto γ_h, and let $\tilde{\psi} = \psi\circ M \in$. Note that

$$\|\tilde{\psi}\|^2_{L^2(\gamma)} \le \int_{-1}^{0} \frac{|\tilde{\psi}(\tau)|^2}{1+\tau}d\tau + \int_0^1 \frac{|\tilde{\psi}(\tau)|^2}{1-\tau}d\tau,$$

which implies

$$\|\tilde{\psi}\|^2_{H_{00}^{1/2}(\gamma)} \le C|\tilde{\psi}|^2_{H_{00}^{1/2}(\gamma)} = C\left(|\tilde{\psi}|^2_{H^{1/2}(\gamma)} + \int_{-1}^{0} \frac{|\tilde{\psi}(\tau)|^2}{1+\tau}d\tau + \int_0^1 \frac{|\tilde{\psi}(\tau)|^2}{1-\tau}d\tau\right).$$

Due to Theorems 7.4–7.5 of [10], there exists an extension $\tilde{\Psi} \in \mathcal{P}_p(T)$ such that

$$\|\tilde{\Psi}\|_{H^1(T)} \le C\|\tilde{\psi}\|_{H_{00}^{1/2}(\gamma)} \le C|\tilde{\psi}|_{H_{00}^{1/2}(\gamma)}.$$

Let $\Psi = \tilde{\Psi}\circ M^{-1}$. Then $\Psi \in \mathcal{P}_p(T_h)$, and by a simple scaling there holds

$$\|\Psi\|_{H^1(T_h)} \le C\|\tilde{\Psi}\|_{H^1(T)} \le C|\tilde{\psi}|_{H_{00}^{1/2}(\gamma)} \le C|\psi|_{H_{00}^{1/2}(\gamma_h)},$$

which leads to (4.7). ∎

LEMMA 4.2. *Let $u \in H^k(\Omega_i)$, $k>1$, where Ω_i is a curved triangular or quadrilateral element of the mesh Δ_h with size h. Then there exists a polynomial $\phi \in \mathcal{P}_p(\Omega_i)$ such that*

$$\|u-\phi\|_{H^1(\Omega_i)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)} \tag{4.8}$$

with $\mu = \min\{p+1,k\}$, and $u(V_l) = \phi(V_l), 1 \le l \le 3$ or 4, are the vertices of Ω_i.

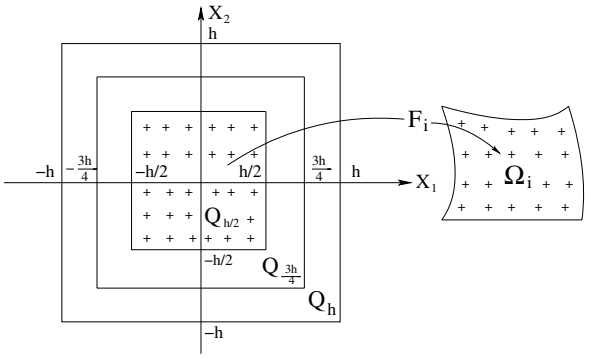


FIG. 4.2. *Mapping of quadrilateral.*

Proof. We assume first that Ω_i is a curved quadrilateral. Let F_i be a mapping of $Q_{h/2} = (-h/2, h/2)^2$ onto Ω_i. Then $\tilde{u} = u \circ F_i \in H^k(Q_{h/2})$, as shown in Figure 4.2, and it can be extended to Q_h such that the extended function has a support contained in $Q_{3h/4}$ and preserves the norm, i.e.,

$$\|\tilde{u}\|_{H^k(Q_h)} \le C\|\tilde{u}\|_{H^k(Q_{h/2})}\|u\|_{H^k(\Omega_i)}.$$

Since $\tilde{u}$ has a compact support in Q_h, $\tilde{u} \in H^{k,\beta}(Q_h)$ with the Jacobi weight $\beta = (-1/2, -1/2)$, and

$$\|\tilde{u}\|_{H^{k,\beta}(Q_h)} \le C\|\tilde{u}\|_{H^k(Q_h)} \le C\|u\|_{H^k(\Omega_i)}. \tag{4.9}$$

By Theorem 3.2, there exists a polynomial $\tilde{\phi} \in \mathcal{P}_p(Q_h)$, $p \ge 1$, such that

$$\|\tilde{u} - \tilde{\phi}\|_{H^{1,\beta}(Q_h)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|\tilde{u}\|_{H^{k,\beta}(Q_h)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)} \tag{4.10}$$

and for $1 \le l \le 4$

$$|\tilde{u}(\tilde{V}_l) - \tilde{\phi}(\tilde{V}_l)| \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)}, \tag{4.11}$$

where $\tilde{V}_l = \tilde{V}_l \circ M_i = (\pm h/2, \pm h/2)$. Let $g_1(x_1) = \frac{h-2x_1}{2h}$ and $g_2(x_1) = \frac{h+2x_1}{2h}$, and let

$$\begin{aligned}\bar{\phi} = \tilde{\phi} + g = \tilde{\phi} &+ (\tilde{u} - \tilde{\phi})(V_1)g_1(x_1)g_1(x_2) + (\tilde{u} - \tilde{\phi})(V_2)g_2(x_1)g_1(x_2)\\ &+ (\tilde{u} - \tilde{\phi})(V_3)g_2(x_1)g_2(x_2) + (\tilde{u} - \tilde{\phi})(V_4)g_1(x_1)g_2(x_2).\end{aligned} \tag{4.12}$$

It is trivial that

$$\|g_m\|_{H^t(\gamma_h)} \le Ch^{1/2-t},\ t = 0, 1, m = 1, 2, \quad \|g_l(x_1)g_m(x_2)\|_{H^1(\Omega)} \le C,\ l, m = 1, 2,$$

which together with (4.10) and (4.11) implies that $\tilde{u}(V_m) = \tilde{\phi}(V_m)$, $1 \le m \le 4$, and

$$\begin{aligned}\|\tilde{u} - \bar{\phi}\|_{H^1(Q_h)} &\le \|\tilde{u} - \tilde{\phi}\|_{H^1(Q_h)} + C\sum_{1\le m\le 4}|(\tilde{u} - \tilde{\phi})(V_m)|\\ &\le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)}.\end{aligned} \tag{4.13}$$

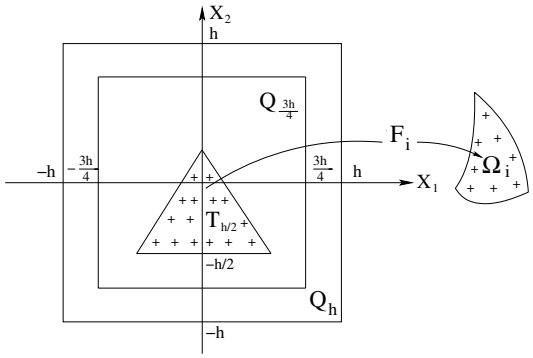


FIG. 4.3. *Mapping of triangle.*

Let $\phi = \bar{\phi} \circ M_i^{-1}$. Then $\phi \in \mathcal{P}_p(\Omega_i)$, $u(V_m) = \tilde{u}(V_m) = \bar{\phi}(V_m) = \phi(V_m)$, $1 \le m \le 4$, and

$$\|u - \phi\|_{H^1(\Omega_i)} \le C\|\tilde{u} - \bar{\phi}\|_{H^1(Q_h)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)}.$$

If Ω_i is a curved triangle, the mapping F_i maps $T_{h/2} = \{x = (x_1, x_2) \mid -\frac{h}{2} + \frac{x_2+h/2}{\sqrt{3}} \le x_1 \le \frac{h}{2} - \frac{x_2+h/2}{\sqrt{3}}, -\frac{h}{2} \le x_2 \le \frac{\sqrt{3}-1}{2}h\}$ onto Ω_i, and $(-h/2, -h/2), (h/2, -h/2)$, and $(0, \frac{\sqrt{3}-1}{2}h)$ are the vertices $\tilde{V}_m, 1 \le m \le 3$, of $T_{h/2}$, which are mapped to the vertices $V_m, 1 \le m \le 3$, of Ω_i, as shown in Figure 4.3. Then $\tilde{u} = u \circ F_i \in H^k(T_{h/2})$, and it can be extended to $H^k(Q_h)$ with a compact support contained in $H^k(Q_{3h/4})$. Then $\tilde{u} \in H^{k,\beta}(Q_h)$ with the Jacobi weight $\beta = (-1/2, -1/2)$, and (4.9) holds. By Theorem 3.2, there exists a polynomial $\tilde{\phi} \in \mathcal{P}_p(Q_h)$, $p \ge 1$, such that (4.10) and (4.11) hold.

Let

$$\bar{\phi} = \tilde{\phi} + g = \tilde{\phi} + \sum_{1 \le m \le 3} (u - \tilde{\phi})(\tilde{V}_m) g_m(x)$$

with

$$g_1(x) = \frac{1}{2} - \frac{x_1}{h} - \frac{x_2 + h/2}{\sqrt{3}}, \quad g_2(x) = \frac{1}{2} + \frac{x_1}{h} - \frac{x_2 + h/2}{\sqrt{3}}, \quad g_3(x) = \frac{2x_2}{\sqrt{3}h} + \frac{1}{\sqrt{3}}.$$

Obviously, $\bar{\phi}(\tilde{V}_m) = \tilde{u}(\tilde{V}_m), 1 \le m \le 3$, and (4.13) holds. Let $\phi = \bar{\phi} \circ M_i^{-1} \in \mathcal{P}_p(\Omega_i)$. Then $\phi(V_m) = u(V_m), 1 \le m \le 3$, and (4.8) holds. ∎

THEOREM 4.3. *Let $\Delta_h = \{\Omega_j,\ 1 \le j \le J\}$ be a quasi-uniform mesh with element size h over Ω containing triangular and quadrilateral elements, and let $S_D^{p,1}(\Omega; \Delta_h; \mathcal{M})$ be the finite element space defined as above. The data functions f and g are assumed such that the solution u of (4.1) is in $H^k(\Omega)$ with $k \ge 1$. Then the finite element solution $u_{hp} \in S_D^{p,1}(\Omega; \Delta_h; \mathcal{M})$ with $p \ge 0$ for the problem (4.1) satisfies*

$$\|u - u_{hp}\|_{H^1(\Omega)} \le C\,\frac{h^{\mu-1}}{(p+1)^{k-1}}\|u\|_{H^k(\Omega)}, \tag{4.14}$$

where $\mu = \min\{p+1, k\}$ and the constant C is independent of p and u.

Proof. We need to construct a polynomial $\varphi \in S_D^p(\Omega;\Delta_h;M)$ such that

$$\|u-\varphi\|_{H^1(\Omega)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega)}. \tag{4.15}$$

By Lemma 4.2, there exists a polynomial $\varphi^{[i]} \in \mathcal{P}_p(\Omega_i)$, $p \ge 1$, such that $\varphi^{[i]} = u$ at the vertices of Ω_i and

$$\|u-\varphi^{[i]}\|_{H^1(\Omega_i)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)}. \tag{4.16}$$

Let $\gamma = \bar\Omega_i \cap \bar\Omega_j$ be a common edge shared by the elements Ω_i and Ω_j, and let $\psi_{ij} = (\varphi^{[i]} - \varphi^{[j]})\,|_\gamma \in \mathcal{P}_p(\gamma)$. Note that $\varphi^{[i]}(V_m) = \varphi^{[j]}(V_m) = u(V_m), m = 1,2$, where V_m denote the end points of γ. Therefore ψ_{ij} vanishes at the end points of γ. By Lemma 4.1 there exists an extension $\Psi \in \mathcal{P}_p(\Omega_i)$ such that $\Psi_{ij}\,|_\gamma = \psi_{ij}$ and vanishes on $\partial\Omega_i\backslash\gamma$, and

$$\|\Psi_{ij}\|_{H^1(\Omega_i)} \le C\,\|\psi_{ij}\|_{H_{00}^{1/2}(\gamma)} \le C\Big(\|u-\varphi^{[i]}\|_{H_{00}^{1/2}(\gamma)} + \|u-\varphi^{[j]}\|_{H_{00}^{1/2}(\gamma)}\Big) \tag{4.17}$$

with the constant C independent of h. If $k \ge \frac{3}{2}$, by Theorem 3.2 and Remark 3.1, there holds for $t = 0, 1$

$$\|u-\varphi^{[i]}\|_{H^t(\gamma)} \le C\,\|u-\varphi^{[i]}\|_{H^{t+1/2}(\Omega_i)} \le C\,\frac{h^{\mu-t-1/2}}{p^{k-t-1/2}}\|u\|_{H^k(\Omega_i)}.$$

Note that $\psi_{ij} \in H_0^1(\gamma)$. Since $H_{00}^{1/2}(\gamma) = \big(L^2(\gamma),\, H_0^1(\gamma)\big)_{\frac{1}{2},2}$, by the standard argument of exact interpolation spaces of θ-exponent [11], it follows that

$$\|u-\varphi^{[i]}\|_{H_{00}^{1/2}(\gamma)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_i)}$$

and

$$\|u-\varphi^{[j]}\|_{H_{00}^{1/2}(\gamma)} \le C\,\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_j)},$$

which imply

$$\|\Psi_{ij}\|_{H^1(\Omega_i)} \le C\,\|\psi_{ij}\|_{H_{00}^{1/2}(\gamma)} \le C\frac{h^{\mu-1}}{p^{k-1}}\Big(\|u\|_{H^k(\Omega_i)} + \|u\|_{H^k(\Omega_j)}\Big). \tag{4.18}$$

Let $\tilde\varphi_p^{[i]} = \varphi_p^{[i]} - \Psi$ on Ω_i and $\tilde\varphi_p^{[j]} = \varphi_p^{[j]}$. Then $\tilde\varphi_p^{[i]}\,|_\gamma = \tilde\varphi_p^{[j]}\,|_\gamma = \varphi_p^{[j]}\,|_\gamma$, and there hold

$$\|u-\tilde\varphi_p^{[j]}\|_{H^1(\Omega_j)} = \|u-\varphi_p^{[j]}\|_{H^1(\Omega_j)} \le C\frac{h^{\mu-1}}{p^{k-1}}\|u\|_{H^k(\Omega_j)}$$

and

$$\begin{aligned}\|u-\tilde\varphi_p^{[i]}\|_{H^1(\Omega_i)} &\le C\,\Big(\|u-\varphi_p^{[i]}\|_{H^1(\Omega_i)} + \|\Psi\|_{H^1(\Omega_i)}\Big)\\ &\le C\frac{h^{\mu-1}}{p^{k-1}}\Big(\|u\|_{H^k(\Omega_i)} + \|u\|_{H^k(\Omega_j)}\Big).\end{aligned} \tag{4.19}$$

Adjusting $\varphi^{[i]}$ and $\varphi^{[j]}$ by Ψ_{ij} on each internal edge $\gamma = \bar{\Omega}_i \cap \bar{\Omega}_j$, we achieve the continuity across γ. $\varphi^{[i]}$ can be adjusted in the same way to satisfy the homogeneous Dirichlet boundary condition on each edge $\gamma \subset \Gamma^D$. Let $\varphi = \tilde{\varphi}^{[i]}$ in $\Omega_i, 1 \le i \le M$. Then $\varphi \in S_D^{p,1}(\Omega; \Delta_h; \mathcal{M})$ and satisfies (4.15).

So far we have proved the theorem for $k \ge \frac{3}{2}$. If $1 < k < \frac{3}{2}, k = (1-\theta) + 2\theta = 1 + \theta$ with $\theta = k - 1 \in (0, 1)$. Then $H_D^k(\Omega) = (H_D^1(\Omega), H^2(\Omega) \cap H_D^1(\Omega))_{\theta,2} = (H^1(\Omega), H^2(\Omega))_{\theta,2} \cap H_D^1(\Omega)$. It was shown in [8] that

$$H_D^k(\Omega) \subset \big(H^1(\Omega), H^2(\Omega)\big)_{\theta,\infty} \cap H_D^1(\Omega) = \mathcal{B}_{2,\infty}^k(\Omega) \cap H_D^1(\Omega) = B^k(\Omega) \cap H_D^1(\Omega).$$

Suppose that $v \in H_D^1(\Omega)$ and $w \in H^2(\Omega) \cap H_D^1(\Omega)$ form a decomposition of $u \in H_D^k(\Omega)$ such that $u = v + w$. Applying (4.18) for $k = 2$, we have a polynomial $\tilde{\varphi}_p^{[i]} \in S_D^{p,1}(\Omega; \Delta_h; \mathcal{M})$ with $p \ge 1$ such that

$$\|w - \varphi_p\|_{H^1(\Omega)} \le C\, \frac{h}{p} \|w\|_{H^2(\Omega)}.$$

Therefore, we have

$$\begin{aligned} \|u - \varphi\|_{H^1(\Omega)} &\le \|v\|_{H^1(\Omega)} + \|w - \varphi_p\|_{H^1(\Omega)} \le C \left(\|v\|_{H^1(\Omega)} + \frac{h}{p} \|w\|_{H^2(\Omega)} \right) \\ &= C \left(\|v\|_{H^1(\Omega)} + t \|w\|_{H^2(\Omega)} \right) \end{aligned}$$

with $t = \frac{h}{p}$. Due to the definition of the Besov spaces $B^k(\Omega) = \mathcal{B}_{2,\infty}^k(\Omega)$, we have

$$\|u - \tilde{\varphi}\|_{H^1(\Omega)} \le t^\theta \, \|u\|_{B^k(\Omega)} \le C \left(\frac{h}{p} \right)^{k-1} \|u\|_{H^k(\Omega)}, \tag{4.20}$$

which is (4.15) for $p \ge 1$, $1 < k < 3/2$, $\mu = \min\{k, p+1\} = k$. ∎

Remark 4.1. The theorem stands for $u \in B^s(\Omega)$ with $s > 1$ by a typical argument of interpolation spaces as we argued for $u \in H^k(\Omega)$ with $1 < k < 3/2$.

Remark 4.2. The convergence of the *h-p* version of FEM with quasi-uniform mesh containing triangular and parallel elements was proved in [10] for problems with smooth solutions. Lemma 4.1 and Theorem 4.3 generalize it to quadrilateral elements and curved elements, and the analysis is conducted in the framework of the Jacobi-weighted Sobolev spaces, which simplifies the proof and make it more robust. It proves that the Jacobi-weighted spaces not only work perfectly for the problems with singular solutions, but also work very well for problems with smooth solutions.

4.2. The *h-p* version finite element method for problems with singular solutions. Let $S_{\delta_i} = \big\{x \in \Omega \mid \operatorname{dist}(x, A_i) < \delta_i\big\}$ be a neighborhood of the vertices A_i, shown in Figure 4.4, with $\delta_i \in (0, 1)$. δ_i is selected such that $S_{\delta_i} \cap S_{\delta_j} = \emptyset$ for $i \ne j$. $\Omega_0 = \Omega \backslash \bigcup_{i \in \mathcal{M}} S_{\delta_i/2}$ contains no vertices of Ω, and $\Omega_0 \cap S_{\delta_i} \ne \emptyset$ for $i \in \mathcal{M}$. Ω_0 is called the regular part of Ω.

We assume that f and g are such that the solution u of (3.1) is in $H^k(\Omega_0), k \ge 1$, and in each neighborhood S_{δ_i}, u has an expansion in terms of singular functions of $r^\gamma \log^\nu r$-type:

$$u = \sum_{m \ge 1,\, 0 < \gamma_m^{[i]} \le k-1} \sum_{l=1}^{L_m^{[i]}} C_m^{[i]} \, r_i^{\gamma_m^{[i]}} \log^{\nu_{l,m}^{[i]}} r_i \Phi_m^{[i]}(\theta_i) \, \chi(r_i) + u_0^{[i]} = v + u_0^{[i]}, \tag{4.21}$$

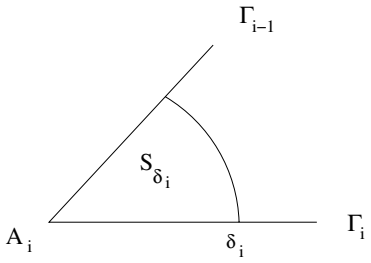


FIG. 4.4. *A neighborhood of the vertex A_i.*

where (r_i, θ_i) are polar coordinates with respect to the vertex A_i, $u_0^{[i]} \in H^k(S_{\delta_i})$ is the smooth part of u, $\gamma_m^{[i]} > 0$ are real, and $\nu_m^{[i]} \geq 0$ are integers. We assume that $\nu_{l,m}^{[i]} > \nu_{l+1,m}^{[i]}$ and $\gamma_m^{[i]} < \gamma_{m+1}^{[i]}$, $\chi(r_i)$ and $\Phi_m^{[i]}(\theta_i)$ are C^∞ functions, $\chi(r_i) = 1$ for $0 < r_i < \delta_i < \frac{1}{2}$, and $\chi(r_i) = 0$ for $r_i > \delta_i$. Let

$$\gamma = \min_i \gamma_1^{[i]}, \qquad \nu_\gamma = \max_{i, \gamma_1^{[i]} = \gamma} \nu_1^{[i]}. \tag{4.22}$$

There exists i_0 such that $\gamma_1^{[i_0]} = \gamma$ and $\nu_\gamma = \nu_1^{[i_0]}$. We will analyze the asymptotic rate of convergence of the h-p version of the FEM for problems with singularities on polygonal domains.

THEOREM 4.4. *Let $\Omega_h = \{\Omega_j,\ 1 \leq j \leq J\}$ be a quasi-uniform mesh over Ω containing triangular and parallelogram elements, and let $S_D^p(\Omega; \Delta_h; \mathcal{M})$ with $p > \gamma$ be the finite element space defined as above. The data functions f and g are assumed such that the solution u of* (4.1) *is in $H^k(\Omega_0)$ with $k > 1 + 2\gamma$, and u has the expansion* (4.21) *with $u_0^{[i]} \in H^k(S_{\delta_i})$ in each neighborhood S_{δ_i}. Then the finite element solution $u_{hp} \in S_D^{p,1}(\Omega; \Delta; \mathcal{M})$ for the problem* (4.1) *satisfies*

$$\|u - u_{hp}\|_{H^1(\Omega)} \leq C_1 \frac{h^\gamma}{p^{2\gamma}}\, F_{\nu_\gamma}(p, h) \tag{4.23}$$

with the constant C_1 depending on u, γ, and ν_γ, but not on p and h, where γ and ν_γ are as given in (4.22) *and $F_{\nu_\gamma}(p, h)$ is as given in* (3.19).

Proof. Due to (4.6), it suffices to construct a polynomial $\varphi \in S_D^p(\Omega; \Delta_h; \mathcal{M})$ with $p > \gamma$ such that

$$\|u - \varphi\|_{H^1(\Omega)} \leq C\, \frac{h^\gamma}{p^{2\gamma}}\, F_{\nu_\gamma}(p, h). \tag{4.24}$$

For elements Ω_i containing no vertices, by Lemma 4.2 there exists a polynomial $\varphi^{[i]} \in \mathcal{P}_p(\Omega_i)$ such that $\varphi^{[i]} = u$ at the vertices of Ω_i, and

$$\|u - \varphi^{[i]}\|_{H^1(\Omega_i)} \leq C\, \frac{h^{\tilde{\mu}-1}}{p^{k-1}} \leq C\, \frac{h^\gamma}{p^{2\gamma}} \|u\|_{H^k(\Omega_i)}$$

with $\tilde{\mu} = \min\{p+1, k\} \geq 1 + \gamma$.

Let the element Ω_j contain a vertex A_1 of Ω. Then (4.21) holds with $i = 1$ in S_{δ_1}. By Lemma 4.2, there exists a polynomial $\psi_0 \in \mathcal{P}_p(\Omega_j)$ such that $\psi_0 = u$ at the vertices of Ω_j, and

$$\|u_0 - \psi_0\|_{H^1(\Omega_j)} \leq C\, \frac{h^{\mu-1}}{p^{k-1}}$$

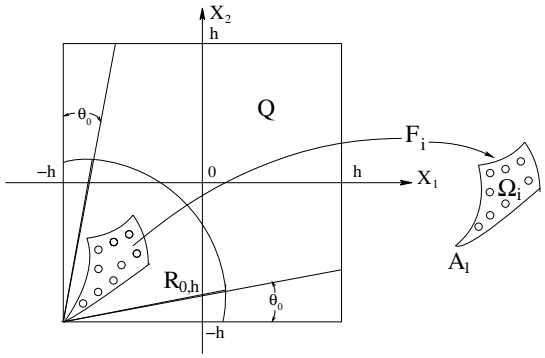


FIG. 4.5. *Mapping of element with a vertex of* Ω.

with $\mu = \min\{p+1, k\} \geq 1+\gamma$. For a sharp approximation to u_1, we map Ω_j into $R_0^h \subset Q_h$ by an affine mapping F_j such that $A_1 \circ F_j = (-h, -h)$ and Ω_j is contained in R_0^h, as shown in Figure 4.5. Without loss of generality we may assume that $A_1 = (-h, -h)$ and $\Omega_j \subseteq R_0^h$. Due to Theorem 3.5, there exists a polynomial $\psi_m \in \mathcal{P}_p(\Omega_j)$ such that $v_m = \psi_m$ at the vertices of Ω_j, and

$$\|v_m - \psi_m\|_{H^1(\Omega_j)} \leq C \, \frac{h^{\gamma^{[1]}}}{p^{2\gamma_m^{[1]}}} \, F_{\nu_{1,m}^{[1]}}(p,h),$$

where $v_m = \sum_{l=1}^{L_m^{[i]}} r_i^{\gamma_m^{[i]}} \log^{\nu_{l,m}^{[i]}} r_i \Phi_m^{[i]}(\theta_i)\, \chi(r_i)$. Let $\psi = \sum_{0<\gamma_m^{[1]} \leq k-1} C_m^{[1]} \psi_m$ and $\varphi_j = \psi + \psi_0$. Then $v = \psi$ at the vertices of Ω_j, and

$$\|v - \psi\|_{H^1(\Omega_j)} \leq C \sum_{0<\gamma_m^{[i]} \leq k-1} \frac{h^{\gamma^{[1]}}}{p^{2\gamma_m^{[1]}}} \, F_{\nu_{1,m}^{[1]}}(p,h) \leq C \frac{h^\gamma}{p^{2\gamma}} \, F_{\nu_\gamma}(p,h),$$

which implies that $u = \varphi^{[j]}$ at the vertices of Ω_j and that

$$\|u - \varphi^{[j]}\|_{H^1(\Omega_j)} \leq C \left(\frac{h^\gamma}{p^{2\gamma}} \, F_{\nu_\gamma}(p,h) + \frac{h^{\tilde{\mu}-1}}{p^{k-1}} \right) \leq C \frac{h^\gamma}{p^{2\gamma}} \, F_{\nu_\gamma}(p,h).$$

Noting that $u \in H^{1+\gamma-\epsilon}(\Omega)$ with $\epsilon > 0$ arbitrary, we can adjust these $\varphi^{[i]}$ as in the proof of Theorem 4.3 to achieve the continuity across internal edges γ of elements and the homogeneous Dirichlet boundary condition on the edges $\gamma \subset \Gamma_D$. Let $\varphi = \varphi^{[i]}$ on each $\Omega_i, 1 \leq i \leq J$; then $\varphi_p \in S_D^p(\Omega; \Delta; \mathcal{M})$ and satisfies (4.24). ∎

Remark 4.3. The convergence of the *h-p* version of the FEM for problems with singular solutions on polygonal domains was derived in [10], which is sharp for $\nu_\gamma = 0$. It is worth indicating that approximation properties of Chebyshev projection were utilized. Unfortunately, the Jacobi-weighted spaces were not introduced then to precisely describe the singularity and to fully explore approximability of singular solutions. Therefore, it was impossible to have the estimation (4.23) in the 1980s.

THEOREM 4.5. *Let* u *be the solution of the problem* (4.1) *such that* $u \in H^k(\Omega_0)$ *and* $u_0^{[i]} \in H^k(S_{\delta_i})$ *with* $k > 1 + 2\gamma$ *in each neighborhood* S_{δ_i}, *and let* u_{hp} *be its finite element solution in* $S_D^p(\Omega; \Delta_h; \mathcal{M})$ *with* $p > \gamma$. *Then*

$$\|u - u_{hp}\|_{H^1(\Omega)} \geq C_2 \left(\frac{h}{p^2} \right)^\gamma F_{\nu_\gamma}(p,h) \tag{4.25}$$

with constant C_2 depending on u, k, γ, and ν_γ, but not on p and h, where γ and ν_γ are as given in (4.22)*, and $F_{\nu_\gamma}(p,h)$ is as given in* (3.19)*.*

Proof. We may assume without loss of generality that the element Ω_1 contains the vertex $A_1 = (0,0)$ where the strongest singularity occurs, i.e., $\gamma = \gamma^{[1]}, \nu_\gamma = \nu_1^{[1]}$. According to (4.21),

$$u = u_1 + \sum_{m\geq 2, 0<\gamma_m\leq k-1} u_m + \sum_{l=2}^{L_m} v_l + u_0 \tag{4.26}$$

with

$$u_1 = c_{1,1}\, r^\gamma \log^{\nu_\gamma} r\Phi_1(\theta)\, \chi(r), \tag{4.27}$$

$$v_l = c_{l,1}\, r^\gamma \log^{\nu_{l,1}} r\Phi_1(\theta)\, \chi(r),\ l \geq 2, \tag{4.28}$$

$$u_m = \sum_{l=1}^{L_m} c_{l,m}\, r^{\gamma_m} \log^{\nu_{l,m}} r\, \Phi_m(\theta)\chi(r),\ m \geq 2. \tag{4.29}$$

Here we omit the index [1].

We now assume that the assertion of the theorem does not hold. Therefore, there exists a function $\delta(p,h)$ such that

$$\|u - u_{hp}\|_{H^1(\Omega_1)} \leq \|u - u_{hp}\|_{H^1(\Omega)} \leq C\left(\frac{h}{p^2}\right)^\gamma F_{\nu_\gamma}(p,h)\delta(p,h) \tag{4.30}$$

with $\delta(p,h) \to 0$ as $p \to \infty$ or $h \to 0$. By the argument of Theorem 4.3, there exist polynomials $w_m, z_l \in \mathcal{P}_p(\Omega_1)$ such that

$$\|u_m - w_m\|_{H^1(\Omega_1)} \leq C\left(\frac{h}{(p+1)^2}\right)^{\gamma_m} F_{\nu_{1,m}}(p,h)$$

and

$$\|v_l - z_l\|_{H^1(\Omega_1)} \leq C\left(\frac{h}{(p+1)^2}\right)^{\gamma} F_{\nu_{l,1}}(p,h).$$

Also, by Theorem 3.2, there exists a polynomial $w_0 \in \mathcal{P}_p(\Omega_1)$ such that

$$\|u_0 - w_0\|_{H^1(\Omega_1)} \leq Ch^{\mu-1}(p+1)^{-(k-1)},$$

where $\mu = \min\{k, p+1\}$. Therefore, combining the above estimates and the assumption (4.30) we obtain

$$\begin{aligned}
&\left\| u_1 - \left(u_{hp} - w_0 - \sum_{m\geq 2, 0<\gamma_m\leq k-1} w_m - \sum_{l=2}^{L_1} z_l \right) \right\|_{H^1(\Omega_1)} \\
&\leq \|u - u_{hp}\|_{H^1(\Omega_1)} + \|u_0 - w_0\|_{H^1(\Omega_1)} + \sum_{m\geq 2, 0<\gamma_m\leq k-1} \|u_m - w_m\|_{H^1(\Omega_1)} \\
&\quad + \sum_{l=2}^{L_1} \|v_l - z_l\|_{H^1(\Omega_1)} \leq C\left(\frac{h}{(p+1)^2}\right)^\gamma F_{\nu_\gamma}(p,h)\ \tilde{\delta}(p,h)
\end{aligned}$$

with

$$\tilde{\delta}(p,h) = \delta(p,h) + \frac{h^{\mu-1-\gamma}}{(p+1)^{k-1-2\gamma}} + \sum_{l=2}^{L_1} \frac{F_{\nu_{l,1}}(p,h)}{F_{\nu_\gamma}(p,h)} + \sum_{m\geq 2, 0<\gamma_m\leq k-1} \left(\frac{h}{(p+1)^2}\right)^{\gamma_m-\gamma} \frac{F_{\nu_{1,m}}(p,h)}{F_{\nu_\gamma}(p,h)}.$$

Note that $\gamma_m > \gamma$ for $m \geq 2$ and $\nu_{l,1} < \nu_\gamma$ for $l \geq 2$. Also note that $p > \gamma$ and $k > 1 + 2\gamma$, which implies $\mu > 1 + \gamma$. Hence $\tilde{\delta}(p,h) \to 0$ as $p \to \infty$ or $h \to 0$.

On the other hand, we have by Theorem 3.8

$$\left\| u_1 - \left(u_{hp} - w_0 - \sum_{m\geq 2, 0<\gamma_m\leq k-1} w_m - \sum_{l=2}^{L_1} z_l \right) \right\|_{H^1(\Omega_1)} \geq \inf_{\phi\in\mathcal{P}_p(\Omega_1)} \|u_1 - \phi\|_{H^1(\Omega_1)} \geq C \left(\frac{h}{(p+1)^2}\right)^\gamma F_{\nu_\gamma}(p,h),$$

which leads to a contradiction. Thus we complete the proof. □

As a corollary of Theorems 4.4 and 4.5, we have the optimal convergence of the finite element solutions of the h-p version with quasi-uniform meshes and quasi-uniform degree for elliptic problems on polygonal domains.

THEOREM 4.6. *Let u and u_{hp} be the solution of the problem* (4.1) *and its finite element solution in $S_D^p(\Omega;\Delta_h;\mathcal{M})$, respectively, as in the previous theorem. Then there exist two constants C_1 and C_2 depending on u, k, γ, and ν_γ but not on p and h such that*

$$C_2\left(\frac{h}{p^2}\right)^\gamma F_{\nu_\gamma}(p,h) \leq \|u - u_{hp}\|_{H^1(\Omega)} \leq C_1\left(\frac{h}{p^2}\right)^\gamma F_{\nu_\gamma}(p,h), \tag{4.31}$$

where γ and ν_γ are as given in (4.22) *and $F_{\nu_\gamma}(p,h)$ is as given in* (3.19).

5. Concluding remarks. Based on the approximabilities of smooth and singular functions in the Jacobi-weighted Besov and Sobolev spaces, we have proved the convergence of the h-p version of the FEM with quasi-uniform meshes for problems with smooth and singular solutions. The analysis is conducted in the framework of the Jacobi-weighted Besov and Sobolev spaces, which proves that the Jacobi-weighted spaces are not only appropriate for problems with singular solutions but also for problems with smooth solutions. Hence this framework is the most powerful tool for analyzing the p and h-p versions of the FEM.

For problems with singular solutions of $r^\gamma \log^\nu r$-type, the optimal convergence for the h-p version of the FEM is established after deriving the sharpest estimation of upper and lower bounds for the approximation errors in FEM solutions. Within the framework of the Jacobi-weighted Besov and Sobolev spaces and by incorporating properly the well-designed scaling arguments, we have proved the optimal rate of convergence of the h-p version of the FEM with quasi-uniform meshes for elliptic problems on polygonal domains where singularities of $r^\gamma \log r$-type occur. The results include the h- and p-versions of the FEM as two special cases. For fixed h, it coincides with the optimal convergence of the p-version of the FEM [6], and for fixed p, it gives the optimal convergence of the h-version. Also the results are parallel to those of the h-p version of the BEM with quasi-uniform meshes [20].

The concepts, methods, and techniques in this paper can be generalized to the *h-p* of the FEM in three-dimensional problems, but such a generalization will be substantial and is feasible only when the analysis for the convergence of the *p*-version of the FEM in three dimensions becomes available in the future.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.

[2] M. Ainsworth and J. Coyle, *Computation of Maxwell eigenvalues on curvilinear domains using hp-version Nédélec elements*, in Numerical Mathematics and Advanced Applications, F. Brezzi et al., eds., Springer-Verlag, Berlin, 2003, pp. 219–231.

[3] M. Ainsworth and K. Kay, *Approximation theory for the hp-version finite element method and application to the non-linear Laplacian*, Appl. Numer. Math, 34 (2000), pp. 329–344.

[4] M. Ainsworth and K. Pinchedez, *hp-MITC finite element method for the Reissner-Mindlin plate problem*, J. Comput. Appl. Math., 148 (2002), pp. 429–462.

[5] I. Babuška and B. Q. Guo, *Direct and inverse approximation theorems for the p-version of the finite element method in the framework of weighted Besov spaces, Part* I: *Approximability of functions in the weighted Besov spaces*, SIAM J. Numer. Anal., 39 (2001), pp. 1512–1538.

[6] I. Babuška and B. Q. Guo, *Direct and inverse approximation theorems of the p-version of the finite element method in the framework of weighted Besov spaces, Part* II: *Optimal rate of convergence of the p-version finite element solutions*, Math. Models Methods Appl. Sci., 12 (2002), pp. 689–719.

[7] I. Babuška and B. Q. Guo, *Optimal estimates for lower and upper bounds of approximation errors in the p-version of the finite element method in two dimensions*, Numer. Math., 85 (2000), pp. 219–255.

[8] I. Babuška, R. B. Kellogg, and J. Pitkäranta, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.

[9] I. Babuška and M. Suri, *The optimal convergence rate of the p-version of the finite element method*, SIAM J. Numer. Anal., 24 (1987), pp. 750–776.

[10] I. Babuška and M. Suri, *The h-p version of the finite element method with quasi-uniform meshes*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 199–238.

[11] J. Bergh and J. Löfström, *Interpolation Spaces*, Springer-Verlag, Berlin, Heidelberg, New York, 1976.

[12] C. Bernardi and Y. Maday, *Spectral methods,* in Handbook of Numerical Analysis, Vol. V, Part 2, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1997, pp. 209–485.

[13] K. S. Bey and J. T. Oden, *hp-version discontinuous Galerkin methods for hyperbolic conservation laws*, Comput. Methods. Appl. Mech. Engrg., 133 (1996), pp. 259–286.

[14] R. Courant and D. Hillbert, *Methods of Mathematical Physics*, Vol. 1, Academic Press, New York, 1953.

[15] L. Demkowicz, P. Monk, L. Vardapetyan, and W. Rachowicz, *de Rham diagram for hp finite element spaces*, Comput. Math. Appl., 39 (2000), pp. 29–38.

[16] W. Gui and I. Babuška, *The h, p, and h-p versions of the finite element method in* 1 *dimension.* I. *The error analysis of the p-version;* II. *The error analysis of the h and h-p versions*, Numer. Math., 49 (1986), pp. 577–612; 613–657.

[17] B. Q. Guo, *Approximation theory for the p-version of the finite element method in three dimensions. Part* 1: *Approximabilities of singular functions in the framework of the Jacobi-weighted Besov and Sobolev spaces*, SIAM J. Numer. Anal., 44 (2006), pp. 246–269.

[18] B. Q. Guo, *The p and h-p Finite Element Analysis, Theory, Algorithm and Computation*, Lecture note, Department of Mathematics, University of Manitoba, Winnipeg, MB, Canada, 2004.

[19] B. Q. Guo and N. Heuer, *The optimal convergence of the p-version of the boundary element method in two dimensions*, Numer. Math., 98 (2004), pp. 499–538.

[20] B. Q. Guo and N. Heuer, *The optimal convergence of the h-p version of the boundary element method with quasiuniform meshes for elliptic problems on polygonal domains*, Adv. Comput. Math. 24 (2006), pp. 353–374.

[21] B. Y. Guo, *Jacobi approximations in certain Hilbert spaces and their applications to singular differential equations*, J. Math. Anal. Appl., 243 (2000), pp. 373–408.

[22] B. Y. Guo, *Gegenbauer approximation and its applications to differential equations on the whole line*, J. Math. Anal. Appl., 226 (1998), pp. 180–206.

[23] B. Y. Guo and L. L. Wang, *Jacobi approximations in non-uniformly Jacobi-weighted Sobolev spaces*, J. Approx. Theory, 128 (2004), pp. 1–41.
[24] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element method for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.
[25] D. Schötzau, C. Schwab, and A. Toselli, *Mixed hp-DGFEM for incompressible flows*, SIAM J. Numer. Anal., 40 (2003), pp. 2171–2194.
[26] C. Schwab and M. Suri, *Mixed hp finite element methods for Stokes and non-Newtonian flow*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 217–241.
[27] P. Seshaiyer and M. Suri, *Uniform hp convergence results for the mortar finite element method*, Math. Comp., 69 (2000), pp. 481–500.
[28] E. P. Stephan and M. Suri, *The h-p version of the boundary element method on polygonal domains with quasiuniform meshes*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 783–807.
[29] L. Vardapetyan and L. Demkowicz, *hp-adaptive finite elements in electromagnetics*, Comput. Methods Appl. Mech. Engrg., 169 (1999), pp. 331–344.
[30] L. Vardapetyan, L. Demkowicz, and D. Neikirk, *hp-vector finite element method for eigenmode analysis of waveguides*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 185–201.
[31] T. Werder, K. Gerdes, D. Schötzau, and C. Schwab, *hp-discontinuous Galerkin time stepping for parabolic problems*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6685–6708.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 731–745

FIRST-ORDER SYSTEM LEAST-SQUARES FOR DARCY–STOKES FLOW*

GARVIN DANISCH† AND GERHARD STARKE†

Abstract. The subject of this paper is a first-order system least-squares formulation for the Stokes equation which remains uniformly valid in the limit of vanishing viscosity. For this so-called Darcy–Stokes flow problem we establish continuity and coercivity of the corresponding least-squares functional in appropriate norms. Two types of finite element spaces for the approximation of the velocity field are investigated in detail: the well-known Raviart–Thomas elements and an element recently introduced by Mardal, Tai, and Winther specifically for mixed approaches to Darcy–Stokes flow. The computational results derived with next-to-lowest order Raviart–Thomas elements as well as the Mardal–Tai–Winther elements confirm our analysis.

Key words. least-squares finite element method, first-order system, Darcy–Stokes flow, small viscosity

AMS subject classifications. 65M60, 65M15

DOI. 10.1137/050638163

1. Introduction.

Our purpose in this paper is to present a least-squares finite element method for Darcy–Stokes flow which remains valid for arbitrarily small viscosity. This type of singular perturbation problem was studied before in [11], where a successful mixed finite element approach was presented. The mixed variational formulation of [11] is of saddle point structure with its well-known limitation on the admissible combinations of finite element spaces. One of the motivations for the development of the least-squares approach presented in this paper is the greater flexibility in the choice of finite element spaces which is not restricted by a compatibility condition.

In the limit of vanishing viscosity, our least-squares formulation turns into the one proposed in [8]. The approach in [8] constructs approximations for the pressure and the velocity in $H^1(\Omega)$ and $H(\mathrm{div}, \Omega)$, respectively. In the viscous case, however, an approximation for the velocity is sought in $H^1(\Omega)^2$ instead. This is achieved by an augmentation with a least-squares functional along the edges of the triangulation over the jump of the tangential component and by introducing the velocity gradient as an additional variable. The case of small viscosity is handled by an appropriate weighting of the components in the least-squares functional.

Our main motivation for this work comes from the treatment of shallow water systems treated with the method of characteristics for time discretization. In this context, linearization of the boundary value problems at each time-step leads to flow problems of Darcy–Stokes type. Shallow water flow is described by the scalar water level and by the velocity field. These process variables are directly approximated by the first-order system least-squares formulation treated in this paper. The extension to shallow water systems including a viscosity term is therefore straightforward. For vanishing viscosity that approach reduces to the first-order system least-squares method investigated in [14].

*Received by the editors August 15, 2006; accepted for publication (in revised form) October 31, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sinum/45-2/63816.html
†Institut für Angewandte Mathematik, Leibniz Universität Hannover, Welfengarten 1, 30167 Hannover, Germany (danisch@ifam.uni-hannover.de, gcs@ifam.uni-hannover.de).

Among the most popular methods for the case $\mu = 0$ is the Raviart–Thomas mixed finite element method which couples, for example, lowest-order Raviart–Thomas elements for the flux with piecewise constant functions for the scalar variable. This approach is well studied in the case of the linear first-order Darcy-type system (see, e.g., [4, section III.5]) as well as for the shallow water system without viscosity (see [10, 13]). For nonvanishing viscosity different pairs of finite element spaces are used. Among the most common approaches is the Taylor–Hood element pair which combines quadratic conforming elements for the velocity field with linear conforming ones on the same mesh for the scalar variable. Again, this approach is well known to be stable for the mixed variational formulation of the Stokes problem (see, e.g., [4, section III.7]). It is also widely used for the numerical treatment of the shallow water equations with viscosity (see, e.g., [12]).

A smooth transition between both of these situations was achieved only recently by Mardal, Tai, and Winther in [11]. Their element is nonconforming for the Stokes system (i.e., with respect to $H^1(\Omega)^2$), and it is conforming in the case of vanishing viscosity (i.e., with respect to $H(\mathrm{div}, \Omega)$). The finite element of the Mardal–Tai–Winther approach is represented by three basis functions per edge, two for the normal component and one for the tangential component of the velocity field. It is shown to be stable if combined with a piecewise constant pressure approximation (cf. [11]). Another nonconforming approach for the Darcy–Stokes problem was studied by Burman and Hansbo in [7] based on a stabilized Crouzeix–Raviart element.

Our approach proposed in this paper is based on a least-squares formulation of the Darcy–Stokes system, which is obtained by introducing the velocity flux as an auxiliary variable. For nonvanishing viscosity it is augmented with an edge functional involving the tangential velocity component. If next-to-lowest order (quadratic) Raviart–Thomas elements are used for the velocity field, then approximation order 2 is obtained for the case of zero viscosity. In the presence of positive viscosity, only approximation order 1 is achieved. More precisely, the behavior for small viscosity is such that the error reduction is of order 2 on coarse meshes and eventually reduces to order 1 on finer meshes where the viscous error components are no longer negligible. Of course, the finite elements of Mardal, Tai, and Winther can also be inserted into our least-squares formulation. This leads to approximation order 1 independently of the size of the viscosity which is also verified by our numerical computations.

The elementwise evaluation of the least-squares functional constitutes an a posteriori error estimator at no additional cost. This a posteriori error estimator gives rise to adaptive refinement strategies which dramatically increase the accuracy and efficiency of numerical methods in many practical situations. However, we do not report on adaptive computations in this paper and refer to the extension for the viscous shallow water equations in [9] instead.

For the above reasons, among others, least-squares finite element methods have become increasingly popular in recent years for a number of different application problems; see [2] for an overview. Several least-squares formulations for the Navier–Stokes equations have been studied in [1, 3], where the partial derivatives of the velocity field are also introduced as additional variables.

The structure of this paper is as follows. The first-order system formulation of Darcy–Stokes flow is presented in the next section. Section 3 investigates the nonconforming least-squares formulation, which is set in the space $H(\mathrm{div}, \Omega)$. It is shown that the least-squares functional satisfies continuity and coercivity estimates with respect to suitable norms. Finite element approximation estimates which are uniform in the limit of vanishing viscosity are derived in section 4. Finally, in section 5

the computational results for a test example with varying viscosity parameter μ are reported.

2. First-order system formulation of Darcy–Stokes flow. For a region $\Omega \subset \mathbb{R}^2$ with boundary $\Gamma = \partial\Omega$ we consider the boundary value problem

$$\begin{aligned} \delta\, p + \operatorname{div} \mathbf{u} &= 0 \quad \text{in } \Omega, \\ \mathbf{u} + \nabla p - \mu \boldsymbol{\Delta}\, \mathbf{u} &= \mathbf{0} \quad \text{in } \Omega, \\ \mathbf{n} \cdot \mathbf{u} &= g \quad \text{on } \Gamma, \\ \mu(\mathbf{t} \cdot \mathbf{u}) &= 0 \quad \text{on } \Gamma, \end{aligned} \tag{2.1}$$

with nonnegative parameters δ, μ. For $(\delta, \mu) = (0, 1)$ this constitutes a stationary Stokes system, while for $(\delta, \mu) = (0, 0)$, (2.1) is simply a first-order reformulation of the Laplace equation. Our aim in this work is the development of a discretization scheme which remains stable uniformly as $\mu \to 0$. This involves that the boundary condition for the tangential velocity component must be smoothly faded out for $\mu \to 0$.

For the solution of (2.1) we propose a least-squares finite element method which starts from the first-order system

$$\mathcal{R}(p, \mathbf{u}, \mathbf{U}) = \begin{pmatrix} \delta\, p + \operatorname{div} \mathbf{u} \\ \mathbf{u} + \nabla p + \mu^{1/2} \operatorname{div} \mathbf{U} \\ \mathbf{U} + \mu^{1/2} \nabla \mathbf{u} \end{pmatrix} = \mathbf{0} \tag{2.2}$$

resulting from the introduction of $\mathbf{U}$ as an additional variable. Our variational formulation will be based on the Sobolev spaces

$$\begin{aligned} H^1_\Gamma(\Omega) &= \{q \in H^1(\Omega) : q = 0 \text{ on } \Gamma\}, \\ H_\Gamma(\operatorname{div}, \Omega) &= \{\mathbf{v} \in H(\operatorname{div}, \Omega) : \mathbf{n} \cdot \mathbf{v} = 0 \text{ on } \Gamma\}. \end{aligned}$$

$H(\operatorname{div}, \Omega)$ denotes the Sobolev space corresponding to the norm

$$\|\mathbf{v}\|_{\operatorname{div},\Omega} = \left(\|\mathbf{v}\|^2_{0,\Omega} + \|\operatorname{div} \mathbf{v}\|^2_{0,\Omega}\right)^{1/2}.$$

If we construct a function $\mathbf{u}^N \in H^1(\Omega)$ which satisfies the boundary conditions in (2.1), then our aim is to solve (2.2) for $p \in H^1(\Omega)$, $\mathbf{u} = \mathbf{u}^N + \hat{\mathbf{u}}$ with $\hat{\mathbf{u}} \in H^1_\Gamma(\Omega)^2$, and $\mathbf{U} \in H(\operatorname{div}, \Omega)^2$. The associated least-squares variational formulation would consist in finding $(p, \hat{\mathbf{u}}, \mathbf{U}) \in H^1(\Omega) \times H^1_\Gamma(\Omega)^2 \times H(\operatorname{div}, \Omega)^2$ such that

$$\|\mathcal{R}(p, \mathbf{u}^N + \hat{\mathbf{u}}, \mathbf{U})\|^2_{0,\Omega} \le \|\mathcal{R}(q, \mathbf{u}^N + \mathbf{v}, \mathbf{V})\|^2_{0,\Omega} \tag{2.3}$$

holds for all $(q, \mathbf{v}, \mathbf{V}) \in H^1(\Omega) \times H^1_\Gamma(\Omega)^2 \times H(\operatorname{div}, \Omega)^2$.

Before we turn to the investigation of this least-squares formulation, we consider the special case $\mu = 0$ in more detail. Let us denote the first-order system (2.2) corresponding to $\mu = 0$ by

$$\mathcal{R}_0(p, \mathbf{u}, \mathbf{U}) = \begin{pmatrix} \delta\, p + \operatorname{div} \mathbf{u} \\ \mathbf{u} + \nabla p \\ \mathbf{U} \end{pmatrix}$$

and observe that

$$\mathcal{R}(p, \mathbf{u}, \mathbf{U}) = \mathcal{R}_0(p, \mathbf{u}, \mathbf{U}) + \mu^{1/2} \begin{pmatrix} 0 \\ \operatorname{div} \mathbf{U} \\ \nabla \mathbf{u} \end{pmatrix} \tag{2.4}$$

holds. For $\mu = 0$, (2.3) reduces to the least-squares minimization of $\|\mathcal{R}_0(p, \mathbf{u}, \mathbf{U})\|_{0,\Omega}$ for $p \in H^1(\Omega)$, $\mathbf{u} \in \mathbf{u}^N + H^1_\Gamma(\Omega)^2$, and $\mathbf{U} \in H(\mathrm{div}, \Omega)^2$. Unfortunately, this formulation is no longer well-posed and needs to be formulated in the larger product space $H^1(\Omega) \times (\mathbf{u}^N + H_\Gamma(\mathrm{div}, \Omega)) \times L^2(\Omega)^2$ instead. Moreover, only the normal component $\mathbf{n}\cdot\mathbf{u}^N$ may be prescribed at the boundary for $\mathbf{u}^N$, which only needs to be in $H(\mathrm{div}, \Omega)$. The well-posedness of this minimization problem may be deduced from coercivity and continuity of the associated variational formulation. In other words, it is required that there be positive constants α_0 and β_0 such that

$$\begin{aligned} \alpha_0 \left(\|q\|^2_{1,\Omega} + \|\mathbf{v}\|^2_{\mathrm{div},\Omega} + \|\mathbf{V}\|^2_{0,\Omega} \right) &\leq \|\mathcal{R}_0(q, \mathbf{v}, \mathbf{V})\|^2_{0,\Omega} \\ &\leq \beta_0 \left(\|q\|^2_{1,\Omega} + \|\mathbf{v}\|^2_{\mathrm{div},\Omega} + \|\mathbf{V}\|^2_{0,\Omega} \right) \end{aligned} \tag{2.5}$$

holds for all $(q, \mathbf{v}, \mathbf{V}) \in H^1(\Omega) \times H_\Gamma(\mathrm{div}, \Omega) \times L^2(\Omega)^2$. As a first step towards such an estimate we observe that

$$\|\mathcal{R}_0(q, \mathbf{v}, \mathbf{V})\|^2_{0,\Omega} \tag{2.6}$$
$$\begin{cases} \geq \min\{\delta, 1\} \left(\|\delta^{1/2} q + \delta^{-1/2} \,\mathrm{div}\, \mathbf{v}\|^2_{0,\Omega} + \|\mathbf{v} + \nabla q\|^2_{0,\Omega} + \|\mathbf{V}\|^2_{0,\Omega} \right) \\ \leq \max\{\delta, 1\} \left(\|\delta^{1/2} q + \delta^{-1/2} \,\mathrm{div}\, \mathbf{v}\|^2_{0,\Omega} + \|\mathbf{v} + \nabla q\|^2_{0,\Omega} + \|\mathbf{V}\|^2_{0,\Omega} \right) \end{cases}$$

holds. For $(q, \mathbf{v}, \mathbf{V}) \in H^1(\Omega) \times H_\Gamma(\mathrm{div}, \Omega) \times L^2(\Omega)^2$,

$$\begin{aligned} &\|\delta^{1/2} q + \delta^{-1/2} \,\mathrm{div}\, \mathbf{v}\|^2_{0,\Omega} + \|\mathbf{v} + \nabla q\|^2_{0,\Omega} + \|\mathbf{V}\|^2_{0,\Omega} \\ &\quad = \delta \|q\|^2_{0,\Omega} + 2(q, \mathrm{div}\, \mathbf{v})_{0,\Omega} + \delta^{-1} \|\mathrm{div}\, \mathbf{v}\|^2_{0,\Omega} \\ &\qquad + \|\nabla q\|^2_{0,\Omega} + 2(\mathbf{v}, \nabla q)_{0,\Omega} + \|\mathbf{v}\|^2_{0,\Omega} + \|\mathbf{V}\|^2_{0,\Omega} \\ &\quad = \delta \|q\|^2_{0,\Omega} + \|\nabla q\|^2_{0,\Omega} + \|\mathbf{v}\|^2_{0,\Omega} + \delta^{-1} \|\mathrm{div}\, \mathbf{v}\|^2_{0,\Omega} + \|\mathbf{V}\|^2_{0,\Omega}, \end{aligned} \tag{2.7}$$

where the mixed terms vanish due to integration by parts. Combined with (2.6), this proves (2.5) with $\alpha_0 = \min\{\delta^2, \delta^{-1}\}$ and $\beta_0 = \max\{\delta^2, \delta^{-1}\}$. We have therefore established continuity and coercivity with respect to $H^1(\Omega) \times H_\Gamma(\mathrm{div}, \Omega) \times L^2(\Omega)^2$ in this case.

For simplicity, we have assumed $\delta > 0$ in our analysis above, and we will continue to do so throughout the rest of this paper. The case $\delta = 0$ in (2.2) may be considered by adding a constraint such as

$$\int_\Gamma p \, ds = 0,$$

which makes p unique. Of course, the boundary conditions must be compatible in this case, which implies that

$$\int_\Gamma g \, ds = 0$$

must hold. Such an example will also be included in our numerical results.

The formulation (2.3) based on the first-order system (2.2) is only meaningful under the assumption $\mu > 0$, while for $\mu \to 0$ the problem does not remain uniformly well-posed with respect to $H^1(\Omega) \times H^1_\Gamma(\Omega)^2 \times H(\mathrm{div}, \Omega)^2$. Our goal in this paper is to derive an approach which is valid for the entire parameter range of $\mu \geq 0$. To this end, a suitable transition from $\mathbf{u} \in H^1_\Gamma(\Omega)^2$ to the space $H_\Gamma(\mathrm{div}, \Omega)$ is required. In the next section, we present an approach which treats the velocity in the space $H_\Gamma(\mathrm{div}, \Omega)$ in all cases and enforces the condition $\mathbf{u} \in H^1_\Gamma(\Omega)^2$ weakly for $\mu > 0$.

3. A nonconforming least-squares formulation in $H(\mathrm{div},\Omega)$. For the nonconforming least-squares formulation we define a family of triangulations $\mathcal{T}_h$ with a parameter h measuring the mesh resolution. The set of edges associated with the triangulation $\mathcal{T}_h$ is denoted by $\mathcal{E}_h$. For a piecewise polynomial velocity field $\mathbf{u} \in L^2(\Omega)^2$ and an edge $E \in \mathcal{E}_h$, the jump term may be defined as

$$[\mathbf{u}]_E = \begin{cases} \mathbf{u}|_{K_{l,E}} - \mathbf{u}|_{K_{r,E}}, & E \in \mathcal{E}_h \cap \Omega, \\ \mathbf{u}|_{K_{l,E}}, & E \in \mathcal{E}_h \cap \Gamma. \end{cases}$$

Here, $K_{l,E}$ and $K_{r,E}$ denote the left and right triangles, respectively, adjacent to E. Similarly, the jump term for the tangential velocity component may be defined as

$$[\mathbf{t}\cdot\mathbf{u}]_E = \begin{cases} \mathbf{t}\cdot\mathbf{u}|_{K_{l,E}} - \mathbf{t}\cdot\mathbf{u}|_{K_{r,E}}, & E \in \mathcal{E}_h \cap \Omega, \\ \mathbf{t}\cdot\mathbf{u}|_{K_{l,E}}, & E \in \mathcal{E}_h \cap \Gamma. \end{cases}$$

For piecewise polynomial $\mathbf{u} \in H_\Gamma(\mathrm{div},\Omega)$, in fact, since the normal jump component vanishes on all edges,

$$[\mathbf{u}]_E = [\mathbf{n}\cdot\mathbf{u}]_E\,\mathbf{n} + [\mathbf{t}\cdot\mathbf{u}]_E\,\mathbf{t} = [\mathbf{t}\cdot\mathbf{u}]_E\,\mathbf{t} \tag{3.1}$$

holds for all $E \in \mathcal{E}_h$.

With this, the least-squares functional may be defined as

$$\mathcal{F}(p,\mathbf{u},\mathbf{U}) = \sum_{K\in\mathcal{T}_h} \|\mathcal{R}(p,\mathbf{u},\mathbf{U})\|_{0,K}^2 + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{u}]_E\|_{0,E}^2. \tag{3.2}$$

The least-squares minimization problem consists of finding $(p,\hat{\mathbf{u}},\mathbf{U}) \in H^1(\Omega) \times H_\Gamma(\mathrm{div},\Omega) \times H(\mathrm{div},\Omega)^2$ such that

$$\mathcal{F}(p,\mathbf{u}^N+\hat{\mathbf{u}},\mathbf{U}) \le \mathcal{F}(q,\mathbf{u}^N+\mathbf{v},\mathbf{V}) \tag{3.3}$$

holds for all $(q,\mathbf{v},\mathbf{V}) \in H^1(\Omega)\times H_\Gamma(\mathrm{div},\Omega)\times H(\mathrm{div},\Omega)^2$. Similarly, the corresponding finite element approximation $(p_h,\hat{\mathbf{u}}_h,\mathbf{U}_h) \in Q_h \times \boldsymbol{\Sigma}_h \times \boldsymbol{\Theta}_h$ satisfies

$$\mathcal{F}(p_h,\mathbf{u}^N+\hat{\mathbf{u}}_h,\mathbf{U}_h) \le \mathcal{F}(q_h,\mathbf{u}^N+\mathbf{v}_h,\mathbf{V}_h) \tag{3.4}$$

for all $(q_h,\mathbf{v}_h,\mathbf{V}_h) \in Q_h \times \boldsymbol{\Sigma}_h \times \boldsymbol{\Theta}_h$. Here, $Q_h \times \boldsymbol{\Sigma}_h \times \boldsymbol{\Theta}_h \subset H^1(\Omega) \times H_\Gamma(\mathrm{div},\Omega) \times H(\mathrm{div},\Omega)^2$ denote appropriate finite-dimensional spaces to be specified in section 4. The bilinear form associated with the functional (3.2) is given by

$$\begin{aligned} &\mathcal{B}(p,\mathbf{u},\mathbf{U};q,\mathbf{v},\mathbf{V}) \\ &\quad= \sum_{K\in\mathcal{T}_h} (\mathcal{R}(p,\mathbf{u},\mathbf{U}),\mathcal{R}(q,\mathbf{v},\mathbf{V}))_{0,K} + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}([\mathbf{t}\cdot\mathbf{u}]_E,[\mathbf{t}\cdot\mathbf{v}]_E)_{0,E}. \end{aligned} \tag{3.5}$$

The solution of (3.3) also satisfies the variational formulation

$$\mathcal{B}(p,\mathbf{u}^N+\hat{\mathbf{u}},\mathbf{U};r,\mathbf{w},\mathbf{W}) = 0 \tag{3.6}$$

for all $(r,\mathbf{w},\mathbf{W}) \in H^1(\Omega) \times H_\Gamma(\mathrm{div},\Omega) \times H(\mathrm{div},\Omega)^2$. Similarly, the solution of (3.4) is characterized by

$$\mathcal{B}(p_h,\mathbf{u}^N+\hat{\mathbf{u}}_h,\mathbf{U}_h;r_h,\mathbf{w}_h,\mathbf{W}_h) = 0 \tag{3.7}$$

for all $(r_h, \mathbf{w}_h, \mathbf{W}_h) \in Q_h \times \mathbf{\Sigma}_h \times \mathbf{\Theta}_h$.

Our aim is to show that the least-squares functional associated with (3.3),

$$\sum_{K\in\mathcal{T}_h} \|\mathcal{R}(p,\mathbf{u},\mathbf{U})\|_{0,K}^2 + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{u}]_E\|_{0,E}^2, \tag{3.8}$$

is also continuous and coercive with respect to suitable norms. Clearly, as in (2.6), we have

$$\begin{aligned}\sum_{K\in\mathcal{T}_h} \|\mathcal{R}(q,\mathbf{v},\mathbf{V})\|_{0,K}^2 \le \max\{\delta,1\}\Bigg(&\|\delta^{1/2}q + \delta^{-1/2}\operatorname{div}\mathbf{v}\|_{0,\Omega}^2 \\ &+ \|\mathbf{v}+\nabla q+\mu^{1/2}\operatorname{div}\mathbf{V}\|_{0,\Omega}^2 + \sum_{K\in\mathcal{T}_h}\|\mathbf{V}+\mu^{1/2}\nabla\mathbf{v}\|_{0,K}^2\Bigg),\end{aligned}$$

and therefore

$$\begin{aligned}&\sum_{K\in\mathcal{T}_h} \|\mathcal{R}(q,\mathbf{v},\mathbf{V})\|_{0,K}^2 + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}]_E\|_{0,E}^2 \\ &\le 2\max\{\delta^2,\delta^{-1}\}\Bigg(\|\mathbf{v}\|_{0,\Omega}^2 + \|\operatorname{div}\mathbf{v}\|_{0,\Omega}^2 + \mu\sum_{K\in\mathcal{T}_h}\|\nabla\mathbf{v}\|_{0,\Omega}^2 + \mu\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}]_E\|_{0,E}^2 \\ &\qquad + \|q\|_{0,\Omega}^2 + \|\mathbf{V}\|_{0,\Omega}^2 + \|\nabla q+\mu^{1/2}\operatorname{div}\mathbf{V}\|_{0,\Omega}^2\Bigg)\end{aligned}$$

for all $(q,\mathbf{v},\mathbf{V}) \in H^1(\Omega)\times H_\Gamma(\operatorname{div},\Omega)\times H(\operatorname{div},\Omega)^2$. The term in brackets clearly defines a norm on $H^1(\Omega)\times H_\Gamma(\operatorname{div},\Omega)\times H(\operatorname{div},\Omega)^2$, which we may abbreviate as

$$\begin{aligned}|||(q,\mathbf{v},\mathbf{V})||| = \Bigg(&\|\mathbf{v}\|_{0,\Omega}^2 + \|\operatorname{div}\mathbf{v}\|_{0,\Omega}^2 + \mu\sum_{K\in\mathcal{T}_h}\|\nabla\mathbf{v}\|_{0,\Omega}^2 + \mu\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}]_E\|_{0,E}^2 \\ &+ \|q\|_{0,\Omega}^2 + \|\mathbf{V}\|_{0,\Omega}^2 + \|\nabla q+\mu^{1/2}\operatorname{div}\mathbf{V}\|_{0,\Omega}^2\Bigg)^{1/2}.\end{aligned}$$

From now on, we use the symbol $\lesssim$ to indicate that an inequality as above holds with constants which remain bounded as μ tends to 0. The above continuity estimate may therefore be rewritten as follows.

THEOREM 3.1. *For the least-squares functional defined in* (3.2),

$$\mathcal{F}(q,\mathbf{v},\mathbf{V}) \lesssim |||(q,\mathbf{v},\mathbf{V})|||^2 \tag{3.9}$$

holds for all $(q,\mathbf{v},\mathbf{V}) \in H^1(\Omega)\times H_\Gamma(\operatorname{div},\Omega)\times H(\operatorname{div},\Omega)^2$.

Naturally, the derivation of a coercivity estimate is more complicated. To this end, we define

$$\hat{\mathcal{R}}(q,\mathbf{v},\mathbf{V}) = \begin{pmatrix}\delta^{1/2}q + \delta^{-1/2}\operatorname{div}\mathbf{v}\\ \mathbf{v}+\nabla q+\mu^{1/2}\operatorname{div}\mathbf{V}\\ \mathbf{V}+\mu^{1/2}\nabla\mathbf{v}\end{pmatrix} \tag{3.10}$$

and observe that

$$\|\mathcal{R}(q,\mathbf{v},\mathbf{V})\|_{0,\Omega}^2 \geq \min\{\delta,1\}\|\hat{\mathcal{R}}(q,\mathbf{v},\mathbf{V})\|_{0,\Omega}^2$$

holds for all $(q,\mathbf{v},\mathbf{V}) \in H^1(\Omega)\times H_\Gamma(\mathrm{div},\Omega)\times H(\mathrm{div},\Omega)^2$ due to (2.6). This leaves us with the task of deriving a lower bound for $\hat{\mathcal{R}}(q,\mathbf{v},\mathbf{V})$. In fact, since our approximation results will be based on Strang's lemma (see, e.g., [4, section III.1] or [5, section 10.1]), it suffices to show coercivity of $\hat{\mathcal{R}}(q,\mathbf{v},\mathbf{V})$ with respect to the finite element spaces $Q_h\times\boldsymbol{\Sigma}_h\times\boldsymbol{\Theta}_h \subset H^1(\Omega)\times H_\Gamma(\mathrm{div},\Omega)\times H(\mathrm{div},\Omega)^2$.

We start with a technical lemma that will be used later in the analysis.

LEMMA 3.2. *Assume that $\boldsymbol{\Theta}_h \subset H(\mathrm{div},\Omega)^2$ and $\boldsymbol{\Sigma}_h \subset H_\Gamma(\mathrm{div},\Omega)$ are piecewise polynomial finite element spaces on $\mathcal{T}_h$. Then there is a constant $C>1$ such that the inequality*

$$2\left|\sum_{E\in\mathcal{E}_h}(\mathbf{V}_h\cdot\mathbf{n},[\mathbf{v}]_E)_{0,E}\right| \leq \frac{1}{2\mu^{1/2}}\|\mathbf{V}_h\|_{0,\Omega}^2 + 2C\mu^{1/2}\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|_{0,E}^2 \tag{3.11}$$

holds for all $\mathbf{V}_h\in\boldsymbol{\Theta}_h$ and $\mathbf{v}\in\boldsymbol{\Sigma}_h$.

Proof. Using (3.1) and the Cauchy–Schwarz inequality, we are led to

$$\begin{aligned}\left|\sum_{E\in\mathcal{E}_h}(\mathbf{V}_h\cdot\mathbf{n},[\mathbf{v}]_E)_{0,E}\right| &= \left|\sum_{E\in\mathcal{E}_h}(\mathbf{t}(\mathbf{V}_h\cdot\mathbf{n}),[\mathbf{t}\cdot\mathbf{v}_h]_E)_{0,E}\right| \\ &\leq \sum_{E\in\mathcal{E}_h}\|\mathbf{t}\cdot(\mathbf{V}_h\cdot\mathbf{n})\|_{0,E}\,\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|_{0,E} \qquad (3.12)\\ &\leq \frac{\rho}{2}\sum_{E\in\mathcal{E}_h}h_E\|\mathbf{t}\cdot(\mathbf{V}_h\cdot\mathbf{n})\|_{0,E}^2 + \frac{1}{2\rho}\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|_{0,E}^2\end{aligned}$$

with $\rho\in(0,1)$ still to be chosen appropriately below. Moreover, for all $\mathbf{V}_h\in\boldsymbol{\Theta}_h$,

$$\begin{aligned}\sum_{E\in\mathcal{E}_h}h_E\|\mathbf{t}\cdot(\mathbf{V}_h\cdot\mathbf{n})\|_{0,E}^2 &\leq \sum_{E\in\mathcal{E}_h}h_E\|\mathbf{V}_h\|_{0,E}^2 \leq \sum_{K\in\mathcal{T}_h}\sum_{E\subset\partial K}h_E\|\mathbf{V}_h\|_{0,E}^2 \\ &\leq C\sum_{K\in\mathcal{T}_h}\|\mathbf{V}_h\|_{0,K}^2 = C\|\mathbf{V}_h\|_{0,\Omega}^2\end{aligned}$$

holds with a constant C ($h_E\|\mathbf{V}_h\|_{0,E}^2 \leq C\|\mathbf{V}_h\|_{0,K}^2$ is due to a scaling argument). Assuming, without loss of generality, $C>1/2$ and setting $\rho=1/(2C)$ in (3.12) finally implies

$$2\sum_{E\in\mathcal{E}_h}(\mathbf{V}_h\cdot\mathbf{n},[\mathbf{v}]_E)_{0,E} \leq \frac{1}{2\mu^{1/2}}\|\mathbf{V}_h\|_{0,\Omega}^2 + 2C\mu^{1/2}\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|_{0,E}^2,$$

which also proves (3.11). ∎

We are now ready to establish our coercivity result.

THEOREM 3.3. *Assume that $Q_h\subset H^1(\Omega)$, $\boldsymbol{\Sigma}_h\subset H_\Gamma(\mathrm{div},\Omega)$, and $\boldsymbol{\Theta}_h\subset H(\mathrm{div},\Omega)^2$ are piecewise polynomial finite element spaces. Then, for the least-squares functional defined in* (3.2),

$$|||(q_h,\mathbf{v}_h,\mathbf{V}_h)|||^2 \lesssim \mathcal{F}(q_h,\mathbf{v}_h,\mathbf{V}_h) \tag{3.13}$$

holds for all $(q_h, \mathbf{v}_h, \mathbf{V}_h) \in Q_h \times \boldsymbol{\Sigma}_h \times \boldsymbol{\Theta}_h$.

Proof. The definition of the operator in (3.10) immediately leads to

$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h} \|\hat{\mathcal{R}}(q,\mathbf{v},\mathbf{V})\|_{0,K}^2 \\
&\quad = \|\delta^{1/2} q + \delta^{-1/2}\operatorname{div}\mathbf{v}\|_{0,\Omega}^2 + \|\mathbf{v} + \nabla q + \mu^{1/2}\operatorname{div}\mathbf{V}\|_{0,\Omega}^2 + \sum_{K\in\mathcal{T}_h} \|\mathbf{V} + \mu^{1/2}\nabla\mathbf{v}\|_{0,K}^2 \\
&\quad = \delta\,\|q\|_{0,\Omega}^2 + \delta^{-1}\|\operatorname{div}\mathbf{v}\|_{0,\Omega}^2 + \|\mathbf{v}\|_{0,\Omega}^2 + \|\nabla q + \mu^{1/2}\operatorname{div}\mathbf{V}\|_{0,\Omega}^2 + \|\mathbf{V}\|_{0,\Omega}^2 \\
&\qquad + \mu \sum_{K\in\mathcal{T}_h} \|\nabla\mathbf{v}\|_{0,K}^2 + 2\mu^{1/2} \sum_{E\in\mathcal{E}_h} (\mathbf{V}\cdot\mathbf{n}, [\mathbf{v}]_E)_{0,E}
\end{aligned}$$

for all $(q,\mathbf{v},\mathbf{V}) \in H^1(\Omega)\times H_\Gamma(\operatorname{div},\Omega)\times H(\operatorname{div},\Omega)^2$, where integration by parts is used at appropriate places. With the constant C from Lemma 3.2, using (3.11) gives

(3.14)
$$\begin{aligned}
&\sum_{K\in\mathcal{T}_h} \|\hat{\mathcal{R}}(q_h,\mathbf{v}_h,\mathbf{V}_h)\|_{0,K}^2 + (2C+1)\mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|_{0,E}^2 \\
&\qquad \geq \|\mathbf{v}_h\|_{0,\Omega}^2 + \delta^{-1}\|\operatorname{div}\mathbf{v}_h\|_{0,\Omega}^2 + \mu \sum_{K\in\mathcal{T}_h} \|\nabla\mathbf{v}_h\|_{0,K}^2 + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{v}]_E\|_{0,E}^2 \\
&\qquad + \delta\,\|q_h\|_{0,\Omega}^2 + \frac{1}{2}\|\mathbf{V}_h\|_{0,\Omega}^2 + \|\nabla q_h + \mu^{1/2}\operatorname{div}\mathbf{V}_h\|_{0,\Omega}^2
\end{aligned}$$

for all $(q_h,\mathbf{v}_h,\mathbf{V}_h) \in Q_h \times \boldsymbol{\Sigma}_h \times \boldsymbol{\Theta}_h$, which completes the proof of (3.13). ∎

THEOREM 3.4. *Let* $(p,\mathbf{u},\mathbf{U}) \in H^1(\Omega)\times H_\Gamma(\operatorname{div},\Omega)\times H(\operatorname{div},\Omega)^2$ *be the exact solution of* (3.3)*, and let* $(p_h,\mathbf{u}_h,\mathbf{U}_h) \in Q_h\times\boldsymbol{\Sigma}_h\times\boldsymbol{\Theta}_h$ *be the nonconforming approximation* (3.4)*. Then,*

$$\begin{aligned}
&|||(p-p_h,\mathbf{u}-\mathbf{u}_h,\mathbf{U}-\mathbf{U}_h)||| \\
&\qquad \lesssim \inf_{(q_h,\mathbf{v}_h,\mathbf{V}_h)\in Q_h\times\boldsymbol{\Sigma}_h\times\boldsymbol{\Theta}_h} |||(p-q_h,\mathbf{u}-\mathbf{v}_h,\mathbf{U}-\mathbf{V}_h)|||.
\end{aligned} \tag{3.15}$$

Proof. Strang's lemma (cf. [5, Lemma 10.1.1]) gives

$$\begin{aligned}
|||(p-p_h,\mathbf{u}-\mathbf{u}_h,\mathbf{U}-\mathbf{U}_h)||| &\lesssim \inf_{(q_h,\mathbf{v}_h,\mathbf{V}_h)\in Q_h\times\boldsymbol{\Sigma}_h\times\boldsymbol{\Theta}_h} |||(p-q_h,\mathbf{u}-\mathbf{v}_h,\mathbf{U}-\mathbf{V}_h)||| \\
&+ \sup_{(r_h,\mathbf{w}_h,\mathbf{W}_h)\in Q_h\times\boldsymbol{\Sigma}_h\times\boldsymbol{\Theta}_h} \frac{\mathcal{B}(p-p_h,\mathbf{u}-\mathbf{u}_h,\mathbf{U}-\mathbf{U}_h;r_h,\mathbf{w}_h,\mathbf{W}_h)}{|||(r_h,\mathbf{w}_h,\mathbf{W}_h)|||}.
\end{aligned}$$

We investigate the second term on the right-hand side, the so-called consistency error, more closely. For its numerator, due to (3.7), we obtain

$$\begin{aligned}
&\mathcal{B}(p-p_h,\mathbf{u}-\mathbf{u}_h,\mathbf{U}-\mathbf{U}_h;r_h,\mathbf{w}_h,\mathbf{W}_h) = \mathcal{B}(p,\mathbf{u},\mathbf{U};r_h,\mathbf{w}_h,\mathbf{W}_h) \\
&\quad = \sum_{K\in\mathcal{T}_h} (\mathcal{R}(p,\mathbf{u},\mathbf{U}),\mathcal{R}(r_h,\mathbf{w}_h,\mathbf{W}_h))_{0,K} + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E}([\mathbf{t}\cdot\mathbf{u}]_E,[\mathbf{t}\cdot\mathbf{w}_h]_E)_{0,E}.
\end{aligned}$$

Since, for the exact solution $(p,\mathbf{u},\mathbf{U})$, $\mathcal{R}(p,\mathbf{u},\mathbf{U}) = 0$ and $\mu[\mathbf{t}\cdot\mathbf{u}]_E = 0$ for all $E\in\mathcal{E}_h$, the consistency error vanishes, which completes the proof. ∎

The norm $|||(\cdot,\cdot,\cdot)|||$ contains a term which, for $\mu > 0$, couples ∇q and $\operatorname{div} \mathbf{V}$. A bound without such a coupling can easily be obtained by noting that

$$\|\mathbf{v}\|_{0,\Omega}^2 + \|\operatorname{div} \mathbf{v}\|_{0,\Omega}^2 + \mu \sum_{K\in\mathcal{T}_h} \|\nabla \mathbf{v}\|_{0,\Omega}^2 + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E} \|[\mathbf{t}\cdot\mathbf{v}]_E\|_{0,E}^2 + \|q\|_{0,\Omega}^2 + \|\mathbf{V}\|_{0,\Omega}^2 \le |||(q,\mathbf{v},\mathbf{V})|||^2$$

holds. Moreover, using

$$|||(q,\mathbf{v},\mathbf{V})|||^2 \le \|\mathbf{v}\|_{0,\Omega}^2 + \|\operatorname{div} \mathbf{v}\|_{0,\Omega}^2 + \mu \sum_{K\in\mathcal{T}_h} \|\nabla \mathbf{v}\|_{0,\Omega}^2 + \mu \sum_{E\in\mathcal{E}_h} \frac{1}{h_E} \|[\mathbf{t}\cdot\mathbf{v}]_E\|_{0,E}^2 + \|q\|_{0,\Omega}^2 + 2\|\nabla q\|_{0,\Omega}^2 + \|\mathbf{V}\|_{0,\Omega}^2 + 2\mu \|\operatorname{div} \mathbf{V}\|_{0,\Omega}^2$$

combined with the interpolation estimates presented in the following section, bounds for the approximation error may be deduced from Theorem 3.15.

4. Finite element approximation estimates. An appropriate choice of the combination of finite element spaces for the approximation of the variables in our least-squares formulation is motivated in the following discussion. To this end, let $m \ge 1$ be an integer. For the approximation p_h of $p \in H^1(\Omega)$, we use standard H^1-conforming finite elements of piecewise polynomials of degree $m+1$. We may denote the corresponding subspace as Q_h and get

$$\inf_{q_h\in Q_h} \|q - q_h\|_{1,\Omega} \lesssim h^{m+1} |q|_{m+2,\Omega} \tag{4.1}$$

from standard finite element interpolation results (cf. [4, section II.6]). For the finite element representation of the velocity field $\mathbf{u}$, $H(\operatorname{div},\Omega)$-conforming elements should be used. One possibility for the construction of an approximation $\mathbf{u}_h$ for the velocity field consists of the use of Raviart–Thomas spaces, which consist of piecewise polynomials of the form

$$\mathbf{u}_h|_K = \begin{pmatrix} p_m^{(I)} \\ p_m^{(II)} \end{pmatrix} + \mathbf{x} p_m^{(III)}$$

on each triangle $K \in \mathcal{T}_h$, where $p_m^{(I)}$, $p_m^{(II)}$, and $p_m^{(III)}$ denote polynomials of degree m. For $m = 1$, this implies

$$\mathbf{u}_h|_K = \begin{pmatrix} \alpha_K + \beta_K x_1 + \gamma_K x_2 \\ \delta_K + \rho_K x_1 + \sigma_K x_2 \end{pmatrix} + (\omega_K x_1 + \xi_K x_2) \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$$

with $\alpha_K, \beta_K, \gamma_K, \delta_K, \rho_K, \sigma_K, \omega_K, \xi_K \in \mathbb{R}$. If we denote the Raviart–Thomas space of degree m by $\mathbf{\Sigma}_h$, then [6, Proposition III.3.9] leads to

$$\inf_{\mathbf{v}_h\in\mathbf{\Sigma}_h} \|\mathbf{u} - \mathbf{v}_h\|_{\operatorname{div},\Omega} \lesssim h^{m+1} \left(|\mathbf{u}|_{m+1,\Omega} + |\operatorname{div} \mathbf{u}|_{m+1,\Omega} \right).$$

In addition, [6, Proposition III.3.6] generalizes this estimate to

$$\inf_{\mathbf{v}_h\in\mathbf{\Sigma}_h} \left(\|\mathbf{u} - \mathbf{v}_h\|_{\operatorname{div},\Omega}^2 + \mu \sum_{K\in\mathcal{T}_h} \|\nabla(\mathbf{u} - \mathbf{v}_h)\|_{0,K}^2 \right)^{1/2} \lesssim h^{m+1} \left(|\mathbf{u}|_{m+1,\Omega} + |\operatorname{div} \mathbf{u}|_{m+1,\Omega} \right) + h^m \mu^{1/2} |\mathbf{u}|_{m+1,\Omega}. \tag{4.2}$$

Let $\mathbf{\Pi}_h : H_\Gamma(\mathrm{div},\Omega) \to \mathbf{\Sigma}_h$ be the interpolation operator used in association with the Raviart–Thomas spaces (see [6, section III.3]). If, furthermore, $\mathbf{\Phi}_h : H^1_\Gamma(\Omega)^2 \to Q^2_h$ denotes the standard finite element interpolation operator, then we have

$$\begin{aligned}\|[\mathbf{t}\cdot(\mathbf{\Pi}_h\mathbf{u})]_E\|^2_{0,E} &= \|[\mathbf{\Pi}_h\mathbf{u}]_E\|^2_{0,E} = \|\,\mathbf{\Pi}_h\mathbf{u}|_{K_{l,E}} - \mathbf{\Pi}_h\mathbf{u}|_{K_{r,E}}\,\|^2_{0,E}\\ &\le 2\|\,\mathbf{\Pi}_h\mathbf{u}|_{K_{l,E}} - \mathbf{\Phi}_h\mathbf{u}\|^2_{0,E} + 2\|\,\mathbf{\Pi}_h\mathbf{u}|_{K_{r,E}} - \mathbf{\Phi}_h\mathbf{u}\|^2_{0,E},\end{aligned}$$

where the first identity follows from (3.1). This leads to

$$\begin{aligned}\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot(\mathbf{\Pi}_h\mathbf{u})]_E\|^2_{0,E} &\le 2\sum_{K\in\mathcal{T}_h}\sum_{E\subset\partial K}\frac{1}{h_E}\|\mathbf{\Pi}_h\mathbf{u}-\mathbf{\Phi}_h\mathbf{u}\|^2_{0,E}\\ &\lesssim \frac{1}{h}\sum_{K\in\mathcal{T}_h}\|\mathbf{\Pi}_h\mathbf{u}-\mathbf{\Phi}_h\mathbf{u}\|^2_{0,\partial K}\\ &\lesssim \frac{1}{h^2}\sum_{K\in\mathcal{T}_h}\|\mathbf{\Pi}_h\mathbf{u}-\mathbf{\Phi}_h\mathbf{u}\|^2_{0,K} = \frac{1}{h^2}\|\mathbf{\Pi}_h\mathbf{u}-\mathbf{\Phi}_h\mathbf{u}\|^2_{0,\Omega}\\ &\le \frac{2}{h^2}\|\mathbf{u}-\mathbf{\Pi}_h\mathbf{u}\|^2_{0,\Omega} + \|\mathbf{u}-\mathbf{\Phi}_h\mathbf{u}\|^2_{0,\Omega} \lesssim h^{2m}|\mathbf{u}|^2_{m+1,\Omega}.\end{aligned}$$

Thus, (4.2) can be augmented to

$$\begin{aligned}\inf_{\mathbf{v}_h\in\mathbf{\Sigma}_h}\left(\|\mathbf{u}-\mathbf{v}_h\|^2_{\mathrm{div},\Omega} + \mu\sum_{K\in\mathcal{T}_h}\|\nabla(\mathbf{u}-\mathbf{v}_h)\|^2_{0,K} + \mu\sum_{E\in\mathcal{E}_h}\frac{\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|^2_{0,E}}{h_E}\right)^{1/2}\\ \lesssim h^{m+1}\left(|\mathbf{u}|_{m+1,\Omega} + |\mathrm{div}\,\mathbf{u}|_{m+1,\Omega}\right) + h^m\mu^{1/2}|\mathbf{u}|_{m+1,\Omega}.\end{aligned} \tag{4.3}$$

The auxiliary variable $\mathbf{U}\in H(\mathrm{div},\Omega)^2$ may be approximated rowwise by Raviart–Thomas elements of order $m-1$, i.e.,

$$\mathbf{U}_h|_K = \begin{pmatrix}\alpha_{K,1}+\gamma_{K,1}x_1 & \beta_{K,1}+\gamma_{K,1}x_2\\ \alpha_{K,2}+\gamma_{K,2}x_1 & \beta_{K,2}+\gamma_{K,2}x_2\end{pmatrix}$$

with $\alpha_{K,i},\beta_{K,i},\gamma_{K,i}\in\mathbb{R}$, $i=1,2$. If $\mathbf{\Theta}_h\subset H(\mathrm{div},\Omega)^2$ denotes the corresponding finite element subspace, then we obtain, again from [6, Proposition III.3.9],

$$\begin{aligned}&\inf_{\mathbf{V}_h\in\mathbf{\Theta}_h}\left(\|\mathbf{U}-\mathbf{V}_h\|^2_{0,\Omega} + \mu\|\mathrm{div}(\mathbf{U}-\mathbf{V}_h)\|^2_{0,\Omega}\right)^{1/2}\\ &\quad\lesssim h^m\left(|\mathbf{U}|_{m,\Omega} + \mu^{1/2}|\mathrm{div}\,\mathbf{U}|_{m,\Omega}\right) = h^m\left(\mu^{1/2}|\nabla\mathbf{u}|_{m,\Omega} + \mu|\mathbf{\Delta}\mathbf{u}|_{m,\Omega}\right)\\ &\quad= h^m\mu^{1/2}\left(|\mathbf{u}|_{m+1,\Omega} + \mu^{1/2}|\mathbf{\Delta}\mathbf{u}|_{m,\Omega}\right).\end{aligned} \tag{4.4}$$

All this finally leads to

$$\begin{aligned}\left(\sum_{K\in\mathcal{T}_h}\|\mathcal{R}(p_h,\mathbf{u}_h,\mathbf{U}_h)\|^2_{0,K} + \mu\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{u}_h]_E\|^2_{0,E}\right)^{1/2}\\ \lesssim h^{m+1}\left(|\mathbf{u}|_{m+1,\Omega} + |\mathrm{div}\,\mathbf{u}|_{m+1,\Omega} + |p|_{m+2,\Omega}\right)\\ + h^m\left(\mu^{1/2}|\mathbf{u}|_{m+1,\Omega} + \mu|\mathbf{\Delta}\mathbf{u}|_{m,\Omega}\right).\end{aligned} \tag{4.5}$$

In particular, for $m = 1$, i.e., using quadratic conforming elements for Q_h, next-to-lowest order Raviart–Thomas elements for $\mathbf{\Sigma}_h$, and lowest-order Raviart–Thomas elements for $\mathbf{\Theta}_h$,

$$\left(\sum_{K \in \mathcal{T}_h} \|\mathcal{R}(p_h, \mathbf{u}_h, \mathbf{U}_h)\|_{0,K}^2 + \mu \sum_{E \in \mathcal{E}_h} \frac{1}{h_E} \|[\mathbf{t} \cdot \mathbf{u}_h]_E\|_{0,E}^2 \right)^{1/2} \lesssim h^2 \left(|\mathbf{u}|_{2,\Omega} + |\operatorname{div} \mathbf{u}|_{2,\Omega} + |p|_{3,\Omega} \right) + h \left(\mu^{1/2} |\mathbf{u}|_{2,\Omega} + \mu |\mathbf{\Delta u}|_{1,\Omega} \right) \tag{4.6}$$

is obtained for the finite element approximation.

As an alternative for the velocity approximation space $\mathbf{\Sigma}_h$, the finite element space by Mardal, Tai, and Winther [11] is considered. The Mardal–Tai–Winther element was introduced specifically for mixed approaches to Darcy–Stokes flow. In this case, each component of the velocity field is represented by a piecewise polynomial of degree 3,

$$\mathbf{u}_h|_K = \begin{pmatrix} p_3^{(I)} \\ p_3^{(II)} \end{pmatrix}$$

on each triangle $K \in \mathcal{T}_h$, with the restriction that $\operatorname{div} \mathbf{u}_h$ is constant on K and that $\mathbf{n} \cdot \mathbf{u}_h$ coincides with a polynomial of degree 1 on each edge. Continuity of $\mathbf{n} \cdot \mathbf{u}_h$ across edges makes this finite element space $H(\operatorname{div})$-conforming. In addition, the mean value of the tangential component is required to be continuous across edges; in other words,

$$\int_E [\mathbf{t} \cdot \mathbf{u}_h] \, ds = 0$$

for all $E \in \mathcal{E}_h$.

In order to analyze the approximation properties of the Mardal–Tai–Winther element in association with our least-squares formulation, let $\mathbf{\Pi}_h : H_\Gamma(\operatorname{div}, \Omega) \to \mathbf{\Sigma}_h$ be the corresponding interpolation operator (see [11, section 4]). Estimate (4.5) in [11] gives

$$\left(\|\mathbf{u} - \mathbf{\Pi}_h \mathbf{u}\|_{\operatorname{div},\Omega}^2 + \mu \sum_{K \in \mathcal{T}_h} \|\nabla(\mathbf{u} - \mathbf{\Pi}_h \mathbf{u})\|_{0,K}^2 \right)^{1/2} \lesssim h \, |\mathbf{u}|_{2,\Omega}. \tag{4.7}$$

Similarly as in the case of the Raviart–Thomas elements, (3.1) gives

$$\begin{aligned} \|[\mathbf{t} \cdot (\mathbf{\Pi}_h \mathbf{u})]_E\|_{0,E}^2 &= \|[\mathbf{\Pi}_h \mathbf{u}]_E\|_{0,E}^2 = \| \, \mathbf{\Pi}_h \mathbf{u}|_{K_{l,E}} - \mathbf{\Pi}_h \mathbf{u}|_{K_{r,E}} \, \|_{0,E}^2 \\ &\le 2 \| \, \mathbf{\Pi}_h \mathbf{u}|_{K_{l,E}} - \mathbf{\Phi}_h \mathbf{u}\|_{0,E}^2 + 2 \| \, \mathbf{\Pi}_h \mathbf{u}|_{K_{r,E}} - \mathbf{\Phi}_h \mathbf{u}\|_{0,E}^2, \end{aligned}$$

where $\mathbf{\Phi}_h$ again denotes the standard finite element interpolation operator. This leads to

$$\begin{aligned} &\sum_{E \in \mathcal{E}_h} \frac{1}{h_E} \|[\mathbf{t} \cdot (\mathbf{\Pi}_h \mathbf{u})]_E\|_{0,E}^2 \\ &\qquad \le 2 \sum_{E \in \mathcal{E}_h} \frac{1}{h_E} \left(\| \, \mathbf{\Pi}_h \mathbf{u}|_{K_{l,E}} - \mathbf{\Phi}_h \mathbf{u}\|_{0,E}^2 + \| \, \mathbf{\Pi}_h \mathbf{u}|_{K_{r,E}} - \mathbf{\Phi}_h \mathbf{u}\|_{0,E}^2 \right) \end{aligned}$$

$$\lesssim \sum_{E\in\mathcal{E}_h} \frac{1}{h_E^2}\left(\|\,\mathbf{\Pi}_h\mathbf{u}|_{K_{l,E}} - \mathbf{\Phi}_h\mathbf{u}\|_{0,K_{l,E}}^2 + \|\,\mathbf{\Pi}_h\mathbf{u}|_{K_{r,E}} - \mathbf{\Phi}_h\mathbf{u}\|_{0,K_{r,E}}^2\right)$$

$$\lesssim \sum_{K\in\mathcal{T}_h} \frac{1}{h_K^2}\|\mathbf{\Pi}_h\mathbf{u} - \mathbf{\Phi}_h\mathbf{u}\|_{0,K}^2$$

$$\lesssim \sum_{K\in\mathcal{T}_h} \frac{1}{h_K^2}\left(\|\mathbf{u} - \mathbf{\Pi}_h\mathbf{u}\|_{0,K}^2 + \|\mathbf{u} - \mathbf{\Phi}_h\mathbf{u}\|_{0,K}^2\right)$$

$$\lesssim \sum_{K\in\mathcal{T}_h} h_K^2|\mathbf{u}|_{2,K}^2 \lesssim h^2|\mathbf{u}|_{2,\Omega}^2,$$

where we used (4.5) in [11] once more for the interpolation estimate associated with $\mathbf{\Pi}_h$. Combined with (4.7), this implies

$$\inf_{\mathbf{v}_h\in\mathbf{\Sigma}_h}\left(\|\mathbf{u}-\mathbf{v}_h\|_{\mathrm{div},\Omega}^2 + \mu\sum_{K\in\mathcal{T}_h}\|\nabla(\mathbf{u}-\mathbf{v}_h)\|_{0,K}^2 + \mu\sum_{E\in\mathcal{E}_h}\frac{\|[\mathbf{t}\cdot\mathbf{v}_h]_E\|_{0,E}^2}{h_E}\right)^{1/2} \lesssim h\,|\mathbf{u}|_{2,\Omega}. \tag{4.8}$$

If the Mardal–Tai–Winther elements for the velocity approximation space $\mathbf{\Sigma}_h$ are combined with standard conforming linears for Q_h and with lowest-order Raviart–Thomas elements for $\mathbf{\Theta}_h$,

$$\left(\sum_{K\in\mathcal{T}_h}\|\mathcal{R}(p_h,\mathbf{u}_h,\mathbf{U}_h)\|_{0,K}^2 + \mu\sum_{E\in\mathcal{E}_h}\frac{1}{h_E}\|[\mathbf{t}\cdot\mathbf{u}_h]_E\|_{0,E}^2\right)^{1/2} \lesssim h\,(|\mathbf{u}|_{2,\Omega} + |p|_{2,\Omega} + \mu|\mathbf{\Delta}\mathbf{u}|_{1,\Omega}) \tag{4.9}$$

is obtained for the overall approximation of the least-squares finite element approach.

5. Computational results. The numerical tests for our least-squares finite element method are performed for the first-order system (2.2) on the square domain $\Omega = [-1,1]\times[-1,1]$ with

$$g(x_1,x_2) = \begin{cases} 1-x_1^2, & \text{if } x_2 = 1,\\ x_1^2 - 1, & \text{if } x_2 = -1,\\ 0, & \text{elsewhere on } \Gamma, \end{cases}$$

for the boundary conditions. Our interest is in the confirmation of the theoretical estimates for the finite element approximation properties derived in the previous sections of this paper. To this end, the computed values of the least-squares functional $\mathcal{F}(p_h,\mathbf{u}_h,\mathbf{U}_h)$ are shown in the following tables for variable sizes of μ and h. The coarsest triangulation ($l = 0$) consists of 12 triangles, 13 nodes, and 24 edges and is uniformly refined five times, resulting in a finest triangulation ($l = 5$) with 12288 triangles, 6337 nodes, and 18624 edges. The dimensions of the finite element spaces Q_h, $\mathbf{\Sigma}_h$, and $\mathbf{\Theta}_h$ are also given in the tables in order to allow a comparison to the computational effort.

The first two sets of results listed in Tables 5.1 and 5.2 are computed with (next-to-lowest order) Raviart–Thomas elements for the velocity approximation space $\mathbf{\Sigma}_h$. In order to achieve quadratic approximation order for $\mu = 0$, piecewise quadratic standard H^1-conforming elements are used for Q_h. On the other hand, lowest-order

TABLE 5.1
$P_2/RT_2/RT_1^2$: Least squares functional for different values of μ, $\delta = 1$.

	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
$\dim Q_h$	37	121	433	1633	6337	24961
$\dim \boldsymbol{\Sigma}_h$	48	216	912	3744	15168	61056
$\dim \boldsymbol{\Theta}_h$	48	168	624	2400	9408	37248
$\mu = 1$	7.24 e-1	2.49 e-1	6.93 e-2	1.81 e-2	4.61 e-3	1.16 e-3
1 e-1	1.14 e-1	4.61 e-2	1.52 e-2	4.27 e-3	1.13 e-3	2.92 e-4
1 e-2	4.22 e-2	1.91 e-2	1.56 e-2	6.85 e-3	1.97 e-3	5.09 e-4
1 e-3	3.02 e-2	5.70 e-3	4.63 e-3	5.96 e-3	4.01 e-3	1.34 e-3
1 e-4	2.88 e-2	3.47 e-3	7.89 e-4	9.99 e-4	1.65 e-3	1.81 e-3
1 e-5	2.87 e-2	3.23 e-3	3.32 e-4	1.25 e-4	2.05 e-4	3.80 e-4
1 e-6	2.87 e-2	3.21 e-3	2.85 e-4	3.26 e-5	2.24 e-5	4.13 e-5
1 e-7	2.87 e-2	3.20 e-3	2.80 e-4	2.33 e-5	3.76 e-6	4.27 e-6
1 e-8	2.87 e-2	3.20 e-3	2.80 e-4	2.24 e-5	1.89 e-6	5.39 e-7
1 e-9	2.87 e-2	3.20 e-3	2.80 e-4	2.23 e-5	1.71 e-6	1.65 e-7
1 e-10	2.87 e-2	3.20 e-3	2.80 e-4	2.23 e-5	1.69 e-6	1.28 e-7
$\mu = 0$	2.87 e-2	3.20 e-3	2.80 e-4	2.23 e-5	1.69 e-6	1.24 e-7

TABLE 5.2
$P_2/RT_2/RT_1^2$: Least squares functional for different values of μ, $\delta = 0$.

	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
$\dim Q_h$	37	121	433	1633	6337	24961
$\dim \boldsymbol{\Sigma}_h$	48	216	912	3744	15168	61056
$\dim \boldsymbol{\Theta}_h$	48	168	624	2400	9408	37248
$\mu = 1$	7.09 e-1	2.87 e-1	8.87 e-2	2.41 e-2	6.21 e-3	1.57 e-3
1 e-1	1.30 e-1	7.61 e-2	2.92 e-2	8.23 e-3	2.13 e-3	5.37 e-4
1 e-2	4.36 e-2	3.02 e-2	2.90 e-2	1.33 e-2	3.82 e-3	9.85 e-4
1 e-3	2.75 e-2	7.00 e-3	7.78 e-3	1.05 e-2	7.27 e-3	2.46 e-3
1 e-4	2.56 e-2	3.36 e-3	1.13 e-3	1.70 e-3	2.86 e-3	3.19 e-3
1 e-5	2.54 e-2	2.97 e-3	3.49 e-4	1.98 e-4	3.50 e-4	6.54 e-4
1 e-6	2.54 e-2	2.94 e-3	2.70 e-4	3.88 e-5	3.71 e-5	7.08 e-5
1 e-7	2.54 e-2	2.93 e-3	2.62 e-4	2.28 e-5	5.17 e-6	7.24 e-6
1 e-8	2.54 e-2	2.93 e-3	2.61 e-4	2.12 e-5	1.97 e-6	8.32 e-7
1 e-9	2.54 e-2	2.93 e-3	2.61 e-4	2.11 e-5	1.65 e-6	1.90 e-7
1 e-10	2.54 e-2	2.93 e-3	2.61 e-4	2.11 e-5	1.61 e-6	1.26 e-7
$\mu = 0$	2.54 e-2	2.93 e-3	2.61 e-4	2.11 e-5	1.61 e-6	1.19 e-7

Raviart–Thomas elements are sufficient for (each row of) the finite element space $\boldsymbol{\Theta}$ representing the velocity gradient.

Table 5.1 shows the results with this combination of finite element spaces for $\delta = 1$. For $\mu = 0$ quadratic convergence (i.e., the functional behaves like h^4 and, consequently, its square root like h^2) can clearly be observed in the last row of Table 5.1. For $\mu = 1$ it is also obvious that the square root of the functional decreases only proportional to h; i.e., only linear convergence is achieved. For intermediate values of μ it appears that there is an initial phase of almost quadratic convergence, which is then slowed down once the viscosity becomes dominant. Note that the functional even increases for certain values during a refinement step. The possibility that this may happen here is due to the nonconformity of the approach and the fact that the spaces are not nested. A closer inspection shows that the tangential jump term of the least-squares functional is actually responsible for this increase.

Table 5.2 shows the results with the same combination of finite element spaces as above for $\delta = 0$. The convergence behavior is quite similar to the case $\delta = 1$.

The next two sets of results listed in Tables 5.3 and 5.4 are obtained with the Mardal–Tai–Winther (MTW) elements for the approximation of the velocity field. Ta-

TABLE 5.3
$P_2/MTW/RT_1^2$: Least squares functional for different values of μ, $\delta = 1$.

	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
$\dim Q_h$	37	121	433	1633	6337	24961
$\dim \mathbf{\Sigma}_h$	48	204	840	3408	13728	55104
$\dim \mathbf{\Theta}_h$	48	168	624	2400	9408	37248
$\mu = 1$	9.06 e-1	3.09 e-1	8.56 e-2	2.22 e-2	5.63 e-3	1.42 e-3
$1e-1$	1.61 e-1	6.01 e-2	1.86 e-2	5.03 e-3	1.31 e-3	3.35 e-4
$1e-2$	7.56 e-2	3.00 e-2	1.68 e-2	6.53 e-3	1.84 e-3	4.72 e-4
$1e-3$	5.97 e-2	1.61 e-2	7.35 e-3	6.09 e-3	3.51 e-3	1.14 e-3
$1e-4$	5.76 e-2	1.36 e-2	3.63 e-3	1.70 e-3	1.72 e-3	1.61 e-3
$1e-5$	5.74 e-2	1.34 e-2	3.16 e-3	8.57 e-4	3.85 e-4	4.07 e-4
$1e-6$	5.74 e-2	1.33 e-2	3.11 e-3	7.64 e-4	2.07 e-4	8.71 e-5
$1e-7$	5.74 e-2	1.33 e-2	3.11 e-3	7.55 e-4	1.89 e-4	5.06 e-5
$1e-8$	5.74 e-2	1.33 e-2	3.11 e-3	7.54 e-4	1.87 e-4	4.69 e-5
$1e-9$	5.74 e-2	1.33 e-2	3.11 e-3	7.54 e-4	1.86 e-4	4.65 e-5
$1e-10$	5.74 e-2	1.33 e-2	3.11 e-3	7.54 e-4	1.86 e-4	4.65 e-5
$\mu = 0$	5.74 e-2	1.33 e-2	3.11 e-3	7.54 e-4	1.86 e-4	4.65 e-5

TABLE 5.4
$P_2/MTW/RT_1^2$: Least squares functional for different values of μ, $\delta = 0$.

	$l=0$	$l=1$	$l=2$	$l=3$	$l=4$	$l=5$
$\dim Q_h$	37	121	433	1633	6337	24961
$\dim \mathbf{\Sigma}_h$	48	204	840	3408	13728	55104
$\dim \mathbf{\Theta}_h$	48	168	624	2400	9408	37248
$\mu = 1$	6.13 e-1	2.46 e-1	7.38 e-2	1.97 e-2	5.03 e-3	1.27 e-3
$1e-1$	1.18 e-1	6.66 e-2	2.46 e-2	6.82 e-3	1.75 e-3	4.40 e-4
$1e-2$	3.77 e-2	2.86 e-2	2.50 e-2	1.09 e-2	3.12 e-3	8.03 e-4
$1e-3$	1.88 e-2	6.21 e-3	7.38 e-3	9.40 e-3	6.00 e-3	2.01 e-3
$1e-4$	1.63 e-2	2.32 e-3	1.04 e-3	1.63 e-3	2.65 e-3	2.75 e-3
$1e-5$	1.60 e-2	1.90 e-3	2.53 e-4	1.89 e-4	3.42 e-4	6.21 e-4
$1e-6$	1.60 e-2	1.85 e-3	1.72 e-4	3.09 e-5	3.64 e-5	6.99 e-5
$1e-7$	1.60 e-2	1.85 e-3	1.63 e-4	1.48 e-5	4.56 e-6	7.19 e-6
$1e-8$	1.60 e-2	1.85 e-3	1.63 e-4	1.32 e-5	1.35 e-6	7.86 e-7
$1e-9$	1.60 e-2	1.85 e-3	1.62 e-4	1.31 e-5	1.03 e-6	1.45 e-7
$1e-10$	1.60 e-2	1.85 e-3	1.62 e-4	1.31 e-5	1.00 e-6	8.08 e-8
$\mu = 0$	1.60 e-2	1.85 e-3	1.62 e-4	1.31 e-5	9.97 e-7	7.37 e-8

ble 5.3 shows that approximation order 2 is not achieved any longer for zero viscosity with this finite element space. The results for the MTW approximation of the velocity field clearly show a linear convergence behavior for all values of μ (i.e., the functional is proportional to h^2). In general, for $\mu > 0$, the computed values of the functional are always somewhat larger for the Mardal–Tai–Winther elements compared to the Raviart–Thomas elements. It seems that this does not compensate for the slight reduction of the size of the system resulting from the fewer degrees of freedom associated with $\mathbf{\Sigma}_h$. Since the overall approximation order is only 1 for the Mardal–Tai–Winther elements, we may as well use piecewise linears for p. This leads to a further reduction of the number of degrees of freedom and leads to almost the same results as in Table 5.3.

The computed results for the Mardal–Tai–Winther elements in the case $\delta = 0$ are shown in Table 5.4. Interestingly, the results indicate that for $\mu = 0$ the approximation order is significantly faster than linear.

Acknowledgment. We would like to thank Zhiqiang Cai for suggesting the ellipticity bound in the form (3.13), which is stronger than the one in an earlier version of this manuscript.

REFERENCES

[1] P. Bochev, Z. Cai, T. A. Manteuffel, and S. F. McCormick, *Analysis of velocity-flux first-order system least-squares principles for the Navier–Stokes equations: Part* I, SIAM J. Numer. Anal., 35 (1998), pp. 990–1009.

[2] P. B. Bochev and M. D. Gunzburger, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.

[3] P. Bochev, T. A. Manteuffel, and S. F. McCormick, *Analysis of velocity-flux least-squares principles for the Navier–Stokes equations: Part* II, SIAM J. Numer. Anal., 36 (1999), pp. 1125–1144.

[4] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd ed., Cambridge University Press, Cambridge, UK, 2001.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer, New York, 2002.

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

[7] E. Burman and P. Hansbo, *Stabilized Crouzeix–Raviart element for the Darcy–Stokes problem*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 986–997.

[8] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[9] G. Danisch, *Least-Squares Mixed Finite Element Methods for the Shallow Water Equations with Small Viscosity*, Ph.D. thesis, Department of Mathematics, Universität Hannover, Hannover, Germany, 2007.

[10] L. Fontana, E. Miglio, A. Quarteroni, and F. Saleri, *A finite element method for* 3*D hydrostatic water flows*, Comput. Vis. Sci., 2 (1999), pp. 85–93.

[11] K. A. Mardal, X.-C. Tai, and R. Winther, *A robust finite element method for Darcy–Stokes flow*, SIAM J. Numer. Anal., 40 (2002), pp. 1605–1631.

[12] M. Marrocu and D. Ambrosi, *Mesh adaptation strategies for shallow water flow*, Internat. J. Numer. Methods Fluids, 31 (1999), pp. 497–512.

[13] E. Miglio, A. Quarteroni, and F. Saleri, *Finite element approximation of quasi-*3*D shallow water equations*, Comput. Methods Appl. Mech. Engrg., 174 (1999), pp. 355–369.

[14] G. Starke, *A first-order system least-squares finite element method for the shallow water equations*, SIAM J. Numer. Anal., 42 (2005), pp. 2387–2407.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 746–769

ON THE ASYMPTOTIC SPECTRUM OF FINITE ELEMENT MATRIX SEQUENCES*

BERNHARD BECKERMANN† AND STEFANO SERRA-CAPIZZANO‡

Abstract. We derive a new formula for the asymptotic eigenvalue distribution of stiffness matrices obtained by applying P_1 finite elements with standard mesh refinement to the semielliptic PDE of second order in divergence form $-\nabla(K\nabla^T u) = f$ on Ω, $u = g$ on $\partial\Omega$. Here $\Omega \subset \mathbb{R}^2$, and K is supposed to be piecewise continuous and pointwise symmetric semipositive definite. The symbol describing this asymptotic eigenvalue distribution depends on the PDE, but also both on the numerical scheme for approaching the underlying bilinear form and on the geometry of triangulation of the domain. Our work is motivated by recent results on the superlinear convergence behavior of the conjugate gradient method, which requires the knowledge of such asymptotic eigenvalue distributions for sequences of matrices depending on a discretization parameter h when $h \to 0$. We compare our findings with similar results for the finite difference method which were published in recent years. In particular we observe that our sequence of stiffness matrices is part of the class of generalized locally Toeplitz sequences for which many theoretical tools are available. This enables us to derive some results on the conditioning and preconditioning of such stiffness matrices.

Key words. finite element methods, matrix sequence, asymptotic eigenvalue distribution

AMS subject classifications. 65F10, 65N22, 15A18, 15A12, 47B65

DOI. 10.1137/05063533X

1. Introduction and statement of the main results.

Consider the semielliptic PDE of second order in divergence form

$$-\nabla(K\nabla^T u) = f \quad \text{on } \Omega, \qquad u = g \quad \text{on } \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^2$ is a bounded open "smooth" set (say, with piecewise $\mathcal{C}^1$ boundary), and $K : \Omega \mapsto \mathbb{R}^{2\times 2}$ is piecewise continuous in Ω and symmetric semipositive definite at each point of Ω. In this paper we are interested in describing the asymptotic distribution of eigenvalues of the matrix of coefficients obtained by approximating the above elliptic PDE by P_1 finite elements in the case where the position of the vertices can be described by some mapping.

The task of finding the asymptotic eigenvalue distribution is motivated by some recent results on the (superlinear) convergence behavior for the method of conjugate gradients (CG) [4, 5, 6]: a discretization of (1) for some sequence of stepsizes h tending to zero leads to a sequence of systems of linear equations $A_n x_n = b_n$ with A_n some symmetric positive definite matrix of order n, where of course n depends on h and tends to ∞ for $h \to 0$. The CG method is a popular method for solving such systems, and its convergence properties have been analyzed by many authors (see, e.g., [3, 41]). For instance, one has a simple upper bound for the CG error in the energy norm in

*Received by the editors July 6, 2005; accepted for publication (in revised form) November 1, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sinum/45-2/63533.html

†Laboratoire de Mathématiques Paul Painlevé, UMR CNRS 8524, Université des Sciences et Technologies de Lille, F-59655 Villeneuve d'Ascq, France (Bernhard.Beckermann@univ-lille1.fr). This author's work was supported in part by INTAS network NeCCA 03-51-6637, and in part by the grant FAR 2004 of the University of Como.

‡Dipartimento di Fisica e Matematica, Università dell'Insubria, Via Valleggio 11, 22100 Como, Italy (stefano.serrac@uninsubria.it, serra@mail.dm.unipi.it). This author's work was supported in part by MIUR grant 2002014121 and in part by the Department of Mathematics, Université des Sciences et Technologies de Lille.

terms of the spectral condition number of A_n, that is, the ratio of the largest divided by the smallest eigenvalue of A_n; see, e.g., [24, (6.106)]. Both for finite difference and finite element approximations, asymptotics for the smallest eigenvalue of A_n in terms of h and the smallest eigenvalue of the differential operator of (1) are known; see, for instance, [20]. By elementary means one also obtains upper bounds for the largest eigenvalue, and hence upper bounds for the CG error.

However, the (linear) upper bound based on the condition number is usually quite rough, especially in the range of superlinear convergence of CG. This superlinear convergence behavior is observed numerically to be quite pronounced in the context of discretized elliptic problems in ≥ 2 dimensions, in particular for small stepsizes h. Here CG convergence is known to be governed by the distribution of the spectrum $\Lambda(A_n)$ of A_n, which at least for very simple model problems can be computed explicitly. Roughly speaking, superlinear CG convergence occurs if the eigenvalue distribution of A_n is far from being a worst case eigenvalue distribution. This qualitative rule of thumb has been known already for some time, but has been quantified only recently in [4, 5, 6]: here the authors give asymptotic error estimates for CG in terms of the asymptotic eigenvalue distribution of $(A_n)_{n\geq 0}$, namely the so-called *asymptotic spectrum* defined as follows.

A sequence of matrices $(A_n)_{n\geq 0}$, A_n Hermitian of order n with spectrum $\Lambda(A_n) \subset \mathbb{R}$, is said to have an *asymptotic spectrum* given by some measure σ if for all functions $f \in \mathcal{C}_c(\mathbb{R})$ (i.e., continuous with compact support) there holds

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\lambda\in\Lambda(A_n)} f(\lambda) = \int f(\lambda)\, d\sigma(\lambda), \tag{2}$$

where each eigenvalue is counted according to its multiplicity (and hence σ is a probability measure supported on the extended real line $\overline{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$). In the case where the limit (2) exists and takes the form

$$\lim_{n\to\infty} \frac{1}{n} \sum_{\lambda\in\Lambda(A_n)} f(\lambda) = \int_D f(\omega(t))\, \frac{dt}{m(D)} \tag{3}$$

with a domain $D \subset \mathbb{R}^d$ having finite Lebesgue measure $m(D) > 0$, the function ω will be referred to as the *symbol* of (A_n).

The probably most classical example of sequences of matrices having an asymptotic spectrum is given by Hermitian Toeplitz matrices $A_n = (t_{j-k})_{j,k=1,\ldots,n}$ obtained from the Fourier coefficients of the Lebesgue integrable generating function $\omega(s) = \sum_{j\in\mathbb{Z}} t_j e^{ijs}$, $i^2 = -1$; see, for instance, [8] and references therein. Here the symbol coincides with the generating function, and $D = (-\pi, \pi)$.

In the present paper, the matrices A_n will result from the same approximation process when using different (decreasing) stepsizes, and thus one might expect that the sequence of matrices (A_n) has an asymptotic spectrum. Indeed, in case of finite difference discretization for differential operators, explicit formulas for an asymptotic spectrum have been given in [23, 38, 33, 26] (one-dimensional setting) and [31, 32, 30, 28, 35] (two-dimensional and multidimensional setting). Each time, the underlying symbol includes information on the coefficients and the domain of the PDE and information on the discretization schemes for the derivatives. To our knowledge, results for finite element approximations are still lacking (except for some preliminary results in [26, 31]).

Before stating our results on stiffness matrices for finite elements in subsection 1.2, we first recall in subsection 1.1 some known examples of asymptotic spectra in the finite difference case.

1.1. The case of finite difference discretizations. Consider the discretization of the one-dimensional boundary value problem

$$\begin{cases} -\dfrac{d}{dx}\left(k(x)\dfrac{d}{dx}u(x)\right) = f(x), & x \in (0,1), \\ u(0), u(1) \ \text{ given numbers}, \end{cases}$$

on a uniformly spaced grid using centered finite differences of precision order 2 and minimal bandwidth. The resulting linear systems are of tridiagonal type with coefficient matrices (A_n) having entries which are weighted samples of k:

$$A_n = \begin{bmatrix} k_{\frac{1}{2}} + k_{\frac{3}{2}} & -k_{\frac{3}{2}} & & & & \\ -k_{\frac{3}{2}} & k_{\frac{3}{2}} + k_{\frac{5}{2}} & -k_{\frac{5}{2}} & & & \\ & -k_{\frac{5}{2}} & \ddots & \ddots & & \\ & & \ddots & \ddots & -k_{\frac{2n-1}{2}} & \\ & & & -k_{\frac{2n-1}{2}} & k_{\frac{2n-1}{2}} + k_{\frac{2n+1}{2}} & \end{bmatrix}, \tag{4}$$

with $k_t = k(t \cdot h)$, $h = (n+1)^{-1}$. When $k(x) \equiv 1$, the matrix A_n reduces to the Toeplitz matrix $T_n(a)$ of size n,

$$T_n(a) = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & -1 & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ & & & -1 & 2 \end{bmatrix}, \tag{5}$$

generated by $a(s) = 2 - 2\cos(s)$: note that the numbers $-1, 2, -1$ are the (nonzero) Fourier coefficients c_1, c_0, c_{-1} of a and represent also the stencil of the finite difference formula. This latter statement is not a coincidence: if we change the stencil (for instance, in order to obtain more precise discretization schemes), then we obtain Toeplitz matrices generated by a new function a having Fourier coefficients given by the entries of this new stencil [33]. A well-known fact from the theory of Toeplitz matrices is that $(T_n(a))_n$ has an asymptotic spectrum given by $\omega(s) = a(s)$ with $D = [-\pi, \pi]$; see, for instance, the seminal work by Grenander and Szegö [17]. In the more general case of variable coefficients, it follows from the locally Toeplitz analysis of [38] that the matrices A_n of (4) have an asymptotic spectrum given by the symbol

$$\omega(x, s) = k(x)a(s)$$

with $D = (0,1) \times [-\pi, \pi]$ (see also [23]). We observe that the result is in some sense natural since the samplings of k move along the diagonals of A_n smoothly (if k is smooth), and therefore also the algebraic structure of A_n looks like a Toeplitz if we restrict our attention to a local portion of the matrix: this nice algebraic behavior has a natural counterpart in the global spectral behavior. As in the constant coefficient case, the change of the discretization scheme, i.e., of the stencil, will change only

the function a in the symbol (compare [33] and [38]). Finally, we observe that the matrices (A_n) are essentially of the same type as those which one encounters when dealing with sequences of orthogonal polynomials with varying coefficients. Here again locally Toeplitz tools have been used for finding the distribution of the zeros of the considered orthogonal polynomials under very weak assumptions (only measurability) on the regularity of the coefficients [22] (see also [40]).

A further variation which could be considered in the discretization of the above one-dimensional boundary value problem is the use of nonequispaced grids. Indeed, if the new grid of size n is obtained as the image under a map $\phi : [0,1] \mapsto [0,1]$ of a uniform grid of the same size n or if the new grid can be approximated in this way (see, e.g., [35, Definition 4.6]), then the corresponding matrix sequence (A_n) has an asymptotic spectrum described by the symbol

$$\omega(x,s) = \frac{k(\phi(x))}{[\phi'(x)]^2} a(s) \quad \text{with} \quad D = (0,1) \times [-\pi,\pi]. \tag{6}$$

For these results, motivated by the use of collocation techniques (see, e.g., [21]) for approximating the solution of one-dimensional and multidimensional boundary value problems, see [35].

In the case of a two-dimensional problem such as (1), the analysis is also quite complete concerning finite difference approximations. For instance, when $\Omega = (0,1)^2$ and $K = I_2$, using the classical 5 point stencil or the 7 point stencil (in this case there is no difference since $K_{1,2} = K_{2,1} = 0$), we obtain the two-level Toeplitz matrix

$$T_N(b) = T_{n_1}(a) \otimes I_{n_2} + I_{n_1} \otimes T_{n_2}(a), \tag{7}$$

where $N = (n_1, n_2)$ (n_1 is the number of internal points in the x_1 direction and n_2 is the number of internal points in the x_2 direction), $n = n_1 n_2$ is the size, and $b(s_1, s_2) = a(s_1) + a(s_2)$ with $a(s) = 2 - 2\cos(s)$. Also in this case the bivariate stencil represents the nonzero Fourier coefficients of the bivariate generating function b, and this property remains valid for other stencils. Moreover, according to relation (3), the asymptotic spectrum of $(T_N(b))_N$ is described by the symbol $\omega(s_1, s_2) = b(s_1, s_2)$ with $D = [-\pi, \pi]^2$ (see, e.g., [39]). We observe that the same matrix, with $n_1 = n_2 = \nu - 1$, is obtained when employing the P_1 finite element approximation with triangles having the vertices

$$\left(\frac{(j,k)}{\nu}, \frac{(j+\epsilon,k)}{\nu}, \frac{(j,k+\epsilon)}{\nu} \right), \qquad \epsilon = \pm 1. \tag{8}$$

More generally, as a consequence of the theory of generalized locally Toeplitz sequences presented in [31, 32], asymptotic spectra can be given for finite difference approximations of (1) for general functions K and a domain Ω, which guarantees the symmetry of the resulting matrix (e.g., a pluri-rectangle that is a connected finite union of rectangles with edges parallel to the main axes; see [36]). For instance, for a 7 point stencil (see the proof of Corollary 1.2(b) below) we know that the resulting matrix sequence has an asymptotic spectrum with symbol

$$\omega(x,s) = \begin{bmatrix} 1 - e^{is_1} \\ 1 - e^{is_2} \end{bmatrix}^* \cdot K(x) \cdot \begin{bmatrix} 1 - e^{is_1} \\ 1 - e^{is_2} \end{bmatrix}, \tag{9}$$

with $D = \Omega \times [-\pi, \pi]^2$. Notice that if $\Omega = (0,1)^2$ and $K(x) = I_2$, then the above symbol reduces to that of (7) since

$$\begin{bmatrix} 1 - e^{is_1} \\ 1 - e^{is_2} \end{bmatrix}^* \begin{bmatrix} 1 - e^{is_1} \\ 1 - e^{is_2} \end{bmatrix} = |1 - e^{is_1}|^2 + |1 - e^{is_2}|^2 = a(s_1) + a(s_2) = b(s).$$

Furthermore, for nonequispaced tensor grids obtained as the image under a bijective map $\phi(x) = (\phi_1(x_1), \phi_2(x_2))^T$ of an equispaced tensor grid, the general structure of the symbol (see [35, 31]) is the natural generalization of (6): denoting by $\nabla\phi$ the (diagonal) Jacobian of $\phi(x) = (\phi_1(x_1), \phi_2(x_2))^T$, we have

$$\omega(x,s) = \begin{bmatrix} 1-e^{is_1} \\ 1-e^{is_2} \end{bmatrix}^* \cdot \widetilde{K}(x) \cdot \begin{bmatrix} 1-e^{is_1} \\ 1-e^{is_2} \end{bmatrix}, \tag{10}$$

$$\widetilde{K}(x) = \nabla\phi(x)^{-1} K(\phi(x)) \nabla\phi(x)^{-T}$$

over $D = \widetilde{\Omega} \times [-\pi,\pi]^2$, $\widetilde{\Omega} := \phi^{-1}(\Omega)$. We notice that (10) is the natural two-dimensional generalization of (6) and that the symbol in (10) reduces to that in (9) if $\phi_1(x_1) = x_1$ and $\phi_2(x_2) = x_2$, i.e., in the case where the grids are uniform.

Finally, recently the above results have been extended to non-Hermitian matrices A_n occurring, e.g., in the finite difference discretization of PDEs containing lower order difference operators: it has been shown in [16, 18] that, provided that the spectral norm of A_n is uniformly bounded in n and that the trace norm of $S_n = (A_n - A_n^*)/(2i)$, the skew-Hermitian part of A_n, grows at most as $o(n)$, then the sequence (A_n) has the same asymptotic spectrum as the sequence $((A_n + A_n^*)/2)$ obtained from the Hermitian part of A_n. This result also implies [18, 19] that (9) remains true for more general domains Ω, even if one uses different approximation schemes for the boundary conditions.

1.2. The case of finite element approximations. Taking into account the results of the previous subsection, the natural question arises of whether similar results on the asymptotic spectrum hold for matrices obtained by applying finite elements to (1). We mentioned already the well-known fact that for the special case $K = I_2$, $\Omega = (0,1)^2$ and a uniform triangulation on the square such as (8), the stiffness matrix for P_1 elements is identical to that obtained by finite differences using a 5 point stencil. However, this connection is no longer true in the general case and is not sufficient for us to fully understand the asymptotic properties of stiffness matrices, since for finite elements, for instance, a triangulation does not need to be of tensor form.

Rather than developing a general theory, we will discuss in this paper only the example of an approximation of (1) using P_1 finite elements, together with triangulations $\mathcal{T}_\nu$ allowing for some a priori mesh refinement. More specifically, in the following we suppose that we have some $\nu \geq 1$, some open bounded set $\widetilde{\Omega}$, and a triangulation $\mathcal{T}_\nu$ of $\mathrm{Clos}(\Omega)$ with vertices described by a bijective mapping $\phi : \mathrm{Clos}(\widetilde{\Omega}) \mapsto \mathrm{Clos}(\Omega)$ of the form

$$(j/\nu, k/\nu)^T \in \mathrm{Clos}(\widetilde{\Omega}) : \qquad P_{j,k} = \phi((j/\nu, k/\nu)) \tag{11}$$

and triangles

$$(P_{j,k}, P_{j+\epsilon,k}, P_{j,k+\epsilon}), \qquad \epsilon = \pm 1. \tag{12}$$

Such a function ϕ allows us to include also graded triangulations which are suitable if our domain Ω has nonconvex vertices (e.g., for L-shaped domains); see Examples 1.3 and 1.4 below. The usual procedure for solving (the variational form of) (1) via P_1 finite elements (see, e.g., [10, 13]) is to consider for $P_{j,k} \in \Omega$ the hat function $\psi_{j,k}$ being linear on each of the triangles, taking the value 1 on the vertex $P_{j,k}$ and 0 on any other vertex (and thus having a support given by the set of the six triangles which

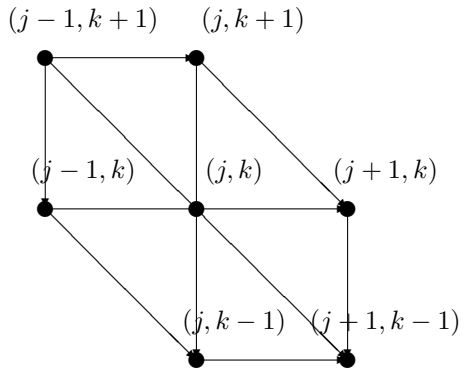


FIG. 1. *The vertex (j,k) and its adjacent vertices for P_1 finite elements.*

share the vertex $P_{j,k}$; see Figure 1), and to solve the system of linear equations

$$A_n x_n = b_n, \qquad A_n = \left(\int_\Omega \nabla \psi_{j,k}(x) K(x) \nabla \psi_{j',k'}(x)^T \, dx \right)_{P_{j,k}, P_{j',k'} \in \Omega} \tag{13}$$

with a suitable right-hand side b_n depending on f and g. The matrix A_n is usually referred to as the stiffness matrix. Notice that the same matrix of coefficients but a different right-hand side is obtained if the Dirichlet boundary conditions are partly replaced by Neumann boundary conditions. In what follows, the letter n will always denote the size of the matrix A_n, i.e., the number of vertices in Ω (which is proportional to ν^2; compare with (18) below).

THEOREM 1.1. *Consider the above triangulation $\mathcal{T}_\nu$ of* $\mathrm{Clos}(\Omega)$ *with vertices* (11) *and triangles* (12). *We suppose that $\phi : \mathrm{Clos}(\widetilde{\Omega}) \mapsto \mathrm{Clos}(\Omega)$ is bijective, $m(\widetilde{\Omega}) > 0$, and that there exists an "exceptional" compact set $\Gamma \subset \mathrm{Clos}(\widetilde{\Omega})$ with $\partial\widetilde{\Omega} \subset \Gamma$ and with Lebesgue measure $m(\Gamma) = 0$ such that $K \circ \phi$ is continuous in $\widetilde{\Omega} \setminus \Gamma$, and ϕ is of class $\mathcal{C}^1$ in $\widetilde{\Omega} \setminus \Gamma$, with nonsingular Jacobian $\nabla\phi$. Then an asymptotic spectrum of the stiffness matrices A_n of* (13) *for $\nu \to \infty$ exists and is given by the formula*

$$\int f \, d\sigma = \frac{1}{(2\pi)^2} \frac{1}{m(\widetilde{\Omega})} \int_{[-\pi,\pi]^2} ds \int_{\widetilde{\Omega}} dx \, f(\omega(x,s)),$$

where

$$\omega(x,s) = \begin{bmatrix} 1 - e^{is_1} \\ 1 - e^{is_2} \end{bmatrix}^* \cdot \widetilde{K}(x) \cdot \begin{bmatrix} 1 - e^{is_1} \\ 1 - e^{is_2} \end{bmatrix},$$

$$\widetilde{K}(x) = |\det \nabla\phi(x)| \nabla\phi(x)^{-1} K(\phi(x)) \nabla\phi(x)^{-T}.$$

Moreover, this formula for the asymptotic spectrum remains valid if one uses numerical integration for evaluating the entries of A_n, as long as the quadrature formula has positive weights and integrates constants exactly.

Some consequences of Theorem 1.1 are summarized in the following result.

COROLLARY 1.2. *With the notations and assumptions of Theorem* 1.1, *the following hold:*

(a) *The sequence of matrices of coefficients (A_n) has the same asymptotic spectrum as the one obtained by applying P_1 elements on the uniform triangulation* (8) *to*

the PDE

$$-\nabla(\widetilde{K}\nabla^T u) = \widetilde{f} \quad \text{on } \widetilde{\Omega}, \qquad u = \widetilde{g} \quad \text{on } \partial\widetilde{\Omega}. \tag{14}$$

Moreover, the bilinear form in the weak formulation of problems (1) *and* (14) *are equivalent via variable transformation.*

(b) *One obtains for* (A_n) *the same asymptotic spectrum as that for matrices obtained by applying finite differences based on a* 7 *point stencil (see Figure* 1*) to* (14). *Moreover,* (A_n) *is a (reduced) generalized locally Toeplitz sequence in the sense of* [32, Definition 3.1], *with the symbol* $\omega(x,s)$ *of Theorem* 1.1.

It is quite instructive to compare the results of Theorem 1.1 and Corollary 1.2 with those of subsection 1.1 for finite difference discretizations. We observe that the symbol in formula (10) and the expression of ω in Theorem 1.1 have a similar structure; in particular, we have the same dependency on the domain Ω and on the matrix-valued coefficient function K. Also, the trigonometric polynomials in s_1, s_2 occurring in Theorem 1.1 are the same as those in (10). These polynomials translate the dependency of the asymptotic spectrum on the discretization scheme (5/7 point stencil or P_1 finite elements). The main difference between the two symbols is the dependency on the triangulation described by our function ϕ: in case of finite elements there is an additional factor $|\det \nabla\phi|$, leading to a smoother symbol in neighborhoods of points $x \in \Gamma$ with $|\det \nabla\phi(x)| = 0$ (corresponding, e.g., to nonconvex edges of Ω; compare with Example 1.3 below), and implying that the finite element matrix of coefficients has fewer eigenvalues of "large" magnitude than the corresponding finite difference matrix of coefficients.

We conclude this section by considering two examples for triangulations $\mathcal{T}_\nu$ induced by some mapping ϕ.

Example 1.3. Suppose that Ω is some nonconvex polygon Ω, with nonconvex vertices given by a_j, $j = 1, \ldots, p$, and corresponding inner angles $\beta_j\pi \in (\pi, 2\pi)$, and let $d > 0$ be sufficiently small. Consider the choice $\Omega = \widetilde{\Omega}$ and

$$\phi(x) = \begin{cases} a_j + (x - a_j) \cdot \left(\frac{||x-a_j||}{d}\right)^{\beta_j - 1} & \text{for } ||x - a_j|| < d, \\ x & \text{else,} \end{cases}$$

where $||\cdot||$ denotes the Euclidean norm. By construction, $\phi : \mathrm{Clos}(\Omega) \mapsto \mathrm{Clos}(\Omega)$ is bijective and of class $\mathcal{C}^1$ in $\widetilde{\Omega} \setminus \Gamma = \{z \in \Omega : ||z - a_j|| \neq d \text{ for } j = 1, 2, \ldots, p\}$. Its Jacobian for $||x - a_j|| < d$ is given by

$$\nabla\phi(x) = \frac{||x - a_j||^{\beta_j - 1}}{d^{\beta_j - 1}} \left[I_2 + (\beta_j - 1)\frac{(x - a_j)(x - a_j)^T}{||x - a_j||^2} \right],$$

and $|\det \nabla\phi(x)| = \beta_j(||x - a_j||/d)^{2\beta_j - 2}$ tends to 0 for $x \to a_j$. For the inverse of the normalized Jacobian occurring in the symbol of Theorem 1.1 we find

$$\sqrt{|\det(\nabla\phi(x))|}\nabla\phi(x)^{-1} = \sqrt{\beta_j}\left[I_2 - \left(1 - \frac{1}{\beta_j}\right)\frac{(x - a_j)(x - a_j)^T}{||x - a_j||^2} \right].$$

Notice also that $||\nabla\phi(x)||$ is bounded uniformly in $\widetilde{\Omega} \setminus \Gamma$, implying that the finesse parameter of the triangulation $\mathcal{T}_\nu$, i.e., the largest of the diameters of the triangles of this triangulation, is of order $\mathcal{O}(1/\nu)$. We finally observe that for triangles where the largest of the distances of the three vertices to a_j is given by $t^{\beta_j} \leq d$ have edges with

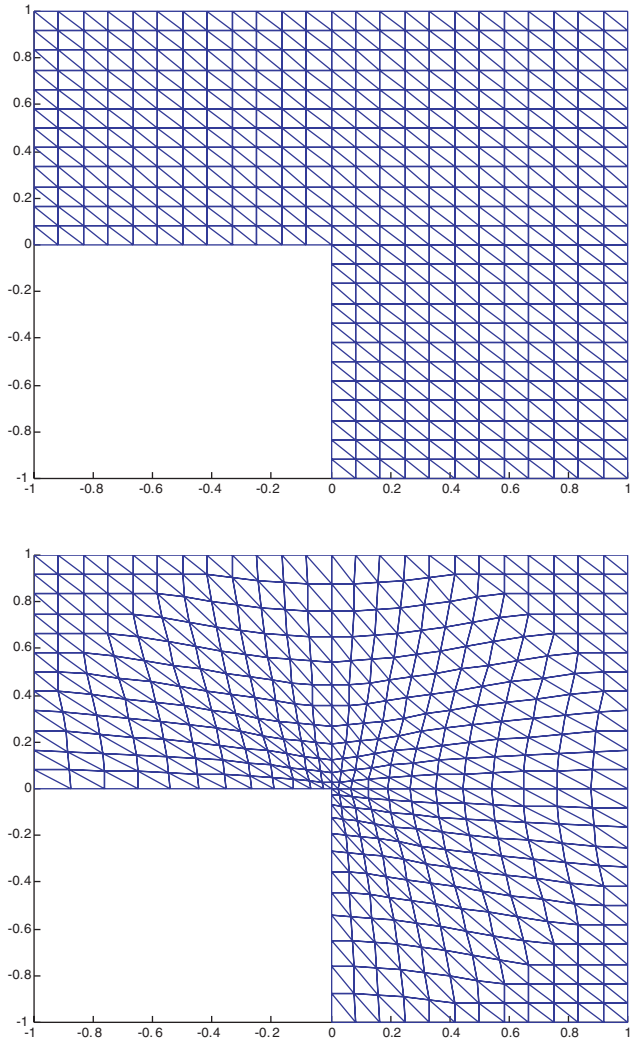


FIG. 2. *Triangulation of an L-shape for* $\nu = 12$. *On the top we find the uniform triangulation, and on the bottom its image under the map* $\phi(x) = x \cdot \min\{1, \sqrt{||x||}\}$ *leading to some grid refinement around the origin.*

size of order $t^{\beta_j - 1}/\nu$: such a mesh refinement based on the grading function $t \mapsto t^{\beta_j}$ is often used in order to keep the classical finite element error estimate also for singular solutions induced by nonconvex vertices.

Example 1.4. A typical example covered by Example 1.3 is a triangulation of an L-shape with vertices $(0,0), (-1,0), (-1,1), (1,1), (1,-1), (0,-1)$, the only nonconvex edge being at the origin $a_1 = 0$, with $\beta_1 := \beta = 3/2$. Here we can choose $d = 1$ in Example 1.3, leading to the function $\phi(x) = x \cdot \min\{1, ||x||^{\beta-1}\}$, with the inverse of the normalized Jacobian given by

$$\sqrt{|\det(\nabla\phi(x))|}\nabla\phi(x)^{-1} = \sqrt{\beta} I_2 - \left(\sqrt{\beta} - \frac{1}{\sqrt{\beta}}\right) \frac{xx^T}{||x||^2}, \qquad ||x|| < 1.$$

In Figure 2 we have drawn both the uniform triangulation and its image under ϕ, leading to some graduation around the origin.

We should notice that in the proof of Theorem 1.1 we do not need any properties of the triangles of $\mathcal{T}_\nu$ having a nonempty intersection with Γ. Thus Theorem 1.1 remains valid if one uses, for instance, curved elements in order to fit more complicated boundaries.

The remainder of the paper is organized as follows: in section 2 we give the proof of Theorem 1.1 and Corollary 1.2. In section 3 we discuss relations between stiffness matrices for different triangulations, in order to design efficient preconditioning strategies. Finally, in section 4 we draw some conclusions.

2. Proof. In what follows we write $\lambda_1(A_n) \leq \lambda_2(A_n) \leq \cdots \leq \lambda_n(A_n)$ for the eigenvalues of some symmetric matrix A_n of order n, and $\mu(A_n) = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(A_n)}$ for the corresponding counting measure. Moreover, we will write $\mu(A_n) \to \sigma$ for $n \to \infty$ if there is weak-star convergence in the sense of (2), i.e., the matrix sequence (A_n) has an asymptotic spectrum described by the measure σ.

For proving the above result we make use of the following result on (reduced) generalized locally Toeplitz matrix sequences (see [31, 32]), which we will not cite in its greatest generality: we will focus instead on a subclass of matrix sequences that are (reduced) generalized locally Toeplitz (see [32, Definition 3.1] for the precise definitions in full generality) and also banded and symmetric. Let (M_n) be a sequence of matrices of size n and of level $\gamma \in \mathbb{N}$ defined according to the multi-index rule

$$M_n = (M_{a,a'})_{a,a' \in \nu D \cap \mathbb{Z}^\gamma}, \tag{15}$$

$$M_{a,a'} = \frac{1}{(2\pi)^\gamma} \int_{[-\pi,\pi]^\gamma} ds e^{-s^T(a'-a)} \omega\left(\frac{a+a'}{2\nu}, s\right),$$

and corresponding to some open $D \subset \mathbb{R}^\gamma$, some integer $\nu \geq 1$, and some symbol $\omega : D \times [-\pi,\pi]^\gamma \to \mathbb{R}$ with $\omega(x,s) = \omega(x,-s)$ being a polynomial in e^{is}, e^{-is} with coefficients continuous in x. We observe that a matrix M_n of such a type and level 1 is just an ordinary banded matrix, where succeeding elements on any diagonal vary only slightly (for large ν and therefore a fortiori for large n) since they are values of some continuous function at arguments differing only by $1/\nu$ (which tends to zero as $n = n(\nu)$ tends to infinity). Also, a matrix M_n of level γ is block banded with blocks being themselves of the same structure as in (15) of level $\gamma - 1$. Finally, if the symbol $\omega(x,s)$ does not depend on x and $D = \bigotimes_{j=1}^{\gamma}(0,\alpha_j)$, we obtain the classical Toeplitz matrices of level γ and order $\prod_{j=1}^{\gamma}[\nu \cdot \alpha_j - 1]$. A basic result on such symmetric banded (reduced) generalized locally Toeplitz matrix sequences is that they have an asymptotic spectrum given by the following formula [31, 32]:

$$\lim_{n\to\infty} \mu(M_n) = \sigma, \qquad \int f\, d\sigma = \frac{1}{(2\pi)^\gamma} \frac{1}{m(D)} \int_{[-\pi,\pi]^\gamma} ds \int_D dx\, f(\omega(x,s)). \tag{16}$$

We will also apply the following statement on the behavior of an asymptotic spectrum under perturbations: the idea relies upon the use of some kind of (matrix) approximation theory for reducing the computation of the symbol of a complicated matrix sequence to the computation of the symbol of simpler matrix sequences (see [29, 31, 32]).

LEMMA 2.1. *Let $A_n \in \mathbb{C}^{n\times n}$ be symmetric, and suppose that there exist probability measures σ, σ' such that, for each $\epsilon > 0$, we may write $A_n = A'_n + A''_n + A'''_n$ with symmetric matrix sequences $A'_n := A'_n(\epsilon), A''_n := A''_n(\epsilon), A'''_n := A'''_n(\epsilon)$, where*

$$\limsup_{n\to\infty} ||A''_n|| \leq \epsilon, \qquad \limsup_{n\to\infty} \frac{\operatorname{rank}(A'''_n)}{n} < \epsilon,$$

and $(A'_n)_n$ *having an asymptotic spectrum* $\mu \le \epsilon\sigma' + \sigma$. *Then* (A_n) *has the asymptotic spectrum* σ.

Proof. Suppose that the assertion of the lemma is not true. Then by Helley's theorem [25, Theorem 0.1.3] there exists an infinite set of natural numbers $\mathcal{N}$ such that $(\mu(A_n))_{n\in\mathcal{N}}$ tends to some probability measure ν different from the probability measure σ. By possibly replacing A_n by $-A_n$ we may conclude that there exists a $b \in \mathbb{R}$ with

$$\nu([-\infty, b)) > \sigma([-\infty, b)) = \sigma([-\infty, b]). \tag{17}$$

Write $r_n = \operatorname{rank}(A'''_n)$. Any $V \subset \mathbb{C}^n$ can be written as direct sum $V' \oplus V''$, V' being a subset of the kernel of A'''_n, V'' being therefore a subset of the image of $(A'''_n)^* = A'''_n$, implying that $\dim(V') \ge \dim(V) - r_n$. Consequently, using the Courant min-max principle, we obtain for any $1 \le j \le n - r_n$

$$\begin{aligned}
\lambda_j(A'_n) &= \max_{V\subset\mathbb{C}^n, \dim(V)=n+1-j} \min_{y\in V} \frac{y^*A'_n y}{y^*y} \\
&\le \max_{V\subset\mathbb{C}^n, \dim(V)=n+1-j} \min_{y\in V} \frac{y^*(A'_n + A''_n)y}{y^*y} + ||A''_n|| \\
&\le \max_{V'\subset \operatorname{Ker}(A'''_n), \dim(V')\ge n+1-j-r_n} \min_{y\in V'} \frac{y^*(A'_n + A''_n)y}{y^*y} + ||A''_n|| \\
&\le \max_{V'\subset\mathbb{C}^n, \dim(V')\ge n+1-j-r_n} \min_{y\in V'} \frac{y^*A_n y}{y^*y} + ||A''_n|| = \lambda_{j+r_n}(A_n) + ||A''_n||.
\end{aligned}$$

Taking into account [25, Theorem 0.1.4], we conclude that

$$\begin{aligned}
\nu([-\infty, b)) &\le \limsup_{n\to\infty} \mu(A_n)([-\infty, b]) = \limsup_{n\to\infty} \frac{\#\{j : \lambda_j(A_n) \le b\}}{n} \\
&\le \limsup_{n\to\infty} \frac{r_n + \#\{j > r_n : \lambda_{j-r_n}(A'_n) \le b + ||A''_n||\}}{n} \\
&\le \epsilon + \limsup_{n\to\infty} \mu(A'_n)([-\infty, b + 2\epsilon]) \le \epsilon + \sigma([-\infty, b + 2\epsilon]).
\end{aligned}$$

For $\epsilon \to 0$, we are left with $\nu([-\infty, b)) \le \sigma([-\infty, b])$, in contradiction with (17). Hence the lemma is shown. ∎

The above lemma is essentially contained in original work by Tilli on (one-level) locally Toeplitz sequences [38] and can be considered an evolution of the low-rank, low-norm splittings used by Tyrtyshnikov [39]. A form which is closer to the present approach can be found in [31], where the main role is played by the symbols of the involved matrix sequences. However, in the present version the language and the tools of Lemma 2.1 are a bit different since the results are expressed in terms of measures (recall formulation (2)) rather than symbols (recall formulation (3)).

Proof of Theorem 1.1. We start by establishing the formula

$$\lim_{\nu\to\infty} \frac{n(\nu)}{\nu^2} = m(\widetilde{\Omega}), \qquad \text{where } n = n(\nu) = \#\left\{\frac{(j,k)}{\nu} \in \widetilde{\Omega}\right\} \tag{18}$$

is the size of the stiffness matrix (13) for the triangulation with parameter ν. For $d > 0$, denote by $\Gamma_d := \{y \in \mathbb{R}^2 : \operatorname{dist}(y, \Gamma) \le d\}$ the closed d-neighborhood of Γ,

where we recall that $\partial\widetilde{\Omega} \subset \Gamma$ by assumption on Γ. For any $\frac{(j,k)}{\nu} \in \widetilde{\Omega}$ we find an open square of Lebesgue measure $1/\nu^2$ being a subset of the $(2/\nu)$-neighborhood of $\widetilde{\Omega}$, any two of such squares having an empty intersection, and thus $n(\nu)/\nu^2 \leq m(\widetilde{\Omega} \cup \Gamma_{2/\nu})$. On the other hand, the set $\widetilde{\Omega} \setminus \Gamma_{2/\nu}$ is a subset of the union of closed squares of Lebesgue measure $1/\nu^2$ centered at $\frac{(j,k)}{\nu} \in \widetilde{\Omega}$, implying that $n(\nu)/\nu^2 \geq m(\widetilde{\Omega} \setminus \Gamma_{2/\nu})$. Taking into account that $m(\Gamma_d) \to m(\Gamma) = 0$ for $d \to 0$ by assumption of Theorem 1.1, we arrive at relation (18).

Let $\epsilon > 0$. We now choose suitable subsets of $\widetilde{\Omega}$. Let $d > 0$ with $m(\widetilde{\Omega} \setminus \Gamma_{3d}) > \left(1 - \frac{\epsilon}{3}\right) m(\widetilde{\Omega})$. By compactness of Γ, we may cover Γ with a finite number of open ∞-neighborhoods $U_d(x_j) = \{y \in \mathbb{R}^2 : ||y - x_j||_\infty < d\}$, $j = 1, \ldots, p$, with $x_j \in \Gamma$. Defining the pluri-rectangles

$$\widetilde{\Omega}' := \widetilde{\Omega} \setminus \bigcup_{j=1}^{p} \mathrm{Clos}(U_{2d}(x_j)), \quad \widetilde{\Omega}'' := \widetilde{\Omega} \setminus \bigcup_{j=1}^{p} U_d(x_j),$$

we find that $\widetilde{\Omega} \setminus \Gamma_{3d} \subset \widetilde{\Omega}' \subset \widetilde{\Omega}'' \subset \widetilde{\Omega} \setminus \Gamma$, with $\widetilde{\Omega}''$ being compact, $\widetilde{\Omega}'$ being open, and

(19)
$$\lim_{\nu\to\infty} \frac{n'(\nu)}{\nu^2} = m(\widetilde{\Omega}') \geq \left(1 - \frac{\epsilon}{3}\right) m(\widetilde{\Omega}), \qquad \text{where } \; n' = n'(\nu) = \#\left\{\frac{(j,k)}{\nu} \in \widetilde{\Omega}'\right\}.$$

Thus, for sufficiently large ν,

$$\frac{n'(\nu)}{n(\nu)} > 1 - \frac{\epsilon}{2}. \tag{20}$$

We are now prepared to introduce a suitable splitting of the stiffness matrix A_n of (13): we first apply a suitable simultaneous permutation of row and columns such that the first $n'(\nu)$ rows and columns of A_n correspond to indices with $(j,k)/\nu \in \widetilde{\Omega}'$. Then the matrix A_n''' defined by

$$A_n - A_n''' = \begin{bmatrix} \widetilde{A}_n & 0 \\ 0 & 0 \end{bmatrix}, \quad \widetilde{A}_n = \left(\int_\Omega \nabla\psi_{j,k}(x) K(x) \nabla\psi_{j',k'}(x)^T \, dx\right)_{(j,k)/\nu,(j',k')/\nu \in \widetilde{\Omega}'}$$

is symmetric and has a rank bounded above by twice the difference of the order $n = n(\nu)$ of A_n minus the order $n' = n'(\nu)$ of $\widetilde{A}_n$. A combination with (20) leads to the relations

$$(A_n''')^* = A_n''', \qquad \mathrm{rank}\,(A_n''') \leq \epsilon n. \tag{21}$$

We want to apply Lemma 2.1 via a splitting $\widetilde{A}_n = \widetilde{A}_n' + \widetilde{A}_n''$, and

$$A_n = A_n' + A_n'' + A_n''', \quad A_n' = \begin{bmatrix} \widetilde{A}_n' & 0 \\ 0 & 0 \end{bmatrix}, \quad A_n'' = \begin{bmatrix} \widetilde{A}_n'' & 0 \\ 0 & 0 \end{bmatrix}, \tag{22}$$

where $\widetilde{A}_n''$ will be a symmetric matrix of small norm, and $\widetilde{A}_n'$ symmetric and banded. Moreover, $(\widetilde{A}_n')$ will be (reduced) generalized locally Toeplitz of level 2 in the sense of (15), and thus we know the existence and the explicit form of the asymptotic spectrum of $(\widetilde{A}_n')$ for $\nu \to \infty$.

We make use of the classical assembling procedure of a P_1 finite element matrix A_n: starting from the zero matrix, the stiffness matrix A_n is obtained after applying for all triangles T of the form $(P_{j,k}, P_{j+\eta,k}, P_{j,k+\eta})$, $\eta = \pm 1$, the updating formula

(23)

$$A_n \begin{pmatrix} (j,k),(j+\eta,k),(j,k+\eta) \\ (j,k),(j+\eta,k),(j,k+\eta) \end{pmatrix}$$
$$\leftarrow A_n \begin{pmatrix} (j,k),(j+\eta,k),(j,k+\eta) \\ (j,k),(j+\eta,k),(j,k+\eta) \end{pmatrix} + \frac{1}{2|\det(C^{-1})|} B^T C^{-1} \frac{\int_T K(x)\,dx}{\int_T dx} C^{-T} B,$$

where the affine mapping $x \mapsto P_{j,k} + Cx$ brings the points $(0,0),(1,0),(0,1)$ to $P_{j,k}, P_{j+\eta,k}, P_{j,k+\eta}$, respectively, and

$$B = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

An important observation in our proof is that the updating term in (23) behaves like $\frac{1}{2} B^T \widetilde{K}(\zeta) B$ for some $\zeta \in \phi^{-1}(T)$ for "most" triangles T. In order to make this claim more precise in (24) below, we notice that, by construction, $\widetilde{\Omega}''$ is a compact subset of $\widetilde{\Omega} \setminus \Gamma$, and hence the Jacobian $\nabla\phi$ of ϕ, its inverse $\nabla\phi(x)^{-1}$, and the function $K \circ \phi$ are uniformly continuous in $\widetilde{\Omega}''$. Let

$$M := \sup_{x \in \widetilde{\Omega}''} \max \left\{ ||\nabla\phi(x)||, ||\nabla\phi(x)^{-1}||, ||K(\phi(x))||, \sqrt{2\epsilon} \right\} \geq 1,$$

and choose ν sufficiently large such that a triangle having at least one vertex in $\widetilde{\Omega}'$ is a subset of $\widetilde{\Omega}''$, and that any of the above functions varies at most by $\epsilon/(4M^5)$ by choosing two arguments in any triangle that is a subset of $\widetilde{\Omega}''$. For the matrix $\widetilde{A}_n$ we need to consider only triangles T having at least one vertex with preimage in $\widetilde{\Omega}'$. Denoting by $\widetilde{T} \subset \widetilde{\Omega}''$ the corresponding triangle with vertices $\frac{(j,k)}{\nu}, \frac{(j+\eta,k)}{\nu}, \frac{(j,k+\eta)}{\nu}$, we may conclude with help of the mean value theorem that, for any $\zeta \in \widetilde{T}$,

$$\left\| \frac{\int_T K(x)\,dx}{\int_T dx} - K(\phi(\zeta)) \right\| \leq \frac{\epsilon}{4M^5} \leq M, \qquad \left\| \frac{\nu}{\eta} C - \nabla\phi(\zeta) \right\| \leq \frac{\epsilon}{M^5} \leq \frac{1}{2||\nabla\phi(\zeta)^{-1}||},$$

and hence

$$\left\| \left(\frac{\nu}{\eta} C \right)^{-1} - \nabla\phi(\zeta)^{-1} \right\| \leq \frac{2\epsilon}{M^3} \leq M, \qquad \left\| \det\left(\frac{\nu}{\eta} C \right) - \det(\nabla\phi(\zeta)) \right\| \leq \frac{4\epsilon}{M^4} \leq M.$$

Applying the triangular inequality several times, we obtain after some elementary computations the (quite rough) estimate

$$\max_{\zeta \in \widetilde{T}} \left\| \frac{1}{|\det(C^{-1})|} C^{-1} \frac{\int_T K(x)\,dx}{\int_T dx} C^{-T} - \widetilde{K}(\zeta) \right\| \leq 80\epsilon, \tag{24}$$

with $\widetilde{K}$ as in the statement of Theorem 1.1. We remark that the same conclusion holds if instead of exact integration one uses a quadrature formula with positive weights for

TABLE 1

The six adjacent vertices of $\frac{(j,k)}{\nu} \in \widetilde{\Omega}'$ and the corresponding off-diagonal entries of $\widetilde{A}'_n$: in the first column we find the index (j',k') of an adjacent vertex, in the second and third column the index of the third vertex of the two triangles giving a nontrivial contribution to the entry in row (j,k) and column (j',k') of A_n, and in the last column the entry of $\widetilde{A}'_n$ at the same position.

(j',k')	(j'',k'')	(j''',k''')	Corresponding entry of $\widetilde{A}'_n$
$(j-1,k)$	$(j,k-1)$	$(j-1,k+1)$	$B_1^T\,\widetilde{K}(\frac{(j+j',k+k')}{2\nu})\,B_2$
$(j,k-1)$	$(j+1,k-1)$	$(j-1,k)$	$B_1^T\,\widetilde{K}(\frac{(j+j',k+k')}{2\nu})\,B_3$
$(j+1,k-1)$	$(j+1,k)$	$(j,k-1)$	$B_2^T\,\widetilde{K}(\frac{(j+j',k+k')}{2\nu})\,B_3$
$(j+1,k)$	$(j,k+1)$	$(j+1,k-1)$	$B_1^T\,\widetilde{K}(\frac{(j+j',k+k')}{2\nu})\,B_2$
$(j,k+1)$	$(j-1,k+1)$	$(j+1,k)$	$B_1^T\,\widetilde{K}(\frac{(j+j',k+k')}{2\nu})\,B_3$
$(j-1,k+1)$	$(j-1,k)$	$(j,k+1)$	$B_2^T\,\widetilde{K}(\frac{(j+j',k+k')}{2\nu})\,B_3$

the entries of the stiffness matrix, provided that this quadrature formula integrates constants exactly.

Notice that, in the updating procedure (23), an off-diagonal entry of A_n is updated twice since a fixed edge of the triangulation is adjacent to two triangles, and a diagonal entry is updated six times since there are six triangles adjacent to a vertex; compare with Figure 1. More precisely, in row labeled (j,k), the matrix $\widetilde{A}_n$ has nonzero off-diagonal entries in columns labeled

$$(j',k') \in \{(j-1,k+1),(j,k+1),(j-1,k),(j+1,k),(j,k-1),(j+1,k-1)\},$$

i.e., the indices of adjacent vertices. For instance, for the entry in column $(j',k') = (j-1,k)$ we have to consider the two triangles T with third vertex labeled $(j'',k'') = (j,k-1)$, and $(j''',k''') = (j-1,k+1)$, respectively, and the corresponding updating quantities can be found at position $(1,2)$ and $(2,1)$, respectively, of the symmetric 3×3 updating matrix on the right-hand side of (23). Thus, defining the corresponding off-diagonal entry of $\widetilde{A}'_n$ by

$$\begin{aligned}\widetilde{A}'_n\begin{pmatrix}(j',k')\\(j,k)\end{pmatrix} &= B_1^T\widetilde{K}\left(\frac{1}{2}\left(\frac{(j,k)}{\nu}+\frac{(j',k')}{\nu}\right)\right)B_2\\ &= (-1,-1)\widetilde{K}\left(\frac{(j+j',k+k')}{2\nu}\right)(1,0)^T,\end{aligned}$$

B_ℓ denoting the ℓth column of B, we find according to (24) that

$$\left|\widetilde{A}'_n\begin{pmatrix}(j',k')\\(j,k)\end{pmatrix}-\widetilde{A}_n\begin{pmatrix}(j',k')\\(j,k)\end{pmatrix}\right| \le 80\epsilon\,||B||^2 = 240\epsilon.$$

The off-diagonal entries of $\widetilde{A}'_n$ for the other five adjacent vertices (j',k') of (j,k) are given in Table 1, and each time we obtain the same estimate for the off-diagonal entries of $\widetilde{A}'_n - \widetilde{A}_n$. We define the diagonal entries of $\widetilde{A}'_n$ by

$$\begin{aligned}\widetilde{A}'_n\begin{pmatrix}(j,k)\\(j,k)\end{pmatrix} &= \operatorname{trace}\Big(B^T\widetilde{K}\left(\frac{(j,k)}{\nu}\right)B\Big)\\ &= -2\left(B_1^T\widetilde{K}\left(\frac{(j,k)}{\nu}\right)B_2+B_1^T\widetilde{K}\left(\frac{(j,k)}{\nu}\right)B_3+B_2^T\widetilde{K}\left(\frac{(j,k)}{\nu}\right)B_3\right)\end{aligned}\qquad (25)$$

and find according to (24) that

$$\left|\widetilde{A}'_n\begin{pmatrix}(j,k)\\(j,k)\end{pmatrix}-\widetilde{A}_n\begin{pmatrix}(j,k)\\(j,k)\end{pmatrix}\right|\le 240\epsilon\,||B||^2=720\epsilon,$$

and thus, by (22),

$$||A''_n||=||\widetilde{A}_n-\widetilde{A}'_n||\le\sqrt{||\widetilde{A}_n-\widetilde{A}'_n||_1||\widetilde{A}_n-\widetilde{A}'_n||_\infty}\le(6\cdot 240+720)\epsilon=2160\,\epsilon.$$

It remains to analyze $\widetilde{A}'_n$. Comparing the definition (15) with the last column of Table 1 and with (25), we see that $(\widetilde{A}'_n)$ is a banded and symmetric generalized locally Toeplitz matrix sequence of level 2 corresponding to the domain $\widetilde{\Omega}'$ and the symbol

$$\begin{aligned}\omega(x,s)&=(2\cos(s_1)-2)\,B_1^T\widetilde{K}(x)B_2+(2\cos(s_2)-2)\,B_1^T\widetilde{K}(x)B_3\\&\quad+(2\cos(s_2-s_1)-2)\,B_2^T\widetilde{K}(x)B_3\\&=4\sin^2\left(\frac{s_1}{2}\right)\widetilde{K}_{1,1}(x)+4\sin^2\left(\frac{s_2}{2}\right)\widetilde{K}_{2,2}(x)\\&\quad+4\left[\sin^2\left(\frac{s_1}{2}\right)+\sin^2\left(\frac{s_2}{2}\right)-\sin^2\left(\frac{s_2-s_1}{2}\right)\right]\widetilde{K}_{1,2}(x),\end{aligned}$$

that is, the same symbol (but a different domain) as in the statement of Theorem 1.1. Using (16), we may conclude that $(\mu(\widetilde{A}'_n))$ has the limit $\widetilde{\sigma}$, with

$$\int f\,d\widetilde{\sigma}=\frac{1}{(2\pi)^2}\frac{1}{m(\widetilde{\Omega}')}\int_{[-\pi,\pi]^2}ds\int_{\widetilde{\Omega}'}dx\,f(\omega(x,s)).$$

According to (22), for the corresponding counting measures for $\nu\to\infty$, we get using (18), (19),

$$\mu(A'_n)=\frac{n(\nu)-n'(\nu)}{n(\nu)}\cdot\delta_0+\frac{n'(\nu)}{n(\nu)}\mu(\widetilde{A}'_n)\to\frac{m(\widetilde{\Omega})-m(\widetilde{\Omega}')}{m(\widetilde{\Omega})}\cdot\delta_0+\frac{m(\widetilde{\Omega}')}{m(\widetilde{\Omega})}\widetilde{\sigma}$$

and

$$\frac{m(\widetilde{\Omega})-m(\widetilde{\Omega}')}{m(\widetilde{\Omega})}\cdot\delta_0+\frac{m(\widetilde{\Omega}')}{m(\widetilde{\Omega})}\widetilde{\sigma}\le\epsilon\cdot\delta_0+\frac{m(\widetilde{\Omega}')}{m(\widetilde{\Omega})}\widetilde{\sigma}\le\epsilon\cdot\delta_0+\sigma,$$

since $\widetilde{\sigma}$ differs from σ by using a different normalization and a smaller set of integration $\widetilde{\Omega}'\subset\widetilde{\Omega}$. Thus we may apply Lemma 2.1, giving the asymptotic spectrum for (A_n) as claimed in Theorem 1.1. □

Proof of Corollary 1.2. The first sentence of part (a) follows immediately by applying the formulas of Theorem 1.1 twice. With respect to the second one, consider the variable transformation $x=\phi(\widetilde{x})$ in (1): with $\widetilde{f}(\widetilde{x})=f(\phi(\widetilde{x}))$, we have $\widetilde{\nabla}\widetilde{f}(\widetilde{x})=(\nabla f)(\phi(\widetilde{x}))\nabla\phi(\widetilde{x})$, and hence

$$\begin{aligned}&\int_\Omega(\nabla u)(x)K(x)(\nabla v)(x)^T\,dx\\&\quad=\int_{\widetilde{\Omega}}(\widetilde{\nabla}\widetilde{u})(\widetilde{x})\nabla\phi(\widetilde{x})^{-1}K(\phi(x))\nabla\phi(\widetilde{x})^{-T}(\widetilde{\nabla}\widetilde{v})(\widetilde{x})^T|\det\nabla\phi(\widetilde{x})|d\widetilde{x}\\&\quad=\int_{\widetilde{\Omega}}(\widetilde{\nabla}\widetilde{u})(\widetilde{x})\widetilde{K}(\widetilde{x})(\widetilde{\nabla}\widetilde{v})(\widetilde{x})^T\,d\widetilde{x}.\end{aligned}$$

For a proof of part (b), we consider

$$y_\nu = (u_{j,k})_{(j,k)/\nu\in\widetilde{\Omega}'}, \qquad \widetilde{u}_{j,k} \approx u\left(\frac{(j,k)}{\nu}\right)$$

and the second order central difference operators using the 7 point stencil of Figure 1,

$$\Delta_1 u_{j,k} = u_{j+1/2,k} - u_{j-1/2,k} \approx \frac{1}{\nu}\frac{\partial}{\partial\widetilde{x}_1}u\left(\frac{(j,k)}{\nu}\right),$$

$$\Delta_2 u_{j,k} = u_{j,k+1/2} - u_{j,k-1/2} \approx \frac{1}{\nu}\frac{\partial}{\partial\widetilde{x}_2}u\left(\frac{(j,k)}{\nu}\right),$$

$$\Delta_3 u_{j,k} = u_{j+1/2,k-1/2} - u_{j-1/2,k+1/2} \approx \frac{1}{\nu}\left(\frac{\partial}{\partial\widetilde{x}_1} - \frac{\partial}{\partial\widetilde{x}_2}\right)u\left(\frac{(j,k)}{\nu}\right).$$

Let $\widetilde{\Omega}'$ and $\widetilde{A}'_n$ be as in the preceding proof, and let C_n be obtained from the matrix $\widetilde{A}'_n$ by replacing the diagonal entries (25) by

$$\begin{aligned} C_n\begin{pmatrix}(j,k)\\(j,k)\end{pmatrix} = &-B_1^T\left(\widetilde{K}\left(\frac{(2j-1,2k)}{2\nu}\right) + \widetilde{K}\left(\frac{(2j+1,2k)}{2\nu}\right)\right)B_2 \\ &- B_1^T\left(\widetilde{K}\left(\frac{(2j,2k-1)}{2\nu}\right) + \widetilde{K}\left(\frac{(2j,2k+1)}{2\nu}\right)\right)B_3 \\ &-B_2^T\left(\widetilde{K}\left(\frac{(2j+1,2k-1)}{2\nu}\right) + \widetilde{K}\left(\frac{(2j-1,2k+1)}{2\nu}\right)\right)B_3, \end{aligned}$$

and hence $||\widetilde{A}'_n - C_n||$ is of order ϵ; compare with (24). For a grid point $\frac{(j,k)}{\nu} \in \widetilde{\Omega}'$ having all its adjacent vertices in $\widetilde{\Omega}'$, the component of $C_n y_\nu$ with index (j,k) can be written as

$$\begin{aligned} &[\widetilde{K}_{1,1} + \widetilde{K}_{1,2}]\left(\frac{(2j-1,2k)}{2\nu}\right)(u_{j,k} - u_{j-1,k}) \\ &\quad + [\widetilde{K}_{1,1} + \widetilde{K}_{1,2}]\left(\frac{(2j+1,2k)}{2\nu}\right)(u_{j,k} - u_{j+1,k}) \\ &\quad + [\widetilde{K}_{2,2} + \widetilde{K}_{1,2}]\left(\frac{(2j,2k-1)}{2\nu}\right)(u_{j,k} - u_{j,k-1}) \\ &\quad + [\widetilde{K}_{2,2} + \widetilde{K}_{2,1}]\left(\frac{(2j,2k+1)}{2\nu}\right)(u_{j,k} - u_{j,k+1}) \\ &\quad + \widetilde{K}_{1,2}\left(\frac{(2j+1,2k-1)}{2\nu}\right)(u_{j+1,k-1} - u_{j,k}) \\ &\quad + \widetilde{K}_{1,2}\left(\frac{(2j-1,2k+1)}{2\nu}\right)(u_{j-1,k+1} - u_{j,k}) \\ &= -\Delta_1[\widetilde{K}_{1,1} + \widetilde{K}_{1,2}]\Delta_1 u_{j,k} - \Delta_2[\widetilde{K}_{2,2} + \widetilde{K}_{1,2}]\Delta_2 u_{j,k} + \Delta_3\widetilde{K}_{1,2}\Delta_3 u_{j,k}. \end{aligned}$$

If some of the vertices $\frac{(j',k')}{\nu}$ adjacent to $\frac{(j,k)}{\nu}$ lie outside of $\widetilde{\Omega}'$, we get a similar expression, where the corresponding values $u_{j',k'}$ have to be dropped. Therefore the matrix C_n describes a finite difference discretization in $\widetilde{\Omega}'$ based on the 7 point stencil of Figure 1 for the differential expression

$$-\frac{\partial}{\partial\widetilde{x}_1}\left([\widetilde{K}_{1,1}+\widetilde{K}_{1,2}]\frac{\partial u}{\partial\widetilde{x}_1}\right)-\frac{\partial}{\partial\widetilde{x}_2}\left([\widetilde{K}_{2,2}+\widetilde{K}_{1,2}]\frac{\partial u}{\partial\widetilde{x}_2}\right)$$
$$+\left(\frac{\partial}{\partial\widetilde{x}_1}-\frac{\partial}{\partial\widetilde{x}_2}\right)\left(\widetilde{K}_{1,2}\left(\frac{\partial u}{\partial\widetilde{x}_1}-\frac{\partial u}{\partial\widetilde{x}_2}\right)\right),$$

coinciding with $-\nabla(\widetilde{K}\nabla u)$, the differential expression of the PDE of Corollary 1.2(a). Using the same limit considerations as in the proof of Theorem 1.1, the first assertion of Corollary 1.2(b) follows. The second assertion now is a simple consequence of the above relationship between A_n and the 7 point stencil finite difference matrix and of the fact that every finite difference discretization of second order PDEs leads to (reduced) generalized locally Toeplitz sequences (see [31, 32]). □

3. Uniform versus nonuniform triangulations and preconditioning. Let us briefly recall some classical terminology concerning finite element triangulations. The *finesse parameter* of a triangulation $\mathcal{T}_\nu$ is the largest among the diameters of the triangles of this triangulation. A family of triangulations $\mathcal{T}_\nu$ for varying ν is called quasi-uniform [2, 20] if the length of the shortest edge in $\mathcal{T}_\nu$ divided by the finesse parameter of $\mathcal{T}_\nu$ is bounded below by some constant uniformly in ν. The family of triangulations $\mathcal{T}_\nu$ is called *shape-regular* [9, Definition II.5.1] if the ratio of the diameter divided by the radius of the largest inscribed disk is bounded uniformly for each triangle $T\in\mathcal{T}_\nu$ and all ν (or, equivalently, if all angles are bounded away from zero uniformly in ν).

In the previous sections we have considered a triangulation $\mathcal{T}_\nu$ of Ω being the image under a bijective map ϕ of a uniform triangulation $\widetilde{\mathcal{T}}_\nu$ of $\widetilde{\Omega}$ with stepsize $1/\nu$. Denote by $A_n(K,\mathcal{T}_\nu)$ the corresponding stiffness matrix (13). Since in general the two triangulations $\mathcal{T}_\nu$ and $\widetilde{\mathcal{T}}_\nu$ lead to stiffness matrices of the same size, we want to discuss in this section in more detail some spectral properties of the matrix $A_n(I_2,\widetilde{\mathcal{T}}_\nu)^{-1}A_n(K,\mathcal{T}_\nu)$ and other related matrices. This analysis is motivated by the task of finding efficient preconditioning strategies for the method of conjugate gradients applied to the stiffness matrix $A_n(K,\mathcal{T}_\nu)$. Our uniform triangulation $(\widetilde{\mathcal{T}}_\nu)_\nu$ is trivially both quasi-uniform and shape-regular, while $(\mathcal{T}_\nu)_\nu$ is not necessarily so. For instance, for the graduated mesh of Example 1.3 we find a finesse parameter $\geq 1/\nu$, but the triangle with vertex a_j has edges of size $d(1/(d\nu))^{\beta_j}$, and hence $(\mathcal{T}_\nu)_\nu$ is not quasi-uniform. In this section we will be particularly interested in the case where $(\mathcal{T}_\nu)_\nu$ is only shape-regular.

The main results of this section are given in subsection 3.2: in Theorem 3.2 we first relate two stiffness matrices with respect to the partial ordering of Hermitian matrices ($M_1\leq M_2$ if M_1,M_2 are Hermitian and M_2-M_1 is semipositive definite). Subsequently, in Corollary 3.4 we deduce bounds for the smallest and the largest eigenvalue of such preconditioned stiffness matrices, and in Theorem 3.5 we give results on the asymptotic spectrum for such matrices. But first we provide in subsection 3.1 a basic proposition (based on the local analysis of finite element matrices), which is the keystone for proving the results of subsection 3.2.

3.1. Local domain analysis of the finite element matrices. In order to better understand the local properties of a stiffness matrix, let us go back to the classical assembling procedure of a P_1 finite element matrix A_n mentioned already in the proof of Theorem 1.1. Starting from the zero matrix, we have the following updating formulas: any triangle T of the form $(P_{j,k}, P_{j+\eta,k}, P_{j,k+\eta})$, $\eta \in \{\pm 1\}$, gives the contribution

(26)

$$A_n \begin{pmatrix} (j,k),(j+\eta,k),(j,k+\eta) \\ (j,k),(j+\eta,k),(j,k+\eta) \end{pmatrix}$$
$$\leftarrow A_n \begin{pmatrix} (j,k),(j+\eta,k),(j,k+\eta) \\ (j,k),(j+\eta,k),(j,k+\eta) \end{pmatrix} + \frac{1}{2|\det(C^{-1})|} B^T C^{-1} \frac{\int_T K(x)\,dx}{\int_T dx} C^{-T} B,$$

where the affine mapping $x \mapsto P_{j,k} + Cx$ maps the points $(0,0),(1,0),(0,1)$ to $P_{j,k}, P_{j+\eta,k}, P_{j,k+\eta}$, respectively, and

$$B = \begin{bmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{bmatrix}.$$

Suppose that the three points $(P_{j,k}, P_{j+\eta,k}, P_{j,k+\eta})$ have positive orientation, and define by α, β, γ, respectively, the angles of the triangle T at these vertices. In addition, define Π to be a rotation matrix mapping the half line $(0, P_{j+\eta,k} - P_{j,k})$ to the half line $((0,0),(1,0))$; then

$$\Pi C = \frac{||P_{j+\eta,k} - P_{j,k+\eta}||}{\sin(\alpha)} \begin{bmatrix} \sin(\gamma) & \sin(\beta)\cos(\alpha) \\ 0 & \sin(\beta)\sin(\alpha) \end{bmatrix},$$

and, in addition,

$$\frac{C^{-1}}{\sqrt{|\det(C^{-1})|}} = \frac{1}{\sqrt{\sin(\alpha)\sin(\beta)\sin(\gamma)}} \begin{bmatrix} \sin(\alpha)\sin(\beta) & -\cos(\alpha)\sin(\beta) \\ 0 & \sin(\gamma) \end{bmatrix} \cdot \Pi.$$

Observe also that $C^{-1}/\sqrt{|\det(C^{-1})|}$ has the singular values $\sqrt{\delta_T}$ and $1/\sqrt{\delta_T}$ and thus a spectral condition number δ_T, which can be computed explicitly in terms of the angles of T:

(27)

$$\delta_T := \operatorname{cond}\left(\frac{C^{-1}}{\sqrt{|\det(C^{-1})|}}\right) = y_T + \sqrt{y_T^2 - 1}, \qquad y_T = \frac{\sin^2(\beta) + \sin^2(\gamma)}{2\sin(\alpha)\sin(\beta)\sin(\gamma)}.$$

Therefore

$$\frac{1}{\delta_T} I_2 \le \frac{1}{|\det(C^{-1})|} C^{-1} C^{-T} \le \delta_T I_2. \tag{28}$$

If the three points $(P_{j,k}, P_{j+\eta,k}, P_{j,k+\eta})$ have negative orientation, then we switch axes; that is, we exchange the role of β and γ, but the conclusions in (27) and (28) are the same. For instance, for a triangle $T \in \widetilde{\mathcal{T}}_\nu$ of a uniform triangulation we get $\alpha = \pi/2$ and $\beta = \gamma = \pi/4$, leading to $\delta_T = 1$, but in general $\delta_T \ge 1$.

The relation (28) enables us to compare the updating matrices in (26) for different meshes and $K = I_2$, and, by a similar argument, for different (pointwise symmetric positive definite) coefficient functions K.

PROPOSITION 3.1. *With*

$$\kappa_{\min} = \operatorname{essinf}_{x\in T}\lambda_{\min}(K(x)) \geq 0, \qquad \kappa_{\max} = \operatorname{esssup}_{x\in T}\lambda_{\max}(K(x)),$$

and B, C as in (26) *we have that*

$$\kappa_{\min}\frac{B^TC^{-1}\,C^{-T}B}{2|\det(C^{-1})|} \leq \frac{1}{2|\det(C^{-1})|}B^TC^{-1}\frac{\int_T K(x)\,dx}{\int_T dx}C^{-T}B \leq \kappa_{\max}\frac{B^TC^{-1}\,C^{-T}B}{2|\det(C^{-1})|},$$

and, with $\delta_T \geq 1$ as in (27),

$$\frac{1}{\delta_T}\frac{B^TB}{2} \leq \frac{B^TC^{-1}C^{-T}B}{2|\det(C^{-1})|} \leq \delta_T\frac{B^TB}{2}.$$

There are many ways of writing the constant δ_T of (27). For instance, if $\beta, \gamma \in (0, \pi/2)$, we find using the relation $\alpha + \beta + \gamma = \pi$ that

$$y_T = \frac{\sin^2(\beta) + \sin^2(\gamma)}{\sin^2(\beta)\sin(2\gamma) + \sin^2(\gamma)\sin(2\beta)} \leq \frac{1}{\min(\sin(2\beta), \sin(2\gamma))},$$

which is quite precise if β or γ is small compared to the other two angles. We also have that δ_T is uniformly bounded for $T \in \mathcal{T}_\nu$ for all ν if and only if all angles occurring in $\mathcal{T}_\nu$ are bounded away from zero uniformly in ν, i.e., $(\mathcal{T}_\nu)_\nu$ is shape-regular. Moreover, there holds

$$\delta_T \leq 2y_T = \frac{b^2 + c^2}{2m(T)} \leq \frac{a+b+c}{2m(T)}\,\max\{a, b, c\},$$

the expression on the right-hand side being bounded above by the ratio of the diameter of the triangle T to the radius of the largest disk contained in T.

For our triangulation $\mathcal{T}_\nu$ obtained as the image of the uniform triangulation, we also know from the proof of Theorem 1.1 that

$$\frac{C}{\sqrt{|\det(C)|}} \approx \eta\,\frac{\nabla\phi(\zeta)}{\sqrt{|\det(\nabla\phi(\zeta))|}}, \quad \zeta \in \phi^{-1}(T), \tag{29}$$

and hence

$$\delta := \sup_\nu \max_{T\in\mathcal{T}_\nu}\delta_T = \sup_\nu \max_{T\in\mathcal{T}_\nu} \operatorname{cond}\left(\frac{C}{\sqrt{|\det(C)|}}\right) \approx \sup_{\zeta\in\tilde{\Omega}\setminus\Gamma} \operatorname{cond}\left(\frac{\nabla\phi(\zeta)}{\sqrt{|\det(\nabla\phi(\zeta))|}}\right).$$

This latter quantity turns out to be very simple for the refined triangulations discussed in Examples 1.3 and 1.4, namely $\delta \approx \beta$, with $\beta\pi \in (\pi, 2\pi)$ being the largest inner angle of Ω. We should notice that these last arguments are not completely rigorous, since in general relation (29) can be shown to be true only for triangles T with $\phi^{-1}(T)$ having a certain distance to Γ. However, there exist similar mesh refinements where the resulting family $(\mathcal{T}_\nu)_\nu$ is shape-regular and where explicit lower bounds for the angles are known.

3.2. Extremal eigenvalues, condition numbers, and preconditioning. The four statements in this section will have a short proof since they are related to previously known results. For our first statement we have been strongly inspired by similar results for so-called matrix-valued linear and positive operators (LPOs) (see [27, 34]). Here we give a short direct proof.

THEOREM 3.2. *Assume that the matrix K is uniformly elliptic and bounded; i.e., there exist positive constant $\kappa_{\min}$ and $\kappa_{\max}$ such that $\kappa_{\min} I_2 \le K(x) \le \kappa_{\max} I_2$ almost everywhere with respect to x (for instance, $\kappa_{\min} = \operatorname{essinf}_x \lambda_{\min}(K(x))$, $\kappa_{\max} = \operatorname{esssup}_x \lambda_{\max}(K(x))$). Then*

$$(A_n(K, \mathcal{T}_\nu))_\nu \text{ and } (A_n(I_2, \mathcal{T}_\nu))_\nu \text{ are uniformly equivalent}$$

$$\text{and more precisely, } \kappa_{\min} A_n(I_2, \mathcal{T}_\nu) \le A_n(K, \mathcal{T}_\nu) \le \kappa_{\max} A_n(I_2, \mathcal{T}_\nu), \tag{30}$$

and the same result is true if one replaces $\mathcal{T}_\nu$ in (30) by $\widetilde{\mathcal{T}}_\nu$.

Assume that the family of triangulations $(\mathcal{T}_\nu)_\nu$ is shape-regular, and define

$$\delta := \sup_\nu \max_{T \in \mathcal{T}_\nu} \delta_T < \infty$$

with δ_T as in (27). Then

$$(A_n(I_2, \mathcal{T}_\nu))_\nu \text{ and } (A_n(I_2, \widetilde{\mathcal{T}}_\nu))_\nu \text{ are uniformly equivalent}$$

$$\text{and more precisely } \tfrac{1}{\delta} A_n(I_2, \widetilde{\mathcal{T}}_\nu) \le A_n(I_2, \mathcal{T}_\nu) \le \delta A_n(I_2, \widetilde{\mathcal{T}}_\nu). \tag{31}$$

Proof. The main work for proving statements (30) and (31) has been done already in subsection 3.1: according to (26), the claimed inequalities in (30) are obtained by summing over all triangles $T \in \mathcal{T}_\nu$ the first inequality of Proposition 3.1. Similarly, relating the triangulations $\mathcal{T}_\nu$ and $\widetilde{\mathcal{T}}_\nu$ for $K = I_2$ means that we have to study how the stiffness matrix changes if C in (26) is replaced by I_2: the answer is obtained by summing the last inequality of Proposition 3.1 for all triangles (after replacing δ_T by δ). ∎

The preceding result enables us to give more precise bounds for the smallest and largest eigenvalue of the different stiffness matrices occurring in Theorem 3.2.

COROLLARY 3.3. *Assume that the matrix K is uniformly elliptic and bounded, and that $(\mathcal{T}_\nu)_\nu$ is shape-regular. Then the largest eigenvalue of $A_n(K, \mathcal{T}_\nu)$ is uniformly bounded in ν, and the smallest behaves like $1/\nu^2$ for $\nu \to \infty$.*

In particular, the spectral condition number of $A_n(K, \mathcal{T}_\nu)$ behaves like n, the number of vertices of $\mathcal{T}_\nu$.

Proof. Since $\widetilde{\Omega}$ is bounded, it is contained in a square with sides of size d_{out} and contains a square of size d_{in}. Then $A_n(I_2, \widetilde{\mathcal{T}}_\nu)$ contains as submatrix the Toeplitz matrix generated by $4 - 2\cos(s_1) - 2\cos(s_2)$ of order $d_{\text{in}}(\nu - 1)^2$, and, in addition, $A_n(I_2, \widetilde{\mathcal{T}}_\nu)$ is a submatrix of a Toeplitz matrix generated by $4 - 2\cos(s_1) - 2\cos(s_2)$ of order $d_{\text{out}}^2 \nu^2$ (see [36]). Since the eigenvalues of Toeplitz matrices generated by linear cosine polynomials are explicitly known, it follows that the smallest eigenvalue of $A_n(I_2, \widetilde{\mathcal{T}}_\nu)$ is of order $1/\nu^2 \sim n^{-1}$, and its maximal eigenvalue is uniformly bounded by 8, which is also its limit for $\nu \to \infty$. Using, for instance, the well-known representation of extremal eigenvalues of Hermitian matrices in terms of Rayleigh quotients, it follows from Theorem 3.2, by combining (30) and (31), that all three matrices $A_n(K, \mathcal{T}_\nu)$, $A_n(I_2, \mathcal{T}_\nu)$, and $A_n(K, \widetilde{\mathcal{T}}_\nu)$ have a smallest eigenvalue of order $1/\nu^2 \sim n^{-1}$ and a maximal eigenvalue bounded uniformly in ν. ∎

Corollary 3.3 has been proved in [2, relation (5.102c), p. 235, and pp. 236–238], [9, Lemma V.2.6], [20, p. 61 and Lemma 2.6, p. 233], and [37, Theorem 5.1], under the additional assumption that $(\mathcal{T}_\nu)_\nu$ is also quasi-uniform. Notice that the proofs given in the above references consist of comparing suitable Sobolev norms, and here the quasi uniformity condition cannot be dropped. The idea contained in subsection 3.1 is to use the updating formulas, i.e., a kind of element-by-element local analysis which is more effective than a global analysis (see, e.g., [1] and the work by Fried [15], where a similar technique has been extensively used).

Let us finally turn to the problem of designing a preconditioner for the CG method applied to the system $A_n(K,\mathcal{T}_\nu)x_n = b_n$. We recall that the matrix $A_n(I_2,\widetilde{\mathcal{T}}_\nu)$ corresponding to the uniform triangulation $\widetilde{\mathcal{T}}_\nu$ coincides with that obtained by applying the classical finite difference 5 point stencil to the Poisson problem $-\Delta u = f$. Thus solving the system $A_n(I_2,\widetilde{\mathcal{T}}_\nu)y_n = c_n$ can be performed in $\mathcal{O}(n)$ operations using, e.g., the method of cyclic reductions [11, 12, 14], and thus such a matrix would be a practical preconditioner. Define also the matrix

$$D_n = \operatorname{diag}\left(\left\|\widetilde{K}\left(\frac{(j,k)}{\nu}\right)\right\|\right)_{\frac{(j,k)}{\nu}\in\widetilde{\Omega}_h},$$

which again is a practical preconditioner. Then, under the assumptions of Proposition 3.1, the condition number of $A_n(I_2,\widetilde{\mathcal{T}}_\nu)^{-1}A_n(K,\mathcal{T}_\nu)$ and of $A_n(I_2,\widetilde{\mathcal{T}}_\nu)^{-1}D_n^{-1/2}\cdot A_n(K,\mathcal{T}_\nu)D_n^{-1/2}$ can be bounded independently of the stepsize $1/\nu$ in terms of the smallest angle used in the triangulation of Ω, plus possibly the norm and the ellipicity constant of K. This means that the associated preconditioned CG (PCG) will achieve a fixed precision in $\mathcal{O}(n)$ operations also in the nonconstant coefficient case with a nonuniform triangulation.

In the following two results we give a complete picture (localization and distribution) of the spectral behavior of preconditioned matrix sequences arising from the use of the above-mentioned preconditioners.

COROLLARY 3.4. *Assume that the matrix K is uniformly elliptic and bounded, i.e., there exist positive constant $\kappa_{\min}$ and $\kappa_{\max}$ such that $\kappa_{\min}I_2 \le K(x) \le \kappa_{\max}I_2$ almost everywhere with respect to x (for instance, $\kappa_{\min} = \operatorname{essinf}_x\lambda_{\min}(K(x))$, $\kappa_{\max} = \operatorname{esssup}_x\lambda_{\max}(K(x))$). Then*

$$\text{(32)}\qquad \text{the eigenvalues of } A_n(I_2,\mathcal{T}_\nu)^{-1}A_n(K,\mathcal{T}_\nu) \text{ belong to } [\kappa_{\min},\kappa_{\max}],$$

and the same result is true if one replaces $\mathcal{T}_\nu$ in (32) by $\widetilde{\mathcal{T}}_\nu$.

Assume also that the family of triangulations $(\mathcal{T}_\nu)_\nu$ is shape-regular such that $\delta := \sup_\nu \max_{T\in\mathcal{T}_\nu}\delta_T < \infty$ with δ_T as in (27). Then

$$\text{(33)}\qquad \text{the eigenvalues of } A_n(I_2,\widetilde{\mathcal{T}}_\nu)^{-1}A_n(I_2,\mathcal{T}_\nu) \text{ belong to } [1/\delta,\delta];$$

$$\text{(34)}\qquad \text{the eigenvalues of } A_n(I_2,\widetilde{\mathcal{T}}_\nu)^{-1}A_n(K,\mathcal{T}_\nu) \text{ belong to } [\kappa_{\min}/\delta,\kappa_{\max}\delta].$$

Proof. Statements (32) and (33) follow from the corresponding statements (30) and (31) in Theorem 3.2 and the fact that, for Hermitian positive definite X, Y, we have for the spectrum $\Lambda(Y^{-1}X)$ the localization

$$\Lambda(Y^{-1}X) \subset \left\{\frac{u^*Xu}{u^*Yu} : u \neq 0\right\}.$$

The claim (34) follows from (30) and (31) by rewriting the Rayleigh quotient as

$$\frac{u^* A_n(K,\mathcal{T}_\nu)u}{u^* A_n(I_2,\widetilde{\mathcal{T}}_\nu)u} = \frac{u^* A_n(K,\mathcal{T}_\nu)u}{u^* A_n(I_2,\mathcal{T}_\nu)u}\frac{u^* A_n(I_2,\mathcal{T}_\nu)u}{u^* A_n(I_2,\widetilde{\mathcal{T}}_\nu)u}. \qquad \square$$

THEOREM 3.5. *Assume that the matrix K is uniformly elliptic in the sense of Corollary* 3.4. *Consider the preconditioned sequences $(Y_n^{-1}X_n)$ with*

$$[Y_n, X_n] \in \{[A_n(I_2,\widetilde{\mathcal{T}}_\nu), A_n(K,\widetilde{\mathcal{T}}_\nu)], [A_n(I_2,\mathcal{T}_\nu), A_n(K,\mathcal{T}_\nu)], \\ [A_n(I_2,\widetilde{\mathcal{T}}_\nu), A_n(I_2,\mathcal{T}_\nu)], [A_n(I_2,\widetilde{\mathcal{T}}_\nu), A_n(K,\mathcal{T}_\nu)]\}.$$

Then, calling ω_X the symbol of (X_n) and calling ω_Y the symbol of (Y_n), we have that the asymptotic spectrum of $(Y_n^{-1}X_n)$ is given by ω_X/ω_Y.

Proof. It is enough to observe that all the involved matrix sequences are such that both X_n and Y_n come from the same matrix-valued LPO for which the distribution is known (see Theorem 1.1) and is sparsely vanishing (i.e., the symbol vanishes in a set of zero Lebesgue measure). The conclusion follows from the general theory of LPOs as in Theorem 2.9 of [28] (compare also Theorem 4.6 in [33] and Theorem 3.7 in [26]). $\square$

With the notation of the above theorem, we remark that the same result could be proved for the matrices $[Y_n, X_n] = [D_n^{1/2} A_n(I_2,\widetilde{\mathcal{T}}_\nu) D_n^{1/2}, A_n(K,\mathcal{T}_\nu)]$. Indeed $D_n^{1/2}$, $A_n(I_2,\widetilde{\mathcal{T}}_\nu)$, and $A_n(K,\mathcal{T}_\nu)$ are all (reduced) generalized locally Toeplitz sequences with sparsely vanishing symbols (i.e., zero on at most a set of zero Lebesgue measure): for D_n the statement is trivial since the matrix is diagonal, while for the remaining two matrix sequences this has been proved in Corollary 1.2. Then our claims follow from the fact that, if the symbols are all sparsely vanishing and sparsely unbounded (the inverse of a sparsely vanishing), then the operation $X_n \odot Y_n$ also gives a sequence in the generalized locally Toeplitz class, with asymptotic spectrum described by the symbol $\omega_X \odot \omega_Y$; this has been shown in [31, Theorem 5.8] for $\odot$ being multiplication, in the same paper for $\odot$ being addition or subtraction, and is known to be true also for inversion, that is, for the sequence $(Y_n^{-1}X_n)$ (see [32, Theorems 2.2 and 3.2]).

In order to illustrate Theorem 3.5 and its link with Theorem 1.1, we mention more explicitly the example that the sequence of matrices $(A_n(I_2,\widetilde{\mathcal{T}}_\nu)^{-1}A_n(K,\mathcal{T}_\nu))$ for $\nu \to \infty$ has an asymptotic spectrum described by the measure σ, with

$$\int f\, d\sigma = \frac{1}{(2\pi)^2}\frac{1}{m(\widetilde{\Omega})}\int_{[-\pi,\pi]^2} ds \int_{\widetilde{\Omega}} dx\, f\left(\omega(x,s)\right),$$

$\widetilde{K}(x) = |\det \nabla\phi(x)| \nabla\phi(x)^{-1} K(\phi(x)) \nabla\phi(x)^{-T}$ as before,

$$\omega(x,s) = \frac{\omega_X(x,s)}{\omega_Y(x,s)} = \frac{\begin{bmatrix} 1-e^{is_1} \\ 1-e^{is_2}\end{bmatrix}^* \cdot \widetilde{K}(x) \cdot \begin{bmatrix} 1-e^{is_1} \\ 1-e^{is_2}\end{bmatrix}}{\begin{bmatrix} 1-e^{is_1} \\ 1-e^{is_2}\end{bmatrix}^* \cdot \begin{bmatrix} 1-e^{is_1} \\ 1-e^{is_2}\end{bmatrix}},$$

and with $\omega_X(x,s)$, $\omega_Y(x,s)$ according to the notation of Theorem 3.5.

In particular (compare with (34)), the most important part of its eigenvalues lies in the interval

$$[\kappa_{\min}, \kappa_{\max}] = \left[\operatorname{essinf}_{x\in\widetilde{\Omega}} \lambda_{\min}(\widetilde{K}(x)), \operatorname{esssup}_{x\in\widetilde{\Omega}} \lambda_{\max}(\widetilde{K}(x))\right].$$

4. Concluding remarks. We have shown the existence of an asymptotic spectrum for the sequence of stiffness matrices, which occur in the P_1 finite element approximation of the two-dimensional model problem (1) with an a priori mesh refinement and varying stepsizes. The underlying symbol ω of this asymptotic spectrum, given in Theorem 1.1, depends not only on the domain and the coefficient functions of the PDE, but also on the particular P_1 approximation scheme (via the dependency on the Fourier variable s) and the map ϕ which describes our mesh refinement. We expect, by analogy with the finite difference case (see [31]), that Theorem 1.1 holds also for other finite elements if one adapts the choice of the trigonometric polynomials in s. It is probably also possible to extend our results to higher dimensions and to other elliptic PDEs, and probably we need only quite weak regularity assumptions on the involved domain and the involved coefficient functions, as in the finite difference case (see [38, 31, 32]). On the other hand, the graded meshes used in modern solvers (especially those generated by a posteriori mesh refinements) in general are not topologically equivalent to the meshes considered in this paper. Notice that, for proving asymptotic spectral results of global type, it is sufficient that the graded meshes are equivalent to an approximation of our meshes (see [35]). These issues should be investigated in more detail in future works, in order to widen the practical impact of our findings.

In the second part of the paper we have analyzed the spectral behavior of some preconditioned finite element matrix sequences in terms of localization, extremal, and, especially, distributional spectral results. The analysis could be used for deducing more precise bounds on the (P)CG convergence, in view of the results in [4, 5, 6]: the related specific study and the related numerical experiments will be part of a subsequent work.

Beside the locally Toeplitz idea, we have used in section 3.1 another purely linear algebra tool, namely the local domain analysis: it consists of decomposing complicated matrix structures in linear combinations of nonnegative definite dyads or low-rank matrices for which the (spectral) analysis is very simple, and then combining these results to deduce properties of the original matrix. (For finite elements see [1] and the beautiful and rich paper by Fried [15, e.g., (47)]; for finite differences compare with [33, section 3.5], [7, Theorem 3.7]; and for general matrices see [26].) We mention that this simple tool is especially useful for preconditioning analysis and for the analysis of extremal eigenvalues asymptotics. As a byproduct we have deduced in Corollary 3.4 that the finite element matrix sequence with uniform triangulation and the nonuniform one (not necessarily verifying the quasi uniformity) are spectrally equivalent. Thus a simpler (projected) two-level Toeplitz structure associated with the uniform triangulation can be employed as preconditioner requiring a constant number of iterations independently of the size of the problem.

REFERENCES

[1] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, Germany, 1999.

[2] O. Axelsson and V. Barker, *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press, New York, 1984.

[3] O. Axelsson and G. Lindskog, *On the rate of convergence of the preconditioned conjugate gradient method*, Numer. Math., 48 (1986), pp. 499–523.

[4] B. Beckermann and A. B. J. Kuijlaars, *Superlinear convergence of conjugate gradients*, SIAM J. Numer. Anal., 39 (2001), pp. 300–329.

[5] B. Beckermann and A. B. J. Kuijlaars, *On the sharpness of an asymptotic error estimate for conjugate gradients*, BIT, 41 (2001), pp. 856–867.

[6] B. Beckermann and A. B. J. Kuijlaars, *Superlinear CG convergence for special right-hand sides*, Electron. Trans. Numer. Anal., 14 (2002), pp. 1–19.
[7] D. Bertaccini, G. Golub, S. Serra Capizzano, and C. Tablino Possio, *Preconditioned HSS method for the solution of non-Hermitian positive definite linear systems and applications to the discrete convection-diffusion equation*, Numer. Math., 99 (2005), pp. 441–484.
[8] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*, Springer, New York, 1999.
[9] D. Braess, *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Cambridge University Press, Cambridge, UK, 2001.
[10] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.
[11] B. L. Buzbee, G. H. Golub, and C. W. Nielson, *On direct methods for solving Poisson's equations*, SIAM J. Numer. Anal., 7 (1970), pp. 627–656.
[12] B. L. Buzbee, F. W. Dorr, J. A. George, and G. H. Golub, *The direct solutions of the discrete Poisson equation on irregular regions*, SIAM J. Numer. Anal., 8 (1971), pp. 722–736.
[13] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.
[14] F. W. Dorr, *The direct solution of the discrete Poisson equation on a rectangle*, SIAM Rev., 12 (1970), pp. 248–263.
[15] I. Fried, *Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices*, Internat. J. Solids Structures, 9 (1973), pp. 1013–1034.
[16] L. Golinskii and S. Serra Capizzano, *The asymptotic properties of the spectrum of non symmetrically perturbed Jacobi matrix sequences*, J. Approx. Theory, 144 (2007), pp. 84–102.
[17] U. Grenander and G. Szegö, *Toeplitz Forms and Their Applications*, 2nd ed., Chelsea, New York, 1984.
[18] S. Holmgren, S. Serra Capizzano, and P. Sundqvist, *Can one hear the composition of a drum?* Mediterranean J. Math., 3 (2006), pp. 227–249.
[19] S. Holmgren, S. Serra Capizzano, and P. Sundqvist, *On the asymptotic spectrum of (non symmetric) finite difference matrix sequences*, 2006, in preparation.
[20] C. Johnson, *Numerical Solutions of Partial Differential Equations by the Finite Elements Methods*, Cambridge University Press, Cambridge, UK, 1988.
[21] S. D. Kim and S. V. Parter, *Preconditioning Chebyshev spectral collocation by finite-differences operators*, SIAM J. Numer. Anal., 34 (1997), pp. 939–958.
[22] A. B. J. Kuijlaars and S. Serra Capizzano, *Asymptotic zero distribution of orthogonal polynomials with discontinuously varying recurrence coefficients*, J. Approx. Theory, 113 (2001), pp. 142–155.
[23] S. Parter, *On the eigenvalues of certain generalizations of Toeplitz matrices*, Arch. Ration. Math. Mech., 3 (1962), pp. 244–257.
[24] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.
[25] E. B. Saff and V. Totik, *Logarithmic Potentials with External Fields*, Springer, Berlin, 1997.
[26] S. Serra Capizzano, *Locally X matrices, spectral distributions, preconditioning, and applications*, SIAM J. Matrix Anal. Appl., 21 (2000), pp. 1354–1388.
[27] S. Serra Capizzano, *Some theorems on linear positive operators and functionals and their applications*, Comput. Math. Appl., 39 (2000), pp. 139–167.
[28] S. Serra Capizzano, *A note on the asymptotic spectra of finite difference discretizations of second order elliptic partial differential equations*, Asian J. Math., 4 (2000), pp. 499–514.
[29] S. Serra Capizzano, *Distribution results on the algebra generated by Toeplitz sequences: A finite dimensional approach*, Linear Algebra Appl., 328 (2001), pp. 121–130.
[30] S. Serra Capizzano, *Spectral behaviour of matrix sequences and discretized boundary value problems*, Linear Algebra Appl., 337 (2001), pp. 37–78.
[31] S. Serra Capizzano, *Generalized locally Toeplitz sequences: Spectral analysis and applications to discretized partial differential equations*, Linear Algebra Appl., 366 (2003), pp. 371–402.
[32] S. Serra Capizzano, *GLT sequences as a generalized Fourier analysis and applications*, Linear Algebra Appl., 419 (2006), pp. 180–233.
[33] S. Serra Capizzano and C. Tablino Possio, *Spectral and structural analysis of high precision finite difference matrices for elliptic operators*, Linear Algebra Appl., 293 (1999), pp. 85–131.
[34] S. Serra Capizzano and C. Tablino Possio, *Finite element matrix-sequences: The case of rectangular domains*, Numer. Algorithms, 28 (2001), pp. 309–327.
[35] S. Serra Capizzano and C. Tablino Possio, *Analysis of preconditioning strategies for collocation linear systems*, Linear Algebra Appl., 369 (2003), pp. 41–75.
[36] S. Serra Capizzano and C. Tablino Possio, *Superlinear preconditioners for finite differences linear systems*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 152–164.

[37] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, Prentice–Hall, Englewood Cliffs, NJ, 1973.
[38] P. Tilli, *Locally Toeplitz sequences: Spectral properties and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.
[39] E. E. Tyrtyshnikov, *A uniform approach to some old and new theorems on distribution and clustering*, Linear Algebra Appl., 232 (1996), pp. 1–43.
[40] E. E. Tyrtyshnikov, *A matrix view on the root distribution for orthogonal polynomials*, in Structured Matrices: Recent Developments in Theory and Computation, D. Bini, E. Tyrtyshnikov, and P. Yalamov, eds., Nova Science, New York, 2001, pp. 149–156.
[41] A. van der Sluis and H. A. van der Vorst, *The rate of convergence of conjugate gradients*, Numer. Math., 48 (1986), pp. 543–560.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 770–786

ULTRASPHERICAL STIELTJES POLYNOMIALS AND GAUSS–KRONROD QUADRATURE BEHAVE NICELY FOR $\lambda < 0$*

B. DE LA CALLE YSERN† AND F. PEHERSTORFER‡

Abstract. We show that the zeros of the Stieltjes polynomials associated with the ultraspherical weight function with parameter $\lambda < 0$ are real and simple and, except for two of them, belong to $(-1, 1)$. We also prove that they strictly interlace with the zeros of the corresponding ultraspherical polynomials. Consequently, a Gauss–Kronrod quadrature formula with two nodes outside the interval $[-1, 1]$ is obtained. We prove that the coefficients of such quadrature rules are positive. Finally, an asymptotic representation of Stieltjes polynomials which converges uniformly on the whole interval $[-1, 1]$ is provided.

Key words. Gauss–Kronrod quadrature, Stieltjes polynomials, ultraspherical polynomials, asymptotics

AMS subject classifications. 33C45, 42C05, 65D32

DOI. 10.1137/060651896

1. Introduction.

Let $p_n^\lambda(x) = \kappa_{\lambda,n} x^n + \ldots$, $\kappa_{\lambda,n} > 0$, be the orthonormal polynomial of degree n with respect to the ultraspherical weight function $w_\lambda(x) = (1-x^2)^{\lambda-1/2}$, $\lambda > -1/2$, i.e.,

$$\int_{-1}^{1} p_n^\lambda(x)\, p_m^\lambda(x)\, w_\lambda(x)\, dx = \delta_{n,m}, \tag{1.1}$$

where $\delta_{n,m}$ is the Kronecker symbol. Furthermore, let

$$q_n^\lambda(z) := \int_{-1}^{1} \frac{p_n^\lambda(x)}{z-x}\, w_\lambda(x)\, dx, \quad z \notin [-1, 1],$$

be the corresponding nth function of the second kind. Note that, by (1.1),

$$q_n^\lambda(z) = \mathcal{O}(1/z^{n+1}), \quad z \to \infty.$$

In 1894, in one of his letters to Hermite, Stieltjes [11] considered the polynomials $E_{n+1}^\lambda(z) = z^{n+1} + \ldots$, now called Stieltjes polynomials, which arise in the series expansion of $1/q_n^\lambda(z)$ at $z = \infty$, that is,

$$1/q_n^\lambda(z) = \kappa_{\lambda,n}\, E_{n+1}^\lambda(z) + \mathcal{O}(1/z), \quad z \to \infty, \tag{1.2}$$

and proved that these polynomials satisfy orthogonality conditions with respect to a sign-changing function, to be exact,

$$\int_{-1}^{1} x^j\, E_{n+1}^\lambda(x)\, p_n^\lambda(x)\, w_\lambda(x)\, dx = 0, \quad j = 0, \ldots, n. \tag{1.3}$$

*Received by the editors February 10, 2006; accepted for publication (in revised form) December 8, 2006; published electronically April 13, 2007.
http://www.siam.org/journals/sinum/45-2/65189.html

†Dpto. de Matemática Aplicada, E. T. S. de Ingenieros Industriales, Universidad Politécnica de Madrid, José G. Abascal 2, 28006 Madrid, Spain (bcalle@etsii.upm.es). The work of this author was supported by the Dirección General de Investigación, Ministerio de Educación y Ciencia, under grants MTM2006-13000-C03-02 and MTM2006-07186 and by UPM under Ayuda Puente Ref. AY05/11263.

‡Abteilung für Dynamische Systeme und Approximationstheorie, Institut für Analysis, Johannes Kepler Universität Linz, Altenbergerstr. 69, A-4040 Linz, Austria (franz.peherstorfer@jku.at). The work of this author was supported by the Austrian Science Fund FWF, project P16390-N04.

In fact, Stieltjes considered only the Legendre weight $w_{1/2}(x) \equiv 1$ and conjectured that all the zeros of $E_{n+1}^{1/2}$ belong to the interval $(-1,1)$ and strictly interlace with the zeros of the Legendre polynomials $p_n^{1/2}$. Hermite [11] declared himself truly enchanted with the new polynomials, but no further progress seems to have been made either by him or Stieltjes.

Forty years later, Szegő [24] proved the conjecture of Stieltjes. Indeed, he proved that for $0 < \lambda \leq 2$ the Stieltjes polynomial E_{n+1}^{λ} has $n+1$ simple zeros in $(-1,1)$ and that the zeros of E_{n+1}^{λ} strictly interlace with the zeros of p_n^{λ}. The case $\lambda > 2$ has been recently studied by Petras and the second author in a series of papers [21, 22]. For $2 < \lambda < 3$ they show that on compact subsets $[-1+\epsilon, 1-\epsilon]$, $\epsilon > 0$ arbitrary but fixed, and for $n \geq n_0(\epsilon)$, the strictly interlacing property of the zeros of E_{n+1}^{λ} and p_n^{λ} still holds, whereas for $\lambda > 3$ they prove that the polynomials E_{n+1}^{λ} have only $o(n)$ real zeros. The case $-1/2 < \lambda < 0$ has remained open up to now.

An asymptotic representation of the Stieltjes polynomials for $\lambda \in (0,2)$ has been obtained by Ehrich [2, 3]. For $\lambda > 2$ asymptotic formulae are given in [21, 22].

The importance of the zeros of the Stieltjes polynomials in numerical integration arose in the sixties when Kronrod [14] suggested the so-called extended Gaussian quadrature formulae. In such quadrature rules, one looks for formulae of the form

$$\int_{-1}^{1} f(x)\, w_{\lambda}(x)\, dx = \sum_{i=1}^{n} A_{n,i}^{\lambda}\, f(x_{n,i}) + \sum_{j=1}^{n+1} B_{n,j}^{\lambda}\, f(y_{n,j}) + R_{2n+1}^{\lambda}(f), \tag{1.4}$$

where nodes $\{x_{n,i}\}_{i=1}^{n}$ are the zeros of p_n^{λ}, i.e., the Gaussian nodes, and the rest of nodes $\{y_{n,j}\}_{j=1}^{n+1}$ and all the quadrature weights $\{A_{n,i}\}_{i=1}^{n}, \{B_{n,j}\}_{j=1}^{n+1}$ are chosen so that

$$R_{2n+1}^{\lambda}(p) = 0 \quad \text{for all} \quad p \in \mathbb{P}_{3n+1} \tag{1.5}$$

(as usual $\mathbb{P}_m$ denotes the set of polynomials of degree less than or equal to m), where the degree of polynomial exactness is $3n+1$ when n is an even number and $3n+2$ if n is odd. Now, it is not difficult to show that condition (1.5) is equivalent to the fact that the nodal polynomial

$$\prod_{j=1}^{n+1} (x - y_{n,j})$$

satisfies the orthogonality relations (1.3) in the same way as E_{n+1}^{λ}. Hence, the nodes $\{y_{n,j}\}_{j=1}^{n+1}$ turn out to be the zeros of the Sieltjes polynomial E_{n+1}^{λ}.

This was the starting point for the renewed interest in Stieltjes polynomials, since the Gauss–Kronrod quadrature has been frequently used in automatic integration processes [7, 23]. For numerical stability reasons, positivity of the quadrature weights is an important and often required property. It may be easily proved that the positivity of the coefficients $\{B_{n,j}^{\lambda}\}_{j=1}^{n+1}$ is equivalent to the strict interlacing property of the zeros of E_{n+1}^{λ} and p_n^{λ}. Using this fact, and with the help of Szegő's results [24], Monegato [16] has proved the positivity of all of the quadrature weights appearing in (1.4) for $0 \leq \lambda \leq 1$. Recall that for $\lambda > 3$ the Gauss–Kronrod quadrature is not possible for sufficiently large n because only $o(n)$ nodes are real.

Surprisingly, except for the numerical work [9] of Gautschi and Notaris from which nice behavior of the zeros of E_{n+1}^{λ} is conjectured, nothing was known until now about the reality of the Kronrod nodes for $-1/2 < \lambda < 0$. In this paper we give a complete

description of the behavior of the nodes and the sign of the quadrature weights of the Gauss–Kronrod rule *for each* $n \in \mathbb{N}$. To be more precise, we show that E_{n+1}^{λ} has exactly two real zeros outside $[-1,1]$ and that, except for that feature, it has all the properties satisfied by the Stieltjes polynomial in the classical case $\lambda \in (0,1)$; that is, the zeros of E_{n+1}^{λ} strictly interlace with the zeros of p_n^{λ} and the quadrature weights are all positive. Furthermore, we provide an asymptotic representation of the Stieltjes polynomials that converges uniformly on the *whole* interval $[-1,1]$.

Finally, we mention that, in recent years, other wide classes of weight functions have been found for which Gauss–Kronrod quadrature with positive quadrature weights is possible for $n \geq n_0$, namely, weight functions of the form $\sqrt{1-x^2}\, v(x)$, where $v \in C^2[-1,1]$ and $v > 0$ on $[-1,1]$ (see [19, 20]). For Jacobi weights with parameters $0 \leq \alpha, \beta < 5/2$ asymptotic representations on $[-1+\epsilon, 1-\epsilon]$, $\epsilon > 0$, for the corresponding Stieltjes polynomials and quadrature weights have been given in [22]. For a more detailed discussion of weight functions admitting the Gauss–Kronrod quadrature rule as well as further references, see the surveys [17, 8, 18, 20, 5].

In the case of real measures $d\mu(x) = k(x)\, dx$, where $k \in L^1[-1,1]$, there exist results [4] concerning product integration based on the Gauss–Kronrod nodes which show that, from the practical point of view, Gauss–Kronrod nodes are as good as the Clenshaw–Curtis ones for constructing interpolatory (degree of polynomial exactness equal to $2n$) quadrature rules. For applications of the Stieltjes polynomials in interpolation theory, see, e.g., [6, 12].

2. Statement of the main results. Let T_n and U_n be the Chebyshev polynomials of degree n of the first and second kind, respectively.

THEOREM 2.1. *For $\lambda \in (-1/2, 0)$ and $n \in \mathbb{N}$, the zeros of $E_{n+1}^{\lambda}(x)$ are real and simple and, except for two of them, belong to the interval $(-1,1)$. Moreover, they strictly interlace with the zeros of the polynomial $(1-x^2)\, U_{n-2}(x)$ and also with the zeros of the orthonormal polynomial $p_n^{\lambda}(x)$.*

From results contained in [1] it follows that the two zeros outside $[-1,1]$ approach the ends of the interval as n tends to ∞. In this respect, it may be pertinent to recall that for $\lambda = 0$ the smallest zero and the largest zero of the Stieltjes polynomials coincide with the boundary points -1 and $+1$, respectively, and move inside the interval $(-1,1)$ as λ gets larger with n fixed.

As a consequence of the last assertion in Theorem 2.1, we can prove the following corollary.

COROLLARY 2.2. *For $\lambda \in (-1/2, 0)$ and $n \in \mathbb{N}$, the quadrature weights of* (1.4) *are all strictly positive.*

Regarding asymptotics of Stieltjes polynomials, we have the following theorem.

THEOREM 2.3. *For $\lambda \in (-1/2, 0)$ and $\theta \in [0, \pi]$, it holds that*

$$2^n E_{n+1}^{\lambda}(\cos\theta) = (2\sin\theta)^{1-\lambda}\, \cos\{(n+1)\theta + (\lambda-1)(\theta - \pi/2)\} + o(1), \tag{2.1}$$

uniformly on $[0,\pi]$.

For $\lambda \in [0,3)$ and $\theta \in [\epsilon, \pi - \epsilon]$, with $\epsilon > 0$ arbitrary but fixed, relation (2.1) has been proved in [2, 3, 22].

Remark 1. Taking a closer look at the right-hand side of (2.1), we obtain, for $\lambda \in (-1/2, 0)$ and $x \in [-1,1]$, the following asymptotic representation of Stieltjes polynomials in terms of Chebyshev polynomials:

$$2^n E_{n+1}^{\lambda}(x) = \sum_{k=0}^{[(n+1)/2]\dagger} \frac{\Gamma(k+\lambda-1)}{\Gamma(\lambda-1)\,\Gamma(k+1)}\, T_{n+1-2k}(x) + o(1), \tag{2.2}$$

uniformly on $[-1,1]$, where the symbol † indicates that the last term should be halved if n is odd. We note that (2.2) holds true for $\lambda \in [0,2)$ and $x \in [-1+\epsilon, 1-\epsilon]$, $\epsilon > 0$, taking into consideration what was mentioned immediately after Theorem 2.3. For more details, see section 4.4.

Remark 2. It is worth pointing out that, as a byproduct of the proof of Theorem 2.3 (cf. equalities (4.10) and (4.12) below), an asymptotic formula is obtained deserving of attention on its own. Namely, for $\lambda \in (-1/2, 0)$,

$$2^n E_{n+1}^{\lambda}(\cos\theta) = (2\sin\theta)^{1-\lambda}\,\cos\{(n+1)\theta + (\lambda-1)(\theta-\pi/2)\} + \mathcal{O}(1/n), \tag{2.3}$$

uniformly on $[\epsilon, \pi-\epsilon]$, with $\epsilon > 0$ arbitrary but fixed. It would be interesting to know whether the rate of convergence in (2.3) may be extended uniformly to the whole interval $[0,\pi]$. Numerical experiments support the possibility of such a result being true.

3. Preliminary results. Let $\lambda \neq 0$. In what follows, G_n^{λ} will denote the ultraspherical polynomial of degree n associated with w_λ (cf. [25, Chapter IV] for details of definitions and properties below) normalized so that

$$G_n^{\lambda}(1) = \frac{\Gamma(n+2\lambda)}{\Gamma(n+1)\Gamma(2\lambda)}.$$

A straightforward calculation shows that the leading coefficient of G_n^{λ}, which will be denoted by $d_{\lambda,n}$, is given by

$$d_{\lambda,n} = \frac{2^n\,\Gamma(n+\lambda)}{\Gamma(n+1)\,\Gamma(\lambda)}, \tag{3.1}$$

whereas $\kappa_{\lambda,n}$, the leading coefficient of the orthonormal polynomial p_n^{λ}, turns out to be

$$\kappa_{\lambda,n} = \frac{2^{n+\lambda}}{\sqrt{2\pi}}\,\Gamma(n+\lambda)\,\sqrt{\frac{n+\lambda}{\Gamma(n+1)\Gamma(n+2\lambda)}}. \tag{3.2}$$

Denote by Q_n^{λ} the ultraspherical function of the second kind normalized so that

$$Q_n^{\lambda}(z) = \frac{1}{2}\,\frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\int_{-1}^{1}\frac{G_n^{\lambda}(t)}{z-t}\,(1-t^2)^{\lambda-1/2}dt, \quad z \notin [-1,1].$$

It is known (see [25, Theorem 4.62.1]) that if $\lambda < 1/2$, then

$$\lim_{x\to 1^+} Q_n^{\lambda}(x)\,(x^2-1)^{1/2-\lambda} = K_{\lambda,n} > 0, \quad x \in \mathbb{R}. \tag{3.3}$$

The behavior of $Q_n^{\lambda}(x)$ at $x=-1$ is similar. As usual, Q_n^{λ} is defined on $[-1,1]$ as the Cauchy principal value, i.e.,

$$Q_n^{\lambda}(x) = \frac{(Q_n^{\lambda})^+(x) + (Q_n^{\lambda})^-(x)}{2} = \lim_{\epsilon\to 0^+}\frac{Q_n^{\lambda}(x+i\epsilon) + Q_n^{\lambda}(x-i\epsilon)}{2}, \tag{3.4}$$

whenever the above limit exists. It is known that (3.4) defines an analytic function on $(-1,1)$. Sometimes, it will be more convenient to consider, instead of Q_n^{λ}, the function

$$\widetilde{Q}_n^{\lambda}(x) = (1-x^2)^{1/2-\lambda}\,Q_n^{\lambda}(x), \quad x \in (-1,1).$$

Additionally, for $\theta \in (0,\pi)$, it holds that

$$(Q_n^\lambda)^+(\cos\theta) - (Q_n^\lambda)^-(\cos\theta) = -\pi i\, \frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\; G_n^\lambda(\cos\theta)\,(\sin\theta)^{2\lambda-1},$$

so

$$(Q_n^\lambda)^-(\cos\theta) = Q_n^\lambda(\cos\theta) + \frac{i\pi}{2}\, \frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\; G_n^\lambda(\cos\theta)\,(\sin\theta)^{2\lambda-1} \tag{3.5}$$

for $\theta \in (0,\pi)$.

Set

$$C_{\lambda,n} = \sqrt{\pi}\, \frac{\Gamma(n+2\lambda)}{\Gamma(n+\lambda+1)}.$$

Following the ideas of Szegő [24], from the representation of Q_n^λ as a hypergeometric function it follows that

$$Q_n^\lambda((z+1/z)/2) = C_{\lambda,n}\, z^{n+1} \sum_{k=0}^{\infty} a_{n,k}\, z^{2k}, \quad |z|<1, \tag{3.6}$$

where the sequence $\{a_{n,k}\}_{k=0}^{\infty}$ is defined by the relations

$$a_{n,0} = 1, \quad a_{n,k} = \left(1-\frac{\lambda}{k}\right)\left(1-\frac{\lambda}{n+k+\lambda}\right) a_{n,k-1}, \quad k\in\mathbb{N}. \tag{3.7}$$

It is clear that $a_{n,k}$ also depends on λ; yet this notation will not be misleading. From (3.5) and (3.6), we obtain

$$\lim_{r\to 1^-} C_{\lambda,n} \sum_{k=0}^{\infty} a_{n,k}\,(re^{i\theta})^{n+1+2k} = Q_n^\lambda(\cos\theta) + \frac{i\pi}{2}\, \frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\; G_n^\lambda(\cos\theta)\,(\sin\theta)^{2\lambda-1}, \tag{3.8}$$

where $\theta \in (0,\pi)$.

It is very easy to prove, using (1.2), that

$$\frac{C_{\lambda,n}\, z^{n+1}}{Q_n^\lambda((z+1/z)/2)} = (2z)^{n+1}\, E_{n+1}^\lambda((z+1/z)/2) + \mathcal{O}(z^{n+2}), \quad z\to 0.$$

Now, if real numbers $b_{n,k}$, $k\in\mathbb{N}\cup\{0\}$, are defined by the recurrence formula

$$b_{n,0} = 1, \qquad \sum_{j=0}^{k} a_{n,k-j}\, b_{n,j} = 0, \quad k\in\mathbb{N}, \tag{3.9}$$

which is equivalent to

$$\left(\sum_{k=0}^{\infty} a_{n,k}\, z^{2k}\right)\left(\sum_{k=0}^{\infty} b_{n,k}\, z^{2k}\right) = 1, \quad |z|<1, \tag{3.10}$$

then we can write

$$\sum_{k=0}^{\infty} b_{n,k}\, z^{2k} = (2z)^{n+1}\, E_{n+1}^\lambda((z+1/z)/2) + \mathcal{O}(z^{n+2}), \quad z\to 0.$$

Hence it may be proved (taking into account the symmetry of $E_{n+1}^{\lambda}((z+1/z)/2)$) that the numbers $b_{n,k}$ are precisely the coefficients of the Chebyshev polynomial representation of E_{n+1}^{λ}. Specifically, we have

$$2^n\, E_{n+1}^{\lambda}(x) = \sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, T_{n+1-2k}(x), \tag{3.11}$$

where, as we mentioned before, the symbol † indicates that the last term should be halved if n is odd.

Finally, we will need some asymptotic formulae for the proof of Theorem 2.3. Namely (cf. [25, Theorem 8.21.8]),

$$\frac{n^{1-\lambda}\,\pi}{2}\,\frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\,G_n^{\lambda}(\cos\theta) = \sqrt{\pi}\,2^{\lambda-1}\,(\sin\theta)^{-\lambda}\,\cos[n\theta+\lambda(\theta-\pi/2)] + \mathcal{O}(1/n), \tag{3.12}$$

where the convergence is uniform on $[\epsilon, \pi-\epsilon]$, with $\epsilon > 0$ arbitrary but fixed.

Also (cf., for instance, [2]), it holds that

$$n^{1-\lambda}\,\widetilde{Q}_n^{\lambda}(\cos\theta) = \sqrt{\pi}\,2^{\lambda-1}(\sin\theta)^{-\lambda}\,\cos[(n+1)\theta+(\lambda-1)\,(\theta-\pi/2)] + \mathcal{O}(1/n), \tag{3.13}$$

again uniformly on $[\epsilon, \pi-\epsilon]$, with ϵ as above.

4. Proofs. In the case that $\lambda \in (0,1)$, by using a theorem on reciprocal power series by Kaluza [13], all the coefficients $b_{n,k}$, $k \geq 1$, are proved to be negative, which turns out to be a key fact in proving the properties satisfied by E_{n+1}^{λ} (cf. [24]). To the best of our knowledge, there exist no general results concerning coefficients of reciprocal power series which can be applied when $\lambda < 0$. The following lemma supplies us with a result analogous to the Kaluza theorem for the specific coefficients (3.7). Due to its rather technical character and the fact that the arguments employed in its proof are not related to the rest, we prefer to delay its proof until the end of the paper; see section 4.5. For the rest of the proofs we have partly followed some ideas from [2, 3, 24].

LEMMA 4.1. *For all $\lambda \in (-1/2, 0)$ and $n \in \mathbb{N}$, the coefficients $b_{n,k}$ defined by relations* (3.9) *and* (3.7) *satisfy*

$$b_{n,1} < 0 \;\; \textit{and, for } k \geq 2, \;\; b_{n,k} > 0.$$

For proving the theorems mentioned above, we will need further knowledge of properties fulfilled by the coefficients $\{b_{n,k}\}_{k\in\mathbb{N}}$, $n \in \mathbb{N}$, which are stated in the following two lemmas.

LEMMA 4.2. *For all $\lambda \in (-1/2, 0)$ and $n \in \mathbb{N}$, it holds that*

$$\sum_{k=0}^{\infty} b_{n,k} = 0 \quad \textit{and} \quad \sum_{k=0}^{\infty} |b_{n,k}| < K,$$

where $K < +\infty$ is an absolute constant.

Proof. Use (3.6) to obtain

$$\lim_{r\to 1^-} \sum_{k=0}^{\infty} a_{n,k}\, r^{2k} = \lim_{r\to 1^-} \frac{Q_n^{\lambda}((r+1/r)/2)}{C_{\lambda,n}\, r^{n+1}} = \frac{1}{C_{\lambda,n}} \lim_{x\to 1^+} Q_n^{\lambda}(x) = +\infty, \tag{4.1}$$

according to (3.3). As we saw in Lemma 4.1, the coefficients $b_{n,k}$ are all positive whenever $k \geq 2$. Therefore,

$$\lim_{r\to 1^-} \sum_{k=0}^{\infty} b_{n,k}\, r^{2k} = \sum_{k=0}^{\infty} b_{n,k},$$

which, together with (3.10) and (4.1), proves the first part of our assertion. Regarding the second one, we have

$$\sum_{k=0}^{\infty} |b_{n,k}| = -b_{n,1} + \sum_{\substack{k=0\\ k\neq 1}}^{\infty} b_{n,k} = -2\, b_{n,1} = 2\, a_{n,1},$$

and the result follows from the fact that (see (3.7)) $\lim_{n\to\infty} a_{n,1} = 1 - \lambda$. □

LEMMA 4.3. *For all $\lambda \in (-1/2, 0)$ and $n \in \mathbb{N}$, it holds that*

$$\sum_{k=m+1}^{\infty} b_{n,k} = \mathcal{O}(1/m), \quad m \to \infty,$$

uniformly on $n \in \mathbb{N}$; i.e., the constants involved in the term $\mathcal{O}(1/m)$ are independent of $n \in \mathbb{N}$.

Proof. For convenience, we will prove the equivalent expression

$$\sum_{k=2m+1}^{\infty} b_{n,k} = \mathcal{O}(1/m), \quad m \to \infty.$$

Let us consider the product

$$\left(-\sum_{k=0}^{2m} b_{n,k}\right)\left(\sum_{k=0}^{2m} a_{n,k}\right) = (-b_{n,1})\,(a_{n,0} + a_{n,2m}) + \sum_{k=1}^{2m-1} (-b_{n,1}\, a_{n,k}) + t_m,$$

where we have included all of the negative terms of the product in t_m. Each of the terms $(-b_{n,1}\, a_{n,k})$, $k = 1, \ldots, 2m-1$, is to be canceled out by part of the negative terms contained in t_m, according to (3.9). Notice that we use each negative term at most once and that we have enough negative terms in t_m since $k = 1, \ldots, 2m-1$. Then, on account of Lemma 4.2, we obtain

$$\sum_{k=2m+1}^{\infty} b_{n,k} = -\sum_{k=0}^{2m} b_{n,k} < a_{n,1}\, \frac{a_{n,0} + a_{n,2m}}{\displaystyle\sum_{k=0}^{2m} a_{n,k}} < \frac{a_{n,1}(2a_{n,2m})}{\displaystyle\sum_{k=m+1}^{2m} a_{n,k}} < K\, \frac{a_{n,2m}}{m\, a_{n,m}}, \tag{4.2}$$

since $a_{n,0} = 1$ and the sequence $\{a_{n,k}\}_{k\in\mathbb{N}}$ is increasing. Let us estimate $a_{n,2m}$ and $a_{n,m}$. For this, we will need the limit (cf., for instance, [10, formula 8.328.2])

$$\lim_{n\to\infty} \frac{\Gamma(n+a)}{\Gamma(n+b)\, n^{a-b}} = 1, \quad a, b \in \mathbb{R}. \tag{4.3}$$

From (3.7), it is clear that

$$a_{n,k} = \frac{\Gamma(k-\lambda+1)}{\Gamma(k+1)}\, \frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\Gamma(1-\lambda)}\, \frac{\Gamma(n+k+1)}{\Gamma(n+k+\lambda+1)}. \tag{4.4}$$

When applied to the numbers (4.4), the limit (4.3) gives

$$(1/2)\,k^{-\lambda} < \frac{\Gamma(k-\lambda+1)}{\Gamma(k+1)} < (3/2)\,k^{-\lambda},$$

$$(1/2)\,(n+k)^{-\lambda} < \frac{\Gamma(n+k+1)}{\Gamma(n+k+\lambda+1)} < (3/2)\,(n+k)^{-\lambda}$$

for all $n \in \mathbb{N}$ and $k \geq k_0$. Then

$$\frac{1}{4}\{k\,(n+k)\}^{-\lambda}\,\frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\,\Gamma(1-\lambda)} < a_{n,k} < \frac{9}{4}\{k\,(n+k)\}^{-\lambda}\,\frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\,\Gamma(1-\lambda)}$$

again for all $n \in \mathbb{N}$ and $k \geq k_0$. So, by virtue of (4.2), we have

$$\sum_{k=2m+1}^{\infty} b_{n,k} < \frac{2^{-\lambda}\,9\,K}{m}\,\left(\frac{n+2m}{n+m}\right)^{-\lambda} < \frac{4^{-\lambda}\,9\,K}{m}$$

for $m \geq k_0$ and any $n \in \mathbb{N}$, which concludes the proof. □

4.1. Proof of Theorem 2.1. As the Stieltjes polynomial E_{n+1}^{λ} is either an even or an odd function depending on n, it is enough to prove that it has one zero greater than 1 to see that it has two zeros outside $[-1,1]$. So

$$2^n\,E_{n+1}^{\lambda}(1) = \sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\,T_{n+1-2k}(1) = \sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k} < \sum_{k=0}^{\infty} b_{n,k} = 0,$$

due to Lemmas 4.1 and 4.2 and (3.11). It is obvious that the leading coefficient of E_{n+1}^{λ} is positive since it is monic. Consequently, E_{n+1}^{λ} must have a zero in $(1,+\infty)$.

Now suppose that $n \geq 2$. It is very well known that the Chebyshev polynomial T_{n-1} has n points in $[-1,1]$ at which it attains its maximum value with alternate sign. Let x_0 be one such point with $T_{n-1}(x_0) = 1$. Then, using the same arguments as in the previous reasoning, we have

$$2^n\,E_{n+1}^{\lambda}(x_0) = b_{n,1} + \sum_{\substack{k=0\\k\neq 1}}^{[(n+1)/2]\dagger} b_{n,k}\,T_{n+1-2k}(x_0) < \sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k} < \sum_{k=0}^{\infty} b_{n,k} = 0.$$

On the other hand, if $T_{n-1}(y_0) = -1$, then

$$2^n\,E_{n+1}^{\lambda}(y_0) = -b_{n,1} + \sum_{\substack{k=0\\k\neq 1}}^{[(n+1)/2]\dagger} b_{n,k}\,T_{n+1-2k}(y_0) > -\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k} > 0.$$

Therefore, E_{n+1}^{λ} has $n-1$ simple zeros inside $(-1,1)$ and two simple zeros outside $[-1,1]$. As $T'_{n-1}(x) = (n-1)\,U_{n-2}(x)$, the zeros of E_{n+1}^{λ} strictly interlace with the zeros of $(1-x^2)\,U_{n-2}(x)$.

Next, we will prove the interlacing property with respect to the zeros of the ultraspherical polynomials. Notice that, due to Lemma 4.2, it holds that

$$\lim_{r\to 1^-} \sum_{k=0}^{\infty} b_{n,k}\,(re^{i\theta})^{2k} = \sum_{k=0}^{\infty} b_{n,k}\,e^{2ki\theta}, \quad \theta \in [0,2\pi]. \tag{4.5}$$

Let $\theta \in (0,\pi)$ and consider

$$\frac{\displaystyle\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta}}{\displaystyle\sum_{k=0}^{\infty} b_{n,k}\, e^{2ki\theta}} = \left(1 + \frac{\displaystyle\sum_{k=[n/2+1]\ddagger}^{\infty} b_{n,k}\, e^{2ki\theta}}{\displaystyle\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta}} \right)^{-1},$$

where the symbol ‡ means that the first term should be halved if n is odd. Then

$$\left| \frac{\displaystyle\sum_{k=[n/2+1]\ddagger}^{\infty} b_{n,k}\, e^{2ki\theta}}{\displaystyle\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta}} \right| < \frac{\displaystyle\sum_{k=[n/2+1]\ddagger}^{\infty} b_{n,k}}{|b_{n,1}| - \left| \displaystyle\sum_{\substack{k=0 \\ k\neq 1}}^{[(n+1)/2]\dagger} b_{n,k} \right|} = \frac{\displaystyle\sum_{k=[n/2+1]\ddagger}^{\infty} b_{n,k}}{\displaystyle\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}} = 1,$$

because of Lemma 4.2. An easy calculation shows that $\Re\,(1/(1+z)) > 1/2$ for $|z| < 1$. Therefore,

$$\Re\left[\left(\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta} \right) \left(\sum_{k=0}^{\infty} b_{n,k}\, e^{2ki\theta} \right)^{-1} \right] > \frac{1}{2}. \tag{4.6}$$

Taking (3.10), (4.5), and (3.8) into account, we obtain

$$\frac{C_{\lambda,n} \displaystyle\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta}}{\displaystyle\sum_{k=0}^{\infty} b_{n,k}\, e^{2ki\theta}} = \left(\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta} \right) \left(\lim_{r\to 1^-} C_{\lambda,n} \sum_{k=0}^{\infty} a_{n,k}\, (re^{i\theta})^{2k} \right)$$

$$= \left(\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{-(n+1-2k)i\theta} \right) \left(\lim_{r\to 1^-} C_{\lambda,n} \sum_{k=0}^{\infty} a_{n,k}\, (re^{i\theta})^{n+1+2k} \right)$$

$$= \left(2^n\, E_{n+1}^{\lambda}(\cos\theta) - i\widetilde{E}_n(\theta) \right) \left(Q_n^{\lambda}(\cos\theta) + \frac{\pi i}{2}\, \frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)\, (\sin\theta)^{2\lambda-1} \right),$$

where $\widetilde{E}_n(\theta)$ is the imaginary part of

$$\sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{(n+1-2k)i\theta}.$$

Thus, the real part of

$$\left(C_{\lambda,n} \sum_{k=0}^{[(n+1)/2]\dagger} b_{n,k}\, e^{2ki\theta} \right) \left(\sum_{k=0}^{\infty} b_{n,k}\, e^{2ki\theta} \right)^{-1}$$

is

$$2^n\,E_{n+1}^\lambda(\cos\theta)\,Q_n^\lambda(\cos\theta)+\frac{\pi}{2}\,\widetilde{E}_n(\theta)\frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\;G_n^\lambda(\cos\theta)\,(\sin\theta)^{2\lambda-1},$$

which, together with (4.6), proves that

$$2^n\,E_{n+1}^\lambda(\cos\theta)\,Q_n^\lambda(\cos\theta)+\frac{\pi}{2}\,\widetilde{E}_n(\theta)\frac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\;G_n^\lambda(\cos\theta)\,(\sin\theta)^{2\lambda-1}>\frac{C_{\lambda,n}}{2}$$

for $\theta\in(0,\pi)$. In particular,

$$2^n\,E_{n+1}^\lambda(x_{n,i}^\lambda)\,Q_n^\lambda(x_{n,i}^\lambda)>\frac{C_{\lambda,n}}{2}>0,\quad i=1,\dots,n, \tag{4.7}$$

where the numbers $x_{n,i}^\lambda$, $i=1,\dots,n$, are the zeros of the ultraspherical polynomial G_n^λ. It is well known that the zeros of Q_n^λ and G_n^λ interlace since the functions $\widetilde{Q}_n^\lambda(x)=(1-x^2)^{1/2-\lambda}\,Q_n^\lambda(x)$ and $G_n^\lambda(x)$ satisfy the same differential equation (see [25, p. 78]). Thus, it follows from (4.7) that the zeros of E_{n+1}^λ and G_n^λ behave similarly.

4.2. Proof of Corollary 2.2. As we mentioned above, the interlacing property stated in Theorem 2.1 implies that the coefficients $B_{n,j}^\lambda$, $j=1,\dots,n+1$, are all positive. Regarding the coefficients $A_{n,i}^\lambda$, we have the representation (see [15, Theorem 2])

$$A_{n,i}^\lambda=\sigma_{n,i}^\lambda+\frac{d_{\lambda,n}}{\kappa_{\lambda,n}^2\,(G_n^\lambda)'(x_{n,i}^\lambda)\,E_{n+1}^\lambda(x_{n,i}^\lambda)},\quad i=1,\dots,n, \tag{4.8}$$

where $\sigma_{n,i}^\lambda$, $i=1,\dots,n$, are the Christoffel numbers corresponding to the weight function w_λ (recall that $d_{\lambda,n}<0$ is the leading coefficient of G_n^λ). Obviously

$$\sigma_{n,i}^\lambda=\frac{2\,\Gamma(\lambda+1/2)}{-\Gamma(2\lambda)}\left|\frac{Q_n^\lambda(x_{n,i}^\lambda)}{(G_n^\lambda)'(x_{n,i}^\lambda)}\right|,\quad i=1,\dots,n. \tag{4.9}$$

As the zeros of E_{n+1}^λ and G_n^λ interlace, it is clear that

$$(G_n^\lambda)'(x_{n,i}^\lambda)\,E_{n+1}^\lambda(x_{n,i}^\lambda)=|(G_n^\lambda)'(x_{n,i}^\lambda)\,E_{n+1}^\lambda(x_{n,i}^\lambda)|,\quad i=1,\dots,n.$$

On account of this equality and (4.7), (4.8), and (4.9) as well as the explicit formulae (3.1) and (3.2), we obtain

$$\begin{aligned}A_{n,i}^\lambda &> \left|\frac{Q_n^\lambda(x_{n,i}^\lambda)}{(G_n^\lambda)'(x_{n,i}^\lambda)}\right|\left(-\frac{2\,\Gamma(\lambda+1/2)}{\Gamma(2\lambda)}+\frac{2^{n+1}\,d_{\lambda,n}}{C_{\lambda,n}\,\kappa_{\lambda,n}^2}\right)\\ &= \left|\frac{Q_n^\lambda(x_{n,i}^\lambda)}{(G_n^\lambda)'(x_{n,i}^\lambda)}\right|\frac{2}{\Gamma(\lambda)}\left(-\frac{\Gamma(\lambda)\,\Gamma(\lambda+1/2)}{\Gamma(2\lambda)}+\frac{2\sqrt{\pi}}{4^\lambda}\right)=0,\end{aligned}$$

where we have used (cf. [25, formula (1.7.6)]) the functional equation

$$\Gamma(z)\,\Gamma(z+1/2)=\sqrt{\pi}\,2^{1-2z}\,\Gamma(2z),\quad z\neq 0,-1/2,-1,-3/2,-2,\dots,$$

in the last equality.

4.3. Proof of Theorem 2.3. For $\theta \in [0,\pi]$, we have

$$\begin{aligned} 2^n\, E_{n+1}^{\lambda}(\cos\theta) &= \Re\left(\sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right) \\ &- \Re\left(\sum_{k=[n/2+1]\ddagger}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right) \end{aligned}$$

because of (3.11). If $n \geq 2$, it is clear that

$$\left|\Re\left(\sum_{k=[n/2+1]\ddagger}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right)\right| \leq \sum_{k=[n/2+1]}^{\infty} b_{n,k} = \mathcal{O}(1/n)$$

due to Lemma 4.3. Thus, we have proved

$$2^n\, E_{n+1}^{\lambda}(\cos\theta) = \Re\left(\sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right) + \mathcal{O}(1/n), \tag{4.10}$$

uniformly on $[0,\pi]$.

Additionally, if $\theta \in (0,\pi)$, reasoning as in the proof of Theorem 2.1, we have

$$\begin{aligned} \sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta} &= \left(\lim_{r\to 1^-} \sum_{k=0}^{\infty} a_{n,k}\,(re^{i\theta})^{n+1+2k}\right)^{-1} \\ &= \frac{C_{\lambda,n}}{Q_n^{\lambda}(\cos\theta) - \dfrac{\pi i}{2}\,\dfrac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)\,(\sin\theta)^{2\lambda-1}} \\ &= \frac{C_{\lambda,n}\,(\sin\theta)^{1-2\lambda}}{\widetilde{Q}_n^{\lambda}(\cos\theta) - \dfrac{\pi i}{2}\,\dfrac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)} \\ &= (\sin\theta)^{1-2\lambda}\,\frac{C_{\lambda,n}\left(\widetilde{Q}_n^{\lambda}(\cos\theta) + \dfrac{\pi i}{2}\,\dfrac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)\right)}{\left[\widetilde{Q}_n^{\lambda}(\cos\theta)\right]^2 + \left[\dfrac{\pi}{2}\,\dfrac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)\right]^2}. \end{aligned}$$

Therefore, we obtain

$$\Re\left(\sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right) = \frac{C_{\lambda,n}\,(\sin\theta)^{1-2\lambda}\,\widetilde{Q}_n^{\lambda}(\cos\theta)}{\left[\widetilde{Q}_n^{\lambda}(\cos\theta)\right]^2 + \left[\dfrac{\pi}{2}\,\dfrac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)\right]^2} \tag{4.11}$$

for $\theta \in (0,\pi)$. Notice that the denominator in (4.11) cannot vanish since the zeros of $\widetilde{Q}_n^{\lambda}$ and G_n^{λ} interlace, as previously mentioned.

Fix $\epsilon > 0$. From (3.12) and (3.13), it readily follows that

$$\frac{1}{n^{2-2\lambda}}\,\frac{(\sin\theta)^{-2\lambda}}{\left[\widetilde{Q}_n^{\lambda}(\cos\theta)\right]^2 + \left[\dfrac{\pi}{2}\,\dfrac{\Gamma(2\lambda)}{\Gamma(\lambda+1/2)}\, G_n^{\lambda}(\cos\theta)\right]^2} = \frac{4^{1-\lambda}}{\pi} + \mathcal{O}(1/n),$$

uniformly on $[\epsilon, \pi - \epsilon]$. Additionally, $\lim_{n\to\infty} n^{1-\lambda} C_{\lambda,n} = \sqrt{\pi}$ is obtained with the help of (4.3). So, using (3.13) again and (4.11), we have

(4.12)
$$\Re\left(\sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right) = (2\sin\theta)^{1-\lambda}\, \cos\{(n+1)\theta + (\lambda-1)(\theta-\pi/2)\} + \mathcal{O}(1/n),$$

uniformly on $[\epsilon, \pi - \epsilon]$.

For all $n \in \mathbb{N}$, let $f_n(\theta)$, $\theta \in [0, \pi]$, denote the continuous function

$$\Re\left(\sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right),$$

whereas $g_n(\theta)$ stands for $(2\sin\theta)^{1-\lambda}\, \cos\{(n+1)\theta + (\lambda-1)(\theta-\pi/2)\}$, $\theta \in [0, \pi]$. We will reason by contradiction. Suppose that the sequence $\{f_n - g_n\}_{n\in\mathbb{N}}$ does not converge uniformly to 0 on $[0, \pi]$. So, there exist a sequence of points $\theta_n, n \in \Lambda \subset \mathbb{N}$, and a number $\epsilon_0 > 0$ such that

$$|f_n(\theta_n) - g_n(\theta_n)| \geq \epsilon_0 \tag{4.13}$$

for all $n \in \Lambda$. As the sequence $\{\theta_n\}_{n\in\Lambda}$ is included in $[0, \pi]$, we can assume that it is a convergent sequence. Moreover, as a result of (4.12), the limit point must be one of the two ends of the interval $[0, \pi]$. So, without loss of generality, we can also assume that

$$\lim_{n\in\Lambda} \theta_n = 0. \tag{4.14}$$

We will prove that given any $\epsilon > 0$ there exists $\delta > 0$ (depending only on ϵ) such that

$$|f_n(\theta) - g_n(\theta)| < \epsilon$$

for all $\theta \in [0, \delta)$ and any $n \in \mathbb{N}$. It is clear that this assertion contradicts (4.13) and (4.14).

Choose $\delta_1 > 0$ such that for all $\theta \in [0, \delta_1)$ it holds that

$$|g_n(\theta)| \leq |2\sin\theta|^{1-\lambda} < \epsilon/4, \quad n \in \mathbb{N}. \tag{4.15}$$

By the use of Lemma 4.3, fix $m \in \mathbb{N}$ such that

$$\sum_{k=m+1}^{\infty} b_{n,k} < \frac{\epsilon}{4}, \quad n \in \mathbb{N}. \tag{4.16}$$

Hence

$$-\frac{\epsilon}{4} < \sum_{k=0}^{m} b_{n,k} < 0, \quad n \in \mathbb{N}. \tag{4.17}$$

Next, choose $\delta_2 > 0$ such that for all $\theta \in [0, \delta_2)$ we have

$$|\sin(2k\theta)| < \epsilon/(8K), \quad k = 1, \ldots, m, \tag{4.18}$$

and

$$\cos(2k\theta) > 1 - \epsilon/(8K), \quad k = 1, \dots, m, \tag{4.19}$$

where K is the constant given by Lemma 4.2. Take $\delta = \min\{\delta_1, \delta_2\}$. Let $\theta \in [0, \delta)$. Then

$$\sum_{k=0}^{m} b_{n,k} \cos(2k\theta) < b_{n,1} \left(1 - \frac{\epsilon}{8K}\right) + \sum_{\substack{k=0 \\ k\neq 1}}^{m} b_{n,k} = \sum_{k=0}^{m} b_{n,k} - \frac{b_{n,1}\,\epsilon}{8K} < \frac{\epsilon}{16}, \tag{4.20}$$

where we have used (4.19) and (4.17) as well as the fact that $-b_{n,1} \leq K/2$ for all $n \in \mathbb{N}$. Analogously,

$$\sum_{k=0}^{m} b_{n,k} \cos(2k\theta) > b_{n,1} + \sum_{\substack{k=0 \\ k\neq 1}}^{m} b_{n,k} \left(1 - \frac{\epsilon}{8K}\right) = \sum_{k=0}^{m} b_{n,k} - \sum_{\substack{k=0 \\ k\neq 1}}^{m} b_{n,k} \frac{\epsilon}{8K} > -\frac{\epsilon}{4} - \frac{\epsilon}{16}. \tag{4.21}$$

Finally, for $\theta \in [0, \delta)$ and arbitrary $n \in \mathbb{N}$, we obtain

$$|f_n(\theta) - g_n(\theta)| \leq |f_n(\theta)| + |g_n(\theta)| < \left|\sum_{k=0}^{\infty} b_{n,k}\, e^{(n+1-2k)i\theta}\right| + \frac{\epsilon}{4}$$

$$\leq \left|\sum_{k=0}^{m} b_{n,k} \cos(2k\theta)\right| + \left|\sum_{k=0}^{m} b_{n,k} \sin(2k\theta)\right| + \sum_{k=m+1}^{\infty} b_{n,k} + \frac{\epsilon}{4} < \epsilon,$$

using (4.15) in the second step and (4.16), (4.18), (4.20), and (4.21) in the last one.

Once the contradiction has been established, we obtain $f_n(\theta) = g_n(\theta) + o(1)$, uniformly on $[0, \pi]$, which, together with (4.10), proves the theorem.

4.4. Proof of formula (2.2). It is well known that

$$\sum_{k=0}^{\infty} \frac{\Gamma(k+\lambda-1)}{\Gamma(k+1)\,\Gamma(\lambda-1)}\, z^k = (1-z)^{1-\lambda},$$

uniformly on $|z| \leq 1$ provided that $\lambda \leq 1$. Thus, if we put $z = e^{-2i\theta}$, we have

$$\sum_{k=0}^{[(n+1)/2]\dagger} \frac{\Gamma(k+\lambda-1)}{\Gamma(k+1)\,\Gamma(\lambda-1)}\, e^{i\theta(n+1-2k)} = \frac{e^{i\theta(n+1)}}{(1-e^{-2i\theta})^{\lambda-1}} + o(1),$$

uniformly on $[0, \pi]$. Now, the proof follows from taking real parts in the above expression and applying Theorem 2.3.

For $\lambda \in (1, 2)$ the proof is similar. In that case, by using Dirichlet's criterion on uniform convergence, we have

$$\sum_{k=0}^{\infty} \frac{\Gamma(k+\lambda-1)}{\Gamma(k+1)\,\Gamma(\lambda-1)}\, e^{-2ik\theta} = (1-e^{-2i\theta})^{1-\lambda},$$

uniformly on $[\epsilon, \pi - \epsilon]$. The rest is analogous.

4.5. Proof of Lemma 4.1. Obviously, $b_{n,0} = 1 > 0$ and $b_{n,1} = -a_{n,1} < -1$. Besides,

$$b_{n,2} = a_{n,1}^2 - a_{n,2} = a_{n,1}(a_{n,1}/a_{n,0} - a_{n,2}/a_{n,1}) > 0,$$

since the sequence $\{a_{n,k}/a_{n,k-1}\}_{k\in\mathbb{N}}$ is decreasing. Now, the idea is to obtain a formula similar to (3.9) without the term corresponding to $j = 1$ (which is negative) in order to carry out a proof by induction on the index $k \geq 2$. Thus, if we rewrite (3.9) using (4.4), we have

$$\sum_{\substack{j=0\\ j\neq 1}}^{k} b_{n,j}\, a_{n,k-j} + b_{n,1}\, \frac{\Gamma(k-\lambda)}{\Gamma(k)}\, \frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\Gamma(1-\lambda)}\, \frac{\Gamma(n+k)}{\Gamma(n+k+\lambda)} = 0. \tag{4.22}$$

Analogously, by replacing k with $k-1$, we obtain

$$\sum_{\substack{j=0\\ j\neq 1}}^{k-1} b_{n,j}\, a_{n,k-j-1} + b_{n,1}\, \frac{\Gamma(k-\lambda-1)}{\Gamma(k-1)}\, \frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\Gamma(1-\lambda)}\, \frac{\Gamma(n+k-1)}{\Gamma(n+k+\lambda-1)} = 0,$$

or, equivalently,

$$\begin{aligned}\sum_{\substack{j=0\\ j\neq 1}}^{k-1} b_{n,j}\, a_{n,k-j-1}\, \frac{(k-\lambda-1)(n+k-1)}{(k-1)(n+k+\lambda-1)} + b_{n,1} \left\{ \frac{\Gamma(k-\lambda-1)}{\Gamma(k-1)} \right.\\ \left. \times\, \frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\Gamma(1-\lambda)}\, \frac{\Gamma(n+k-1)}{\Gamma(n+k+\lambda-1)}\, \frac{(k-\lambda-1)(n+k-1)}{(k-1)(n+k+\lambda-1)} \right\} = 0.\end{aligned} \tag{4.23}$$

Subtracting (4.23) from (4.22), it follows that

$$\begin{aligned} b_{n,k} + \frac{\Gamma(n+\lambda+1)}{\Gamma(n+1)\Gamma(1-\lambda)} \sum_{\substack{j=0\\ j\neq 1}}^{k-1} \left[b_{n,j}\, \frac{\Gamma(k-j-\lambda)}{\Gamma(k-j)}\, \frac{\Gamma(n+k-j)}{\Gamma(n+k-j+\lambda)} \right.\\ \left. \times \left\{ \frac{(k-j-\lambda)(n+k-j)}{(k-j)(n+k-j+\lambda)} - \frac{(k-\lambda-1)(n+k-1)}{(k-1)(n+k+\lambda-1)} \right\} \right] = 0. \end{aligned}$$

The above expression may be written as

$$\begin{aligned} &\frac{\Gamma(-\lambda)(k-1)(n+k+\lambda-1)\Gamma(n+1)}{\Gamma(n+\lambda+1)}\, b_{n,k}\\ &+ \sum_{j=0}^{k-1} b_{n,j}\, (j-1)\, \frac{\Gamma(k-j-\lambda)}{\Gamma(k-j+1)}\, \frac{\Gamma(n+k-j)}{\Gamma(n+k-j+\lambda+1)}\, P_{n,\lambda}(k,j) = 0, \end{aligned}$$

where $P_{n,\lambda}(k,j) = n(n+k+\lambda-1) + (k-j)(n+2k-2)$. Therefore,

$$\sum_{j=0}^{k} b_{n,j}\, (j-1)\, \frac{\Gamma(k-j-\lambda)}{\Gamma(k-j+1)}\, \frac{\Gamma(n+k-j)}{\Gamma(n+k-j+\lambda+1)}\, P_{n,\lambda}(k,j) = 0. \tag{4.24}$$

As the term corresponding to $j = 0$ in (4.24) is now the only one which is negative, we then repeat the above argument in order to delete it. Replacing k with $k-1$ in (4.24), we obtain

$$\sum_{j=0}^{k-1} b_{n,j}\,(j-1)\,\frac{\Gamma(k-j-\lambda-1)}{\Gamma(k-j)}\,\frac{\Gamma(n+k-j-1)}{\Gamma(n+k-j+\lambda)}\,P_{n,\lambda}(k-1,j)=0.$$

Then

$$\begin{aligned}\sum_{j=0}^{k-1}\Bigg\{ & b_{n,j}\,(j-1)\,\frac{\Gamma(k-j-\lambda-1)}{\Gamma(k-j)}\,\frac{\Gamma(n+k-j-1)}{\Gamma(n+k-j+\lambda)}\\ & \times\frac{(k-\lambda-1)(n+k-1)}{k\,(n+k+\lambda)}\,P_{n,\lambda}(k-1,j)\,\frac{P_{n,\lambda}(k,0)}{P_{n,\lambda}(k-1,0)}\Bigg\}=0.\end{aligned}\tag{4.25}$$

Subtracting (4.25) from (4.24) gives

$$\begin{aligned}\frac{\Gamma(-\lambda)(k-1)(n+k+\lambda-1)\Gamma(n+1)}{\Gamma(n+\lambda+1)}\,b_{n,k} &= \sum_{j=2}^{k-1}\Bigg[\,b_{n,j}\,(j-1)\,\frac{\Gamma(k-j-\lambda-1)}{\Gamma(k-j)}\\ &\times\frac{\Gamma(n+k-j-1)}{\Gamma(n+k-j+\lambda)}\Bigg\{\frac{(k-\lambda-1)(n+k-1)}{k\,(n+k+\lambda)}\,\frac{P_{n,\lambda}(k-1,j)\,P_{n,\lambda}(k,0)}{P_{n,\lambda}(k-1,0)}\\ &-\frac{(k-j-\lambda-1)(n+k-j-1)}{(k-j)(n+k-j+\lambda)}P_{n,\lambda}(k,j)\Bigg\}\Bigg],\end{aligned}$$

which may be written as

$$\begin{aligned}&\frac{\Gamma(-\lambda)\,k\,(k-1)(n+k+\lambda)(n+k+\lambda-1)\Gamma(n+1)P_{n,\lambda}(k-1,0)}{\Gamma(n+\lambda+1)}\;b_{n,k}\\ &=\sum_{j=2}^{k-1}b_{n,j}\,(j-1)\,\frac{\Gamma(k-j-\lambda-1)}{\Gamma(k-j+1)}\,\frac{\Gamma(n+k-j-1)}{\Gamma(n+k-j+\lambda+1)}\,H_{n,\lambda}(k,j),\end{aligned}\tag{4.26}$$

where

$$\begin{aligned}H_{n,\lambda}(k,j) &= (k-j)(n+k-j+\lambda)(k-\lambda-1)(n+k-1)P_{n,\lambda}(k,0)P_{n,\lambda}(k-1,j)\\ &\quad -k\,(n+k+\lambda)(k-j-\lambda-1)(n+k-j-1)P_{n,\lambda}(k,j)P_{n,\lambda}(k-1,0).\end{aligned}$$

Therefore, a proof by induction may be carried out by the use of (4.26), provided that $H_{n,\lambda}(k,j) > 0$ for all $\lambda \in (-1/2, 0)$, $n \in \mathbb{N}$, and $k \geq 3$. The expression $H_{n,\lambda}(k,j)$ is a polynomial of degree 8 in the variables k, j, n, and λ which, apparently, cannot be factorized except for a factor j. It has 163 terms, about half of which are negative, and, additionally, it takes values arbitrarily close to zero. Despite these features, it is possible to show that $H_{n,\lambda}(k,j) > 0$ performing the following change of variables:

$$\left.\begin{array}{lcl} 1+2\lambda & = & \mu \\ n-1 & = & N \\ j-2 & = & i \\ k-j-1 & = & m \end{array}\right\} \quad\Leftrightarrow\quad \left\{\begin{array}{lcl} \lambda & = & (\mu-1)/2 \\ n & = & N+1 \\ j & = & i+2 \\ k & = & m+i+3 \end{array}\right.$$

$$(i+2)(144\,m^2N^4\mu+512\,m^2i\,\mu N^3+1296\,m^2N^3+384\,i^3m^2\mu\,N+32\,i\,N^5+48\,m\,N^5+960\,m^3N^2+640\,m^4i\,\mu\,N\;+\;768\,m^3i^2\mu\,N\;+\;56\,i^2N^4\;+\;256\,m^3N^3\mu\;+\;1305\,N\;+\;2880\,\mu\;+\;20032\,i\,m\,N\,\mu\;+15648\,m^2i\,\mu\,N+7984\,i^2m\,\mu\,N+2000\,\mu^2m\,i\,N+5248\,m^3i\,\mu\,N+4384\,i^2m^2\,\mu\,N+912\,\mu^2m^2i\,N+10208\,\mu\,N^2i\,m+2256\,i\,m\,N^3\mu+4944\,\mu\,m^2N^2i+2736\,\mu\,m\,N^2i^2+864\,\mu^2m\,N^2i+480\,\mu^2m\,N\,i^2+16\,\mu^3m\,i\,N+1280\,i^3m\,\mu\,N+9792\,\mu\,m+2244\,m^2N+2832\,m\,N+1752\,i\,N+13456\,\mu\,m^2+6768\,\mu\,N+4512\,\mu\,i+2712\,N^2+576\,\mu^2+1344\,\mu^2m+3396\,m^2N^2+768\,m^3N+5076\,m\,N^2+3148\,i\,N^2+872\,i^2N+2624\,i^2\mu+9600\,\mu\,m^3+1104\,\mu^2m^2+6050\,\mu\,N^2+1286\,\mu^2N+672\,\mu^2i+2816\,i\,m\,N+17632\,m\,N+13680\,\mu\,m\,i+4304\,i\,m\,N^2+11940\,\mu\,m\,N^2+2496\,\mu^2m\,N+1264\,\mu^2m\,i+17748\,m^2N\,\mu+1488\,m^2i\,N+15952\,m^2i\,\mu+912\,i^2m\,N+7024\,i^2m\,\mu+9252\,i\,N\,\mu+1872\,m^2N^2i+8484\,m^2N^2\mu\;+256\,m^3N\,i+240\,m^2N\,i^2+1152\,i^2m\,N^2+96\,i^3m\,N+6692\,i\,N^2\mu+4636\,i^2\mu\,N+1320\,\mu^2i\,N+1548\,\mu^2m\,N^2+1692\,\mu^2m^2N+48\,\mu^3m\,N+720\,\mu^2m^2i+368\,\mu^2m\,i^2+384\,m^2N^3i+1776\,m^2N^3\mu+256\,m^3N^2i+240\,m^2N^2i^2+2592\,m^3N^2\mu+1680\,m\,N^3i+3956\,m\,N^3\mu+2344\,i\,N^3\mu+2564\,i^2N^2\mu+1008\,i^3\mu\,N+440\,i^2\mu^2N+1568\,i^3\mu\,m+684\,\mu^2m^2N^2+480\,\mu^2m^3N+12\,\mu^3m^2N+128\,\mu^2m^3i+112\,\mu^2m^2i^2+892\,\mu^2N^2i+28\,\mu^3i\,N+1961\,N^3+96\,m^4N+2820\,m\,N^3+1756\,i\,N^3+1324\,i^2N^2+192\,i^3N+2820\,\mu\,N^3+1110\,\mu^2N^2+672\,\mu\,i^3+48\,\mu^3N+256\,\mu^2i^2+384\,\mu^2m^3+144\,m^2N^4+192\,m^3N^3+96\,m^4N^2+624\,m\,N^4+392\,i\,N^4+508\,i^2N^3+240\,i^3N^2+16\,i^4N+64\,i^4\mu+48\,\mu^2m^4+770\,\mu\,N^4+510\,\mu^2N^3+654\,N^4+78\,\mu^3N^2+32\,\mu^2i^3+3760\,m^4\mu+192\,i\,m\,N^4+240\,i^2m\,N^3+96\,i^3m\,N^2+672\,\mu\,m\,N^4+444\,\mu^2m\,N^3+60\,\mu^3m\,N^2+32\,\mu^2m\,i^3+2064\,m^4N\,\mu+2432\,m^4i\,\mu+2688\,m^3i^2\mu+1152\,m^2i^3\mu+128\,i^4m\,\mu+8672\,m^3N\,\mu+8960\,m^3i\,\mu+6704\,i^2m^2\mu+768\,m^3i\,\mu\,N^2+688\,m^2i^2\mu\,N^2+192\,m^2i\,\mu^2N^2+304\,i^2m\,\mu\,N^3+224\,i^3m\,\mu\,N^2+112\,i^2m\,\mu^2N^2+128\,\mu^2m\,i\,N^3+16\,\mu^3m\,i\,N^2+192\,\mu N^4i\,m+128\,\mu^2m^3i\,N+112\,\mu^2m^2i^2N+32\,\mu^2m\,i^3N+96\,m^2N^3\mu^2+424\,i\,N^4\mu+608\,i^2N^3\mu+276\,i\,N^3\mu^2+384\,i^3\mu\,N^2+220\,i^2\mu^2N^2+36\,\mu^3i\,N^2+48\,\mu^2m\,N^4+12\,\mu^3m\,N^3+96\,\mu^2m^3N^2+12\,\mu^3m^2N^2+288\,m^4N^2\mu+80\,\mu\,i^4N+48\,\mu^2i^3N+4\,\mu^3i^2N+64\,i^4m\,\mu\,N+768\,\mu\,m^5+48\,\mu^2m^4N+48\,m\,N^5\mu+32\,i\,N^5\mu+56\,i^2N^4\mu+32\,i\,N^4\mu^2+48\,i^3\mu\,N^3+16\,i^4\mu\,N^2+16\,i^3\mu^2N^2+36\,i^2\mu^2N^3+4\,i^2\mu^3N^2+8\,\mu^3N^3i+108\,N^5+8\,N^6+48\,i^3N^3+16\,i^4N^2+120\,\mu\,N^5+122\,\mu^2N^4+36\,\mu^3N^3+2\,\mu^4N^2+8\,\mu\,N^6+12\,\mu^2N^5+6\,\mu^3N^4+\mu^4N\;+\mu^4N^3+192\,m^5\mu\,N+256\,i^3\mu\,m^3+64\,i^4\mu\,m^2+256\,m^5i\,\mu+384\,m^4i^2\mu+64\,m^6\mu)/16$$

FIG. 4.1. $H_{n,\lambda}(k,j) = H_{N+1,(\mu-1)/2}(m+i+3, i+2)$.

Note that $\mu \in (0,1)$, $N \geq 0$, $i \geq 0$, $m \geq 0$. Thus

$$\begin{aligned}
H_{n,\lambda}(k,j) &= H_{N+1,(\mu-1)/2}(m+i+3, i+2)\\
&= (m+i+5/2-\mu/2)(N+m+i+3)(m+1)(N+m+3/2+\mu/2)\\
&\times\{(N+1)(N+m+i+5/2+\mu/2)+(m+i+3)(N+2m+2i+5)\}\\
&\times\{(N+1)(N+m+i+3/2+\mu/2)+m\,(N+2m+2i+3)\}\\
&-(m+1/2-\mu/2)\,(m+i+3)\,(N+m+1)\,(N+m+i+7/2+\mu/2)\\
&\times\{(N+1)(N+m+i+5/2+\mu/2)+(m+1)(N+2m+2i+5)\}\\
&\times\{(N+1)(N+m+i+3/2+\mu/2)+(m+i+2)\,(N+2m+2i+3)\}\,.
\end{aligned}$$

The resulting polynomial is of a similar complexity as the previous one. Nevertheless, a cumbersome calculation shows that all its coefficients are positive, which proves that $H_{n,\lambda}(k,j) \geq 0$. In order to help those readers interested in checking this last step of the proof, we include Figure 4.1 displaying the above-mentioned polynomial which has 197 terms as well as the factor $(i+2)/16$. The presence of the summand $2880\,\mu$ (second line in Fig. 4.1) guarantees that $H_{n,\lambda}(k,j) > 0$ for $\lambda \in (-1/2, 0)$, $n \in \mathbb{N}$, and $k \geq 3$.

Acknowledgment. The second author thanks Prof. Miodrag M. Spalević for providing numerical evidence of the results at an early stage of the work.

REFERENCES

[1] M. Bello Hernández, B. de la Calle Ysern, J. J. Guadalupe Hernández, and G. López Lagomasino, *Asymptotics for Stieltjes polynomials, Padé approximants, and Gauss–Kronrod quadrature*, J. Anal. Math., 86 (2002), pp. 1–23.

[2] S. Ehrich, *Asymptotic properties of Stieltjes polynomials and Gauss–Kronrod quadrature formulae*, J. Approx. Theory, 82 (1995), pp. 287–303.

[3] S. Ehrich, *Asymptotic behaviour of Stieltjes polynomials for ultraspherical weight functions*, J. Comput. Appl. Math., 65 (1995), pp. 135–144.

[4] S. Ehrich, *On product integration with Gauss–Kronrod nodes*, SIAM J. Numer. Anal., 35 (1998), pp. 78–92.

[5] S. Ehrich, *Stieltjes polynomials and the error of Gauss–Kronrod quadrature formulas*, in Applications and Computation of Orthogonal Polynomials, W. Gautschi, G. Golub, and G. Opfer, eds., Proc. Conf. Oberwolfach, Internat. Ser. Numer. Math. 131, Birkhäuser, Basel, 1999, pp. 57–77.
[6] S. Ehrich and G. Mastroianni, *Stieltjes polynomials and Lagrange interpolation*, Math. Comp., 66 (1997), pp. 311–331.
[7] P. Favati, G. Lotti, and F. Romani, *Algorithm* 691: *Improving QUADPACK automatic integration routines*, ACM Trans. Math. Software, 17 (1991), pp. 218–232.
[8] W. Gautschi, *Gauss–Kronrod quadrature—a survey*, in Numerical Methods and Approximation Theory III, G. V. Milovanović, ed., University of Niš, Niš, 1988, pp. 39–66.
[9] W. Gautschi and S. E. Notaris, *An algebraic study of Gauss–Kronrod quadrature formulae for Jacobi weight functions*, Math. Comp., 51 (1988), pp. 231–248.
[10] I. S. Gradshteyn and I. M. Ryzhik, *Table of Integrals, Series, and Products*, 6th ed., A. Jeffrey, ed., Academic Press, San Diego, 2000.
[11] Ch. Hermite and T. J. Stieltjes, *Correspondance d'Hermite et de Stieltjes*, Prentice–Hall, Englewood Cliffs, NJ, 1966.
[12] H. S. Jung, *Hermite and Hermite–Fejér interpolation for Stieltjes polynomials*, Math. Comp., 75 (2006), pp. 743–766.
[13] T. Kaluza, *Über die Koeffizienten reziproker Potenzreihen*, Math. Z., 28 (1928), pp. 161–170.
[14] A. S. Kronrod, *Nodes and Weights for Quadrature Formulae. Sixteen-Place Tables*, Nauka, Moscow, 1964 (in Russian).
[15] G. Monegato, *A note on extended Gaussian quadrature rules*, Math. Comp., 30 (1976), pp. 812–817.
[16] G. Monegato, *Positivity of weights of extended Gauss–Legendre quadrature rules*, Math. Comp., 32 (1978), pp. 243–245.
[17] G. Monegato, *Stieltjes polynomials and related quadrature rules*, SIAM Rev., 24 (1982), pp. 137–158.
[18] S. E. Notaris, *An overview of results on the existence or nonexistence and the error term of Gauss–Kronrod quadrature formulae*, in Approximation and Computation: A Festschrift in Honor of Walter Gautschi, R. V. M. Zahar, ed., Internat. Ser. Numer. Math. 119, Birkhäuser, Basel, 1994, pp. 485–496.
[19] F. Peherstorfer, *On the asymptotic behaviour of functions of second kind and Stieltjes polynomials, and on Gauss–Kronrod quadrature formulas*, J. Approx. Theory, 70 (1992), pp. 156–190.
[20] F. Peherstorfer, *Stieltjes polynomials and functions of second kind*, J. Comput. Appl. Math., 65 (1995), pp. 319–338.
[21] F. Peherstorfer and K. Petras, *Ultraspherical Gauss–Kronrod quadrature is not possible for* $\lambda > 3$, SIAM J. Numer. Anal., 37 (2000), pp. 927–948.
[22] F. Peherstorfer and K. Petras, *Stieltjes polynomials and Gauss–Kronrod quadrature for Jacobi weight functions*, Numer. Math., 95 (2003), pp. 689–706.
[23] R. Piessens, E. de Doncker-Kapenga, C. W. Überhuber, and D. K. Kahaner, *QUADPACK: A Subroutine Package for Automatic Integration*, Springer Ser. Comput. Math. 1, Springer-Verlag, Berlin, 1983.
[24] G. Szegő, *Über gewisse orthogonale Polynome, die zu einer oszillierenden Belegungsfunktion gehören*, Math. Ann., 110 (1935), pp. 501–513.
[25] G. Szegő, *Orthogonal Polynomials*, 4th ed., Coll. Publ. XXIII, AMS, Providence, RI, 1975.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 787–809

FOSLL* METHOD FOR THE EDDY CURRENT PROBLEM WITH THREE-DIMENSIONAL EDGE SINGULARITIES*

EUNJUNG LEE† AND THOMAS A. MANTEUFFEL†

Abstract. In the case that the domain has reentrant edges, the standard finite element method loses its global accuracy because of singularities on the boundary. To overcome this difficulty, FOSLL* is applied in this paper. FOSLL* is a methodology for solving PDEs using the dual operator. Here, a modified FOSLL* method is developed that employs a partially weighted functional and allows the use of a standard finite element scheme without losing global accuracy.

Key words. eddy current problem, singularities, least squares, finite element methods

AMS subject classification. 65N30

DOI. 10.1137/050647001

1. Introduction. The Maxwell equations are a set of fundamental equations governing all macroscopic electromagnetic phenomena. It is known that the numerical resolution of the full system of the Maxwell equations can be very expensive. However, it is possible to use a simplified model that approximates the Maxwell equations and explains particular problems encountered in electromagnetism. In many cases, one can use the so-called eddy current model, which is obtained by neglecting the displacement current in the Maxwell equations. Here, we consider the following two basic laws of electricity and magnetism, which form the eddy current model:

$$\text{Faraday's Law}: \quad \frac{\partial \mu \mathbf{H}}{\partial t} + \nabla \times \mathbf{E} = \mathbf{0},$$

$$\text{Ampère's Law}: \quad \nabla \times \mathbf{H} - \sigma \mathbf{E} = \mathbf{0},$$

where $\mathbf{E}$ is the electric field intensity, $\mathbf{H}$ is the magnetic field intensity, μ is the permeability, and σ is the conductivity. We consider two types of boundary conditions

$$\mathbf{n} \times \mathbf{E} = \mathbf{0}, \quad \mathbf{n} \cdot \mathbf{H} = 0, \qquad \text{and} \qquad \mathbf{n} \cdot \mathbf{E} = 0, \quad \mathbf{n} \times \mathbf{H} = \mathbf{0},$$

where $\mathbf{n}$ is the unit external normal vector. The electric and magnetic field intensities, $\mathbf{E}$ and $\mathbf{H}$, which follow Faraday's and Ampère's laws with homogeneous boundary conditions, satisfy

$$\mathbf{E} \in H_0(\nabla\times) \cap H(\nabla \cdot \sigma), \quad \mathbf{H} \in H(\nabla\times) \cap H_0(\nabla \cdot \mu)$$

or

$$\mathbf{E} \in H(\nabla\times) \cap H_0(\nabla \cdot \sigma), \quad \mathbf{H} \in H_0(\nabla\times) \cap H(\nabla \cdot \mu).$$

For a precise definition of the above Sobolev spaces, see section 2. In addition, if

*Received by the editors December 7, 2005; accepted for publication (in revised form) October 4, 2006; published electronically April 20, 2007. This work was sponsored by the National Science Foundation under grant number DMS–0420873, and the Department of Energy under grant numbers DE–FC02-01ER25479 and DE–FG02-03ER25574.

http://www.siam.org/journals/sinum/45-2/64700.html

†Department of Applied Mathematics, University of Colorado at Boulder, Campus Box 526, Boulder, CO 80309-0526 (eunjung@colorado.edu, tmanteuf@colorado.edu).

- μ and σ are smooth,
- either the domain is a convex polyhedron or the boundary is $\mathcal{C}^{1,1}$, and
- different types of boundary conditions do not meet at an edge with the internal angle $> \pi/2$,

then $\mathbf{E} \in (H^1)^3$ and $\mathbf{H} \in (H^1)^3$. Standard numerical techniques can be used to approximately solve the equations under the above smoothness assumptions. For example, first-order system least squares (FOSLS) with H^1-finite element spaces and multigrid methods can be used to solve these equations efficiently (cf. [3], [4], [14]). The FOSLS method is based on minimization of the squared residual norm, $||L\mathbf{V} - \mathbf{F}||_0^2$, of the system $L\mathbf{U} = \mathbf{F}$, where L represents a system of linear first-order equations, $\mathbf{U}$ a vector of unknowns, and $\mathbf{F}$ a vector of known functions. The standard least squares method approximates unknown $\mathbf{U}$ in the given H^1-finite element space when the bilinear form corresponding to $||L\mathbf{V} - \mathbf{F}||_0^2$ is equivalent to the product H^1-norm, and this H^1-equivalence is provided under sufficient smoothness assumptions on the domain, coefficients, and data of the original problem.

In the presence of discontinuous coefficients, nonsmooth, nonconvex domain, or certain irregular boundary conditions, the solution may not be in H^1. This precludes the use of H^1-conforming finite element spaces in least squares and Galerkin formulations of the Maxwell equations.

A partial list of the remedies for this loss of H^1-regularity in FOSLS can be found in [1], [5], [18], and [24]. In [5], the first-order system LL* (FOSLL*) method was introduced to overcome the difficulty that arises from discontinuous coefficients. The basic idea of the FOSLL* method can be explained by looking at a linear system of equations, $A\mathbf{x} = \mathbf{b}$. The least squares method minimizes $||A\mathbf{x} - \mathbf{b}||_0^2$, which leads to the normal equations $A^tA\mathbf{x} = A^t\mathbf{b}$. The dual of this method involves the system, $AA^t\mathbf{y} = \mathbf{b}$, where $\mathbf{x} = A^t\mathbf{y}$. FOSLL* solves $AA^t\mathbf{y} = \mathbf{b}$ by minimizing the functional $\langle A^t\mathbf{y}, A^t\mathbf{y}\rangle - 2\langle \mathbf{y}, \mathbf{b}\rangle$ which is equivalent to minimizing $||A^t\mathbf{y} - \mathbf{x}||_0^2$. For a given first-order linear system of PDEs, $L\mathbf{U} = \mathbf{F}$, the FOSLL* method solves the system, $LL^*\mathbf{U}^* = \mathbf{F}$, by minimizing the functional, $||L^*\mathbf{U}^* - \mathbf{U}||_0^2$, with the dual variable, $\mathbf{U}^*$, and the L^2-adjoint operator, L^*, of L. Minimizing $||L^*\mathbf{U}^* - \mathbf{U}||_0^2$ over $\mathbf{U}^*$ in the domain of L^* is accomplished by solving the weak problem of finding $\mathbf{U}^*$ such that

$$\langle\, L^*\mathbf{U}^*, L^*\mathbf{V}\,\rangle \;=\; \langle\, \mathbf{U}, L^*\mathbf{V}\,\rangle \;=\; \langle\, L\mathbf{U}, \mathbf{V}\,\rangle \;=\; \langle\, \mathbf{F}, \mathbf{V}\,\rangle \tag{1.1}$$

for every $\mathbf{V}$ in the domain of L^*. Then, the solution we seek is $\mathbf{U} = L^*\mathbf{U}^*$. The equation in (1.1) shows that we can solve the dual problem with the given data (right-hand side) of the original problem without knowing the exact solution, $\mathbf{U}$.

In [18], a modified FOSLL* method was developed that allows an accurate approximation using H^1-conforming finite elements for the equations having singular boundary points in two dimension. The results in [24] established a modification of the FOSLS method for the problem in a two-dimensional nonconvex domain having irregular boundary conditions. A weighted norm was used in [24] in order to reduce the difficulties from dealing with the absence of the smoothness of the problem. As a different type of remedy, one of the most common approaches is to use Raviart–Thomas or Nédélec edge elements as a finite element space [20]. These Raviart–Thomas and Nédélec edge element spaces are in $H(\nabla\cdot)$ and $H(\nabla\times)$, respectively, but not in $(H^1)^3$. Another potential form to reduce the difficulties from low regularity of the solution was introduced in [2]. The analysis in [2] is based on a weak variational formulation; the authors employ an H^{-1}-norm least-squares approach in discrete space to avoid dealing with the inf-sup condition. In [10], weighted regularization of time

harmonic Maxwell equations in a polyhedral domain using a Galerkin formulation was investigated. Introducing special weights inside the divergence integral allows the approximation of nonsmooth solutions by an H^1-conforming finite element. Error estimates under the assumptions of special finite element spaces were established in [10].

As mentioned above, modifications to FOSLL* were developed that effectively handle discontinuous coefficients in two and three dimensions and irregular boundary points in two dimensions. However, there has not been any previous attempt to use FOSLL* to handle the difficulty from reentrant edges in three dimensions. First, we use standard FOSLL* to abate the difficulties from discontinuous coefficients, and then modify it to deal with the reentrant edges. We develop a modified FOSLL* using partially weighted norms in the functional to be minimized, so that we can use H^1-conforming finite elements. We do not consider the case that different types of boundary conditions meet at an edge with an internal angle greater than $\pi/2$ or the case in which the domain has conical points and vertices, where several reentrant edges meet. However, we believe that the approach developed here can be easily extended to those cases.

The approximate solution that the FOSLL* approach produces is of the form L^*U^h, where U^h is an H^1-conforming finite element. This approximation contains the curl-free Nédélec edge elements. Our approach involves a substantial decrease in computational cost over the curl-curl formulation because it is easy to implement and the resulting linear systems are easily solved by algebraic multigrid methods [23] even with higher order elements. We obtain the same error estimates as the Nédélec element approach in the L^2-norm and we can easily extend our approach to obtain the $H(\nabla\times)$-norm, while the approach in [2] provides only an L^2-error estimate. Moreover, we obtain error estimates in $H(\nabla\cdot\mu)$- and $H(\nabla\cdot\sigma)$-norms, too.

There are similarities between the FOSLL* approach developed here and the Galerkin formulation with the weighted regularization presented in [10]. While FOSLL* differs in many respects from a Galerkin formulation, under special circumstances we show that they are equivalent (see section 5). In [10], σ was assumed to be constant and E was approximated. It is easy to see that if μ were assumed constant, the same approach could be used to approximate for H. FOSLL* allows both σ and μ to be discontinuous in a natural way. We obtain the same error estimates as the approach in [10] while employing any standard H^1-conforming finite element spaces.

In this paper, we consider the Maxwell equations with discontinuous coefficients and irregular boundary. The error estimates established here hold for standard H^1-conforming finite element spaces and provide convergence rates that depend on the power of the weighting used. Numerical tests show surprising agreement with the theory. The model problem is given in section 2. In section 3, we introduce the FOSLS and FOSLL* methods briefly and explain the difficulties arising from singularities. In section 4, we modify standard FOSLL* and show that H^1-conforming elements can be used. A scaling is introduced and the connection to the Galerkin formulation in a special case is explored in section 5. In section 6, the discretization error estimates are obtained. The numerical results are given in section 7.

2. Model problem. Let Q be a polygon in $\mathbb{R}^2$ with a reentrant corner, that is, a corner that has inner angle bigger than π. Let $I \in \mathbb{R}$ be a bounded interval, and consider the prototype domain, $\Omega := Q \times I \subset \mathbb{R}^3$, which is a polyhedral cylinder. In this paper, we restrict ourselves to the case where the domain has one reentrant edge; however, the general case follows easily. By translation and rotation, we may suppose

that the reentrant edge on the boundary that induces the singularity is on the z-axis.

Throughout this paper, we use $\langle\cdot,\cdot\rangle$ and $||\cdot||$ to denote the L^2-inner product and norm, respectively. We use $||\cdot||_k$ to denote the standard Sobolev H^k-norm and $|\cdot|_k$ to denote the seminorm in $H^k(\Omega)$. Let $b \in L^\infty(\Omega)$ be a scalar function, and define

$$H_0(\nabla\times)\cap H(\nabla\cdot\, b) := \{\mathbf{u}\in L^2(\Omega)^3 \mid ||\nabla\times\mathbf{u}||^2+||\nabla\cdot b\mathbf{u}||^2<\infty,\ \mathbf{n}\times\mathbf{u}=\mathbf{0} \text{ on } \partial\Omega\},$$

$$H(\nabla\times)\cap H_0(\nabla\cdot\, b) := \{\mathbf{u}\in L^2(\Omega)^3 \mid ||\nabla\times\mathbf{u}||^2+||\nabla\cdot b\mathbf{u}||^2<\infty,\ \mathbf{n}\cdot\mathbf{u}=0 \text{ on } \partial\Omega\}.$$

Define $H^k_\beta(\Omega)$ as the weighted Sobolev space of functions u such that

$$||u||^2_{k,\beta} = \sum_{|m|=0}^{k}\int_\Omega r^{2(\beta+|m|-k)}\,|D^m u|^2\,d\Omega<\infty,$$

where $r := r(\mathbf{x})$ is the distance of $\mathbf{x}\in\Omega$ from the reentrant edge. We define partially weighted norms to use in our modification of the FOSLL* functional, for $\mathbf{u},\mathbf{v}\in L^2(\Omega)^3$ and $p,q\in H^0_\beta(\Omega)$, as

$$||(\mathbf{u}^t,p)^t||^2_\beta := ||\mathbf{u}||^2+||p||^2_{0,\beta}, \tag{2.1}$$

$$||(\mathbf{u}^t,p,\mathbf{v}^t,q)^t||^2_\beta := ||\mathbf{u}||^2+||p||^2_{0,\beta}+||\mathbf{v}||^2+||q||^2_{0,\beta}. \tag{2.2}$$

In the above, note that only the scalar terms, p and q, involve weighted norms.

Now, consider the following eddy current problem:

$$\begin{aligned}\frac{\partial\mu\mathbf{H}}{\partial t}+\nabla\times\mathbf{E}&=\mathbf{0}\quad\text{in}\quad\Omega,\\ \nabla\times\mathbf{H}-\sigma\mathbf{E}&=\mathbf{0}\quad\text{in}\quad\Omega,\end{aligned} \tag{2.3}$$

with $\mathbf{E}(\mathbf{x},t)$ the electric field intensity, $\mathbf{H}(\mathbf{x},t)$ the magnetic field intensity, $\mu(\mathbf{x})$ the permeability, and $\sigma(\mathbf{x})$ the conductivity. We assume that coefficients $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are piecewise smooth, positive real valued, and bounded; that is, they satisfy

$$\mu_0\le\mu(\mathbf{x})\le\mu_1,\quad \sigma_0\le\sigma(\mathbf{x})\le\sigma_1\quad\text{for all }\mathbf{x}\in\overline{\Omega}, \tag{2.4}$$

for positive constants μ_0, μ_1, σ_0, and σ_1. We consider two types of boundary conditions,

$$\begin{aligned}\text{type I}:&\quad \mathbf{n}\times\mathbf{E}=\mathbf{0},\quad \mathbf{n}\cdot\mathbf{H}=0,\\ \text{type II}:&\quad \mathbf{n}\cdot\mathbf{E}=0,\quad \mathbf{n}\times\mathbf{H}=\mathbf{0}.\end{aligned}$$

Type I corresponds to perfectly conducting walls, while type II corresponds to perfectly insulating walls. Using the backward Euler approximation in time gives

$$\frac{\mu}{\delta t}\mathbf{H}+\nabla\times\mathbf{E}=\frac{\mu}{\delta t}\mathbf{H}_{\text{old}},$$

where $\mathbf{H}_{\text{old}}$ is the solution at the previous time step. Equation (2.3) implies

$$\nabla\cdot\sigma\mathbf{E}=0,\qquad \nabla\cdot\mu\mathbf{H}=\nabla\cdot\mu\mathbf{H}_{\text{old}}. \tag{2.5}$$

Without loss of generality, we assume $\nabla \cdot \mu \mathbf{H}_{\text{old}} = 0$. The resulting system then is

$$\begin{aligned} -\sigma\mathbf{E} + \nabla \times \mathbf{H} &= \mathbf{0}, \quad \nabla \times \mathbf{E} + \tilde{\mu}\mathbf{H} = \tilde{\mu}\mathbf{H}_{\text{old}}, \\ \nabla \cdot \sigma\mathbf{E} &= 0, \quad \nabla \cdot \mu\mathbf{H} = 0, \end{aligned} \tag{2.6}$$

where $\tilde{\mu} = \mu/\delta t$. Since δt is a constant, $\nabla \cdot \tilde{\mu}\mathbf{H} = 0$. Let δt^{-1} be absorbed into μ and $\tilde{\mu}$ be replaced with μ. It is known that there exists a solution, $(\mathbf{E}, \mathbf{H})$, of the system (2.6) in $(H(\nabla\times) \cap H(\nabla \cdot \sigma)) \times (H(\nabla\times) \cap H(\nabla \cdot \mu))$ satisfying type I or II boundary conditions (cf. [14]). From now on, we consider only the case that the domain is surrounded by perfectly conducting walls, since the procedure is the same for perfectly insulating walls. Moreover, the case of mixed boundary conditions can be handled in a similar fashion.

In this paper, c is a generic term that is used to denote various constants. Its dependence on other quantities is indicated when necessary. For convenience of notation, superscript t for the vector transpose is omitted.

3. FOSLS and FOSLL*. In this section, we give a brief introduction to FOSLS and FOSLL* to explain the basic ideas and to show how they suffer in the presence of singularities. First, we introduce slack variables. Even though system (2.6) can be solved by itself, we extend the system since the extended system provides H^1-equivalence to the bilinear form of $||L^*\mathbf{U}^* - \mathbf{U}||$ in FOSLL* under sufficient smoothness assumptions. We extend system (2.6) by adding slack variables, s and k, to yield

$$\begin{array}{rcrcrcll} -\sigma\mathbf{E} & & + & \nabla \times \mathbf{H} & - & \nabla k & = \mathbf{0} & \text{in } \Omega, \\ & - a_1 s & + & \nabla \cdot \mu\mathbf{H} & & & = 0 & \text{in } \Omega, \\ \nabla \times \mathbf{E} & - \nabla s & + & \mu\mathbf{H} & & & = \mu\mathbf{H}_{\text{old}} & \text{in } \Omega, \\ \nabla \cdot \sigma\mathbf{E} & & & & + & a_2 k & = 0 & \text{in } \Omega, \end{array}$$

$$\mathbf{n} \times \mathbf{E} = \mathbf{0}, \quad \mathbf{n} \cdot \mathbf{H} = 0, \quad k = 0 \quad \text{on } \partial\Omega,$$

with nonnegative constants a_1 and a_2. The above system can be rewritten as

$$\mathcal{L}\mathbf{U} = \mathcal{L}(\mathbf{E}, s, \mathbf{H}, k) = \mathbf{F} \quad \text{in } \Omega,$$

where

$$\mathcal{L}\mathbf{U} = \begin{bmatrix} -\sigma I & 0 & \nabla\times & -\nabla \\ 0 & -a_1 & \nabla \cdot \mu & 0 \\ \nabla\times & -\nabla & \mu I & 0 \\ \nabla \cdot \sigma & 0 & 0 & a_2 \end{bmatrix} \begin{bmatrix} \mathbf{E} \\ s \\ \mathbf{H} \\ k \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \\ \mu\mathbf{H}_{\text{old}} \\ 0 \end{bmatrix} = \mathbf{F}. \tag{3.1}$$

The domain of $\mathcal{L}$ is

$$D(\mathcal{L}) = (H_0(\nabla\times) \cap H(\nabla \cdot \sigma)) \times H^1(\Omega)/\mathbb{R} \times (H(\nabla\times) \cap H_0(\nabla \cdot \mu)) \times H_0^1(\Omega),$$

which is a Hilbert space under the norm

$$\begin{aligned} ||(\mathbf{E}, s, \mathbf{H}, k)||_{\mathcal{L}}^2 &:= ||\mathbf{E}||^2 + ||\nabla \times \mathbf{E}||^2 + ||\nabla \cdot \sigma\mathbf{E}||^2 + ||s||_1^2 \\ &\quad + ||\mathbf{H}||^2 + ||\nabla \times \mathbf{H}||^2 + ||\nabla \cdot \mu\mathbf{H}||^2 + ||k||_1^2. \end{aligned} \tag{3.2}$$

The range of $\mathcal{L}$ is $L^2(\Omega)^8$. It is easily shown that $s = 0$ and $k = 0$ if $(\mathbf{E}, s, \mathbf{H}, k)$ is the solution of (3.1) in $D(\mathcal{L})$ as long as the constants a_1 and a_2 are nonnegative. The FOSLS method minimizes the least-squares functional

$$\mathcal{F}(\mathbf{U}; \mathbf{F}) = || \ \mathcal{L}\mathbf{U} - \mathbf{F} \ ||^2$$

in the weak sense, that is, we look for the solution of the corresponding weak form as follows: Find $\mathbf{U} \in D(\mathcal{L})$ satisfying

$$\langle \ \mathcal{L}\mathbf{U}, \mathcal{L}\mathbf{V} \ \rangle \ = \ \langle \ \mathbf{F}, \mathcal{L}\mathbf{V} \ \rangle \qquad \text{for all } \ \mathbf{V} \in D(\mathcal{L}). \tag{3.3}$$

The FOSLL* approach solves the corresponding dual problem

$$\mathcal{L}^*\mathbf{U}^* = \mathcal{L}^*(\mathcal{U}, p, \mathcal{V}, q) = \mathbf{U} \quad \text{in } \Omega, \tag{3.4}$$

where the L^2-adjoint operator $\mathcal{L}^*$ of $\mathcal{L}$ is defined by

$$\mathcal{L}^*\mathbf{U}^* = \begin{bmatrix} -\sigma I & 0 & \nabla\times & -\sigma\nabla \\ 0 & -a_1 & \nabla\cdot & 0 \\ \nabla\times & -\mu\nabla & \mu I & 0 \\ \nabla\cdot & 0 & 0 & a_2 \end{bmatrix} \begin{bmatrix} \mathcal{U} \\ p \\ \mathcal{V} \\ q \end{bmatrix} = \begin{bmatrix} \mathbf{E} \\ s \\ \mathbf{H} \\ k \end{bmatrix} = \mathbf{U}, \tag{3.5}$$

and $\mathcal{L}^* : D(\mathcal{L}^*) \to L^2(\Omega)^8$ with

$$D(\mathcal{L}^*) = (H_0(\nabla\times) \cap H(\nabla\cdot)) \times H^1(\Omega)/\mathbb{R} \times (H(\nabla\times) \cap H_0(\nabla\cdot)) \times H_0^1(\Omega).$$

To solve the dual problem we minimize the dual functional

$$\mathcal{F}^*(\mathbf{U}^*; \mathbf{U}) = || \ \mathcal{L}^*\mathbf{U}^* - \mathbf{U} \ ||^2 \tag{3.6}$$

on $D(\mathcal{L}^*)$. The corresponding weak form is the following: Find $\mathbf{U}^* \in D(\mathcal{L}^*)$ satisfying

$$\langle \ \mathcal{L}^*\mathbf{U}^*, \mathcal{L}^*\mathbf{V}^* \ \rangle = \langle \ \mathbf{U}, \mathcal{L}^*\mathbf{V}^* \ \rangle = \langle \ \mathbf{F}, \mathbf{V}^* \ \rangle \qquad \text{for all } \ \mathbf{V}^* \in D(\mathcal{L}^*), \tag{3.7}$$

where $\mathbf{U}$ is the solution of (3.1) for given $\mathbf{F}$. Equation (3.7) shows that we can solve the dual problem with the given data, $\mathbf{F}$, of the original problem without knowing the solution, $\mathbf{U}$. Then, we obtain the solution from (3.4), $\mathbf{U} = \mathcal{L}^*\mathbf{U}^*$.

LEMMA 3.1. *There exists a unique solution,* $\mathbf{U} \in D(\mathcal{L})$*, satisfying* (3.3).

Proof. Let $(E, e, H, h) \in D(\mathcal{L})$. Using the same manner which was used to prove Lemmas 3.4 and 3.6 in [12] for E, H and the Poincaré inequality for e, h, we have

$$||(E, e, H, h)||_{\mathcal{L}}^2 \le c \left(||\nabla \times E||^2 + ||\nabla\cdot \sigma E||^2 + |e|_1^2 + ||\nabla \times H||^2 + ||\nabla\cdot \mu H||^2 + |h|_1^2 \right).$$

Since $\mathbf{n} \times E = \mathbf{0}$ and $h = 0$ on the boundary, the conditions in (2.4) provide

$$\begin{aligned} \mu_1^{-1}||\nabla \times E||^2 \ &\le \ \left\langle \mu^{-1}\nabla \times E, \nabla \times E - \nabla e + \mu H \right\rangle - \langle \nabla \times E, H \rangle, \\ {\sigma_1}^{-1}||\nabla \times H||^2 \ &\le \ \left\langle \sigma^{-1}\nabla \times H, -\sigma E + \nabla \times H - \nabla h \right\rangle + \langle \nabla \times H, E \rangle. \end{aligned}$$

The above two inequalities, together with Hölder's inequality and the ϵ-inequality, give

$$||\nabla \times E||^2 + ||\nabla \times H||^2 \le c \left(||\nabla \times E - \nabla e + \mu H||^2 + || - \sigma E + \nabla \times H - \nabla h||^2 \right).$$

Consider the following several different cases for a_1 and a_2:

(*a*) If $a_1 \neq 0$ and $a_2 \neq 0$, then, by Green's formula,

$$||\nabla \cdot \sigma E||^2 = \langle \nabla \cdot \sigma E, \nabla \cdot \sigma E + a_2 h \rangle + a_2 \langle \sigma E, \nabla h \rangle , \tag{3.8}$$

$$||\nabla \cdot \mu H||^2 = \langle \nabla \cdot \mu H, \nabla \cdot \mu H - a_1 e \rangle - a_1 \langle \mu H, \nabla e \rangle , \tag{3.9}$$

$$||\nabla e||^2 = \langle \nabla e, -\nabla \times E + \nabla e - \mu H \rangle + \langle \nabla e, \mu H \rangle , \tag{3.10}$$

$$||\nabla h||^2 = \langle \nabla h, \sigma E - \nabla \times H + \nabla h \rangle - \langle \nabla h, \sigma E \rangle . \tag{3.11}$$

Multiply (3.10) by a_1 and (3.11) by a_2 and add to (3.9) and (3.8), respectively. Again use Hölder's inequality and the ϵ-inequality to obtain

$$||\nabla \cdot \sigma E||^2 + ||\nabla \cdot \mu H||^2 + ||\nabla e||^2 + ||\nabla h||^2 \leq c \, ||\mathcal{L}(E, e, H, h)||^2.$$

(*b*) If $a_1 = a_2 = 0$, taking Hölder's and Poincaré inequalities in (3.10) and (3.11) implies $||\nabla e||^2 + ||\nabla h||^2 \leq c \, ||\mathcal{L}(E, e, H, h)||^2$.

(*c*) If only one of a_1 and a_2 is 0, for example $a_1 = 0$ and $a_2 \neq 0$, then we use the same calculation in case (*a*) for ∇e. Multiply (3.11) by a_2 and add it to (3.8) to get $||\nabla \cdot \sigma E||^2 + ||\nabla h||^2 \leq c \, (||\nabla \cdot \sigma E + a_2 h||^2 + || - \sigma E + \nabla \times H - \nabla h||^2)$.

Thus, $||(E, e, H, h)||_{\mathcal{L}}^2 \leq c \, ||\mathcal{L}(E, e, H, h)||^2$, so that $\mathcal{L}$ is coercive. It is easy to prove the continuity of $\mathcal{L}$ by using the triangle inequality. Therefore, by the Lax–Milgram theorem, there exists the solution of (3.3). ∎

Now, we consider the dual weak problem (3.7). In a similar manner, we can show the existence and uniqueness of the solution for (3.7).

LEMMA 3.2. *There exists a unique solution,* $\mathbf{U}^* \in D(\mathcal{L}^*)$*, satisfying* (3.7).

COROLLARY 3.3. *The operator* $\mathcal{L} : D(\mathcal{L}) \to L^2(\Omega)^8$*, defined in* (3.1)*, is bijective.*

Proof. In Lemmas 3.1 and 3.2, it is proved that $\mathcal{L}$ and $\mathcal{L}^*$, defined in (3.1) and (3.5), respectively, are coercive. Therefore, $\mathcal{L}$ and $\mathcal{L}^*$ are injective. The coercivity and continuity of $\mathcal{L}$ provide that $\mathcal{L}$ is a closed operator. Then, by the closed range theorem (cf. [25]), the injectivity of $\mathcal{L}^*$ induces the surjectivity of $\mathcal{L}$. Thus, $\mathcal{L}$ is bijective. ∎

COROLLARY 3.4. *The operator* $\mathcal{L}^* \colon D(\mathcal{L}^*) \to L^2(\Omega)^8$*, defined in* (3.5)*, is bijective.*

Proof. Since $\mathcal{L}$ is a closed operator, by Lemma 2.1 in [5] and Corollary 3.3, $\mathcal{L}^*$ is bijective. ∎

Remark 3.5. Corollary 3.4 implies that, for given $\mathbf{F} \in L^2(\Omega)^8$, there exists a unique solution $\mathbf{U}^* \in D(\mathcal{L}^*)$ satisfying the weak form (3.7).

We consider several cases that incur difficulties in approximately solving the eddy current problem with H^1-conforming finite elements. Suppose that there are no boundary singularities but μ and σ are not smooth. Because the coefficients are not smooth, $D(\mathcal{L})$ is not imbedded into $H^1(\Omega)^8$. In fact, $H^1(\Omega)^8$ is a closed, proper subspace of $D(\mathcal{L})$. Therefore, H^1-conforming finite element spaces cannot be used to approximate the solution of system (3.1). The FOSLL* method may be used to overcome this difficulty. The efficiency of FOSLL* in this context can be seen by observing the dual operator $\mathcal{L}^*$ in (3.5). All of the discontinuous coefficients inside the derivatives in the $\mathcal{L}$ system are outside the differential operators in the $\mathcal{L}^*$ system. Accordingly, we have $D(\mathcal{L}^*)$ imbedded into $H^1(\Omega)^8$. Now, we suppose that μ and σ are not smooth and there is a boundary singularity. Although we can resolve the difficulty with the discontinuous coefficients by applying the standard FOSLL* method, the boundary singularity still leads to

$$H_0(\nabla\times) \cap H(\nabla\cdot) \not\subset H^1(\Omega)^3 \quad \text{and} \quad H(\nabla\times) \cap H_0(\nabla\cdot) \not\subset H^1(\Omega)^3.$$

In [18], a modification of the FOSLL* method was developed that overcomes this difficulty for the general scalar elliptic PDEs in the plane. In this paper, we introduce a different type of modification of FOSLL* to mitigate the difficulties with boundary singularities in three space dimensions.

4. The modified FOSLL* method. In this section, we present a modified FOSLL* functional in which the second and fourth equations in (3.5) involve weighted norms, that is, the functional is given by $\|\mathcal{L}^*\mathbf{U}^* - \mathbf{U}\|_\alpha^2$. Note that we have used the partially weighted norm that was introduced in (2.2). In subsection 4.2, we show how this modified FOSLL* functional works in the presence of singularities. Before getting into the details about the modified FOSLL* functional, we first show several Poincaré-type inequalities which are useful in many places. The first lemma appears in [15].

LEMMA 4.1. *Let* $\Omega = \Omega_1 \times (a,b)$ *with* $\Omega_1 = \{(r,\theta)|0 < r < R < 1, 0 < \theta < \omega, 0 < \omega \le 2\pi\}$. *If* $q \in H^1_{\beta+1}(\Omega)$ *vanishes on* $\partial\Omega$, *then, for any* β,

$$||q||_{0,\beta} \;\le\; c\,||\nabla q||_{0,\beta+1}. \tag{4.1}$$

Using the above lemma, we show the following.

LEMMA 4.2. *Assume that* Ω *is bounded, Lipschitz continuous, and simply connected. Let* $\phi \in H_0(\nabla\times) \cap H(\nabla\cdot)$; *then there exists a constant* c *such that, for any* $0 \le \alpha \le 1$,

$$||\phi|| \;\le\; c\;\left(||\nabla \times \phi|| + ||r^\alpha \nabla \cdot \phi||\right).$$

Proof. Let $\phi \in H_0(\nabla\times) \cap H(\nabla\cdot)$. By Lemma 3.4 in [12], ϕ can be written as

$$\phi = \varphi + \nabla\xi, \tag{4.2}$$

where $\varphi \in H_0(\nabla\times) \cap H(\nabla\cdot)$, $\nabla\cdot\varphi = 0$, and $\xi \in H^1_0(\Omega)$ satisfies $\Delta\xi = \nabla\cdot\phi$. Using the Cauchy–Schwarz inequality, Lemma 4.1, and the assumption on α yields

$$\begin{aligned} ||\nabla\xi||^2 = \langle\nabla\xi, \nabla\xi\rangle \;&=\; \langle -\nabla\cdot\phi, \xi\rangle \;\le\; ||r^\alpha\nabla\cdot\phi||\;||r^{-\alpha}\xi|| \\ &\le c||r^\alpha\nabla\cdot\phi||\;||r^{1-\alpha}\nabla\xi|| \;\le\; c||r^\alpha\nabla\cdot\phi||\;||\nabla\xi||. \end{aligned} \tag{4.3}$$

Now, (4.2), (4.3), and Lemma 3.4 in [12] imply

$$||\phi|| \le c(||\varphi|| + ||\nabla\xi||) \le c(||\nabla\times\varphi|| + ||r^\alpha\nabla\cdot\phi||) = c(||\nabla\times\phi|| + ||r^\alpha\nabla\cdot\phi||). \qquad \square$$

Lemmas 4.3 and 4.4 basically claim the same inequality in Lemma 4.1 without the zero boundary condition.

LEMMA 4.3. *Assume* Ω *is the same as in Lemma* 4.1 *and* $\beta > -1$. *For* $p \in H^1_{\beta+1}(\Omega)$, *there exists a constant* c *such that*

$$||p||_{0,\beta} \;\le\; c\;(||p||_{0,\beta+1} + ||\nabla p||_{0,\beta+1}).$$

Proof. Let $R_0 = \frac{R}{4}$, and let χ be a smooth function defined in Ω such that $\chi(r) = 1$ when $r < R_0$ and $\chi(r) = 0$ when $r > 2R_0$ and $|\chi'| \le cR_0^{-1}$ for some constant c. Since $1 = \chi + 1 - \chi$,

$$\int_0^R r^{2\beta}|p|^2 r dr = \int_0^R r^{2\beta}\left|\chi p + (1-\chi)p\right|^2 r dr \le 2\int_0^R r^{2\beta}\left(|\chi p|^2 + |(1-\chi)p|^2\right) r dr.$$

By the modified Hardy's inequality in [16], for $\beta > -1$,

$$\int_0^R r^{2\beta}|\chi p|^2 r\ dr \le c\int_0^R r^{2\beta+2}\left|\frac{\partial(\chi p)}{\partial r}\right|^2 r\ dr \le c\int_0^{2R_0} r^{2\beta+2}\left(\frac{1}{R_0^2}|p|^2+\left|\frac{\partial p}{\partial r}\right|^2\right) r\ dr.$$

Since $(1-\chi)p$ has nonzero values only on (R_0, R),

$$\int_0^R r^{2\beta}|(1-\chi)p|^2 r\ dr = \int_{R_0}^R r^{2\beta}|(1-\chi)p|^2 r\ dr = \int_{R_0}^R r^{-2}\ r^{2\beta+2}|(1-\chi)p|^2 r\ dr$$

$$\le {R_0}^{-2}\ \int_{R_0}^R r^{2\beta+2}|(1-\chi)p|^2 r\ dr \le {R_0}^{-2}\ \int_0^R r^{2\beta+2}|p|^2 r\ dr.$$

Hence

$$\int_\Omega r^{2\beta}|p|^2 d\Omega \le c\ R^{-2}\int_\Omega r^{2\beta+2}|p|^2 d\Omega + c\int_\Omega r^{2\beta+2}|\nabla p|^2 d\Omega. \qquad \square$$

To handle $||p||_{0,\beta+1}$ in Lemma 4.3, we prove the following lemma.

LEMMA 4.4. *Let $p \in H^1(\Omega)$ satisfying $||\nabla p||_{\beta+1-\epsilon} < \infty$; then there exist constants b and c such that, for any $\beta > -1$ and $\epsilon > 0$,*

$$||p-b||_{0,\beta}\ \le\ c\ ||\nabla p||_{0,\beta+1-\epsilon}.$$

Proof. Here, we show an outline of the proof. The details can be found in [17]. Let $p \in H^1(\Omega)$ satisfy the assumption and consider the following expression for P:

$$\begin{aligned} &p(r,\theta,z) - p(r_0,\theta_0,z_0)\\ &= p(r,\theta,z) - p(r,\theta_0,z) + p(r,\theta_0,z) - p(r_0,\theta_0,z) + p(r_0,\theta_0,z) - p(r_0,\theta_0,z_0)\\ &= \int_{\theta_0}^{\theta}\frac{\partial p}{\partial\tilde\theta}(r,\tilde\theta,z)\ d\tilde\theta + \int_{r_0}^{r}\frac{\partial p}{\partial\tilde r}(\tilde r,\theta_0,z)\ d\tilde r + \int_{z_0}^{z}\frac{\partial p}{\partial\tilde z}(r_0,\theta_0,\tilde z)\ d\tilde z. \end{aligned}$$

Multiply by $r_0^{\beta+\frac12}$ and perform the integration $\int_\Omega r_0 dr_0 d\theta_0 dz_0$ on both sides:

$$\begin{aligned} c_1 p(r,\theta,z) = &\int_\Omega r_0^{\beta+\frac12} p(r_0,\theta_0,z_0) r_0 dr_0 d\theta_0 dz_0 + \int_\Omega r_0^{\beta+\frac12}\left\{\int_{\theta_0}^{\theta}\frac{\partial p}{\partial\tilde\theta}(r,\tilde\theta,z)\ d\tilde\theta\right.\\ &\left. + \int_{r_0}^{r}\frac{\partial p}{\partial\tilde r}(\tilde r,\theta_0,z)\ d\tilde r + \int_{z_0}^{z}\frac{\partial p}{\partial\tilde z}(r_0,\theta_0,\tilde z)\ d\tilde z\right\} r_0 dr_0 d\theta_0 dz_0, \end{aligned} \tag{4.4}$$

where $c_1 = \int_\Omega r_0^{\beta+\frac12} r_0 dr_0 d\theta_0 dz_0$. Let

$$b = \frac{1}{c_1}\int_\Omega r_0^{\beta+\frac12} p(r_0,\theta_0,z_0)\ r_0 dr_0 d\theta_0 dz_0;$$

then $|b| \le c||p|| < \infty$. Subtracting b from both sides in (4.4), changing the order of integration, inserting $\tilde r^{\frac{-1+\epsilon}{2}}\cdot\tilde r^{\frac{1-\epsilon}{2}} = 1$ in order to group $\tilde r^{\frac{1-\epsilon}{2}}$ with the $\frac{\partial p}{\partial\tilde r}$ term, using the Cauchy–Schwarz inequality, and squaring both sides yield

$$\begin{aligned} |p(r,\theta,z)-b|^2 \le c\Bigg\{&\int_0^\omega\left|\frac{\partial p}{\partial\tilde\theta}(r,\tilde\theta,z)\right|^2 d\tilde\theta + \int_0^\omega\int_0^R \tilde r^{2\beta+3}\left|\frac{\partial p}{\partial\tilde r}(\tilde r,\theta_0,z)\right|^2 d\tilde r d\theta_0\\ &+\int_0^\omega\left\{\frac{R^\epsilon}{\epsilon}\int_r^R \tilde r^{1-\epsilon}\left|\frac{\partial p}{\partial\tilde r}(\tilde r,\theta_0,z)\right|^2 d\tilde r + \int_0^R r_0^{2\beta+3}\int_a^b\left|\frac{\partial p}{\partial\tilde z}(r_0,\theta_0,\tilde z)\right|^2 d\tilde z dr_0\right\} d\theta_0\Bigg\}. \end{aligned}$$

To establish the weighted L^2-norm of $|p-b|$, multiply by $r^{2\beta}$ and take an integration over Ω. Then, we have

$$\int_\Omega r^{2\beta}|p(r,\theta,z)-b|^2 d\Omega \le c\int_\Omega r^{2\beta+2}\left(\left|\frac{1}{r}\frac{\partial p}{\partial\theta}\right|^2+\left|\frac{\partial p}{\partial r}\right|^2+\left|\frac{\partial p}{\partial z}\right|^2\right)+r^{2\beta+2-\epsilon}\left|\frac{\partial p}{\partial r}\right|^2 d\Omega$$

$$\le c\int_\Omega r^{2\beta+2-\epsilon}|\nabla p\,|^2 d\Omega,$$

where $c=c(\Omega,\beta,\epsilon,(\beta+1)^{-1},\epsilon^{-1})\to\infty$ as $\epsilon\to 0$ and $\beta\to -1$. ∎

LEMMA 4.5. *Assume that Ω is bounded, Lipschitz continuous, and simply connected. Let $\psi\in H(\nabla\times)\cap H_0(\nabla\cdot)$; then there exists a constant c such that, for any $0\le\alpha<1$,*

$$||\psi|| \;\le\; c\,(||\nabla\times\psi||+||r^\alpha\nabla\cdot\psi||).$$

Proof. The proof follows similarly to Lemma 4.2 using Lemmas 4.3 and 4.4. ∎

If a vector function is in H^1 and satisfies certain boundary conditions, then the sum of norms of div and curl is equal to the semi-H^1-norm.

LEMMA 4.6. *Let Ω be a bounded polyhedral domain in $\mathbb{R}^3$. If $\mathbf{v}\in H^1(\Omega)^3$ and satisfies $\mathbf{n}\cdot\mathbf{v}=0$ or $\mathbf{n}\times\mathbf{v}=\mathbf{0}$ on the boundary $\partial\Omega$, then*

$$||\nabla\cdot\mathbf{v}||^2+||\nabla\times\mathbf{v}||^2=||\nabla\mathbf{v}||^2.$$

Proof. See [7] and [8]. ∎

The basic idea of the modification here is to use a weighted norm in certain terms of the least squares functional in (3.6). Using a weighted norm allows the existence of a sequence, $\{\mathbf{U}_n\}\subset D(\mathcal{L}^*)\cap H^1(\Omega)^8$, converging to the nonsmooth solution, $\mathbf{U}^*\in D(\mathcal{L}^*)$, in the functional norm. Consider the operator $\mathcal{L}^*$ blockwise. Let $D_A=H_0(\nabla\times)\cap H(\nabla\cdot)$ and $D_B=H(\nabla\times)\cap H_0(\nabla\cdot)$ and define

$$A=\begin{bmatrix}\nabla\times & -\mu\nabla\\ \nabla\cdot & 0\end{bmatrix}\quad\text{and}\quad B=\begin{bmatrix}\nabla\times & -\sigma\nabla\\ \nabla\cdot & 0\end{bmatrix}. \tag{4.5}$$

We first show that there exist sequences $\{\mathcal{X}_n\}$ and $\{\mathcal{Y}_n\}$ in $H^1(\Omega)^4$ such that

$$||A\mathcal{X}_n-F||_\alpha\longrightarrow 0\;\text{ and }\;||B\mathcal{Y}_n-G||_\alpha\longrightarrow 0\;\text{ as }\;n\longrightarrow\infty, \tag{4.6}$$

for given $F,G\in L^2(\Omega)^4$ and the norm $||\cdot||_\alpha$ defined in (2.1). We again emphasize that this notation implies that only the scalar term, the term involving $\nabla\cdot$, is weighted. Then, we discuss the density of H^1-functions in D_A and D_B under weighted norms.

4.1. The density arguments in D_A and D_B. As a first step to show the existence of H^1-sequences satisfying (4.6), we apply the well-known L^2-decomposition and show several lemmas. The next lemma provides the decomposition of $L^2(\Omega)^3$.

LEMMA 4.7. *Every function $\mathbf{w}\in L^2(\Omega)^3$ has the orthogonal decomposition*

$$\mathbf{w}=\nabla\times\mathbf{u}+\nabla\psi,$$

where $\psi\in H^1(\Omega)/\mathbb{R}$ is the only solution of $\langle\nabla\psi,\nabla\xi\rangle=\langle\mathbf{w},\nabla\xi\rangle$, for any $\xi\in H^1(\Omega)$, and $\mathbf{u}\in H^1(\Omega)^3$ satisfies $\nabla\cdot\mathbf{u}=0$.

Proof. See [12] for details. ∎

LEMMA 4.8. *For given $F \in L^2(\Omega)^4$, there exists a unique solution, $\mathcal{X} \in D_A \times H^1(\Omega)/\mathbb{R}$, of $A\mathcal{X} = F$.*

Proof. The result follows from a proof similar to the proofs of section 3. ∎

Analogously, we show the following lemma.

LEMMA 4.9. *For given $G \in L^2(\Omega)^4$, there exists a unique solution, $\mathcal{Y} \in D_B \times H_0^1(\Omega)$, of $B\mathcal{Y} = G$.*

Now, we provide some decompositions in D_A and D_B.

THEOREM 4.10. *Given $\tilde{\mathbf{u}} \in D_A$, there exists $\mathbf{u} \in H^1(\Omega)^3 \cap D_A$ and $\phi \in H_0^1(\Omega)$ such that*

$$\tilde{\mathbf{u}} = \mathbf{u} + \nabla\phi.$$

Proof. Use Lemma 4.7 to write $\nabla \times \tilde{\mathbf{u}} = \nabla \times \mathbf{u}_0 + \nabla\psi$ with $\mathbf{u}_0 \in H^1(\Omega)^3$ and $\psi \in H^1(\Omega)/\mathbb{R}$. Taking the divergence of the above equation leads to the conclusion that $\psi = 0$. Thus, $\nabla \times (\tilde{\mathbf{u}} - \mathbf{u}_0) = \mathbf{0}$, which implies that $\tilde{\mathbf{u}} = \mathbf{u}_0 + \nabla\phi_0$, for some $\phi_0 \in H^1(\Omega)/\mathbb{R}$. Now, $\mathbf{0} = \mathbf{n} \times \tilde{\mathbf{u}} = \mathbf{n} \times \mathbf{u}_0 + \mathbf{n} \times \nabla\phi_0$. Since $\mathbf{u}_0 \in H^1(\Omega)^3$, we have $\mathbf{n} \times \mathbf{u}_0 \in H^{\frac{1}{2}}(\partial\Omega)^3$. Thus, $\mathbf{n} \times \nabla\phi_0 = -\mathbf{n} \times \mathbf{u}_0$ on $\partial\Omega$, which implies $trace(\phi_0) \in H^{\frac{3}{2}}(\partial\Omega)$. Let $\phi_2 \in H^2(\Omega)$ satisfy $trace(\phi_0) = trace(\phi_2)$. Then, let

$$\mathbf{u} = \mathbf{u}_0 + \nabla\phi_2, \quad \phi = \phi_0 - \phi_2.$$

Since $\mathbf{n} \times \nabla\phi = \mathbf{0}$, the theorem is proved. ∎

THEOREM 4.11. *Given $\tilde{\mathbf{v}} \in D_B$, there exists $\mathbf{v} \in H^1(\Omega)^3 \cap D_B$ and $\psi \in H^1(\Omega)$ with $\mathbf{n} \cdot \nabla\psi = 0$ on $\partial\Omega$ such that*

$$\tilde{\mathbf{v}} = \mathbf{v} + \nabla\psi.$$

Proof. The proof is similar to the proof of Theorem 4.10. Here, we construct ψ satisfying $\mathbf{n} \cdot \nabla\psi = 0$ on the boundary. ∎

In the domain with a reentrant edge, the solution of the Poisson equation

$$-\Delta\phi = f$$

for $f \in L^2(\Omega)$, with a Dirichlet or Neumann boundary condition is, in general, not in $H^2(\Omega)$. It is in $H^2_{loc}(\Omega)$; that is, $\phi \in H^2(S)$ for any open subset S of Ω such that its closure $\overline{S}$ does not meet the reentrant edge (cf. [13]). The solution, ϕ, is also in $H^{1+\gamma}(\Omega)$ for some $\gamma \in (0, 1)$. A more precise measure is given by the weighted Sobolev space. This solution ϕ is in $H^2_\beta(\Omega)$ with β related to the angle of the reentrant edge (cf. [15], [21]). In the following theorems, we establish H^1-sequences satisfying (4.6).

From this point forward, if not mentioned explicitly, Ω is the prototype domain which was defined in section 2.

THEOREM 4.12. *For given $F \in L^2(\Omega)^4$ and an operator A defined in* (4.5), *there exists a sequence $\{\mathcal{X}_n\} \subset H^1(\Omega)^4 \cap \big(D_A \times H^1(\Omega)/\mathbb{R}\big)$ such that*

$$\|A\mathcal{X}_n - F\|_\alpha \longrightarrow 0 \quad \text{as} \quad n \to \infty,$$

where $\alpha > 1 - \lambda$, $\lambda = \pi/\omega$, and ω is the angle of the reentrant edge.

Proof. Let $F = (\mathbf{f}_1, f_2) \in L^2(\Omega)^4$. From Lemma 4.8, we have $\tilde{\mathbf{u}} \in D_A$ and $\tilde{p} \in H^1(\Omega)/\mathbb{R}$ satisfying $\nabla \times \tilde{\mathbf{u}} - \mu\nabla\tilde{p} = \mathbf{f}_1$ and $\nabla \cdot \tilde{\mathbf{u}} = f_2$. By Theorem 4.10, $\tilde{\mathbf{u}}$ is decomposed of $\tilde{\mathbf{u}} = \mathbf{u} + \nabla\phi$, where $\mathbf{u} \in H^1(\Omega)^3 \cap D_A$ and $\phi \in H_0^1(\Omega)$. Here, ϕ satisfies

$$\left\{ \begin{array}{rcll} \nabla \cdot \nabla\phi & = & -\nabla \cdot \mathbf{u} + f_2 & \text{in } \Omega, \\ \phi & = & 0 & \text{on } \partial\Omega. \end{array} \right. \tag{4.7}$$

Given $\alpha > 1-\lambda$, choose β such that $\beta < \alpha$ and $|\beta - 1| < \lambda$. It is known that the solution ϕ is in $H^1(\Omega) \cap H^2_\beta(\Omega)$. Define $\Omega_n = (\{(x,y)|\ 1/(2n) \le r \le 1/n\} \times \mathbb{R}) \cap \Omega$ with $r = \sqrt{x^2+y^2}$ and $\delta_n(r)$ a smooth function satisfying

$$\delta_n(r) = \begin{cases} 0 & \text{if } r < 1/(2n), \\ 1 & \text{if } r > 1/n, \end{cases} \tag{4.8}$$

where $|\ \delta_n'\ | \le c_1 n$ and $|\ \delta_n''\ | \le c_2 n^2$, for some positive constants c_1 and c_2. Define $\phi_n = \delta_n \phi$; then $\phi_n \in H^2(\Omega)$ (cf. [13]). Therefore, $\mathbf{u}_n := \mathbf{u} + \nabla\phi_n$ is in $H^1(\Omega)^3 \cap D_A$ and satisfies

$$\nabla \times \mathbf{u}_n - \mu\nabla\tilde{p} = \nabla \times \mathbf{u} - \mu\nabla\tilde{p} = \nabla \times \tilde{\mathbf{u}} - \mu\nabla\tilde{p} = \mathbf{f}_1.$$

Using the triangle inequality several times and the properties of δ_n yields

$$\|\nabla\cdot\mathbf{u}_n - f_2\|^2_{0,\alpha} = \|\nabla\cdot(\mathbf{u}+\nabla\phi_n) - \nabla\cdot(\mathbf{u}+\nabla\phi)\|^2_{0,\alpha} = \int_\Omega r^{2\alpha}|\Delta((\delta_n(r)-1)\phi)|^2\ d\Omega$$

$$= \int\int\left(\int_0^{\frac{1}{2n}} r^{2\alpha}|\Delta\phi|^2\ r dr + \int_{\frac{1}{2n}}^{\frac{1}{n}} r^{2\alpha}|\Delta((\delta_n(r)-1)\phi)|^2\ r dr\right) d\theta dz$$

$$\le\ c\left(\frac{1}{2n}\right)^{2(\alpha-\beta)} ||\Delta\phi||^2_{0,\beta} + c\int_{\Omega_n} r^{2\alpha}\left(|\Delta\phi|^2 + n^4|\phi|^2 + n^2\left(|\partial_x\phi|^2 + |\partial_y\phi|^2\right)\right) d\Omega$$

$$\le\ c\ n^{-2(\alpha-\beta)}|\phi|^2_{2,\beta} + c\ n^{-2(\alpha-\beta)}||\phi||^2_{2,\beta} = c\ n^{-2(\alpha-\beta)}||\phi||^2_{2,\beta}.$$

The right-hand side goes to 0 as n goes to infinity. By letting $\mathcal{X}_n := (\mathbf{u}_n, \tilde{p})$, the proof is completed. ∎

We can show the next theorem in the same manner.

THEOREM 4.13. *For given $G \in L^2(\Omega)^4$ and an operator B defined in* (4.5), *there exists a sequence $\{\mathcal{Y}_n\} \subset H^1(\Omega)^4 \cap \left(D_B \times H^1_0(\Omega)\right)$ such that*

$$\|B\mathcal{Y}_n - G\|_\alpha \longrightarrow 0 \quad \text{as} \quad n \to \infty,$$

where $\alpha > 1-\lambda$, $\lambda = \pi/\omega$, and ω is the angle of the reentrant edge.

Now, we state some density results. Define

$$D_{A_\alpha} := \{\mathbf{u} \in L^2(\Omega)^3 \mid ||\nabla\times\mathbf{u}|| + ||\nabla\cdot\mathbf{u}||_{0,\alpha} < \infty,\ \mathbf{n}\times\mathbf{u} = \mathbf{0} \text{ on } \partial\Omega\}, \tag{4.9}$$

$$D_{B_\alpha} := \{\mathbf{u} \in L^2(\Omega)^3 \mid ||\nabla\times\mathbf{u}|| + ||\nabla\cdot\mathbf{u}||_{0,\alpha} < \infty,\ \mathbf{n}\cdot\mathbf{u} = 0 \text{ on } \partial\Omega\}, \tag{4.10}$$

which are Hilbert spaces under the norm $||\mathbf{u}||_{D_{A_\alpha}} = ||\mathbf{u}||_{D_{B_\alpha}} := (||\mathbf{u}||^2 + ||\nabla\times\mathbf{u}||^2 + ||\nabla\cdot\mathbf{u}||^2_{0,\alpha})^{\frac{1}{2}}$. The density statement for D_{A_α} can be found in [8], [10], and [11] for $\alpha \in (1-\lambda, 1)$. Here, we extend the density results to $\alpha > 1-\lambda$.

THEOREM 4.14. *$D_{A_\alpha} \cap H^1(\Omega)^3$ is dense in D_{A_α} when $\alpha > 1-\lambda$, and $D_{B_\alpha} \cap H^1(\Omega)^3$ is dense in D_{B_α} when $\alpha > 1-\lambda$.*

Proof. We separate the proof into two cases. First, we consider $1-\lambda < \alpha < 1$. Let the operator A be defined as in (4.5) and let $(\mathbf{u}, p) \in D_{A_\alpha} \times H^1(\Omega)/\mathbb{R}$; then,

$$\|A(\mathbf{u},p)\|^2_\alpha = \|\nabla\times\mathbf{u} - \mu\nabla p\|^2 + ||\nabla\cdot\mathbf{u}||^2_{0,\alpha} \ge \mu_0||(1/\sqrt{\mu})\nabla\times\mathbf{u} - \sqrt{\mu}\nabla p||^2 + ||\nabla\cdot\mathbf{u}||^2_{0,\alpha}$$

$$\ge c(||\nabla\times\mathbf{u}||^2 + ||\nabla p||^2 + ||\nabla\cdot\mathbf{u}||^2_{0,\alpha}) \ge c(||\mathbf{u}||^2_{D_{A_\alpha}} + ||p||^2_1).$$

In the above, Lemma 4.2 and Theorem 4.12 imply the density in D_{A_α} for $1-\lambda < \alpha < 1$.

Now, we consider the case $\alpha \geq 1$. Let $\mathbf{u} \in D_{A_\alpha}$; then, similarly to Theorem 4.10, we can show that $\mathbf{u}$ is decomposed in the form of $\mathbf{u} = \mathbf{u}_0 + \nabla\phi$, where $\mathbf{u}_0 \in H^1(\Omega)^3 \cap D_{A_\alpha}$ and $\phi \in H_0^1(\Omega)$. Let Ω_n and the smooth cut-off function $\delta_n(r)$ be defined as in the proof of Theorem 4.12, and define $\Omega_{\tilde{n}} = (\{(x,y)|r \leq 1/n\} \times \mathbb{R}) \cap \Omega$. Define $\mathbf{u}_n = \mathbf{u}_0 + \nabla(\delta_n(r)\phi)$; then $\mathbf{u}_n$ is in $H^1(\Omega)^3 \cap D_{A_\alpha}$. Since $\phi \in H_0^1(\Omega)$ and $\|\nabla\cdot\nabla\phi\|_{0,\alpha} < \infty$, it is easy to see that

$$\|\Delta\phi\|_{0,\alpha,\Omega_{\tilde{n}}} \to 0 \quad \text{and} \quad \|\phi\|_{1,\Omega_{\tilde{n}}} \to 0 \tag{4.11}$$

as $n \to \infty$, where the subscript $\Omega_{\tilde{n}}$ means the integration over $\Omega_{\tilde{n}}$. Therefore, the triangle inequality and the property of $\delta_n(r)$ yield

$$\begin{aligned}
&\|\mathbf{u}-\mathbf{u}_n\|^2_{D_{A_\alpha}} = \|\mathbf{u}-\mathbf{u}_n\|^2 + \|\nabla\times(\mathbf{u}-\mathbf{u}_n)\|^2 + \|\nabla\cdot(\mathbf{u}-\mathbf{u}_n)\|^2_{0,\alpha}\\
&= \|\nabla((1-\delta_n(r))\phi)\|^2 + \|\nabla\cdot\nabla((1-\delta_n(r))\phi)\|^2_{0,\alpha}\\
&\leq c\left(\|\delta_n'\phi\|^2_{0,\Omega_n} + \|\nabla\phi\|^2_{0,\Omega_{\tilde{n}}} + \|\Delta\phi\|^2_{0,\alpha,\Omega_{\tilde{n}}} + \|\delta_n''\phi\|^2_{0,\alpha,\Omega_n} + \|r^{-1}\delta_n'\phi\|^2_{0,\alpha,\Omega_n} + \|\delta_n'\nabla\phi\|^2_{0,\alpha,\Omega_n}\right).
\end{aligned}$$

We have the second and third terms in the last line of the above go to 0 by (4.11) and we have $\|r^{-1}\delta_n'\phi\|_{0,\alpha,\Omega_n} \leq c\|\delta''\phi\|_{0,\alpha,\Omega_n}$ by the property of δ_n. Since $|\delta_n'(r)| \leq cn$, $1/(2n) \leq r \leq 1/n$ on Ω_n, and $\alpha \geq 1$, the sixth term in the above is

$$\|\delta_n'\nabla\phi\|^2_{0,\alpha,\Omega_n} \leq c\|r^\alpha n\nabla\phi\|^2_{0,\Omega_n} \leq c\|r^{\alpha-1}\nabla\phi\|^2_{0,\Omega_n} = c\|\nabla\phi\|^2_{0,\Omega_n} \to 0.$$

We now focus on the following two terms: By Lemma 4.3 and $\alpha \geq 1$, for $\epsilon > 0$,

$$\begin{aligned}
&\|\delta_n'\phi\|^2_{0,\Omega_n} + \|\delta_n''\phi\|^2_{0,\alpha,\Omega_n} \leq c\left(\|n\phi\|^2_{0,\Omega_n} + \|r^\alpha n^2\phi\|^2_{0,\Omega_n}\right) \leq c\left(\|r^{-1}\phi\|^2_{0,\Omega_n} + \|r^{\alpha-2}\phi\|^2_{0,\Omega_n}\right)\\
&\leq c(1/2n)^{-2\epsilon}\left(\|r^{-1+\epsilon}\phi\|^2_{0,\Omega_n} + \|r^{\alpha-2+\epsilon}\phi\|^2_{0,\Omega_n}\right) \leq cn^{2\epsilon}\left(\|r^{-1+\epsilon}\phi\|^2_{0,\Omega_{\tilde{n}}} + \|r^{\alpha-2+\epsilon}\phi\|^2_{0,\Omega_{\tilde{n}}}\right)\\
&\leq cn^{2\epsilon}\left(\|r^\epsilon\phi\|^2_{0,\Omega_{\tilde{n}}} + \|r^\epsilon\nabla\phi\|^2_{0,\Omega_{\tilde{n}}} + \|r^{\alpha-1+\epsilon}\phi\|^2_{0,\Omega_{\tilde{n}}} + \|r^{\alpha-1+\epsilon}\nabla\phi\|^2_{0,\Omega_{\tilde{n}}}\right)\\
&\leq cn^{2\epsilon}\left(n^{-2\epsilon}(\|\phi\|^2_{0,\Omega_{\tilde{n}}} + \|\nabla\phi\|^2_{0,\Omega_{\tilde{n}}}) + n^{-2(\alpha-1+\epsilon)}(\|\phi\|^2_{0,\Omega_{\tilde{n}}} + \|\nabla\phi\|^2_{0,\Omega_{\tilde{n}}})\right) \leq c\|\phi\|^2_{1,\Omega_{\tilde{n}}}.
\end{aligned}$$

Hence, we proved that $\|\mathbf{u}-\mathbf{u}_n\|^2_{D_{A_\alpha}} \to 0$ as long as $\alpha > 1-\lambda$. The density $D_{B_\alpha} \cap H^1(\Omega)^3$ in D_{B_α} follows the same process. ∎

4.2. The existence of H^1-sequences. So far, we have obtained H^1-sequences satisfying (4.6). For given $(\mathbf{E}, s, \mathbf{H}, k)$, we consider the minimization of the functional (3.6) in the partially weighted norm from (2.2),

$$\mathcal{F}^*_\alpha(\mathbf{U}^*; (\mathbf{E}, s, \mathbf{H}, k)) = ||\mathcal{L}^*\mathbf{U}^* - (\mathbf{E}, s, \mathbf{H}, k)||^2_\alpha \tag{4.12}$$

for all $(\mathcal{U}, p, \mathcal{V}, q) \in D(\mathcal{L}^*)$, where the weighted norms involve only the equations corresponding to slack variables s and k. Since s and k are slack variables of the original system, we may assume that $s = 0$ and $k = 0$. Then the corresponding weak form is as follows: Find $\mathbf{U}^* \in D(\mathcal{L}^*)$ satisfying

$$\langle\mathcal{L}^*\mathbf{U}^*, \mathcal{L}^*\mathbf{V}^*\rangle_\alpha = \langle(\mathbf{E}, 0, \mathbf{H}, 0), \mathcal{L}^*\mathbf{V}^*\rangle_\alpha = \langle\mathcal{L}(\mathbf{E}, 0, \mathbf{H}, 0), \mathbf{V}^*\rangle = \langle\mathbf{F}, \mathbf{V}^*\rangle$$

for all $\mathbf{V}^* \in D(\mathcal{L}^*)$, where $\langle\cdot,\cdot\rangle_\alpha = \langle J_\alpha\cdot, J_\alpha\cdot\rangle$ with J_α the diagonal matrix $J_\alpha = \text{diag}[\ 1,\ 1,\ 1,\ r^\alpha,\ 1,\ 1,\ 1,\ r^\alpha\ ]$. As an important step in achieving the goal of this paper, we show that there exists an H^1-sequence, $\{\mathbf{U}_n\}$, satisfying the following.

THEOREM 4.15. *Assume $\alpha > 1-\lambda$. For given $\mathbf{U} = (\mathbf{E}, s, \mathbf{H}, k) \in L^2(\Omega)^8$, there exists a sequence $\mathbf{U}_n \in D(\mathcal{L}^*) \cap H^1(\Omega)^8$ such that*

$$||\mathcal{L}^*\mathbf{U}_n - \mathbf{U}||_\alpha \longrightarrow 0 \tag{4.13}$$

as $n \longrightarrow \infty$.

Proof. By surjectivity of $\mathcal{L}^*$, there exists $\mathbf{U}^* = (\mathcal{U}, \tilde{p}, \mathcal{V}, \tilde{q}) \in D(\mathcal{L}^*)$ such that $\mathcal{L}^*\mathbf{U}^* = \mathbf{U}$. From Theorems 4.12 and 4.13, we have $\mathbf{U}_n \in D(\mathcal{L}^*) \cap H^1(\Omega)^8$ satisfying

(4.14)

$$\nabla \times \mathcal{V}_n - \sigma\nabla q = \mathbf{E} + \sigma\mathcal{U} = \nabla \times \mathcal{V} - \sigma\nabla\tilde{q}\ , \quad ||\nabla \cdot \mathcal{V}_n - a_1\tilde{p} - s||_{0,\alpha} \longrightarrow 0,$$

$$\nabla \times \mathcal{U}_n - \mu\nabla p = \mathbf{H} - \mu\mathcal{V} = \nabla \times \mathcal{U} - \mu\nabla\tilde{p}\ , \quad ||\nabla \cdot \mathcal{U}_n + a_2\tilde{q} - k||_{0,\alpha} \longrightarrow 0,$$

as n goes to infinity, where $\mathbf{U}_n = (\mathcal{U}_n, p, \mathcal{V}_n, q)$ and a_1, a_2 are nonnegative constants. First, consider the case $1-\lambda < \alpha < 1$. By substituting $\mathcal{L}^*\mathbf{U}^* = \mathbf{U}$ into (4.14) and using Lemmas 4.2 and 4.5, we have the first inequality in the following equation:

$$\begin{aligned} ||\mathcal{L}^*\mathbf{U}_n - \mathbf{U}||_\alpha^2 &\le c\,\big(\ ||\nabla \times (\mathcal{U}_n - \mathcal{U})||^2 + ||\nabla \cdot (\mathcal{U}_n - \mathcal{U})||_{0,\alpha}^2 \\ &\qquad + ||\nabla \times (\mathcal{V}_n - \mathcal{V})||^2 + ||\nabla \cdot (\mathcal{V}_n - \mathcal{V})||_{0,\alpha}^2\ \big) \\ &\le c\ (||\nabla \cdot (\mathcal{U}_n - \mathcal{U})||_{0,\alpha}^2 + ||\nabla \cdot (\mathcal{V}_n - \mathcal{V})||_{0,\alpha}^2), \end{aligned}$$

where $c = c(\Omega, \mu, \sigma, \alpha)$. Boundary conditions and orthogonality properties provide the second inequality in the above. By (4.14), the right-hand side converges to 0.

Now consider $\alpha \ge 1$. Since $|r| < 1$, it is easy to see that, when $\alpha_1 \ge \alpha_2$, $||\cdot||_{0,\alpha_1} \le ||\cdot||_{0,\alpha_2}$. Therefore, for $\alpha \ge 1$,

$$||\mathcal{L}^*\mathbf{U}_n - \mathbf{U}||_\alpha \ \le\ ||\mathcal{L}^*\mathbf{U}_n - \mathbf{U}||_{1-\epsilon}$$

for $\epsilon > 0$. Hence, the result holds. ∎

COROLLARY 4.16. *Let $\mathbf{U}^* = (\mathcal{U}, \tilde{p}, \mathcal{V}, \tilde{q}) \in D(\mathcal{L}^*)$ satisfying $\mathcal{L}^*\mathbf{U}^* = \mathbf{U}$ and let $\mathbf{U}_n = (\mathcal{U}_n, p, \mathcal{V}_n, q) \in D(\mathcal{L}^*) \cap H^1(\Omega)^8$ satisfying* (4.14)*, where $\mathcal{U}_n = \mathbf{u} + \nabla\delta_n\phi$ and $\mathcal{V}_n = \mathbf{v} + \nabla\delta_n\psi$ with δ_n defined as in* (4.8)*, $\mathbf{u} \in H^1(\Omega)^3 \cap D_A$, $\mathbf{v} \in H^1(\Omega)^3 \cap D_B$, $\phi \in H^1(\Omega)/\mathbb{R}$, and $\psi \in H_0^1(\Omega)$ from Theorems* 4.10 *and* 4.11*. Then*

$$\mathcal{U} = \mathbf{u} + \nabla\phi, \quad \mathcal{V} = \mathbf{v} + \nabla\psi, \quad \tilde{p} = p, \quad \textit{and} \quad \tilde{q} = q.$$

Proof. By taking divergence on the first and third equations in (4.14), we obtain $\tilde{p} = p$ and $\tilde{q} = q$. Then, we have

$$\mathbf{0} = \nabla \times (\mathcal{U} - \mathcal{U}_n) = \nabla \times (\mathcal{U} - (\mathbf{u} + \nabla\delta_n\phi)) = \nabla \times (\mathcal{U} - (\mathbf{u} + \nabla\phi)),$$

$$0 = \nabla \cdot \mathcal{U} + a_2 q - k = \nabla \cdot \mathcal{U} - (\nabla \cdot \mathbf{u} + \Delta\phi) = \nabla \cdot (\mathcal{U} - (\mathbf{u} + \nabla\phi)),$$

which imply $\mathcal{U} = \mathbf{u} + \nabla\phi$. Similarly, $\mathcal{V} = \mathbf{v} + \nabla\psi$. ∎

The singularity on the boundary implies that the solution, $\mathbf{U}^* \in D(\mathcal{L}^*)$, of $\mathcal{L}^*\mathbf{U}^* = \mathbf{U}$ is not in H^1. However, we have shown that there is an H^1-sequence, $\mathbf{U}_n$, satisfying (4.13). This allows us to use the standard H^1-conforming finite elements, as we demonstrate in section 7. In the next theorem we establish the coercivity and continuity of $\mathcal{F}^*$ in the partially weighted norm.

THEOREM 4.17. *If $(\mathcal{U}, p, \mathcal{V}, q) \in D(\mathcal{L}^*)$, then there exist c and C such that*

$$\begin{aligned} & c(||\mathcal{U}|| + ||\nabla \times \mathcal{U}|| + ||r^\alpha \nabla \cdot \mathcal{U}|| + ||p||_1 + ||\mathcal{V}|| + ||\nabla \times \mathcal{V}|| + ||r^\alpha \nabla \cdot \mathcal{V}|| + ||q||_1) \\ & \leq \ ||\mathcal{L}^*(\mathcal{U}, p, \mathcal{V}, q)||_\alpha \\ & \leq \ C(||\mathcal{U}|| + ||\nabla \times \mathcal{U}|| + ||r^\alpha \nabla \cdot \mathcal{U}|| + ||p||_1 + ||\mathcal{V}|| + ||\nabla \times \mathcal{V}|| + ||r^\alpha \nabla \cdot \mathcal{V}|| + ||q||_1), \end{aligned}$$

where $1 - \lambda < \alpha < 1$.

Proof. It is clear that $||r^\alpha \nabla \cdot \mathcal{U}|| + ||r^\alpha \nabla \cdot \mathcal{V}|| \leq ||\mathcal{L}^*(\mathcal{U}, p, \mathcal{V}, q)||_\alpha$. By Lemmas 4.2 and 4.5, and the Poincaré inequality, it is enough to show that

$$||\nabla \times \mathcal{U}|| + ||\nabla p|| + ||\nabla \times \mathcal{V}|| + ||\nabla q|| \leq c||\mathcal{L}^*(\mathcal{U}, p, \mathcal{V}, q)||_\alpha.$$

Using orthogonality and Hölder's inequality, we can easily show the lower inequality. The upper inequality follows by the triangle inequality. For more details, see [17]. ∎

5. Scaling in FOSLL*. In this section, we briefly introduce scaling in FOSLS and FOSLL*. From [18], it is known that using a scaling in FOSLS and FOSLL* sometimes has computational advantages. Here, we are particularly interested in scaling with $\sqrt{\mu}$ and $\sqrt{\sigma}$ since it gives orthogonality between $\nabla\times$ and ∇ in FOSLL*.

The eddy current equations (3.1) can be rewritten as

$$\mathcal{L}_s \mathbf{U} = \begin{bmatrix} -\sqrt{\sigma} I & 0 & \nabla \times \frac{1}{\sqrt{\mu}} & -\nabla \frac{1}{\sqrt{\mu}} \\ 0 & -\frac{1}{\sqrt{\sigma}} a_1 & \nabla \cdot \sqrt{\mu} & 0 \\ \nabla \times \frac{1}{\sqrt{\sigma}} & -\nabla \frac{1}{\sqrt{\sigma}} & \sqrt{\mu} I & 0 \\ \nabla \cdot \sqrt{\sigma} & 0 & 0 & \frac{1}{\sqrt{\mu}} a_2 \end{bmatrix} \begin{bmatrix} \sqrt{\sigma}\mathbf{E} \\ \sqrt{\sigma} s \\ \sqrt{\mu}\mathbf{H} \\ \sqrt{\mu} k \end{bmatrix} = \mathbf{F}. \tag{5.1}$$

Then, the corresponding dual problem has the form

$$\mathcal{L}_s^* \mathbf{U}^* = \begin{bmatrix} -\sqrt{\sigma} I & 0 & \frac{1}{\sqrt{\sigma}} \nabla \times & -\sqrt{\sigma} \nabla \\ 0 & -\frac{1}{\sqrt{\sigma}} a_1 & \frac{1}{\sqrt{\sigma}} \nabla \cdot & 0 \\ \frac{1}{\sqrt{\mu}} \nabla \times & -\sqrt{\mu} \nabla & \sqrt{\mu} I & 0 \\ \frac{1}{\sqrt{\mu}} \nabla \cdot & 0 & 0 & \frac{1}{\sqrt{\mu}} a_2 \end{bmatrix} \begin{bmatrix} \mathcal{U} \\ p \\ \mathcal{V} \\ q \end{bmatrix} = \begin{bmatrix} \sqrt{\sigma}\mathbf{E} \\ \sqrt{\sigma} s \\ \sqrt{\mu}\mathbf{H} \\ \sqrt{\mu} k \end{bmatrix}. \tag{5.2}$$

FOSLL* for the scaled system minimizes the dual functional $\mathcal{F}_s^*(\mathbf{U}^*; \mathbf{U}) = ||\mathcal{L}_s^* \mathbf{U}^* - \mathbf{U}||^2$ in the weak sense as follows: Find $\mathbf{U}^* \in D(\mathcal{L}^*)$ that satisfies

$$\langle \mathcal{L}_s^* \mathbf{U}^*, \mathcal{L}_s^* \mathbf{V}^* \rangle = \langle \mathbf{U}, \mathcal{L}_s^* \mathbf{V}^* \rangle = \langle \mathcal{L}_s \mathbf{U}, \mathbf{V}^* \rangle = \langle \mathbf{F}, \mathbf{V}^* \rangle \tag{5.3}$$

for all $\mathbf{V}^* \in D(\mathcal{L}^*)$. To gain insight into the effectiveness of the scaled approach in FOSLL*, we observe the formal normal, $\mathcal{L}_s \mathcal{L}_s^*$, of (5.3):

$$\begin{bmatrix} \sigma I + \nabla \times \frac{1}{\mu} \nabla \times - \nabla \frac{1}{\mu} \nabla \cdot & 0 & 0 & \sigma \nabla - \nabla \frac{a_2}{\mu} \\ 0 & \frac{a_1^2}{\sigma} - \nabla \cdot \mu \nabla & \nabla \cdot \mu - \frac{a_1}{\sigma} \nabla \cdot & 0 \\ 0 & \nabla \frac{a_1}{\sigma} - \mu \nabla & \nabla \times \frac{1}{\sigma} \nabla \times - \nabla \frac{1}{\sigma} \nabla \cdot + \mu I & 0 \\ \frac{a_2}{\mu} \nabla \cdot - \nabla \cdot \sigma & 0 & 0 & \frac{a_2^2}{\mu} - \nabla \cdot \sigma \nabla \end{bmatrix}.$$

Compare the above to the formal normal, $\mathcal{L}\mathcal{L}^*$, of the original system (3.1). The formal normal of the scaled system provides two small systems, each totally separated,

corresponding to the variables $(\mathcal{U}, q)$ and $(\mathcal{V}, p)$, respectively:

$$\begin{bmatrix} \sigma I + \nabla \times \frac{1}{\mu} \nabla \times - \nabla \frac{1}{\mu} \nabla \cdot & \sigma \nabla - \nabla \frac{a_2}{\mu} \\ \frac{a_2}{\mu} \nabla \cdot - \nabla \cdot \sigma & \frac{a_2^2}{\mu} - \nabla \cdot \sigma \nabla \end{bmatrix} \begin{bmatrix} \mathcal{U} \\ q \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix} \tag{5.4}$$

and

$$\begin{bmatrix} \mu I + \nabla \times \frac{1}{\sigma} \nabla \times - \nabla \frac{1}{\sigma} \nabla \cdot & \nabla \frac{a_1}{\sigma} - \mu \nabla \\ \nabla \cdot \mu - \frac{a_1}{\sigma} \nabla \cdot & \frac{a_1^2}{\sigma} - \nabla \cdot \mu \nabla \end{bmatrix} \begin{bmatrix} \mathcal{V} \\ p \end{bmatrix} = \begin{bmatrix} \mu \mathbf{H}_{\text{old}} \\ 0 \end{bmatrix}. \tag{5.5}$$

The weak form also separates and we solve two smaller systems. For the eddy current problem, it is clear that $(\mathcal{U}, q) = (\mathbf{0}, 0)$. For more general formulations, both systems might have a nontrivial solution.

Remark 5.1. If σ, μ are constants and $a_1 = a_2 = \sigma \cdot \mu$, then (5.5) is reduced to

$$\begin{bmatrix} \mu I + \frac{1}{\sigma} (\nabla \times \nabla \times - \nabla \nabla \cdot) & 0 \\ 0 & \sigma \mu^2 - \mu \nabla \cdot \nabla \end{bmatrix} \begin{bmatrix} \mathcal{V} \\ p \end{bmatrix} = \begin{bmatrix} \mu \mathbf{H}_{\text{old}} \\ 0 \end{bmatrix}.$$

Clearly, $p = 0$ and $\mathcal{V}$ satisfies

$$\mu \mathcal{V} + \frac{1}{\sigma} \nabla \times \nabla \times \mathcal{V} - \frac{1}{\sigma} \nabla \nabla \cdot \mathcal{V} \;=\; \mu \mathbf{H}_{\text{old}}.$$

The above equation is the same as a modified Galerkin formulation for the magnetic field, $\mathbf{H}$. In the context of constant σ and μ, using FOSLL* with the square root scaling described in (5.1) and certain values for a_1, a_2 is equivalent to solving the original problem (2.6) by eliminating the electric field, $\mathbf{E}$, and using a modified Galerkin formulation on $\mathbf{H}$. However, it is the case of nonconstant σ and μ and the presence of reentrant edges that we consider in this paper.

Remark 5.2. In the modified FOSLL*, the formal normal of (5.3) is

$$\begin{bmatrix} \sigma I + \nabla \times \frac{1}{\mu} \nabla \times - \nabla \frac{r^{2\alpha}}{\mu} \nabla \cdot & 0 & 0 & \sigma \nabla - \nabla \frac{r^{2\alpha} a_2}{\mu} \\ 0 & \frac{r^{2\alpha} a_1^2}{\sigma} - \nabla \cdot \mu \nabla & \nabla \cdot \mu - \frac{r^{2\alpha} a_1}{\sigma} \nabla \cdot & 0 \\ 0 & \nabla \frac{r^{2\alpha} a_1}{\sigma} - \mu \nabla & \mu I + \nabla \times \frac{1}{\sigma} \nabla \times - \nabla \frac{r^{2\alpha}}{\sigma} \nabla \cdot & 0 \\ \frac{r^{2\alpha} a_2}{\mu} \nabla \cdot - \nabla \cdot \sigma & 0 & 0 & \frac{r^{2\alpha} a_2^2}{\mu} - \nabla \cdot \sigma \nabla \end{bmatrix}.$$

Because of the weighting terms, there is no simple way to further decouple the equations through a choice of a_1 and a_2. The term in the (3,3) position in the above is similar to the formal normal associated with the partially weighted modified Galerkin described in [10].

6. Discrete approximation. Let $\mathcal{T}_h$ be a partition of the domain $\overline{\Omega} = \cup_{K \in \mathcal{T}_h} K$, and each finite element $K \in \mathcal{T}_h$ be a closed subset of $\overline{\Omega}$ with $h := \max\{h_K := diam(K) : K \in \mathcal{T}_h\}$. Assume that the partition $\mathcal{T}_h$ is regular so that we can choose a finite element basis that is conforming and satisfies the approximation property (see [6]). We also assume that there exists a constant, ρ, satisfying $h \leq \rho h_K$. Define by P_k the space of all polynomials of degree $\leq k$ with respect to each variable. Let the standard polynomial interpolation operator, $I^h \in \mathcal{L}((H^1(\Omega))^8; (H^1(\Omega))^8)$, be such that $I^h p = p$ for all $p \in (P_1)^8$, and let the finite dimensional subspace, $\mathcal{W}^h \subset D(\mathcal{L}^*) \cap H^1(\Omega)^8$, have $I^h(D(\mathcal{L}^*) \cap H^1(\Omega)^8 \cap \mathcal{C}^0(\Omega)) \subset \mathcal{W}^h$.

From section 3, we know that, for given $\mathbf{U} \in L^2(\Omega)^8$, there exists the solution $\mathbf{U}^* \in D(\mathcal{L}^*)$ satisfying $\mathcal{L}^*\mathbf{U}^* = \mathbf{U}$, that is,

$$\mathbf{U}^* = \arg \min_{\mathcal{X} \in D(\mathcal{L}^*)} ||\mathcal{L}^*\mathcal{X} - \mathbf{U}||_\alpha. \tag{6.1}$$

Here, we minimize in (6.1) over a finite-dimensional subspace $\mathcal{W}^h$ which yields the corresponding weak form as follows: Find $\mathbf{U}^h \in \mathcal{W}^h$ satisfying

$$\left\langle \mathcal{L}^*\mathbf{U}^h, \mathcal{L}^*\mathbf{X}^h \right\rangle_\alpha = \left\langle \mathbf{U}, \mathcal{L}^*\mathbf{X}^h \right\rangle_\alpha = \left\langle (\mathbf{0}, 0, \mu\mathbf{H}_{\text{old}}, 0), \mathbf{X}^h \right\rangle \tag{6.2}$$

for all $\mathbf{X}^h \in \mathcal{W}^h$. By computing $\mathcal{L}^*\mathbf{U}^h$, we obtain the approximations for $\mathbf{E}$ and $\mathbf{H}$:

$$\mathbf{E}^h = -\sigma\mathcal{U}^h + \nabla \times \mathcal{V}^h - \sigma\nabla\tilde{q}^h, \quad \mathbf{H}^h = \nabla \times \mathcal{U}^h - \mu\nabla\tilde{p}^h + \mu\mathcal{V}^h, \tag{6.3}$$

where $\mathbf{U}^h = \left(\mathcal{U}^h, \tilde{p}^h, \mathcal{V}^h, \tilde{q}^h\right)$.

The following theorem provides the L^2-error estimates for the solution $\mathbf{E}$ and $\mathbf{H}$ of (2.6) with the approximation $\mathcal{L}^*\mathbf{U}^h$. Here, we use Theorem 4.15 to accomplish the L^2-error estimates by adopting the standard finite element approximation property. Vectors $(\mathbf{E}, s, \mathbf{H}, k)$, $(\mathcal{U}, p, \mathcal{V}, q)$, $(\mathcal{U}_n, p, \mathcal{V}_n, q)$, and $(\mathcal{U}_n^h, p^h, \mathcal{V}_n^h, q^h)$ are abbreviated to $\mathbf{U}$, $\mathbf{U}^*, \mathbf{U}_n$, and $\mathbf{U}_n^h$, respectively.

THEOREM 6.1. *Assume $\mathbf{U} \in D(\mathcal{L})$ and $\alpha > 1 - \lambda$. Let $\mathbf{U}^* = (\mathcal{U}, p, \mathcal{V}, q) \in D(\mathcal{L}^*)$ such that $\mathcal{L}^*\mathbf{U}^* = \mathbf{U}$. Then, Corollary* 4.16 *leads the decompositions $\mathcal{U} = \mathbf{u} + \nabla\phi$ and $\mathcal{V} = \mathbf{v} + \nabla\psi$. Assume $\mathbf{u}, \mathbf{v} \in H^{1+\eta_1}(\Omega)^3$ and $p, q \in H^{1+\eta_2}(\Omega)$ for some $\eta_1, \eta_2 > 0$. If $\mathbf{U}^h \in \mathcal{W}^h$ satisfies* (6.2), *then there exists a constant c such that*

$$||\mathbf{U} - \mathcal{L}^*\mathbf{U}^h||_\alpha^2 \le c\, h^{2\tau}\left(|\mathbf{u}|_{1+\eta_1}^2 + |\mathbf{v}|_{1+\eta_1}^2 + ||\phi||_{3,1+\beta}^2 + ||\psi||_{3,1+\beta}^2 + |p|_{1+\eta_2}^2 + |q|_{1+\eta_2}^2\right)$$

for any $\tau < \min\left\{\eta_1, \eta_2, \frac{\alpha-1+\lambda}{\alpha+1}\right\}$ and some $\beta \in (1-\lambda, 1)$, $\beta < \alpha$.

Proof. Let $\mathbf{U}_n \in D(\mathcal{L}^*) \cap H^1(\Omega)^8$ satisfying Theorem 4.15, and let

$$\mathbf{U}_n^h = (\mathcal{U}_n^h, p^h, \mathcal{V}_n^h, q^h) = \arg \min_{\mathcal{X}_n^h \in \mathcal{W}^h} ||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*\mathcal{X}_n^h||_\alpha. \tag{6.4}$$

By the triangle inequality,

$$||\mathbf{U} - \mathcal{L}^*\mathbf{U}^h||_\alpha^2 \le 3\left(||\mathcal{L}^*\mathbf{U}^* - \mathcal{L}^*\mathbf{U}_n||_\alpha^2 + ||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*\mathbf{U}_n^h||_\alpha^2 + ||\mathcal{L}^*\mathbf{U}_n^h - \mathcal{L}^*\mathbf{U}^h||_\alpha^2\right).$$

From Theorems 4.12, 4.13, and 4.15, we have

$$||\mathbf{U} - \mathcal{L}^*\mathbf{U}_n||_\alpha^2 \;<\; c\, n^{-2(\alpha-\beta)}\left(||\phi||_{2,\beta}^2 + ||\psi||_{2,\beta}^2\right). \tag{6.5}$$

The linearity of $\mathcal{L}^*$ and the optimality on the finite-dimensional space imply

$$\begin{aligned} ||\mathcal{L}^*\mathbf{U}_n^h - \mathcal{L}^*\mathbf{U}^h||_\alpha^2 &= \left\langle \mathcal{L}^*(\mathbf{U}_n^h - \mathbf{U}_n + \mathbf{U}_n - \mathbf{U}^* + \mathbf{U}^* - \mathbf{U}^h), \mathcal{L}^*(\mathbf{U}_n^h - \mathbf{U}^h)\right\rangle_\alpha \\ &\le ||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*\mathbf{U}^*||_\alpha \, ||\mathcal{L}^*\mathbf{U}_n^h - \mathcal{L}^*\mathbf{U}^h||_\alpha. \end{aligned} \tag{6.6}$$

Thus, (6.5) and (6.6) yield

$$||\mathbf{U} - \mathcal{L}^*\mathbf{U}^h||_\alpha^2 \le c\, n^{-2(\alpha-\beta)}\left(||\phi||_{2,\beta}^2 + ||\psi||_{2,\beta}^2\right) + c\, ||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*\mathbf{U}_n^h||_\alpha^2. \tag{6.7}$$

Since $\mathbf{U}_n^h$ satisfies (6.4), by Céa's lemma, $||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*\mathbf{U}_n^h||_\alpha^2 \leq c\,||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*I^h\mathbf{U}_n||_\alpha^2$. Using the triangle inequality, we have

$$\begin{aligned}||\mathcal{L}^*\mathbf{U}_n - \mathcal{L}^*I^h\mathbf{U}_n||_\alpha^2 \leq c\,\big(&||\mathcal{U}_n - I^h\mathcal{U}_n||^2 + ||\mathcal{V}_n - I^h\mathcal{V}_n||^2\\ &+ ||\nabla\times(\mathcal{U}_n - I^h\mathcal{U}_n)||^2 + ||\nabla\cdot(\mathcal{U}_n - I^h\mathcal{U}_n)||_{0,\alpha}^2 + ||\nabla(p - I^hp)||^2\\ &+||\nabla\times(\mathcal{V}_n - I^h\mathcal{V}_n)||^2 + ||\nabla\cdot(\mathcal{V}_n - I^h\mathcal{V}_n)||_{0,\alpha}^2 + ||\nabla(q - I^hq)||^2\big)\,.\end{aligned} \tag{6.8}$$

First, we consider $\mathcal{U}_n$-terms. By [12], we have

$$\begin{aligned}&||\mathcal{U}_n - I^h\mathcal{U}_n||^2 + ||\nabla\times(\mathcal{U}_n - I^h\mathcal{U}_n)||^2 + ||\nabla\cdot(\mathcal{U}_n - I^h\mathcal{U}_n)||_{0,\alpha}^2\\ &\qquad\leq c||\nabla(\mathcal{U}_n - I^h\mathcal{U}_n)||^2 = c\sum_{K\in\mathcal{T}_h}||\nabla(\mathcal{U}_n - I^h\mathcal{U}_n)||_K^2,\end{aligned}$$

where $||\cdot||_K$ means an integration over K. Since ϕ satisfies

$$\left\{\begin{array}{rcll}\nabla\cdot\nabla\phi &=& -\nabla\cdot\mathbf{u} - a_2\tilde{q} + k & \text{in } \Omega,\\ \phi &=& 0 & \text{on } \partial\Omega,\end{array}\right. \tag{6.9}$$

and $\nabla\cdot\mathbf{u} + a_2\tilde{q} - k \in H_\beta^1(\Omega) \subset H_{1+\beta}^1(\Omega)$, the solution, ϕ, of (6.9) is in $H_{1+\beta}^3(\Omega)$ (see [19]). From Theorems 4.12 and 4.15, $\mathcal{U}_n$ is decomposed of $\mathbf{u}+\nabla\delta_n\phi$, where δ_n is defined as in (4.8). The fact that $\phi \in H_{1+\beta}^3(\Omega)$ and the definition of δ_n yield $\delta_n\phi \in H^3(\Omega)$. On each element K, we use the triangle inequality and standard interpolation error estimates to obtain

$$\begin{aligned}||\nabla(\mathcal{U}_n - I^h\mathcal{U}_n)||_K^2 &\leq c\,\big(||\nabla(\mathbf{u} - I^h\mathbf{u})||_K^2 + ||\nabla(\nabla\phi_n - I^h\nabla\phi_n)||_K^2\big)\\ &\leq c\;h^{2\eta_1}|\mathbf{u}|_{1+\eta_1,K}^2 + c\;h^2|\phi_n|_{3,K}^2.\end{aligned}$$

Since $\delta_n = 0$ when $r \leq (1/2n)$ and $\delta_n' = 0$ when $r \notin (1/2n, 1/n)$,

$$\sum_K |\phi_n|_{3,K}^2 = |\phi_n|_3^2 \leq c\int |\delta_n'''\phi|^2 + |\delta_n''\nabla\phi|^2 + |\delta_n'\nabla^2\phi|^2 + |\delta_n\nabla^3\phi|^2 d\Omega$$

$$\leq c\iiint_{\frac{1}{2n}}^{\frac{1}{n}} |n^3\phi|^2 + |n^2\nabla\phi|^2 + |n\nabla^2\phi|^2 r dr d\theta dz + c\iiint_{\frac{1}{2n}}^{R(\theta)} |\nabla^3\phi|^2 r dr d\theta dz$$

$$\leq cn^{2(1+\beta)}\left(\iiint_{\frac{1}{2n}}^{\frac{1}{n}}\sum_{k=0}^{2}|r^{\beta-k}\nabla^{2-k}\phi|^2 d\Omega + \iiint_{\frac{1}{2n}}^{R(\theta)}|r^{1+\beta}\nabla^3\phi|^2 d\Omega\right) \leq cn^{2(1+\beta)}||\phi||_{3,1+\beta}^2.$$

Thus, we have $||\nabla(\mathcal{U}_n - I^h\mathcal{U}_n)||^2 \leq\; c\big(h^{2\eta_1}|\mathbf{u}|_{1+\eta_1}^2 + h^2n^{2(1+\beta)}||\phi||_{3,1+\beta}^2\big)$. Choose n such that $\frac{1}{2n} < \sqrt{2}h^{\frac{1}{\alpha+1}} < \frac{11}{20n}$ to balance with (6.7). Then, the optimal choice of β is $1-\lambda+\epsilon$ and this yields $hn^{1+\beta} = h^{\frac{\alpha-1+\lambda}{\alpha+1}-\epsilon}$. Then,

$$||\nabla(\mathcal{U}_n - I^h\mathcal{U}_n)||^2 \leq\; c\left(h^{2\eta_1}|\mathbf{u}|_{1+\eta_1}^2 + h^{2\frac{\alpha-1+\lambda}{\alpha+1}-\epsilon}||\phi||_{3,1+\beta}^2\right).$$

The above calculation can be applied to $\mathcal{V}_n$ analogously. For p and q, the standard error estimates yields $||\nabla(p-I^hp)||^2 + ||\nabla(q-I^hq)||^2 \leq ch^{2\eta_2}(|p|_{1+\eta_2}^2 + |q|_{1+\eta_2}^2)$. ∎

COROLLARY 6.2. *If μ, σ are constants, then η_1, η_2 are any real values $< \lambda$.*

Proof. If μ and σ are constants, then $\nabla \times \nabla \times (\mathbf{u} + \nabla\phi) \in L^2(\Omega)$, where $\mathbf{u}$ and ϕ are from the proof of Theorem 6.1. Also, $\nabla \cdot (\mathbf{u} + \nabla\phi) = 0$ and $\mathbf{n} \times (\mathbf{u} + \nabla\phi) = 0$ on $\partial\Omega$. Thus, by [9], we have $\mathbf{u} \in H^{1+\eta_1}(\Omega)^3$ for any $\eta_1 < \lambda$. The variable p is the solution of the Poisson equation with a Dirichlet boundary condition. Thus, $p \in H^{1+\eta_2}(\Omega)$, where $\eta_2 < \lambda$. Similarly, we have $\mathbf{v} \in H^{1+\eta_1}(\Omega)^3$ and $q \in H^{1+\eta_2}(\Omega)$ for any $\eta_1, \eta_2 < \lambda$. ∎

Remark 6.3. In [10], error estimates in the D_{A_α}-norm (see (4.9)) with higher regularity in $\mathbf{E}$ were developed. They used H^1-conforming finite element spaces which include $\nabla\Phi^h$, where Φ^h is an almost affine family of $\mathcal{C}^1$ elements and has good approximation properties in the H^2_β-norm. In this paper, we use H^1-conforming finite elements to approximately solve the problem and develop L^2-error estimates. Our approximation to the electric field is of the form $E^h = -\sigma\mathcal{U}^h + \nabla \times \mathcal{V}^h - \sigma\nabla\tilde{q}^h$, where $\mathcal{U}^h, \mathcal{V}^h$, and $\tilde{q}^h$ are chosen from H^1-conforming finite element spaces, which means we explicitly present the solution as a combination of such terms, and thus, do not need to construct special finite element spaces.

In the following section, we present several numerical examples. The results show clearly that the convergence rate is related to α values as well as to the regularity of the dual solution in agreement with the above theorem.

7. Numerical results. In this section, we report on numerical results of applying the modified FOSLL* method to problem (3.1). We choose the prototype domain described by

$$\Omega = (-0.5, 0.5)^3 \backslash \{(x, y, z) | 0 \le x \le 0.5, -0.5 \le y \le 0, -0.5 < z < 0.5\}.$$

The domain has a reentrant edge along the z-axis with interior angle $\frac{3\pi}{2}$. Thus, we expect the solution to have a singularity of the form $r^{-\frac{1}{3}}$, where r is the distance to the z-axis. The square root scaling described in section 5 was used for all three tests. This requires solving for only four dependent variables, denoted by $(\mathcal{V}, p)$, since the other four variables $(\mathcal{U}, q)$ are known to be zero. Trilinear finite elements were used for all variables. In this context, we minimize $||\mathcal{L}^*\mathcal{X}^h - (\mathbf{E}, 0, \mathbf{H}, 0)||_\alpha$ over $\mathcal{X}^h = (\mathcal{U}^h, p^h, \mathcal{V}^h, q^h)$ in the finite-dimensional subspace $\mathcal{W}^h$, holding $(\mathcal{U}^h, q^h) = (\mathbf{0}, 0)$, in order to get the approximation, $\mathbf{U}^h$, for the dual solution, $\mathbf{U}^*$, of (3.5). Then, we compute $\mathcal{L}^*\mathbf{U}^h$ as the approximation for $(\mathbf{E}, 0, \mathbf{H}, 0)$ and observe the L^2-errors $||\mathbf{E} - \mathbf{E}^h||$ and $||\mathbf{H} - \mathbf{H}^h||$.

The software package FOSPACK [22] was used to construct the discrete systems and to solve them by a conjugate gradient iteration preconditioned by algebraic multigrid (AMG) using W(1,1)-cycles. Problems with given exact solutions were constructed so that the error could be monitored. The constants a_1 and a_2 were fixed at 0. However, the results are similar to those achieved when they are fixed as positive constants. A residual reduction 10^{-10} was used as the AMG W-cycle stopping criterion. While this level of error is excessive in practice, we employ it here to remove algebraic error from the calculation of the convergence of the discrete solution.

Example 7.1. We choose the exact solutions $\mathbf{E}$ and $\mathbf{H}$ to be

$$\mathbf{E} = \frac{1}{\sigma}\nabla \times \mathbf{H} \quad \text{and} \quad \mathbf{H} = (\ \partial_y g,\ -\partial_x g,\ 0\ ),$$

where

$$g = \delta(r) r^{\frac{2}{3}} \sin\left(\frac{2}{3}\theta\right) \sin(2\pi z) \quad \text{and} \quad \delta(r) = \begin{cases} 1, & r \le 0.25, \\ 0, & r \ge 0.375 \end{cases}$$

TABLE 7.1
The L^2-norm of the errors and observed convergence rates, τ, for Example 7.1 ($\times 10^{-1}$ *means that the values in the table divide by* 10), $\|\mathbf{E}\| \sim 10.290$, $\|\mathbf{H}\| \sim 0.55727$.

$\|\|\mathbf{E}-\mathbf{E}^h\|\|$										
	$\alpha=0$		$\alpha=2/3$		$\alpha=4/3$		$\alpha=2$		$\alpha=3$	
1/8	4.67	τ	4.65	τ	4.64	τ	4.64	τ	4.63	τ
1/16	3.91	0.26	3.84	0.28	3.80	0.29	3.79	0.29	3.79	0.29
1/32	2.16	0.85	1.97	0.96	1.91	0.99	1.89	1.00	1.88	1.01
1/64	1.45	0.57	1.08	0.86	1.00	0.94	0.97	0.96	0.96	0.97
$\|\|\mathbf{H}-\mathbf{H}^h\|\|$ ($\times 10^{-1}$)										
	$\alpha=0$		$\alpha=2/3$		$\alpha=4/3$		$\alpha=2$		$\alpha=3$	
1/8	2.32	τ	2.18	τ	2.11	τ	2.06	τ	2.02	τ
1/16	1.74	0.41	1.36	0.68	1.11	0.93	2.06	1.06	2.02	1.11
1/32	1.57	0.16	0.92	0.56	0.55	1.00	0.45	1.14	0.41	1.19
1/64	1.52	0.04	0.69	0.42	0.31	0.85	0.23	0.94	0.21	0.94

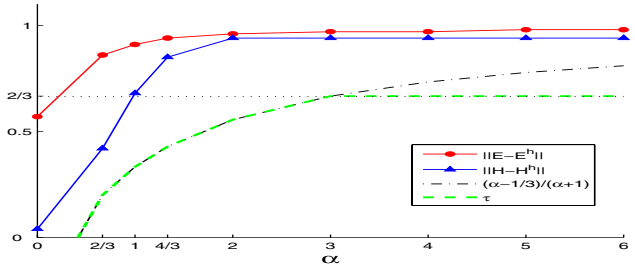


FIG. 7.1. *Finite element convergence rate, τ, as a function of α for Example* 7.1.

TABLE 7.2
AMG convergence factors for Example 7.1.

	$\alpha=0$	$\alpha=2/3$	$\alpha=4/3$	$\alpha=2$	$\alpha=3$	$\alpha=4$	$\alpha=5$	$\alpha=6$
1/8	0.03	0.03	0.04	0.05	0.05	0.06	0.07	0.07
1/16	0.03	0.05	0.09	0.14	0.28	0.23	0.20	0.20
1/32	0.03	0.17	0.20	0.29	0.33	0.37	0.42	0.44
1/64	0.03	0.14	0.32	0.40	0.44	0.51	0.54	0.54

with $r = \sqrt{x^2+y^2}$, $\theta = \arctan(\frac{y}{x})$, and $\delta(r) \in \mathcal{C}^3$ cut-off function. Then, the solution satisfies type II boundary conditions. We fix the $\mu = 1$ and $\sigma = 1$.

Table 7.1 displays the L^2-errors of $\mathbf{E}$ and $\mathbf{H}$. The rate, τ, represents the value of the observed convergent factor, h^τ, when the mesh decreases from h to $h/2$. As shown in Table 7.1, standard FOSLL* ($\alpha = 0$) gives poor convergence. The declines in convergence factors are dramatic in this case. This is to be expected because the exact dual solutions $\mathcal{U}$ and $\mathcal{V}$ are not in H^1, but rather in H^γ for any $\gamma < \frac{2}{3}$. The results in Table 7.1 for $\alpha > 1 - \lambda = \frac{1}{3}$ show that partial unweighting of the functional produces improved convergence in all terms of the functional. By Theorem 6.1, the L^2-errors of $\mathbf{E}$ and $\mathbf{H}$ are expected to exhibit $O(h^\tau)$, for any $\tau < \min\{\frac{2}{3}, \frac{\alpha-\frac{1}{3}}{\alpha+1}\}$ (dashed line in Figure 7.1) as long as $\alpha > \frac{1}{3}$, that is, the bound τ, on the convergence rate stays at $\frac{2}{3}$ for $\alpha > 3$. In fact, the results show better convergence than expected. In Figure 7.1, we compare convergence rates for the L^2-errors in $\mathbf{E}$ and $\mathbf{H}$ while the mesh moves from 1/32 to 1/64 with more α values than are showed in Table 7.1. We observe in Table 7.2 that increasing α results in an increasing convergence factor for

TABLE 7.3

The L^2-norm of the errors and observed convergence rates, τ, for Example 7.2, $\|\mathbf{E}\| \sim 1.9302$, $\|\mathbf{H}\| \sim 0.55727$.

$\|\|\mathbf{E}-\mathbf{E}^h\|\|$ $(\times 10^{-1})$										
	$\alpha=0$		$\alpha=2/3$		$\alpha=4/3$		$\alpha=2$		$\alpha=3$	
1/8	9.56	τ	9.27	τ	9.13	τ	9.04	τ	8.96	τ
1/16	8.57	0.16	7.75	0.26	7.34	0.31	7.20	0.33	7.14	0.33
1/32	6.77	0.34	5.06	0.62	4.49	0.71	4.26	0.76	4.04	0.82
1/64	6.20	0.13	3.77	0.43	3.07	0.55	2.70	0.66	2.38	0.76
$\|\|\mathbf{H}-\mathbf{H}^h\|\|$ $(\times 10^{-1})$										
	$\alpha=0$		$\alpha=2/3$		$\alpha=4/3$		$\alpha=2$		$\alpha=3$	
1/8	2.34	τ	2.24	τ	2.18	τ	2.14	τ	2.09	τ
1/16	1.83	0.36	1.60	0.48	1.38	0.65	1.22	0.81	1.09	0.94
1/32	1.71	0.10	1.33	0.26	0.93	0.58	0.68	0.84	0.52	1.05
1/64	1.68	0.02	1.18	0.18	0.64	0.53	0.39	0.81	0.26	0.98

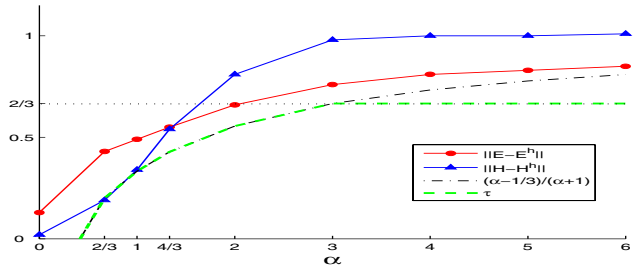


FIG. 7.2. *Finite element convergence rate, τ, as a function of α for Example* 7.2.

the AMG algorithm. This behavior is dependent on the particular AMG algorithm that was used in the test. An improved AMG would change the picture.

Example 7.2. In this example, we take a smooth function for the coefficient σ. Let $\mathbf{E}$ and $\mathbf{H}$ be the same as in Example 7.1 and let $\mu = 0.5$ and $\sigma = 100(x^2+y^2)+1$.

Table 7.3 shows the L^2-errors of $\mathbf{E}$ and $\mathbf{H}$ and the convergence rates. More convergence rates corresponding to α values when the mesh moves from 1/32 to 1/64 appear in Figure 7.2. Note that the observed convergence rates are slightly worse than the ones in Example 7.1. The AMG convergence factor behaves essentially the same as in the first example.

In the next example, we examine the case having discontinuous coefficients as well as a reentrant edge on the boundary.

Example 7.3. Let $\mathbf{E}$ and $\mathbf{H}$ be the same as in Example 7.1. Let $\mu = \sigma = 1$ if $r = \sqrt{x^2+y^2} \le 0.25$ and $\mu = 25$, $\sigma = 100$ otherwise.

In this example, we need to be careful about the regularity of $\mathbf{E}$ and $\mathbf{H}$. Since μ and σ have jumps at $r = 0.25$, $\mathbf{E}$ is not in $H(\nabla\times)$ but in $H(\nabla \times \sigma)$, and $\mathbf{H}$ is not in $H(\nabla \cdot \mu)$ but in $H(\nabla\cdot)$. $\mathbf{E}$ and $\mathbf{H}$ do not satisfy the eddy current equations, but are useful as a test to observe how modified FOSLL* would work for a problem with both discontinuous coefficients and a reentrant edge. Numerical results in Table 7.4 show great convergence with modified FOSLL* approximation even though the problem has both nongrid-aligned discontinuities in the coefficients and a boundary singularity. Convergence rates of the L^2-errors for $\mathbf{E}$ and $\mathbf{H}$ are greater than both of $\frac{2}{3}$ and $\frac{\alpha-\frac{1}{3}}{\alpha+1}$ for $\alpha > 3$. Figure 7.3 shows convergence rates for more values of α using grid size $h = 1/64$.

TABLE 7.4
The L^2-norm of the errors and observed convergence rates, τ, for Example 7.3, $\|\mathbf{E}\| \sim 8.1056$, $\|\mathbf{H}\| \sim 0.55727$.

$\|\mathbf{E}-\mathbf{E}^h\|$										
	$\alpha=0$		$\alpha=2/3$		$\alpha=4/3$		$\alpha=2$		$\alpha=3$	
1/8	2.37	τ	2.35	τ	2.34	τ	2.34	τ	2.34	τ
1/16	2.16	0.14	2.10	0.16	2.08	0.17	2.07	0.18	2.07	0.18
1/32	1.19	0.86	1.06	0.99	1.02	1.02	1.02	1.03	1.01	1.03
1/64	0.82	0.54	0.59	0.83	0.55	0.89	0.54	0.91	0.53	0.92
$\|\mathbf{H}-\mathbf{H}^h\|$ $(\times 10^{-2})$										
	$\alpha=0$		$\alpha=2/3$		$\alpha=4/3$		$\alpha=2$		$\alpha=3$	
1/8	20.7	τ	20.3	τ	20.1	τ	19.9	τ	19.8	τ
1/16	11.9	0.80	10.4	0.97	9.60	1.07	9.23	1.11	8.99	1.14
1/32	9.35	0.35	6.22	0.74	4.89	0.97	4.40	1.07	4.09	1.14
1/64	9.02	0.05	4.77	0.38	3.17	0.62	2.53	0.80	2.17	0.92

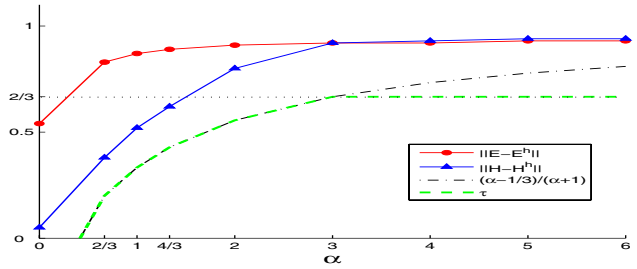


FIG. 7.3. *Finite element convergence rate, τ, as a function of α for Example* 7.3.

TABLE 7.5
AMG convergence factors for Example 7.3.

	$\alpha=0$	$\alpha=2/3$	$\alpha=4/3$	$\alpha=2$	$\alpha=3$	$\alpha=4$	$\alpha=5$	$\alpha=6$
1/8	0.64	0.66	0.66	0.63	0.66	0.66	0.66	0.65
1/16	0.68	0.67	0.68	0.67	0.66	0.67	0.66	0.68
1/32	0.67	0.68	0.68	0.66	0.67	0.66	0.68	0.68
1/64	0.63	0.65	0.65	0.65	0.66	0.67	0.67	0.67

The AMG convergence factors are slightly worse, but still quite acceptable, for discontinuous coefficients, as indicated in Table 7.5. Again, we believe that an improved AMG algorithm may overcome this difficulty.

8. Conclusion. In this paper, we developed a FOSLL* method with a partially weighted norm for the eddy current approximation to Maxwell's equations on a three-dimensional domain with a reentrant edge. We have shown the existence of an H^1-sequence converging to the solution of the eddy current problem in the partially weighted functional norm. This allows accurate approximation using standard H^1-conforming finite element spaces. An L^2-error estimate was established that depends continuously on the weight parameter, α. Numerical tests support our theory. In the future, we will apply our theory to other problems, like full Maxwell's equations, elasticity equations, and Navier–Stokes equations. Also, the reentrant corners (e.g., the Fichera cube) will be considered. We don't anticipate the results, but we believe that our theory can be easily extended to these problems.

REFERENCES

[1] M. Berndt, T. A. Manteuffel, S. F. McCormick, and G. Starke, *Analysis of first-order system least squares (fosls) for elliptic problems with discontinuous coefficients: Part* I, SIAM J. Numer. Anal., 43 (2005), pp. 386–408.

[2] J. H. Bramble, T. V. Kolev, and J. E. Pasciak, *A least-squares approximation method for the time-harmonic Maxwell equations*, J. Numer. Math., 13 (2005), pp. 237–263.

[3] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[4] Z. Cai, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* II, SIAM J. Numer. Anal., 34 (1997), pp. 425–454.

[5] Z. Cai, T. A. Manteuffel, S. F. McCormick, and J. Ruge, *First-order system $\mathcal{LL}^*$ ($FOSLL^*$): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Classics in Appl. Math. 40, SIAM, Philadelphia, 2002; reprint of the 1978 original (North–Holland, Amsterdam).

[7] M. Costabel, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.

[8] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.

[9] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.

[10] M. Costabel and M. Dauge, *Weighted regularization of Maxwell equations in polyhedral domains. A rehabilitation of nodal finite elements*, Numer. Math., 93 (2002), pp. 239–277.

[11] M. Costabel, M. Dauge, and S. Nicaise, *Singularities of Maxwell interface problems*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 627–649.

[12] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations. Theory and Algorithms*, Springer Ser. Comput. Math. 5, Springer-Verlag, Berlin, 1986.

[13] P. Grisvard, *Singularities in Boundary Value Problems*, Rech. Math. Appl. 22, Masson, Paris, 1992.

[14] B.-N. Jiang, *The Least-Squares Finite Element Method. Theory and Applications in Computational Fluid Dynamics and Electromagnetics,* Sci. Comput., Springer-Verlag, Berlin, 1998.

[15] V. A. Kondrat'ev, *The smoothness of a solution of Dirichlet's problem for second-order elliptic equations in a region with a piecewise-smooth boundary*, Differential Equations, 6 (1970), pp. 1392–1401.

[16] V. A. Kondrat'ev, *Boundary value problems for elliptic equations in domains with conical or angular points*, Trudy Moskov. Mat. Obšč., 16 (1967), pp. 209–292 (in Russian).

[17] E. Lee, *Fosll* for Eddy Current Problems with Three-Dimensional Edge Singularities*, Ph.D. thesis, University of Colorado, Boulder, 2005.

[18] T. A. Manteuffel, S. F. McCormick, J. Ruge, and J. G. Schmidt, *First-order system $\mathcal{LL}^*$ ($FOSLL^*$) for general scalar elliptic problems in the plane*, SIAM J. Numer. Anal., 43 (2005), pp. 2098–2120.

[19] V. Mazýa, S. Nazarov, and B. Plamenevskij, *Asymptotic Theory of Elliptic Boundary Value Problems in Singularly Perturbed Domains. Vol.* II, Oper. Theory Adv. Appl. 112, Birkhäuser Verlag, Basel, 2000.

[20] S. Nicaise, *Edge elements on anisotropic meshes and approximation of the Maxwell equations*, SIAM J. Numer. Anal., 39 (2001), pp. 784–816.

[21] V. A. Nikiškin, *Singularities of the solution of the Dirichlet problem for a second-order equation in the neighborhood of an edge*, Vestnik Moskov. Univ. Ser. I Mat. Mekh., 2 (1979), pp. 51–62, 103.

[22] J. Ruge, *Fospack: A first-order system least-squares (fosls) code*, in preparation.

[23] J. Ruge and K. Stüben, *Efficient solution of finite difference and finite element equations*, in Multigrid Methods for Integral and Differential Equations (Bristol, 1983), Inst. Math. Appl. Conf. Ser. New Ser. 3, Oxford University Press, New York, 1985, pp. 169–212.

[24] C. R. Westphal, *First-Order System Least Squares for Geometrically-Nonlinear Elasticity in Nonsmooth Domain*, Ph.D. thesis, University of Colorado, Boulder, 2004.

[25] K. Yosida, *Functional analysis*, reprint of the sixth (1980) edition, Classics Math., Springer-Verlag, Berlin, 1995.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 810–836

DETECTING INTERFACES IN A PARABOLIC-ELLIPTIC PROBLEM FROM SURFACE MEASUREMENTS*

FLORIAN FRÜHAUF†, BASTIAN GEBAUER‡, AND OTMAR SCHERZER†

Abstract. Assuming that the heat capacity of a body is negligible outside certain inclusions the heat equation degenerates to a parabolic-elliptic interface problem. In this work we aim to detect these interfaces from thermal measurements on the surface of the body. We deduce an equivalent variational formulation for the parabolic-elliptic problem and give a new proof of the unique solvability based on Lions's projection lemma. For the case that the heat conductivity is higher inside the inclusions, we develop an adaptation of the factorization method to this time-dependent problem. In particular this shows that the locations of the interfaces are uniquely determined by boundary measurements. The method also yields to a numerical algorithm to recover the inclusions and thus the interfaces. We demonstrate how measurement data can be simulated numerically by a coupling of a finite element method with a boundary element method, and finally we present some numerical results for the inverse problem.

Key words. parabolic-elliptic equation, inverse problems, factorization method

AMS subject classifications. 65J20, 35K65

DOI. 10.1137/050641545

1. Introduction.

We consider the heat equation in a domain $B \subset \mathbb{R}^n$

$$\partial_t(c(x)u(x,t)) - \nabla \cdot (\kappa(x)\,\nabla u(x,t)) = 0 \quad \text{in } B\times]0,T[, \tag{1.1}$$

with (spatially dependent) heat capacity c and conductivity κ. The special case we are studying here is that the heat capacity $c(x)$ is bounded from below inside an inclusion $\overline{\Omega} \subset B$, and negligibly small on the outside $Q := B \setminus \overline{\Omega}$ (cf. Figure 1.1 for a sketch of the geometry). Throughout this work Ω is allowed to be disconnected; thus the case of multiple inclusions is covered as well.

If we assume for simplicity that $c(x) = \chi_\Omega(x)$ is the characteristic function of Ω, then the evolution equation (1.1) can be rewritten as a parabolic-elliptic equation,

$$\partial_t u(x,t) - \nabla \cdot (\kappa(x)\,\nabla u(x,t)) = 0 \text{ in } \Omega\times]0,T[, \tag{1.2}$$

$$\nabla \cdot (\kappa(x)\,\nabla u(x,t)) = 0 \text{ in } Q\times]0,T[, \tag{1.3}$$

together with appropriate interface conditions on $\partial\Omega$.

For the case $B = \mathbb{R}^2$ and $\kappa = 1$ this problem also arises in the study of two-dimensional eddy currents and was studied by MacCamy and Suri in [23] and by Costabel, Ervin, and Stephan in [9]. In both papers boundary integral operators are used to replace the Laplace equation in the exterior of Ω by a nonlocal boundary condition for the parabolic equation inside Ω. This problem is then solved by

*Received by the editors September 29, 2005; accepted for publication (in revised form) November 21, 2006; published electronically April 20, 2007.
http://www.siam.org/journals/sinum/45-2/64154.html

†Department of Computer Science, University of Innsbruck, Technikerstr. 21a, 6020 Innsbruck, Austria (florian.fruehauf@uibk.ac.at, otmar.scherzer@uibk.ac.at, http://informatik.uibk.ac.at/infmath). The research of these authors was supported by the Austria Science Foundation (FWF) Projects Y-123INF, FSP 9203-N12, and FSP 9207-N12.

‡Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstr. 69, 4040 Linz, Austria (bastian.gebauer@ricam.oeaw.ac.at, http://www.ricam.oeaw.ac.at).

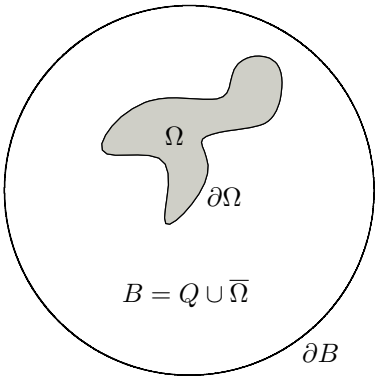


FIG. 1.1. *Sketch of geometry.*

a Galerkin method. In [8] Costabel uses boundary integral operators to solve the resulting interior problem also.

In this work we study the problem for general $\kappa \in L^\infty_+(B)$ in a bounded domain B with given Neumann boundary values on ∂B. By considering (1.1) in the sense of distributions we deduce (1.2), (1.3) together with natural interface conditions (that would otherwise have to be postulated). Moreover, we prove that the weak formulation in appropriate Sobolev spaces is equivalent to (1.1). We show existence of a unique solution using Lions's projection lemma; cf. section 2.

In section 3 we study the inverse problem of locating the interface $\partial\Omega$, resp., the inclusion Ω, from surface measurements on ∂B. If the conductivity is larger inside Ω than in the exterior Q, we show that the points belonging to Ω can be characterized using a variant of the so-called factorization method introduced by Kirsch in [16], generalized by Brühl and Hanke in [6, 5], and since then adapted to various stationary and time-harmonic problems; cf. [1, 2, 7, 15, 17, 18, 19] for more recent contributions. To our knowledge this is the first successful extension of this method to a time-dependent problem.

In section 4 we show how the direct problem can be solved numerically with a coupling of finite element methods and boundary element methods similar to [23]. Using simulated measurements we demonstrate the numerical realization of the factorization method following the ideas of Brühl and Hanke in [6, 5].

2. The direct problem.

2.1. A parabolic-elliptic problem. Let $T > 0$ and Ω, $B \subset \mathbb{R}^n$, $n \geq 2$, be bounded domains with smooth boundaries, $\overline{\Omega} \subset B$, and connected complement $Q := B \setminus \overline{\Omega}$.

In this section we study the parabolic-elliptic problem

$$\partial_t(\chi_\Omega(x)u(x,t)) - \nabla \cdot (\kappa(x)\,\nabla u(x,t)) = 0 \quad \text{in } B\times]0,T[, \tag{2.1}$$

with $\kappa \in L^\infty_+(B)$, where we denote by L^∞_+ the space of L^∞-functions with positive essential infima, and χ_Ω is the characteristic function of Ω.

A standard way to treat an equation like (2.1) is to multiply both sides with a test function followed by a formal partial integration. Assuming additional (also

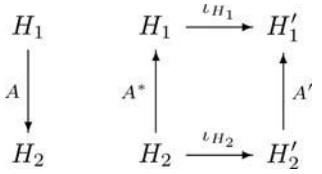


FIG. 2.1. *Relation between dual and adjoint operator.*

formal) boundary and initial conditions, this leads to a variational formulation, which is mathematically meaningful in some Sobolev spaces (and thus no longer formal). Instead of (2.1) one would then study this variational formulation, the so-called weak form of the equation.

In this work we proceed in a slightly different way. We start by noting that the left-hand side of (2.1) does have a mathematical meaning for every $u \in L^2(0,T,H^1(B))$ if the derivatives are interpreted in the sense of (scalar-valued) distributions.

We denote by $\mathcal{D}(B\times]0,T[)$ the space of infinitely often differentiable functions with support in $B\times]0,T[$ and by $\mathcal{D}'(B\times]0,T[)$ its dual space, i.e., the space of distributions on $B\times]0,T[$. By the definition of distributional derivatives, (2.1) is equivalent to

$$-\int_0^T \int_\Omega u(x,t)\partial_t\varphi(x,t)\,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_B \kappa(x)\,\nabla u(x,t)\cdot\nabla\varphi(x,t)\,\mathrm{d}x\,\mathrm{d}t = 0 \tag{2.2}$$

for all $\varphi \in \mathcal{D}(B\times]0,T[)$.

We will show in this section that (2.1) (together with appropriate boundary and initial conditions) has a unique solution in $L^2(0,T,H^1(B))$. In Theorem 2.6 we give an equivalent variational formulation in Sobolev spaces, using the time-derivative in the sense of vector-valued distributions (which we denote by u'). This variational formulation is the same that one would have obtained as the weak generalization of (2.1) using the above-mentioned formal arguments.

We denote by ν the exterior normal on ∂B, resp., the exterior normal on $\partial\Omega$, and by $\mathcal{D}(\overline{Q}\times]0,T[)$ the restrictions of functions from $\mathcal{D}(\mathbb{R}^n\times]0,T[)$ to $Q\times]0,T[$. Analogous notation is used for Ω and B, and $\mathcal{D}(\overline{B}\times[0,T[)$ is the space of restrictions of functions from $\mathcal{D}(\mathbb{R}^n\times]-\infty,T[)$ to $B\times]0,T[$.

We use the anisotropic Sobolev spaces from [22]. For $r,s\geq 0$ we write

$$H^{r,s}(\mathcal{X}) := L^2(0,T,H^r(\mathcal{X}))\cap H^s(0,T,L^2(\mathcal{X})) \text{ for } \mathcal{X}\in\{B,\Omega,Q,\partial B,\partial\Omega\},$$

and for $s<\frac{1}{2}$ and $\mathcal{X}\in\{\partial B,\partial\Omega\}$

$$H^{-r,-s}(\mathcal{X}) := \left(H^{r,s}(\mathcal{X})\right)'.$$

The inner product on a real Hilbert space H is denoted by $(\cdot,\cdot)$ and the dual pairing on $H'\times H$ by $\langle\cdot,\cdot\rangle$. They are related by the isometry $\iota_H : H\to H'$ that "identifies H with its dual"; i.e., $\langle\iota_H u,\cdot\rangle := (u,\cdot)$ for all $u\in H$. Throughout this work we rigorously distinguish between the dual operator (denoted by A') and the adjoint operator (denoted by A^*) of an operator $A\in\mathcal{L}(H_1,H_2)$ between real Hilbert spaces H_1, H_2. They satisfy the identity $A^* = \iota_{H_1}^{-1}A'\iota_{H_2}$; cf. Figure 2.1.

We summarize some known properties of the Dirichlet and Neumann traces for solutions of the Laplace, resp., heat equation. On the boundary $\partial\Omega$ we use the

superscript $^-$ when the trace is taken from inside the inclusion Ω and the superscript $^+$ when it is taken from the outside.

THEOREM 2.1. (a) *The trace mapping*

$$v \mapsto v|_{\partial B}, \text{ resp., } v \mapsto v^+|_{\partial\Omega}, \quad v \in \mathcal{D}(\overline{Q}\times]0,T[),$$

can be extended to a continuous mapping from $H^{1,0}(Q)$ *to* $H^{\frac{1}{2},0}(\partial B)$, *resp., to* $H^{\frac{1}{2},0}(\partial\Omega)$, *that has a continuous right inverse. The same holds for* $H^{1,0}(\Omega) \to H^{\frac{1}{2},0}(\partial\Omega)$, $v \mapsto v^-|_{\partial\Omega}$.

(b) *The Neumann traces* $\kappa\partial_\nu v|_{\partial B}$ *and* $\kappa\partial_\nu v^+|_{\partial\Omega}$ *are defined for every* $v \in H^{1,0}(Q)$ *that solves*

$$\nabla\cdot(\kappa\,\nabla v) = 0 \text{ in } Q\times]0,T[\tag{2.3}$$

by setting

$$\langle\kappa\partial_\nu v|_{\partial B}, f\rangle := \int_0^T\int_Q \kappa\,\nabla v\cdot\nabla v_f\,\mathrm{d}x\,\mathrm{d}t,$$
$$\langle\kappa\partial_\nu v^+|_{\partial\Omega}, \phi\rangle := -\int_0^T\int_Q \kappa\,\nabla v\cdot\nabla v_\phi\,\mathrm{d}x\,\mathrm{d}t$$

for every function f *on* ∂B *and every function* ϕ *on* $\partial\Omega$ *that have extensions* v_f, $v_\phi \in \mathcal{D}(\overline{Q}\times]0,T[)$ *with* $v_f|_{\partial B} = f$, $v_f|_{\partial\Omega} = 0$, *resp.,* $v_\phi|_{\partial B} = 0$, $v_\phi|_{\partial\Omega} = \phi$.

The Neumann traces can be extended to continuous mappings from the subspace of solutions of (2.3) *(equipped with the* $H^{1,0}(Q)$*-norm) to* $H^{-\frac{1}{2},0}(\partial B)$, *resp.,* $H^{-\frac{1}{2},0}(\partial\Omega)$.

(c) *The Neumann trace* $\kappa\partial_\nu v^-|_{\partial\Omega}$ *is defined for every* $v \in H^{1,0}(\Omega)$ *that solves*

$$\partial_t v - \nabla\cdot(\kappa\,\nabla v) = 0 \text{ in } \Omega\times]0,T[\tag{2.4}$$

by setting

$$\langle\kappa\partial_\nu v^-|_{\partial\Omega}, \phi\rangle := \int_0^T\int_\Omega \kappa\,\nabla v\cdot\nabla v_\phi\,\mathrm{d}x\,\mathrm{d}t - \int_0^T\int_\Omega v\,\partial_t v_\phi\,\mathrm{d}x\,\mathrm{d}t$$

for every function ϕ *on* $\partial\Omega$ *that has an extension* $v_\phi \in \mathcal{D}(\overline{\Omega}\times]0,T[)$ *with* $v_\phi|_{\partial\Omega} = \phi$.

The Neumann trace can be extended to a continuous mapping from the subspace of solutions of (2.4) *(equipped with the* $H^{1,0}(\Omega)$*-norm) to* $H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)$.

Proof. (a), (b) immediately follow from the classical trace theorems on H^1. For (c) we refer the reader to [8]. □

Denoting

$$[v]_{\partial\Omega} := v^+|_{\partial\Omega} - v^-|_{\partial\Omega} \text{ and } [\kappa\partial_\nu v]_{\partial\Omega} := \kappa\partial_\nu v^+|_{\partial\Omega} - \kappa\partial_\nu v^-|_{\partial\Omega}$$

we can write (2.1) as a diffraction problem.

LEMMA 2.2. $u \in H^{1,0}(B)$ *solves* (2.1) *if and only if* $u \in H^{1,0}(B\setminus\partial\Omega)$ *solves*

$$\partial_t u - \nabla\cdot(\kappa\,\nabla u) = 0 \text{ in } \Omega\times]0,T[, \tag{2.5}$$
$$\nabla\cdot(\kappa\,\nabla u) = 0 \text{ in } Q\times]0,T[, \tag{2.6}$$
$$[\kappa\partial_\nu u]_{\partial\Omega} = 0, \tag{2.7}$$
$$[u]_{\partial\Omega} = 0. \tag{2.8}$$

In particular, (2.6) *and* (2.7) *imply that* $\kappa\partial_\nu u^-|_{\partial\Omega}$ *can be extended by continuity to* $H^{-\frac{1}{2},0}(\partial\Omega)$.

Proof. Like in the stationary case we have $u \in H^{1,0}(B)$ if and only if $u \in H^{1,0}(B \setminus \partial\Omega)$ and u satisfies (2.8). The rest immediately follows from the definition of distributional derivatives and the Neumann traces. ∎

The next lemma shows uniqueness for the diffraction problem with a Neumann boundary condition and an initial condition on Ω. With respect to the Gelfand triple $H^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^1(\Omega)'$ we denote by

$$W := W(0, T, H^1(\Omega), H^1(\Omega)')$$

the space of functions $u \in L^2(0, T, H^1(\Omega))$ with vector-valued distributional time derivative $u' \in L^2(0, T, H^1(\Omega)')$. From [10, Chp. XVIII], it follows that

$$W \subset C^0([0, T], L^2(\Omega)).$$

LEMMA 2.3. *Let* $u \in H^{1,0}(B \setminus \partial\Omega)$ *solve* (2.5), (2.6), *and*

$$[\kappa\partial_\nu u]_{\partial\Omega} = \psi \in H^{-\frac{1}{2},0}(\partial\Omega), \tag{2.9}$$

$$[u]_{\partial\Omega} = f \in H^{\frac{1}{2},0}(\partial\Omega), \tag{2.10}$$

$$\kappa\partial_\nu u|_{\partial B} = g \in H^{-\frac{1}{2},0}(\partial B). \tag{2.11}$$

Then $u|_\Omega \in W$ *and* u *is uniquely determined by* ψ, f, g, *and the initial condition*

$$u(x, 0) = 0 \text{ on } \Omega. \tag{2.12}$$

Proof. Again (2.9) implies that the Neumann trace $\kappa\partial_\nu u^-|_{\partial\Omega}$ can be extended by continuity to $H^{-\frac{1}{2},0}(\partial\Omega)$.

Thus we can define $w \in L^2(0, T, H^1(\Omega)')$ by setting for every $t \in]0, T[$ and $v \in H^1(\Omega)$

$$\langle w(t), v\rangle := \left\langle \kappa\partial_\nu u^-(t)|_{\partial\Omega}, v^-|_{\partial\Omega}\right\rangle - \int_\Omega \kappa\nabla u(t) \cdot \nabla v \,\mathrm{d}x.$$

We have

$$\begin{aligned}
&\int_\Omega \left(-\int_0^T u\,\partial_t\varphi \,\mathrm{d}t\right) v \,\mathrm{d}x \\
&= \int_0^T \langle \kappa\partial_\nu u^-|_{\partial\Omega}, v^-|_{\partial\Omega}\rangle\varphi \,\mathrm{d}t - \int_0^T \int_\Omega \kappa\nabla u \cdot \nabla v \,\mathrm{d}x\,\varphi \,\mathrm{d}t \\
&= \left\langle \int_0^T w\varphi \,\mathrm{d}t, v \right\rangle
\end{aligned}$$

for all $v(x)\varphi(t) \in \mathcal{D}(\overline{\Omega}\times]0, T[)$ and thus by continuous extension for all $v\varphi \in H^1(\Omega) \otimes \mathcal{D}(]0, T[)$. Thus in the sense of vector-valued distributions

$$w = (u|_\Omega)' \text{ with respect to } H^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^1(\Omega)',$$

and hence $u|_\Omega \in W \subset C^0([0, T], L^2(\Omega))$.

To show uniqueness let $f = 0$, $\psi = 0$, $g = 0$, and (2.12) hold. Since Green's formula holds for functions in W we have

$$\begin{aligned}\frac{1}{2}\int_\Omega |u(T)|^2 \,\mathrm{d}x &= \int_0^T \langle u'(t), u(t)\rangle \,\mathrm{d}t \\ &= \int_0^T \langle \kappa \partial_\nu u^-|_{\partial\Omega}, u^-|_{\partial\Omega}\rangle \,\mathrm{d}t - \int_0^T \int_\Omega \kappa |\nabla u|^2 \,\mathrm{d}x\,\mathrm{d}t \\ &= -\int_0^T \int_B \kappa |\nabla u|^2 \,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

This implies that $u(x,t) = c(t)$, where $c \in C^0([0,T],\mathbb{R})$ solves $c' = 0$ and $c(0) = 0$. Thus $u = 0$. ∎

To show existence of a solution we proceed analogously to [8, Lemma 2.3] by using Lions's projection lemma.

LEMMA 2.4 (Lions's projection lemma). *Assume that H is a Hilbert space and Φ is a subspace of H. Moreover let $a : H \times \Phi \to \mathbb{R}$ be a bilinear form satisfying the following properties:*

(a) *For every $\varphi \in \Phi$, the linear form $u \mapsto a(u,\varphi)$ is continuous on H.*

(b) *There exists $\alpha > 0$ such that $a(\varphi,\varphi) \geq \alpha \|\varphi\|_H^2$ for all $\varphi \in \Phi$.*

Then for each continuous linear form $l \in H'$, there exists $u_0 \in H$ such that

$$a(u_0,\varphi) = \langle l, \varphi\rangle \text{ for all } \varphi \in \Phi \quad \text{and} \quad \|u_0\|_H \leq \frac{1}{\alpha}\|l\|_{H'}.$$

Proof. The lemma is proven in [20]. We repeat the proof for the sake of completeness.

From assumption (a) and the Riesz representation theorem it follows that for every $\varphi \in \Phi$ there exists $K\varphi \in H$ with

$$(u, K\varphi) = a(u,\varphi) \text{ for all } u \in H.$$

This defines a linear (possibly unbounded) operator $K : \Phi \to V := K(\Phi) \subseteq H$. From assumption (b) it follows that K is injective and thus possesses an inverse $R_0 : V \to \Phi$. Again using assumption (b) we have

$$\|R_0 v\|^2 \leq \frac{1}{\alpha} a(R_0 v, R_0 v) = \frac{1}{\alpha}(R_0 v, v) \leq \frac{1}{\alpha}\|R_0 v\|\,\|v\|,$$

which yields $\|R_0 v\| \leq \frac{1}{\alpha}\|v\|$. Thus R_0 can be extended by continuity to the closure $\overline{V}$ of V. If we denote this extension by $\overline{R}_0$ then we have $\overline{R}_0 : \overline{V} \to \overline{\Phi}$.

$\overline{\Phi}$ is a closed subspace of the Hilbert space H and thus also a Hilbert space. Using the Riesz representation theorem on $\overline{\Phi}$ we obtain a $\xi_l \in \overline{\Phi}$ with

$$l(\varphi) = (\xi_l, \varphi) \text{ for all } \varphi \in \overline{\Phi}.$$

Finally, let $P : H \to \overline{V}$ be the orthogonal projection onto $\overline{V}$; then $u_0 := P^* \overline{R}_0{}^* \xi_l$ has the desired properties. ∎

We prove existence of a solution of the parabolic-elliptic diffraction problem (2.5), (2.6), (2.9)–(2.12) under the additional assumption that g and ψ have vanishing integral mean. For $\mathcal{X} \in \{\partial B, \partial\Omega\}$ we define

$$H_\diamond^{-\frac{1}{2}}(\mathcal{X}) := \{g \in H^{-\frac{1}{2}}(\mathcal{X}) \,:\, \langle g, 1_{\mathcal{X}}\rangle = 0\} \quad \text{and} \quad H_\diamond^{-\frac{1}{2},0}(\mathcal{X}) := L^2(0,T,H_\diamond^{-\frac{1}{2}}(\mathcal{X})).$$

Again they are Hilbert spaces because they are closed subspaces of $H^{-\frac{1}{2}}(\mathcal{X})$, resp., $H^{-\frac{1}{2},0}(\mathcal{X})$.

LEMMA 2.5. *For every*

$$g \in H_{\diamond}^{-\frac{1}{2},0}(\partial B),\ f \in H^{\frac{1}{2},0}(\partial\Omega),\ \text{and } \psi \in H_{\diamond}^{-\frac{1}{2},0}(\partial\Omega),$$

there exists $u \in H^{1,0}(B \setminus \partial\Omega)$ *that solves* (2.5), (2.6), *and* (2.9)–(2.12).

u *depends continuously on* g, f, *and* ψ, *and it fulfills*

$$\int_{\Omega} u(x,t)\,\mathrm{d}x = 0 \quad \text{for } t \in [0,T] \text{ a.e.}$$

Proof. Let $\gamma^{-}_{\partial\Omega} : H^{\frac{1}{2}}(\partial\Omega) \to H^1(Q)$ be a lifting operator, i.e., a continuous right inverse of the trace operator $\cdot|_{\partial\Omega}$ with $(\gamma^{-}_{\partial\Omega}h)|_{\partial B} = 0$ for all $h \in H^{\frac{1}{2}}(\partial\Omega)$, and set $u_f = \gamma^{-}_{\partial\Omega} f \in H^{1,0}(Q)$.

We define the spaces

$$H^1_{\square}(B) := \left\{ v \in H^1(B) \ :\ \int_{\Omega} v\,\mathrm{d}x = 0 \right\}, \qquad H \ := \ L^2(0,T,H^1_{\square}(B)),$$
$$\Phi := \left\{ \varphi \in \mathcal{D}([0,T[\times\overline{B}) \ :\ \int_{\Omega} \varphi\,\mathrm{d}x = 0 \right\},$$

and we set for all $v \in H$ and $\varphi \in \Phi$

$$a(v,\varphi) := \int_0^T \int_B \kappa \nabla v \cdot \nabla\varphi\,\mathrm{d}x\,\mathrm{d}t - \int_0^T \int_{\Omega} v\,\partial_t\varphi\,\mathrm{d}x\,\mathrm{d}t,$$
$$\langle l, v\rangle := -\int_0^T \int_Q \kappa \nabla u_f \cdot \nabla v\,\mathrm{d}x\,\mathrm{d}t + \int_0^T \langle g, v|_{\partial B}\rangle\,\mathrm{d}t - \int_0^T \langle \psi, v|_{\partial\Omega}\rangle\,\mathrm{d}t.$$

Since H is a closed subspace of $H^{1,0}(B)$, it is a Hilbert space. $\Phi \subset H$ and for every $\varphi \in \Phi$, the linear form $v \to a(v,\varphi)$ is continuous on H.

Poincaré's inequality yields that $\left(\int_B |\nabla v|^2\,\mathrm{d}x\right)^{1/2}$ is an equivalent norm on $H^1_{\square}(B)$; thus there exists $\alpha > 0$ such that for all $\varphi \in \Phi$

$$\begin{aligned} a(\varphi,\varphi) &= \int_0^T \int_B \kappa |\nabla\varphi(x,t)|^2\,\mathrm{d}x\,\mathrm{d}t + \frac{1}{2}\int_{\Omega} |\varphi(0,x)|^2\,\mathrm{d}x \\ &\geq \int_0^T \int_B \kappa |\nabla\varphi|^2\,\mathrm{d}x\,\mathrm{d}t \geq \alpha \|\varphi\|_H^2 . \end{aligned}$$

Moreover, the continuity of the trace and lifting operators yields the existence of a constant C that does not depend on g, f, and ψ such that for all $v \in H$

$$\begin{aligned} \langle l, v\rangle &\leq C\left(\|g\|_{H^{-\frac{1}{2},0}(\partial B)} + \|f\|_{H^{\frac{1}{2},0}(\partial\Omega)} + \|\psi\|_{H^{-\frac{1}{2},0}(\partial\Omega)} \right) \|v\|_{H^{1,0}(B)} \\ &= C\left(\|g\|_{H_{\diamond}^{-\frac{1}{2},0}(\partial B)} + \|f\|_{H^{\frac{1}{2},0}(\partial\Omega)} + \|\psi\|_{H_{\diamond}^{-\frac{1}{2},0}(\partial\Omega)} \right) \|v\|_H , \end{aligned}$$

and thus $l \in H'$ with $\|l\|_{H'} \leq C\big(\|g\|_{H_{\diamond}^{-\frac{1}{2},0}(\partial B)} + \|f\|_{H^{\frac{1}{2},0}(\partial\Omega)} + \|\psi\|_{H_{\diamond}^{-\frac{1}{2},0}(\partial\Omega)} \big)$.

Now Lemma 2.4 gives existence of $\tilde{u} \in H$ that solves

$$\begin{aligned}(2.13)\quad &\int_0^T \int_B \kappa \nabla \tilde{u} \cdot \nabla \varphi \,\mathrm{d}x \,\mathrm{d}t - \int_0^T \int_\Omega \tilde{u}\, \partial_t \varphi \,\mathrm{d}x \,\mathrm{d}t \\ &\quad = -\int_0^T \int_Q \kappa \nabla u_f \cdot \nabla \varphi \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \langle g, \varphi|_{\partial B} \rangle \,\mathrm{d}t - \int_0^T \langle \psi, \varphi|_{\partial\Omega} \rangle \,\mathrm{d}t\end{aligned}$$

for all $\varphi \in \Phi$ and $\tilde{u}$ depends continuously on l (and therefore on g, f, and ψ).

We define $u \in H^{1,0}(B \setminus \partial\Omega)$ by setting $u|_\Omega := \tilde{u}|_\Omega$ and $u|_Q := \tilde{u}|_Q + u_f$. Then u solves (2.10) and there exist constants $C', C'' > 0$ such that

$$\begin{aligned}\|u\|_{H^{1,0}(B\setminus\partial\Omega)} &\le C' \left(\|\tilde{u}|_\Omega\|_{H^{1,0}(\Omega)} + \|\tilde{u}|_Q\|_{H^{1,0}(Q)} + \|u_f\|_{H^{1,0}(Q)} \right) \\ &\le C'' \left(\|\tilde{u}\|_H + \|u_f\|_{H^{1,0}(Q)} \right),\end{aligned}$$

and thus u depends continuously on g, f, and ψ.

Since $\int_\Omega \tilde{u}(x,t) \,\mathrm{d}x = 0$ for $t \in [0,T]$ a.e., the left side of (2.13) vanishes for all $\varphi(x,t) = c(t) \in \mathcal{D}([0,T[\times\overline{B})$. Due to our additional assumptions on g and ψ, the right side of (2.13) also vanishes for those φ. Thus (2.13) holds for all $\varphi \in \Phi$ and for all $\varphi(x,t) = c(t)$, which shows that (2.13) holds for all $\varphi \in \mathcal{D}([0,T[\times\overline{B})$, and we immediately obtain that u solves (2.5), (2.6), (2.9), and (2.11).

From Lemma 2.3 it follows that $\tilde{u}|_\Omega = u|_\Omega \in W$ and thus Green's formula holds. We obtain that for every $\varphi \in \mathcal{D}([0,T[\times B)$ with support in $[0,T[\times\Omega$

$$\begin{aligned}&-\int_\Omega u(0)\varphi(0) \,\mathrm{d}x \\ &= \int_0^T \int_\Omega u\, \partial_t \varphi \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \langle u', \varphi \rangle \,\mathrm{d}t \\ &= \int_0^T \int_\Omega \tilde{u}\, \partial_t \varphi \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \langle \kappa \partial_\nu u^-|_{\partial\Omega}, \varphi|_{\partial\Omega} \rangle \,\mathrm{d}t - \int_0^T \int_\Omega \kappa \nabla u \cdot \nabla \varphi \,\mathrm{d}x \,\mathrm{d}t \\ &= \int_0^T \int_B \kappa \nabla \tilde{u} \cdot \nabla \varphi \,\mathrm{d}x \,\mathrm{d}t - \int_0^T \int_\Omega \kappa \nabla u \cdot \nabla \varphi \,\mathrm{d}x \,\mathrm{d}t \\ &= 0,\end{aligned}$$

where we used that the right side of (2.13) vanishes for $\operatorname{supp}\varphi \in [0,T[\times\Omega$. As $\mathcal{D}(\Omega)$ is dense in $L^2(\Omega)$ this yields that $u|_\Omega(0) = 0$. ∎

We summarize the results of this section and give a useful variational formulation in Sobolev spaces.

THEOREM 2.6. *Let $g \in H_\diamond^{-\frac{1}{2},0}(\partial B)$, $f \in H^{\frac{1}{2},0}(\partial\Omega)$, and $\psi \in H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$, and let $u_f \in H^{1,0}(B \setminus \partial\Omega)$ be such that $u_f|_{\partial B} = 0$, $u_f|_{\partial\Omega} = f$, and $u_f|_\Omega = 0$.*

For $u \in H^{1,0}(B \setminus \partial\Omega)$ the following three problems are equivalent and possess the same unique solution. The solution depends continuously on g, f, and ψ and it fulfills $\int_\Omega u(x,t) \,\mathrm{d}x = 0$ for $t \in [0,T]$ a.e.

(a) *u solves*

$$(2.14)\qquad \partial_t u - \nabla \cdot (\kappa \nabla u) = 0 \text{ in } \Omega\times]0,T[,$$

$$(2.15)\qquad \nabla \cdot (\kappa \nabla u) = 0 \text{ in } Q\times]0,T[,$$

$$(2.16)\qquad [\kappa \partial_\nu u]_{\partial\Omega} = \psi,$$

$$[u]_{\partial\Omega} = f, \tag{2.17}$$

$$\kappa\partial_\nu u|_{\partial B} = g, \tag{2.18}$$

$$u(x,0) = 0 \text{ in } \Omega. \tag{2.19}$$

(b) $u|_\Omega \in W$, $u(x,0) = 0$ *in* Ω, *and* $\tilde{u} := u - u_f$ *solves*

$$\begin{aligned}\int_0^T \langle (\tilde{u}|_\Omega)', v|_\Omega\rangle \,\mathrm{d}t + \int_0^T \int_B \kappa\nabla\tilde{u}\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t \\ = \int_0^T \langle g, v|_{\partial B}\rangle \,\mathrm{d}t - \int_0^T \langle \psi, v|_{\partial\Omega}\rangle \,\mathrm{d}t - \int_0^T \int_Q \kappa\nabla u_f\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t\end{aligned} \tag{2.20}$$

for all $v \in H^{1,0}(B)$.

(c) $\tilde{u} := u - u_f$ *solves*

$$\begin{aligned}&\int_0^T \int_B \kappa\nabla\tilde{u}\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t - \int_0^T \langle (v|_\Omega)', \tilde{u}|_\Omega\rangle \,\mathrm{d}t \\ &= \int_0^T \langle g, v|_{\partial B}\rangle \,\mathrm{d}t - \int_0^T \langle \psi, v|_{\partial\Omega}\rangle \,\mathrm{d}t - \int_0^T \int_Q \kappa\nabla u_f\nabla v \,\mathrm{d}x\,\mathrm{d}t\end{aligned}$$

for all $v \in H^{1,0}(B)$ *with* $v|_\Omega \in W$ *and* $v(x,T) = 0$ *on* Ω.

Proof. We showed the unique solvability of the equations in (a) and the properties of the solution in Lemmas 2.3 and 2.5. Thus it remains only to prove the equivalence of (a), (b), and (c).

(a) $\Rightarrow$ (b). Note that $\tilde{u} \in H^{1,0}(B)$, $\kappa\partial_\nu u^-|_{\partial\Omega} \in H^{-\frac{1}{2},0}(\partial\Omega)$, and, by Lemma 2.3, $\tilde{u}|_\Omega = u|_\Omega \in W$.

It suffices to show (2.20) for $v \in \mathcal{D}(]0,T[\times\overline{B})$. Equations (2.14) and (2.15) imply that

$$\begin{aligned}0 &= \int_0^T \langle (\tilde{u}|_\Omega)', v|_\Omega\rangle \,\mathrm{d}t - \int_0^T \langle \nabla\cdot(\kappa\,\nabla u|_\Omega), v|_\Omega\rangle \,\mathrm{d}t \\ &= \int_0^T \langle (\tilde{u}|_\Omega)', v|_\Omega\rangle \,\mathrm{d}t - \int_0^T \langle \kappa\,\partial_\nu u^-|_{\partial\Omega}, v|_{\partial\Omega}\rangle \,\mathrm{d}t + \int_0^T \int_\Omega \kappa\,\nabla u\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t\end{aligned}$$

and

$$\begin{aligned}0 &= \int_0^T \langle \nabla\cdot(\kappa\,\nabla u|_Q), v|_Q\rangle \,\mathrm{d}t \\ &= -\int_0^T \langle \kappa\,\partial_\nu u^+|_{\partial\Omega}, v|_{\partial\Omega}\rangle \,\mathrm{d}t + \int_0^T \langle \kappa\,\partial_\nu u|_{\partial B}, v|_{\partial B}\rangle \,\mathrm{d}t \\ &\quad - \int_0^T \int_Q \kappa\,\nabla u\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Subtracting these two equations and using (2.16) and (2.18) give

$$\begin{aligned}0 = \int_0^T \langle (\tilde{u}|_\Omega)', v|_\Omega\rangle \,\mathrm{d}t + \int_0^T \langle \psi, v|_{\partial\Omega}\rangle \,\mathrm{d}t - \int_0^T \langle g, v|_{\partial B}\rangle \,\mathrm{d}t \\ + \int_0^T \int_\Omega \kappa\,\nabla u\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t + \int_0^T \int_Q \kappa\,\nabla u\cdot\nabla v \,\mathrm{d}x\,\mathrm{d}t.\end{aligned}$$

Now (2.20) follows from

$$\int_0^T \int_\Omega \kappa \nabla u \cdot \nabla v \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \int_Q \kappa \nabla u \cdot \nabla v \,\mathrm{d}x \,\mathrm{d}t$$
$$= \int_0^T \int_B \kappa \nabla \tilde{u} \cdot \nabla v \,\mathrm{d}x \,\mathrm{d}t + \int_0^T \int_Q \kappa \nabla u_f \cdot \nabla v \,\mathrm{d}x \,\mathrm{d}t.$$

(b) $\Rightarrow$ (c). This part of the proof follows from Green's formula on W.
(c) $\Rightarrow$ (a). This part of the proof was shown in the proof of Lemma 2.5. ∎

2.2. Boundary measurements and a reference problem. We assume that the inclusion not only has a higher heat capacity but also has a higher conductivity κ than the background. For simplicity we fix $\kappa = 1$ on Q and therefore require that $\kappa|_\Omega - 1 \in L^\infty_+(\Omega)$.

We introduce the measurement operator

$$\Lambda_1 : \; g \mapsto u_1|_{\partial B}, \text{ where } u_1 \text{ solves (2.1) with } \partial_\nu u_1|_{\partial B} = g, \; u_1|_\Omega = 0 \text{ at } t = 0.$$

Using the results from section 2.1 we know that Λ_1 is a continuous linear operator from $H_\diamond^{-\frac{1}{2},0}(\partial B)$ to $H^{\frac{1}{2},0}(\partial B)$.

To locate the inclusion Ω we compare Λ_1 with boundary measurements of a domain without inclusions, i.e., with the measurement operator

$$\Lambda_0 : \; g \mapsto u_0|_{\partial B}, \text{ where } \Delta u_0 = 0 \text{ on } B\times]0,T[\text{ and } \partial_\nu u_0|_{\partial B} = g.$$

The Lax–Milgram theorem shows that u_0 is uniquely determined up to addition of a spatially constant function $u(x,t) = c(t) \in L^2(0,T,\mathbb{R})$ and that Λ_0 is a continuous linear operator from $H_\diamond^{-\frac{1}{2},0}(\partial B)$ to $H_\diamond^{\frac{1}{2},0}(\partial B) := L^2(0,T,H_\diamond^{\frac{1}{2}}(\partial B))$, where the quotient space $H_\diamond^{\frac{1}{2}}(\partial B) := H^{\frac{1}{2}}(\partial B)/\mathbb{R}$ can be identified with the dual space of $H_\diamond^{-\frac{1}{2}}(\partial B)$ and $H_\diamond^{\frac{1}{2},0}(\partial B)$ with the dual space of $H_\diamond^{-\frac{1}{2},0}(\partial B)$.

Analogously we define quotient spaces on B, Q, and $\partial\Omega$ and note that in the case that $\partial\Omega$ is disconnected the quotient space $H_\diamond^{\frac{1}{2}}(\partial\Omega)$ is still obtained by factoring out the one-dimensional space of functions that are constant on $\partial\Omega$, and not the multidimensional space of functions that are constant on each connected component.

Mathematically the elements of the quotient spaces $H_\diamond^{r,0}$, $r \geq 0$, are equivalence classes; i.e., all functions that differ only by a spatially constant function are called equivalent and combined into one class. For the sake of readability we write an equivalence class as a function and keep in mind that it is a representant of its class. We also note that the space $H_\diamond^{-\frac{1}{2},0}$, which we defined earlier, is not a quotient space.

Without changing notation we use the canonical epimorphism to restrict Λ_1 to the spaces of the reference problem. Thus we will investigate the inverse problem of locating the inclusion Ω from knowledge of

$$\Lambda_0, \Lambda_1 : \; H_\diamond^{-\frac{1}{2},0}(\partial B) \to H_\diamond^{\frac{1}{2},0}(\partial B).$$

3. The inverse problem. We use the factorization method to reconstruct Ω from the boundary measurements. To this end we show that the difference of the measurement operators $\Lambda_0 - \Lambda_1$ can be factorized into the product

$$\Lambda_0 - \Lambda_1 = L(F_0 - F_1)L' \tag{3.1}$$

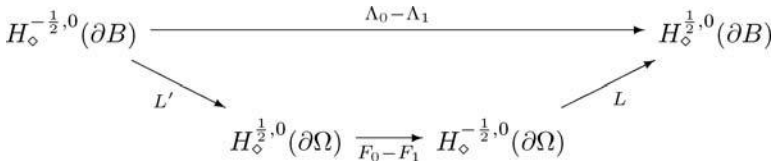


FIG. 3.1. *Factorization of* $\Lambda_0 - \Lambda_1$.

(cf. Figure 3.1), where the operator L corresponds to virtual measurements on the complement Q of the inclusion, and its range contains all information about Q and thus about the location of Ω.

Unlike previously known applications of the factorization method, the explicit time-dependence of the problem prevents us from calculating $\mathcal{R}(L)$ from the boundary measurements, but using a new approach we can show that the knowledge of $\Lambda_0 - \Lambda_1$ still suffices to determine Ω.

3.1. Factorization of the boundary measurements. We define a virtual measurement operator that corresponds to inducing a heat flux on the inclusion's boundary

$$L:\ H_\diamond^{-\frac{1}{2},0}(\partial\Omega) \to H_\diamond^{\frac{1}{2},0}(\partial B), \quad L\psi := v|_{\partial B},$$

where $v \in H_\diamond^{1,0}(Q)$ solves

$$\Delta v = 0 \text{ in } Q\times]0,T[, \quad \partial_\nu v = \begin{cases} -\psi & \text{on } \partial\Omega, \\ 0 & \text{on } \partial B. \end{cases} \tag{3.2}$$

We also need the two auxiliary operators

$$F_0:\ H^{\frac{1}{2},0}(\partial\Omega) \to H^{-\frac{1}{2},0}(\partial\Omega), \quad F_0\phi := \partial_\nu v_0^+|_{\partial\Omega},$$
$$F_1:\ H^{\frac{1}{2},0}(\partial\Omega) \to H^{-\frac{1}{2},0}(\partial\Omega), \quad F_1\phi := \partial_\nu v_1^+|_{\partial\Omega},$$

where $v_0, v_1 \in H^{1,0}(B \setminus \partial\Omega)$ solve

$$\begin{aligned} \Delta v_0 &= 0 \text{ in } (Q\cup\Omega)\times]0,T[, & [\partial_\nu v_0]_{\partial\Omega} &= 0, \\ \partial_\nu v_0|_{\partial B} &= 0, & [v_0]_{\partial\Omega} &= \phi, \end{aligned} \tag{3.3}$$

and

$$\begin{aligned} \Delta v_1 &= 0 \text{ in } Q\times]0,T[, & [\kappa\partial_\nu v_1]_{\partial\Omega} &= 0, \\ \partial_t v_1 - \nabla\cdot(\kappa\nabla v_1) &= 0 \text{ in } \Omega\times]0,T[, & [v_1]_{\partial\Omega} &= \phi, \\ v_1(x,0) &= 0 \text{ in } \Omega, & \partial_\nu v_1|_{\partial B} &= 0. \end{aligned} \tag{3.4}$$

Note that F_0 is well defined even though (3.3) determines v_0 only up to addition of a spatially constant function. Since the ranges of F_0 and F_1 are contained in $H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$ and their kernels contain $L^2(0,T,\mathbb{R})$, we will consider them as operators from

$$H_\diamond^{\frac{1}{2},0}(\partial\Omega) \text{ into } H_\diamond^{-\frac{1}{2},0}(\partial\Omega).$$

THEOREM 3.1. *The difference of the boundary measurements can be factorized into*

$$\Lambda_0 - \Lambda_1 = L(F_0 - F_1)L'.$$

The operators L *and* L' *are injective.*

Proof. For given $g \in H_\diamond^{-\frac{1}{2},0}(\partial B)$ let $w \in H_\diamond^{1,0}(Q)$ solve

$$\Delta w = 0 \text{ in } Q\times]0,T[, \text{ with } \partial_\nu w = \begin{cases} 0 & \text{on } \partial\Omega, \\ g & \text{on } \partial B. \end{cases}$$

Let $\psi \in H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$ and $v \in H_\diamond^{1,0}(Q)$ be the solution of (3.2) in the definition of $L\psi$. Then

$$\begin{aligned} \langle \psi, L'g\rangle = \langle g, L\psi\rangle = \langle \partial_\nu w|_{\partial B}, v|_{\partial B}\rangle = \int_0^T \int_Q \nabla w \cdot \nabla v \,\mathrm{d}x\,\mathrm{d}t \\ = \left\langle -\partial_\nu v^+|_{\partial\Omega}, w^+|_{\partial\Omega}\right\rangle = \left\langle \psi, w^+|_{\partial\Omega}\right\rangle, \end{aligned}$$

and thus $L'g = w^+|_{\partial\Omega}$.

Now let $v_0, v_1 \in H^{1,0}(B \setminus \partial\Omega)$ be the solutions of (3.3), resp., (3.4), from the definition of $F_0 w^+|_{\partial\Omega}$, resp., $F_1 w^+|_{\partial\Omega}$. We define $u_0, u_1 \in H^{1,0}(B \setminus \partial\Omega)$ by setting $u_i|_\Omega := -v_i|_\Omega$ and $u_i|_Q := w - v_i|_Q$, $i = 0, 1$. Then $u_0, u_1 \in H^{1,0}(B)$ and solve the equations in the definitions of $\Lambda_0 g$ and $\Lambda_1 g$. Thus

$$(\Lambda_0 - \Lambda_1)g = (u_0 - u_1)|_{\partial B} = -(v_0 - v_1)|_{\partial B}.$$

Since $\Delta(v_1 - v_0) = 0$ in $Q\times]0,T[$ and $\partial_\nu(v_1 - v_0)|_{\partial B} = 0$ we also have

$$L(\partial_\nu(v_0^+ - v_1^+)|_{\partial\Omega}) = -(v_0 - v_1)|_{\partial B},$$

and thus

$$(\Lambda_0 - \Lambda_1)g = L(\partial_\nu(v_0^+ - v_1^+)|_{\partial\Omega}) = L(F_0 - F_1)w^+|_{\partial\Omega} = L(F_0 - F_1)L'g.$$

To show injectivity of L' let $L'g = 0$ with some $g \in H_\diamond^{-\frac{1}{2},0}(\partial B)$. Then we obtain from the above characterization of L' a solution $w \in H^{1,0}(Q)$ of

$$\Delta w = 0 \quad \text{in } Q\times]0,T[, \qquad w^+|_{\partial\Omega} = 0, \quad \text{and} \quad \partial_\nu w = \begin{cases} 0 & \text{on } \partial\Omega, \\ g & \text{on } \partial B. \end{cases}$$

We set w to zero on $\Omega\times]0,T[$ and denote this continuation by $\tilde{w} \in H^{1,0}(B \setminus \partial\Omega)$. Then we have

$$\Delta\tilde{w} = 0 \quad \text{in } (B \setminus \partial\Omega)\times]0,T[, \qquad [\tilde{w}]_{\partial\Omega} = 0, \qquad [\kappa\partial_\nu\tilde{w}]_{\partial\Omega} = 0,$$

and thus $\tilde{w} \in H^{1,0}(B)$ and $\Delta\tilde{w} = 0$ in $B\times]0,T[$. Hence $\tilde{w}(\cdot,t)$ is analytic for $t \in]0,T[$ a.e. Since $\tilde{w}$ disappears on Ω and B is connected, we obtain that $w = \tilde{w} = 0$ in Q so that $g = 0$. Thus L' is injective.

The injectivity of L follows from the same arguments, when the function from the definition of L is set to zero in $(\mathbb{R}^n \setminus \overline{B})\times]0,T[$. Since Q is connected, $\mathbb{R}^n \setminus \overline{\Omega}$ is also connected. ∎

The injectivity of L and L' yields that they have dense ranges. The operator $F_0 - F_1$ satisfies a coerciveness condition; to show this we introduce the operators λ_1 and λ that correspond to measurements on the inclusion, resp., on its complement.

$$\begin{aligned} \lambda_1 :\ & H_\diamond^{-\frac{1}{2},0}(\partial\Omega) \to H^{\frac{1}{2},0}(\partial\Omega), \quad \lambda_1\psi := u_1^-|_{\partial\Omega}, \\ \lambda :\ & H_\diamond^{-\frac{1}{2},0}(\partial\Omega) \to H_\diamond^{\frac{1}{2},0}(\partial\Omega), \quad \lambda\psi := u^+|_{\partial\Omega}, \end{aligned}$$

where $u_1 \in W$ solves

$$\partial_t u_1 - \nabla \cdot (\kappa \nabla u_1) = 0 \text{ in } \Omega\times]0,T[, \quad \kappa \partial_\nu u_1^-|_{\partial\Omega} = \psi, \quad u_1(x,0) = 0 \text{ on } \Omega, \tag{3.5}$$

and $u \in H^{1,0}_\diamond(Q)$ solves

$$\Delta u = 0 \text{ in } Q\times]0,T[, \quad \partial_\nu u = \begin{cases} -\psi & \text{on } \partial\Omega, \\ 0 & \text{on } \partial B. \end{cases}$$

The unique solvability of (3.5) is shown in [8, Cor. 3.17] for general $\psi \in H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)$. In our case it can also be proven analogously to Lemmas 2.3 and 2.5.

Again we use the canonical epimorphism to restrict λ_1 to the same spaces as λ; i.e., from now on we consider it as an operator

$$\lambda_1 : \ H^{-\frac{1}{2},0}_\diamond(\partial\Omega) \to H^{\frac{1}{2},0}_\diamond(\partial\Omega).$$

LEMMA 3.2. (a) *For every $\psi \in H^{-\frac{1}{2},0}_\diamond(\partial\Omega)$ we have the identity*

$$\langle \psi, \lambda_1 \psi \rangle = \int_0^T \langle (u_1|_\Omega)', u_1|_\Omega \rangle \,\mathrm{d}t + \int_0^T \int_\Omega \kappa |\nabla u_1|^2 \,\mathrm{d}x\,\mathrm{d}t,$$

where $u_1 \in W$ is the solution of (3.5) *in the definition of λ_1.*

(b) *λ_1 is coercive with respect to $H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)$; i.e., there exists $c > 0$ such that*

$$\langle \psi, \lambda_1 \psi \rangle \geq c \, \|\psi\|^2_{H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)} \quad \textit{for all } \psi \in H^{-\frac{1}{2},0}_\diamond(\partial\Omega). \tag{3.6}$$

Proof. By setting it to zero on Q, every solution $u_1 \in W$ of (3.5) can be extended to a solution of (2.14), (2.15), and (2.19) in Theorem 2.6(a), with

$$[\kappa \partial_\nu u_1]_{\partial\Omega} = -\psi, \qquad [u_1]_{\partial\Omega} = -\lambda_1 \psi, \quad \text{and} \quad \kappa \partial_\nu u_1|_{\partial B} = 0.$$

It follows that

$$\int_\Omega u_1(x,t) \,\mathrm{d}x = 0 \quad \text{for } t \in [0,T] \text{ a.e.}, \tag{3.7}$$

and with $u_f \in H^{1,0}(B \setminus \partial\Omega)$ such that $u_f|_{\partial B} = 0$, $u_f|_{\partial\Omega} = -\lambda_1\psi$, and $u_f|_\Omega = 0$ we obtain from the variational formulation for $\tilde{u} := u_1 - u_f$ in Theorem 2.6(b)

$$\int_0^T \langle (\tilde{u}|_\Omega)', \tilde{u}|_\Omega \rangle \,\mathrm{d}t + \int_0^T \int_B \kappa |\nabla \tilde{u}|^2 \,\mathrm{d}x\,\mathrm{d}t = \int_0^T \langle \psi, \tilde{u}|_{\partial\Omega} \rangle \,\mathrm{d}t - \int_0^T \int_Q \kappa \nabla u_f \cdot \nabla \tilde{u} \,\mathrm{d}x\,\mathrm{d}t.$$

Using $\tilde{u}|_{\partial\Omega} = \lambda_1\psi$, $u_f|_\Omega = 0$, and $u_1|_Q = 0$ we conclude that

$$\langle \psi, \lambda_1 \psi \rangle = \int_0^T \langle (u_1|_\Omega)', u_1|_\Omega \rangle \,\mathrm{d}t + \int_0^T \int_\Omega \kappa |\nabla u_1|^2 \,\mathrm{d}x\,\mathrm{d}t,$$

and thus (a) holds.

Because of (3.7) Poincaré's inequality yields the existence of a $c' > 0$ such that

$$\langle \psi, \lambda_1 \psi \rangle \geq c' \, \|u_1\|^2_{H^{1,0}(\Omega)},$$

and so assertion (b) follows from the continuity of the Neumann trace in Theorem 2.1(c). □

LEMMA 3.3. *There exists $c' > 0$ such that*

$$\langle (F_0 - F_1)\phi, \phi \rangle \geq c' \, \|F_1 \phi\|^2_{H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)},$$

and F_1 is bijective with $F_1^{-1} = -\lambda - \lambda_1$.

Proof. For given $\phi \in H^{\frac{1}{2},0}(\partial\Omega)$ let $v_0, v_1 \in H^{1,0}(B \setminus \partial\Omega)$ be the solutions of (3.3) and (3.4) in the definition of F_0 and F_1, and let $v_\phi \in H^{1,0}(B \setminus \partial\Omega)$ be such that $v_\phi^+|_{\partial\Omega} = \phi$, $v_\phi|_{\partial B} = 0$, and $v_\phi|_\Omega = 0$.

Then $\tilde{v}_i := v_i - v_\phi$, $i = 0, 1$, solve

$$\int_0^T \int_B \nabla \tilde{v}_0 \cdot \nabla w \, \mathrm{d}x \, \mathrm{d}t = -\int_0^T \int_Q \nabla v_\phi \cdot \nabla w \, \mathrm{d}x \, \mathrm{d}t,$$

$$\int_0^T \langle (\tilde{v}_1|_\Omega)', w|_\Omega \rangle \, \mathrm{d}t + \int_0^T \int_B \kappa \nabla \tilde{v}_1 \cdot \nabla w \, \mathrm{d}x \, \mathrm{d}t = -\int_0^T \int_Q \nabla v_\phi \cdot \nabla w \, \mathrm{d}x \, \mathrm{d}t$$

for all $w \in H^{1,0}(B)$ (cf. Theorem 2.6 for the second equation). From the Lax–Milgram theorem it follows that for $t \in]0, T[$ a.e. $\tilde{v}_0(\cdot, t)$ minimizes the functional

$$w \mapsto \frac{1}{2} \int_B |\nabla w(x)|^2 \, \mathrm{d}x + \int_Q \nabla v_\phi(x,t) \cdot \nabla w(x) \, \mathrm{d}x$$

in $H^1(B)$ so that

$$\begin{aligned}
&\int_0^T \int_B |\nabla \tilde{v}_0|^2 \, \mathrm{d}x \, \mathrm{d}t \\
&= -2 \left(-\frac{1}{2} \int_0^T \int_B |\nabla \tilde{v}_0|^2 \, \mathrm{d}x \, \mathrm{d}t + \int_0^T \int_Q \nabla v_\phi \cdot \nabla \tilde{v}_0 \, \mathrm{d}x \, \mathrm{d}t \right) \\
&\geq -2 \left(\frac{1}{2} \int_0^T \int_B |\nabla \tilde{v}_1|^2 \, \mathrm{d}x \, \mathrm{d}t + \int_0^T \int_Q \nabla v_\phi \cdot \nabla \tilde{v}_1 \, \mathrm{d}x \, \mathrm{d}t \right) \\
&= -\int_0^T \int_B |\nabla \tilde{v}_1|^2 \, \mathrm{d}x \, \mathrm{d}t + 2 \int_0^T \int_B \kappa |\nabla \tilde{v}_1|^2 \, \mathrm{d}x \, \mathrm{d}t + 2 \int_0^T \langle (\tilde{v}_1|_\Omega)', \tilde{v}_1|_\Omega \rangle \, \mathrm{d}t
\end{aligned}$$

and thus

$$\begin{aligned}
&\langle (F_0 - F_1)\phi, \phi \rangle \\
&= \langle \partial_\nu v_0^+, \phi \rangle - \langle \partial_\nu v_1^+, \phi \rangle = \int_0^T \int_Q \nabla v_1 \cdot \nabla v_\phi \, \mathrm{d}x \, \mathrm{d}t - \int_0^T \int_Q \nabla v_0 \cdot \nabla v_\phi \, \mathrm{d}x \, \mathrm{d}t \\
&= \int_0^T \int_Q \nabla v_\phi \cdot \nabla \tilde{v}_1 \, \mathrm{d}x \, \mathrm{d}t - \int_0^T \int_Q \nabla v_\phi \cdot \nabla \tilde{v}_0 \, \mathrm{d}x \, \mathrm{d}t \\
&= \int_0^T \int_B |\nabla \tilde{v}_0|^2 \, \mathrm{d}x \, \mathrm{d}t - \int_0^T \int_B \kappa |\nabla \tilde{v}_1|^2 \, \mathrm{d}x \, \mathrm{d}t - \int_0^T \langle (\tilde{v}_1|_\Omega)', \tilde{v}_1|_\Omega \rangle \, \mathrm{d}t \\
&\geq \int_0^T \int_\Omega (\kappa - 1) |\nabla \tilde{v}_1|^2 \, \mathrm{d}x \, \mathrm{d}t + \int_0^T \langle (\tilde{v}_1|_\Omega)', \tilde{v}_1|_\Omega \rangle \, \mathrm{d}t.
\end{aligned}$$

Using $\kappa|_\Omega - 1 \in L^\infty_+(\Omega)$, $\int_0^T \langle (\tilde{v}_1|_\Omega)', \tilde{v}_1|_\Omega \rangle \, \mathrm{d}t \geq 0$, $\tilde{v}_1|_\Omega = v_1|_\Omega$, and Lemma 3.2(a) we conclude that there exists $c_\kappa > 0$ such that

$$\begin{aligned}\langle (F_0 - F_1)\phi, \phi \rangle &\geq c_\kappa \left(\int_0^T \int_\Omega \kappa |\nabla \tilde{v}_1|^2 + \int_0^T \langle (\tilde{v}_1|_\Omega)', \tilde{v}_1|_\Omega \rangle \, \mathrm{d}t \right) \\ &= c_\kappa \left\langle \left(\kappa \partial_\nu \tilde{v}_1^-|_{\partial\Omega}\right), \lambda_1 \left(\kappa \partial_\nu \tilde{v}_1^-|_{\partial\Omega}\right) \right\rangle = c_\kappa \left\langle \left(\partial_\nu v_1^+|_{\partial\Omega}\right), \lambda_1 \left(\partial_\nu v_1^+|_{\partial\Omega}\right) \right\rangle \\ &= c_\kappa \left\langle F_1\phi, \lambda_1 F_1 \phi \right\rangle,\end{aligned}$$

and so the first assertion follows from Lemma 3.2(b). To show surjectivity of F_1 let $\psi \in H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$ and denote by $u \in H^{1,0}(Q)$, $u_1 \in W$ the functions from the definition of $\lambda\psi$ and $\lambda_1\psi$.

Define $v_1 \in H^{1,0}(B \setminus \partial\Omega)$ by setting $v_1 := -u$ on Q and $v_1 := u_1$ on Ω. Then v_1 solves the equations in the definition of F_1 with $[v_1]_{\partial\Omega} = (-\lambda - \lambda_1)\psi$ (up to a spatially constant function) and thus $F_1(-\lambda - \lambda_1)\psi = \partial_\nu u_1^+|_{\partial\Omega} = \psi$.

It remains to show injectivity of F_1. To this end let $F_1\phi = 0$ and $v_1 \in H^{1,0}(B\setminus\partial\Omega)$ be the function from the definition of F_1. Then v_1 solves the Laplace equation on Q and the heat equation on Ω each with zero Neumann boundary values. Thus it vanishes on Ω and is spatially constant on Q, which implies that $\phi \in L^2(0,T,\mathbb{R})$. ∎

3.2. Range characterization. Lemma 3.3 implies that the symmetric part of $F_0 - F_1$ is positive and thus also the symmetric part of $\Lambda_0 - \Lambda_1$ is positive. Identifying Hilbert spaces with their duals, these operators have positive square roots, and their ranges can be related. The key to provide this relation is the following lemma that has been used by Brühl to extend the factorization method to the case of nonconstant conductivities in EIT [4, Satz 4.9]. We state it in the form in which it is called the "14th important property of Banach spaces" in [3] and give an elementary proof for the sake of completeness.

LEMMA 3.4. *Let X,Y be two Banach spaces, and let $A \in \mathcal{L}(X;Y)$ and $x' \in X'$. Then*

$$x' \in \mathcal{R}(A') \text{ if and only if } \exists C > 0 \,:\, |\langle x', x\rangle| \leq C \, \|Ax\| \text{ for all } x \in X.$$

Proof. If $x' \in \mathcal{R}(A')$ then there exists $y' \in Y'$ such that $x' = A'y'$. Thus

$$|\langle x', x\rangle| = |\langle y', Ax\rangle| \leq \|y'\| \, \|Ax\| \text{ for all } x \in X,$$

and the assertion holds with $C = \|y'\|$.

Now let $x' \in X'$ such that there exists $C > 0$ with $|\langle x', x\rangle| \leq C \, \|Ax\|$ for all $x \in X$. Define

$$f(z) := \langle x', x\rangle \text{ for every } z = Ax \in \mathcal{R}(A).$$

Then f is a well-defined, continuous, linear functional, with $\|f(z)\| \leq C \, \|z\|$. Using the Hahn–Banach theorem there exists $y' \in Y'$ with $y'|_{\mathcal{R}(A)} = f$. For all $x \in X$ we have

$$\langle A'y', x\rangle = \langle y', Ax\rangle = f(A(x)) = \langle x', x\rangle$$

and thus $x' = A'y' \in \mathcal{R}(A')$. ∎

We will make use of the following simple corollary.

COROLLARY 3.5. *Let H_i, $i = 1, 2$, be Hilbert spaces with norms $\|\cdot\|_i$, X be a third Hilbert space, and $A_i \in \mathcal{L}(X, H_i)$.*

If $\|A_1 x\|_1 \le \|A_2 x\|_2$ *for all* $x \in X$, *then* $\mathcal{R}(A_1^*) \subseteq \mathcal{R}(A_2^*)$.

Proof. Since $A_i' \iota_{H_1} = \iota_X A_i^*$, $i = 1, 2$, $y \in \mathcal{R}(A_1^*)$ implies $\iota_X y \in \mathcal{R}(A_1')$. Using Lemma 3.4 there exists $C > 0$ such that

$$|\langle \iota_X y, x\rangle| \le C \|A_1 x\|_1 \le C \|A_2 x\|_2 \text{ for all } x \in X,$$

and thus $\iota_X y \in \mathcal{R}(A_2')$, which implies $y \in \mathcal{R}(A_2^*)$. ∎

Note that in particular $A_1^* A_1 = A_2^* A_2$ implies $\mathcal{R}(A_1^*) = \mathcal{R}(A_2^*)$ (cf. [11]). Following the argument in [12] we can use Corollary 3.5 to characterize the range of the virtual measurement operator L by reformulating the symmetric part of (3.1) using adjoint operators.

We set

$$\begin{array}{rclcrcl} \Lambda & := & \Lambda_0 - \frac{1}{2}(\Lambda_1 + \Lambda_1'), & & \tilde{\Lambda} & = & \Lambda \iota_{H_\diamond^{1/2,0}(\partial B)}, \\ F & := & F_0 - \frac{1}{2}(F_1 + F_1'), & & \tilde{F} & = & \iota^{-1}_{H_\diamond^{1/2,0}(\partial\Omega)} F. \end{array}$$

LEMMA 3.6. *$\tilde{\Lambda}$ and $\tilde{F}$ are self-adjoint and positive operators and their square roots satisfy*

$$\mathcal{R}(\tilde{\Lambda}^{1/2}) = \mathcal{R}(L \iota_{H_\diamond^{1/2,0}(\partial\Omega)} \tilde{F}^{1/2}).$$

Proof. By construction $\tilde{\Lambda}$ and $\tilde{F}$ are self-adjoint and positive. From Theorem 3.1 it follows that

$$\begin{aligned} \tilde{\Lambda}^{1/2}\tilde{\Lambda}^{1/2} = \tilde{\Lambda} &= L \iota_{H_\diamond^{1/2,0}(\partial\Omega)} \tilde{F} L' \iota_{H_\diamond^{1/2,0}(\partial B)} \\ &= \left(L \iota_{H_\diamond^{1/2,0}(\partial\Omega)}\right) \tilde{F} \left(L \iota_{H_\diamond^{1/2,0}(\partial\Omega)}\right)^* \\ &= \left(L \iota_{H_\diamond^{1/2,0}(\partial\Omega)}\right) \tilde{F}^{1/2} \tilde{F}^{1/2} \left(L \iota_{H_\diamond^{1/2,0}(\partial\Omega)}\right)^*. \end{aligned}$$

The assertion now follows from Corollary 3.5. ∎

If F were coercive with respect to the space $H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$, we would obtain surjectivity of $\tilde{F}^{1/2}$ and thus the range characterization $\mathcal{R}(\tilde{\Lambda}^{1/2}) = \mathcal{R}(L)$ that was used in previous applications of the factorization method. In our situation we have only the weaker coercivity condition from Lemma 3.3. The next theorem shows that this weaker condition is still enough to guarantee that $\mathcal{R}(\tilde{F}^{1/2})$ contains all functions of a certain time regularity, which turns out to be sufficient for the method to work.

THEOREM 3.7.

$$\mathcal{R}(\tilde{\Lambda}^{1/2}) \subseteq \mathcal{R}(L) = L\left(H_\diamond^{-\frac{1}{2},0}(\partial\Omega)\right), \tag{3.8}$$

$$\mathcal{R}(\tilde{\Lambda}^{1/2}) \supseteq \qquad L\left(H^{\frac{1}{4}}(0,T,H_\diamond^{-\frac{1}{2}}(\partial\Omega))\right). \tag{3.9}$$

Proof. Equation (3.8) immediately follows from Lemma 3.6.

Denote by $j\colon H_\diamond^{-\frac{1}{2},0}(\partial\Omega) \hookrightarrow H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)$ the imbedding operator. Using Lemma 3.3 we have for all $\phi \in H_\diamond^{\frac{1}{2},0}(\partial\Omega)$

$$\left\|\tilde{F}^{1/2}\phi\right\|^2_{H_\diamond^{\frac{1}{2},0}(\partial\Omega)} = \left(\tilde{F}\phi, \phi\right)_{H_\diamond^{\frac{1}{2},0}(\partial\Omega)} \ge c' \|jF_1\phi\|^2_{H^{-\frac{1}{2},-\frac{1}{4}}(\partial\Omega)}.$$

Since $F_1^* j^* = \iota^{-1}_{H^{1/2,0}_\diamond(\partial\Omega)} F_1' j' \iota_{H^{-1/2,-1/4}(\partial\Omega)}$ we obtain from Corollary 3.5

$$\mathcal{R}\left(\tilde{F}^{1/2}\right) \supseteq \mathcal{R}(F_1^* j^*) = \mathcal{R}\left(\iota^{-1}_{H^{1/2,0}_\diamond(\partial\Omega)} F_1' j'\right)$$

and from Lemma 3.6

$$\mathcal{R}(\tilde{\Lambda}^{1/2}) = \mathcal{R}(L\iota_{H^{1/2,0}_\diamond(\partial\Omega)}\tilde{F}^{1/2}) \supseteq \mathcal{R}(LF_1' j').$$

Using Lemma 3.4 it is easily seen that

$$\mathcal{R}(j') = H^{\frac{1}{2},\frac{1}{4}}_\diamond(\partial\Omega) := (H^{\frac{1}{2},\frac{1}{4}}(\partial\Omega) + L^2(0,T,\mathbb{R}))/L^2(0,T,\mathbb{R}) \subset H^{\frac{1}{2},0}_\diamond(\partial\Omega).$$

(Note that by this definition $H^{\frac{1}{2},\frac{1}{4}}_\diamond(\partial\Omega)$ is isomorphic to $H^{\frac{1}{2},\frac{1}{4}}(\partial\Omega)/H^{\frac{1}{4}}(0,T,\mathbb{R})$.)

Using Lemma 3.3 we have $(F_1')^{-1} = -\lambda' - \lambda_1'$. Since $\lambda = \lambda'$ and $\mathcal{R}(\lambda_1') \subseteq H^{\frac{1}{2},\frac{1}{4}}_\diamond(\partial\Omega)$ (cf. [8]) it remains only to show that

$$\lambda\left(H^{\frac{1}{4}}(0,T,H_\diamond^{-\frac{1}{2}}(\partial\Omega))\right) \subseteq H^{\frac{1}{2},\frac{1}{4}}_\diamond(\partial\Omega). \tag{3.10}$$

To this end denote by $\overline{\lambda}:\ H_\diamond^{-\frac{1}{2}}(\partial\Omega) \to H^{\frac{1}{2}}_\diamond(\partial\Omega)$, $\overline{\psi} \mapsto \overline{u}^+|_{\partial\Omega}$, where $\overline{u} \in H^1_\diamond(Q)$ solves

$$\Delta\overline{u} = 0 \text{ in } Q, \quad \partial_\nu \overline{u} = \begin{cases} -\overline{\psi} & \text{on } \partial\Omega, \\ 0 & \text{on } \partial B. \end{cases}$$

Then for every $\psi \in H^1(0,T,H_\diamond^{-\frac{1}{2}}(\partial\Omega))$ and $\varphi \in \mathcal{D}(]0,T[)$

$$\begin{aligned}\int_0^T (-1)(\lambda\psi)\varphi'\,\mathrm{d}t &= \overline{\lambda}\left(\int_0^T (-1)\psi\varphi'(t)\,\mathrm{d}t\right) = \overline{\lambda}\left(\int_0^T \psi'\varphi(t)\,\mathrm{d}t\right) \\ &= \int_0^T (\lambda\psi')\varphi\,\mathrm{d}t \in H^{\frac{1}{2}}_\diamond(\partial\Omega).\end{aligned}$$

Thus $\lambda\psi \in H^1(0,T,H^{\frac{1}{2}}_\diamond(\partial\Omega))$ with $(\lambda\psi)' = \lambda(\psi')$, which shows that λ is a continuous operator not only from $L^2(0,T,H_\diamond^{-\frac{1}{2}}(\partial\Omega))$ to $L^2(0,T,H^{\frac{1}{2}}_\diamond(\partial\Omega))$ but also from $H^1(0,T,H_\diamond^{-\frac{1}{2}}(\partial\Omega))$ to $H^1(0,T,H^{\frac{1}{2}}_\diamond(\partial\Omega))$.

By interpolation (cf. [21]) λ is a continuous operator from

$$H^{\frac{1}{4}}(0,T,H_\diamond^{-\frac{1}{2}}(\partial\Omega)) \to H^{\frac{1}{4}}(0,T,H^{\frac{1}{2}}_\diamond(\partial\Omega)) \subset H^{\frac{1}{2},\frac{1}{4}}_\diamond(\partial\Omega).$$

Thus (3.10) holds and the assertion follows. ∎

3.3. Characterization of the inclusion. The composition of time integration and the (compact) imbedding $H^{\frac{1}{2}}_\diamond(\partial B) \hookrightarrow L^2_\diamond(\partial B) := L^2(\partial B)/\mathbb{R}$ defines the operator

$$I:\ H^{\frac{1}{2},0}_\diamond(\partial B) \to L^2_\diamond(\partial B), \quad u \mapsto \int_0^T u(\cdot,t)\,\mathrm{d}t.$$

Identifying $L^2_\diamond(\partial B)$ with its dual we have

$$I\tilde{\Lambda}I^* = I\Lambda I', \tag{3.11}$$

where $I' : L^2_\diamond(\partial B) \to H^{-\frac{1}{2},0}_\diamond(\partial B)$ is given by

$$I'v = w, \text{ with } w(\cdot,t) = v(\cdot) \text{ for } t \in [0,T] \text{ a.e.}$$

The operator $I\Lambda I'$ corresponds to measurements of applying temporal constant (and spatially square integrable) heat fluxes to a body and measuring time integrals of the resulting temperature on the boundary.

We use the same dipole functions as Brühl and Hanke used in [13] for the implementation of the factorization method in EIT. For a direction $d \in \mathbb{R}^n$, $|d| = 1$, and a point $z \in B$ let

$$D_{z,d}(x) := \frac{(z-x)\cdot d}{|z-x|^n}.$$

Then $D_{z,d}(x)$ is analytic and $\Delta D_{z,d}(x) = 0$ in $\mathbb{R}^n \setminus \{z\}$. Moreover, using a ball $B_\epsilon(z)$ centered at z with such small radius $\epsilon > 0$ such that $\overline{B_\epsilon(z)} \subset B$,

$$\int_{\partial B} \partial_\nu D_{z,d}(x)\,\mathrm{d}x = \int_{\partial B_\epsilon(z)} \partial_\nu D_{z,d}(x)\,\mathrm{d}x = 0,$$

so in particular $\partial_\nu D_{z,d} \in H^{-\frac{1}{2}}_\diamond(\partial B)$ and there exists $v_{z,d} \in H^1(B)$ that solves

$$\Delta v_{z,d} = 0 \text{ in } B \quad \text{ and } \quad \partial_\nu v_{z,d} = -\partial_\nu D_{z,d} \text{ on } \partial B.$$

Now $H_{z,d} := D_{z,d} + v_{z,d}$ is harmonic (and thus analytic) in $B \setminus \{z\}$ with $\partial_\nu H_{z,d}|_{\partial B} = 0$ but $H_{z,d} \notin L^2(B \setminus \{z\})$. The inclusion can now be characterized by the traces $h_{z,d} := H_{z,d}|_{\partial B} \in H^{\frac{1}{2}}_\diamond(\partial B)$ (again we use the same notation for the equivalence class of functions that are identical up to addition of constant functions as we used for the original function).

THEOREM 3.8. *For every $d \in \mathbb{R}^n$, $|d| = 1$, and $z \in B$*

$$z \in \Omega \text{ if and only if } h_{z,d} \in \mathcal{R}\left((I\Lambda I')^{1/2}\right).$$

Proof. From Corollary 3.5 and (3.11) it follows that $\mathcal{R}\left((I\Lambda I')^{1/2}\right) = \mathcal{R}(I\tilde{\Lambda}^{1/2})$ and consequently from Theorem 3.7 we obtain

$$\begin{aligned}
\mathcal{R}\left((I\Lambda I')^{1/2}\right) &\subseteq IL\left(H^{-\frac{1}{2},0}_\diamond(\partial\Omega)\right),\\
\mathcal{R}\left((I\Lambda I')^{1/2}\right) &\supseteq IL\left(H^{\frac{1}{4}}(0,T,H^{-1/2}_\diamond(\partial\Omega))\right).
\end{aligned}$$

First let $z \in \Omega$; then we define $w \in H^{1,0}_\diamond(Q)$ by $w(x,t) := H_{z,d}(x)/T$ for $t \in [0,T]$ a.e. Then $-\partial_\nu w^+|_{\partial\Omega} \in H^{\frac{1}{4}}(0,T,H^{-1/2}_\diamond(\partial\Omega))$ and w solves (3.2) in the definition of L, so

$$\begin{aligned}
h_{z,d} = Iw|_{\partial B} &= IL(-\partial_\nu w^+|_{\partial\Omega})\\
&\in IL\left(H^{\frac{1}{4}}(0,T,H^{-1/2}_\diamond(\partial\Omega))\right) \subseteq \mathcal{R}\left((I\Lambda I')^{1/2}\right).
\end{aligned}$$

To show the converse let $h_{z,d} \in \mathcal{R}\left((I\Lambda I')^{1/2}\right) \subseteq IL(H^{-\frac{1}{2},0}_\diamond(\partial\Omega))$. Then $h_{z,d}$ coincides with the integral of the trace of a solution of the Laplace equation on Q with vanishing

Neumann boundary values. Taking the integral of that solution we have that $h_{z,d} = w|_{\partial B}$, with some

$$w \in H^1_\diamond(Q) \text{ that solves } \Delta w = 0 \text{ on } Q,\ \partial_\nu w = 0 \text{ on } \partial B.$$

As $H_{z,d}$ and w are both harmonic on $Q \setminus \{z\}$ with the same Cauchy data on ∂B, they coincide near ∂B and thus by analytic continuation on $Q \setminus \{z\}$. If $z \notin \Omega$ this leads to the contradiction that $w \in L^2_\diamond(Q \setminus \{z\})$ but $H_{z,d} \notin L^2_\diamond(Q \setminus \{z\})$. ∎

By construction $I\Lambda I'$ is a compact and self-adjoint operator and from the factorization and the positiveness of F it follows that it is positive. Since $I\Lambda I'g = 0$ implies that $\langle FL'I'g, L'I'g\rangle = 0$ and thus $L'I'g = 0$, we also obtain injectivity of $I\Lambda I'$ from the injectivity of L' and I'. Hence there exists an orthonormal basis $(v_k)_{k\in\mathbb{N}}$ of eigenfunctions with associated positive eigenvalues $(\lambda_k)_{k\in N}$. Following [13] we use this spectral decomposition to reformulate Theorem 3.8 with the Picard criterion.

COROLLARY 3.9. *For every $d \in \mathbb{R}^n$, $|d| = 1$, and $z \in B$*

$$z \in \Omega \text{ if and only if } \sum_{k\in\mathbb{N}} \frac{1}{\lambda_k} \left(\int_{\partial B} h_{z,d} v_k \,\mathrm{d}x \right)^2 < \infty.$$

We remark that the results of this subsection remain valid with identical proofs when I is replaced by

$$I_S :\ H^{\frac{1}{2},0}_\diamond(\partial B) \to L^2_\diamond(S), \quad u \mapsto \int_0^T u|_S(\cdot,t)\,\mathrm{d}t,$$

where S is a relatively open subset of the boundary ∂B. Thus Ω is uniquely determined by $I_S \Lambda I_S'$, i.e., by measurements of applying (temporal constant) heat fluxes on a part of the boundary and measuring (time integrals of) the resulting temperature on the same part; cf., e.g., [14, 24, 25] for corresponding results in impedance tomography and the effect of partial boundary data on numerical reconstructions.

4. Numerics.

4.1. The direct problem. In this section we show how the direct problem can be solved numerically with a coupling of a finite element method and a boundary element method similar to [9]. We start by reformulating the direct problem.

4.1.1. Reformulation of the direct problem. Recall that λ was defined by

$$\lambda :\ H^{-\frac{1}{2},0}_\diamond(\partial\Omega) \to H^{\frac{1}{2},0}_\diamond(\partial\Omega), \quad \lambda\psi := \eta^+|_{\partial\Omega}\,,$$

where $\eta \in H^{1,0}_\diamond(Q)$ solves

$$\Delta\eta = 0 \text{ in } Q, \quad \partial_\nu \eta = \begin{cases} -\psi & \text{on } \partial\Omega\,, \\ 0 & \text{on } \partial B\,. \end{cases}$$

We use the same notation for the time-independent Neumann–Dirichlet operator

$$\lambda :\ H^{-\frac{1}{2}}_\diamond(\partial\Omega) \to H^{\frac{1}{2}}_\diamond(\partial\Omega).$$

Note that λ is linear, continuous, and coercive, i.e., $\langle \psi, \lambda\psi\rangle \geq c\|\psi\|^2_{H^{-\frac{1}{2}}_\diamond(\partial\Omega)}$.

For the rest of this section we assume that $g \in H_\diamond^{-\frac{1}{2},0}(\partial B)$ and $\xi = \xi(g) \in H_\diamond^{1,0}(Q)$ solves

$$\Delta\xi = 0 \text{ in } Q, \quad \partial_\nu \xi = \begin{cases} 0 & \text{on } \partial\Omega\,, \\ g & \text{on } \partial B\,. \end{cases}$$

THEOREM 4.1. *If $u \in H^{1,0}(B)$ solves (2.5)–(2.8), (2.11), and (2.12), then $v := u|_\Omega$ and $\phi := -\kappa\partial_\nu u^-|_{\partial\Omega}$ satisfy*

$$\partial_t v - \nabla\cdot(\kappa\,\nabla v) = 0 \text{ in } \Omega\times]0,T[, \tag{4.1}$$

$$v^-|_{\partial\Omega} - \lambda\phi = \xi^+|_{\partial\Omega} \text{ in } H_\diamond^{\frac{1}{2},0}(\partial\Omega), \tag{4.2}$$

$$v(x,0) = 0 \text{ in } \Omega\,. \tag{4.3}$$

On the other hand, if $(v,\phi) \in H^{1,0}(\Omega)\times H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$ solves (4.1)–(4.3) and

$$\kappa\partial_\nu v^-|_{\partial\Omega} = -\phi, \tag{4.4}$$

then there exists $u \in H^{1,0}(B)$ that solves (2.5)–(2.8), (2.11), (2.12), and $v = u|_\Omega$. Moreover $u|_Q \in H^{1,0}(Q)$ is the representant of $\xi+\eta \in H_\diamond^{1,0}(Q)$ with $\int_{\partial\Omega} u^+|_{\partial\Omega}\,\mathrm{d}s = \int_{\partial\Omega} u^-|_{\partial\Omega}\,\mathrm{d}s$, where η is as in the definition of $\lambda\phi$.

Proof. The proof immediately follows from the definitions of ξ and λ. □

THEOREM 4.2. *The following problems are equivalent:*

(a) *$(u,\phi) \in H^{1,0}(\Omega)\times H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$ solves (4.1)–(4.4).*

(b) *$(u,\phi) \in W \times H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$, $u(x,0) = 0$ in Ω, and (u,ϕ) solves*

$$\begin{aligned} &\int_0^T \langle u', v\rangle\,\mathrm{d}t + \int_0^T\int_\Omega \kappa\nabla u\cdot\nabla v\,\mathrm{d}x\,\mathrm{d}t + \int_0^T \langle \phi, v^-|_{\partial\Omega}\rangle\,\mathrm{d}t \\ &- \int_0^T \langle\tilde\psi, \lambda\phi\rangle\,\mathrm{d}t + \int_0^T \langle\tilde\psi, u^-|_{\partial\Omega}\rangle\,\mathrm{d}t = \int_0^T \langle\tilde\psi, \xi^+|_{\partial\Omega}\rangle\,\mathrm{d}t \end{aligned} \tag{4.5}$$

for all $v \in H^{1,0}(\Omega)$ and for all $\tilde\psi \in H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$.

(c) *$(u,\phi) \in W \times H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$, $u(x,0) = 0$ in Ω, and (u,ϕ) solves*

$$\begin{aligned} &\langle u'(t), v\rangle + \int_\Omega \kappa\nabla u(t)\cdot\nabla v\,\mathrm{d}x + \langle\phi(t), v^-|_{\partial\Omega}\rangle - \langle\tilde\psi, \lambda\phi(t)\rangle + \langle\tilde\psi, u(t)^-|_{\partial\Omega}\rangle \\ &\qquad = \langle\tilde\psi, \xi(t)^+|_{\partial\Omega}\rangle \end{aligned} \tag{4.6}$$

for $t \in [0,T]$ a.e. and for all $v \in H^1(\Omega)$ and for all $\tilde\psi \in H_\diamond^{-\frac{1}{2}}(\partial\Omega)$.

Proof. (a)⇔(b) can be shown analogously to the proof of Theorem 2.6.

To show (b)⇔(c), note that (4.6) is fulfilled for $t \in [0,T]$ a.e. if and only if it is fulfilled in the sense of $L^2([0,T])$. The equivalence then follows from the fact that $L^2([0,T])\otimes H^1(\Omega)$, resp., $L^2([0,T])\otimes H_\diamond^{-\frac{1}{2}}(\partial\Omega)$, are dense in $H^{1,0}(\Omega)$, resp., $H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$. □

4.1.2. Implementation and convergence analysis of the reformulated problem. Let $\{H_h,\, h>0\}$ and $\{B_h,\, h>0\}$ be families of finite dimensional subspaces of $H^1(\Omega)$ and $H_\diamond^{-\frac{1}{2}}(\partial\Omega)$, respectively. Accordingly the family of L^2-projections $P_h : H^1(\Omega) \to H_h$ is defined by $\int_\Omega P_h v w_h\,\mathrm{d}x = \int_\Omega v w_h\,\mathrm{d}x$ for all $w_h \in H_h$. We assume

that P_h satisfies the following estimate: There exists a constant $\gamma > 0$, independent of h, such that

$$\sup_{v \in H^1(\Omega)} \frac{\|P_h v\|_{H^1(\Omega)}}{\|v\|_{H^1(\Omega)}} \leq \gamma \quad \text{for all } h > 0\,. \tag{4.7}$$

For example, let $\mathcal{T}$ be a regular triangulation of Ω with generic mesh spacing h and H_h be a space of piecewise linear polynomials on $\mathcal{T}$. Then following [23] the operator P_h fulfills (4.7).

We consider the following Galerkin scheme.

Find $u_h : [0,T] \to H_h$, $\phi_h : [0,T] \to B_h$ such that

$$\begin{aligned} \langle u_h', v_h \rangle + \int_\Omega \kappa \nabla u_h \cdot \nabla v_h \,\mathrm{d}x + \langle \phi_h, v_h^-|_{\partial\Omega} \rangle \\ - \langle \psi_h, \lambda \phi_h \rangle + \langle \psi_h, u_h^-|_{\partial\Omega} \rangle \quad = \quad \langle \psi_h, \xi^+|_{\partial\Omega} \rangle \end{aligned} \tag{4.8}$$

for all $(v_h, \psi_h) \in H_h \times B_h$, $t \in [0,T]$ a.e., and $u_h(0) = 0$.

LEMMA 4.3. *For every $h > 0$ the Galerkin scheme* (4.8) *has a unique solution in $H_h^T \times B_h^T$, where*

$$\begin{aligned} H_h^T &:= \left\{ u \in L^2(0,T,H_h) : u' \in L^2(0,T,H_h), u(x,0) = 0 \right\} \subset W, \\ B_h^T &:= L^2(0,T,B_h) \subset H_\diamond^{-\frac{1}{2},0}(\partial\Omega). \end{aligned}$$

Proof. Let $(w_k)_{k=1}^{n_h}$ be a basis of H_h which is orthonormal with respect to the $L^2(\Omega)$ scalar product and $(\psi_j)_{j=1}^{m_h}$ be a basis of B_h. Moreover, if we write $u_h(x,t) = \sum_{k=1}^{n_h} \alpha_k(t) w_k(x)$ and $\phi_h(x,t) = \sum_{j=1}^{m_h} \beta_j(t) \psi_j(x)$, then (4.8) is equivalent to

$$\partial_t \alpha(t) + K\alpha(t) + B\beta(t) = 0\,, \quad \alpha(0) = 0\,, \tag{4.9}$$

and

$$D\beta(t) - B^T \alpha(t) = d(t), \tag{4.10}$$

where $\alpha = (\alpha_1, \ldots, \alpha_{n_h})^T$ and $\beta = (\beta_1, \ldots, \beta_{m_h})^T$. Since λ is linear and coercive we can solve (4.10) for β in terms of α and substitute into (4.9) to obtain a system of ODEs for α. According to standard existence theory for ODEs, there exists a unique absolutely continuous solution α. ∎

LEMMA 4.4. *Assume that $(w_h, \zeta_h) \in H_h^T \times B_h^T$, $\zeta \in H_\diamond^{-\frac{1}{2},0}(\partial\Omega)$, $w \in W$ with $w(x,0) = 0$ in Ω. Moreover assume that the following equation is fulfilled:*

$$\begin{aligned} &\langle w_h', v_h \rangle + \int_\Omega \kappa \nabla w_h \cdot \nabla v_h \,\mathrm{d}x + \langle \zeta_h, v_h^-|_{\partial\Omega} \rangle - \langle \psi_h, \lambda \zeta_h \rangle + \langle \psi_h, w_h^-|_{\partial\Omega} \rangle \\ &= \langle w', v_h \rangle + \int_\Omega \kappa \nabla w \cdot \nabla v_h \,\mathrm{d}x + \langle \zeta, v_h^-|_{\partial\Omega} \rangle - \langle \psi_h, \lambda \zeta \rangle + \langle \psi_h, w^-|_{\partial\Omega} \rangle \end{aligned} \tag{4.11}$$

for $t \in [0,T]$ a.e. and for all $(v_h, \psi_h) \in H_h \times B_h$. Then there exists $c > 0$ independent of h such that

$$\|w_h\|_W + \|\zeta_h\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \leq c \left(\|w\|_W + \|\zeta\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \right).$$

Proof. 1. We use $(v_h, \psi_h) = (w_h, -\zeta_h)$ in (4.11) and obtain for $t \in [0,T]$ a.e.

$$\begin{aligned}
&\langle w_h', w_h\rangle + \int_\Omega \kappa|\nabla w_h|^2 \,\mathrm{d}x + \langle \zeta_h, \lambda\zeta_h\rangle \\
&= \langle w', w_h\rangle + \int_\Omega \kappa\nabla w\cdot\nabla w_h \,\mathrm{d}x + \langle \zeta, w_h^-|_{\partial\Omega}\rangle + \langle \zeta_h, \lambda\zeta\rangle - \langle \zeta_h, w^-|_{\partial\Omega}\rangle \\
&\le c_1\left(\|w'\|_{H^1(\Omega)'} + \|w\|_{H^1(\Omega)} + \|\zeta\|_{H_\diamond^{-\frac12}(\partial\Omega)}\right)\|w_h\|_{H^1(\Omega)} \\
&\quad + c_2\left(\|\zeta\|_{H_\diamond^{-\frac12}(\partial\Omega)} + \|w\|_{H^1(\Omega)}\right)\|\zeta_h\|_{H_\diamond^{-\frac12}(\partial\Omega)},
\end{aligned}$$

where c_i, $i = 1,\dots,8$, are not depending on h. Note that $\|\zeta\|_{H^{-\frac12}(\partial\Omega)} = \|\zeta\|_{H_\diamond^{-\frac12}(\partial\Omega)}$ and $\|\zeta_h\|_{H^{-\frac12}(\partial\Omega)} = \|\zeta_h\|_{H_\diamond^{-\frac12}(\partial\Omega)}$. Now we integrate the left- and the right-hand sides of this inequality from 0 to t and get for $t \in [0,T]$ a.e.

$$\begin{aligned}
&\frac12\|w_h(t)\|_{L^2(\Omega)}^2 + \int_0^t\int_\Omega \kappa|\nabla w_h|^2 \,\mathrm{d}x\,\mathrm{d}t + \int_0^t \langle \zeta_h, \lambda\zeta_h\rangle \,\mathrm{d}t \\
&\le c_3\left(\|w\|_W + \|\zeta\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right)\left(\|w_h\|_{H^{1,0}(\Omega)} + \|\zeta_h\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right). \qquad (4.12)
\end{aligned}$$

Again integrating both sides of this inequality from 0 to T yields

$$\begin{aligned}
&\|w_h\|_{L^2(0,T,L^2(\Omega))}^2 \\
&\le c_4\left(\|w\|_W + \|\zeta\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right)\left(\|w_h\|_{H^{1,0}(\Omega)} + \|\zeta_h\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right). \qquad (4.13)
\end{aligned}$$

2. Since λ is coercive, using (4.12) and (4.13) we get

$$\begin{aligned}
&\|w_h\|_{H^{1,0}(\Omega)}^2 + \|\zeta_h\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}^2 \\
&\le c_5\left(\|w\|_W + \|\zeta\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right)\left(\|w_h\|_{H^{1,0}(\Omega)} + \|\zeta_h\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right).
\end{aligned}$$

Therefore, we have

$$\|w_h\|_{H^{1,0}(\Omega)} + \|\zeta_h\|_{H_\diamond^{-\frac12,0}(\partial\Omega)} \le c_6\left(\|w\|_W + \|\zeta\|_{H_\diamond^{-\frac12,0}(\partial\Omega)}\right). \qquad (4.14)$$

3. Since $w_h' \in H_h$ for $t \in [0,T]$ a.e., using the L^2-projection P_h we have for $t \in [0,T]$ a.e.

$$\|w_h'\|_{H^1(\Omega)'} = \sup_{w\in H^1(\Omega)} \frac{\langle w_h', w\rangle}{\|w\|_{H^1(\Omega)}} = \sup_{w\in H^1(\Omega)} \frac{\langle w_h', P_h w\rangle}{\|w\|_{H^1(\Omega)}}. \qquad (4.15)$$

Now we use $(v_h, \psi_h) = (P_h w, 0)$ in (4.11) and obtain for $t \in [0,T]$ a.e.

$$\begin{aligned}
\langle w_h', P_h w\rangle &= -\int_\Omega \kappa\nabla w_h\cdot\nabla P_h w \,\mathrm{d}x - \langle \zeta_h, P_h w^-|_{\partial\Omega}\rangle + \langle w', P_h w\rangle \\
&\quad + \int_\Omega \kappa\nabla w\cdot\nabla P_h w \,\mathrm{d}x + \langle \zeta, P_h w^-|_{\partial\Omega}\rangle \\
&\le c_7\|P_h w\|_{H^1(\Omega)}\left(\|w_h\|_{H^1(\Omega)} + \|\zeta_h\|_{H_\diamond^{-\frac12}(\partial\Omega)}\right. \\
&\quad \left. + \|w\|_W + \|\zeta\|_{H_\diamond^{-\frac12}(\partial\Omega)}\right).
\end{aligned}$$

Squaring and then integrating (4.15) from 0 to T and combining it with the inequality above, (4.7), and (4.14), we obtain

$$\|w_h'\|_{L^2(0,T,H^1(\Omega)')} \le c_8 \left(\|w\|_W + \|\zeta\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \right). \qquad \square$$

In particular, Lemma 4.4 holds for (u_h, ϕ_h) and (u, ϕ).

We proof a variant of Céa's lemma for this time-dependent problem.

THEOREM 4.5. *Assume that (u, ϕ) and (u_h, ϕ_h) are solutions of (4.5) with $u(x, 0) = 0$ and of the Galerkin scheme, respectively. Then there exists $c > 0$ such that*

$$\begin{aligned} &\|u - u_h\|_W + \|\phi - \phi_h\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \\ &\quad \le c \inf \left\{ \|u - z_h\|_W + \|\phi - \chi_h\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} : z_h \in H_h^T,\, \chi_h \in B_h^T \right\}. \end{aligned}$$

Proof. (u, ϕ) and (u_h, ϕ_h) obviously satisfy (4.11).

Let $(z_h, \chi_h) \in H_h^T \times B_h^T$. We set $(e_1, e_2) := (u_h, \phi_h) - (z_h, \chi_h)$ and $(\epsilon_1, \epsilon_2) := (u, \phi) - (z_h, \chi_h)$. Then (4.11) yields

$$\begin{aligned} &\langle e_1', v_h \rangle + \int_\Omega \kappa \nabla e_1 \cdot \nabla v_h \, dx + \langle e_2, v_h^- |_{\partial\Omega} \rangle - \langle \psi_h, \lambda e_2 \rangle + \langle \psi_h, e_1^- |_{\partial\Omega} \rangle \\ &= \langle \epsilon_1', v_h \rangle + \int_\Omega \kappa \nabla \epsilon_1 \cdot \nabla v_h \, dx + \langle \epsilon_2, v_h^- |_{\partial\Omega} \rangle - \langle \psi_h, \lambda \epsilon_2 \rangle + \langle \psi_h, \epsilon_1^- |_{\partial\Omega} \rangle \end{aligned}$$

for all $(v_h, \psi_h) \in H_h \times B_h$ and for $t \in [0, T]$ a.e. Lemma 4.4 shows that

$$\|e_1\|_W + \|e_2\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \le c_1 \left(\|\epsilon_1\|_W + \|\epsilon_2\|_{H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \right).$$

Hence

$$\begin{aligned} &\|(u_h, \phi_h) - (u, \phi)\|_{W \times H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} \\ &\quad \le c_2 \inf \left\{ \|(u, \phi) - (z_h, \chi_h)\|_{W \times H_\diamond^{-\frac{1}{2},0}(\partial\Omega)} : z_h \in H_h^T,\, \chi_h \in B_h^T \right\}, \end{aligned}$$

which is the desired estimate. $\square$

For our numerical examples we choose the same subspaces as in [9] and [23]; i.e., H_h consists of continuous functions, which are piecewise linear on a finite element grid, and B_h consists of piecewise constant functions. Equations (4.9) and (4.10) are solved numerically by a Crank–Nicolson method; i.e., we solve in each time-step the linear system of equations

$$\begin{bmatrix} I + \frac{\Delta t}{2} K & \frac{\Delta t}{2} B \\ -\frac{1}{2} B^T & \frac{1}{2} D \end{bmatrix} \begin{bmatrix} \alpha(t + \Delta t) \\ \beta(t + \Delta t) \end{bmatrix} = \begin{bmatrix} I - \frac{\Delta t}{2} K & -\frac{\Delta t}{2} B \\ \frac{1}{2} B^T & -\frac{1}{2} D \end{bmatrix} \begin{bmatrix} \alpha(t) \\ \beta(t) \end{bmatrix} + \begin{bmatrix} 0 \\ d(t) \end{bmatrix},$$

with $\alpha(0) = 0$ and $\beta(0) = D^{-1} d(0)$.

For the calculation of ξ and λ we use a boundary element method.

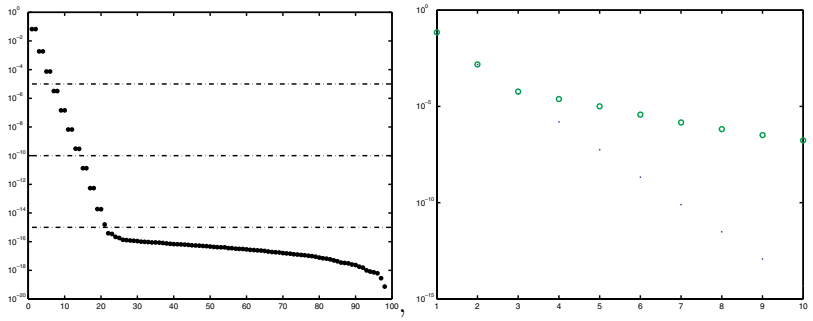


FIG. 4.1. *Left: The exact eigenvalues $\tilde{\lambda}_k$ from A. Right: Exact $(\cdot)$ and perturbed $(\circ)$ averaged eigenvalues.*

4.2. Implementation of the inverse problem. In this subsection we demonstrate how the factorization method can be used to solve the inverse problem, i.e., to locate the inclusion Ω from the knowledge of $I\Lambda I'$. We assume that we are given a finite dimensional approximation of $I\Lambda I'$ and thus a matrix $A \in \mathbb{R}^{m\times m}$. Let $(v_k)_{k\in\mathbb{N}}$, resp., $(\tilde{v}_k)_{k=1}^m$, be the eigenfunctions of $I\Lambda I'$, resp., A, with associated eigenvalues $(\lambda_k)_{k\in\mathbb{N}}$, resp., $(\tilde{\lambda}_k)_{k\in\mathbb{N}}$. Since $I\Lambda I'$ is self-adjoint and positive, the matrix A is symmetric and positive, too.

According to Corollary 3.9 a point $z \in B$ belongs to the inclusion Ω if and only if the infinite series

$$\sum_{k\in\mathbb{N}} \frac{(h_{z,d}, v_k)^2_{L^2(\partial B)}}{\lambda_k}$$

converges. For the numerical realization we have to decide about the convergence of this series from the knowledge of the finite sum

$$\sum_{k=1}^{m} \frac{(h_{z,d}, \tilde{v}_k)^2_{L^2(\partial B)}}{\tilde{\lambda}_k}\,.$$

For that we carry forward the ideas from [4]. Numerical examples show that the numerator and the denominator of the above series decay more or less exponentially and that every two eigenvalues have approximately the same value; cf. the left picture of Figure 4.1. Motivated by the examples and the method from [4], we compare the slopes of the least squares fitting straight lines of $h_1(k) = \log\left(\sqrt{\tilde{\lambda}_{2k-1}\tilde{\lambda}_{2k}}\right)$ and of

$$h_2(k) = \log\left(\frac{1}{2}\left((h_{z,d}, \tilde{v}_{2k-1})^2_{L^2(\partial B)} + (h_{z,d}, \tilde{v}_{2k})^2_{L^2(\partial B)}\right)\right)\,, \quad k = 1, \ldots, r\,.$$

We mark a sampling point z as inside the inclusion if h_1 decays slower than h_2. On the right side of Figure 4.2 the algorithm is demonstrated for two test points. If we apply this method to a large number of points, the black area on the left side of Figure 4.2 illustrates the reconstruction of the inclusion (dashed curve).

The number of the eigenvalues and Fourier coefficients which are used in the reconstruction procedure depends on the quality of the data. If A is known up to a perturbation of $\delta > 0$ (with respect to the spectral norm), then we trust in those

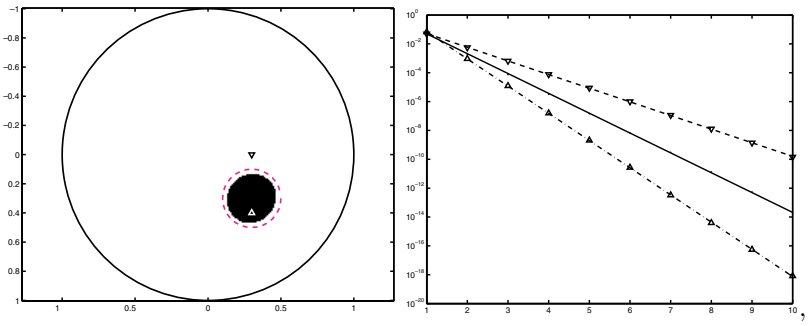


FIG. 4.2. *Least squares fitting straight lines of $h_2(k)$ for a point inside ($\triangle$) and outside ($\triangledown$) the inclusion compared with the least squares fitting straight line of $h_1(k)$ ($\cdot$).*

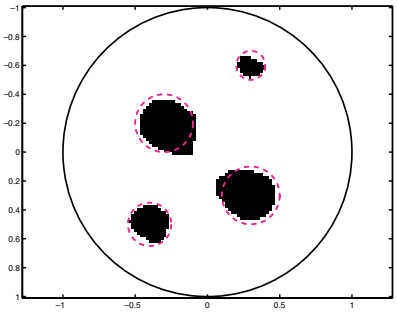


FIG. 4.3. *Reconstruction of four inclusions (dashed curves).*

eigenvalues which are larger than δ. On the left side of Figure 4.1, δ corresponds to the computational accuracy. The right side of Figure 4.1 shows the effect of a relative noise of 0.1% on the eigenvalues, and thus $\delta = 0.1\% \cdot \tilde{\lambda}_1$. The first three averaged pairs of the perturbed eigenvalues have nearly the exact values and they show the same exponential decay rates.

4.3. Numerical examples. To test this reconstruction algorithm we simulate the direct problem to produce the data. For this purpose we calculate the Dirichlet boundary data $f_k = I\Lambda I' g_k$, where $(g_k)_{k=1}^m$ are orthogonal input patterns. In the first examples this data was used for inversion. In the final example this data was perturbed with noise.

We restrict our attention to the case where $\kappa(x) = 2$ for $x \in \Omega$ and B is the unit disc in $\mathbb{R}^2$. For this case the function $h_{z,d}$ is known explicitly:

$$h_{z,d}(x) = \frac{1}{\pi} \frac{(z-x)\cdot d}{|z-x|^2}\,.$$

First we aim to reconstruct a single circle in the interior of B. The result is shown in the left picture of Figure 4.2. The location of Ω is detected but the size is underestimated.

In the second example four inclusions of different size should be located. In Figure 4.3 we demonstrate the possibility of the method to reconstruct nonconnected inclusions. The position and the different size of each are detected.

Our next example is to detect a nonconvex moon-like inclusion; cf. Figure 4.4. The top left picture shows the reconstruction with exact data. The shape of the moon

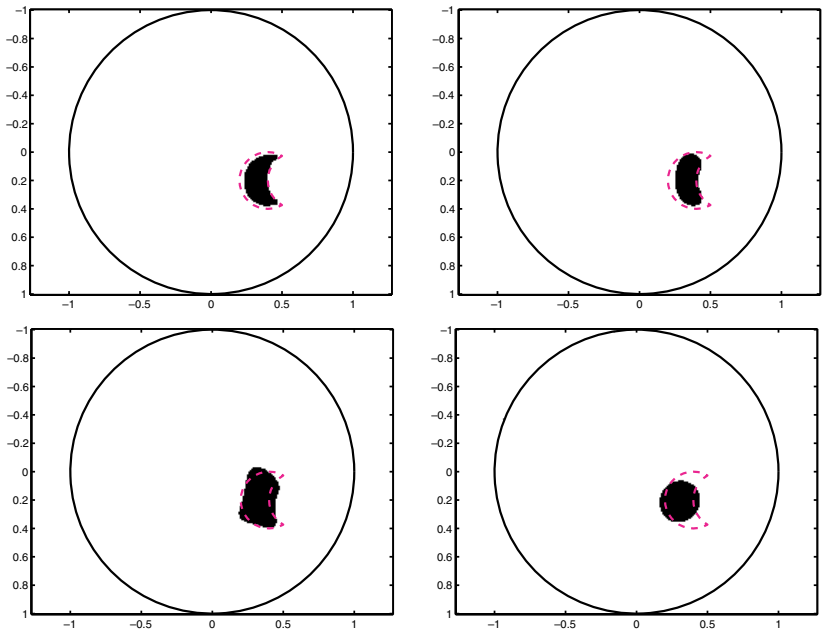


FIG. 4.4. *Reconstruction of a nonconvex inclusion (dashed curve). Top left: With exact data, and with perturbed data. Top right:* 0.05% *noise. Bottom left:* 0.1% *noise. Bottom right:* 1% *noise.*

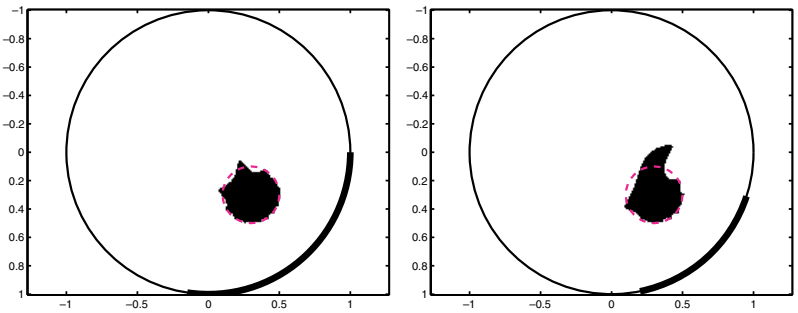


FIG. 4.5. *Reconstruction of an inclusion (dashed curve) by using partial boundary measurements (bold boundary).*

is recovered but the size is underestimated. Next we show the influence of noise on the reconstructions. By adding 0.05%, 0.1%, resp., 1%, noise the position of the inclusion is found, but the quality decreases with increasing noise level; cf. the top right and bottom pictures in Figure 4.4.

The last example shows the reconstruction of a single circle by partial boundary measurements (cf. our remark at the end of section 3). The location of the inclusion is detected and the shape next to the measuring boundary is recovered; see Figure 4.5.

Acknowledgments. We thank Prof. Martin Hanke-Bourgeois and Prof. Olaf Steinbach for stimulating discussion and constructive remarks.

REFERENCES

[1] H. Ammari, R. Griesmaier, and M. Hanke, *Identification of small inhomogeneities: asymptotic factorization*, Math. Comp., to appear.

[2] G. Bal, *Reconstructions in impedance and optical tomography with singular interfaces*, Inverse Problems, 21 (2005), pp. 113–131.

[3] N. Bourbaki, *Topological Vector Spaces, Chapters* 1–5, Elem. Math. (Berlin), Springer-Verlag, Berlin, 1987.

[4] M. Brühl *Gebietserkennung in der elektrischen Impedanztomographie*, Dissertation, Universität Karlsruhe, Karlsruhe, Germany, 1999.

[5] M. Brühl and M. Hanke, *Numerical implementation of two noniterative methods for locating inclusions by impedance tomography*, Inverse Problems, 16 (2000), pp. 1029–1042.

[6] M. Brühl, *Explicit characterization of inclusions in electrical impedance tomography*, SIAM J. Math. Anal., 32 (2001), pp. 1327–1341.

[7] M. Brühl, M. Hanke, and M. S. Vogelius, *A direct impedance tomography algorithm for locating small inhomogeneities*, Numer. Math., 93 (2003), pp. 635–654.

[8] M. Costabel, *Boundary integral operators for the heat equation*, Integral Equations Operator Theory, 13 (1990), pp. 488–552.

[9] M. Costabel, V. J. Ervin, and E. P. Stephan, *Symmetric coupling of finite elements and boundary elements for a parabolic-elliptic interface problem*, Quart. Appl. Math., 48 (1990), pp. 265–279.

[10] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology. Volume* 5: *Evolution Problems* I, Springer-Verlag, Berlin, 1992.

[11] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic Publishers, Dordrecht, The Netherlands, 2000.

[12] B. Gebauer, *The factorization method for real elliptic problems*, Z. Anal. Anwend., 25 (2006), pp. 81–102.

[13] M. Hanke and M. Brühl, *Recent progress in electrical impedance tomography*, Inverse Problems, 19 (2003), pp. S65–S90.

[14] M. Hanke and B. Schappel, *The Factorization Method for Electrical Impedance Tomography in the Half Space*, submitted.

[15] N. Hyvönen, *Characterizing inclusions in optical tomography*, Inverse Problems, 20 (2004), pp. 737–751.

[16] A. Kirsch, *Characterization of the shape of a scattering obstacle using the spectral data of the far field operator*, Inverse Problems, 14 (1998), pp. 1489–1512.

[17] A. Kirsch, *The factorization method for a class of inverse elliptic problems*, Math. Nachr., 278 (2005), pp. 258–277.

[18] A. Kirsch, *The factorization method for Maxwell's equations*, Inverse Problems, 20 (2004), pp. S117–134.

[19] R. Kress, *A factorisation method for an inverse Neumann problem for harmonic vector fields*, Georgian Math. J., 10 (2003), pp. 549–560.

[20] J.-L. Lions, *Equations Differentielles Operationnelles et Problèmes aux Limites*, Springer-Verlag, Berlin, 1961.

[21] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications* I, Springer-Verlag, Berlin, 1972.

[22] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications* II, Springer-Verlag, Berlin, 1972.

[23] R. C. MacCamy and M. Suri, *A time-dependent interface problem for two-dimensional eddy currents*, Quart. Appl. Math., 44 (1987), pp. 675–690.

[24] B. Schappel *Electrical impedance tomography in the half space: Locating obstacles by electrostatic measurements on the boundary*, in Proceedings of the 3rd World Congress on Industrial Process Tomography, Banff, VCIPT, 2003, pp. 788–793.

[25] B. Schappel *Die Faktorisierungsmethode für die elektrische Impedanztomographie im Halbraum*, Dissertation, Joh. Gutenberg-Universität Mainz, Mainz, Germany, 2005. Available online at http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:hebis:77-7427.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 837–869

STABILITY AND CONVERGENCE OF THE CRANK–NICOLSON/ADAMS–BASHFORTH SCHEME FOR THE TIME-DEPENDENT NAVIER–STOKES EQUATIONS*

YINNIAN HE† AND WEIWEI SUN‡

Abstract. In this paper, we study the stability and convergence of the Crank–Nicolson/Adams–Bashforth scheme for the two-dimensional nonstationary Navier–Stokes equations. A finite element method is applied for the spatial approximation of the velocity and pressure. The time discretization is based on the Crank–Nicolson scheme for the linear term and the explicit Adams–Bashforth scheme for the nonlinear term. Moreover, we present optimal error estimates and prove that the scheme is almost unconditionally stable and convergent, i.e., stable and convergent when the time step is less than or equal to a constant.

Key words. Navier–Stokes equations, mixed finite element, Adams–Bashforth scheme, Crank–Nicolson scheme

AMS subject classifications. 35L70, 65N30, 76D06

DOI. 10.1137/050639910

1. Introduction.

Let Ω be a bounded domain in R^2 assumed to have a Lipschitz continuous boundary $\partial\Omega$ and to satisfy a further condition stated in (A1) below. We consider the time-dependent Navier–Stokes problem

$$\begin{cases} u_t - \nu\Delta u + (u\cdot\nabla)u + \nabla p = f,\ \text{div } u = 0, \quad (x,t)\in\Omega\times(0,T], \\ u(x,0) = u_0(x), \quad x\in\Omega; \quad u(x,t)|_{\partial\Omega} = 0, \quad t\in[0,T], \end{cases} \tag{1.1}$$

where $u = u(x,t) = (u_1(x,t), u_2(x,t))$ represents the velocity vector, $p = p(x,t)$ represents the pressure, $f = f(x,t)$ represents the prescribed body force, $u_0(x)$ represents the initial velocity, $\nu > 0$ represents the viscosity, and $T > 0$ represents a finite time.

There are numerous works devoted to the development of efficient schemes for the Navier–Stokes equations [3, 4, 8, 9, 12, 13, 16, 20, 5, 23, 25, 24, 28, 30]: fully implicit, semi-implicit (semiexplicit), and explicit. Among them, high-order schemes are of more interest since first-order schemes are not sufficiently accurate for large time approximations. A key issue is the stability condition of schemes. Usually fully implicit schemes are (almost) unconditionally stable. However, at each time step, one has to solve a system of nonlinear equations. An explicit scheme is much easier in computation. But it suffers a severely restricted time step size from stability requirement. A popular approach is based on an implicit scheme for the linear term and a semi-implicit scheme or an explicit scheme for the nonlinear term. A semi-implicit scheme for the nonlinear term results in a linear system with a variable coefficient matrix of time, and an explicit treatment for the nonlinear term gives a constant matrix. Stability and convergence conditions of schemes have been studied

*Received by the editors September 8, 2005; accepted for publication (in revised form) August 7, 2006; published electronically April 27, 2007. This research was subsidized by the NSF of China (10671154) and the Research Grants Council of the Hong Kong Special Administrative Region, China (project CityU 102005.

http://www.siam.org/journals/sinum/45-2/63991.html

†Faculty of Science, Xi'an Jiaotong University, Xi'an 710049, People's Republic of China (heyn@mail.xjtu.edu.cn).

‡Department of Mathematics, City University of Hong Kong, Hong Kong (maweiw@math.cityu.edu.hk).

by many authors. The main results are summarized below, where we set $\Omega \subset R^d$ with $d = 2, 3$; $0 < h < 1$ denotes the mesh size in the spatial direction; and $0 < \tau = \frac{T}{N} < 1$ denotes the step size in the time direction.

- For the fully implicit Crank–Nicolson scheme (implicit for both linear and nonlinear terms), Heywood and Rannacher [16] proved that it is almost unconditionally stable and convergent, i.e., stable and convergent when

$$\tau \le C_0 \tag{1.2}$$

  for some positive constant C_0 depending on the data (ν, Ω, T, u_0, f) in the case of $d = 2, 3$.
- For a two-step scheme with a semi-implicit treatment for the nonlinear term, He and Li [12] gave the following convergence condition:

$$\tau h^{-1/2} \le C_0. \tag{1.3}$$

- For the Crank–Nicolson extrapolation scheme in which the discretization for the nonlinear term is semi-implicit, He [13] has proved that (1.2) is the stability and convergence condition of the scheme in the case of $d = 2$.
- For the Crank–Nicolson/Adams–Bashforth scheme in which the nonlinear term is treated explicitly, Marion and Temam provided in [25] the stability condition

$$\tau h^{-d} \le C_0, \quad d = 2, 3, \tag{1.4}$$

  and recently, Tone [30] proved the convergence under the condition

$$\tau h^{-2-d/2} \le C_0, \quad d = 2,\ 3. \tag{1.5}$$

- A modified Crank–Nicolson/Adams–Bashforth scheme was proposed by Johnston and Liu [19] in which the nonlinear term and pressure term are discretized explicitly. They claimed from their numerical simulations that the scheme is stable under the standard stability condition

$$\|u\|_{L^\infty} \tau h^{-1} \le 1, \quad d = 2, 3\,. \tag{1.6}$$

  No theoretical analysis has been given.
- For a three-step backward extrapolating scheme (explicit for the nonlinear term), Baker, Dougalis, and Karakashian [4] gave the convergence condition

$$\tau h^{-4/7} \le C_0 \tag{1.7}$$

  in the case of $d = 2, 3$.

Clearly, the time step condition

$$\tau h^{-\alpha} \le C_0 \tag{1.8}$$

for some $\alpha > 0$ was imposed in these previous works when a semi-implicit or an explicit scheme is applied for the nonlinear term, except for the Crank–Nicolson extrapolation scheme in [12] in which a semi-implicit scheme is used for the nonlinear term.

This paper focuses on the second-order Crank–Nicolson/Adams–Bashforth scheme with a finite element approximation in spatial direction for solving the time-dependent Navier–Stokes equations in the case of $d = 2$, which were studied by Marion and

Temam [25], Tone [30], Kim and Moin [20], and Issacson and Keller [18]. The scheme consists of using a finite element pair (X_h, M_h) for the spatial discretization of the Navier–Stokes equations, the Crank–Nicolson scheme for the linear term, and the Adams–Bashforth scheme for the nonlinear term. Under the assumption of $u_0 \in H^2(\Omega)^2 \cap H_0^1(\Omega)^2$ with $\mathrm{div} u_0 = 0$ and $f,\ f_t,\ f_{tt} \in L^\infty(0,T;L^2(\Omega)^2)$, we prove that the scheme is almost unconditionally stable, i.e.,

$$\|d_t u_h^m\|_{L^2}^2 + \nu \|A_h u_h^m\|_{L^2}^2 \le \kappa_2, \quad 1 \le m \le N, \tag{1.9}$$

when the condition (1.2) is satisfied. Moreover, we also provide the optimal error estimates

$$\|u(t_m) - u_h^m\|_{L^2} \le \kappa(\sigma^{-1}(t_m)\tau^2 + h^2), \quad 1 \le m \le N, \tag{1.10}$$

$$\|u(t_m) - u_h^m\|_{H^1} \le \kappa(\sigma^{-1/2}(t_m)\tau + h), \quad 1 \le m \le N, \tag{1.11}$$

$$\|p(t_m) - p_h^m\|_{L^2} \le \kappa(\sigma^{-1}(t_m)\tau + \sigma^{-1/2}(t_m)h), \quad 1 \le m \le N, \tag{1.12}$$

where the finite element space pair (X_h, M_h) satisfies the approximation assumption (A3); $\sigma(t) = \min\{1, t\}$; $\kappa_0,\ \kappa_1,\ \kappa_2$, and κ are some positive constants depending on the data (ν, Ω, T, u_0, f); and A_h is a discrete Stokes operator. The error bound (1.12) of the pressure is better than the error bound

$$\|p_h^m - p(t_m)\|_{L^2} \le \kappa(\sigma^{-3/2}(t_m)\tau + \sigma^{-1/2}(t_m)h), \quad 1 \le m \le N, \tag{1.13}$$

obtained by Heywood and Rannacher [16].

This paper is organized as follows. In section 2 an abstract functional setting of the Navier–Stokes problem is given together with some basic assumptions (A1) and (A2). In section 3 we set out our assumption (A3) concerning the finite element spaces X_h and M_h, finite element Galerkin approximation in space, and some properties on the trilinear form $b(\cdot,\cdot,\cdot)$. Section 3 contains the optimal error estimate and a priori estimate results of the finite element solution $(u_h(t), p_h(t))$. In section 4 we describe the Crank–Nicolson/Adams–Bashforth scheme and prove a stability result of this scheme. In section 5 we describe a second-order dual scheme and prove its stability result. In section 6 we provide optimal error estimates for the numerical solution (u_h^n, p_h^n) with $1 \le n \le N$.

2. Functional setting of the Navier–Stokes equations. For the mathematical setting of problem (1.1), we introduce the following Hilbert spaces:

$$X = H_0^1(\Omega)^2, \quad Y = L^2(\Omega)^2, \quad M = L_0^2(\Omega) = \left\{ q \in L^2(\Omega); \int_\Omega q dx = 0 \right\}.$$

The space $L^2(\Omega)^d$, $d = 1, 2, 4$, is equipped with the usual L^2-scalar product $(\cdot,\cdot)$ and L^2-norm $\|\cdot\|_{L^2}$ or $\|\cdot\|_0$. The spaces $H_0^1(\Omega)$ and X are equipped with their usual scalar product and equivalent norm

$$((u,v)) = (\nabla u, \nabla v), \quad \|u\|_{H_0^1} = \|\nabla u\|_{L^2}.$$

Next, let the closed subset V of X be given by

$$V = \{v \in X; \mathrm{div}\ v = 0\},$$

and denote by H the closed subset of Y, i.e.,

$$H = \{v \in Y; \mathrm{div}\ v = 0, v \cdot n|_{\partial\Omega} = 0\}.$$

We refer readers to [1, 9, 15, 29] for details on these spaces. We denote the Stokes operator by $A = -P\Delta$, where P is the L^2-orthogonal projection of Y onto H. As mentioned above, we need a further assumption on Ω provided in [16].

(A1). Assume that Ω is smooth so that the unique solution $(v, q) \in (X, M)$ of the steady Stokes problem

$$-\nu\Delta v + \nabla q = g, \quad \mathrm{div} v = 0 \quad \text{in } \Omega, \quad v|_{\partial\Omega} = 0,$$

for any prescribed $g \in Y$ exists and satisfies

$$\|v\|_{H^2} + \|q\|_{H^1} \le c\|g\|_{L^2},$$

where $c > 0$ is a generic constant depending on Ω and ν which may stand for different values at its different occurrences.

We remark that the validity of assumption (A1) is known (see [9, 15, 21, 29]) if $\partial\Omega$ is of C^2 or if Ω is a two-dimensional convex polygon. From the assumption (A1), it is well known [1, 15, 22] that

$$\|v\|_{H^2} \le c\|Av\|_{L^2}, \quad v \in D(A) = H^2(\Omega)^2 \cap V, \tag{2.1}$$

$$\|v\|_{L^2} \le \gamma_0\|v\|_{H_0^1}, \quad v \in X, \quad \|v\|_{H_0^1} \le \gamma_0\|Av\|_{L^2}, \quad v \in D(A), \tag{2.2}$$

where γ_0 is a positive constant depending only on Ω. We usually make the following assumption about the prescribed data for problem (1.1).

(A2). The initial velocity $u_0(x)$ and the force $f(x, t)$ satisfy that $u_0 \in D(A)$, $f \in L^\infty(0, T; H^1(\Omega)^2)$, f_t, and $f_{tt} \in L^\infty(0, T; Y)$ with

$$\|Au_0\|_{L^2} + \sup_{0\le t\le T}\{\|f(t)\|_{H^1} + \|f_t(t)\|_{L^2} + \|f_{tt}(t)\|_{L^2}\} \le C$$

for some positive constant C. We also introduce the following bilinear operator:

$$B(u, v) = (u \cdot \nabla)v + \frac{1}{2}(\mathrm{div} u)v, \quad u, \ v \in X.$$

Moreover, we define the continuous bilinear forms $a(\cdot, \cdot)$ and $d(\cdot, \cdot)$ on $X \times X$ and $X \times M$, respectively, by

$$a(u, v) = \nu((u, v)), \quad u, v \in X,$$

and

$$d(v, q) = (q, \mathrm{div} v), \quad v \in X, \quad q \in M,$$

and a trilinear form on $X \times X \times X$ by

$$b(u, v, w) = \langle B(u, v), w\rangle_{X',X} = ((u \cdot \nabla)v, w) + \frac{1}{2}((\mathrm{div} u)v, w)$$

$$= \frac{1}{2}((u \cdot \nabla)v, w) - \frac{1}{2}((u \cdot \nabla)w, v), \quad u, v, w \in X.$$

With the above notations, the variational formulation of problem (2.1) reads as follows: find $(u, p) \in (X, M)$ for all $t \in (0, T]$ such that, for all $(v, q) \in (X, M)$,

$$(u_t, v) + a(u, v) - d(v, p) + d(u, q) + b(u, u, v) = (f, v), \tag{2.3}$$

$$u(0) = u_0. \tag{2.4}$$

3. Finite element Galerkin approximation. Let $h > 0$ be a real positive parameter. The finite element subspace (X_h, M_h) of (X, M) is characterized by $J_h = J_h(\Omega)$, a partitioning of $\bar{\Omega}$ into triangles K or quadrilaterals K, assumed to be uniformly regular as $h \to 0$. For further details, readers can refer to Ciarlet [6] and Girault and Raviart [9].

We define the subspace V_h of X_h given by

$$V_h = \left\{ v_h \in X_h \,;\, d(v_h, q_h) = 0 \ \ \forall q_h \in M_h \right\}. \tag{3.1}$$

Let $P_h : Y \to V_h$ denote the L^2-orthogonal projection defined by

$$(P_h v,\ v_h) = (v, v_h), \quad v \in Y,\ v_h \in V_h.$$

We assume that the couple (X_h, M_h) satisfies the following approximation properties.

(A3) For each $v \in H^2(\Omega)^2 \cap X$ and $q \in H^1(\Omega) \cap M$, there exist approximations $\pi_h v \in X_h$ and $\rho_h q \in M_h$ such that

$$\|v - \pi_h v\|_{H_0^1} \le ch\|v\|_{H^2}, \qquad \|q - \rho_h q\|_{L^2} \le ch\|q\|_{H^1} \tag{3.2}$$

together with the inverse inequality

$$\|v_h\|_{H_0^1} \le \alpha h^{-1} \|v_h\|_{L^2}, \quad v_h \in X_h, \tag{3.3}$$

and we have the so-called inf-sup inequality: for each $q_h \in M_h$, there exists $v_h \in X_h, v_h \neq 0$, such that

$$d(v_h, q_h) \ge \beta \|q_h\|_{L^2} \|v_h\|_{H_0^1}, \tag{3.4}$$

where α and β are positive constants depending on Ω.

The following properties are classical (see [2, 9, 15, 17]):

$$\|P_h v\|_{H_0^1} \le \gamma \|v\|_{H_0^1}, \quad v \in X, \tag{3.5}$$

$$\|v - P_h v\|_{L^2} + h\|v - P_h v\|_{H_0^1} \le \gamma h^2 \|Av\|_{L^2}, \quad v \in D(A), \tag{3.6}$$

$$\|v - P_h v\|_{L^2} \le \gamma h \|v - P_h v\|_{H_0^1}, \quad v \in X \tag{3.7}$$

for some positive constant γ.

The standard finite element Galerkin approximation of (2.3)–(2.4) based on (X_h, M_h) reads as follows: find $(u_h, p_h) \in (X_h, M_h)$ such that, for all $0 < t \le T$ and $(v_h, q_h) \in (X_h, M_h)$,

$$(u_{ht}, v_h) + a(u_h, v_h) - d(v_h, p_h) + d(u_h, q_h) + b(u_h, u_h, v_h) = (f, v_h), \tag{3.8}$$

$$u_h(0) = u_{0h} = P_h u_0. \tag{3.9}$$

With the above statements, a discrete analogue $A_h = -P_h \Delta_h$ of the Stokes operator A is defined through the condition that $(-\Delta_h u_h, v_h) = ((u_h, v_h))$ for all $u_h, v_h \in X_h$. The restriction of A_h to V_h is invertible, with the inverse A_h^{-1}. Since A_h^{-1} is self-adjoint and positive definite, we may define "discrete" Sobolev norms on V_h, of any order $r \in R$, by setting

$$\|v_h\|_r = \|A_h^{r/2} v_h\|_{L^2}, \quad v_h \in V_h.$$

These norms will be assumed to have various properties similar to their continuous counterparts, an assumption that implicitly imposes conditions on the structure of the spaces X_h and M_h. In particular, there holds

$$\|v_h\|_0 = \|v_h\|_{L^2}, \quad \|v_h\|_1 = \|\nabla v_h\|_0, \quad \|v_h\|_2 = \|A_h v_h\|_0, \quad v_h \in V_h.$$

By the way, we derive from (2.2) that

$$\|v_h\|_0 \le \gamma_0 \|\nabla v_h\|_0, \quad \|\nabla v_h\|_0 \le \gamma_0 \|A_h v_h\|_0, \quad v_h \in V_h, \tag{3.10}$$

where $\gamma_0 > 0$ is a constant depending only on Ω.

Remark 3.1. The space V_h is introduced only for theoretical analysis. The practical computation should be based on the finite element space pair (X_h, M_h). For the details of the construction of (X_h, M_h), we refer readers to Heywood and Rannacher [15, 16] and to Hill and Süli [17]. Recently, Nochetto and Pyo [26] proposed a projection method for time-dependent Navier–Stokes equations, in which V_h is a discrete divergence free space which is discontinuous on the boundary of each element and $u \ne 0$ on $\partial\Omega$.

Under the conditions above, and with some further assumptions about the structure of the spaces X_h and M_h, it has been shown in Heywood and Rannacher [15] that

$$\|u(t) - u_h(t)\|_0 + h\|\nabla(u(t) - u_h(t))\|_0 + \sigma^{1/2}(t)h\|p(t) - p_h(t)\|_0 \le \kappa h^2 \tag{3.11}$$

for all $t \in (0, T]$.

This section considers preliminary estimates which are useful in the error estimates of finite element solution. Some estimates of the trilinear form b are given in the following lemma and in the proof can be found in [13, 14].

LEMMA 3.1. *The trilinear form b satisfies the following estimates:*

$$b(u_h, v_h, w_h) = -b(u_h, w_h, v_h), \tag{3.12}$$

$$\begin{aligned} |b(u_h, v_h, w_h)| &+ |b(v_h, u_h, w_h)| + |b(w_h, u_h, v_h)| \\ &\le \frac{c_0}{2}\|u_h\|_0^{1/2}\|u_h\|_1^{1/2}\|v_h\|_1\|w_h\|_0^{1/2}\|w_h\|_1^{1/2} \\ &+ \frac{c_0}{2}\|u_h\|_1\|v_h\|_0^{1/2}\|v_h\|_1^{1/2}\|w_h\|_0^{1/2}\|w_h\|_1^{1/2} \end{aligned} \tag{3.13}$$

for all u_h, v_h, and $w_h \in X_h$;

$$\begin{aligned} |b(u_h, v_h, w_h)| &+ |b(v_h, u_h, w_h)| + |b(w_h, u_h, v_h)| \\ &\le \frac{1}{2}c_0\|A_h v_h\|_0^{1/2}\|v_h\|_1^{1/2}\|u_h\|_0^{1/2}\|u_h\|_1^{1/2}\|w_h\|_0 \\ &+ \frac{1}{2}c_0\|A_h v_h\|_0^{1/2}\|v_h\|_0^{1/2}\|u_h\|_1\|w_h\|_0 \end{aligned} \tag{3.14}$$

for all $u_h,\ v_h \in V_h$, and $w_h \in X_h$; and

$$\begin{aligned} |b(u_h, v_h, w_h)| &+ |b(u_h, v_h, w_h)| + |b(w_h, u_h, v_h)| \\ &\le \frac{1}{3}c_0(\|u_h\|_0^{1/2}\|A_h u_h\|_0^{1/2}\|A_h v_h\|_0 + \|v_h\|_0^{1/2}\|A_h v_h\|_0^{1/2}\|A_h u_h\|_0)\|w_h\|_{-1} \\ &+ \frac{1}{3}c_0\|u_h\|_1^{1/2}\|A_h u_h\|_0^{1/2}\|A_h v_h\|_0^{1/2}\|v_h\|_1^{1/2}\|w_h\|_{-1} \end{aligned} \tag{3.15}$$

for all u_h, v_h, and $w_h \in V_h$, where $c_0 > 0$ is a constant depending only on Ω.

Proof. (3.12)–(3.14) can be found in [13, 14, 15, 17].

Before we proceed further with (3.15), we need some continuous and discrete Gadliardo-Nirenberg estimates (see Temam [29] and Hill and Süli [17]):

$$\|\nabla v\|_{L^4} \le c\|\nabla v\|_0^{1/2}\|Av\|_0^{1/2} \quad \forall v \in D(A), \tag{3.16}$$

$$\|v_h\|_{L^4} \le c\|v_h\|_0^{1/2}\|v_h\|_1^{1/2}, \quad \|v_h\|_{L^\infty} \le c\|v_h\|_0^{1/2}\|A_h v_h\|_0^{1/2},$$

$$\|\nabla v_h\|_{L^4} \le c\|\nabla v_h\|_0^{1/2}\|A_h v_h\|_0^{1/2} \quad \forall v_h \in V_h. \tag{3.17}$$

Moreover, let the map $A^{-1}PA_h : V_h \to D(A)$. Then Heywood and Rannacher [15] showed

$$\|A^{-1}PA_h v_h - v_h\|_0 + h\|\nabla(A^{-1}PA_h v_h - v_h)\|_0 \le ch^2\|A_h v_h\|_0. \tag{3.18}$$

Let $\phi_h = A_h^{-1} w_h$ for each $w_h \in V_h$; then (3.3) gives

$$\|w_h\|_0^2 = (w_h, A_h\phi_h) = (\nabla w_h, \nabla\phi_h) \le \|w_h\|_1\|w_h\|_{-1} \le ch^{-1}\|w_h\|_0\|w_h\|_{-1}. \tag{3.19}$$

Furthermore, we write the bilinear form b into the following:

$$\begin{aligned} b(u_h, v_h, w_h) \le{}& ((u_h\cdot\nabla)(v_h - PA_h v_h), w_h) + \frac{1}{2}(\mathrm{div}(u_h - A^{-1}PA_h u_h)v_h, w_h) \\ &+ \|\nabla P_h[(u_h\cdot\nabla)A^{-1}PA_h v_h]\|_0\|w_h\|_{-1}. \end{aligned} \tag{3.20}$$

Using (3.3), (3.5), and (3.16)–(3.19), we have

$$\begin{aligned} |((u_h\cdot\nabla)(v_h - PA_h v_h), w_h)| &\le ch\|u_h\|_{L^\infty}\|A_h v_h\|_0\|w_h\|_0 \\ &\le c\|u_h\|_{L^\infty}\|A_h v_h\|_0\|w_h\|_{-1}, \\ |(\mathrm{div}(u_h - A^{-1}PA_h u_h)v_h, w_h)| &\le ch\|A_h u_h\|_0\|v_h\|_{L^\infty}\|w_h\|_0 \\ &\le c\|A_h u_h\|_0\|v_h\|_{L^\infty}\|w_h\|_{-1}, \\ \|\nabla P_h[(u_h\cdot\nabla)A^{-1}PA_h v_h]\|_0 &\le c\|\nabla[(u_h\cdot\nabla)A^{-1}PA_h v_h]\|_0 \\ &\le c\|\nabla u_h\|_{L^4}\|\nabla A^{-1}PA_h v_h\|_{L^4} \\ &\quad + c\|u_h\|_{L^\infty}\|\nabla\nabla A^{-1}PA_h v_h\|_0 \\ &\le c\|u_h\|_1^{1/2}\|A_h u_h\|_0^{1/2}\|v_h\|_1^{1/2}\|A_h v_h\|_0^{1/2} \\ &\quad + c\|u_h\|_{L^\infty}\|A_h v_h\|_0. \end{aligned}$$

Combining these inequalities with (3.20), we have deduced

$$\begin{aligned} |b(u_h, v_h, w_h)| \le{}& c(\|u_h\|_{L^\infty}\|A_h v_h\|_0 + c\|A_h u_h\|_0\|v_h\|_{L^\infty})\|w_h\|_{-1} \\ &+ c\|u_h\|_1^{1/2}\|A_h u_h\|_0^{1/2}\|v_h\|_1^{1/2}\|A_h v_h\|_0^{1/2}\|w_h\|_{-1}. \end{aligned} \tag{3.21}$$

Also, we can obtain the above estimate for the trilinear terms $b(v_h, u_h, w_h)$ and $b(w_h, u_h, v_h)$. Hence, we deduce (3.15) by using the above estimates and (3.17). ∎

Remark 3.2. (3.13) and (3.14) are valid in two-dimensional space. One has to seek a different approach for problems in three-dimensional space, e.g., refer to E and Liu [7] and Nochetto and Pyo [26].

In order to obtain our error analysis for time discretization, we recall the following smooth properties of (u_h, p_h) given in [16].

THEOREM 3.2. *Assume that assumptions* (A1)–(A3) *are valid. Then the finite element solution* (u_h, p_h) *satisfies the following estimates:*

$$\|u_h(t)\|_2^2 + \|p_h(t)\|_0^2 \le \kappa, \quad \sigma^r(t)\|u_{ht}(t)\|_r^2 \le \kappa, \quad r = 0, 1, 2, \tag{3.22}$$

$$\sigma^{r+2}\|u_{htt}(t)\|_r^2 \le \kappa, \quad r = -1, 0, 1, \tag{3.23}$$

$$\int_0^t \sigma^r(s)\|u_{ht}(s)\|_{r+1}^2 ds \le \kappa, \quad r = 0, 1, \tag{3.24}$$

$$\int_0^t \sigma^{r+1}(s)\|u_{htt}(s)\|_r^2 ds \le \kappa, \quad r = -1, 0, 1, \tag{3.25}$$

$$\int_0^t \sigma^{r+2}(s)\|u_{http}(s)\|_{r-1}^2 ds \le \kappa, \quad r = -1, 0, 1 \tag{3.26}$$

for all $t \in (0, T]$.

THEOREM 3.3. *Under the assumptions of Theorem* 3.2*, there holds*

$$\int_0^t \sigma^3(s)\|A_h u_{htt}(s)\|_0^2 ds \le \kappa, \quad \sigma^2(t)\|p_{ht}(t)\|_0^2 \le \kappa \tag{3.27}$$

for all $t \in (0, T]$.

Proof. From (3.8), we have

$$\begin{aligned}(u_{http}, v_h) + a(u_{htt}, v_h) + b(u_{htt}, u_h, v_h) + b(u_h, u_{htt}, v_h) + 2b(u_{ht}, u_{ht}, v_h)\\ = (f_{tt}, v_h), \quad v_h \in V_h.\end{aligned} \tag{3.28}$$

Taking $v_h = 2A_h u_{htt} \in V_h$ in (3.28), we obtain

$$\begin{aligned}2(u_{http}, A_h u_{htt}) + 2\nu\|A_h u_{htt}\|_0^2 + 2b(u_{htt}, u_h, A_h u_{htt}) + 2b(u_h, u_{htt}, A_h u_{htt})\\ + 2b(u_{ht}, u_{ht}, A_h u_{htt}) = 2(f_{tt}, A_h u_{htt}).\end{aligned} \tag{3.29}$$

It follows from (3.10), (3.14), and the Young inequality that

$$\begin{aligned}2|b(u_{htt}, u_h, A_h u_{htt})| + 2|b(u_h, u_{htt}, A_h u_{htt})| &\le 2c_0\gamma_0\|A_h u_h\|_0\|u_{htt}\|_1\|A_h u_{htt}\|_0\\ &\le \frac{\nu}{4}\|A_h u_{htt}\|_0^2 + 4\nu^{-1}c_0^2\gamma_0^2\|A_h u_h\|_0^2\|u_{htt}\|_1^2,\\ 4|b(u_{ht}, u_{ht}, A_h u_{htt})| &\le 2c_0\|A_h u_{ht}\|_0^{1/2}\|u_{ht}\|_0^{1/2}\|u_{ht}\|_1\|A_h u_{htt}\|_0\\ &\le \frac{\nu}{4}\|A_h u_{htt}\|_0^2 + 4\nu^{-1}c_0^2\|A_h u_{ht}\|_0\|u_{ht}\|_0\|u_{ht}\|_1^2,\\ 2(u_{http}, A_h u_{htt}) &\le \frac{\nu}{4}\|A_h u_{htt}\|_0^2 + 4\nu^{-1}\|u_{http}\|_0^2,\\ 2|(f_{tt}, A_h u_{htt})| &\le \frac{\nu}{4}\|A_h u_{htt}\|_0^2 + 4\nu^{-1}\|f_{tt}\|_0^2.\end{aligned}$$

Combining these inequalities with (3.29) yields

$$\begin{aligned}\nu\|A_h u_{htt}\|_0^2 &\le 4\nu^{-1}c_0^2\gamma_0^2\|A_h u_h\|_0^2\|u_{htt}\|_0^2\\ &\quad + 4\nu^{-1}c_0^2\|A_h u_{ht}\|_0\|u_{ht}\|_0\|u_{ht}\|_1^2 + 4\nu^{-1}\|u_{http}\|_0^2 + 4\nu^{-1}\|f_{tt}\|_0^2.\end{aligned} \tag{3.30}$$

Multiplying (3.30) by $\sigma^3(t)$ and integrating from 0 to t, we have

$$\begin{aligned}\nu\int_0^t \sigma^3(s)\|A_h u_{htt}\|_0^2 ds &\le 4\nu^{-1}c_0^2\gamma_0^2\int_0^t \sigma^3(s)\|A_h u_h\|_0^2\|u_{htt}\|_1^2 ds\\ &\quad + 4\nu^{-1}c_0^2\int_0^t \sigma^3(s)\|A_h u_{ht}\|_0\|u_{ht}\|_0\|u_{ht}\|_1^2 ds\\ &\quad + \nu^{-1}\int_0^t \sigma^3(s)(\|u_{httt}\|_0^2 + \|f_{tt}\|_0^2) ds,\end{aligned}$$

which together with (3.22) and (3.24)–(3.26) leads to

$$\int_0^t \sigma^3(s)\|A_h u_{htt}\|_0^2 ds \le \kappa, \quad 0 < t \le T. \tag{3.31}$$

Moreover, by using (3.4), (3.10), Lemma 3.1, and (3.8), we have

$$\begin{aligned}\sigma^2(t)\|p_{ht}(t)\|_0^2 &\le c\sigma^2(t)\|u_{htt}(t)\|_0^2 + c\sigma^2(t)\|A_h u_{ht}(t)\|_0^2\\ &\quad + c\|A_h u_h(t)\|_0^2\|u_{ht}(t)\|_0^2,\end{aligned}$$

which with Theorem 3.2 results in

$$\sigma^2(t)\|p_{ht}(t)\|_0^2 \le \kappa, \quad 0 < t \le T.$$

Combining this inequality with (3.31) implies (3.27). ∎

We will frequently use a discrete version of the Gronwall lemmas used in [11] and [27].

LEMMA 3.4. *Let C, τ, a_n, b_n, c_n, and d_n for integers $n \ge 0$ be nonnegative numbers such that*

$$a_m + \tau\sum_{n=1}^{m} b_n \le \tau\sum_{n=0}^{m-1} a_n d_n + \tau\sum_{n=0}^{m-1} c_n + C, \quad m \ge 1. \tag{3.32}$$

Then

$$a_m + \tau\sum_{n=1}^{m} b_n \le \exp\left(\tau\sum_{n=0}^{m-1} d_n\right)\left(\tau\sum_{n=0}^{m-1} c_n + C\right), \quad m \ge 1. \tag{3.33}$$

4. Second-order fully discrete finite element method. In this section we consider the time discretization of the finite element Galerkin approximation (3.8)–(3.9). Let $t_n = n\tau (n = 0, 1, \ldots, N)$, $\tau = \frac{T}{N}$ be the time step size, and N be an integer. Due to the nature of the Adams–Bashforth scheme of three levels in time, we define $u_h^0 = u_{0h} = P_h u_0$ and $(u_h^1, p_h^1) \in (X_h, M_h)$ by the Euler-backward scheme:

$$(d_t u_h^1, v_h) + a(u_h^1, v_h) - d(v_h, p_h^1) + d(u_h^1, q_h) + b(u_h^0, u_h^0, v_h) = (f(t_1), v_h) \tag{4.1}$$

for all $(v_h, q_h) \in (X_h, M_h)$, while $d_t u_h^0 = \lim_{t\to 0} u_{ht}(t)$ is defined so that

$$(d_t u_h^0, v_h) + a(u_h^0, v_h) + b(u_h^0, u_h^0, v_h) = (f(t_0), v_h) \tag{4.2}$$

for all $v_h \in V_h$.

Now, we define recursively the finite element solutions $(u_h^n, p_h^n) \in (X_h, M_h)$, $n = 2, \ldots, N$, by setting

$$\begin{aligned}(d_t u_h^n, v_h) + a(\bar{u}_h^n, v_h) - d(v_h, p_h^n) + d(\bar{u}_h^n, q_h) \\ + \frac{3}{2} b(u_h^{n-1}, u_h^{n-1}, v_h) - \frac{1}{2} b(u_h^{n-2}, u_h^{n-2}, v_h) = (\bar{f}(t_n), v_h)\end{aligned} \tag{4.3}$$

or

$$\begin{aligned}&(d_t u_h^n, v_h) + a(\bar{u}_h^n, v_h) - d(v_h, p_h^n) + d(\bar{u}_h^n, q_h) + b(\bar{u}_h^{n-1}, \bar{u}_h^{n-1}, v_h) \\ &\quad + b(d_t u_h^{n-1}, \bar{u}_h^{n-1}, v_h)\tau + b(\bar{u}_h^{n-1}, d_t u_h^{n-1}, v_h)\tau + \frac{1}{4} b(d_t u_h^{n-1}, d_t u_h^{n-1}, v_h)\tau^2 \\ &\quad = (\bar{f}(t_n), v_h)\end{aligned} \tag{4.4}$$

for all $(v_h, q_h) \in (X_h, M_h)$.

Here and after, we often use the following notations:

$$\bar{u}_h^n = \frac{1}{2}(u_h^n + u_h^{n-1}), \quad \bar{u}_h(t_n) = \frac{1}{2}(u_h(t_n) + u_h(t_{n-1})), \quad d_t u_h^n = \frac{1}{\tau}(u_h^n - u_h^{n-1}).$$

We see from (3.3), (3.6), and the definition of A_h that

$$\begin{aligned}(A_h u_h^0, v_h) &= ((u_h^0 - u_0, v_h)) + ((u_0, v_h)) \le (\alpha h^{-1} \|\nabla(P_h u_0 - u_0)\|_0 + \|Au_0\|_0)\|v_h\|_0 \\ &\le (1 + \alpha\gamma)\|Au_0\|_0 \|v_h\|_0, \quad v_h \in V_h,\end{aligned}$$

which with (3.5) yields

$$\|u_h^0\|_0 \le \|u_0\|_0, \quad \|u_h^0\|_1 \le \gamma \|\nabla u_0\|_0, \quad \|A_h u_h^0\|_0 \le (1 + \alpha\gamma)\|Au_0\|_0. \tag{4.5}$$

We deduce from (4.2) and Lemma 3.1 that

$$\|d_t u_h^0\|_0 \le \nu \|A_h u_h^0\|_0 + c_0\gamma_0 \|A_h u_h^0\|_0 \|u_h^0\|_1 + \|f(t_0)\|_0$$

and by (4.5) that

$$\|d_t u_h^0\|_0 \le (1 + \alpha\gamma)(\nu + c_0\gamma_0\gamma\|\nabla u_0\|_0)\|Au_0\|_0 + \|f(t_0)\|_0. \tag{4.6}$$

LEMMA 4.1. *Suppose that the assumptions* (A1)–(A3) *are valid. Then* (u_h^1, p_h^1) *satisfies the following stabilities:*

$$\|u_h^1\|_0^2 + \nu \|\bar{u}_h^1\|_1^2 \tau \le \kappa_0', \tag{4.7}$$

$$\|u_h^1\|_1^2 + \nu \|A_h \bar{u}_h^1\|_0^2 \tau \le \kappa_1', \tag{4.8}$$

$$\|d_t u_h^1\|_0^2 + \nu \|A_h u_h^1\|_0^2 + \|p_h^1\|_0^2 + \nu \|d_t u_h^1\|_1^2 \tau \le \kappa_2' \tag{4.9}$$

for some positive constants κ_0', κ_1', *and* κ_2' *depending on the data* (ν, Ω, T, u_0, f).

Proof. Taking $v_h = 2A_h^r u_h^1 \tau \in V_h$ and $q_h = 0$ in (4.1) for $r = 0, 1$, we obtain

$$\begin{aligned}&\|u_h^1\|_r^2 - \|u_h^0\|_r^2 + \|d_t u_h^1\|_r^2 \tau^2 + 2\nu \|u_h^1\|_{r+1}^2 \tau + 2b(u_h^0, u_h^0, A_h^r u_h^1)\tau \\ &\quad = 2(f(t_1), A_h^r u_h^1)\tau.\end{aligned} \tag{4.10}$$

In view of Lemma 3.1 and (3.10), there hold

$$\begin{aligned}2|b(u_h^0, u_h^0, A_h^r u_h^1)|\tau &\le c_0\gamma_0 \|u_h^0\|_1 \|u_h^0\|_{r+1} \|u_h^1\|_{r+1}\tau \\ &\le \frac{\nu}{2}\|u_h^1\|_{r+1}^2 \tau + \nu^{-1} c_0^2 \gamma_0^2 \|u_h^0\|_1^2 \|u_h^0\|_{r+1}^2 \tau, \\ 2|(f(t_1), A_h^r u_h^1)|\tau &\le \frac{\nu}{2}\|u_h^1\|_{r+1}^2 + 2\nu^{-1}\gamma_0^{2(1-r)} \|f(t_1)\|_0^2 \tau.\end{aligned}$$

Combining these inequalities with (4.10) yields

$$\begin{aligned}&\|u_h^1\|_r^2 + \|d_t u_h^1\|_r^2\tau^2 + \nu\|u_h^1\|_{r+1}^2\tau\\ &\qquad \le \|u_h^0\|_r^2 + 2\nu^{-1}\gamma_0^{2(1-r)}\|f(t_1)\|_0^2\tau + \nu^{-1}c_0^2\gamma_0^2\|u_h^0\|_1^2\|u_h^0\|_{r+1}^2\tau,\end{aligned} \tag{4.11}$$

which with (4.5)–(4.6) and the triangle inequality implies (4.7)–(4.8).

Again, we deduce from (4.1)–(4.2) that

$$(d_{tt}u_h^1, v_h) + a(d_t u_h^1, v_h) - d(v_h, d_t p_h^1) + d(d_t u_h^1, q_h) = \frac{1}{\tau}\int_{t_0}^{t_1}(f_t(t), v_h)dt. \tag{4.12}$$

By taking $v_h = 2d_t u_h^1\tau$ and $q_h = 2d_t p_h^1\tau$ in (4.12), we have

$$\begin{aligned}&\|d_t u_h^1\|_0^2 + \|d_{tt}u_h^1\|_0^2\tau^2 + 2\nu\|d_t u_h^1\|_1^2\tau \le \|d_t u_h^0\|_0^2 + 2\gamma_0\int_{t_0}^{t_1}\|f_t(t)\|_0 dt\|d_t u_h^1\|_1\\ &\qquad \le \|d_t u_h^0\|_0^2 + \nu\|d_t u_h^1\|_1^2\tau + \nu^{-1}\gamma_0^2\int_{t_0}^{t_1}\|f_t(t)\|_0^2 dt.\end{aligned} \tag{4.13}$$

Moreover, it follows from (4.1) and Lemma 3.1 that

$$\nu\|A_h u_h^1\|_0 \le \|d_t u_h^1\|_0 + c_0\gamma_0\|u_h^0\|_1\|A_h u_h^0\|_0 + \|f(t_1)\|_0$$

and

$$\nu\|A_h u_h^1\|_0^2 \le 3\nu^{-1}(\|d_t u_h^1\|_0^2 + \|f(t_1)\|_0^2 + c_0^2\gamma_0^2\|u_h^0\|_1^2\|A_h u_h^0\|_0^2). \tag{4.14}$$

Finally, using (4.1), (3.4), (3.10), and Lemma 3.1, we arrive at

$$\begin{aligned}\|p_h^1\|_0 &\le \beta^{-1}\sup_{v_h\in X_h}\frac{d(v_h, p_h^1)}{\|v_h\|_1}\\ &\le \beta^{-1}(\|d_t u_h^1\|_0 + \nu\|A_h u_h^1\|_0 + c_0\gamma_0\|u_h^0\|_1^2 + \|f(t_1)\|_0)\end{aligned}$$

or equivalently

$$\|p_h^1\|_0^2 \le 4\beta^{-2}(\|d_t u_h^1\|_0^2 + \nu^2\|A_h u_h^1\|_0^2 + \|f(t_1)\|_0^2 + c_0^2\gamma_0^2\|u_h^0\|_1^2). \tag{4.15}$$

Combining (4.14)–(4.15) with (4.13) and using (4.5)–(4.6) yields (4.9). ∎

The following theorem provides the stability of the scheme in (4.1) and (4.3).

THEOREM 4.2. *Suppose that the assumptions* (A1)–(A3) *are valid and* $0 < \tau < 1$ *satisfies the following stability condition:*

$$160c_0^2\gamma_0^2\nu^{-2}\kappa_2\max\{1, \nu, \kappa_1^{1/2}\}\tau \le 1. \tag{4.16}$$

Then there hold

$$\|u_h^m\|_0^2 + \nu\tau\sum_{n=1}^{m}\|\bar{u}_h^n\|_1^2 \le \kappa_0, \tag{4.17}$$

$$\|u_h^m\|_1^2 + \nu\tau\sum_{n=1}^{m}\|A_h\bar{u}_h^n\|_0^2 \le \kappa_1, \tag{4.18}$$

$$\|d_t u_h^m\|_0^2 + \nu\|A_h u_h^m\|_0^2 + \|p_h^m\|_0^2 + \nu\|d_t u_h^m\|_1^2\tau \le \kappa_2 \tag{4.19}$$

for all $1 \le m \le N$, *where* $\kappa_0 \ge \kappa_0'$, $\kappa_1 \ge \kappa_1'$, *and* $\kappa_2 \ge \kappa_2'$ *are some positive constants depending on the data* (ν, Ω, T, u_0, f).

Proof. We prove (4.17)–(4.19) by the induction method. From Lemma 4.1, (4.17)–(4.19) are true for $m = 1$. We assume that (4.17)–(4.19) are true for $m = 1, \dots, J-1$ with $2 \le J \le N$. We need to prove that (4.17)–(4.19) are true for $m = J$.

First, taking $v_h = 2u_h^n\tau \in V_h$ and $q_h = 0$ in (4.3) and using (3.12) and the formulas

$$u_h^n = \bar{u}_h^n + \frac{1}{2}d_t u_h^n\tau, \quad u_h^n = 2\bar{u}^n - u_h^{n-1}, \quad u_h^n = \bar{u}_h^{n-1} + d_t u_h^n\tau + \frac{1}{2}d_t u_h^{n-1}\tau,$$
$$2(d_t u_h^n, u_h^n)\tau = \|u_h^n\|_0^2 - \|u_h^{n-1}\|_0^2 + \|d_t u_h^n\|_0^2\tau^2,$$
$$2a(\bar{u}_h^n, u_h^n)\tau = \frac{\nu}{2}(\|u_h^n\|_1^2 - \|u_h^{n-1}\|_1^2 + 4\|\bar{u}_h^n\|_1^2)\tau,$$

we obtain

$$\begin{aligned}
&\|u_h^n\|_0^2 - \|u_h^{n-1}\|_0^2 + \|d_t u_h^n\|_0^2\tau^2 + \frac{\nu}{2}(\|u_h^n\|_1^2 - \|u_h^{n-1}\|_1^2 + 4\|\bar{u}_h^n\|_1^2)\tau \\
&\quad + 2b\left(\bar{u}_h^{n-1}, \bar{u}_h^{n-1}, d_t u_h^n + \frac{1}{2}d_t u_h^{n-1}\right)\tau^2 + 2b\left(d_t u_h^{n-1}, \bar{u}_h^{n-1}, \bar{u}_h^n + \frac{1}{2}d_t u_h^n\tau\right)\tau^2 \\
&\quad + 2b\left(\bar{u}_h^{n-1}, d_t u_h^{n-1}, \bar{u}_h^n + \frac{1}{2}d_t u_h^n\tau\right)\tau^2 + b\left(d_t u_h^{n-1}, d_t u_h^{n-1}, \bar{u}_h^n - \frac{1}{2}u_h^{n-1}\right)\tau^3 \\
&= (\bar{f}(t_n), 2\bar{u}_h^n + d_t u_h^n\tau)\tau. \qquad (4.20)
\end{aligned}$$

In view of Lemma 3.1 and (3.10), there hold

$$\begin{aligned}
2\left|b\left(\bar{u}_h^{n-1}, \bar{u}_h^{n-1}, d_t u_h^n + \frac{1}{2}d_t u_h^{n-1}\right)\right|\tau^2 &\le 2c_0\gamma_0\|A_h\bar{u}_h^{n-1}\|_0\|\bar{u}_h^{n-1}\|_1\left\|d_t u_h^n + \frac{1}{2}d_t u_h^{n-1}\right\|_0\tau^2 \\
&\le \frac{1}{4}\|d_t u_h^n\|_0^2\tau^2 + \frac{1}{8}\|d_t u_h^{n-1}\|_0^2\tau^2 \\
&\quad + 6c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-1}\|_0^2\|\bar{u}_h^{n-1}\|_1^2\tau^2, \\
2|b(d_t u_h^{n-1}, \bar{u}_h^{n-1}, \bar{u}_h^n)|\tau^2 + 2|b(\bar{u}_h^{n-1}, d_t u_h^{n-1}, \bar{u}_h^n)|\tau^2 &\le \frac{\nu}{4}\|\bar{u}_h^n\|_1^2\tau + 4\nu^{-1}c_0^2\gamma_0^2\|d_t u_h^{n-1}\|_0^2\|A_h\bar{u}_h^{n-1}\|_0^2\tau^3, \\
|b(d_t u_h^{n-1}, \bar{u}_h^{n-1}, d_t u_h^n)|\tau^3 + |b(\bar{u}_h^{n-1}, d_t u_h^{n-1}, d_t u_h^n)|\tau^3 &\le \frac{1}{4}\|d_t u_h^n\|_0^2\tau^2 + c_0^2\gamma_0^2\|d_t u^{n-1}\|_1^2\|A_h\bar{u}_h^{n-1}\|_0^2\tau^4, \\
|b(d_t u_h^{n-1}, d_t u_h^{n-1}, \bar{u}_h^n)|\tau^3 &\le \frac{\nu}{4}\|\bar{u}_h^n\|_1^2\tau \\
&\quad + \nu^{-1}c_0^2\gamma_0^2\|d_t u_h^{n-1}\|_0^2\|A_h(u_h^{n-1} - u_h^{n-2})\|_0^2\tau^3, \\
\frac{1}{2}|b(d_t u_h^{n-1}, d_t u_h^{n-1}, u_h^{n-1})|\tau^3 &\le \frac{1}{8}\|d_t u_h^{n-1}\|_0^2\tau^2 + \frac{1}{2}c_0^2\gamma_0^2\|d_t u_h^{n-1}\|_1^2\|A_h u_h^{n-1}\|_0^2\tau^4, \\
|(\bar{f}(t_n), 2\bar{u}_h^n + d_t u_h^n\tau)|\tau &\le \frac{\nu}{4}\|\bar{u}_h^n\|_1^2\tau + \frac{1}{8}\|d_t u_h^n\|_0^2\tau^2 \\
&\quad + (4\nu^{-1}\gamma_0^2 + 2\tau)\|\bar{f}(t_n)\|_0^2\tau.
\end{aligned}$$

Combining these inequalities with (4.20) yields

$$\begin{aligned}
&\left(\|u_h^n\|_0^2+\frac{\nu}{2}\|u_h^n\|_1^2\tau\right)-\left(\|u_h^{n-1}\|_0^2+\frac{\nu}{2}\|u_h^{n-1}\|_1^2\tau\right)+\nu\|\bar u_h^n\|_1^2\tau\\
&\quad+\frac{\nu}{4}\|\bar u_h^n\|_1^2\tau-6c_0^2\gamma_0^2\|A_h\bar u_h^{n-1}\|_0^2\|\bar u_h^{n-1}\|_1^2\tau^2+\frac{3}{8}\|d_tu_h^n\|_0^2\tau^2-\frac{2}{8}\|d_tu_h^{n-1}\|_0^2\tau^2\\
&\quad-10\nu^{-1}c_0^2\gamma_0^2(\|A_hu_h^{n-1}\|_0^2+\|A_hu_h^{n-2}\|_0^2)\|d_tu_h^{n-1}\|_0^2\tau^3\\
&\quad\le 3\nu^{-1}c_0^2\gamma_0^2(\|A_hu_h^{n-1}\|_0^2+\|A_hu_h^{n-2}\|_0^2)\nu\|d_tu_h^{n-1}\|_1^2\tau^4\\
&\quad+(4\nu^{-1}\gamma_0^2+2\tau)\|\bar f(t_n)\|_0^2\tau. \qquad (4.21)
\end{aligned}$$

Using (4.16) and the induction assumption with $m=1,\dots,J-1$, we have

$$\begin{aligned}
6c_0^2\gamma_0^2\|A_h\bar u_h^{n-1}\|_0^2\tau&\le 6c_0^2\gamma_0^2\kappa_2\tau\le\frac{\nu}{4},\\
10\nu^{-1}c_0^2\gamma_0^2(\|A_hu_h^{n-1}\|_0^2+\|A_hu_h^{n-2}\|_0^2)\tau&\le 20\nu^{-2}c_0^2\gamma_0^2\kappa_2\tau\le\frac{1}{8},\\
3\nu^{-1}c_0^2\gamma_0^2(\|A_hu_h^{n-1}\|_0^2+\|A_hu_h^{n-2}\|_0^2)\nu\|d_tu_h^{n-1}\|_1^2\tau^4&\le 6\nu^{-2}c_0^2\gamma_0^2\kappa_2^2\tau^3
\end{aligned}$$

for all $n=2,\dots,J$. Summing (4.21) from 2 to J and using the above estimates, we have

$$\begin{aligned}
\|u_h^J\|_0^2+\nu\tau\sum_{n=1}^{J}\|\bar u_h^n\|_1^2&\le\|u_h^1\|_0^2+\frac{\nu}{2}\|u_h^1\|_1^2\tau+\frac{3}{8}\|d_tu_h^1\|_0^2\tau^2+\frac{\nu}{4}\|\bar u_h^1\|_1^2\tau\\
&\quad+6\nu^{-2}c_0^2\gamma_0^2T\kappa_2^2\tau^2+(4\nu^{-1}\gamma_0^2+2\tau)T\sup_{0\le t\le T}\|f(t)\|_0^2 \qquad (4.22)
\end{aligned}$$

for all $1\le J\le N$. Using Lemma 4.1 and (4.16) in (4.22) yields (4.17) with $m=J$.

Next, by taking $v_h=2A_h\bar u_h^n\tau\in V_h$ and $q_h=0$ in (4.4), we obtain

$$\begin{aligned}
&\|u_h^n\|_1^2-\|u_h^{n-1}\|_1^2+2\nu\|A_h\bar u_h^n\|_0^2\tau+2b(\bar u_h^{n-1},\bar u_h^{n-1},A_h\bar u_h^n)\tau\\
&\quad+2b(\bar u_h^{n-1},d_tu_h^{n-1},A_h\bar u_h^n)\tau^2+2b(d_tu_h^{n-1},\bar u_h^{n-1},A_h\bar u_h^n)\tau^2\\
&\quad+\frac{1}{2}b(d_tu_h^{n-1},d_tu_h^{n-1},A_h\bar u_h^n)\tau^3=2(\bar f(t_n),A_h\bar u_h^n)\tau. \qquad (4.23)
\end{aligned}$$

In view of Lemma 3.1 and (3.10), there hold

$$\begin{aligned}
2|b(\bar u_h^{n-1},\bar u_h^{n-1},A_h\bar u_h^n)|\tau&\le c_0\|A_h\bar u_h^{n-1}\|_0^{1/2}\|\bar u_h^{n-1}\|_0^{1/2}\|\bar u_h^{n-1}\|_1\|A_h\bar u_h^n\|_0\tau\\
&\le\frac{\nu}{8}\|A_h\bar u_h^n\|_0^2\tau+2\nu^{-1}c_0^2\|A_h\bar u_h^{n-1}\|_0\|\bar u_h^{n-1}\|_0\|\bar u_h^{n-1}\|_1^2\tau\\
&\le\frac{\nu}{8}(\|A_h\bar u_h^n\|_0^2+2\|A_h\bar u_h^{n-1}\|_0^2)\tau\\
&\quad+2\nu^{-3}c_0^4\|\bar u_h^{n-1}\|_1^4\|\bar u_h^{n-1}\|_0^2\tau,\\
2|b(\bar u_h^{n-1},d_tu_h^{n-1},A_h\bar u_h^n)|\tau^2&+2|b(d_tu_h^{n-1},\bar u_h^{n-1},A_h\bar u_h^n)|\tau^2\\
&\le 2c_0\gamma_0^{1/2}\|A_h\bar u_h^{n-1}\|_0^{1/2}\|\bar u_h^{n-1}\|_1^{1/2}\|d_tu_h^{n-1}\|_1\|A_h\bar u_h^n\|_0\tau^2\\
&\le\frac{\nu}{8}\|A_h\bar u_h^n\|_0^2\tau\\
&\quad+8\nu^{-1}c_0^2\gamma_0\|A_h\bar u_h^{n-1}\|_0\|\bar u_h^{n-1}\|_1\|d_tu_h^{n-1}\|_1^2\tau^3,
\end{aligned}$$

$$\begin{aligned}\frac{1}{2}|b(d_tu_h^{n-1},d_tu_h^{n-1},A_h\bar{u}_h^n)|\tau^3 &\le \frac{1}{2}c_0\|A_hd_tu_h^{n-1}\|_0^{1/2}\|d_tu_h^{n-1}\|_0^{1/2}\|d_tu_h^{n-1}\|_1\|A_h\bar{u}_h^n\|_0\tau^3\\ &\le \frac{\nu}{8}\|A_h\bar{u}_h^n\|_0^2\tau\\ &\quad+\nu^{-1}c_0^2\|A_h(u_h^{n-1}-u_h^{n-2})\|_0\|d_tu_h^{n-1}\|_0\|d_tu_h^{n-1}\|_1^2\tau^4,\\ 2|(\bar{f}(t_n),A_h\bar{u}_h^n)|\tau &\le \frac{\nu}{8}\|A_h\bar{u}_h^n\|_0^2\tau+8\nu^{-1}\|\bar{f}(t_n)\|_0^2\tau.\end{aligned}$$

Combining these inequalities with (4.23) gives

$$\begin{aligned}&(\|u_h^n\|_1^2-\|u_h^{n-1}\|_1^2+\frac{\nu}{2}(3\|A_h\bar{u}_h^n\|_0^2-\|A_h\bar{u}_h^{n-1}\|_0^2)\tau\\ &\quad\le 2\nu^{-3}c_0^4\|\bar{u}_h^{n-1}\|_1^4\|\bar{u}_h^{n-1}\|_0^2\tau+8\nu^{-1}\|\bar{f}(t_n)\|_0^2\tau\\ &\quad+8\nu^{-1}c_0^2\gamma_0\|A_h\bar{u}_h^{n-1}\|_0\|\bar{u}_h^{n-1}\|_1\|d_tu_h^{n-1}\|_1^2\tau^3\\ &\quad+\nu^{-1}c_0^2(\|A_hu_h^{n-1}\|_0+\|A_hu_h^{n-2}\|_0)\|d_tu_h^{n-1}\|_0\|d_tu_h^{n-1}\|_1^2\tau^4.\end{aligned}\tag{4.24}$$

Summing (4.24) from 2 to J and using (4.16) and the induction assumption with $m=0,1,\dots,J-1$, we obtain

$$\begin{aligned}\|u_h^J\|_1^2+\nu\tau\sum_{n=1}^{J}\|A_h\bar{u}_h^n\|_0^2 &\le \tau\sum_{n=1}^{J-1}d_n\|u_h^n\|_1^2+\frac{3}{2}\nu\|A_h\bar{u}_h^1\|_0^2\tau\\ &\quad+\|u_h^1\|_1^2+8\nu^{-1}T\sup_{0\le t\le T}\|f(t)\|_0^2\\ &\quad+\nu^{-1}\gamma_0^{-1}\kappa_0^{1/2}+\nu^{3/2}c_0^{-2}T\gamma_0^{-4}\end{aligned}\tag{4.25}$$

for all $1\le J\le N$, where

$$d_n=2\nu^{-3}c_0^4(\|\bar{u}_h^n\|_1^2\|\bar{u}_h^n\|_0^2+\|\bar{u}_h^{n+1}\|_1^2\|\bar{u}_h^{n+1}\|_0^2),\quad \bar{u}_h^0=u_h^0.$$

We set

$$\begin{aligned}a_n&=\|u_h^n\|_1^2,\quad b_n=\nu\|A_h\bar{u}_h^n\|_0^2,\\ C&=\|u_h^1\|_1^2+\frac{3}{2}\nu\|A_h\bar{u}_h^1\|_0^2\tau+8\nu^{-1}T\sup_{0\le t\le T}\|f(t)\|_0^2\\ &\quad+\nu^{-1}\gamma_0^{-1}\kappa_0^{1/2}+\nu^{3/2}c_0^{-2}T\gamma_0^{-4}.\end{aligned}$$

Applying Lemma 3.4 to (4.25) and using (4.16)–(4.17) and Lemma 4.1, we obtain (4.18) with $m=J$.

Now, we prove that (4.19) holds for $m=J$. For $n=2$, we deduce from (4.1) and (4.3) that

$$\begin{aligned}(d_{tt}u_h^2,v_h)+\frac{1}{2}a(d_tu_h^2,v_h)-d(v_h,d_tp_h^2)+(d_tu_h^2,q_h)\\ +\frac{3}{2}b(d_tu_h^1,\bar{u}_h^1,v_h)+\frac{3}{2}b(\bar{u}_h^1,d_tu_h^1,v_h)\\ =\frac{1}{2\tau}\int_{t_1}^{t_2}(f_t(t),v_h)dt\,.\end{aligned}\tag{4.26}$$

By taking $v_h=2d_tu_h^2\tau$ and $q_h=2d_tp_h^2\tau$ in (4.26) and using (3.10) and Lemma 3.1, we have

$$\begin{aligned}&\|d_tu_h^2\|_0^2-\|d_tu_h^1\|_0^2+\|d_{tt}u_h^2\|_0^2\tau^2+\frac{\nu}{2}\|d_tu_h^2\|_1^2\tau\\ &\quad\le\nu^{-1}\gamma_0^2\int_{t_1}^{t_2}\|f_t(t)\|_0^2dt+\frac{9}{4}\nu^{-1}c_0^2\gamma_0^2\|A_h\bar{u}_h^1\|_0^2\|d_tu_h^1\|_0^2\tau.\end{aligned}\tag{4.27}$$

Moreover, it follows from (4.3) that for $3 \le n \le J$ there holds

$$\begin{aligned}
&(d_{tt}u_h^n, v_h) + a(d_t\bar{u}_h^n, v_h) - d(v_h, d_tp_h^n) + (d_t\bar{u}_h^n, q_h)\\
&\qquad + \frac{3}{2}b(d_tu_h^{n-1}, \bar{u}_h^{n-1}, v_h) + \frac{3}{2}b(\bar{u}_h^{n-1}, d_tu_h^{n-1}, v_h)\\
&\qquad - \frac{1}{2}b(d_tu_h^{n-2}, \bar{u}_h^{n-2}, v_h) - \frac{1}{2}b(\bar{u}_h^{n-2}, d_tu_h^{n-2}, v_h)\\
&\qquad = \frac{1}{2\tau}\int_{t_{n-2}}^{t_n}(f_t(t), v_h)dt. \qquad (4.28)
\end{aligned}$$

Taking $v_h = 2d_tu_h^n\tau$ in (4.28), using (3.12), and noting $d_tu_h^n = d_t\bar{u}_h^n + \frac{1}{2}d_{tt}u_h^n\tau$ yields

$$\begin{aligned}
&\|d_tu_h^n\|_0^2 - \|d_tu_h^{n-1}\|_0^2 + \|d_{tt}u_h^n\|_0^2\tau^2 + \frac{\nu}{2}(\|d_tu_h^n\|_1^2 - \|d_tu_h^{n-1}\|_1^2 + 4\|d_t\bar{u}_h^n\|_1^2)\tau\\
&\quad + 3b(\bar{u}_h^{n-1}, d_tu_h^{n-1}, d_{tt}u_h^n)\tau^2 + 3b\left(d_tu_h^{n-1}, \bar{u}_h^{n-1}, d_t\bar{u}_h^n + \frac{1}{2}d_{tt}u_h^n\tau\right)\tau\\
&\quad - b\left(d_tu_h^{n-2}, \bar{u}_h^{n-2}, d_t\bar{u}_h^n + \frac{1}{2}d_{tt}u_h^n\tau\right)\tau - b\left(\bar{u}_h^{n-2}, d_tu_h^{n-2}, d_t\bar{u}_h^n + \frac{1}{2}d_{tt}u_h^n\tau\right)\tau\\
&\quad = \left(\int_{t_{n-2}}^{t_n} f_t(t), d_t\bar{u}_h^n + \frac{1}{2}d_{tt}u_h^n\tau\right)dt. \qquad (4.29)
\end{aligned}$$

In view of Lemma 3.1 and (3.10), there hold

$$\begin{aligned}
3|b(d_tu_h^{n-1}, \bar{u}_h^{n-1}, d_t\bar{u}_h^n)|\tau &\le \frac{\nu}{4}\|d_t\bar{u}_h^n\|_1^2\tau + 9\nu^{-1}c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-1}\|_0^2\|d_tu_h^{n-1}\|_0^2\tau,\\
3|b(\bar{u}_h^{n-1}, d_tu_h^{n-1}, d_{tt}u_h^n)|\tau^2 &+ 3|b(d_tu_h^{n-1}, \bar{u}_h^{n-1}, d_{tt}u_h^n)|\tau^2\\
&\le \frac{1}{4}\|d_{tt}u_h^n\|_0^2\tau^2 + 9c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-1}\|_0^2\|d_tu_h^{n-1}\|_1^2\tau^2,\\
|b(d_tu_h^{n-2}, \bar{u}_h^{n-2}, d_t\bar{u}_h^n)|\tau &+ |b(\bar{u}_h^{n-2}, d_tu_h^{n-2}, d_t\bar{u}_h^n)|\tau\\
&\le \frac{\nu}{4}\|d_tu_h^n\|_1^2\tau + \nu^{-1}c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-2}\|_0^2\|d_tu_h^{n-2}\|_0^2\tau,\\
|b(d_tu_h^{n-2}, \bar{u}_h^{n-2}, d_{tt}u_h^n)|\tau^2 &+ |b(\bar{u}_h^{n-2}, d_tu_h^{n-2}, d_{tt}u_h^n)|\tau^2\\
&\le \frac{1}{4}\|d_{tt}u_h^n\|_0^2\tau^2 + c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-2}\|_0^2\|d_tu_h^{n-2}\|_1^2\tau^2,\\
\left|\int_{t_{n-2}}^{t_n}(f_t(t), d_t\bar{u}_h^n + \frac{1}{2}d_{tt}u_h^n\tau)dt\right| &\le \frac{\nu}{4}\|d_t\bar{u}_h^n\|_1^2\tau + \frac{1}{4}\|d_{tt}u_h^n\|_0^2\tau^2\\
&\quad + 2(\nu^{-1}\gamma_0^2 + \tau)\int_{t_{n-2}}^{t_n}\|f_t(t)\|_0^2dt.
\end{aligned}$$

Combining these inequalities with (4.29) and using Lemma 4.1 yields

$$\begin{aligned}
&\left(\|d_tu_h^n\|_0^2 + \frac{\nu}{2}\|d_tu_h^n\|_1^2\tau\right) - \left(\|d_tu_h^{n-1}\|_0^2 + \frac{\nu}{2}\|d_tu_h^{n-1}\|_1^2\tau\right)\\
&\quad \le 9\nu^{-1}c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-1}\|_0^2\|d_tu_h^{n-1}\|_0^2\tau + \nu^{-1}c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-2}\|_0^2\|d_tu_h^{n-2}\|_0^2\tau\\
&\quad + 9c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-1}\|_0^2\|d_tu_h^{n-1}\|_1^2\tau^2 + c_0^2\gamma_0^2\|A_h\bar{u}_h^{n-2}\|_0^2\|d_tu_h^{n-2}\|_1^2\tau^2\\
&\quad + 2(\nu^{-1}\gamma_0^2 + \tau)\int_{t_{n-2}}^{t_n}\|f_t(t)\|_0^2dt. \qquad (4.30)
\end{aligned}$$

Summing this inequality from 3 to J, adding (4.27) in, and using the induction assumption with $m = 1, \ldots, J-1$ and (4.17)–(4.18), we arrive at

$$\begin{aligned}\|d_t u_h^J\|_0^2 + \frac{\nu}{2}\|d_t u_h^J\|_1^2\tau \le \tau\sum_{n=0}^{J-1} d_n\left(\|d_t u_h^n\|_0^2 + \frac{\nu}{2}\|d_t u_h^n\|_1^2\tau\right)\\ + \|d_t u_h^1\|_0^2 + \frac{\nu}{2}\|d_t u_h^1\|_1^2\tau + 2(\nu^{-1}\gamma_0^2+\tau)\int_0^T \|f_t(t)\|_0^2 dt \end{aligned} \tag{4.31}$$

for $1 \le J \le N$, where

$$d_n = 13\nu^{-1}c_0^2\gamma_0^2\|A_h\bar{u}_h^n\|_0^2, \quad 1 \le n \le J-1, \quad d_0 = 0.$$

We set

$$\begin{aligned} a_n &= \|d_t u_h^n\|_0^2 + \frac{\nu}{2}\|d_t u_h^n\|_1^2\tau, \quad b_n = 0, \quad c_n = 0,\\ C &= \|d_t u_h^1\|_0^2 + \frac{\nu}{2}\|d_t u_h^1\|_1^2\tau + 2(\nu^{-1}\gamma_0^2+\tau)\int_0^T \|f_t(t)\|_0^2 dt. \end{aligned}$$

Applying Lemma 3.4 to (4.31) and using (4.16)–(4.18) and Lemma 4.1, we obtain

$$\|d_t u_h^J\|_0^2 + \frac{\nu}{2}\|d_t u_h^J\|_1^2\tau \le \exp(13\nu^{-2}c_0^2\gamma_0^2\kappa_1)C. \tag{4.32}$$

Next, by taking $v_h = 2A_h d_t u_h^n\tau \in V_h$ and $q_h = 0$ in (4.4), we obtain

$$\begin{aligned}&2\|d_t u_h^n\|_1^2\tau + \nu\|A_h u_h^n\|_0^2 - \nu\|A_h u_h^{n-1}\|_0^2 + 2b(\bar{u}_h^{n-1}, \bar{u}_h^{n-1}, A_h d_t u_h^n)\tau\\ &\quad + 2b(\bar{u}_h^{n-1}, d_t u_h^{n-1}, A_h d_t u_h^n)\tau^2 + 2b(d_t u_h^{n-1}, \bar{u}_h^{n-1}, A_h d_t u_h^n)\tau^2\\ &\quad + \frac{1}{2}b(d_t u_h^{n-1}, d_t u_h^{n-1}, A_h d_t u_h^n)\tau^3 = 2(\bar{f}(t_n), A_h d_t u_h^n)\tau. \end{aligned} \tag{4.33}$$

In view of Lemma 3.1 and (3.10), there hold

$$\begin{aligned} 2|b(\bar{u}_h^{n-1}, \bar{u}_h^{n-1}, A_h d_t u_h^n)|\tau &\le c_0\gamma_0\|A_h\bar{u}_h^{n-1}\|_0^2\|d_t u_h^n\|_1\tau\\ &\le \frac{1}{8}\|d_t u_h^n\|_1^2\tau + c\|A_h\bar{u}_h^{n-1}\|_0^4\tau\\ 2|b(\bar{u}_h^{n-1}, d_t u_h^{n-1}, A_h d_t u_h^n)|\tau^2 &+ 2|b(d_t u_h^{n-1}, \bar{u}_h^{n-1}, A_h d_t u_h^n)|\tau^2\\ &\le 2c_0\gamma_0\|A_h\bar{u}_h^{n-1}\|_0\|A_h d_t u_h^{n-1}\|_0\|d_t u_h^n\|_1\tau^2\\ &\le \frac{1}{8}\|d_t u_h^n\|_1^2\tau + c\|A_h\bar{u}_h^{n-1}\|_0^2\|A_h u_h^{n-1} - A_h u_h^{n-2}\|_0^2\tau,\\ \frac{1}{2}|b(d_t u_h^{n-1}, d_t u_h^{n-1}, A_h d_t u_h^n)|\tau^3 &\le c\|d_t u_h^{n-1}\|_0^{1/2}\|A_h d_t u_h^{n-1}\|_0^{3/2}\|d_t u_h^n\|_1\tau^3\\ &\quad + c\|d_t u_h^{n-1}\|_1\|A_h d_t u_h^{n-1}\|_0\|d_t u_h^n\|_1\tau^3\\ &\le \frac{1}{8}\|d_t u_h^n\|_1^2\tau\\ &\quad + c(\|d_t u_h^{n-1}\|_1^2\tau^2 + \|d_t u_h^{n-1}\|_0\|A_h d_t u_h^{n-1}\|_0\tau^2)\\ &\quad \times \|A_h(u_h^{n-1} - u_h^{n-2})\|_0^2\tau,\\ 2|(\bar{f}(t_n), A_h d_t u_h^n)|\tau &\le \frac{1}{8}\|d_t u_h^n\|_1^2\tau + c\|\nabla\bar{f}(t_n)\|_0^2\tau. \end{aligned}$$

Combining these inequalities with (4.33) gives

$$\begin{aligned}&\nu\|A_h u_h^n\|_0^2 - \nu\|A_h u_h^{n-1}\|_0^2 + \|d_t u_h^n\|_1^2\tau\\ &\quad\le c(\|A_h\bar{u}_h^{n-1}\|_0^2 + c\|d_t u_h^{n-1}\|_1^2\tau^2 + \|d_t u_h^{n-1}\|_0\|A_h d_t u_h^{n-1}\|_0\tau^2)\\ &\quad\times(\|A_h u_h^{n-1}\|_0^2 + \|A_h u_h^{n-2}\|_0^2)\tau + c\|\nabla\bar{f}(t_n)\|_0^2\tau.\end{aligned}\tag{4.34}$$

Summing (4.34) from 2 to J and using (4.16)–(4.32) and the induction assumption with $m = 0, 1, \ldots, J-1$, we obtain

$$\begin{aligned}\nu\|A_h u_h^J\|_1^2 + \tau\sum_{n=1}^{J}\|d_t u_h^n\|_1^2 &\le \nu\tau\sum_{n=1}^{J-1} d_n\|A_h u_h^n\|_1^2\\ &\quad + \|A_h u_h^1\|_0^2 + c\sup_{0\le t\le T}\|\nabla f(t)\|_0^2\end{aligned}\tag{4.35}$$

for all $1 \le J \le N$, where

$$\begin{aligned}d_n &= c(\|A_h\bar{u}_h^{n+1}\|_0^2 + c\|d_t u_h^{n+1}\|_1^2\tau^2 + \|d_t u_h^{n+1}\|_0\|A_h d_t u_h^{n+1}\|_0\tau^2)\\ &\quad + c(\|A_h\bar{u}_h^{n}\|_0^2 + c\|d_t u_h^{n}\|_1^2\tau^2 + \|d_t u_h^{n}\|_0\|A_h d_t u_h^{n}\|_0\tau^2).\end{aligned}$$

Using (4.16)–(4.18), (4.32), and the induction assumption with $m = 0, 1, \ldots, J-1$, we obtain

$$\tau\sum^{J-1} d_n \le \kappa, \quad \kappa \text{ is independent of } \kappa_2.\tag{4.36}$$

We set

$$\begin{aligned}a_n &= \nu\|A_h u_h^n\|_0^2, \quad b_n = \|d_t u_h^n\|_1^2,\\ C &= \nu\|A_h u_h^1\|_1^2 + c\nu^{-1}T\sup_{0\le t\le T}\|\nabla f(t)\|_0^2.\end{aligned}$$

Applying the discrete Gronwall lemma to (4.35) and using (4.36), we obtain

$$\nu\|A_h u_h^J\|_1^2 + \tau\sum_{n=1}^{J}\|d_t u_h^n\|_1^2 \le \kappa, \quad \kappa \text{ is independent of } \kappa_2.\tag{4.37}$$

Finally, using (4.3), (3.4), (3.10), and Lemma 3.1, we arrive at

$$\begin{aligned}\|p_h^J\|_0 &= \beta^{-1}\sup_{v_h\in X_h}\frac{d(v_h, p_h^J)}{\|v_h\|_1}\\ &\le \beta^{-1}\left(\gamma_0\|d_t u_h^J\|_0 + \nu\|\bar{u}_h^J\|_1 + \gamma_0\|\bar{f}(t_n)\|_0 + \frac{3}{2}c_0\gamma_0\|u_h^{J-1}\|_1^2 + \frac{1}{2}c_0\gamma_0\|u_h^{J-2}\|_1^2\right),\end{aligned}$$

which yields

$$\begin{aligned}\|p_h^J\|_0^2 &\le 5\beta^{-2}\left(\gamma_0^2\|d_t u_h^J\|_0^2 + \nu^2\gamma_0^2\|A_h\bar{u}_h^J\|_0^2 + \gamma_0^2\sup_{0\le t\le T}\|f(t)\|_0^2\right)\\ &\quad + 5\beta^{-2}\left(\frac{9}{4}c_0^2\gamma_0^2\|u_h^{J-1}\|_1^4 + \frac{1}{4}c_0^2\gamma_0^2\|u_h^{J-2}\|_1^4\right).\end{aligned}\tag{4.38}$$

Combining (4.37) and (4.38) with (4.32), we get (4.19) with $m = J$. So, we have completed the proof of Theorem 4.2. □

LEMMA 4.3. *Suppose that the assumptions* (A1)–(A2) *are valid and the couple* (X_h, M_h) *satisfies the approximate properties* (3.2)–(3.4). *Then there hold*

$$\|e^1\|_\alpha^2 + \|d_t e^1\|_\alpha^2 \tau^2 + \nu \|e^1\|_{\alpha+1}^2 \tau \le \kappa \tau^{2-\alpha}, \quad \alpha = -2, \ -1, \ 0, \ 1, \tag{4.39}$$

$$\|\eta^1\|_0^2 \le \kappa, \tag{4.40}$$

where

$$e^0 = 0, \quad e^n = u_h(t_n) - u_h^n, \quad \eta^n = \frac{1}{\tau}\int_{t_{n-1}}^{t_n} p_h(t)dt - p_h^n, \quad n = 1, \ldots, N.$$

Proof. By integrating (3.8) from t_0 to t_1,

$$\begin{aligned}(d_t u_h(t_1), v_h) + \frac{1}{\tau}\int_{t_0}^{t_1} a(u_h(t), v_h)dt - \frac{1}{\tau}\int_{t_0}^{t_1} d(v_h, p_h(t))dt \\ + \frac{1}{\tau}\int_{t_0}^{t_1} b(u_h(t), u_h(t), v_h)dt = \frac{1}{\tau}\int_{t_0}^{t_1} (f(t), v_h)dt.\end{aligned} \tag{4.41}$$

Subtracting (4.1) from (4.41) and using the integral formula by parts, we obtain

$$(d_t e^1, v_h) + a(e^1, v_h) - d(v_h, \eta^1) = (E_1, v_h), \tag{4.42}$$

where

$$\begin{aligned}(E_1, v_h) = -\frac{1}{\tau}\int_{t_0}^{t_1} (t - t_0)(f_t(t), v_h)dt + \frac{1}{\tau}\int_{t_0}^{t_1} (t - t_0) a(u_{ht}(t), v_h)dt \\ + \frac{1}{\tau}\int_{t_0}^{t_1} (t - t_1) b_t(u_h(t), u_h(t), v_h)dt\end{aligned} \tag{4.43}$$

and

$$b_t(u_h(t), u_h(t), v_h) = b(u_{ht}(t), u_h(t), v_h) + b(u_h(t), u_{ht}(t), v_h).$$

Hereafter, we need the L^2-orthogonal projection $R_h : Y \to X_h$ defined by

$$(R_h v, v_h) = (v, v_h) \ \ \forall v \in Y, \quad v_h \in X_h.$$

We see from (3.10), (4.43), Lemma 3.1, and Theorem 3.2 that

$$\begin{aligned}\|A_h^{\frac{\alpha-1}{2}} P_h E_1\|_0 &= \sup_{v_h \in V_h} \frac{(E_1, v_h)}{\|A_h^{\frac{1-\alpha}{2}} v_h\|_0} \\ &\le \left(\gamma_0^{1-\alpha}\tau \sup_{0\le t\le t_1} \|f_t(t)\|_0 + \nu\tau^{\frac{1-\alpha}{2}} \sup_{0\le t\le t_1} \sigma^{\frac{\alpha+1}{2}}(t)\|u_{ht}(t)\|_{\alpha+1}\right) \\ &+ c_0\gamma_0^{1-\alpha}\tau^{1-\frac{\alpha(\alpha+1)}{2}} \sup_{0\le t\le t_1} \sigma^{\frac{\alpha(\alpha+1)}{2}}(t)\|A_h u_h(t)\|_0\|u_{ht}(t)\|_0 \le \kappa\tau^{\frac{1-\alpha}{2}}\end{aligned} \tag{4.44}$$

for $\alpha = -1, \ 0, \ 1$ and

$$\begin{aligned}\|R_h E_1\|_0 = \sup_{v_h \in X_h} \frac{(E_1, v_h)}{\|v_h\|_0} \le \tau \sup_{0\le t\le t_1} \|f_t(t)\|_0 + \nu \sup_{0\le t\le t_1} \sigma(t)\|A_h u_{ht}(t)\|_0 \\ + c_0\gamma_0\|A_h u_h(t_0)\|_0\|u_h(t_0)\|_1 + c_0\gamma_0 \sup_{0\le t\le t_1} \|A_h u_h(t)\|_0\|u_h(t)\|_1 \le \kappa.\end{aligned} \tag{4.45}$$

Taking $v_h = 2A_h^{\alpha} e^1 \tau \in V_h$ in (4.42), we have

$$\|e^1\|_{\alpha}^2 + \|d_t e^1\|_{\alpha}^2 \tau^2 + \nu \|e^1\|_{\alpha+1}^2 \tau \le \nu^{-1} \|A_h^{\frac{\alpha-1}{2}} P_h E_1\|_0^2 \tau \le \kappa \tau^{2-\alpha}. \tag{4.46}$$

Then, we take $v_h = 2A_h^{-2} e^1 \tau \in V_h$ in (4.42) and use (4.44) with $\alpha = -1$ to obtain

$$\frac{1}{2}\|e^1\|_{-2}^2 + \|d_t e^1\|_{-2}^2 \tau^2 + \nu \|e^1\|_{-1}^2 \tau \le 2\|A_h^{-1} P_h E_1\|_0^2 \tau^2 \le \kappa \tau^4. \tag{4.47}$$

Finally, using (4.42), (3.4), (3.10), and Lemma 3.1, we arrive at

$$\begin{aligned} \|\eta^1\|_0 &\le \beta^{-1} \sup_{v_h \in X_h} \frac{d(v_h, \eta^1)}{\|v_h\|_1} \\ &\le \beta^{-1}(\gamma_0 \|d_t e^1\|_0 + \nu \|e^1\|_1 + \gamma_0 \|R_h E_1\|_0), \end{aligned}$$

which gives

$$\|\eta^1\|_0^2 \le 3\beta^{-2}(\|d_t e^1\|_0^2 + \nu^2 \|e^1\|_1^2 + \|P_h E_1\|_{-1}^2) \le \kappa. \tag{4.48}$$

Combining (4.48) with (4.46)–(4.47) has completed the proof of Lemma 4.3. ∎

5. Second-order dual scheme: Stability analysis. In order to derive the L^2-bound on the error $u_h(t_n) - u_h^n$, we employ a parabolic argument that has already been used in [16] for the Crank–Nicolson scheme of the time-dependent Navier–Stokes equation. Let $t_m \in (0, t]$ be given. We consider the linearized "backward" counterpart of the discrete Navier–Stokes equations (4.3): for $\xi^n \in V_h$, $2 \le n \le m$, find $\Phi_h^{n-1} \in V_h$ such that

$$\begin{aligned} &(v_h, d_t \Phi_h^n) - a(v_h, \bar{\Phi}_h^n) - b\left(\frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}, v_h, \Phi_h^n\right) \\ &\qquad - b\left(v_h, \frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}, \Phi_h^n\right) = (v_h, \bar{\xi}^n), \quad v_h \in V_h, \end{aligned} \tag{5.1}$$

with an initial value $\Phi_h^m \in V_h$.

Here, we need to introduce the following discrete dual Gronwall lemma provided in [10].

LEMMA 5.1. *Let C and a_n, b_n, d_n, h_n for integers $n_0 \le n \le m$ be nonnegative numbers such that*

$$a_k + \tau \sum_{n=k}^{m} b_n \le \tau \sum_{n=k}^{m} d_n a_n + C, \quad n_0 \le k \le m \tag{5.2}$$

and $d_n \tau < 1$ for all $n_0 \le n \le m$; then

$$a_k + \tau \sum_{n=k}^{m} b_n \le C \exp\left(\tau \sum_{n=k}^{m} (1 - d_n \tau)^{-1} d_n\right), \quad n_0 \le k \le m. \tag{5.3}$$

The following lemma provides the stability of the scheme (5.1).

LEMMA 5.2. *Under the assumptions of Theorem 4.2, the following a priori estimates hold:*

$$\|\Phi_h^k\|_0^2 + \tau \sum_{n=k+1}^{m} (\nu \|\bar{\Phi}_h^n\|_1^2 + \|d_t \Phi_h^n\|_{-1}^2) \le \kappa \left(\|\Phi_h^m\|_0^2 + \tau \sum_{n=2}^{m} \|\bar{\xi}^n\|_{-1}^2\right), \tag{5.4}$$

$$\|\Phi_h^k\|_1^2 + \nu\tau \sum_{n=k+1}^{m} \|A_h \bar{\Phi}_h^n\|_0^2 \le \kappa \left(\|\Phi_h^m\|_1^2 + \tau \sum_{n=2}^{m} \|\bar{\xi}^n\|_0^2\right) \tag{5.5}$$

for all $1 \le k \le m$.

Proof. The proof follows the line of argument used in the proofs of Theorem 4.2. First, taking $v_h = -2\bar{\Phi}_h^n\tau$ in (5.1) and using (3.12), we get

$$\begin{aligned}&\|\Phi_h^{n-1}\|_0^2 - \|\Phi_h^n\|_0^2 + 2\nu\|\bar{\Phi}_h^n\|_1^2\tau + b(\bar{\Phi}_h^n, 3u_h^{n-1} - u_h^{n-2}, \bar{\Phi}_h^n)\tau\\ &\qquad \le \frac{\nu}{4}\|\bar{\Phi}_h^n\|_1^2\tau + 4\nu^{-1}\|\bar{\xi}^n\|_{-1}^2\tau. \end{aligned}\tag{5.6}$$

It follows from Lemma 3.1 and (3.10) that

$$\begin{aligned}|b(\bar{\Phi}_h^n, 3u_h^{n-1} - u_h^{n-2}, \bar{\Phi}_h^n)|\tau &\le c_0\gamma_0\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0\|\bar{\Phi}_h^n\|_1\|\bar{\Phi}_h^n\|_0\\ &\le \frac{3}{4}\nu\|\bar{\Phi}_h^n\|_1^2\tau + \frac{2}{3}\nu^{-1}c_0^2\gamma_0^2\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2(\|\Phi_h^n\|_0^2 + \|\Phi_h^{n-1}\|_0^2)\tau.\end{aligned}$$

Combining (5.6) with the above estimate gives

$$\begin{aligned}&\|\Phi_h^{n-1}\|_0^2 - \|\Phi_h^n\|_0^2 + \nu\|\bar{\Phi}_h^n\|_1^2\tau\\ &\qquad \le \frac{2}{3}\nu^{-1}c_0^2\gamma_0^2\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2(\|\Phi_h^n\|_0^2 + \|\Phi_h^{n-1}\|_0^2)\tau + 4\nu^{-1}\|\xi^n\|_{-1}^2\tau\end{aligned}$$

for all $2 \le n \le m$. Summing the above inequality from $k+1$ to m, we obtain

$$\begin{aligned}&\|\Phi_h^k\|_0^2 + \nu\tau\sum_{n=k+1}^{m}\|\bar{\Phi}_h^n\|_1^2\\ &\qquad \le \tau\sum_{n=k}^{m} d_n\|\Phi_h^n\|_0^2 + \|\Phi_h^m\|_0^2 + 4\nu^{-1}\tau\sum_{n=2}^{m}\|\bar{\xi}^n\|_{-1}^2, \quad 1 \le k \le m.\end{aligned}\tag{5.7}$$

Let

$$\begin{aligned}a_n &= \|\Phi_h^n\|_0^2, \quad b_n = \nu\|\bar{\Phi}_h^n\|_1^2, \quad C = \|\Phi_h^m\|_0^2 + 4\nu^{-1}\tau\sum_{n=2}^{m}\|\bar{\xi}^n\|_{-1}^2,\\ d_n &= \frac{2}{3}\nu^{-1}c_0^2\gamma_0^2(\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2 + \|A_h(3u_h^n - u_h^{n-1})\|_0^2).\end{aligned}$$

Hence, by using Theorem 4.2 and (4.16), we find

$$d_n\tau \le 27\nu^{-2}c_0^2\gamma_0^2\kappa_2\tau \le \frac{7}{11}, \quad (1 - d_n\tau)^{-1} \le \frac{11}{4}, \quad 1 \le n \le m.\tag{5.8}$$

Then, applying Lemma 5.1 to (5.7) and using Theorem 4.2 to obtain

$$\|\Phi_h^k\|_0^2 + \nu\tau\sum_{n=k+1}^{m}\|\bar{\Phi}_h^n\|_1^2 \le C\exp\left(\frac{11}{4}\tau\sum_{n=k}^{m} d_n\right) \le \kappa\left(\|\Phi_h^m\|_0^2 + \tau\sum_{n=2}^{m}\|\bar{\xi}^n\|_{-1}^2\right)\tag{5.9}$$

for all $1 \le k \le m$.

Moreover, we obtain from (3.10), (5.1), and Lemma 3.1 that

$$\|d_t\Phi_h^n\|_{-1} \le \nu\|\bar{\Phi}_h^n\|_1 + \frac{1}{2}c_0\gamma_0\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0\|\bar{\Phi}_h^n\|_1 + \|\bar{\xi}^n\|_{-1}, \quad 2 \le n \le m,$$

and by Theorem 4.2 that

$$\begin{aligned}\tau\sum_{n=k+1}^{m}\|d_t\Phi_h^n\|_{-1}^2 &\le 3\tau\sum_{n=k+1}^{m}(\nu^2\|\bar{\Phi}_h^n\|_1^2 + c_0^2\gamma_0^2\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2\|\bar{\Phi}_h^n\|_1^2 + \|\bar{\xi}^n\|_{-1}^2)\\ &\le \kappa\tau\sum_{n=k+1}^{m}(\|\bar{\Phi}_h^n\|_1^2 + \|\bar{\xi}^n\|_{-1}^2), \quad 1 \le k \le m.\end{aligned}\tag{5.10}$$

Combining this inequality with (5.9) yields (5.4).

Furthermore, by taking $v_h = -2A_h\bar{\Phi}_h^n\tau$ in (5.1), we obtain

$$\begin{aligned}&\|\Phi_h^{n-1}\|_1^2 - \|\Phi_h^n\|_1^2 + 2\nu\|A_h\bar{\Phi}_h^n\|_0^2\tau\\&\quad + b(A_h\bar{\Phi}_h^n, 3u_h^{n-1} - u_h^{n-2}, \bar{\Phi}_h^n)\tau + b(3u_h^{n-1} - u_h^{n-2}, A_h\bar{\Phi}_h^n, \bar{\Phi}_h^n)\tau\\&\quad \le \frac{\nu}{4}\|A_h\Phi_h^{n-1}\|_0^2\tau + \frac{4}{\nu}\|\bar{\xi}^n\|_0^2\tau.\end{aligned} \tag{5.11}$$

From Lemma 3.1 and (3.10), we have

$$\begin{aligned}&|b(A_h\bar{\Phi}_h^n, 3u_h^{n-1} - u_h^{n-2}, \bar{\Phi}_h^n)|\tau + |b(3u_h^{n-1} - u_h^{n-2}, A_h\bar{\Phi}_h^n, \bar{\Phi}_h^n)|\tau\\&\qquad \le c_0\gamma_0\|\bar{\Phi}_h^n\|_1\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0\|A_h\bar{\Phi}_h^n\|_0\tau\\&\qquad \le \frac{\nu}{4}\|A_h\bar{\Phi}_h^n\|_0^2\tau + \nu^{-1}c_0^2\gamma_0^2\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2\|\bar{\Phi}_h^n\|_1^2\tau.\end{aligned}$$

Combining (5.11) with the above estimate gives

$$\begin{aligned}&\|\Phi_h^{n-1}\|_1^2 - \|\Phi_h^n\|_1^2 + \nu\|A_h\bar{\Phi}_h^n\|_0^2\tau\\&\qquad \le \nu^{-1}c_0^2\gamma_0^2\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2\|\bar{\Phi}_h^n\|_1^2\tau + 4\nu^{-1}\|\xi^n\|_0^2\tau\end{aligned} \tag{5.12}$$

for all $2 \le n \le m$. Summing (5.12) from $k+1$ to m and using (3.10), (5.5), and Theorem 4.2, we obtain (5.5). ∎

LEMMA 5.3. *Under the assumptions of Theorem* 4.2*, the following a priori estimate holds:*

$$\nu\|A_h\Phi_h^1\|_0^2 + \tau\sum_{n=2}^{m}\|d_t\Phi_h^n\|_1^2 \le \kappa\left(\nu\|A_h\Phi_h^m\|_0^2 + \tau\sum_{n=2}^{m}\|\bar{\xi}^n\|_1^2\right). \tag{5.13}$$

Proof. To prove (5.13), we take $v_h = 2A_hd_t\Phi_h^n\tau$ in (5.1) and obtain

$$\begin{aligned}&\nu(\|A_h\Phi_h^{n-1}\|_2^2 - \|A_h\Phi_h^n\|_0^2) + 2\|d_t\Phi_h^n\|_1^2\tau\\&\quad - b(A_hd_t\Phi_h^n, 3u_h^{n-1} - u_h^{n-2}, \bar{\Phi}_h^n)\tau - b(3u_h^{n-1} - u_h^{n-2}, A_hd_t\Phi_h^n, \bar{\Phi}_h^n)\tau\\&\quad \le \frac{1}{4}\|d_t\Phi_h^n\|_1^2\tau + 4\|\bar{\xi}^n\|_1^2\tau.\end{aligned} \tag{5.14}$$

It follows from Lemma 3.1 and (3.10) that

$$\begin{aligned}&|b(3u_h^{n-1} - u_h^{n-2}, A_hd_t\Phi_h^n, \bar{\Phi}_h^n)|\tau + |b(d_t\Phi_h^n, 3u_h^{n-1} - u_h^{n-2}, A_h\bar{\Phi}_h^n)|\tau\\&\qquad \le c_0\gamma_0\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0\|A_h\bar{\Phi}_h^n\|_1\|d_t\Phi_h^n\|_1\\&\qquad \le \frac{1}{4}\|d_t\Phi_h^n\|_1^2\tau + c_0^2\gamma_0^2\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2\|A_h\bar{\Phi}_h^n\|_0^2\end{aligned}$$

and by (5.14) that

$$\begin{aligned}&\nu(\|A_h\Phi_h^{n-1}\|_2^2 - \|A_h\Phi_h^n\|_0^2) + \|d_t\Phi_h^n\|_1^2\tau\\&\qquad \le c_0\gamma_0\|A_h(3u_h^{n-1} - u_h^{n-2})\|_0^2\|A_h\bar{\Phi}_h^n\|_0^2\tau + 4\|\bar{\xi}^n\|_1^2\tau\end{aligned} \tag{5.15}$$

for all $2 \le n \le m$. Summing (5.15) from 2 to m and using Theorem 4.2, Lemma 5.2, and (3.10), we obtain (5.13). ∎

6. Error analysis. In this section, we establish the H^1- and L^2-bound of the error $e^n = u_h(t_n) - u_h^n$ and the L^2-bound of the error $\eta^n = \frac{1}{\tau}\int_{t_{n-1}}^{t_n} p_h(t)dt - p_h^n$ for all $1 \le n \le N$. To do this, we integrate (3.8) from t_{n-1} to t_n. By noting that $u_h(t) \in V_h, \quad t \in [0,T]$, we obtain

$$\begin{aligned}(d_t u_h(t_n), v_h) + \frac{1}{\tau}\int_{t_{n-1}}^{t_n} a(u_h(t), v_h)dt - \frac{1}{\tau}\int_{t_{n-1}}^{t_n} d(v_h, p_h(t))dt + d(\bar{u}_h(t_n), q_h)\\ + \frac{1}{\tau}\int_{t_{n-1}}^{t_n} b(u_h(t), u_h(t), v_h)dt = \frac{1}{\tau}\int_{t_{n-1}}^{t_n} (f(t), v_h)dt.\end{aligned} \tag{6.1}$$

Subtracting (4.3) from (6.1) and using the integral formula

$$\bar{\phi}(t_n) - \frac{1}{\tau}\int_{t_{n-1}}^{t_n} \phi(t)dt = \frac{1}{2\tau}\int_{t_{n-1}}^{t_n} (t - t_{n-1})(t_n - t)\phi_{tt}(t)dt \tag{6.2}$$

for all $\phi \in L^2(t_{n-1}, t_n; H^2(t_{n-1}, t_n))$, we obtain

$$\begin{aligned}(d_t e^n, v_h) + a(\bar{e}^n, v_h) - d(v_h, \eta^n) + d(\bar{e}^n, q_h) + \frac{3}{2}b(e^{n-1}, u_h(t_{n-1}), v_h)\\ + \frac{3}{2}b(u_h^{n-1}, e^{n-1}, v_h) - \frac{1}{2}b(e^{n-2}, u_h(t_{n-2}), v_h) - \frac{1}{2}b(u_h^{n-2}, e^{n-2}, v_h)\\ = (E_n, v_h),\end{aligned} \tag{6.3}$$

with

$$\begin{aligned}(E_n, v_h) = \frac{1}{2\tau}\int_{t_{n-1}}^{t_n} (t - t_{n-1})(t_n - t)(f_{tt}(t), v_h)dt\\ + \frac{1}{\tau}\int_{t_{n-1}}^{t_n} (t - t_{n-1})(t_n - t)a(u_{htt}(t), v_h)dt\\ + \frac{1}{2\tau}\int_{t_{n-1}}^{t_n} (t - t_{n-1})(t_n - t)b_{tt}(u_h(t), u_h(t), v_h)dt\\ + \frac{1}{2}\int_{t_{n-1}}^{t_n} (t - t_n)b_{tt}(u_h(t), u_h(t), v_h)dt\\ - \frac{1}{2}\int_{t_{n-2}}^{t_{n-1}} (t - t_{n-2})b_{tt}(u_h(t), u_h(t), v_h)dt,\end{aligned} \tag{6.4}$$

where

$$b_{tt}(u_h(t), u_h(t), v_h) = b(u_{htt}(t), u_h(t), v_h) + b(u_h(t), u_{htt}(t), v_h) + 2b(u_{ht}, u_{ht}, v_h).$$

In order to derive a bound on the error e^n, we need to provide the following estimates on E_n.

LEMMA 6.1. *Under the assumptions of Theorem* 4.2, *the error* E_n *satisfies the following bounds:*

$$\tau\sum_{n=2}^{m} \sigma^i(t_n)\|A_h^{-1}P_h E_n\|_0^2 \le \kappa\ \tau^{3+i}, \quad i = 0, 1, \tag{6.5}$$

$$\tau\sum_{n=2}^{m} \sigma^i(t_n)\|A_h^{-1/2}P_h E_h\|_0^2 \le \kappa\tau^{2+i}, \quad i = 0, 1, 2, \tag{6.6}$$

$$\tau\sum_{n=2}^{m}\sigma^i(t_n)\|P_hE_n\|_0^2\le\kappa\tau^{1+i},\quad i=0,1,2,3, \tag{6.7}$$

$$\sigma^2(t_n)\|R_hE_n\|_0^2\le\kappa\tau^2, \tag{6.8}$$

for all $2\le m\le N$.

Proof. In view of (3.10) and Lemma 3.1, we deduce from (6.4) that

$$\begin{aligned}\|A_h^{(\alpha-1)/2}P_hE_n\|_0&=\sup_{v_h\in V_h}\frac{|(E_n,v_h)|}{\|A_h^{(1-\alpha)/2}v_h\|_0}\le\frac{1}{2}\gamma_0^{1-\alpha}\tau^{3/2}\left(\int_{t_{n-1}}^{t_n}\|f_{tt}(t)\|_0^2dt\right)^{1/2}\\&+\frac{\nu}{2}\tau^{-1/2}\left(\int_{t_{n-1}}^{t_n}(t_n-t)^2(t-t_{n-1})^2\|u_{htt}(t)\|_{\alpha+1}^2dt\right)^{1/2}\\&+c_0\gamma_0\tau^{1/2}\left(\int_{t_{n-1}}^{t_n}(t_n-t)^2(\|A_hu_h(t)\|_0^2\|u_{htt}(t)\|_\alpha^2+\|u_{ht}\|_1^2\|u_{ht}\|_{\alpha+1}^2)dt\right)^{1/2}\\&+c_0\gamma_0\tau^{1/2}\left(\int_{t_{n-2}}^{t_{n-1}}(t-t_{n-2})^2(\|A_hu_h(t)\|_0^2\|u_{htt}(t)\|_\alpha^2+\|u_{ht}\|_1^2\|u_{ht}\|_{\alpha+1}^2)dt\right)^{1/2},\end{aligned} \tag{6.9}$$

for $\alpha=-1,\ 0,\ 1$. In view of Theorems 3.2 and 3.3, (6.9), and the inequalities

$$t-t_{n-1}\le\sigma(t),\quad t\in[t_{n-1},t_n],\quad t-t_{n-2}\le\sigma(t),\quad t\in[t_{n-2},t_{n-1}], \tag{6.10}$$

$$\sigma(t_n)\le\sigma(t_{n-1})+\tau\le2\sigma(t),\quad t\in[t_{n-1},t_n], \tag{6.11}$$

$$\sigma(t_n)(t-t_{n-2})\le(\sigma(t_{n-2})+2\tau)(t-t_{n-2})\le3\sigma(t)\tau,\quad t\in[t_{n-2},t_{n-1}], \tag{6.12}$$

we have the estimates

$$\begin{aligned}\sigma^i(t_n)\|A_h^{-1}P_hE_n\|_0^2\tau&\le2\tau^{3+i}\int_{t_{n-1}}^{t_n}(\gamma_0^4\|f_{tt}\|_0^2+\nu^2\sigma(t)\|u_{htt}\|_0^2)dt\\&+4c_0^2\gamma_0^2\tau^{3+i}\int_{t_{n-2}}^{t_n}(\|A_hu_h(t)\|_0^2\|u_{htt}(t)\|_{-1}^2+\|u_{ht}\|_1^2\|u_{ht}\|_0^2)dt\end{aligned} \tag{6.13}$$

for $i=0,1$,

$$\begin{aligned}\sigma^i(t_n)\|A_h^{-1/2}P_hE_n\|_0^2\tau&\le4\tau^{2+i}\int_{t_{n-1}}^{t_n}(\gamma_0^2\|f_{tt}\|_0^2+\nu^2\sigma^2(t)\|u_{htt}\|_1^2)dt\\&+8c_0^2\gamma_0^2\tau^{2+i}\int_{t_{n-2}}^{t_n}(\sigma(t)\|u_{htt}(t)\|_0^2\|A_hu_h(t)\|_0^2+\sigma(t)\|u_{ht}\|_1^4)dt\end{aligned} \tag{6.14}$$

for $i=0,1,2$, and

$$\begin{aligned}\sigma^i(t_n)\|P_hE_n\|_0^2\tau&\le4\tau^{1+i}\int_{t_{n-1}}^{t_n}(\|f_{tt}\|_0^2+\nu^2\sigma^3(t)\|A_hu_{htt}\|_0^2)dt\\&+16c_0^2\gamma_0^2\tau^{1+i}\int_{t_{n-2}}^{t_n}(\sigma^2(t)\|u_{htt}(t)\|_1^2\|A_hu_h(t)\|_0^2+\sigma^2(t)\|A_hu_{ht}\|_0^2\|u_{ht}(t)\|_1^2)dt\end{aligned} \tag{6.15}$$

for $i = 0, 1, 2, 3$. Summing (6.13)–(6.15) from 2 to m, respectively, and using Theorem 3.2, we deduce (6.5)–(6.7). Similarly, we deduce from (6.4) and Lemma 3.1 that

$$\begin{aligned}
&\|R_h E_n\|_0 = \sup_{v_h \in X_h} \frac{|(E_n, v_h)|}{\|v_h\|_0} \\
&\le \frac{1}{2}\tau^{3/2}\left(\int_{t_{n-1}}^{t_n} \|f_{tt}(t)\|_0^2 dt\right)^{1/2} + \frac{\nu}{2}\left(\int_{t_{n-1}}^{t_n} (t_n - t)(t - t_{n-1})^2 \|\Delta_h u_{htt}(t)\|_0^2 dt\right)^{1/2} \\
&+ 2c_0\gamma_0\tau^{1/2}\left(\int_{t_{n-1}}^{t_n} (t - t_{n-1})^2 (\|u_{htt}(t)\|_1^2 \|A_h u_h(t)\|_0^2 + \|A_h u_{ht}\|_0^2 \|u_{ht}\|_1^2) dt\right)^{1/2} \\
&+ c_0\gamma_0\tau^{1/2}\left(\int_{t_{n-2}}^{t_{n-1}} (t - t_{n-2})^2 (\|u_{htt}(t)\|_1^2 \|A_h u_h(t)\|_0^2 + \|A_h u_{ht}\|_0^2 \|u_{ht}\|_1^2) dt\right)^{1/2}.
\end{aligned} \tag{6.16}$$

Because of the equivalent relation

$$\|A_h v_h\|_0 \le \|\Delta_h v_h\|_0 \le c_3 \|A_h v_h\|_0, \quad v_h \in V_h, \tag{6.17}$$

for some constant $c_3 > 0$, we deduce from (6.16) and (6.10)–(6.12) that

$$\begin{aligned}
\sigma^2(t_n)\|R_h E_n\|_0^2 \le{}& \tau^3 \int_{t_{n-1}}^{t_n} \|f_{tt}(t)\|_0^2 dt + \nu^2\tau^2 \int_{t_{n-1}}^{t_n} \sigma^3(t)\|A_h u_{htt}(t)\|_0^2 dt \\
&+ 4^2 c_0^2\gamma_0^2\tau^3 \int_{t_{n-1}}^{t_n} \sigma^2(t)(\|u_{htt}(t)\|_1^2 \|A_h u_h(t)\|_0^2 + \|A_h u_{ht}\|_0^2 \|u_{ht}\|_1^2) dt \\
&+ 4c_0^2\gamma_0^2\tau^3 \int_{t_{n-2}}^{t_{n-1}} \sigma^2(t)(\|u_{htt}(t)\|_1^2 \|A_h u_h(t)\|_0^2 + \|A_h u_{ht}\|_0^2 \|u_{ht}\|_1^2) dt.
\end{aligned} \tag{6.18}$$

Combining (6.18) with Theorem 3.2 yields (6.8). ∎

LEMMA 6.2. *Under the assumptions of Theorem 4.2, we have*

$$\|e^m\|_\alpha^2 + \nu\|e^m\|_{\alpha+1}^2\tau + \tau\sum_{n=1}^{m}\left(\frac{1}{2}\|d_t e^n\|_\alpha^2\tau + \nu\|\bar{e}^n\|_{\alpha+1}^2\right) \le \kappa\tau^{2-\alpha}, \quad \alpha = -1, 0, 1, \tag{6.19}$$

for all $1 \le m \le N$.

Proof. Taking $v_h = 2A_h^\alpha e^n\tau \in V_h$ and $q_h = 0$ in (6.3) and noting $e^n = \bar{e}^n + \frac{1}{2}d_t e^n$, we obtain

$$\begin{aligned}
&\|e^n\|_\alpha^2 - \|e^{n-1}\|_\alpha^2 + \|d_t e^n\|_\alpha^2\tau^2 + \frac{\nu}{2}(\|e^n\|_{\alpha+1}^2 - \|e^{n-1}\|_{\alpha+1}^2 + 4\|\bar{e}^n\|_{\alpha+1}^2)\tau \\
&\quad + 3b\left(e^{n-1}, u_h(t_{n-1}), A_h^\alpha\bar{e}^n + \frac{1}{2}A_h^\alpha d_t e^n\tau\right) + 3b\left(u_h^{n-1}, e^{n-1}, A_h^\alpha\bar{e}^n + \frac{1}{2}A_h^\alpha d_t e^n\tau\right)\tau \\
&\quad - b\left(e^{n-2}, u_h(t_{n-2}), A_h^\alpha\bar{e}^n + \frac{1}{2}A_h^\alpha d_t e^n\tau\right)\tau - b\left(u_h^{n-2}, e^{n-2}, A_h^\alpha\bar{e}^n + \frac{1}{2}A_h^\alpha d_t e^n\tau\right)\tau \\
&\quad = 2\left(E_n, A_h^\alpha\bar{e}^n + \frac{1}{2}A_h^\alpha d_t e^n\tau\right)\tau,
\end{aligned} \tag{6.20}$$

and by Lemma 3.1 and (3.10),

$$\begin{aligned}
&\frac{3}{2}|b(e^{n-1},u_h(t_{n-1}),A_h^\alpha d_t e^n)|\tau^2+\frac{3}{2}|b(u_h^{n-1},e^{n-1},A_h^\alpha d_t e^n)|\tau^2\\
&\le \frac{3}{2}c_0\gamma_0\|e^{n-1}\|_{\alpha+1}(\|A_h u_h(t_{n-1})\|_0+\|A_h u_h^{n-1}\|_0)\|d_t e^n\|_\alpha\tau^2\\
&\le \frac{1}{4}\|d_t e^n\|_\alpha^2\tau^2+\frac{9}{2}c_0^2\gamma_0^2(\|A_h u_h(t_{n-1})\|_0^2+\|A_h u_h^{n-1}\|_0^2)\|e^{n-1}\|_{\alpha+1}^2\tau^2,\\
&\frac{1}{2}|b(e^{n-2},u_h(t_{n-2}),A_h^\alpha d_t e^n)|\tau^2+\frac{1}{2}|b(u_h^{n-2},e^{n-2},A_h^\alpha d_t e^n)|\tau^2\\
&\le \frac{1}{8}\|d_t e^n\|_\alpha^2\tau^2+c_0^2\gamma_0^2(\|A_h u_h(t_{n-2})\|_0^2+\|A_h u_h^{n-2}\|_0^2)\|e^{n-2}\|_{\alpha+1}^2\tau^2,\\
&3|b(e^{n-1},u_h(t_{n-1}),A_h^\alpha\bar e^n)|\tau+3|b(u_h^{n-1},e^{n-1},A_h^\alpha\bar e^n)|\tau\\
&\le 3c_0\gamma_0\|e^{n-1}\|_\alpha(\|A_h u_h(t_{n-1})\|_0+\|A_h u_h^{n-1}\|_0)\|\bar e^n\|_{\alpha+1}\tau\\
&\le \frac{\nu}{4}\|\bar e^n\|_{\alpha+1}^2\tau+18\nu^{-1}c_0^2\gamma_0^2(\|A_h u_h(t_{n-1})\|_0^2+\|A_h u_h^{n-1}\|_0^2)\|e^{n-1}\|_\alpha^2\tau,\\
&|b(e^{n-2},u_h(t_{n-2}),A^\alpha\bar e^n)|\tau+|b(u_h^{n-2},e^{n-2},A^\alpha\bar e^n)|\tau\\
&\le \frac{\nu}{4}\|\bar e^n\|_{\alpha+1}^2\tau+18\nu^{-1}c_0^2\gamma_0^2(\|A_h u_h(t_{n-2})\|_0^2+\|A_h u_h^{n-2}\|_0^2)\|e^{n-2}\|_\alpha^2\\
&2\left(E_n,A^\alpha\bar e^n+\frac{1}{2}A^\alpha d_t e^n\tau\right)\tau\le\frac{\nu}{8}\|\bar e^n\|_{\alpha+1}^2\tau+\frac{1}{8}\|d_t e^n\|_\alpha^2\tau^2\\
&+8\nu^{-1}\|A_h^{\frac{\alpha-1}{2}}P_h E_n\|_0^2\tau+8\|A^{\frac{\alpha}{2}}P_h E_n\|_0^2\tau^2.
\end{aligned}$$

Hence, by combining the above inequalities with (6.20), we obtain

$$\begin{aligned}
&\left(\|e^n\|_\alpha^2+\frac{\nu}{2}\|e^n\|_{\alpha+1}^2\tau\right)-\left(\|e^{n-1}\|_\alpha^2+\frac{\nu}{2}\|e^{n-1}\|_{\alpha+1}^2\tau\right)+\frac{1}{2}\|d_t e^n\|_\alpha^2\tau^2+\nu\|\bar e^n\|_{\alpha+1}^2\tau\\
&\quad\le 18\nu^{-1}c_0^2\gamma_0^2(\|A_h u_h(t_{n-1})\|_0^2+\|A_h u_h^{n-1}\|_0^2)\left(\|e^{n-1}\|_\alpha^2+\frac{\nu}{2}\|e^{n-1}\|_{\alpha+1}^2\tau\right)\tau\\
&\quad+2\nu^{-1}c_0^2\gamma_0^2(\|A_h u_h(t_{n-2})\|_0^2+\|A_h u_h^{n-2}\|_0^2)\left(\|e^{n-2}\|_\alpha^2+\frac{\nu}{2}\|e^{n-2}\|_{\alpha+1}^2\tau\right)\tau\\
&\quad+8\nu^{-1}\|A_h^{\frac{\alpha-1}{2}}P_h E_n\|_0^2\tau+8\|A^{\frac{\alpha}{2}}P_h E_n\|_0^2\tau^2 \qquad (6.21)
\end{aligned}$$

for all $2\le n\le N$. Moreover, summing (6.21) from 2 to m and using Lemmas 4.3 and 6.1, we have

$$\begin{aligned}
&\|e^m\|_\alpha^2+\frac{\nu}{2}\|e^m\|_{\alpha+1}^2\tau+\tau\sum_{n=2}^m\left(\frac{1}{2}\|d_t e^n\|_\alpha^2\tau+\nu\|\bar e^n\|_{\alpha+1}^2\right)\\
&\quad\le\tau\sum_{n=1}^{m-1}d_n\left(\|e^n\|_\alpha^2+\frac{\nu}{2}\|e^n\|_{\alpha+1}^2\tau\right)+\|e^1\|_\alpha^2+\frac{\nu}{2}\|e^1\|_{\alpha+1}^2\tau\\
&\quad+8\nu^{-1}\tau\sum_{n=2}^m(\|A_h^{\frac{\alpha-1}{2}}P_h E_n\|_0^2+\nu\|A^{\frac{\alpha}{2}}P_h E_n\|_0^2\tau)\\
&\quad\le\tau\sum_{n=1}^m d_n\left(\|e^n\|_\alpha^2+\frac{\nu}{2}\|e^n\|_{\alpha+1}^2\tau\right)+\kappa\tau^{2-\alpha}, \qquad (6.22)
\end{aligned}$$

where

$$d_n=44\nu^{-1}c_0^2\gamma_0^2(\|A_h u_h(t_n)\|_0^2+\|A_h u_h^n\|_0^2).$$

We set

$$a_n = \|e^n\|_\alpha^2 + \frac{\nu}{2}\|e^n\|_{\alpha+1}^2\tau, \quad b_n = \frac{1}{2}\|d_t e^n\|_0^2\tau + \nu\|\bar{e}^n\|_1^2, \; C = \kappa\tau^{2-\alpha}$$

in (6.22), apply Lemma 3.4 to (6.22), and use Theorems 3.2 and 4.2 to deduce

$$\|e^m\|_\alpha^2 + \frac{\nu}{2}\|e^m\|_{\alpha+1}^2\tau + \tau\sum_{n=2}^{m}\left(\frac{1}{2}\|d_t e^n\|_\alpha^2\tau + \nu\|\bar{e}^n\|_{\alpha+1}^2\right) \le \kappa\tau^{2-\alpha} \tag{6.23}$$

for all $2 \le m \le N$. Combining (6.23) with Lemma 4.3 gives (6.19). ∎

With the aid of Lemma 6.2, we obtain the following preliminary lower-order smoothing error estimate.

LEMMA 6.3. *Under the assumptions of Theorem* 4.2*, we have*

$$\sigma(t_m)\|e^m\|_0^2 + \nu\sigma(t_m)\|e^m\|_1^2\tau + \tau\sum_{n=1}^{m}\sigma(t_n)\left(\frac{1}{2}\|d_t e^n\|_0^2\tau + \nu\|\bar{e}^n\|_1^2\right) \le \kappa\tau^3 \tag{6.24}$$

for all $1 \le m \le N$.

Proof. Multiplying (6.21) with $\alpha = 0$ by $\sigma(t_n)$ and using (6.11) gives

$$\begin{aligned}
&\sigma(t_n)\left(\|e^n\|_0^2 + \frac{\nu}{2}\|e^n\|_1^2\tau\right)\\
&\quad - \sigma(t_{n-1})\left(\|e^{n-1}\|_0^2 + \frac{\nu}{2}\|e^{n-1}\|_1^2\tau\right)\\
&\quad + \frac{1}{2}\sigma(t_n)\|d_t e^n\|_0^2\tau^2 + \nu\sigma(t_n)\|\bar{e}^n\|_1^2\tau\\
&\quad \le \left(\|e^{n-1}\|_0^2 + \frac{\nu}{2}\|e^{n-1}\|_1^2\tau\right)\tau\\
&\quad + 18\nu^{-1}c_0^2\gamma_0^2\sigma(t_n)(\|A_h u_h(t_{n-1})\|_0^2 + \|A_h u_h^{n-1}\|_0^2)\left(\|e^{n-1}\|_0^2 + \frac{\nu}{2}\|e^{n-1}\|_1^2\tau\right)\tau\\
&\quad + 2\nu^{-1}c_0^2\gamma_0^2\sigma(t_n)(\|A_h u_h(t_{n-2})\|_0^2 + \|A_h u_h^{n-2}\|_0^2)\left(\|e^{n-2}\|_0^2 + \frac{\nu}{2}\|e^{n-2}\|_1^2\tau\right)\tau\\
&\quad + 8\nu^{-1}\sigma(t_n)\|A_h^{\frac{-1}{2}}P_h E_n\|_0^2\tau + 8\sigma(t_n)\|P_h E_n\|_0^2\tau^2
\end{aligned} \tag{6.25}$$

for all $2 \le n \le N$. Summing (6.25) from 2 to m and using (6.11), Lemmas 4.3, 6.1, and 6.2 and Theorems 3.2 and 4.2, we obtain

$$\begin{aligned}
&\sigma(t_m)\left(\|e^m\|_0^2 + \frac{\nu}{2}\|e^m\|_1^2\tau\right) + \tau\sum_{n=2}^{m}\sigma(t_n)\left(\frac{1}{2}\|d_t e^n\|_0^2\tau + \nu\|\bar{e}^n\|_1^2\right)\\
&\le \tau\sum_{n=2}^{m}\left(\|e^{n-1}\|_0^2 + \frac{\nu}{2}\|e^{n-1}\|_1^2\tau\right) + \sigma(t_1)\left(\|e^1\|_0^2 + \frac{\nu}{2}\|e^1\|_1^2\tau\right)\\
&\quad + \tau\sum_{n=1}^{m-1} d_n\sigma(t_n)\left(\|e^n\|_0^2 + \frac{\nu}{2}\|e^n\|_1^2\tau\right)\\
&\quad + 8\tau\sum_{n=2}^{m}\sigma(t_n)(\nu^{-1}\|A_h^{-1/2}P_h E_n\|_0^2 + \|P_h E_n\|_0^2\tau)\\
&\quad + 18\nu^{-1}c_0^2\gamma_0^2\tau^2\sum_{n=2}^{m}(\|A_h u_h(t_{n-1})\|_0^2 + \|A_h u_h^{n-1}\|_0^2)\left(\|e^{n-1}\|_0^2 + \frac{\nu}{2}\|e^{n-1}\|_1^2\tau\right)
\end{aligned}$$

$$
\begin{aligned}
&+ 4\nu^{-1}c_0^2\gamma_0^2\tau^2\sum_{n=2}^{m}(\|A_hu_h(t_{n-2})\|_0^2+\|A_hu_h^{n-2}\|_0^2)\left(\|e^{n-2}\|_0^2+\frac{\nu}{2}\|e^{n-2}\|_1^2\tau\right)\\
&\le \tau\sum_{n=2}^{m}\left(\|e^{n-1}\|_0^2+\frac{\nu}{2}\|e^{n-1}\|_1^2\tau\right)+\kappa\tau^3\\
&+ \tau\sum_{n=1}^{m-1}d_n\sigma(t_n)\left(\|e^{n}\|_0^2+\frac{\nu}{2}\|e^{n}\|_1^2\tau\right). \qquad (6.26)
\end{aligned}
$$

Because $e^{n-1}=\bar{e}^n-\frac{1}{2}d_te^n$, we deduce from Lemma 6.2 that

$$
\tau\sum_{n=2}^{m}\|e^{n-1}\|_0^2\le 2\tau\sum_{n=2}^{m}\|\bar{e}^n\|_0^2+\tau\sum_{n=2}^{m}\|d_te^n\|_0^2\tau^2\le\kappa\tau^3 \qquad (6.27)
$$

and

$$
\tau\sum_{n=2}^{m}\|e^{n-1}\|_1^2\le 2\tau\sum_{n=2}^{m}\|\bar{e}^n\|_1^2+\tau\sum_{n=2}^{m}\|d_te^n\|_1^2\tau^2\le\kappa\tau^2. \qquad (6.28)
$$

Combining (6.27)–(6.28) with (6.26) gives

$$
\begin{aligned}
&\sigma(t_m)\left(\|e^m\|_0^2+\frac{\nu}{2}\|e^m\|_1^2\tau\right)+\tau\sum_{n=2}^{m}\sigma(t_n)\left(\frac{1}{2}\|d_te^n\|_0^2\tau+\nu\|\bar{e}^n\|_1^2\right)\\
&\qquad\le\kappa\tau^3+\tau\sum_{n=1}^{m-1}d_n\sigma(t_n)\left(\|e^{n}\|_0^2+\frac{\nu}{2}\|e^{n}\|_1^2\tau\right). \qquad (6.29)
\end{aligned}
$$

Applying Lemma 3.4 to (6.29) yields (6.24). ∎

LEMMA 6.4. *Under the assumptions of Theorem* 4.2*, we have*

$$
\tau\sum_{n=1}^{m}\|\bar{e}^n\|_{-1}^2\le\kappa\tau^4 \qquad (6.30)
$$

for all $1\le m\le N$.

Proof. Let $\{\Phi_h^n\}$ be the solution of (5.1), corresponding to the initial value $\Phi_h^m=0$ and the right-hand side $\{\xi^n\}=\{A_h^{-1}e^n\}$. Then, by construction, there holds

$$
\begin{aligned}
\|\bar{e}^n\|_{-1}^2\tau&=(\bar{e}^n,d_t\Phi_h^n)\tau-a(\bar{e}^n,\bar{\Phi}_h^n)\tau-b\left(\frac{3}{2}u_h^{n-1}-\frac{1}{2}u_h^{n-2},\bar{e}^n,\bar{\Phi}_h^n\right)\tau\\
&\quad-b\left(\bar{e}^n,\frac{3}{2}u_h^{n-1}-\frac{1}{2}u_h^{n-2},\bar{\Phi}_h^n\right)\tau\\
&=d_t(e^n,\Phi_h^n)\tau-(d_te^n,\bar{\Phi}_h^n)\tau-a(\bar{e}^n,\bar{\Phi}_h^n)\tau-b\left(\frac{3}{2}u_h^{n-1}-\frac{1}{2}u_h^{n-2},\bar{e}^n,\bar{\Phi}_h^n\right)\tau\\
&\quad-b\left(\bar{e}^n,\frac{3}{2}u_h^{n-1}-\frac{1}{2}u_h^{n-2},\bar{\Phi}_h^n\right)\tau\\
&=(e^n,\Phi^n)-(e^{n-1},\Phi_h^{n-1})-(E_n,\bar{\Phi}_h^n)\tau+\frac{3}{2}b(e^{n-1},e^{n-1},\bar{\Phi}^n)\tau\\
&\quad-\frac{1}{2}b(e^{n-2},e^{n-2},\bar{\Phi}^n)\tau-\frac{3}{2}b\left(\frac{1}{2}d_te^n\tau,u_h^{n-1},\bar{\Phi}_h^n\right)\tau-\frac{3}{2}b\left(u_h^{n-1},\frac{1}{2}d_te^n\tau,\bar{\Phi}_h^n\right)\tau\\
&\quad+\frac{1}{2}b\left(\frac{1}{2}d_te^n\tau+d_te^{n-1}\tau,u_h^{n-2},\bar{\Phi}_h^n\right)\tau+\frac{1}{2}b\left(u_h^{n-2},\frac{1}{2}d_te^n\tau+d_te^{n-1}\tau,\bar{\Phi}_h^n\right)\tau. \qquad (6.31)
\end{aligned}
$$

Taking $v_h = u_{htt}$ in (5.1) and using (3.10) and Lemma 3.1, we see that

$$\begin{aligned}|a(u_{htt},\bar{\Phi}_h^n)| &\le |(u_{htt}, d_t\Phi_h^n - \bar{\xi}^n)| + \left|b\left(\frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}, u_{htt}, \bar{\Phi}_h^n\right)\right| \\ &\quad + \left|b\left(u_{htt}, \frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}, \bar{\Phi}_h^n\right)\right| \\ &\le \|u_{htt}\|_{-1}\left(\|d_t\Phi_h^n - \bar{\xi}^n\|_1 + c_0\gamma_0\left\|A_h\left(\frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}\right)\right\|_0 \|\bar{\Phi}_h^n\|_2\right). \end{aligned} \tag{6.32}$$

Using (6.4), (6.32), (3.10), and Lemma 3.1, we deduce that

$$\begin{aligned}&|(E_n,\bar{\Phi}_h^n)|\tau \le \tau^2\gamma_0^2\int_{t_{n-1}}^{t_n}\|f_{tt}\|_0 dt \|A_h\bar{\Phi}_h^n\|_0 \\ &+ c_0\gamma_0\tau^2\int_{t_{n-2}}^{t_n}(\|u_{htt}\|_{-1}\|A_hu_h\|_0 + 2\|u_{ht}\|_1\|u_{ht}\|_0)dt\|A_h\bar{\Phi}_h^n\|_0 \\ &+ \tau^2\int_{t_{n-1}}^{t_n}\|u_{htt}\|_{-1}dt\left(\|d_t\Phi_h^n - \bar{\xi}^n\|_1 + c_0\gamma_0\left\|A_h\left(\frac{3}{2}u_h^{n-1} - \frac{1}{2}u_h^{n-2}\right)\right\|_0\|A_h\bar{\Phi}_h^n\|_0\right). \end{aligned} \tag{6.33}$$

Summing (6.33) for $2 \le n \le m$ and using Theorem 3.2, we have

$$\begin{aligned}\tau\sum_{n=2}^{m}|(E_n,\bar{\Phi}_h^n)| &\le \tau^2\kappa\left(\int_{t_0}^{t_m}(\|f_{tt}\|_0^2 + \|u_{htt}\|_{-1}^2 + \|u_{ht}\|_1^2)dt\right)^{1/2} \\ &\quad\times\left(\tau\sum_{n=2}^{m}(\|A_h\bar{\Phi}_h^n\|^2 + \|d_t\Phi_h^n - \bar{\xi}^n\|_1^2)\right)^{1/2}. \end{aligned} \tag{6.34}$$

Moreover, by using (3.10) and Lemma 3.1, we have

$$\begin{aligned}\frac{3}{2}|b(e^{n-1},e^{n-1},\bar{\Phi}^n)|\tau + \frac{1}{2}|b(e^{n-2},e^{n-2},\bar{\Phi}_h^n)|\tau &\le \frac{3}{2}c_0\gamma_0(\|e^{n-1}\|_1\|e^{n-1}\|_0 \\ &\quad + \|e^{n-2}\|_1\|e^{n-2}\|_0)\|A_h\bar{\Phi}_h^n\|_0\tau, \\ \frac{3}{2}\left|b\left(\frac{1}{2}d_te^n\tau, u_h^{n-1},\bar{\Phi}_h^n\right)\right|\tau + \frac{3}{2}\left|b\left(u_h^{n-1},\frac{1}{2}d_te^n\tau,\bar{\Phi}_h^n\right)\right|\tau &\le c_0\gamma_0\|A_hu_h^{n-1}\|_0\|d_te^n\|_{-1}\|A_h\bar{\Phi}_h^n\|_0\tau^2, \\ \frac{1}{2}\left|b\left(\frac{1}{2}d_te^n\tau + d_te^{n-1}\tau, u_h^{n-2},\bar{\Phi}_h^n\right)\right|\tau + \frac{1}{2}\left|b\left(u_h^{n-2},\frac{1}{2}d_te^n\tau + d_te^{n-1}\tau,\bar{\Phi}_h^n\right)\right|\tau &\le c_0\gamma_0\|A_hu_h^{n-2}\|_0(\|d_te^n\|_{-1} \\ &\quad + \|d_te^{n-1}\|_{-1})\|A_h\bar{\Phi}_h^n\|_0\tau^2. \end{aligned}$$

Summing (6.31) from 2 to m and using the above estimates, (6.34), and Theorem 4.2, we arrive at

$$\begin{aligned}
&\tau\sum_{n=2}^{m}\|\bar{e}^n\|_{-1}^2 \le -(e^1,\Phi_h^1)+\tau\sum_{n=2}^{m}|(E_n,\bar{\Phi}_h^n)|\\
&+\frac{3}{2}c_0\gamma_0\tau\sum_{n=2}^{m}(\|e^{n-1}\|_1\|e^{n-1}\|_0+\|e^{n-2}\|_1\|e^{n-2}\|_0)\|A_h\bar{\Phi}_h^n\|_0\\
&+c_0\gamma_0\tau\sum_{n=2}^{m}(\|A_hu_h^{n-1}\|_0+\|A_hu_h^{n-2}\|_0)\|d_te^n\|_{-1}\|A_h\bar{\Phi}_h^n\|_0\tau^2\\
&+c_0\gamma_0\tau\sum_{n=2}^{m}\|A_hu_h^{n-2}\|_0\|d_te^{n-1}\|_{-1}\|A_h\bar{\Phi}_h^n\|_0\tau^2\\
&\le\|e^1\|_{-2}\|A_h\Phi_h^1\|_0+\kappa\tau^2\left(\tau\sum_{n=2}^{m}(\|A_h\bar{\Phi}_h^n\|^2+\|d_t\Phi_h^n-\bar{\xi}^n\|_1^2)\right)^{1/2}\\
&+\kappa\left(\tau\sum_{n=2}^{m}\|e^{n-1}\|_1^2\|e^{n-1}\|_0^2+\|e^{n-2}\|_1^2\|e^{n-2}\|_0^2\right)^{1/2}\left(\tau\sum_{n=2}^{m}\|A_h\bar{\Phi}_h^n\|_0^2\right)^{1/2}\\
&+\kappa\left(\tau\sum_{n=1}^{m}\|d_te^n\|_{-1}^2\tau^2\right)^{1/2}\left(\tau\sum_{n=2}^{m}\|A_h\bar{\Phi}_h^n\|_0^2\right)^{1/2}. \qquad (6.35)
\end{aligned}$$

Using Lemmas 5.2 and 5.3 in (6.35) yields

$$\begin{aligned}
&\tau\sum_{n=2}^{m}\|\bar{e}^n\|_{-1}^2\\
&\le\left(\|e^1\|_{-2}^2+\kappa\tau^4+\kappa\tau\sum_{n=2}^{m}(\|e^{n-1}\|_1^2\|e^{n-1}\|_0^2+\|e^{n-2}\|_1^2\|e^{n-2}\|_0^2)+\kappa\tau\sum_{n=1}^{m}\|d_te^n\|_{-1}^2\tau^2\right)^{1/2}\\
&\times\left(\|A_h\Phi_h^1\|_0^2+\tau\sum_{n=2}^{m}(\|A_h\bar{\Phi}_h^n\|_0^2+\|d_t\Phi_h^n-\bar{\xi}^n\|_1^2)\right)^{1/2}\\
&\le\kappa\left(\|e^1\|_{-2}^2+\tau^4+\tau\sum_{n=2}^{m}(\|e^{n-1}\|_1^2\|e^{n-1}\|_0^2+\|e^{n-2}\|_1^2\|e^{n-2}\|_0^2)+\tau\sum_{n=1}^{m}\|d_te^n\|_{-1}^2\tau^2\right)^{1/2}\\
&\times\left(\tau\sum_{n=2}^{m}\|\bar{\xi}^n\|_1^2\right)^{1/2},
\end{aligned}$$

which leads to

$$\begin{aligned}
&\tau\sum_{n=2}^{m}\|\bar{e}^n\|_{-1}^2\le\kappa(\|e^1\|_{-2}^2+\tau^4)\\
&+\kappa\tau\sum_{n=2}^{m}(\|e^{n-1}\|_1^2\|e^{n-1}\|_0^2+\|e^{n-2}\|_1^2\|e^{n-2}\|_0^2)+\kappa\tau\sum_{n=1}^{m}\|d_te^n\|_{-1}^2\tau^2 \qquad (6.36)
\end{aligned}$$

for all $2\le m\le N$. Using (6.28) and Lemmas 4.3, 6.2, and 6.3 in (6.36), we obtain (6.30). ∎

Next we prove (6.19) for the case $\alpha = -2$.

LEMMA 6.5. *Under the assumptions of Theorem* 4.2*, we have*

$$\sigma^2(t_m)\|e^m\|_0^2 \le \kappa\tau^4 \tag{6.37}$$

for all $1 \le m \le N$.

Proof. Let $\{\Phi_h^n\}$ be the solution of (5.1), corresponding to the initial value $\Phi_h^m = e^m$ and the right-hand side $\{\xi^n\} = \{0\}$. Then, by construction, there holds

$$\begin{aligned} d_t(e^n, \Phi_h^n) &= (d_t e^n, \bar{\Phi}^n) + (\bar{e}^n, d_t\Phi_h^n) \\ &= (E_n, \bar{\Phi}_h^n) - \frac{3}{2}b(e^{n-1}, e^{n-1}, \bar{\Phi}_h^n) + \frac{1}{2}b(e^{n-2}, e^{n-2}, \bar{\Phi}_h^n) \\ &\quad + \frac{3}{2}b\left(u_h^{n-1}, \frac{1}{2}d_t e^n\tau, \bar{\Phi}_h^n\right) - \frac{1}{2}b\left(u_h^{n-2}, d_t e^n\tau + \frac{1}{2}d_t e^{n-1}\tau, \bar{\Phi}_h^n\right)\tau \\ &\quad + \frac{3}{2}b\left(\frac{1}{2}d_t e^n\tau, u_h^{n-1}, \bar{\Phi}_h^n\right) - \frac{1}{2}b\left(d_t e^n\tau + \frac{1}{2}d_t e^{n-1}\tau, u_h^{n-2}, \bar{\Phi}_h^n\right). \end{aligned} \tag{6.38}$$

Using (3.10) and Lemma 3.1, we see that

$$\begin{aligned} &|(E_n, \bar{\Phi}_h^n)| \le \|A_h^{-1/2}P_h E_n\|_0\|\bar{\Phi}_h^n\|_1, \\ &\frac{3}{2}|b(e^{n-1}, e^{n-1}, \bar{\Phi}_h^n)| + \frac{1}{2}|b(e^{n-2}, e^{n-2}, \bar{\Phi}_h^n)| \\ &\le c_0(\|e^{n-1}\|_1\|e^{n-1}\|_0 + \|e^{n-2}\|_1\|e^{n-2}\|_0)\|\bar{\Phi}^n\|_1 + c_0(\|e^{n-1}\|_1^2 + \|e^{n-2}\|_1^2)\|\bar{\Phi}^n\|_0, \\ &\frac{3}{2}\left|b\left(u_h^{n-1}, \frac{1}{2}d_t e^n\tau, \bar{\Phi}_h^n\right)\right| + \frac{3}{2}\left|b\left(\frac{1}{2}d_t e^n\tau, u_h^{n-1}, \bar{\Phi}_h^n\right)\right| \\ &\le c_0\gamma_0\|A_h u_h^{n-1}\|_0\|d_t e^n\|_0\|\bar{\Phi}_h^n\|_1\tau, \\ &\frac{1}{2}\left|b\left(u_h^{n-2}, d_t e^n\tau + \frac{1}{2}d_t e^{n-1}\tau, \bar{\Phi}_h^n\right)\right|\tau + \frac{1}{2}\left|b\left(d_t e^n\tau + \frac{1}{2}d_t e^{n-1}\tau, u_h^{n-2}, \bar{\Phi}_h^n\right)\right| \\ &\le c_0\gamma_0\|A_h u_h^{n-2}\|_0\|d_t e^n + \frac{1}{2}d_t e^{n-1}\|_0\|\bar{\Phi}_h^n\|_1\tau. \end{aligned}$$

Multiplying (6.38) by $t_n\tau$, summing for $2 \le n \le m$, and using the above inequalities, we deduce that

$$\begin{aligned} t_m(e^m, \Phi_h^m) &= \tau(e^1, \Phi_h^1) + \tau\sum_{n=2}^{m}(e^{n-1}, \Phi_h^{n-1}) \\ &\quad + \left(\tau\sum_{n=2}^{m} t_n^2\|A_h^{-1/2}P_h E_n\|_0^2\right)^{1/2}\left(\tau\sum_{n=2}^{m}\|\bar{\Phi}_h^n\|_1^2\right)^{1/2} \\ &\quad + 2c_0\left(\tau\sum_{n=2}^{m} t_n^2(\|e^{n-1}\|_1^2\|e^{n-1}\|_0^2 + \|e^{n-2}\|_1^2\|e^{n-2}\|_0^2)\right)^{1/2}\left(\tau\sum_{n=2}^{m}\|\bar{\Phi}_h^n\|_1^2\right)^{1/2} \\ &\quad + 2c_0\tau\sum_{n=2}^{m} t_n^2(\|e^{n-1}\|_1^2 + \|e^{n-2}\|_1^2)\sup_{1\le n\le m}\|\Phi_h^n\|_0 \\ &\quad + 2c_0\left(\tau\sum_{n=2}^{m} t_n^2\|A_h u_h^{n-1}\|_0^2\|d_t e^n\|_0^2\tau^2\right)^{1/2}\left(\tau\sum_{n=2}^{m}\|\bar{\Phi}_h^n\|_1^2\right)^{1/2} \\ &\quad + 2c_0\left(\tau\sum_{n=2}^{m} t_n^2\|A_h u_h^{n-2}\|_0^2\left\|d_t e^n + \frac{1}{2}d_t e^{n-1}\right\|_0^2\tau^2\right)^{1/2}\left(\tau\sum_{n=2}^{m}\|\bar{\Phi}_h^n\|_1^2\right)^{1/2}. \end{aligned} \tag{6.39}$$

Using Theorem 4.2, Lemmas 6.1, 6.2, and 6.3, (5.4) and (6.28) in (6.39) and noting

$$t_n \le \sigma(t_n)T, \quad t_n^2 \le 4t_{n-1}^2 \le 4\sigma^2(t_{n-1})T^2,$$

we arrive at

$$t_m(e^m, \Phi_h^m) = \tau(e^1, \Phi_h^1) + \tau \sum_{n=2}^{m} (e^{n-1}, \Phi_h^{n-1}) + \kappa\tau^2 \|\Phi_h^m\|_0^2. \tag{6.40}$$

Because

$$(e^{n-1}, \Phi_h^{n-1}) + (e^n, \Phi_h^n) = 2(\bar{e}^n, \bar{\Phi}_h^n) + \frac{1}{4}(d_t e^n, d_t \Phi_h^n)\tau^2,$$

there holds

$$\begin{aligned} \tau \sum_{n=2}^{m} (e^{n-1}, \Phi_h^{n-1}) &= \frac{\tau}{2} \sum_{n=2}^{m} [(e^{n-1}, \Phi_h^{n-1}) + (e^n, \Phi_h^n)] + \frac{\tau}{2}[(e^1, \Phi_h^1) - (e^m, \Phi_h^m)] \\ &= \tau \sum_{n=2}^{m} \left[(\bar{e}^n, \bar{\Phi}_h^n) + \frac{1}{4}(d_t e^n, d_t \Phi_h^n) \right] + \frac{\tau}{2}[(e^1, \Phi_h^1) - (e^m, \Phi_h^m)]. \end{aligned} \tag{6.41}$$

Furthermore, by using Lemmas 5.2, 6.2, and 6.4, we have

$$\begin{aligned} \tau \sum_{n=2}^{m} \left[(\bar{e}^n, \bar{\Phi}_h^n) + \frac{1}{4}(d_t e^n, d_t \Phi_h^n) \right] &\le \left(\tau \sum_{n=2}^{m} \|\bar{e}^n\|_{-1}^2 \right)^{1/2} \left(\tau \sum_{n=2}^{m} \|\bar{\Phi}_h^n\|_1^2 \right)^{1/2} \\ &\quad + \tau^2 \left(\tau \sum_{n=2}^{m} \|d_t e^n\|_1^2 \right)^{1/2} \left(\tau \sum_{n=2}^{m} \|d_t \Phi_h^n\|_{-1}^2 \right)^{1/2} \\ &\le \kappa\tau^2 \|\Phi_h^m\|_0. \end{aligned} \tag{6.42}$$

Combining (6.41)–(6.42) with (6.40) and using Lemmas 4.3 and 5.2 yields

$$\frac{\tau}{2}(e^m, \Phi_h^m) + t_m(e^m, \Phi_h^m) \le \frac{3}{2}\tau \|e^1\|_0 \|\Phi_h^1\|_0 + \kappa\tau^2 \|\Phi_h^m\|_0 \le \kappa\tau^2 \|\Phi_h^m\|_0. \tag{6.43}$$

The assertion follows $\Phi_h^m = e^m$ and $\sigma(t_m) \le t_m$. □

It remains to prove the error estimate for the approximate pressure p_h^m.

By (3.4), (3.10), (6.3), and Lemma 3.1,

$$\begin{aligned} \|\eta^m\|_0 \le \beta^{-1} &\left(\gamma_0 \|d_t e^m\|_0 + \nu \|\bar{e}^m\|_1 + \frac{3}{2} c_0 \gamma_0 \|e^{m-1}\|_0 (\|A_h u_h(t_{m-1})\|_0 + \|A_h u_h^{m-1}\|_0) \right) \\ &+ \frac{1}{2}\beta^{-1} c_0 \gamma_0 \|e^{m-2}\|_0 (\|A_h u_h(t_{m-2})\|_0 + \|A_h u_h^{m-2}\|_0) + \beta^{-1}\gamma_0 \|R_h E_m\|_0, \end{aligned}$$

which with Theorems 3.2 and 4.2 yields

$$\sigma(t_m)\|\eta^m\|_0 \le \kappa\sigma(t_m)(\|d_t e^m\|_0 + \|\bar{e}^m\|_1 + \|e^{m-1}\|_0 + \|e_h^{m-2}\|_0 + \|R_h E_m\|_0). \tag{6.44}$$

Using (6.11) and Lemmas 6.1, 6.3, and 6.5 in (6.44) yields

$$\sigma(t_m)\|\eta^m\|_0 \le \kappa\tau, \quad 2 \le m \le N. \tag{6.45}$$

Finally, by using Theorem 3.3 and the integral by part, we have

$$\begin{aligned}\sigma(t_m)\|p_h(t_m)-p_h^m\|_0 &\le \sigma(t_m)\|\eta^m\|_0+2\sigma(t_{m-1})\left\|p_h(t_m)-\frac{1}{\tau}\int_{t_{m-1}}^{t_m}p_h(t)dt\right\|_0\\ &\le \sigma(t_m)\|\eta^m\|_0+2\int_{t_{m-1}}^{t_m}\sigma(t)\|p_{ht}(t)\|_0dt\\ &\le \sigma(t_m)\|\eta^m\|_0+\kappa\tau,\quad 1\le m\le N.\end{aligned}$$

Combining this inequality with (6.45) and using Lemma 4.3 yields

$$\sigma(t_m)\|p_h(t_m)-p_h^m\|_0\le\kappa\tau,\quad 1\le m\le N. \tag{6.46}$$

THEOREM 6.6. *Under the assumptions of Theorem* 4.2*, the following error estimates hold:*

$$\sigma(t_m)\|u_h(t_m)-u_h^m\|_0+\sigma^{1/2}(t_m)\tau\|u_h(t_m)-u_h^m\|_1\le\kappa\tau^2,\quad t_m\in(0,T], \tag{6.47}$$

$$\sigma(t_m)\|p_h(t_m)-p_h^m\|_0\le\kappa\tau,\quad t_m\in(0,T]. \tag{6.48}$$

This proof is completed by combining (6.45) with Lemmas 6.3 and 6.5.

Remark 6.1. Combining Theorem 6.6 with (3.11) yields (1.10)–(1.12).

REFERENCES

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

[2] A. AIT OU AMMI AND M. MARION, *Nonlinear Galerkin methods and mixed finite elements: Two-grid algorithms for the Navier-Stokes equations*, Numer. Math., 68 (1994), pp. 189–213.

[3] G. A. BAKER, *Galerkin Approximations for the Navier-Stokes Equations*, manuscript, Harvard University, Cambridge, MA, 1976.

[4] G. A. BAKER, V. A. DOUGALIS, AND O. A. KARAKASHIAN, *On a high order accurate fully discrete Galerkin approximation to the Navier-Stokes equations*, Math. Comp., 39 (1982), pp. 339–375.

[5] J. R. CANNON AND Y. LIN, *A priori L^2 error estimates for finite-element methods for nonlinear diffusion equations with memory*, SIAM J. Numer. Anal., 27 (1990), pp. 595–607.

[6] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[7] W. E AND J.-G. LIU, *Projection methods* I: *Convergence and numerical boundary layers*, SIAM J. Numer. Anal., 32(1995), pp. 1017–1057.

[8] G. FAIRWEATHER, H. MA, AND W. SUN, *Orthogonal Spline Collocation Methods for the Navier-Stokes Equations in Stream Function and Vorticity Formulation*, submitted.

[9] V. GIRAULT AND P. A. RAVIART, *Finite Element Method for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1987.

[10] Y. HE, *Stability and error analysis for a spectral Galerkin method for the Navier-Stokes equations with H^2 or H^1 initial data*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 875–904.

[11] Y. HE AND K. LI, *Convergence and stability of finite element nonlinear Galerkin method for the Navier-Stokes equations*, Numer. Math., 79 (1998), pp. 77–106.

[12] Y. HE AND K. LI, *Nonlinear Galerkin method and two-step method for the Navier-Stokes equations*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 283–305.

[13] Y. HE, *Two-level method based on finite element and Crank-Nicolson extrapolation for the time-dependent Navier-Stokes equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1263–1285.

[14] Y. HE AND K. M. LIU, *A multi-level finite element method for the time-dependent Navier-Stokes equations*, Numer. Methods for Partial Differential Equations, 21 (2005), pp. 1052–1068.

[15] J. G. HEYWOOD AND R. RANNACHER, *Finite-element approximations of the nonstationary Navier–Stokes problem. Part* I: *Regularity of solutions and second-order spatial discretization*, SIAM J. Numer. Anal., 19 (1982), pp. 275–311.

[16] J. G. Heywood and R. Rannacher, *Finite-element approximations of the nonstationary Navier–Stokes problem. Part* IV: *Error estimates for second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.
[17] A. T. Hill and E. Süli, *Approximation of the global attractor for the incompressible Navier-Stokes equations*, IMA J. Numer. Anal., 20 (2000), pp. 633–667.
[18] E. Issacson and H. B. Keller, *Analysis of Numerical Methods*, Wiley, New York, 1966.
[19] H. Johnston and J.-G. Liu, *Accurate, stable and efficient Navier-Stokes slovers based on explicit treatment of the pressure term*, J. Comput. Phys., 199 (2004), pp. 221–259.
[20] J. Kim and P. Moin, *Application of a fractional-step method to incompressible Navier-Stokes equations*, J. Comput. Phys., 59 (1985), pp. 308–323.
[21] R. B. Kellogg and J. E. Osborn, *A regularity result for the Stokes problem in a convex polygon*, J. Funct. Anal., 21 (1976), pp. 397–431.
[22] S. Larsson, *The long-time behavior of finite-element approximations of solutions to semilinear parabolic problems*, SIAM J. Numer. Anal., 26 (1989), pp. 348–365.
[23] Y. Lin, *Galerkin methods for nonlinear parabolic integrodifferential equations with nonlinear boundary conditions*, SIAM J. Numer. Anal., 27 (1990), pp. 608–621.
[24] H. Ma and W. Sun, *Optimal error estimates of the Legendre Petro-Galerkin and pseudospectral methods for the generalized Korteweg-de Vries equation*, SIAM J. Numer. Anal., 39 (2001), pp. 1380–1394.
[25] M. Marion and R. Temam, *Navier-Stokes equations: Theory and approximation*, Handb. Numer. Anal. VI, North–Holland, Amsterdam, 1998, pp. 503–688.
[26] R. H. Nochetto and J.-H. Pyo, *A finite element Gauge-Uzawa method Part I: Navier-Stokes equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1043–1068.
[27] J. Shen, *Long time stability and convergence for fully discrete nonlinear Galerkin methods*, Appl. Anal., 38 (1990), pp. 201–229.
[28] J. C. Simo and F. Armero, *Unconditional stability and long-term behavior of transient algorithms for the incompressible Navier-Stokes and Euler equations*, Comput. Methods Appl. Mech. Engrg., 111 (1994), pp. 111–154.
[29] R. Temam, *Navier-Stokes Equations, Theory and Numerical Analysis*, 3rd ed., North–Holland, Amsterdam, 1983.
[30] F. Tone, *Error analysis for a second scheme for the Navier-Stokes equations*, Appl. Numer. Math., 50 (2004), pp. 93–119.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 870–889

POSTPROCESSING FOR STOCHASTIC PARABOLIC PARTIAL DIFFERENTIAL EQUATIONS*

GABRIEL J. LORD† AND TONY SHARDLOW‡

Abstract. We investigate the strong approximation of stochastic parabolic partial differential equations with additive noise. We introduce postprocessing in the context of a standard Galerkin approximation, although other spatial discretizations are possible. In time, we follow [G. J. Lord and J. Rougemont, *IMA J. Numer. Anal.*, 24 (2004), pp. 587–604] and use an exponential integrator. We prove strong error estimates and discuss the best number of postprocessing terms to take. Numerically, we evaluate the efficiency of the methods and observe rates of convergence. Some experiments with the implicit Euler–Maruyama method are described.

Key words. stochastic exponential integrator, postprocessing, numerical solution of stochastic PDEs

AMS subject classifications. 60H15, 65M12, 65M15, 65M60

DOI. 10.1137/050640138

1. Introduction.

We consider the numerical approximation of the stochastic evolution equation

$$du = \Big[\Delta u + F(u)\Big]\,dt + dW(t) \qquad \text{given} \quad u(0) = u_0, \tag{1.1}$$

with periodic boundary conditions on $[0, 2\pi)$, where $W(t)$ is a Q Wiener process [3] on $L^2(0, 2\pi)$ and F is nonlinear (precise assumptions are given in section 3.1).

Suppose that ϕ_n are eigenvectors of the Laplacian Δ with periodic boundary conditions so that $\Delta\phi_n = -n^2\phi_n$, $n \in \mathbb{Z}$. We assume that Q has eigenfunctions ϕ_n with corresponding eigenvalues $\lambda_n \geq 0$, in which case

$$W(t) = \sum_{n\in\mathbb{Z}} \lambda_n^{1/2}\phi_n\beta_n(t), \tag{1.2}$$

for independent Brownian motions β_n. We do not consider the existence of solutions to (1.1) here; instead we call on [3]. We will investigate the effect on numerics of the spatial regularity of the noise determined from the decay of λ_n.

There is a growing literature on numerical methods for stochastic PDEs, and the majority of these analyze convergence in the strong or root mean squared sense. Finite difference approximations have been examined by a number of authors (see, for example, [26, 11, 12, 4]) and finite element methods have also been considered, e.g., [29]. Galerkin approximations and strong Taylor schemes were considered in [10] with a scalar Wiener process. Strong convergence of the implicit Euler–Maruyama

*Received by the editors September 12, 2005; accepted for publication (in revised form) October 10, 2006; published electronically April 27, 2007.
http://www.siam.org/journals/sinum/45-2/64013.html

†Department of Mathematics and the Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK (gabriel@ma.hw.ac.uk). This author's work was supported in part by EPSRC grant GR/S60921/01.

‡School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, UK (shardlow@maths.man.ac.uk). This author's work was supported in part by EPSRC grant GR/R78725/01.

method was investigated in [19]. A more general analysis is found in [14], which considers different types of spatial discretizations (Galerkin as well as collocation, finite differences, finite elements, and wavelet-based schemes) for similar forms of noise considered here. Mueller-Gronbach and Ritter [25] analyze convergence and complexity through the number of random samples of the Wiener process. Spatially smooth noise is considered in [21] and [27], and these papers also consider Fourier-based spatial discretizations. In [27], a Taylor-based discretization is taken, and efficient methods for approximating the Wiener process are considered. In [21] strong convergence of an exponential integrator (see also [24]) is examined, and we consider this scheme further in this paper; see section 3.

The purpose of this paper is to study Galerkin postprocessing methods for (1.1), prove their convergence in the strong sense, and evaluate their efficiency. We have restricted attention to reaction diffusion equations with homogeneous diffusion and additive noise, and plan to return to the general case in further work.

Section 2 is an introduction to postprocessing methods for deterministic PDEs. Section 3 describes our Galerkin postprocessing scheme, and section 3.1 a theorem on the convergence of the method. Section 4 investigates the numerical behavior of the method for the stochastic Allen–Cahn equation. We evaluate the efficiency of the methods, compare the rates of convergence to those predicted by the theorem, and illustrate numerically that postprocessing is efficient for other time stepping algorithms by experimenting with implicit Euler–Maruyama methods. We summarize our results and conclude in section 5. The proof of the theorem is given in section 6, with the proof of two lemmas left to the appendix.

2. A review of deterministic postprocessing. Postprocessing methods originate from analytical results on inertial manifolds for PDEs; see, for example, [6], where it can be shown that the dynamics of infinite dimensional PDEs converge to a finite dimensional system in large time. Typically, a graph Φ is obtained that "enslaves" the high Fourier modes (fine scale dynamics) to a finite number of low Fourier modes (large scale dynamics). For example, if P denotes the projection onto the first N Fourier modes and $u = p + q = Pu + (I - P)u$, we can write the deterministic PDE

$$u_t = \Delta u + F(u) \qquad \text{as} \qquad p_t = \Delta p + PF(p+q), \qquad q_t = \Delta q + (I-P)F(p+q).$$

The dynamics on the inertial manifold can be rewritten as

$$p_t = \Delta p + PF(p+q), \qquad q(t) = \Phi(p).$$

Numerically, the nonlinear Galerkin methods, also called approximate inertial manifolds (AIM) methods, make an approximation to the graph. In these methods, the evolution on a coarse mesh (i.e., low Fourier modes) uses information from the fine scale (i.e., high modes) at each time step, where a simpler form of equation is solved.

To deal with deterministic PDEs with nonsmooth initial data, long transients, or highly oscillatory time dependent forcing, Yinnian and Mattheij [15] introduced a dynamic form of postprocessing, where the following system is approximated:

$$p_t = \Delta p + PF(p), \qquad q_t = \Delta q + (1-P)F(p).$$

It extends the approach of [7], where a fine mesh solution is found at the end of the computations. For the dynamic postprocessing approach, both the coarse and fine mesh approximations are evolved in time and, unlike a traditional AIM approach, there is no communication from fine to coarse mesh until the end of the computation.

Indeed, this communication was one of the main reasons that the AIM approach was computationally less efficient than a standard Galerkin method; see [7, 8].

He and Mattheij [15] discretized the PDEs in space by a Galerkin method and in time by an implicit Euler method, examined stability and convergence of the scheme, and propose this as a computationally more efficient method. In [22] the postprocessing method is examined from a truncation analysis point of view. From a perturbation expansion for the high modes and by keeping terms to different orders, they obtain systems that correspond to the postprocessed Galerkin method and this yields convergence theory. Furthermore, from numerics based on Burgers' equation with highly oscillatory forcing, they show that postprocessing methods are more efficient and have an improved rate of convergence. These results suggest that postprocessing may be advantageous for a stochastically forced PDE.

Although inertial manifolds have been shown to exist for stochastic PDEs [2], we do not attempt to approximate this directly here. Instead we base our method on the postprocessing approaches of [15] and [22].

3. Numerical scheme. We will consider a Fourier-based Galerkin discretization, although other spatial discretizations are possible. The time discretization may be thought of as a stochastic version of an exponential integrator proposed by [20]; for a review of these methods in the deterministic case, see [23], and for an application using a finite difference spatial discretization, see [17]. In the stochastic context such schemes are considered in [21, 24] and related schemes by [28, 18] which are of the exponential time differencing type.

We describe our numerical scheme for (1.1). Represent $u(t)$ as a Fourier series $u(t) = \sum_n u_n(t)\phi_n$ and obtain the infinite system of coupled equations

$$u_n(t) = e^{-tn^2}u_n(0) + \int_0^t e^{-(t-s)n^2}F_n\big(u(s)\big)\,ds + \int_0^t e^{-(t-s)n^2}\lambda_n^{1/2}\,d\beta_n(s), \tag{3.1}$$

where F_n is the nth component of F, so that $F(u) = \sum_n F_n(u)\phi_n$. Let $\Delta t > 0$ denote the time step and N the size of the Galerkin truncation. Consider the discretization of (1.1) at times $t_k = k\Delta t$ given by

$$u_n^N\big(t_{k+1}\big) = e^{-\Delta t n^2}\left(u_n^N(t_k) + \Delta t F_n\big(u^N(t_k)\big) + \lambda_n^{1/2}\Delta B_{k,n}\right), \tag{3.2}$$

where $|n| \le N$, the noise terms $\Delta B_{k,n} = \beta_n(t_{k+1}) - \beta_n(t_k)$, and initial data $u_n^N(0) = u_n(0)$. The relationship between (3.2) and (3.1) is quite obvious when we iterate (3.2): for $t = k\Delta t$,

$$u_n^N(t) = e^{-tn^2}u_n^N(0) + \sum_{k=0}^{\lfloor t/\Delta t\rfloor - 1} e^{-(t-t_k)n^2}\left(\Delta t F_n(u^N(t_k)) + \lambda_n^{1/2}\Delta B_{k,n}\right) \tag{3.3}$$

(no terms in the sum for $0 \le t < \Delta t$). This approximation has been studied in detail in [21] for Gevrey (exponentially smooth) noise.

We study a generalization of this method, which incorporates postprocessing terms and flexibility in the approximation of $W(t)$. The generalized method has the following form: for $|n| \le N$,

$$u_n^N(t_{k+1}) = e^{-n^2\Delta t}\Big(u_n^N(t_k) + \Delta t F_n(u^N(t_k)) + \mathbf{1}_{\{|n|\le N_w\}}\lambda_n^{1/2}\Delta B_{k,n}\Big), \tag{3.4}$$

with initial data $u_n^N(0) = u_n(0) = u_{0,n}$, where $\mathbf{1}_X$ equals 1 if X holds, 0 otherwise. The constant N_w describes the number of modes used to approximate $W(t)$; this is the first generalization and we will show the advantages in taking $N_w < N$ in certain applications. As in [21], the analysis depends on an interpolant of $u_n^N(t_k)$ in time: let

$$u_n^N(t) = e^{-n^2 t} u_n^N(0) + \sum_{k=0}^{\lfloor t/\Delta t \rfloor - 1} e^{-(t-t_k)n^2} \Big(\Delta t F_n(u^N(t_k)) + \lambda_n^{1/2} \mathbf{1}_{\{|n| \le N_w\}} \Delta B_{k,n} \Big), \tag{3.5}$$

and note that the two definitions of $u_n^N(t_k)$ agree.

Now we introduce postprocessing. Given knowledge of u^N, the following are efficiently computed:

$$q_n^N(t_{k+1}) = e^{-n^2 \Delta t} \Big(q_n^N(t_k) + \Delta t \mathbf{1}_{\{|n| \le N_p\}} F_n(u^N(t_k)) + \lambda_n^{1/2} \mathbf{1}_{\{|n| \le N_w\}} \Delta B_{k,n} \Big), \tag{3.6}$$

with initial data $q_n^N(0) = u_n(0)$ for $N < |n| \le N_p$, where N_p describes the number of nonlinear terms. Again in the analysis in section 6, we use an interpolant

$$\begin{aligned} q_n^N(t) &= e^{-n^2 t} q_n^N(0) \\ &\quad + \sum_{k=0}^{\lfloor t/\Delta t \rfloor - 1} e^{-(t-t_k)n^2} \Big(\Delta t \mathbf{1}_{\{|n| \le N_p\}} F_n(u^N(t_k)) + \lambda_n^{1/2} \mathbf{1}_{\{|n| \le N_w\}} \Delta B_{k,n} \Big). \end{aligned} \tag{3.7}$$

We seek to estimate the error in approximating $u(t)$ by $u^N(t) + q^N(t)$, where $u^N = \sum_{|n| \le N} \phi_n u_n^N$ and $q^N = \sum_{N < |n| \le \max\{N_p, N_w\}} \phi_n q_n^N$, and, in particular, to understand the best choice of N_w and N_p.

3.1. Statement of main theorem. Let $\|\cdot\|$ denote the standard $L^2(0, 2\pi)$ norm. Denote the $H^m(0, 2\pi)$ Sobolev norm for $u = \sum_n u_n \phi_n$ by

$$\|u\|_m = \|(I - \Delta)^{m/2} u\| = \left(\sum_{n \in \mathbb{Z}} (1 + n^2)^m u_n^2 \right)^{1/2}.$$

We make the following assumption of f and Q.

ASSUMPTION 3.1. *For $u_1, u_2, u \in L^2(0, 2\pi)$, for some constant K_0 and some $m, r \ge 0$,*

$$\|F(u_1) - F(u_2)\|_r \le K_0 \|u_1 - u_2\|_r, \tag{3.8}$$

$$\|F(u)\|_r \le K_0 (1 + \|u\|_r) \tag{3.9}$$

and

$$\|F(u_1) - F(u_2)\|_m \le K_0 \|u_1 - u_2\|_m, \tag{3.10}$$

$$\|F(u)\|_m \le K_0 (1 + \|u\|_m). \tag{3.11}$$

There exists a constant K_1 such that for $u \in L^2(0, 2\pi)$ and $\delta, \delta_1, \delta_2 \in H^m(0, 2\pi)$,

$$\|dF(u)\delta\|_m \le K_1 \|\delta\|_m, \tag{3.12}$$

$$\|d^2 F(u)(\delta_1, \delta_2)\|_m \le K_1 \|\delta_1\|_m \|\delta_2\|_m. \tag{3.13}$$

The covariance Q of $W(t)$ satisfies $\mathrm{Tr}(I-\Delta)^\gamma Q<\infty$*; i.e.,*

$$\sum_{n\in\mathbb{Z}}(1+n^2)^\gamma\lambda_n<\infty. \tag{3.14}$$

We have introduced three regularity parameters: γ describes regularity of the noise, r gives the regularity of the solution $u(t)$, m indicates the norm for our error analysis.

THEOREM 3.2. *Let $u_0\in H^2(0,2\pi)$, $m<\min\{r,2\}$, $0\le r\le\gamma+1$, and $\gamma>-1$. For some $\nu>0$, consider $\Delta t\to 0$ and $N\to\infty$ with $\Delta t N^2\le\nu$. For each $T>0$, there exists $K>0$ such that*

$$\left(\mathbf{E}\left[\sup_{0<t_k\le T}\|u(t_k)-u^N(t_k)-q^N(t_k)\|_m^2\right]\right)^{1/2}$$
$$\le K\Big(\Delta t+N^{-2}+\mathbf{1}_{N\le N_w}N^{-1-\gamma}+N_p^{-2-r+m}+\Delta t N_w^{1-\gamma+m}+N_w^{-1-\gamma+m}\Big),$$

where $u^N=\sum_{|n|\le N}\phi_n u_n^N$ and $q^N=\sum_{N<|n|\le\max\{N_p,N_w\}}\phi_n q_n^N$ with components defined by (3.5)–(3.7).

Proof. This is given in section 6. ∎

Note that we take limits in $\Delta t, N$ with $\Delta t N^2\le\nu$ but employ no restriction on ν. If an explicit Euler time integrator was used, we would require $\nu\le\frac{1}{2}$ [11], and the absence of this restriction is a clear advantage to the exponential time integrator.

The theorem is stated under the global Lipschitz assumption on the nonlinearity. This is the simplest setting in which to work and allows us to focus attention on postprocessing. The global Lipschitz assumption excludes many important cases, including the Allen–Cahn equation we discuss in section 4. The first approach to this problem is to change the nonlinearity without affecting the underlying model: for example, in the Allen–Cahn equation, the variable u describes the phase of some material and is only physically meaningful inside a bounded set. If we smooth out the nonlinear term at infinity, the essential features of the model remain. In section 4, we discover that our results are demonstrated without such a modification. The second approach is to develop the mathematics to include ever wider classes of nonlinearities. Approaches of this type include [16] for finite dimensional SDEs, which uses moment conditions to control the behavior of u at infinity and gain rates of convergence, and [11], which shows convergence in probability, without rates, for very general classes of f. The inclusion of these approaches in the present paper would obscure the main idea, which is postprocessing.

To understand postprocessing, we state two corollaries (using that $\Delta t N^2=\nu$). The first describes convergence for the method (3.2) for nonsmooth problems (extending work done in [21]). The second gives the values N_w, N_p that yield the best convergence rates.

COROLLARY 3.3 (no postprocessing). *Under the assumptions of Theorem* 3.2 *with $N=N_w=N_p$,*

$$\left(\mathbf{E}\left[\sup_{0<t_k\le T}\|u(t_k)-u^N(t_k)-q^N(t_k)\|_m^2\right]\right)^{1/2}\le K\Big(N^{-2}+N^{-2-r+m}+N^{-1-\gamma+m}\Big).$$

For example, with $\gamma=-1/2$ (space-time white noise), the $L^2(0,2\pi)$ error (case $m=0$) converges like $N^{-1/2}$. This is consistent with related results in the literature

(e.g., [13, 19]). For Gevrey noise and a smooth nonlinearity, the parameters r and γ may be chosen arbitrarily large, and we recover the result of [21]: for any $z > 0$, there exists a constant K such that

$$\mathbf{E}\left[\sup_{0<t_k\le T}\|u(t_k)-u^N(t_k)\|_1\right] \le K(N^{-z}+\Delta t). \tag{3.15}$$

This is faster convergence than any polynomial, although not the exponential rate found [5] for the deterministic case.

Now we turn to postprocessing.

COROLLARY 3.4 (postprocessing). *Let the assumptions of Theorem* 3.2 *hold.*

1. *If* $\gamma \ge 1$ *and* $m < \gamma - 1$, *then*

$$\left(\mathbf{E}\left[\sup_{0<t_k\le T}\|u(t_k)-u^N(t_k)-q^N(t_k)\|_m^2\right]\right)^{1/2} \le KN^{-2}$$

   with $N_p = N$ *and* $N_w = \lceil N^{2/(1+\gamma-m)}\rceil$.
2. *If* $\gamma \ge 1$ *and* $m \ge \gamma - 1$, *then*

$$\left(\mathbf{E}\left[\sup_{0<t_k\le T}\|u(t_k)-u^N(t_k)-q^N(t_k)\|_m^2\right]\right)^{1/2} \le KN^{-1-\gamma+m}$$

   with $N_p = N$ *and* $N_w = N$.
3. *If* $-1 < \gamma < 1$, *then*

$$\left(\mathbf{E}\left[\sup_{0<t_k\le T}\|u(t_k)-u^N(t_k)-q^N(t_k)\|_m^2\right]\right)^{1/2} \le KN^{-1-\gamma}$$

   with $N_p = N$ *and* $N_w = \lceil N^{(1+\gamma)/(1+\gamma-m)}\rceil$.

These choices of N_p *and* N_w *provide the best convergence rate (up to scalar multiplication).*

Proof. We wish to choose N_p and N_w in terms of N to achieve the best convergence rate with N by balancing terms in the estimate provided in Theorem 3.2. We ignore multiplying constants which do not affect the rate.

1. We can achieve an N^{-2} convergence rate by balancing N_p^{-2-r+m}, $\Delta t N_w^{1-\gamma-m}$, and $N_w^{-1-\gamma+m}$ with N^{-2}. The condition $N_p^{-2-r+m} = N^{-2}$ yields $N_p = N^{2/(2+r-m)}$ and as $N^{2/(2+r-m)} \le N$ for $m \le r$, we choose $N_p = N$. Under assumption $m < \gamma - 1$, $\Delta t N_w^{1-\gamma-m} < N^{-2}$, and so the value N_w is found by solving $N^{-2} = N_w^{-1-\gamma+m}$.
2. In the case $m > \gamma - 1$, the accuracy is limited by the term $\Delta t N_w^{1-\gamma+m}$. The condition $\Delta t N_w^{1-\gamma+m} = N_w^{-1-\gamma+m}$ implies $N_w = N$. The condition $N_p^{-2-r+m} = N_w^{-1-\gamma+m}$ implies $N_p = N^{(2+r-m)/(1+\gamma-m)}$. Because we have $N^{(2+r-m)/(1+\gamma-m)} > N$ for $r > \gamma - 1$, the choice $N_p = N$ terms is optimal.
3. We achieve an $N^{-1-\gamma}$ rate by choosing $\Delta t N_w^{1-\gamma-m}$ and $N_w^{-1-\gamma+m}$ less than $N^{-1-\gamma}$. This is achieved by taking

$$N_w \ge \max\{N^{(1-\gamma)/(1-\gamma+m)}, N^{(1+\gamma)/(1+\gamma-m)}\}.$$

   As $m \ge 0$, we take $N_w = N^{(1+\gamma)/(1+\gamma-m)}$. Balancing the terms N_p^{-2-r+m} and $N^{-1-\gamma}$ provides $N_p = N^{(1+\gamma)/(2+r-m)}$. As $m < r$ and $\gamma < 1$, we have $N^{(1+\gamma)/(2+r-m)} < N$ and choose $N_p = N$. ∎

There are a number of issues to consider: the rate of convergence, the constant for this rate, and the efficiency of the scheme. We can improve the rate of convergence by choice of N_w and there are two cases to consider. For smooth noise $\gamma \geq 1+m$, the optimal value is $N_w < N$, which saves computing random numbers for many of the components u_n^N. This has been used with good effect in [27] for a Gevrey smooth noise. Note that $N_w \to 1$ as $\gamma \to \infty$. In practice, it is important for $N_w \to \infty$ as we ask for more accuracy and to take to enough modes to resolve the noise.

For nonsmooth noise ($\gamma < 1$), the optimal $N_w > N$, which implies that the postprocessing corrections q_n^N are Gaussian processes

$$q_n^N(t_{k+1}) = e^{-n^2\Delta t}\Big(q_n^N(t_k) + \lambda_n^{1/2}\mathbf{1}_{\{|n|\leq N_w\}}\Delta B_{k,n}\Big). \tag{3.16}$$

Thus, computing the postprocessing update is straightforward and cheap. To compare solutions for a single realization of $W(t)$, q_n^N must be found by time stepping. For weak approximation, it will be more efficient to compute and sample from the Gaussian distribution at the final time.

Our analysis predicts no improvement in the rate of convergence from postprocessing the nonlinear term. This contrasts with results on postprocessing in the deterministic case, where there is a gain in the rate of convergence [7, 8] (though this gain is often outweighed by extra computational cost).

4. Numerics. Consider the one-dimensional Allen–Cahn equation with noise:

$$du = \Big[\alpha u_{xx} + u - u^3\Big]\,dt + dW(t), \qquad u(0) = u_0, \tag{4.1}$$

with periodic boundary conditions on $[0, 2\pi)$. For numerical calculations, we take the diffusion coefficient $\alpha = 1/36$. We always take noise white in time and vary the spatial regularity γ; see (3.14).

To test the numerics, "true" solutions were computed by a standard Galerkin approximation with $N = 2^{11}$ modes and a time step $\Delta t = 5 \times 10^{-6}$. To avoid aliasing errors, the nonlinear term was computed with $2N$ terms (more than the optimal number of terms suggested by the 2/3 rule [1]). For a discussion of the role of aliasing in postprocessing (in the deterministic case), see [9].

Sample "true" solutions are plotted in Figure 4.1; this shows (left) the effect of different spatial regularity in real space and (right) the corresponding log-log plot in Fourier space. In real space, the solutions are smoother as the regularity of the noise increases. This is confirmed by the decay of the Fourier modes, and we see numerically that $r = \gamma + 1$, consistent with the results of Lemma A.1. Essentially, we gain a derivative on the regularity of the solution over the noise.

Let $\hat{N}$ denote a parameter for postprocessing (either $2N$, $4N$, $8N$, or N^2 in experiments). The "true" solutions were used to compute errors for the following approximations:

***Galerkin*:** A standard Galerkin approximation, from solving (3.4) with $N_w = N$.

***PP Full*:** A full postprocessed solution, from solving (3.4) and (3.6) with $N_w = N_p = \hat{N}$.

***PP Noise*:** A postprocessed solution on noise only, from solving (3.4) and (3.16) with $N_p = N$, $N_w = \hat{N}$.

We examine the rate of convergence and efficiency by a mean cpu time. From a practical point of view, plots of cpu time versus error can be interpreted in two ways: either fix a desired accuracy and see how long it would take to achieve, or fix a time

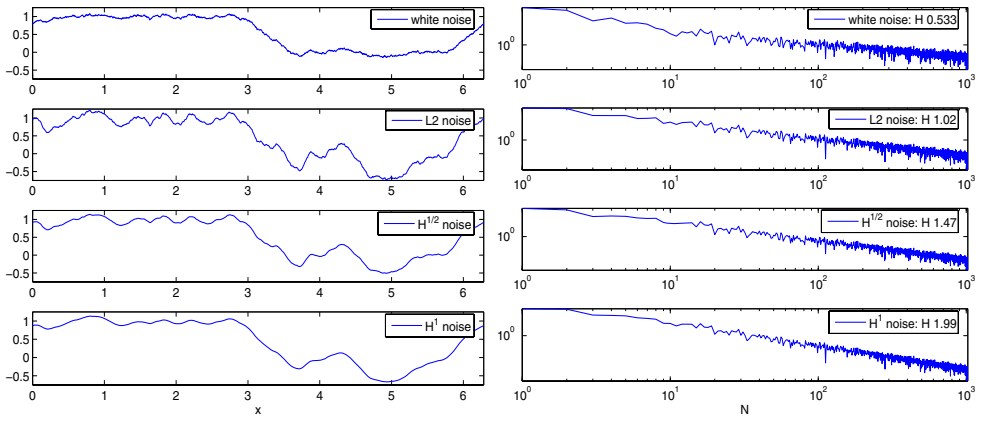


FIG. 4.1. *Plot (left) of "true" solutions at time $t = 1$ for $\gamma = -0.5, 0, 0.5, 1.0$ for one realization of the noise. Plot (right) is the corresponding log-log plot of the Fourier coefficients at time $t = 1$, which shows that for $\gamma > 0$ the solutions are in a Sobolev space H^r with $r = 0.5, 1, 1.5, 2$.*

and see how accurate a solution can be computed in that time. The expectation is computed from 10 samples, and we examine the root mean square of the error at time $t = 1$ in an appropriate norm. Normally we take the L^2 norm ($m = 0$) or H^1 norm ($m = 1$).

On the plots below we draw a line with slope equal to the predicted rate of convergence for *Galerkin*. We also report in the legend the observed slope from the data for the rate of convergence.

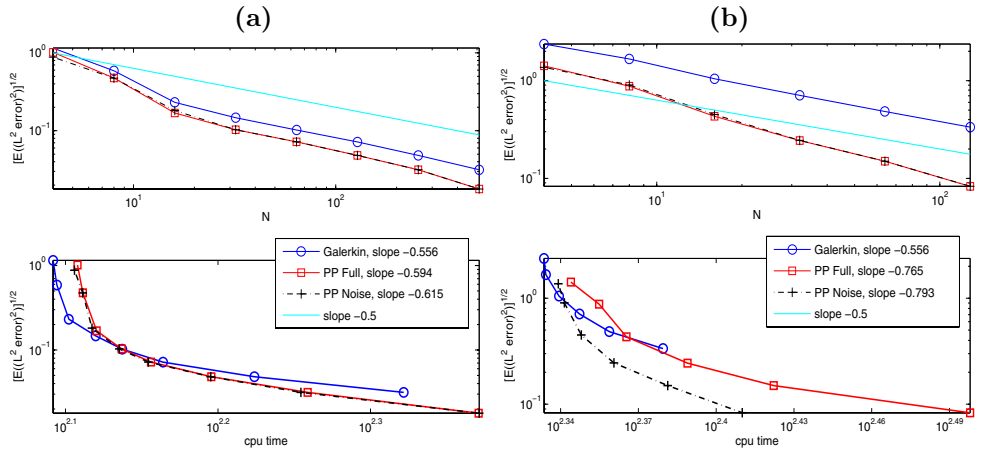


FIG. 4.2. *Space-time white noise* (a) *with $\hat{N} = 2N$ and* (b) *with $\hat{N} = 8N$. Plots show the L^2 error (top) rate of convergence and (below) plot of efficiency (cpu time).*

We examine the rates of convergence and computational efficiency for $W(t)$ defined by (1.2) with $\lambda_n = (1 + n^2)^{-\gamma}|n|^{-1}$, $n \neq 0$, and $\lambda_0 = 0$. We consider $\gamma = -1/2$ (*space-time white noise*), $\gamma = 0$ (L^2 *noise*), and $\gamma = 1/2, 1, 2$ (H^γ *noise*). Our predictions for the numerics are based on Theorem 3.2 where, motivated by Lemma A.1, we assume that $r = \gamma + 1$.

4.1. Space-time white noise: $\gamma = -\frac{1}{2}$. We observe in Figure 4.2 (top) the theoretically predicted rates of convergence for *Galerkin*: the L^2 error decays like $N^{-1/2}$. There is no convergence for H^1 error.

Postprocessing is not expected to improve the rate of convergence in the L^2 norm, as $N_w = N$ in Corollary 3.4. With $\hat{N} = 2N$, this is supported by computations; see Figure 4.2 (a) (top) where the postprocessing has no beneficial effect and the observed rate is the same as for *Galerkin*. However, there is an improvement in the error constant, and for $N > 32$ modes postprocessing is more efficient; see Figure 4.2 (a) (bottom). Taking this further and using more modes for the postprocessing, Figure 4.2 (b) shows *PP Full* and *PP Noise* with $\hat{N} = 8N$. The numerics suggest a rate of convergence faster than the theoretical one. This is encouraging, although the resolution is coarse and the theoretical rate may reappear for larger N.

We clearly see the computational advantage of *PP Noise* compared to *PP Full* and *Galerkin* in Figure 4.2 (a) and (b) bottom. Postprocessing on the noise terms only is far more efficient.

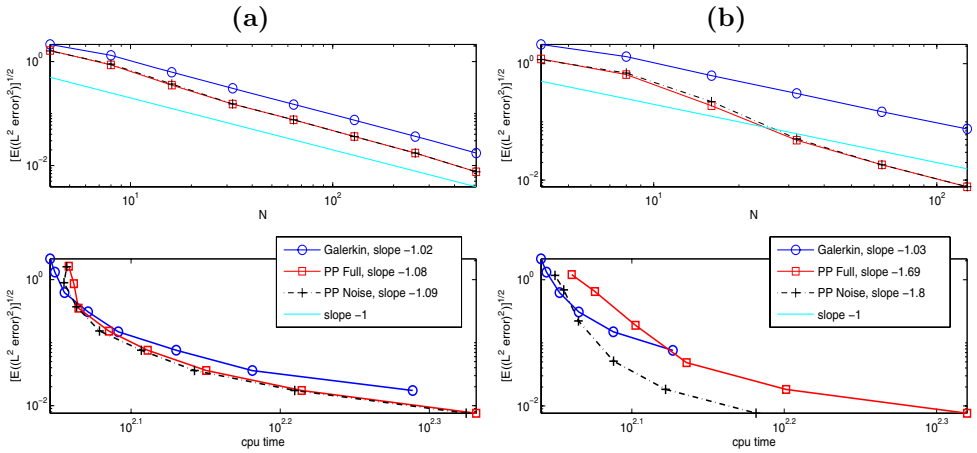


FIG. 4.3. *For L^2 noise, we plot* (a) *the L^2 error with $\hat{N} = 2N$ and* (b) *the L^2 error with $\hat{N} = 8N$. Top shows error against N, and bottom error against average cpu time.*

4.2. L^2 noise. This is similar to white noise: for *Galerkin*, the L^2 error decays like N^{-1}, which is observed in Figure 4.3 (a) and (b), and the H^1 error does not converge. In theory, postprocessing offers no improvement. In practice, there is an improvement in the error constant and an improvement in efficiency for $\hat{N} = 2N$ and further improvement for $\hat{N} = 8N$. See Figure 4.3 (a) and (b).

4.3. $H^{1/2}$ noise. Corollary 3.3 predicts convergence of the L^2 error like $N^{-3/2}$ and the H^1 error like $N^{-1/2}$ for *Galerkin*, and these rates are observed in Figure 4.4 (a) and (b). With postprocessing, the optimal rate for the L^2 error is not changed, and the H^1 error is like $N^{-3/2}$ if $N_w = N^3$. It is impractical to calculate with N^3 postprocessing terms for large N, and instead we look at $\hat{N} = 2N, 4N, 8N$. Figure 4.4 shows the effect of increasing $\hat{N}$ for L^2 error (left) and H^1 error (right) with $\hat{N}$ increasing top to bottom. For L^2 and H^1 errors, increasing $\hat{N}$ improves the error and seems to improve the rate of convergence—although this is not expected from the analysis for L^2 and we are a long way from taking the predicted N^3 modes for H^1. We clearly see that *PP Noise* is the most efficient method.

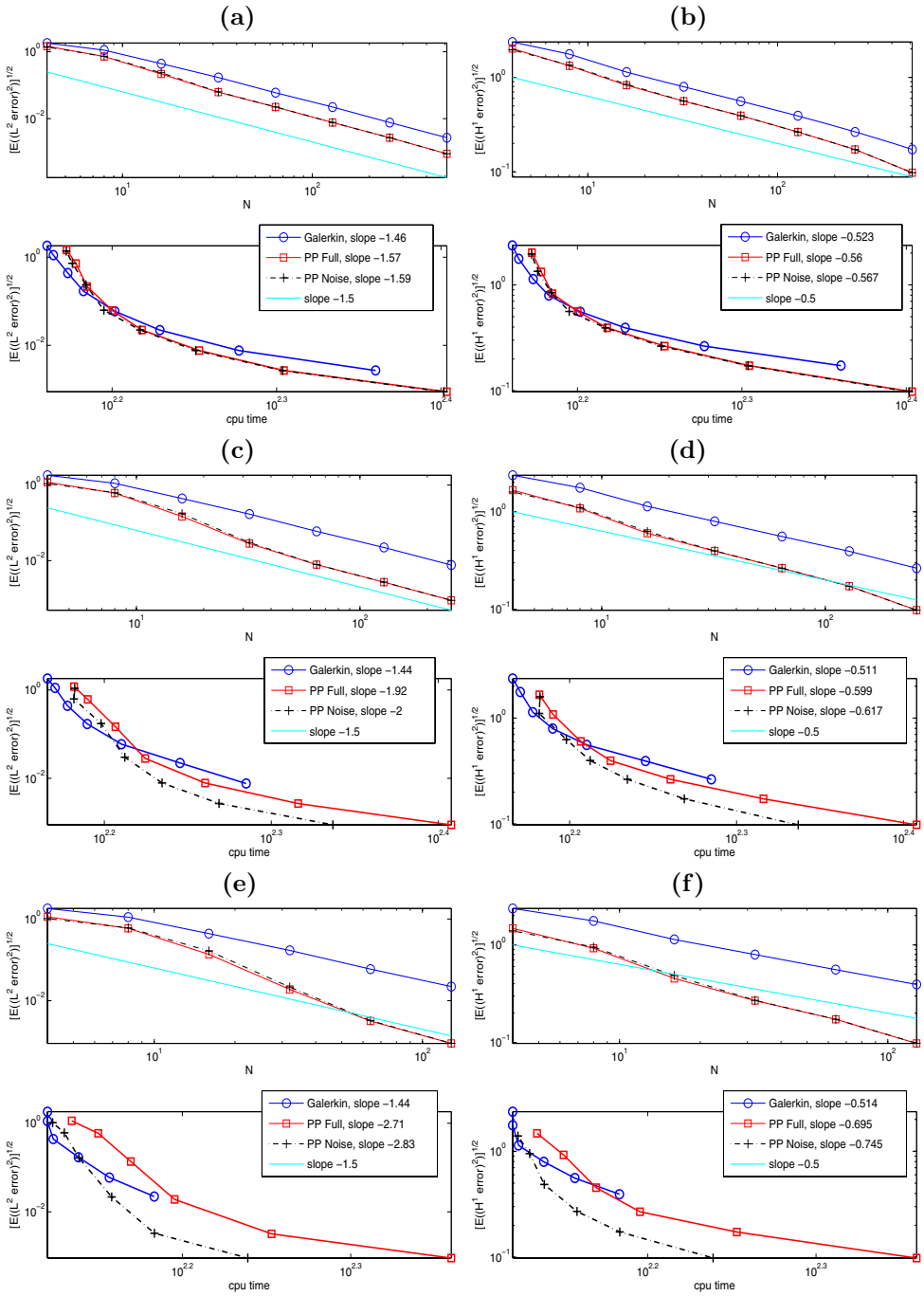


FIG. 4.4. *For $H^{1/2}$ noise, we examine the number of postprocessing terms $\hat{N}$. In* (a) *and* (b) *we take $\hat{N} = 2N$,* (c) *and* (d) *$\hat{N} = 4N$,* (e) *and* (f) *$\hat{N} = 8N$ with L_2 error (left) and H^1 error (right). For each case, we show plots of error against N (above) and error against cpu time (below).*

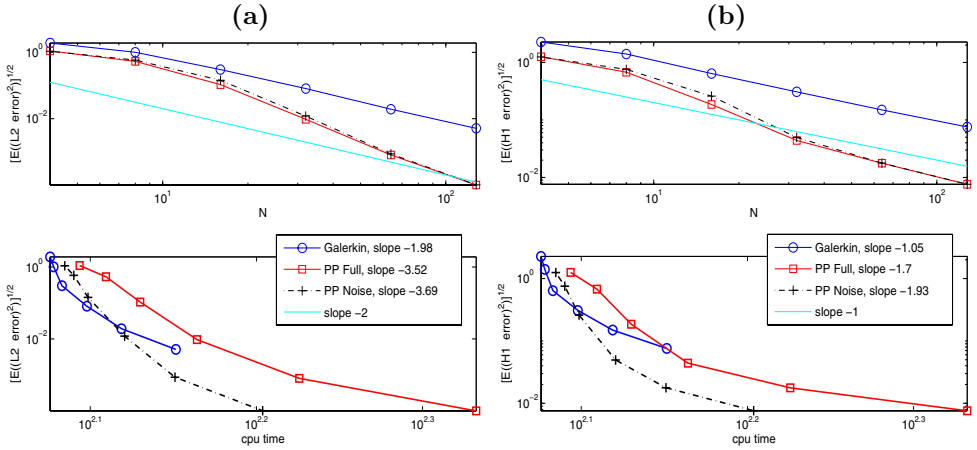


FIG. 4.5. *For H^1 noise with $\hat{N} = 8N$, we plot* (a) *L^2 error and* (b) *H^1 error. The top shows error against N, and the bottom shows error against cputime.*

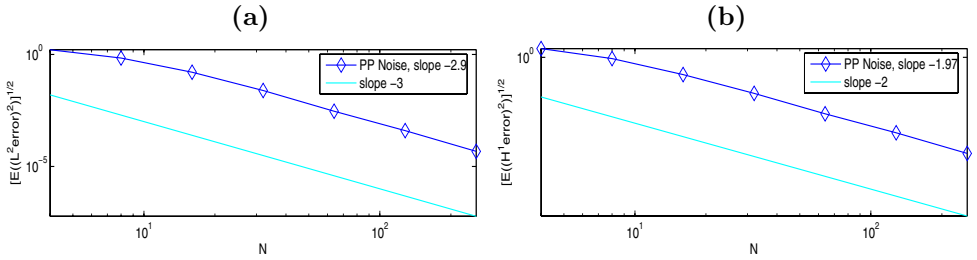


FIG. 4.6. *For H^2 noise, we see* (a) *faster than the predicted rate of convergence for the L^2 error with the optimal value $\hat{N} = N^{2/3}$ and* (b) *the N^{-2} convergence rate is achieved for the H^1 error, with $\hat{N} = N^{2/3}$ rather than the theoretical rate of $N_w = N$.*

4.4. H^1 noise. Corollary 3.3 predicts that the *Galerkin* L^2 error decays like N^{-2} and H^1 error decays like N^{-1}, as observed in Figure 4.5. This is the limiting case in Corollary 3.4, where we find N^{-2} convergence by taking $N_w = N$ for L^2 error and $N_w = N^2$ for H^1 error; the solution is smooth in space, and accuracy is now limited by time stepping. It is impractical to calculate with N^2 postprocessing terms for large N, and instead we look at $\hat{N} = 8N$; Figure 4.5 shows (a) the L^2 error and (b) the H^1 error. The postprocessing methods give smaller errors and are more efficient, in particular *PP Noise*.

4.5. H^2 noise. The optimal number of modes is $N_w = N^{2/3}$ for the L^2 error, giving N^{-2} convergence. We see in Figure 4.6 (a) that the L^2 error is converging faster than the theoretical rate, close to N^{-3}. Here we see a limitation of the analysis: the theoretical convergence rate is limited to an N^{-2} rate because of time stepping and regularity of the initial data. In this case, the error is dominated by the spatial approximation of smooth problems, which may decrease like N^{-3}, similar to rates described in (3.15).

The H^1 error in (b) shows N^{-2} convergence—although only $\hat{N} = N^{2/3}$ modes are used rather than the theoretical optimum value N. In this case, the accuracy is determined by approximation of the deterministic terms, and we are unable to

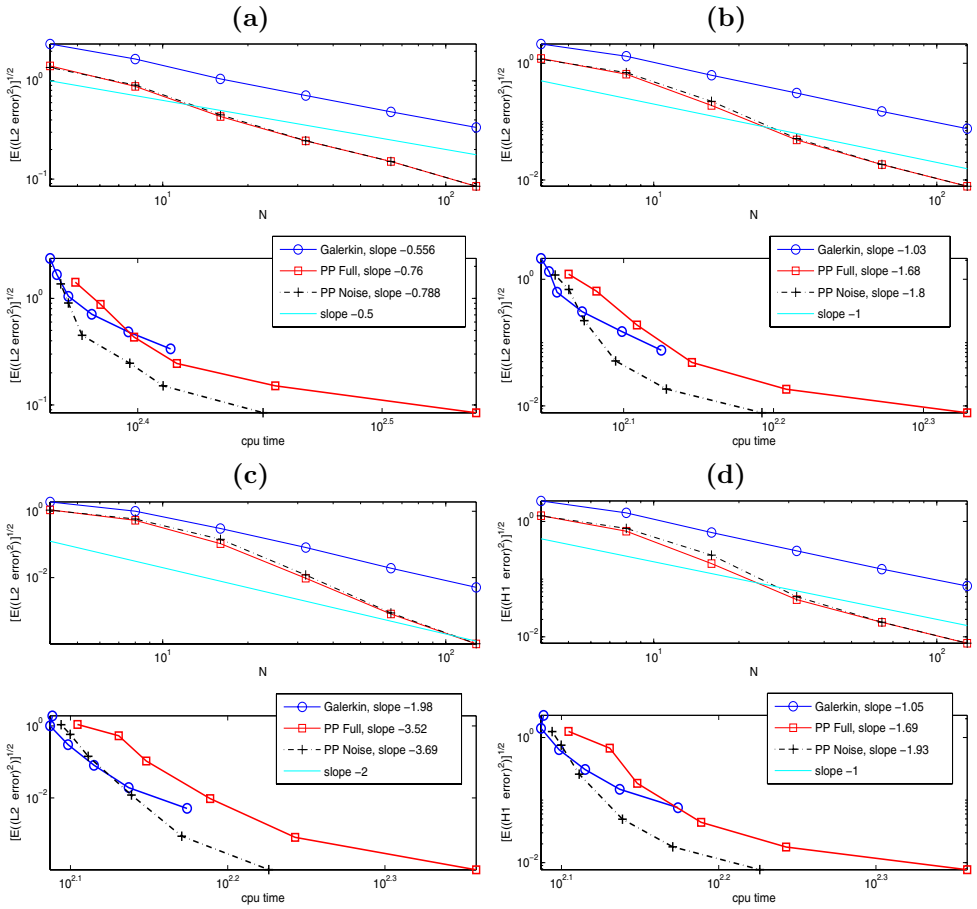


FIG. 4.7. *Postprocessing for the implicit Euler–Maruyama method. In* (a) *white noise and* L^2 *error,* (b) L^2 *noise and* L^2 *error,* (c) H^1 *noise and* L^2 *error,* (d) *again* H^1 *noise but with* H^1 *error.*

increase the number of modes to see the theoretical optimal number for N_w bite.

4.6. Postprocessing implicit Euler–Maruyama. Postprocessing is effective for other time stepping algorithms. In Figure 4.7, we plot results of experiments with the implicit Euler–Maruyama scheme. We take $\hat{N} = 8N$ and plot (a) the L^2 error for white noise, (b) the L^2 error with L^2 noise, (c) L^2 error with H^1 noise, and (d) H^1 error with H^1 noise. Again *PP Noise* is the most efficient of the methods, and there appears to be an improvement in the rate of convergence in addition to the constant. These trends are identical to those found in Theorem 3.2 and shown in Figures 4.2–4.6.

5. Conclusions. Theorem 3.2 shows that postprocessing can improve the rate of convergence over a standard Galerkin method for stochastic PDEs. For nonsmooth forcing, the best number of modes is greater than the standard Galerkin method. For smooth noise, as observed in [27], the optimal number of modes is smaller. With the smooth nonlinearity in (4.1), it is flexibility in the number of modes that approximate $W(t)$ that is key. This was confirmed in numerics. We found that postprocessing on the noise improves on the convergence and efficiency of the standard Galerkin

approximation and that the contribution from the (smooth) nonlinearity in the post-processing is negligible. This improvement in efficiency over the standard Galerkin method holds true for all spatial regularities of the noise that we tested.

It is often computationally prohibitive to use the number of modes suggested by the theorem. From a practical point of view, improvements were noted with $N_w = 2N$ even when the theoretical optimum number of nodes is $N_w = N^2$. For nonsmooth noise, we found numerically that taking $N_w = 8N$ gave a good compromise between the extra effort involved and accuracy. Indeed it seems we get a rate of convergence not predicted by the theory.

For smooth noise, our numerics suggest a convergence rate faster than that predicted by the theorem. From [21], it is known that for exponentially smooth noise a faster than polynomial convergence is available for smooth problems. Such techniques have not been used in the present paper, and the results we give are optimal for the H^2 initial data and time stepping method studied.

Finally, although our analysis is for the scheme (3.2), this approach works equally well for other time stepping methods, such as the implicit Euler–Maruyama time stepping scheme. Our presentation is for a Galerkin-based approximation; however, postprocessing can easily be extended to other spatial discretizations using, for example, two grids.

6. Proof of main theorem. We prove Theorem 3.2 by estimating

$$\mathbf{E}\left[\sup_{0\le t_j\le t'}\|u(t_j)-u^N(t_j)-q^N(t_j)\|_m^2\right]$$

for $0 \le t' \le T$ and applying Gronwall's lemma. To estimate terms, we use a generic constant K which varies between instances but is independent of Δt and N (it may depend on (1.1) and the length of time integration T and constant ν). Consider the difference of the variation of constants formulae (3.1), (3.5), and (3.7). Split into Fourier modes with $|n| \le N_p$ and $|n| > N_p$ and by nonlinear and noise terms.

Nonlinear terms: Modes $|n| \le N_p$.

$$\begin{aligned}
&\mathbf{E}\sup_{0\le t_j\le t'}\sum_{|n|\le N_p}(1+n^2)^m\left|\sum_{k=0}^{j-1}\times\int_{t_k}^{t_{k+1}}e^{-(t_j-t_k)n^2}(e^{(s-t_k)n^2}F_n(u(s))-F_n(u^N(t_k)))\,ds\right|^2\\
&\qquad=\sum_{|n|\le N_p}\mathbf{E}\Bigg[\sup_{0\le t_j\le t'}\sum_{k=0}^{j-1}\int_{t_k}^{t_{k+1}}e^{-(t_j-t_k)n^2}(1+n^2)^{m/2}\\
&\qquad\Big(\Big(F_n(u(s))-F_n(u(t_k))\Big)+\Big(F_n(u(t_k))-F_n(u^N(t_k)+q^N(t_k))\Big)\\
&\qquad+\Big(F_n(u^N(t_k)+q^N(t_k))-F_n(u^N(t_k))\Big)+\Big((e^{(s-t_k)n^2}-1)F_n(u(s))\Big)\Big)\,ds\Bigg]^2\\
&\qquad\le K(\mathbf{NL}_1+\cdots+\mathbf{NL}_4),
\end{aligned}$$

where the four terms $\mathbf{NL}_i$ are analyzed below.

The first term. Fix t_j and consider $k \le j-1$. Define

$$L_{k,n}=\int_{t_k}^{t_{k+1}}e^{-(t_j-t_k)n^2}(1+n^2)^{m/2}\Big(F_n(u(s))-F_n(u(t_k))\Big)\,ds,$$

and let

$$\mathbf{NL}'_1 = \sum_{|n|\le N_p} \mathbf{E}\left[\sum_{k=0}^{j-1} L_{k,n}\right]^2. \tag{6.1}$$

Write $U_k = u(t_k)$ and $u(s) = u(t_k) + \delta_s$ for $t_k \le s < t_{k+1}$; then

$$F_n(u(s)) - F_n(u(t_k)) = dF_n(U_k)\delta_s + \int_0^1 \int_0^\eta d^2F_n(U_k + \xi\delta_s)(\delta_s, \delta_s)\, d\xi\, d\eta.$$

In the following argument we neglect the remainder term, which can be dealt with easily under (3.13). Denote by $\mathcal{F}_t$ the filtration for the Wiener process $W(t)$. For $k > i$, under (3.12), the cross terms in (6.1)

$$\begin{aligned}\sum_{|n|\le N_p} \mathbf{E}L_{k,n}L_{i,n} = &\sum_{|n|\le N_p} (1+n^2)^m \mathbf{E}\bigg[\int_{t_k}^{t_{k+1}} e^{-(t_j-t_k)n^2}\mathbf{E}\Big[dF_n(U_k)\delta_s|\mathcal{F}_{t_k}\Big]\, ds \\ &\times \int_{t_i}^{t_{i+1}} e^{-(t_j-t_i)n^2} dF_n(U_i)\delta_s\, ds\bigg] + \text{higher order terms (h.o.t.)} \\ \le\ & K\Delta t^4,\end{aligned}$$

because $dF_n(U_i)\delta_s$ is $\mathcal{F}_{t_k}$ measurable and $\|\mathbf{E}\big[dF(U_k)\delta_s|\mathcal{F}_{t_k}\big]\|_m \le K\Delta t$. As

$$\left[\int_{t_k}^{t_{k+1}} \phi(s)\, ds\right]^2 \le (t_{k+1}-t_k)\int_{t_k}^{t_{k+1}} \phi(s)^2\, ds, \text{ for } \phi \in L^2(0,T),$$

$$\sum_{|n|\le N_p} \mathbf{E}L_{k,n}^2 \le \Delta t \sum_{|n|\le N_p} \int_{t_k}^{t_{k+1}} \mathbf{E}\big[e^{-(t_j-t_k)n^2}(1+n^2)^{m/2} dF_n(U_k)\delta_s\big]^2\, ds + \text{h.o.t.}$$

Here

$$\begin{aligned}&\sum_{|n|\le N_p} \int_{t_k}^{t_{k+1}} \mathbf{E}\big[e^{-(t_j-t_k)n^2}(1+n^2)^{m/2} dF_n(U_k)\delta_s\big]^2\, ds \\ &\qquad \le \int_{t_k}^{t_{k+1}} \mathbf{E}\big[\|dF_n(U_k)\|_m^2 \cdot \|\delta_s\|_m^2\big]\, ds.\end{aligned}$$

Because $\mathbf{E}\|u^N(t) - u^N(s)\|_m^2 \le K|t-s|\|u_0\|_m^2$ and (3.12) holds, we conclude that

$$\sum_{|n|\le N_p} \int_{t_k}^{t_{k+1}} \mathbf{E}\big[e^{-(t_j-t_k)n^2}(1+n^2)^{m/2} dF_n(U_k)\delta_s\big]^2\, ds \le K\Delta t^2.$$

Thus, we may estimate

$$\mathbf{NL}'_1 \le \sup_{0\le t_j\le t'} \sum_{|n|\le N_p} \left\{\sum_{k=0}^{j-1} \mathbf{E}\big[L_{k,n}\big]^2 + \sum_{k,i=0,\, k\ne i}^{j-1} \mathbf{E}L_{k,n}L_{i,n}\right\} \le K\Delta t^2.$$

Apply the Doob martingale inequality to get

$$\mathbf{E}\Big[\sup_{0\le t_j\le t'} NL_1'\Big] \le 4K\Delta t^2.$$

The second term.

$$\begin{aligned}\mathbf{NL}_2 &= \sum_{|n|\le N_p}(1+n^2)^m\mathbf{E}\Big[\sup_{0\le t_j\le t'}\sum_{k=0}^{j-1}\int_{t_k}^{t_{k+1}} e^{-(t_j-t_k)n^2}\\ &\qquad\qquad\times\Big(|F_n(u(t_k))-F_n(u^N(t_k)+q^N(t_k))|\Big)\,ds\Big]^2\\ &\le \int_0^{t'}\sum_{|n|\le N_p}\mathbf{E}\Big[\sup_{0\le t_k\le t}(1+n^2)^m|F_n(u(t_k))-F_n(u^N(t_k)+q^N(t_k))|^2\Big]\,dt.\end{aligned}$$

Using (3.10),

$$\mathbf{NL}_2 \le K\int_0^{t'}\mathbf{E}\Big[\sup_{0\le t_k\le t}\|u(t_k)-u^N(t_k)-q^N(t_k)\|_m^2\Big]\,dt.$$

The third nonlinear term.

$$\begin{aligned}\mathbf{NL}_3 &= \sum_{|n|\le N_p}(1+n^2)^m\mathbf{E}\Big[\sup_{0\le t_j\le t'}\sum_{k=0}^{j-1}\int_{t_k}^{t_{k+1}} e^{-(t_j-t_k)n^2}\\ &\qquad\qquad\times\Big(|F_n(u^N(t_k)+q^N(t_k))-F_n(u^N(t_k))|\Big)\,ds\Big]^2\\ &\le \sum_{|n|\le N_p}(1+n^2)^m\mathbf{E}\Big[\sup_{0\le t_j\le t'}|F_n(u^N(t_j)+q^N(t_j))-F_n(u^N(t_j))|\\ &\qquad\qquad\times\sum_{k=0}^{j-1}\int_{t_k}^{t_{k+1}} e^{-(t_j-t_k)n^2}\,ds\Big]^2\\ &\le \sum_{0<|n|\le N_p}(1+n^2)^m\mathbf{E}\Big[\sup_{0\le t_k\le t'}|F_n(u^N(t_k)+q^N(t_k))-F_n(u^N(t_k))|\frac{1}{n^2}\Big]^2\\ &+\ \mathbf{E}\Big[\sup_{0\le t_k\le t'}|F_0(u^N(t_k)+q^N(t_k))-F_0(u^N(t_k))|\Big].\end{aligned}$$

Choose $m\le 2$, then using (3.8),

$$\begin{aligned}\mathbf{NL}_3 &\le \sum_{|n|\le N_p}\mathbf{E}\Big[\sup_{0\le t_k\le t'}(1+n^2)^m|F_n(u^N(t_k)+q^N(t_k))-F_n(u^N(t_k))|^2\Big]\\ &\le\ K\int_0^{t'}\mathbf{E}\Big[\sup_{0<t_k\le t}\|q^N(t_k)\|^2\Big]\,dt.\end{aligned}$$

Finally, from Lemma A.2,

$$\mathbf{NL}_3 \le K(N^{2(-2)}+\mathbf{1}_{N\le N_w}N^{2(-1-\gamma)}+N^{2(-2-r)}).$$

The fourth nonlinear term.

$$\mathbf{NL}_4 = \sum_{|n|\le N_p} (1+n^2)^m \mathbf{E}\Big[\sup_{0\le t_j\le t'} \sum_{k=0}^{j-1} \int_{t_k}^{t_{k+1}} e^{-(t_j-t_k)n^2}\Big(|(e^{(s-t_k)n^2}-1)F_n(u(s))|\Big)\,ds\Big]^2$$
$$\le \sum_{|n|\le N_p} (1+n^2)^m \mathbf{E}\Big[\sup_{0\le t_j\le t'} |F_n(u(t_j))|^2 \sum_{k=0}^{\lfloor t/\Delta t\rfloor -1} e^{-(t_j-t_k)n^2} K\Delta t^2 n^2\Big]^2.$$

Note that for $0 \le \Delta t n^2 \le \nu$

$$\int_{t_k}^{t_{k+1}} |e^{(s-t_k)n^2}-1|\,ds \le \Big(\frac{e^{\Delta t n^2}-1}{n^2} - \Delta t\Big) \le n^{-2}(K\Delta t^2 n^4 e^{n^2\Delta t}) \le K\Delta t^2 n^2 e^{\nu}$$

and

$$\sum_{k=1}^{\infty} e^{-kn^2\Delta t} \le \frac{1}{1-e^{-n^2\Delta t}} \le \frac{K}{n^2\Delta t}.$$

Thus, using (3.11),

$$\mathbf{NL}_4 \le K \sum_{|n|\le N_p} (1+n^2)^m \mathbf{E}\Big[\sup_{0\le s\le t'} |F_n(u(s))|^2 \Delta t\Big]^2$$
$$\le K\Delta t^2 \Big(1 + \mathbf{E}\Big[\sup_{0\le s\le t'} \|u(s)\|_m\Big]^2\Big).$$

By (3.10) and Lemma A.1,

$$\mathbf{NL}_4 \le K\,\Delta t^2.$$

Nonlinear terms: Modes $|n| > N_p$. Consider now the tail of the expansion of $u(t)$; i.e., the modes not included in either u^N or q^N. If $r > m$,

$$\mathbf{TAIL} = \mathbf{E}\Big[\sup_{0\le t_j\le t'} \sum_{|n|>N_p} (1+n^2)^m \Big| \int_0^{t_j} e^{-(t_j-s)n^2} F_n(u(s))\,ds\Big|^2\Big]$$
$$\le K\Big(\int_0^{t'} (1+N_p^2)^{-(r-m)/2} e^{-(t_j-s)N_p^2}\,ds\Big)^2 \mathbf{E}\Big[\sup_{0\le s\le t'} \|F(u(s))\|_r^2\Big].$$

By (3.9) and Lemma A.1,

$$\mathbf{TAIL} \le K\,N_p^{2(m-2-r)}.$$

Noise with modes $|n| \le N_w$.

$$\mathbf{NOISE}_1 = \mathbf{E}\Big[\sup_{0<t_j\le t'} \sum_{|n|\le N_w} (1+n^2)^m$$
$$\times \Big|\sum_{k=0}^{j-1}\Big(\int_{t_k}^{t_{k+1}} e^{-(t_j-s)n^2}\lambda_n^{1/2} d\beta_n(s) - e^{-(t-t_k)n^2}\lambda_n^{1/2}\Delta B_{k,n}\Big)\Big|^2\Big]$$
$$\le \sum_{|n|\le N_w} (1+n^2)^m |\lambda_n| \mathbf{E}\Big[\sup_{0<t_j\le t'} \int_0^{t_j} (e^{-(t_j-s)n^2} - e^{-(t_j-\lfloor s/\Delta t\rfloor \Delta t)n^2}) d\beta_n(s)\Big]^2.$$

By Doob's martingale inequality

$$\begin{aligned}\mathbf{NOISE}_1 \leq\ & 4\sum_{|n|\leq N_w}(1+n^2)^m|\lambda_n|\int_0^{t'}(e^{-(t_j-s)n^2}-e^{-(t_j-\lfloor s/\Delta t\rfloor\Delta t)n^2})^2\,ds\\ =\ & 4\sum_{|n|\leq N_w}(1+n^2)^m|\lambda_n|\int_0^{t'}e^{-2(t_j-s)n^2}(1-e^{-(s-\lfloor s/\Delta t\rfloor\Delta t)n^2})^2\,ds.\end{aligned}$$

Note that $1-e^{-tn^2}\leq tn^2$ for $0\leq t\leq\Delta t$ and

$$\int_0^{t'}e^{-2(t_j-s)n^2}(1-e^{-(s-\lfloor s/\Delta t\rfloor\Delta t)n^2})^2\,ds\leq(\Delta tn^2)^2\int_0^{t'}e^{-2(t_j-s)n^2}\,ds\leq K\Delta t^2n^2.$$

Hence

$$\begin{aligned}\mathbf{NOISE}_1 \leq\ & 4\sum_{|n|\leq N_w}(1+n^2)^m|\lambda_n|\Delta t^2n^2\\ \leq\ & K\Delta t^2(1+N_w^2)^{(1+m-\gamma)}\sum_{|n|\leq N_w}(1+n^2)^\gamma|\lambda_n|\\ \leq\ & K\Delta t^2(1+N_w^2)^{(1+m-\gamma)},\end{aligned}$$

under (3.14).

Noise with modes $|n|>N_w$.

$$\begin{aligned}\mathbf{NOISE}_2 =\ & \mathbf{E}\left[\sup_{0\leq t_j\leq t'}\sum_{|n|>N_w}(1+n^2)^m\left|\int_0^{t_j}e^{-(t_j-s)n^2}\lambda_n^{1/2}d\beta_n(s)\right|^2\right]\\ \leq\ & 4(1+N_w^2)^{m-\gamma}\frac{1-e^{-t'N_w^2}}{N_w^2}\sum_{|n|\geq N_w}\lambda_n(1+n^2)^\gamma\leq K\,N_w^{2(m-1-\gamma)},\end{aligned}$$

using (3.14).

Conclusion. We have achieved the following inequality:

$$\begin{aligned}&\mathbf{E}\left[\sup_{0\leq t_j\leq t'}\|u(t_j)-u^N(t_j)-q^N(t_j)\|_m^2\right],\\ &\leq K\Bigg(\Delta t^2+(N^{2(-2)}+\mathbf{1}_{N\leq N_w}N^{2(-1-\gamma)}+N^{2(-2-r)})+N_p^{2(-2-r+m)}+\Delta t^2N_w^{2(-\gamma+1+m)}\\ &+N_w^{2(-1-\gamma+m)}+\int_0^T\mathbf{E}\left[\sup_{0<t_k\leq t}\|u(t_k)-u^N(t_k)-q^N(t_k)\|^2\right]dt\Bigg).\end{aligned}$$

Note $N^{2(-2-r)}\leq N^{2(-2)}$, and then Gronwall's lemma provides

$$\begin{aligned}\mathbf{E}\left[\sup_{0\leq t\leq t'}\|u(t)-u^N(t)-q^N(t)\|_m^2\right]\leq\ & K\Big(\Delta t^2+N^{2(-2)}+\mathbf{1}_{N\leq N_w}N^{2(-1-\gamma)}\\ &+N_p^{2(-2-r+m)}+\Delta t^2N_w^{2(-\gamma+1+m)}+N_w^{2(-\gamma-1+m)}\Big).\end{aligned}$$

This completes the proof of Theorem 3.2. □

Appendix (lemmas). We collect two elementary lemmas used in the proof of the main theorem.

LEMMA A.1. *For $r \le \gamma + 1$,*

$$\mathbf{E} \sup_{0\le t\le T} \|u(t)\|_r^2 \le K(1 + \|u_0\|_r^2).$$

Proof. Examine the nonlinear term in (3.1) under (3.9):

$$\mathbf{E}\left[\sup_{0\le t\le t'} \sum_n \left|(1+n^2)^{r/2} \int_0^t e^{-(t-s)n^2} F_n\big(u(s)\big)\, ds\right|^2\right] \le K \int_0^{t'} \left(1 + \mathbf{E}\left[\sup_{0\le s\le t} \|u(s)\|_r^2\right]\right) dt$$

and the noise term (modes with $n \ne 0$)

$$\begin{aligned}\mathbf{E}&\left[\sup_{0\le t\le t'} \sum_{n\ne 0} (1+n^2)^{r/2} \left| \int_0^t e^{-(t-s)n^2} \lambda_n^{1/2} d\beta(s)\right|^2\right] \\ &\le 4\mathbf{E}\left[\sum_{n\ne 0} (1+n^2)^{(r-\gamma)} \left| \int_0^{t'} e^{-2(t-s)n^2} (1+n^2)^{\gamma} \lambda_n ds\right|\right] \\ &\le \sum_{n\ne 0} \frac{(1+n^2)^{(r-\gamma)}}{n^2} (1+n^2)^{\gamma} \lambda_n\end{aligned}$$

using (3.14). This is finite if $r - \gamma \le 1$, so that the Gronwall lemma completes the proof. □

LEMMA A.2. *Under the assumptions of Lemma* A.1,

$$\mathbf{E} \sup_{0\le t\le T} \|q^N(t)\|^2 \le K(N^{2(-2)} + \mathbf{1}_{N\le N_w} N^{2(-1-\gamma)} + N^{2(-2-r)}).$$

Proof. We seek upper estimates on

$$\mathbf{E}\left[\sup_{0\le t\le T} \|q^N(t)\|^2\right].$$

To do this, estimate the influence of the initial data

$$\begin{aligned}\sum_{N<|n|\le \max N_p, N_w} \mathbf{E}\left[\sup_{0\le t_k\le T} |e^{-t_k n^2} u_n(0)|^2\right] &= \sum_{N<|n|\le \max N_p, N_w} u_{0,n}^2 \\ &\le KN^{-4} \sum_{N<|n|\le \max N_p, N_w} (1+n^2)^2 u_{0,n}^2.\end{aligned}$$

If $u_0 \in H^2(0, 2\pi)$, this term is bounded by $KN^{2(-2)}$.

Now the nonlinear terms

$$\mathbf{E}\left[\sup_{0\le t_j\le T}\sum_{N<|n|\le N_p}\left|\sum_{k=0}^{j-1}\int_{t_k}^{t_{k+1}} e^{-(t_j-t_k)n^2}F_n(u^N(t_k))\,ds\right|^2\right]$$

$$\le \mathbf{E}\left[\sup_{0\le t_j\le T}\sum_{N<|n|\le N_p}(1+n^2)^r|F_n(u^N(t_k))|^2\left|\sum_{k=0}^{j-1}\int_{t_k}^{t_{k+1}}(1+n^2)^{-r/2}e^{-(t_j-t_k)n^2}\,ds\right|^2\right]$$

$$\le \mathbf{E}\left[\sup_{0\le t\le T}\|u^N(t)\|_r^2\right].\left|\sum_{k=0}^{\lfloor t/\Delta t\rfloor-1}\int_{t_k}^{t_{k+1}}(1+n^2)^{-r/2}e^{-(t_j-t_k)n^2}\,ds\right|^2$$

$$\le \mathbf{E}\left[\sup_{0\le t\le T}\|u^N(t)\|_r^2\right]\frac{(1+N^2)^{-r}}{N^4}.$$

This term is bounded by $K\,N^{2(-r-2)}$ by applying Lemma A.1. The noise term is

$$\mathbf{E}\left[\sup_{0\le t_j\le T}\sum_{N<|n|\le N_w}\left|\sum_{k=0}^{j-1}\left(\int_{t_k}^{t_{k+1}} e^{-(t-t_k)n^2}\lambda_n^{1/2}\Delta B_{k,n}\right)\right|^2\right]$$

$$= 4\sum_{N<|n|\le N_w}(1+n^2)^{\gamma}\lambda_n\int_0^T(1+n^2)^{-\gamma}e^{-2(t_j-t_k)n^2}\,ds$$

$$\le 4\,\mathbf{1}_{N\le N_w}\,N^{2(-1-\gamma)}\sum_{N<|n|\le N_w}(1+n^2)^{\gamma}\lambda_n.$$

This completes the proof. □

Acknowledgements. We are grateful to Yubin Yan for helpful discussions on this work. We would also like to thank Edriss Titi and Jacques Rougemont for their comments early on in this work.

REFERENCES

[1] C. Canuto, M. Y. Hussaini, Q. Quateroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Computational Physics, Springer-Verlag, Berlin, 1988.

[2] G. Da Prato and A. Debussche, *Construction of stochastic inertial manifolds using backward integration*, Stochastics Stochastics Rep., 59 (1996), pp. 305–324.

[3] G. Da Prato and J. Zabczyk, *Stochastic Equations in Infinite Dimensions*, Encyclopedia of Mathematics and Its Applications 44, Cambridge University Press, Cambridge, UK, 1992.

[4] A. M. Davie and J. G. Gaines, *Convergence of numerical schemes for the solution of parabolic stochastic partial differential equations*, Math. Comp., 70 (2001), pp. 121–134.

[5] A. Doelman and E. Titi, *Regularity of solutions and the convergence of the Galerkin method in the complex Ginzburg–Landau equation*, Numer. Funct. Anal. Optim., 14 (1993), pp. 299–321.

[6] C. Foias, O. Manley, and R. Temam, *Modelling of the interaction of small and large eddies in two-dimensional turbulent flows*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 93–114.

[7] B. García-Archilla, J. Novo, and E. S. Titi, *Postprocessing the Galerkin method: A novel approach to approximate inertial manifolds*, SIAM J. Numer. Anal., 35 (1998), pp. 941–972.

[8] B. García-Archilla, J. Novo, and E. S. Titi, *An approximate inertial manifolds approach to postprocessing the Galerkin method for the Navier–Stokes equations*, Math. Comp., 68 (1999), pp. 893–911.

[9] B. García-Archilla, J. Novo, and E. S. Titi, *Postprocessing Fourier spectral methods: The case of smooth solutions*, Appl. Numer. Math., 43 (2002), pp. 191–209.

[10] W. Grecksch and P. Kloeden, *Time-discretized Galerkin approximations of parabolic stochastic PDEs*, Bull. Austral. Math. Soc., 54 (1996), pp. 79–85.

[11] I. Gyöngy, *Lattice approximations for stochastic quasi-linear parabolic partial differential equations driven by space-time white noise* II, Potential Anal., 11 (1999), pp. 1–37.
[12] I. Gyöngy and D. Nualart, *Implicit scheme for quasi-linear parabolic partial differential equations perturbed by space-time white noise*, Stochastic Processes Appl., 58 (1995), pp. 57–72.
[13] E. Hausenblas, *Approximation for semilinear stochastic evolution equations*, Potential Anal., 18 (2003), pp. 141–186.
[14] E. Hausenblas, *Numerical analysis of semilinear stochastic evolution equations in Banach spaces*, J. Comput. Appl. Math., 147 (2002), pp. 485–516.
[15] Y. He and R. M. M. Mattheij, *Stability and convergence for the reform postprocessing Galerkin method*, Nonlinear Anal. Real World Appl., 1 (2000), pp. 517–533.
[16] D. J. Higham, X. Mao, and A. M. Stuart, *Strong convergence of Euler-type methods for nonlinear stochastic differential equations*, SIAM J. Numer. Anal., 40 (2002), pp. 1041–1063.
[17] M. Hochbruck and A. Ostermann, *Explicit exponential Runge–Kutta methods for semilinear parabolic problems*, SIAM J. Numer. Anal, 43 (2005), pp. 1069–1090.
[18] J. C. Jiminez, I. Shoji, and T. Ozaki, *Simulation of stochastic differential equations through the local linearization method. A comparative study*, J. Statist. Phys., 94 (1999), pp. 587–602.
[19] P. Kloeden and S. Shott, *Linear-implicit strong schemes for Ito–Galerkin approximations of stochastic PDEs*, J. Appl. Math. Stoch. Anal., 14 (2001), pp. 47–53.
[20] J. D. Lawson, *Generalized Runge–Kutta processes for stable systems with large Lipschitz constants*, SIAM J. Numer. Anal., 4 (1967), pp. 372–380.
[21] G. J. Lord and J. Rougemont, *A numerical scheme for stochastic PDEs with Gevrey regularity*, IMA J. Numer. Anal., 24 (2004), pp. 587–604.
[22] L. G. Margolin, E. S. Titi, and S. Wynne, *The postprocessing Galerkin and nonlinear Galerkin methods—A truncation analysis point of view*, SIAM J. Numer. Anal., 41 (2003), pp. 695–714.
[23] B. Minchev and W. Wright, *A Review of Exponential Integrators for First Order Semi-linear Problems*, Technical report, Norwegian University of Science and Technology, Trondheim, Norway, 2005.
[24] C. M. Mora, *Weak exponential schemes for stochastic differential equations with additive noise*, IMA J. Numer. Anal., 25 (2005), pp. 486–506.
[25] T. Mueller-Gronbach and K. Ritter, *Lower bounds and nonuniform time discretization for approximation of stochastic heat equations*, Found. Comput. Math., to appear.
[26] T. Shardlow, *Numerical methods for stochastic parabolic PDEs*, Numer. Funct. Anal. Optim., 20 (1999), pp. 121–145.
[27] T. Shardlow, *Numerical simulation of stochastic PDEs for excitable media*, J. Comput. Appl. Math., 175 (2005), pp. 429–446.
[28] I. Shoji, *Approximation of continuous time stochastic processes by a local linearization method*, Math. Comp., 67 (1998), pp. 287–298.
[29] Y. Yubin, *Semidiscrete Galerkin approximation for a linear stochastic parabolic partial differential equation driven by an additive noise*, BIT, 44 (2004), pp. 829–847.

SIAM J. NUMER. ANAL.
Vol. 45, No. 2, pp. 890–904

MONTE CARLO METHODS IN PAGERANK COMPUTATION: WHEN ONE ITERATION IS SUFFICIENT*

K. AVRACHENKOV†, N. LITVAK‡, D. NEMIROVSKY§, AND N. OSIPOVA†

Abstract. PageRank is one of the principle criteria according to which Google ranks Web pages. PageRank can be interpreted as a frequency of visiting a Web page by a random surfer, and thus it reflects the popularity of a Web page. Google computes the PageRank using the power iteration method, which requires about one week of intensive computations. In the present work we propose and analyze Monte Carlo-type methods for the PageRank computation. There are several advantages of the probabilistic Monte Carlo methods over the deterministic power iteration method: Monte Carlo methods already provide good estimation of the PageRank for relatively important pages after one iteration; Monte Carlo methods have natural parallel implementation; and finally, Monte Carlo methods allow one to perform continuous update of the PageRank as the structure of the Web changes.

Key words. Google, PageRank, Monte Carlo methods, absorbing Markov chains

AMS subject classifications. 60J10, 60J20, 65C05

DOI. 10.1137/050643799

1. Introduction.

Surfers on the Internet frequently use search engines to find pages satisfying their query. However, there are typically hundreds or thousands of relevant pages available on the Web. Thus, listing them in a proper order is a crucial and nontrivial task. The original idea of Google presented in [5] is to list pages according to their PageRank, which reflects the popularity of a page. The PageRank is defined in the following way. Denote by n the total number of pages on the Web, and define the $n \times n$ hyperlink matrix P as follows. Suppose that page i has $k > 0$ outgoing links. Then $p_{ij} = 1/k$ if j is one of the outgoing links, and $p_{ij} = 0$ otherwise. If a page does not have outgoing links, the probability is spread among all pages of the Web, namely, $p_{ij} = 1/n$. In order to make the hyperlink graph connected, it is assumed that a random surfer goes with some probability to an arbitrary Web page with uniform distribution. Thus, the PageRank is defined as a stationary distribution of a Markov chain whose state space is the set of all Web pages, and the transition matrix is

$$\tilde{P} = cP + (1-c)(1/n)E, \tag{1.1}$$

where E is a matrix whose all entries are equal to one and $c \in (0,1)$ is the probability of not jumping to a random page (it is chosen by Google to be 0.85). The Google matrix $\tilde{P}$ is stochastic, aperiodic, and irreducible, so there exists a unique row vector

*Received by the editors October 30, 2005; accepted for publication (in revised form) October 30, 2006; published electronically May 1, 2007. This work was supported by the Meervoud grant 632.002.401 from the Dutch National Research Council (NWO) and by the French-Russian ECO-NET grant from EGIDE.

http://www.siam.org/journals/sinum/45-2/64379.html

†INRIA Sophia Antipolis, MAESTRO team, 2004 Route des Lucioles, B. P. 93, 06902 Sophia Antipolis, France (k.avrachenkov@sophia.inria.fr, natalia.osipova@sophia.inria.fr).

‡Department of Applied Mathematics, Faculty EEMCS, University of Twente, P. O. Box 217, 7500 AE Enschede, The Netherlands (n.litvak@ewi.utwente.nl).

§Department of Information Technology, St. Petersburg State University, 35 Universitetskii Prospekt, Peterhof, 198504 St. Petersburg, Russia (danil.nemirovsky@gmail.com).

π such that

$$\pi \tilde{P} = \pi, \quad \pi \underline{1} = 1, \tag{1.2}$$

where $\underline{1}$ is a column vector of ones. The row vector π satisfying (1.2) is called a PageRank vector, or simply PageRank. If a surfer follows a hyperlink with probability c and jumps to a random page with probability $1-c$, then π_i can be interpreted as a stationary probability that the surfer is at page i. The PageRank also allows several different interpretations through expectations. For instance, in [2], the PageRank is seen as the average number of surfers navigating a given page at a given time instant, provided that at each time instant $t \geq 0$ a surfer can cease from navigating with probability $(1-c)$ and on average $(1-c)$ surfers start navigating from each page. This interpretation is helpful for deeper understanding of the PageRank, but it is hard to use in practice because it involves the time component. The interpretation via absorbing Markov chains that we explore in the present paper is easier and leads naturally to simple simulation algorithms for the computation of PageRank. The endpoint of a random walk that starts from a random page and can be terminated at each step with probability $1-c$ appears to be a sample from the distribution π [4, 8, 10]. Thus, after repeating the process many times, the estimate of π_j for $j = 1, \ldots, n$ is determined as the number of times when a run terminated at j, divided by the total number of runs.

In order to keep up with constant modifications of the Web structure, Google updates its PageRank at least once per month. According to publicly available information Google still uses a simple power iteration (PI) method to compute the PageRank. Starting from the initial approximation as the uniform distribution vector $\pi^{(0)} = (1/n)\underline{1}^T$, the kth approximation vector is calculated by

$$\pi^{(k)} = \pi^{(k-1)} \tilde{P}, \quad k \geq 1. \tag{1.3}$$

The method stops when the required precision ε is achieved. The number of flops needed for the method to converge is of the order $\frac{\log \varepsilon}{\log c}\, nnz(P)$, where $nnz(P)$ is the number of nonzero elements of the matrix P [15]. We note that the relative error decreases uniformly for all pages. Several proposals [9, 12, 13, 16] (see also an extensive survey paper [15]) have recently been put forward to accelerate the PI algorithm.

In contrast, here we study Monte Carlo (MC)-type methods for the PageRank computation. To the best far knowledge, in only two works [3, 8] are the MC methods applied to the PageRank computation. The principle advantages of the probabilistic MC-type methods over the deterministic methods are that the PageRank of important pages is determined with high accuracy already after the first iteration; MC methods have natural parallel implementation; and MC methods allow continuous update of the PageRank as the structure of the Web changes.

The structure and the contributions of the paper are as follows. In section 2, we describe different MC algorithms. In particular, we propose an algorithm that takes into account the information not only about the last visited page (as in [3, 8]), but also about all pages visited during the simulation run. In section 3, we analyze and compare the convergence of MC algorithms in terms of confidence intervals. We show that the PageRank of relatively important pages can be determined with high accuracy even after the first iteration. In section 4, we show that experiments with real data from the Web confirm our theoretical analysis. Finally, we summarize the results of the present work in section 5.

2. Monte Carlo algorithms. MC algorithms are motivated by the following convenient formula, which follows directly from the definition of the PageRank:

$$\pi = \frac{1-c}{n}\underline{1}^T[I-cP]^{-1} = \frac{1-c}{n}\underline{1}^T\sum_{k=0}^{\infty}c^kP^k. \tag{2.1}$$

This formula suggests a simple way of sampling from the PageRank distribution [4, 8, 10]. Consider a random walk $\{X_t\}_{t\geq 0}$ that starts from a randomly chosen page. Assume that at each step the random walk terminates with probability $(1-c)$ and makes a transition according to the matrix P with probability c. It follows from (2.1) that the end-point of such a random walk has a distribution π. Hence, one can suggest the following algorithm employed in [3].

ALGORITHM 1. MC END-POINT WITH RANDOM START. *Simulate N runs of the random walk $\{X_t\}_{t\geq 0}$ initiated at a randomly chosen page. Evaluate π_j as a fraction of N random walks which end at page $j=1,\dots,n$.*

Let $\hat{\pi}_{j,N}$ be the estimator of π_j obtained by Algorithm 1. It is straightforward that $\mathbb{E}(\hat{\pi}_{j,N})=\pi_j$ and $Var(\hat{\pi}_{j,N})=N^{-1}\pi_j(1-\pi_j)$. A rough estimate $Var(\hat{\pi}_{j,N})<1/(4N)$ given in [3] results in a conclusion that the number of samples (random walks) needed to achieve a good relative accuracy with high probability, is of the order $O(n^2)$. In the ensuing sections 3 and 4 we will show that this complexity evaluation is quite pessimistic. The number of required samples turns out to be linear in n. Moreover, a reasonable evaluation of the PageRank for popular pages can be obtained even with $N=n$; that is, one needs only as little as one run per page!

In order to improve the estimator $\hat{\pi}$, one can think of various methods of variance reduction. For instance, denoting $Z=[I-cP]^{-1}$ and writing π_j in (2.1) as $\pi_j=\frac{1-c}{n}\sum_{i=1}^n z_{ij}$ for $j=1,\dots,n$, we can view π_j as a given number $(1/n)$ multiplied by a sum of conditional probabilities $p_{ij}=(1-c)z_{ij}$ that the random walk ends at j given that it started at i. Since n is known, an unnecessary randomness in experiments can be avoided by taking $N=mn$ and initiating the random walk exactly m times from each page in a cyclic fashion, rather than jumping N times to a random page. This results in the following algorithm, whose version was used in [8] for computing personalized PageRank.

ALGORITHM 2. MC END-POINT WITH CYCLIC START. *Simulate $N=mn$ runs of the random walk $\{X_t\}_{t\geq 0}$ initiated at each page exactly m times. Evaluate π_j as a fraction of N random walks which end at page $j=1,\dots,n$.*

Let $\hat{p}_{ij}$ be a fraction of m random walks initiated at i that ended at j. Then the estimator for π_j suggested by Algorithm 2 can be expressed as $\hat{\hat{\pi}}_j=\frac{1}{n}\sum_{i=1}\hat{p}_{ij}$. For this estimator, we have $\mathbb{E}(\hat{\hat{\pi}}_j)=\pi_j$ and $Var(\hat{\hat{\pi}}_j)=(N)^{-1}[\pi_j-n^{-1}\sum_{i=1}^n p_{ij}^2]<Var(\hat{\pi}_j)$. Besides the variance reduction, the estimator $\hat{\hat{\pi}}_i$ has important advantages in implementation, because picking a page at random from a huge database is not a trivial problem [11]. This difficulty is completely avoided if the pages are visited in a cyclic fashion.[1] As the only advantage of the MC with random start, note that it does not require the number of samples N to be a multiple of n.

Another and probably more promising way of reducing the variance is to look at formula (2.1) from yet another angle. Note that for all $i,j=1,\dots,n$, the element z_{ij}

[1] When referring to MC algorithms with cyclic start, we shall use the words "cycle" and "iteration" interchangeably.

of the matrix

$$Z = [I - cP]^{-1} = \sum_{k=0}^{\infty} c^k P^k \tag{2.2}$$

can be regarded as the average number of times that the random walk $\{X_t\}_{t\geq 0}$ visits a page j, given that this random walk started at page i. Thus, we can propose an estimator based on a complete path of the random walk $\{X_t\}_{t\geq 0}$ instead of taking into account only its end-point. The complete path version of the MC method can be described as follows.

ALGORITHM 3. MC COMPLETE PATH. *Simulate the random walk $\{X_t\}_{t\geq 0}$ exactly m times from each page. For any page i, evaluate π_j as the total number of visits to page j multiplied by $(1-c)/(n\cdot m)$.*

Algorithm 3 can be further improved by getting rid of artifacts in the matrix P related to pages without outgoing links (so-called dangling nodes). When a random walk reaches a dangling node, it jumps with uniform probability to an arbitrary page. Clearly, it is more efficient just to terminate the random walk once it reaches a dangling node. Thus, we aim to rewrite (2.1) in terms of the original hyperlink matrix Q with elements defined as $Q_{ij} = 1/k$ if page i has $k > 0$ outgoing links and a link points to page j, and 0 otherwise. The artificial links from dangling pages are not taken into account in hyperlink matrix Q. Denote by $\mathcal{I}_0$ a set of dangling pages and by $\mathcal{I}_1 = \{1,\dots,n\}\backslash\mathcal{I}_0$ a set of pages which have at least one outgoing link. For all $j = 1,\dots,n$, it follows from (1.1) and (1.2) that

$$\pi_j = c\sum_{i=1}^{n} P_{ij}\pi_i + \frac{(1-c)}{n}\sum_{i=1}^{n}\pi_i = c\sum_{i=1}^{n} Q_{ij}\pi_i + \gamma, \tag{2.3}$$

where γ is the same for each j:

$$\gamma = \frac{c}{n}\sum_{i\in\mathcal{I}_0}\pi_i + \frac{(1-c)}{n} < \frac{1}{n}. \tag{2.4}$$

Now, we rewrite (2.3) in the matrix form $\pi = \pi cQ + \gamma \underline{1}^T$, which leads to the new expression for π:

$$\pi = \gamma \underline{1}^T [I - cQ]^{-1}. \tag{2.5}$$

Note that the above equation is in accordance with the original definition of PageRank presented by Brin et al. [5]. The definition via the matrix P appeared later in order to develop the Markov chain formulation of the PageRank problem. The one-to-one correspondence between (2.1) and (2.5) was noticed and proved in [2], but we find the proof presented above more insightful in our context.

Consider now a random walk $\{Y_t\}_{t\geq 0}$ which follows hyperlinks exactly as $\{X_t\}_{t\geq 0}$ except that the transitions are governed by the matrix Q instead of the matrix P. Thus, the random walk $\{Y_t\}_{t\geq 0}$ can be terminated at each step either with probability $(1-c)$ or when it reaches a dangling node. For all $i,j = 1,\dots,n$, the element w_{ij} of the matrix $W = [I - cQ]^{-1}$ is the average number of visits of $\{Y_t\}_{t\geq 0}$ to page j, given that the random walk started at page i. Denote $w_{\cdot j} = \sum_{i=1}^{n} w_{ij}$. Since the coordinates of π in (2.5) sum up to one, we have

$$\gamma = \left[\sum_{i,j=1}^{n} w_{ij}\right]^{-1} = \left[\sum_{j=1}^{n} w_{\cdot j}\right]^{-1} \tag{2.6}$$

and

$$\pi_j = w_{\cdot j} \left[\sum_{j=1}^{n} w_{\cdot j} \right]^{-1}. \tag{2.7}$$

This calls for another version of the complete path method.

ALGORITHM 4. MC COMPLETE PATH STOPPING AT DANGLING NODES. *Simulate the random walk $\{Y_t\}_{t\geq 0}$ starting exactly m times from each page. For any page j, evaluate π_j as the total number of visits to page j divided by the total number of visited pages.*

Let W_{ij} be a random variable distributed as a number of visits to page $j = 1, \ldots, n$ by the random walk $\{Y_t\}_{t\geq 0}$ given that the random walk initiated at state $i = 1, \ldots, n$. Formally, $\mathbb{P}(W_{ij} = x) = \mathbb{P}\left(\left[\sum_{t=0}^{\infty} \mathbf{1}_{\{Y_t = j\}}\right] = x | Y_0 = i\right)$ for $x = 0, 1, \ldots$, where $\mathbf{1}_{\{\cdot\}}$ is the indicator function. Let $W_{ij}^{(l)}$, $l \geq 1$, be independent random variables distributed as W_{ij}. Then the estimator produced by Algorithm 4 can be written as

$$\bar{\pi}_j = \left[\sum_{l=1}^{m} \sum_{i=1}^{n} W_{ij}^{(l)} \right] \left[\sum_{l=1}^{m} \sum_{i,j=1}^{n} W_{ij}^{(l)} \right]^{-1}. \tag{2.8}$$

In the next section we present the analysis of this estimator.

We note that the complete path versions of the MC methods also admit a random start. The corresponding algorithm is as follows.

ALGORITHM 5. MC COMPLETE PATH WITH RANDOM START. *Simulate N samples of the random walk $\{Y_t\}_{t\geq 0}$ started at a random page. For any page j, evaluate π_j as the total number of visits to page i divided by the total number of visited pages.*

One can show, however, that Algorithm 4 provides an estimator with a smaller variance than Algorithm 5. Indeed, let W_{Uj} be the number of visits to page j from a randomly chosen page $U \in \{1, \ldots, n\}$. Then, we have

$$Var(W_{Uj}) = \frac{1}{n} \sum_{i=1}^{n} Var(W_{ij}) + \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}^2(W_{ij}) - \left[\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(W_{ij}) \right]^2 > \frac{1}{n} \sum_{i=1}^{n} Var(W_{ij}).$$

Now note that in n simulation runs, Algorithm 4 generates one sample of the sum $\sum_{i=1}^{n} W_{ij}$, whereas Algorithm 5 generates n samples of W_{Uj}. Hence, Algorithm 4 provides random variables with smaller variance in both numerator and denominator of (2.8).

3. Convergence analysis. From the preliminary analysis of the previous section, we can already conclude that MC algorithms with cyclic start are preferable to the analogous MC algorithms with random start. In the present section we thoroughly analyze and compare MC complete path stopping at dangling nodes with MC end-point. We show that under natural conditions MC complete path stopping at dangling nodes outperforms MC end-point.

We start by studying the properties of W_{ij}'s. Denote by q_{ij} the probability that starting from page i, the random walk $\{Y_t\}_{t\geq 0}$ reaches page j:

$$q_{ij} = \mathbb{P}\left(\bigcup_{t\geq 1} \{Y_t = j\} | Y_0 = i \right), \quad i, j = 1, \ldots, n.$$

Note that in this definition, $q_{jj} < 1$ is the probability of returning to state j if the process started at j. It follows from the strong Markov property that W_{jj} has a geometric distribution with parameter $1-q_{jj} \geq 1-c$: $\mathbb{P}(W_{jj} = x) = q_{jj}^{x-1}(1-q_{jj})$, $x = 1, 2, \ldots$, which implies $\mathbb{E}(W_{jj}) = 1/(1-q_{jj}), Var(W_{jj}) = q_{jj}/(1-q_{jj})^2$. Further, applying again the strong Markov property, one can show that for all $i, j = 1, \ldots, n$, W_{ij} has a shifted geometric distribution:

$$\mathbb{P}(W_{ij} = x) = \begin{cases} 1 - q_{ij}, & x = 0, \\ q_{ij}\mathbb{P}(W_{jj} = x), & x = 1, 2, \ldots . \end{cases}$$

Consequently,

$$\mathbb{E}(W_{ij}) = w_{ij} = q_{ij}\mathbb{E}(W_{jj}) = \frac{q_{ij}}{1 - q_{jj}} \tag{3.1}$$

and

$$Var(W_{ij}) = \frac{1 + q_{jj}}{1 - q_{jj}}\, w_{ij} - w_{ij}^2. \tag{3.2}$$

Now, define $W_{\cdot j} = \sum_{i=1}^n W_{ij}$ for $j = 1, \ldots, n$, and $W = \sum_{j=1}^n W_{\cdot j}$. Assuming that all W_{ij}'s are independent, we immediately obtain

$$\mathbb{E}(W_{\cdot j}) = \sum_{i=1}^n w_{ij} = w_{\cdot j},$$

$$Var(W_{\cdot j}) = \frac{1 + q_{jj}}{1 - q_{jj}}\, w_{\cdot j} - \sum_{i=1}^n w_{ij}^2 < \frac{1 + q_{jj}}{1 - q_{jj}}\, w_{\cdot j},$$

$$\mathbb{E}(W) = \sum_{j=1}^n w_{\cdot j} = \gamma^{-1}.$$

For $i, j = 1, \ldots, n$, let the empirical mean $\bar{W}_{ij} = \frac{1}{m}\sum_{l=1}^m W_{ij}^{(l)}$ be the estimator of w_{ij}, and view $\bar{W}_{\cdot j} = \sum_{i=1}^n \bar{W}_{ij}$, $j = 1, \ldots, n$, and $\bar{W} = \sum_{j=1}^n \bar{W}_{\cdot j}$ as estimators of $w_{\cdot j}$ and γ^{-1}, respectively. The estimator (2.8) can be then written as

$$\bar{\pi}_j = \bar{W}_{\cdot j}\bar{W}^{-1}. \tag{3.3}$$

Since the second multiplier in (3.3) is the same for all $j = 1, \ldots, n$, the estimator $\bar{\pi}_j$ is completely determined by $\bar{W}_{\cdot j}$. The following theorem states that the relative errors of $\bar{\pi}$ and $\bar{W}_{\cdot j}$ are similar.

THEOREM 3.1. *Given the event that the estimator $\bar{W}_{\cdot j}$ satisfies*

$$|\bar{W}_{\cdot j} - w_{\cdot j}| \leq \varepsilon w_{\cdot j}, \tag{3.4}$$

the event

$$|\bar{\pi}_j - \pi_j| \leq \varepsilon_{n,\beta}\pi_j$$

occurs with probability at least $1-\beta$ *for any* $\beta > 0$ *and* $\varepsilon_{n,\beta}$ *satisfying*

$$|\varepsilon - \varepsilon_{n,\beta}| < \frac{C(\beta)(1+\varepsilon)}{\sqrt{nm}}.$$

The factor $C(\beta)$ *can be approximated as*

$$C(\beta) \approx x_{1-\beta/2}\sqrt{\frac{n-n_0}{n}\,(1+c^3)}\frac{c}{1-c},$$

where $x_{1-\beta/2}$ *is a* $(1-\beta/2)$*-quantile of the standard normal distribution and* n_0 *is the number of dangling nodes.*

Proof. See the appendix.

Theorem 3.1 has two important consequences. First, it states that the estimator $\bar{\pi}_j$ converges to π_j in probability when m goes to infinity. Thus, the estimator $\bar{\pi}_j$ is consistent. Second, Theorem 3.1 states that the error in the estimate of π_j originates mainly from estimating $w_{\cdot j}$. The additional relative error caused by estimating γ as $\left[\sum \bar{W}_{\cdot j}\right]^{-1}$ is of the order $1/\sqrt{mn}$ with arbitrarily high probability, and thus this error can essentially be neglected.

It follows from the above analysis that the quality of the estimator $\bar{\pi}_j$ as well as the complexity of the algorithm can be evaluated by the estimator $\bar{W}_{\cdot j}$. We proceed by analyzing the confidence intervals. Consider the confidence interval for $\bar{W}_{\cdot j}$ defined as

$$\mathbb{P}(|\bar{W}_{\cdot j} - w_{\cdot j}| < \varepsilon w_{\cdot j}) \geq 1-\alpha. \tag{3.5}$$

From (3.1) and (3.2), we have $\mathbb{E}(\bar{W}_{\cdot j}) = w_{\cdot j}$ and $Var(\bar{W}_{\cdot j}) \leq \frac{1}{m}\frac{1+q_{jj}}{1-q_{jj}} w_{\cdot j}$. Since $\bar{W}_{\cdot j}$ is the sum of a large number of terms, the random variable $[\bar{W}_{\cdot j} - w_{\cdot j}]/\sqrt{Var(\bar{W}_{\cdot j})}$ has approximately a standard normal distribution. Thus, from (3.5) we deduce that $\varepsilon w_{\cdot j}/\sqrt{Var(\bar{W}_{\cdot j})} \geq x_{1-\alpha/2}$, which results in

$$m \geq \frac{1+q_{jj}}{1-q_{jj}}\frac{x^2_{1-\alpha/2}}{\varepsilon^2 w_{\cdot j}}.$$

Now applying $w_{\cdot j} = \gamma^{-1}\pi_j$, we get

$$m \approx \frac{1+q_{jj}}{1-q_{jj}}\frac{\gamma x^2_{1-\alpha/2}}{\varepsilon^2 \pi_j}. \tag{3.6}$$

Note that $\pi_j \geq \gamma$ for all $j = 1, \ldots, n$. Thus, with a high probability, a couple of hundred iterations allow us to evaluate the PageRank of all pages with relative error at most 0.1. In practice, however, it is essential to evaluate well the PageRank of important pages in a short time. We argue that a typical user of a search engine does not check more than a dozen of the first answers to his/her query. Therefore, let us evaluate the relative error ε for a given value of π_j. Using (2.4), from (3.6) we derive

$$\varepsilon \approx x_{1-\alpha/2}\sqrt{\frac{1+q_{jj}}{1-q_{jj}}}\frac{\sqrt{1-c+c\sum_{i\in\mathcal{I}_0}\pi_i}}{\sqrt{\pi_j}\sqrt{mn}}. \tag{3.7}$$

Strikingly, it follows from (3.7) that the MC method gives good results for important pages in one iteration only, that is, when $m = 1$. From the examples of PageRank

values presented in [5], it follows that the PageRank of popular pages is at least 10^4 times greater than the PageRank of an average page. Since the PageRank value is bounded from below by $(1-c)/n$, the formula (3.7) implies that if the important pages have PageRank 10^4 times larger than the PageRank of the pages with the minimal PageRank value, the MC method achieves an error of about 1% for the important pages already after the first iteration. In contrast, the power iteration method takes into account only the weighted sum of the number of incoming links after the first iteration.

Let us now compare the precision of the end-point version and the complete path version of the MC method. According to Algorithm 1, the end-point version estimates π_j simply as a fraction of $N = mn$ random walks that end at page j. Using standard techniques for such an estimate, we construct a confidence interval $\mathbb{P}(|\hat{\pi}_{j,N} - \pi_{j,N}| < \varepsilon\pi_{j,N}) = 1 - \alpha$. Using again the standard normal distribution, we get

$$\varepsilon = x_{1-\alpha/2}\frac{\sqrt{1-\pi_j}}{\sqrt{\pi_j}\sqrt{mn}}. \tag{3.8}$$

Forgetting for a moment about slight corrections caused by the trade-off between random and cyclic start, we see that the choice between the end-point version and the complete-path version essentially depends on two factors: the total PageRank of dangling nodes and the probability of a cycle when a random walk started from j returns back to j. If the Web graph has many short cycles, then the extra information from registering visits to every page is obtained at cost of a high extra variability, which leads to a worse precision. If the total rank of dangling nodes is high, the random walk will often reach dangling nodes and stop. This can have a negative impact on the complete path algorithm. The above mentioned two phenomena, if present, can make the difference between the end-point and the complete-path versions negligible. The experiments of the next section on the real data, however, indicate that the real Web structure is such that the complete path version is more efficient than the end-point version.

We remark that if the results of the first iteration are not satisfactory, it is hard to improve them by increasing m. After m iterations, the relative error of the MC method will reduce on average only by the factor $1/\sqrt{m}$, whereas the error of the power iteration method decreases exponentially with m. However, because of simplicity in implementation (in particular, simplicity in parallel implementation), the MC algorithms can still be advantageous even if a high precision is required.

Let us also evaluate a magnitude of π_j's for which a desired relative error ε is achieved. Rewriting (3.7), we get

$$\pi_j \approx x^2_{1-\alpha/2}\frac{1+q_{jj}}{1-q_{jj}}\frac{(1-c+c\sum_{i\in\mathcal{I}_0}\pi_i)}{\varepsilon^2 mn}. \tag{3.9}$$

Finally, we would like to discuss the implementation issues of the MC algorithms and to compare MC and PI methods. Each available processor can run an independent MC simulation. At the end of computations one central normalization procedure is needed. In contrast, any parallel implementation of the PI method requires the exchange of information at every iteration. Thus, if the Web graph does not fit into the main memory, there is a trade-off between the frequent request of MC to the database and the exchange of information in the parallel implementation of the PI

method. The problem of frequent requests of MC to the database can be partially mitigated by using hashing mechanisms. It is hard to say in advance which method (MC or PI) will work faster in the case of parallel implementation. However, it is possible to compare the performance of PI and MC methods in the case when the Web graph fits into the main memory. In such a case, one iteration of MC will be faster than one iteration of PI, if the expected length of the MC run is smaller than the ratio between the number of links and number of pages. If there were no dangling nodes, the expected length of the MC run would be $1/(1-c) \approx 6.7$. When dangling nodes are present, this value is smaller. According to [6, 14], the ratio between the number of links and number of pages of the Web is between 7.2 and 7.5. If one computes the PageRank for a Web graph with the latter property and the Web graph fits into the main memory, the MC method should outperform the PI method.

Moreover, the MC algorithms allow one to perform a continuous update of the PageRank vector. Since the PageRank vector changes significantly during one month, Google prefers to recompute the PageRank vector starting from the uniform distribution rather than to use the PageRank vector of the previous month as the initial approximation [15]. Then, it takes about a week to compute a new PageRank vector. It is possible to update the PageRank vector using linear algebra methods [16]. However, one needs first to separate new nodes and links from the old ones. This is not necessary if one uses MC algorithms. Specifically, we suggest running MC algorithms continuously while the database is updated with new data, and hence having an up-to-date estimation of the PageRank for relatively important pages with high accuracy. Then, once in a while one can run the power iteration method to have a good PageRank estimation for all pages. In particular, the continuous update should eliminate the negative reaction of users to the so-called "Google dance" [17].

4. Experiments. For our numerical experiments we have used the Web site of INRIA Sophia Antipolis (http://www-sop.inria.fr). It is a typical Web site with about 50000 Web pages and 200000 hyperlinks. Since the Web has a fractal structure [7], we expect that our dataset is sufficiently representative. Accordingly, datasets of similar sizes have been extensively used in experimental studies of novel algorithms for PageRank computation [1, 15, 16]. To collect the Web graph data, we construct our own Web crawler which works with the Oracle database. The crawler consists of two parts: the first part is realized based on Java and is responsible for downloading pages from the Internet, parsing the pages, and inserting their hyperlinks into the database; the second part is realized with the help of stored procedures written in PL/SQL language and is responsible for the data management. The program allows one to run several crawlers in parallel to efficiently use the network and computer resources. Since multi user access is already realized in the Oracle database management system, it is relatively easy to organize the information collection by several crawlers and parallel implementation of MC algorithms. We have also implemented the power iteration method and the following three MC algorithms in PL/SQL language:

- MC complete path stopping in dangling nodes,
  MC comp path dangl nodes, for short;
- MC end-point with cyclic start,
  MC end-point cycl start, for short;
- MC complete path with random start,
  MC comp path rand start, for short.

First, we performed a sufficient number of power iterations to obtain the value of PageRank with 20 digits accuracy. We sorted the PageRank vector in the decreasing order and plotted it in the log-log scale (see Figure 4.1). It is interesting to observe

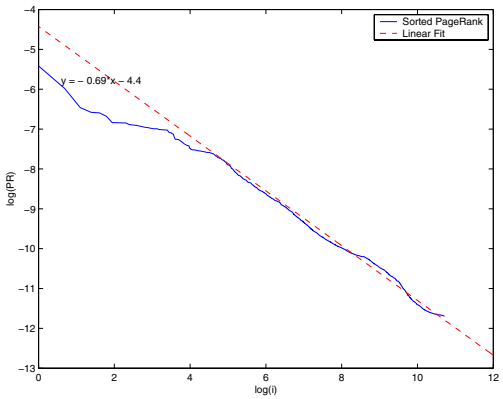


FIG. 4.1. *Sorted PageRank in log-log scale.*

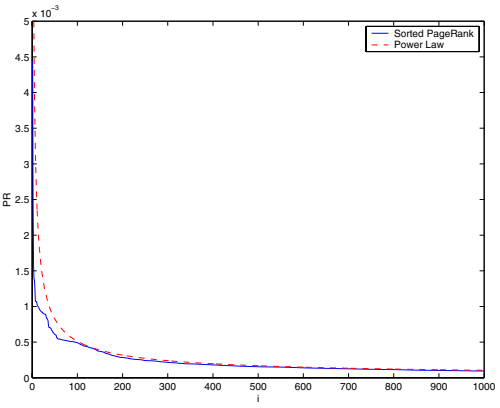


FIG. 4.2. *Sorted PageRank in linear scale.*

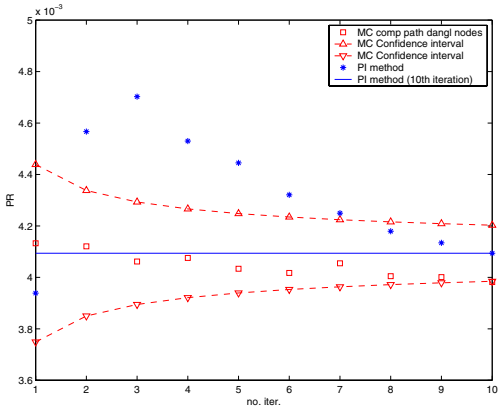


FIG. 4.3. *PI versus MC:* π_1.

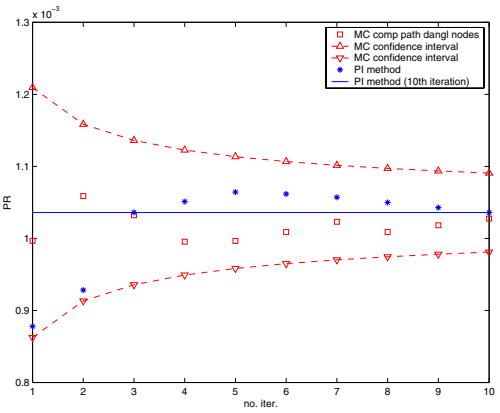


FIG. 4.4. *PI versus MC:* π_{10}.

that the PageRank vector very closely follows a power law. One can also see in Figure 4.2 how well the power law approximates the PageRank vector in linear scale starting from approximately the 100th largest element. Then, we have chosen four elements from the sorted PageRank vector:

$$\begin{aligned} \pi_1 &= 0.004093834, \\ \pi_{10} &= 0.001035867, \\ \pi_{100} &= 0.000546446, \\ \pi_{1000} &= 0.000097785. \end{aligned} \tag{4.1}$$

We have performed ten iterations of the PI method and ten iterations of the three implemented MC algorithms. In Figures 4.3–4.6, we compare the results of ten iterations of PI method and MC complete path stopping in dangling nodes method for the four chosen pages (4.1). Indeed, as predicted by formula (3.7), already the first iteration of the MC complete path stopping in dangling nodes algorithm gives a small error for important Web pages. In fact, from Figures 4.3–4.6 one can see that MC complete path stopping in dangling nodes algorithm outperforms the PI method

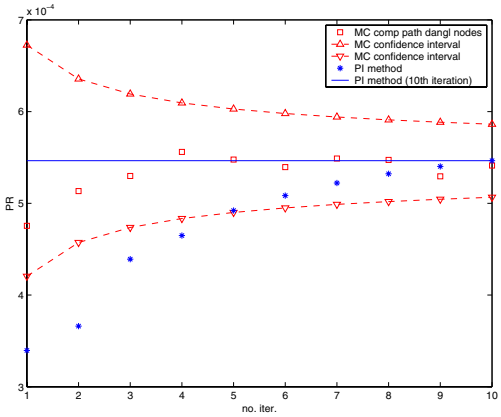


FIG. 4.5. *PI versus MC:* π_{100}.

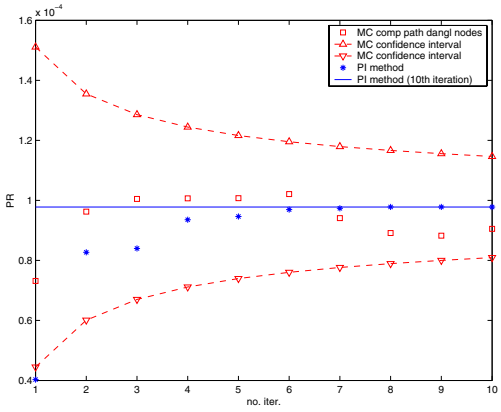


FIG. 4.6. *PI versus MC:* π_{1000}.

even for the first 1000 most important pages. In Figures 4.3–4.6, we also plotted 95% confidence intervals for the MC method. As expected, there is some randomness in the convergence pattern of the MC method, and some points might fall outside of confidence intervals. However, as one can see from Figures 4.3–4.4, the PI method does not converge in a monotone fashion for the first few iterations either.

At first sight, it looks surprising that one iteration gives a relative error of only 7% with 95% confidence for pages with high PageRank. On the other hand, such a result is to be expected. Roughly speaking, we use $5 \cdot 10^4$ independent samples in order to estimate the probability $\pi = 0.004$. A binomial random variable B with parameters $n = 5 \cdot 10^4$, $p = 0.004$ has mean 200 and standard deviation 14.1, and thus, with a high probability, a relative error of a standard estimator $\tilde{\pi} = B/n$ will be less than 11%. The additional gain that we get in (3.7) is due to regular visits to every page and the usage of the complete path information.

Next, in Figures 4.7–4.10 we compare three versions of the MC method: MC complete path stopping in dangling nodes, MC end-point with cyclic start, and MC complete path with random start. We plotted actual relative error and the estimated 95% confidence intervals. It turns out that on our dataset MC complete path stopping in dangling nodes performs the best, followed by MC complete path with random start. MC end-point with cyclic start has the worst performance. The better performance of MC with cyclic start in respect to MC with random start was expected from the preliminary analysis of section 2. MC is not trapped in cycles in our instance of the Web graph and the total PageRank of dangling nodes is relatively small

$$\sum_{i \in \mathcal{I}_0} \pi_i = 0.23;$$

hence, we have

$$\varepsilon_{comp.path} \approx \sqrt{1 - c + c \sum_{i \in \mathcal{I}_0} \pi_i}\, \varepsilon_{end-point} \approx 0.59 \varepsilon_{end-point}.$$

To check whether the presence of cycles hinders the convergence of the MC methods, we took into account the intra-page hyperlinks. On the modified graph the MC

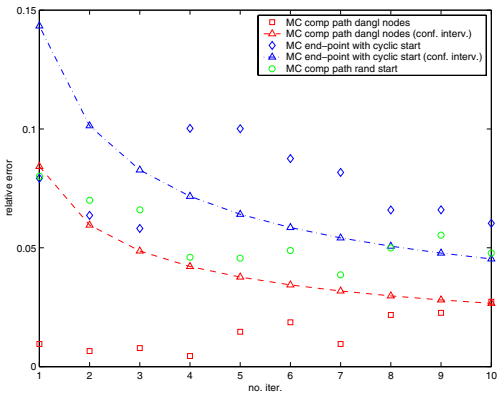


FIG. 4.7. *Comparison of MC algorithms:* π_1.

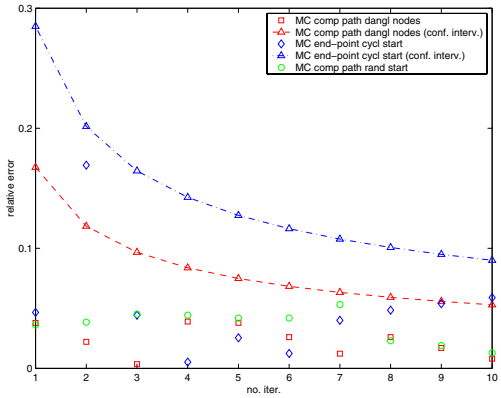


FIG. 4.8. *Comparison of MC algorithms:* π_{10}.

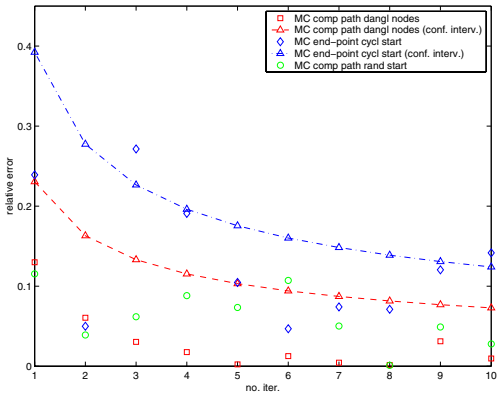


FIG. 4.9. *Comparison of MC algorithms:* π_{100}.

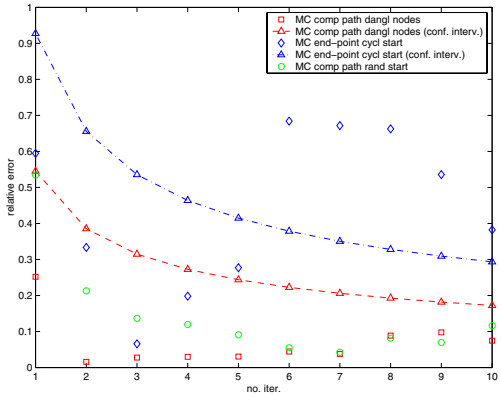


FIG. 4.10. *Comparison of MC algorithms:* π_{1000}.

methods have shown a very slow convergence. It is thus fortunate for MC methods that the original definition of the PageRank excludes the intra-page hyperlinks.

We may conclude that on the web data, MC algorithms with cyclic start outperform MC algorithms with random start. Besides, our theoretical and experimental results have demonstrated that the MC algorithms determine the PageRank of relatively important pages already after the first iteration. Here is a sharp contrast with the PI method, which approximates the PageRank vector with the uniform relative error and takes into account only the weighted sum of the number of incoming links after the first iteration. The other advantages of MC algorithms are natural parallel implementation and the possibility of continuous PageRank update while the crawler brings new data from the Web.

5. Conclusions. We have considered several MC algorithms for PageRank computation. In particular, we have proposed a new MC algorithm that takes into account not only the information about the last visited page, but about all visited pages during the simulation run. We have shown that MC algorithms with cyclic start outperform MC algorithms with random start. Our theoretical and experimental results have demonstrated that the MC algorithms determine the PageRank of relatively impor-

tant pages already after the first iteration. Here is a sharp contrast with the PI method that approximates the PageRank vector with the uniform relative error and takes into account only the weighted sum of the number of incoming links after the first iteration. The other advantages of MC algorithms are natural parallel implementation and the possibility of continuous PageRank update while the crawler brings new data from the Web. As a future research direction, it is necessary to address the question of how fast the continuous update version of the MC method can adapt to the perpetual changes in Web structure. One more promising future research direction is the development of MC methods for the other link-based ranking criteria such as HITS and SALSA.

Appendix: The proof of Theorem 3.1. To prove Theorem 3.1 we need the following lemma.

LEMMA A.1. *Let* $W_{i\cdot} = \sum_{j=1}^{n} W_{ij}$ *be the length of the random walk* $\{Y_t\}_{t\geq 0}$ *initiated at page* $i = 1, \ldots, n$. *Then for all dangling nodes* $i \in \mathcal{I}_0$, $W_{i\cdot} \equiv 1$ *holds, and for nondangling nodes* $i \in \mathcal{I}_1$,

$$\mathbb{E}(W_{i\cdot}) \leq \frac{1}{1-c}, \qquad Var(W_{i\cdot}) \leq \frac{c(1+c^3)}{(1-c)^2}. \tag{A.1}$$

Proof. The statement for dangling nodes is obvious. For nondangling nodes, (A.1) essentially follows from the distributional identity

$$W_{i\cdot} \stackrel{d}{=} \min\{X, N_i\}, \quad i = 1, \ldots, n, \tag{A.2}$$

where N_i is a number of transitions needed to reach a dangling node from page i, and X has a geometric distribution with parameter $1-c$. The mean and variance of X are given by

$$\mathbb{E}(X) = \frac{1}{1-c}, \quad Var(X) = \frac{c}{(1-c)^2}.$$

The upper bound for the expectation of $W_{i\cdot}$ now follows directly from (A.2). For the variance, we write

$$Var(W_{i\cdot}) = \mathbb{E}[Var(W_{i\cdot}|N_i)] + Var[\mathbb{E}(W_{i\cdot}|N_i)].$$

Conditioning on events $[N_i = k]$ and computing $Var(W_i|k)$ for $k = 1, 2, \ldots$, one can show that

$$\mathbb{E}[Var(W_{i\cdot}|N_i)] < Var(X).$$

Furthermore, we derive

$$\mathbb{E}(W_{i\cdot}|N_i) = \sum_{k=1}^{N_i} \mathbb{P}(X \geq k) = \sum_{k=1}^{N_i} c^k = \frac{c(1-c^{N_i})}{1-c},$$

and thus the variance of $\mathbb{E}(W_{i\cdot}|N_i)$ satisfies

$$Var(\mathbb{E}(W_{i\cdot}|N_i)) = \frac{c^2 Var(c^{N_i})}{(1-c)^2} \leq \frac{c^4}{(1-c)^2},$$

because for nondangling nodes, the random variable c^{N_i} takes values only in the interval $[0, c]$. This completes the proof of the lemma. ∎

We are now ready to prove Theorem 3.1.

Proof of Theorem 3.1. Using (2.6) and (2.7), we derive

$$\begin{aligned}\bar{\pi}_j - \pi_j &= \bar{W}_{\cdot j}\bar{W}^{-1} - \pi_j \\ &= \gamma(\bar{W}_{\cdot j} - w_{\cdot j})(\gamma\bar{W})^{-1} + \left((\gamma\bar{W})^{-1} - 1\right)\pi_j.\end{aligned}$$

Given the event (3.4), the last equation together with (2.6) and (2.7) yields

$$|\bar{\pi}_j - \pi_j| \le \varepsilon\pi_j + \left|(\gamma\bar{W})^{-1} - 1\right|(1+\varepsilon)\pi_j. \tag{A.3}$$

Let us now investigate the magnitude of the term $(\gamma\bar{W})^{-1}$. First, note that the random variables

$$\bar{W}_{i\cdot} = \sum_{j=1}^{n} \bar{W}_{ij}, \quad i \in \mathcal{I}_1,$$

are independent because they are determined by simulation runs initiated at different pages. Further, for a nondangling node i, using Lemma A.1, we find

$$\begin{aligned}\mathbb{E}(\bar{W}_{i\cdot}) &= \sum_{j=1}^{n} w_{ij}, \\ Var(\bar{W}_{i\cdot}) &= \frac{1}{m} Var(W_{i\cdot}) \le \frac{1}{m}\,\frac{c(1+c^3)}{(1-c)^2}.\end{aligned}$$

Thus, $\bar{W}$ equals the number of dangling nodes n_0 plus the sum of $n-n_0$ independent random variables $\bar{W}_{i\cdot}$, $i \in \mathcal{I}_1$. Since the number $n-n_0$ is obviously very large, $\bar{W}$ is approximately normally distributed with mean γ^{-1} and variance

$$Var(\bar{W}) = \sum_{i\in\mathcal{I}_1} Var(\bar{W}_{i\cdot}) \le (n-n_0)\frac{c(1+c^3)}{m(1-c)^2}.$$

Hence, $\gamma\bar{W}$ is approximately normally distributed with mean 1 and variance

$$Var(\gamma\bar{W}) \le \gamma^2(n-n_0)\frac{c(1+c^3)}{m(1-c)^2} < \frac{n-n_0}{n^2}\,\frac{c(1+c^3)}{m(1-c)^2}, \tag{A.4}$$

which is a value of the order $(nm)^{-1}$. Now, let us consider a $(1-\beta)$-confidence interval defined as

$$\mathbb{P}\left(\left|(\gamma\bar{W})^{-1} - 1\right| < \epsilon\right) > 1-\beta \tag{A.5}$$

for some small positive β and ϵ. If ϵ is small enough so that $1/(1-\epsilon) \approx 1+\epsilon$ and $1/(1+\epsilon) \approx 1-\epsilon$, then the above probability approximately equals $\mathbb{P}\left(\left|\gamma\bar{W} - 1\right| < \epsilon\right)$, and because of (A.4), the inequality (A.5) holds for all ϵ satisfying

$$\epsilon \ge x_{1-\beta/2}\frac{c}{1-c}\sqrt{\frac{n-n_0}{n}\,(1+c^3)}\,\frac{1}{\sqrt{nm}}. \tag{A.6}$$

The right-hand side of (A.6) constitutes the additional relative error in estimating π_j. For any $\beta > 0$, this additional error can be exceeded with probability at most β. This completes the proof of the theorem. □

REFERENCES

[1] S. Abiteboul, M. Preda, and G. Cobena, *Adaptive on-line page importance computation*, in Proceedings of the 12th International World Wide Web Conference, Budapest, 2003.

[2] M. Bianchini, M. Gori, and F. Scarselli, *Inside PageRank*, ACM Trans. Internet Technology, 5 (2005), pp. 92–128.

[3] L. A. Breyer, *Markovian Page Ranking Distributions: Some Theory and Simulations*, Technical report, 2002; available online at http://www.lbreyer.com/preprints.html.

[4] L. A. Breyer and G. O. Roberts, *Catalytic perfect simulation*, Methodol. Comput. Appl. Probab., 3 (2001), pp. 161–177.

[5] S. Brin, L. Page, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*, Stanford University Technical Report, Stanford, CA, 1998; available online at http://dbpubs.stanford.edu:8090/pub/1999-66.

[6] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, *Graph structure in the Web*, Computer Networks, 33 (2000), pp. 309–320.

[7] S. Dill, R. Kumar, K. McCurley, S. Rajagopalan, D. Sivakumar, and A. Tomkins, *Self-similarity in the Web*, ACM Trans. Internet Technol., 2 (2002), pp. 205–223.

[8] D. Fogaras and B. Racz, *Towards scaling fully personalized PageRank*, in Proceedings of the 3rd International Workshop on Algorithms and Models for the Web-Graph, New York, 2004.

[9] I. C. F. Ipsen and S. Kirkland, *Convergence Analysis of an Improved PageRank Algorithm*, North Carolina State University Technical Report CRSC-TR04-02, 2004; available online at http://www.ncsu.edu/crsc/reports/reports04.html.

[10] G. Jeh and J. Widom, *Scaling personalized web search*, in Proceedings of the 12th World Wide Web Conference, Budapest, 2003.

[11] M. R. Henzinger, *Algorithmic challenges in web search engines*, Internet Mathematics, 1 (2003), pp. 115–126.

[12] S. D. Kamvar, T. H. Haveliwala, and G. H. Golub, *Adaptive methods for the computation of PageRank*, Linear Algebra Appl., 386 (2004), pp. 51–65.

[13] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, *Extrapolation methods for accelerating PageRank computations*, in Proceedings of the 12th International World Wide Web Conference, Budapest, 2003.

[14] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins, *The Web as a Graph: Measurements, Models, and Methods*, Invited Survey at the International Conference on Combinatorics and Computing, Tokyo, 1999; available online at http://www.cs.cornell.edu/home/kleinber/web-graph.ps.

[15] A. N. Langville and C. D. Meyer, *Deeper Inside PageRank*, Internet Math., 1 (2004), pp. 335–400; also available online at http://www4.ncsu.edu/~anlangvi/.

[16] A. N. Langville and C. D. Meyer, *Updating PageRank with iterative aggregation*, in Proceedings of the 13th World Wide Web Conference, New York, 2004.

[17] *Google Dance—The Index Update of the Google Search Engine*, http://dance.efactory.de.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 905–927

A CONVERGENT AND CONSTRAINT-PRESERVING FINITE ELEMENT METHOD FOR THE P-HARMONIC FLOW INTO SPHERES*

JOHN W. BARRETT†, SÖREN BARTELS‡, XIAOBING FENG§, AND ANDREAS PROHL¶

Abstract. An explicit fully discrete finite element method, which satisfies the nonconvex side constraint at every node, is developed for approximating the p-harmonic flow for $p \in (1, \infty)$. Convergence of the method is established under certain conditions on the domain and mesh. Computational examples are presented to demonstrate finite-time blow-ups and qualitative geometric changes of weak solutions of the p-harmonic flow.

Key words. p-harmonic map, singular and degenerate PDE, finite element method, convergence analysis

AMS subject classifications. 35K55, 65M12, 68U10, 94A08

DOI. 10.1137/050639429

1. Introduction and summary.

Minimizing the energy

$$E_p(\mathbf{u}) := \frac{1}{p} \int_\Omega |\nabla \mathbf{u}|^p \, d\mathbf{x}, \quad 1 \le p < \infty, \tag{1.1}$$

for maps $\mathbf{u} : \Omega \to S^{m-1}$ ($m \ge 2$), where $\Omega \subset \mathbb{R}^n$ ($n \ge 1$) is bounded and $S^{m-1} \subset \mathbb{R}^m$ is the unit sphere, gives rise to p-harmonic maps. Such maps have natural applications such as micromagnetics [12, 28], liquid crystal theory [1, 24, 29, 7] ($p = 2$), or color image denoising [33, 34, 36, 11, 22] ($p = 1$). At present, there are not many schemes available to reliably approximate such maps. The main numerical difficulties are the nonconvexity of the constraint, $|\mathbf{u}| = 1$ a.e. in Ω, and the limited regularity and nonuniqueness of minimizers.

The first numerical schemes to approximate (1.1) in the case $p = 2$ were proposed in [17, 18, 23, 29]. The idea is in each search direction, first to reduce the energy functional ignoring the sphere constraint; then renormalize this solution $\mathbf{V}^j$ to obtain $\mathbf{U}^j = \frac{\mathbf{V}^j}{|\mathbf{V}^j|}$. However, the question is then whether the energy is still decreased during the renormalization step. This problem has been elegantly solved in [1], where an interesting convergent algorithm is proposed. Given an admissible $\mathbf{U}^j$, the strategy there is to decrease the energy $E_2(\mathbf{U}^j - \mathbf{V}^j)$ for $\mathbf{V}^j$ belonging to the tangential plane $\{\mathbf{w} \in H^1(\Omega, \mathbb{R}^m) : \langle \mathbf{w}, \mathbf{U}^j \rangle_{\mathbb{R}^m} = 0 \text{ a.e. in } \Omega\}$; then perform the renormalization $\mathbf{U}^{j+1} = \frac{\mathbf{U}^j - \mathbf{V}^j}{|\mathbf{U}^j - \mathbf{V}^j|} \in H^1(\Omega, S^{m-1})$. By construction, it follows that $E_2(\mathbf{U}^j - \mathbf{V}^j) = \min_{\mathbf{w}} E_2(\mathbf{U}^j - \mathbf{w}) \le E_2(\mathbf{U}^j)$, since $\mathbf{w} = \mathbf{0}$ is admissible. Moreover, $|\mathbf{U}^j - \mathbf{V}^j| \ge 1$

*Received by the editors September 2, 2005; accepted for publication (in revised form) September 21, 2006; published electronically May 4, 2007.
http://www.siam.org/journals/sinum/45-3/63942.html

†Department of Mathematics, Imperial College, London, SW7 2AZ, UK (jwb@ic.ac.uk).

‡Department of Mathematics, Humboldt-Universität zu Berlin, Unter den Linden 6, D-10099 Berlin, Germany (sba@mathematik.hu-berlin.de). Part of this work was written when this author visited Forschungsinstitut für Mathematik (ETH Zürich) in January, 2005.

§Department of Mathematics, The University of Tennessee, Knoxville, TN 37996 (xfeng@math.utk.edu). The work of this author is partially supported by NSF grant DMS-0410266.

¶Mathematisches Institut, Universität Tübingen, Auf der Morgenstelle 10, D-72076 Tübingen, Germany (prohl@na.uni-tuebingen.de).

a.e. in Ω, which is sufficient to guarantee decrease of the energy in the renormalization step. Recently, convergence of a finite element realization of this algorithm has been verified for restricted (acute) mesh partitions [6]. A generalization of this (Alouges') strategy to the degenerate regime $p \neq 2$ is easily possible, but convergence behavior seems unclear to the authors for the singular cases $p < 2$.

Another discretization approach is based on the convergent penalization strategy; see [30]. Here the nonconvex constraint is approximated by adding the penalty term $\varepsilon^{-1} \int_\Omega (|\mathbf{u}|^2 - 1)^2 \, \mathrm{d}\mathbf{x}$ to $E_p(\mathbf{u})$, leading to the unconstrained Ginzburg–Landau energy $E_{p,\varepsilon}(\mathbf{u})$ for an $\varepsilon > 0$. However, a numerical approximation of $E_{p,\varepsilon}(\mathbf{u})$ requires that the penalization parameter ε and the mesh parameter h be tuned. In [36] a different approach is proposed. This is based on the unconstrained minimization of

$$F_p(\mathbf{v}) := \int_\Omega \left| \nabla \left(\frac{\mathbf{v}}{|\,\mathbf{v}\,|} \right) \right|^p \mathrm{d}\mathbf{x}, \quad 1 \leq p < \infty, \tag{1.2}$$

for maps $\mathbf{v} : \Omega \to \mathbb{R}^m \backslash \{0\}$. A parametrization of the sphere then yields an efficient unconstrained numerical scheme, which is consistent with the nonconvex side-constraint and leads to energy decay. However, this approach restricts possible minimizers of (1.1) and leaves convergence properties of (1.2) unclear.

An alternative strategy to study minimizers of (1.1) is to consider the long-time behavior of the p-harmonic flow into spheres:

$$\mathbf{u}_t - \Delta_p \mathbf{u} = |\,\nabla \mathbf{u}\,|^p \mathbf{u} \quad \text{on } \Omega_T, \qquad \frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega_T, \tag{1.3}$$

$$|\,\mathbf{u}(\cdot,\cdot)\,| = 1 \quad \text{a.e. in } \Omega_T, \qquad \mathbf{u}(0,\cdot) = \mathbf{u}_0 \quad \text{on } \Omega, \tag{1.4}$$

for any $T > 0$. Here $\Omega_T := (0,T) \times \Omega$, $\partial\Omega_T := (0,T) \times \partial\Omega$ with $\partial\Omega$ being the boundary of Ω with normal $\mathbf{n}$. The system (1.3)–(1.4) characterizes the L^2-gradient flow of (1.1) with $\Delta_p \mathbf{u} \equiv \nabla \cdot (|\nabla \mathbf{u}|^{p-2} \nabla \mathbf{u})$. Solutions to this problem have been studied intensively over the last fifteen years, starting with the case $p = 2$ [14], and followed by $p > 2$ (existence [15, 27], nonuniqueness [25]), and $1 < p < 2$ (existence and nonuniqueness [19, 31]). Weak solutions to (1.3)–(1.4) satisfy (1.3) in a distributional sense and the initial condition in (1.4) in the sense of traces for $\mathbf{u}_0 \in W^{1,p}(\Omega, S^{m-1})$. Weak solutions that also satisfy the energy inequality

$$\int_0^t \| \, \mathbf{u}_t(s) \, \|_{L^2}^2 \, \mathrm{d}s + E_p\big(\mathbf{u}(t)\big) \leq E_p(\mathbf{u}_0) \quad \text{for a.e. } t \in (0,T) \tag{1.5}$$

are sometimes referred to as Struwe weak solutions (cf. [32]). This energy decay motivates the conjecture that there exists a subsequence $\{t_{k'}\} \subset \{t_k\}$, for $t_k \to \infty$, such that $\mathbf{u}^* = \lim_{k' \to \infty} \mathbf{u}(t_{k'}, \cdot)$ is a p-harmonic map, which is known for the case $p = 2$, and for any $p > 1$ in the case of small initial data [20]. We remark that there exist weak solutions to (1.3)–(1.4), which do not satisfy (1.5); cf. [8, 9, 35] for the case $p = 2$.

In order to verify existence of a weak solution to (1.3)–(1.4), the problem is modified to first finding a solution $\mathbf{u}^\varepsilon : \Omega_T \to \mathbb{R}^m$ to the following unconstrained penalized formulation [14, 16]: for $\varepsilon > 0$ and $T > 0$,

$$\mathbf{u}^\varepsilon_t - \Delta_p \mathbf{u}^\varepsilon + \frac{1}{2\varepsilon} \big(|\,\mathbf{u}^\varepsilon\,|^2 - 1 \big) \mathbf{u}^\varepsilon = 0 \quad \text{on } \Omega_T, \qquad \frac{\partial \mathbf{u}^\varepsilon}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega_T, \tag{1.6}$$

$$\mathbf{u}^\varepsilon(0,\cdot) = \mathbf{u}_0 \quad \text{on } \Omega. \tag{1.7}$$

Tracing the limit as $\varepsilon \to 0$ for solutions to (1.6)–(1.7) then leads to weak solutions of (1.3)–(1.4) satisfying (1.5) for the cases $1 < p < \infty$. When $p = 1$, local strong solutions are proved by Giga, Kashima, and Yamazaki in [22]. Apart from its use as an analytical tool, problem (1.6)–(1.7) is often the starting point to construct convergent discretizations for which the computed (discrete) solutions $\mathbf{U}_{k,h}^{\varepsilon}$ converge to solutions of (1.3)–(1.4) as the time step k, the mesh parameter h, and the penalization parameter ε tend to zero. Popularization of this approach is partially due to the fact that the direct construction of a convergent discretization of (1.3)–(1.4) is a nontrivial task.

The goal of this paper is to propose a convergent fully discrete finite element approximation of (1.3)–(1.4). Its construction is inspired by the recent work [2] for the Landau–Lifshitz equations. Our numerical scheme is based on the following equivalent reformulation of (1.3)–(1.4): find $\mathbf{u}$ satisfying the constraint and the initial condition such that

$$\int_0^T (\mathbf{u}_t(t), \mathbf{w})\,\mathrm{d}t + \int_0^T (|\nabla \mathbf{u}(t)|^{p-2}\nabla \mathbf{u}(t), \nabla \mathbf{w})\,\mathrm{d}t = 0 \quad \forall\, T > 0, \tag{1.8}$$

for all $\mathbf{w} \in L^2\big((0,T); W^{1,p}(\Omega, \mathbb{R}^m)\big) \cap L^\infty(\Omega_T, \mathbb{R}^m)$, such that $\langle \mathbf{w}, \mathbf{u}\rangle_{\mathbb{R}^m} = 0$ a.e. in Ω_T, where $(\boldsymbol{\eta}_1, \boldsymbol{\eta}_2) := \int_\Omega \langle \boldsymbol{\eta}_1, \boldsymbol{\eta}_2\rangle_{\mathbb{R}^{\ell_1\times\ell_2}}\,\mathrm{d}\mathbf{x}$ for $\boldsymbol{\eta}_i(t,\cdot) \in \mathbb{R}^{\ell_1\times\ell_2}$ and $\langle\cdot,\cdot\rangle_{\mathbb{R}^{\ell_1\times\ell_2}}$ is the standard inner product on $\mathbb{R}^{\ell_1\times\ell_2}$.

To introduce our finite element scheme and state the main convergence result, we need to make the following assumptions on the finite element partitioning:

(A1) Assuming that Ω is either polygonal ($n = 2$) or polyhedral ($n = 3$), let $\mathcal{T}_h$ be a quasi-uniform partitioning of Ω into disjoint open simplices K with $h_K := \mathrm{diam}(K)$ and $h := \max_{K\in\mathcal{T}_h} h_K$, so that $\overline{\Omega} = \cup_{K\in\mathcal{T}_h} \overline{K}$.

We require the quasi-uniformity constraint on the partitioning, as many of the proofs in this paper use the inverse inequalities on functions in $\boldsymbol{\mathcal{V}}_h$. The convergence proof of our finite element approximation for $p > n$ or $p = 2$ is fairly straightforward. In order to prove convergence if $p \le n$ and $p \ne 2$, our proof requires the denseness of $C^\infty(\overline{\Omega}, S^{m-1})$ in $W^{1,p}(\Omega, S^{m-1})$, which imposes the restrictions of either $p = n$ or $p < m - 1$; see [10]. Moreover, in this case we have to place a further restriction on the partitioning for a monotonicity argument to hold.

(A2) In addition to the assumption (A1) above, we assume that all simplices $K \in \mathcal{T}_h$ are right-angled. (For $n = 3$ this means that all tetrahedrons have one vertex with exactly one right angle, one vertex with exactly two right angles, and all other angles are strictly acute; see section 4 for more details. We note that a cube is easily partitioned into such tetrahedrons. Sufficient for our analysis is to assume that each element has n mutually perpendicular edges; the case that a tetrahedron has a vertex with three right angles is unrealistic in practice and therefore, for ease of exposition, excluded.)

Let $\mathcal{P}_1$ be the space of linear polynomials. We then introduce the following sets of functions:

$$\begin{aligned}
\boldsymbol{\mathcal{V}}_h &:= \big\{\mathbf{W} \in C(\overline{\Omega}, \mathbb{R}^m);\ \mathbf{W}\big|_K \in \boldsymbol{\mathcal{P}}_1(K, \mathbb{R}^m)\ \forall\ K \in \mathcal{T}_h\big\},\\
\boldsymbol{\mathcal{M}}_h &:= \big\{\mathbf{W} \in \boldsymbol{\mathcal{V}}_h;\ |\,\mathbf{W}(\mathbf{q}_i)\,| = 1\ \forall \text{ nodes } \mathbf{q}_i \text{ of } \mathcal{T}_h\big\},\\
\boldsymbol{\mathcal{F}}_h(\boldsymbol{\chi}) &:= \big\{\mathbf{W} \in \boldsymbol{\mathcal{V}}_h;\ \big\langle \mathbf{W}(\mathbf{q}_i), \boldsymbol{\chi}(\mathbf{q}_i)\big\rangle_{\mathbb{R}^m} = 0 \quad \forall \text{ nodes } \mathbf{q}_i \text{ of } \mathcal{T}_h\big\}, \quad \text{where } \boldsymbol{\chi} \in \boldsymbol{\mathcal{M}}_h.
\end{aligned}$$

Let $I_h : C(\overline{\Omega}, \mathbb{R}) \to V_h$ be the linear interpolation operator, where $V_h \equiv \boldsymbol{\mathcal{V}}_h$ with $m = 1$, such that $(I_h\, v)(\mathbf{q}_i) = v(\mathbf{q}_i)$ for all nodes $\mathbf{q}_i$ of $\mathcal{T}_h$. We then set $(\cdot,\cdot)_h$:

$C(\overline{\Omega},\mathbb{R}^m)\times C(\overline{\Omega},\mathbb{R}^m)\to\mathbb{R}$ to be

$$(\boldsymbol{\chi},\boldsymbol{\mathfrak{Z}})_h := \int_\Omega I_h\big(\langle\boldsymbol{\chi},\boldsymbol{\mathfrak{Z}}\rangle_{\mathbb{R}^m}\big)\,\mathrm{d}\mathbf{x} \equiv \sum_{K\in\mathcal{T}_h}\frac{|K|}{n+1}\sum_{\mathbf{q}_i\in K}\langle\boldsymbol{\chi}(\mathbf{q}_i),\boldsymbol{\mathfrak{Z}}(\mathbf{q}_i)\rangle_{\mathbb{R}^m}, \tag{1.9}$$

where $|K|$ is the area/volume of K.

Let k be the time step such that $Jk=T$ and $d_tv^j=k^{-1}\,(v^j-v^{j-1})$. Then a fully discrete implicit approximation of (1.8) reads: For $j=0\to J-1$, given $\widehat{\mathbf{U}}^j\in\boldsymbol{\mathcal{M}}_h$, find $\widehat{\mathbf{U}}^{j+1}\in\boldsymbol{\mathcal{M}}_h$ such that

$$(d_t\widehat{\mathbf{U}}^{j+1},\mathbf{W})+(\,|\,\nabla\widehat{\mathbf{U}}^{j+1}\,|^{p-2}\nabla\widehat{\mathbf{U}}^{j+1},\nabla\mathbf{W})=0\quad\forall\,\mathbf{W}\in\boldsymbol{\mathcal{F}}_h(\widehat{\mathbf{U}}^j), \tag{1.10}$$

where $\widehat{\mathbf{U}}^0$ is an approximation of $\mathbf{u}_0\in W^{1,p}(\Omega,S^{m-1})$. This problem is clearly too difficult to solve because of the imposed nonconvex constraint on $\boldsymbol{\mathcal{M}}_h$. However, since $\langle\mathbf{u}_t,\mathbf{u}\rangle_{\mathbb{R}^m}=0$, we may assume that $d_t\widehat{\mathbf{U}}^{j+1}$ is almost an element of $\boldsymbol{\mathcal{F}}_h(\widehat{\mathbf{U}}^j)$. This motivates our explicit scheme, which adapts the algorithm in [2] for the Landau–Lifshitz equations to the p-harmonic flow with $p\in(1,\infty)$.

SCHEME.

Step 1: Start with an initial vector field $\mathbf{U}^0\in\boldsymbol{\mathcal{M}}_h$.

Step 2: For $j=0\to J-1$, given $\mathbf{U}^j\in\boldsymbol{\mathcal{M}}_h$, find $\mathbf{V}^j\in\boldsymbol{\mathcal{F}}_h(\mathbf{U}^j)$ which solves

$$(\mathbf{V}^j,\mathbf{W})_h=-(\,|\,\nabla\mathbf{U}^j\,|^{p-2}\nabla\mathbf{U}^j,\nabla\mathbf{W})\qquad\forall\,\mathbf{W}\in\boldsymbol{\mathcal{F}}_h(\mathbf{U}^j).$$

Step 3: Define $\mathbf{U}^{j+1}\in\boldsymbol{\mathcal{M}}_h$ via

$$\mathbf{U}^{j+1}(\mathbf{q}_i)=\frac{\mathbf{U}^j(\mathbf{q}_i)+k\mathbf{V}^j(\mathbf{q}_i)}{|\,\mathbf{U}^j(\mathbf{q}_i)+k\mathbf{V}^j(\mathbf{q}_i)\,|}\qquad\forall\text{ nodes }\mathbf{q}_i\text{ of }\mathcal{T}_h.$$

We note that Step 2 is explicit, due to the use of numerical integration on the left-hand side, but remark that our analysis also holds if exact integration is used. For the fully discrete finite element solution $\{\mathbf{U}^j\}_{j\ge1}$ we define its constant and linear interpolations in time as follows:

$$\begin{aligned}&\underline{\mathbf{U}}(t,\cdot):=\mathbf{U}^{j-1}(\cdot)&&\forall\,t\in[t_{j-1},t_j),\quad 1\le j\le J,\\ &\mathbf{U}(t,\cdot):=\frac{t-t_{j-1}}{k}\mathbf{U}^j(\cdot)+\frac{t_j-t}{k}\mathbf{U}^{j-1}(\cdot)&&\forall\,t\in[t_{j-1},t_j],\quad 1\le j\le J.\end{aligned} \tag{1.11}$$

In this paper we will prove the following theorem.

THEOREM 1.1. *If $p=2$ or $p\in(n,\infty)$, let the assumption* (A1) *hold. If $p=n$ or $p\in(1,m-1)$, let the assumption* (A2) *hold. In addition, we assume that $\mathbf{u}_0\in W^{1,p}(\Omega,S^{m-1})$ and $\mathbf{U}^0\in\boldsymbol{\mathcal{M}}_h$ satisfies $\mathbf{U}^0\to\mathbf{u}_0$ strongly in $W^{1,p}(\Omega,\mathbb{R}^m)$ as $h\to0$, and*

$$k\le\begin{cases}o\big(\min\{h^{\frac{p}{p-1}},h^{p+\frac{n}{2}}\}\big) & \text{for } 1<p<2,\\ o\big(\min\{h^p,h^{1+n\,(1-\frac{1}{p})}\}\big) & \text{for } 2\le p<\infty.\end{cases} \tag{1.12}$$

Then there exists a subsequence of $\{\mathbf{U}\}_h$ such that as $h\to0$

$$\mathbf{U}\rightharpoonup\mathbf{u}\quad\text{weakly* in } L^\infty\big(0,T;W^{1,p}(\Omega,\mathbb{R}^m)\big),\quad \mathbf{U}_t\rightharpoonup\mathbf{u}_t\quad\text{weakly in } L^2(\Omega_T,\mathbb{R}^m),$$

where $\mathbf{u}\in H^1\big((0,T);L^2(\Omega,\mathbb{R}^m)\big)\cap L^\infty\big((0,T);W^{1,p}(\Omega,\mathbb{R}^m)\big)$ is a weak solution to (1.3)–(1.4).

To summarize: we prove convergence of our finite element approximation when

$$\begin{aligned} &n = 2, \text{ if either (i) } m = 2 \text{ and } p \in [2, \infty) \text{ or (ii) } m \geq 3 \text{ and } p \in (1, \infty); \\ &n = 3, \text{ if either (i) } m = 2 \text{ and } p \in \{2\} \cup [3, \infty) \text{ or (ii) } m = 3 \\ &\qquad \text{and } p \in (1, 2] \cup [3, \infty) \text{ or (iii) } m \geq 4 \text{ and } p \in (1, \infty). \end{aligned} \tag{1.13}$$

We also remark that the above theorem does not hold for $p = 1$, in which case the weak solutions are only BV-functions, instead of Sobolev functions. Moreover, computational experiments suggest that the constraint on the time step k is sharp as $p \to 1$.

The remainder of this paper is organized as follows. In section 2, we give a precise weak formulation of problem (1.3)–(1.4). In section 3, we establish the stability of the numerical solution and the mesh conditions on k described in Theorem 1.1. In section 4, we prove the convergence result of Theorem 1.1. Finally, in section 5, we present some numerical experiments, which show discrete finite-time blow-up and other qualitative behaviors of solutions of the p-harmonic flow for various values of p.

2. Preliminaries. With $\Omega \subset \mathbb{R}^n$ bounded, we define the nonlinear Sobolev space

$$W^{1,p}(\Omega, S^{m-1}) = \left\{ \mathbf{v} \in W^{1,p}(\Omega, \mathbb{R}^m) \,\middle|\, \mathbf{v} \in S^{m-1} \text{ a.e. in } \Omega \right\}, \quad 1 < p < \infty.$$

Critical points $\mathbf{u} \in W^{1,p}(\Omega, S^{m-1})$ of $E_p(\mathbf{u})$ for $p \in (1, \infty)$ can be characterized as solutions to the Euler–Lagrange equation

$$-\Delta_p \mathbf{u} = |\nabla \mathbf{u}|^p \mathbf{u} \quad \text{on } \Omega, \qquad \frac{\partial \mathbf{u}}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega. \tag{2.1}$$

If a map $\mathbf{u} \in W^{1,p}(\Omega, S^{m-1})$ satisfies (2.1) in the sense of distributions, $\mathbf{u}$ is called a weakly p-harmonic map. The p-harmonic flow (1.3)–(1.4) was first studied in [15, 26]. We now make precise what we mean by a weak solution to (1.3)–(1.4).

DEFINITION 2.1. *Let $\mathbf{u}_0 \in W^{1,p}(\Omega, \mathbb{R}^m)$, $p > 1$; then $\mathbf{u}$ is a weak solution to* (1.3)–(1.4) *if $\mathbf{u}$ is a function defined a.e. on $\Omega \times \mathbb{R}^+$ such that*

1. $\mathbf{u} \in L^\infty\big((0,T); W^{1,p}(\Omega, \mathbb{R}^m)\big) \cap H^1\big((0,T); L^2(\Omega, \mathbb{R}^m)\big)$ *for all $T > 0$,*
2. $\mathbf{u}$ *is weakly continuous for $t > 0$ with values in $W^{1,p}(\Omega, \mathbb{R}^m)$, i.e., for any test function $\mathbf{g} \in C^\infty(\Omega, \mathbb{R}^m)$,*

$$f_1(t) = \int_\Omega \langle \mathbf{u}, \mathbf{g} \rangle_{\mathbb{R}^m} \, \mathrm{d}\mathbf{x}, \quad f_2(t) = \int_\Omega \langle \nabla \mathbf{u}, \nabla \mathbf{g} \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}$$

   are continuous for $t > 0$, with possible modification on a set of measure zero on $(0, \infty)$,
3. $|\mathbf{u}| = 1$ *a.e. on $\Omega \times \mathbb{R}^+$,*
4. (1.3) *holds in the sense of distributions,*
5. *the initial condition holds in the sense of traces.*

Verification of the existence of a weak solution to (1.3)–(1.4) uses monotonicity arguments for a penalization approach to approximate the p-harmonic flow on the space $W^{1,p}(\Omega, \mathbb{R}^m)$. A parabolic version of Murat's lemma then gives enough compactness to identify limits of terms of a wedged version of the penalized problem as a wedged version of (1.3), which holds in distributional sense. This weak solution is known to satisfy the energy law (1.5), and we refer to [32, 24] for further details in this direction. Also, weak solutions to (1.3)–(1.4) are not unique; see, e.g., [31] and [25].

Of course, the subsequent proof of Theorem 1.1 can be considered as an alternative way to construct weak solutions to (1.3)–(1.4).

REMARK 2.1. *In* [31] (*see also* [27], *for* $p > 2$), *Misawa demonstrates existence of weak solutions to* (1.3)–(1.4) *by the Rothe method: set* $\mathbf{u}^0 = \mathbf{u}_0$; *then for* $j \geq 1$ *minimizers* $\mathbf{u}^j = \operatorname{argmin}_{W^{1,p}(\Omega,S^{m-1})} E_p(\mathbf{v})$, *of* $E_p(\mathbf{v}) := E_p(\mathbf{v}) + \frac{1}{2k}\int_\Omega |\,\mathbf{v}-\mathbf{u}^{j-1}\,|^2\,\mathrm{d}\mathbf{x}$, *exist, and solve*

$$d_t\mathbf{u}^j - \Delta_p\mathbf{u}^j = \left(|\,\nabla\mathbf{u}^j\,|^p + \frac{k}{2}\,|\,d_t\mathbf{u}^j\,|^2\right)\mathbf{u}^j \quad \text{on } \Omega, \qquad \frac{\partial \mathbf{u}^j}{\partial \mathbf{n}} = 0 \quad \text{on } \partial\Omega. \tag{2.2}$$

In addition, they satisfy a semidiscrete version of energy inequality (1.5) *on the equidistant time mesh* $\{t_j\}_{j\geq 0}$. *Then a compactness argument as in* [15] *together with a parabolic version of Murat's lemma* (*cf.* [27]) *proves subsequence convergence to a weak solution of* (1.3)–(1.4) *as* $k \to 0$. *Unfortunately, the scheme* (2.2) *is not practically useful, due to the nonconvex constraint.*

We end this section by introducing some notation and stating a few useful results. Let $1 < p < \infty$. For all $\mathbf{P}, \mathbf{Q} \in \mathbb{R}^{m\times n}$, $m, n \geq 1$, and $\delta \geq 0$ there exist positive constants $C_i(p, m, n)$ such that

$$\text{(i)}\ \big||\,\mathbf{P}\,|^{p-2}\mathbf{P} - |\,\mathbf{Q}\,|^{p-2}\mathbf{Q}\big| \leq C_1\big(|\,\mathbf{P}\,| + |\,\mathbf{Q}\,|\big)^{p-2+\delta}|\,\mathbf{P}-\mathbf{Q}\,|^{1-\delta},$$
$$\text{(ii)}\ \big\langle |\,\mathbf{P}\,|^{p-2}\mathbf{P} - |\,\mathbf{Q}\,|^{p-2}\mathbf{Q}, \mathbf{P}-\mathbf{Q}\big\rangle_{\mathbb{R}^{m\times n}} \geq C_2\big(|\,\mathbf{P}\,| + |\,\mathbf{Q}\,|\big)^{p-2-\delta}|\,\mathbf{P}-\mathbf{Q}\,|^{2+\delta}. \tag{2.3}$$

For example, these results were proved in [5] for the case $\mathbb{R}^{n\times n}$, and that proof easily transfers to the present case. We recall the following results concerning $(\cdot,\cdot)_h$:

$$\|\boldsymbol{\chi}\|_{L^2}^2 \leq |\boldsymbol{\chi}|_h^2 := (\boldsymbol{\chi},\boldsymbol{\chi})_h \leq (n+2)\|\boldsymbol{\chi}\|_{L^2}^2 \qquad \forall \boldsymbol{\chi} \in \boldsymbol{\mathcal{V}}_h\,; \tag{2.4}$$

$$|(\boldsymbol{\chi},\boldsymbol{\mathfrak{Z}}) - (\boldsymbol{\chi},\boldsymbol{\mathfrak{Z}})_h| \leq Ch\|\boldsymbol{\chi}\|_{L^2}\|\nabla\boldsymbol{\mathfrak{Z}}\|_{L^2} \leq C\|\boldsymbol{\chi}\|_{L^2}\|\boldsymbol{\mathfrak{Z}}\|_{L^2} \qquad \forall \boldsymbol{\chi}, \boldsymbol{\mathfrak{Z}} \in \boldsymbol{\mathcal{V}}_h. \tag{2.5}$$

For later purposes, we also introduce the linear interpolation operator $\mathcal{I}_h : C(\overline{\Omega},\mathbb{R}^m) \to \boldsymbol{\mathcal{V}}_h$ such that $(\mathcal{I}_h\mathbf{v})(\mathbf{q}_i) = \mathbf{v}(\mathbf{q}_i)$ for all nodes $\mathbf{q}_i$ of $\mathcal{T}_h$. Finally, throughout the paper we adopt the standard notation for Sobolev spaces and their associated norms. For notational convenience, we drop the domain from the norm subscript if the domain is Ω, that is, $\|\cdot\|_{L^2} \equiv \|\cdot\|_{L^2(\Omega)}$.

3. Stability. As a first step toward showing the convergence of our numerical scheme to a weak solution of problem (1.3)–(1.4), we shall establish a discrete version of the energy inequality (1.5).

LEMMA 3.1. *Let the assumption* (A1) *hold. Let* $\mathbf{u}_0 \in W^{1,p}(\Omega, S^{m-1})$ *and* $\mathbf{U}^0 \in \boldsymbol{\mathcal{M}}_h$ *satisfy* $\mathbf{U}^0 \to \mathbf{u}_0$ *strongly in* $W^{1,p}(\Omega,\mathbb{R}^m)$ *as* $h \to 0$, *and let* k *satisfy* (1.12). *Then the iterates* $\{\mathbf{V}^{j-1}, \mathbf{U}^j\}_{j=1}^J$ *computed from our scheme satisfy for* $j = 1 \to J$

$$\begin{aligned}(1-c_0)k\sum_{\ell=1}^{j}\|\,d_t\mathbf{U}^\ell\,\|_{L^2}^2 + \frac{1}{p}\|\,\nabla\mathbf{U}^j\,\|_{L^p}^p &\leq (1-c_1)k\sum_{\ell=0}^{j-1}\|\,\mathbf{V}^\ell\,\|_{L^2}^2 + \frac{1}{p}\|\,\nabla\mathbf{U}^j\,\|_{L^p}^p \\ &\leq \frac{1}{p}\|\,\nabla\mathbf{U}^0\,\|_{L^p}^p + c_2,\end{aligned} \tag{3.1}$$

where c_i *are* $o(1)$.

Proof. First, we choose $\mathbf{W} = \mathbf{V}^j$ in Step 2 of the scheme. As $\mathbf{U}^j \in \boldsymbol{\mathcal{M}}_h$, on noting (2.4) and on applying an inverse inequality, we conclude that

$$\begin{aligned}\|\,\mathbf{V}^j\,\|_{L^2}^2 \leq |\mathbf{V}^j|_h^2 &= -\big(|\,\nabla\mathbf{U}^j\,|^{p-2}\nabla\mathbf{U}^j, \nabla\mathbf{V}^j\big) \leq \int_\Omega |\,\nabla\mathbf{U}^j\,|^{p-1}|\,\nabla\mathbf{V}^j\,|\,\mathrm{d}\mathbf{x} \\ &\leq \|\,\nabla\mathbf{U}^j\,\|_{L^{2(p-1)}}^{p-1}\|\,\nabla\mathbf{V}^j\,\|_{L^2} \leq Ch^{-p}\,\|\,\mathbf{U}^j\,\|_{L^{2(p-1)}}^{p-1}\|\,\mathbf{V}^j\,\|_{L^2} \leq Ch^{-2p}.\end{aligned} \tag{3.2}$$

We note that if $\|\nabla \mathbf{U}^j\|_{L^p} \le C$, then we have, via (2.4) and inverse inequalities, the improved bound

$$\begin{aligned} \| \mathbf{V}^j \|_{L^2}^2 \le |\mathbf{V}^j|_h^2 &= -\big(|\nabla \mathbf{U}^j|^{p-2}\nabla \mathbf{U}^j, \nabla \mathbf{V}^j\big) \le \| \nabla \mathbf{U}^j \|_{L^p}^{p-1} \| \nabla \mathbf{V}^j \|_{L^p} \\ &\le Ch^{-1} \| \mathbf{V}^j \|_{L^p} \le \begin{cases} Ch^{-2} & \text{for } 1 \le p \le 2, \\ Ch^{-2-n(1-\frac{2}{p})} & \text{for } 2 \le p < \infty. \end{cases} \end{aligned} \tag{3.3}$$

The following argument is adapted from [2]. On defining $\mathbf{R}^j := \mathbf{U}^{j+1} - \mathbf{U}^j - k\mathbf{V}^j \in \boldsymbol{\mathcal{V}}_h$, then Step 3 of the scheme yields for all nodes $\mathbf{q}_i$ of $\mathcal{T}_h$ that

$$|\,\mathbf{R}^j(\mathbf{q}_i)\,| = \left| \frac{\mathbf{U}^j(\mathbf{q}_i) + k\mathbf{V}^j(\mathbf{q}_i)}{|\,\mathbf{U}^j(\mathbf{q}_i) + k\mathbf{V}^j(\mathbf{q}_i)\,|} - \mathbf{U}^j(\mathbf{q}_i) - k\mathbf{V}^j(\mathbf{q}_i) \right| = \Big| 1 - \big|\mathbf{U}^j(\mathbf{q}_i) + k\mathbf{V}^j(\mathbf{q}_i)\big| \Big|,$$

and since $1 \le |\,\mathbf{U}^j(\mathbf{q}_i) + k\mathbf{V}^j(\mathbf{q}_i)\,| = \sqrt{1 + k^2|\,\mathbf{V}^j(\mathbf{q}_i)\,|^2} \le 1 + \frac{k^2}{2}|\,\mathbf{V}^j(\mathbf{q}_i)\,|^2$, we conclude that

$$|\,\mathbf{R}^j(\mathbf{q}_i)\,| \le \frac{k^2}{2}|\,\mathbf{V}^j(\mathbf{q}_i)\,|^2. \tag{3.4}$$

Therefore, on recalling (2.4), we have that

$$\int_\Omega |\mathbf{R}^j|\,\mathrm{d}\mathbf{x} \le \int_\Omega I_h[\,|\mathbf{R}^j|\,]\,\mathrm{d}\mathbf{x} \le \frac{k^2}{2}\int_\Omega I_h[\,|\mathbf{V}^j|^2\,]\,\mathrm{d}\mathbf{x} \le \frac{k^2(n+2)}{2}\int_\Omega |\mathbf{V}^j|^2\,\mathrm{d}\mathbf{x}. \tag{3.5}$$

Similarly, it follows from (2.4) and an inverse inequality that

$$\|\mathbf{R}^j\|_{L^2}^2 \le |\mathbf{R}^j|_h^2 \le \frac{k^4}{4}\|\mathbf{V}^j\|_{L^\infty}^2 |\mathbf{V}^j|_h^2 \le Ck^4h^{-n}\|\mathbf{V}^j\|_{L^2}^4; \tag{3.6}$$

and hence we have that

$$\|d_t\mathbf{U}^{j+1}\|_{L^2}^2 \le [\|\mathbf{V}^j\|_{L^2} + k^{-1}\|\mathbf{R}^j\|_{L^2}]^2 \le [1 + Ckh^{-\frac{n}{2}}\|\mathbf{V}^j\|_{L^2}]^2\|\mathbf{V}^j\|_{L^2}^2. \tag{3.7}$$

Now, choosing $\mathbf{W} = \mathbf{V}^j = d_t\mathbf{U}^{j+1} - k^{-1}\mathbf{R}^j$ in Step 2 of our scheme, noting the convexity of $|\nabla \cdot|^p$, that $\mathbf{U}^j$, $\mathbf{U}^{j+1} \in \boldsymbol{\mathcal{M}}_h$ and applying (2.3)(i) with $\delta = 2 - p$ if $p \in (1, 2]$ and $\delta = 0$ if $p \in [2, \infty)$, together with inverse estimates and (3.5), we arrive at

$$\begin{aligned} \| \mathbf{V}^j \|_{L^2}^2 + \frac{1}{p}d_t\| \nabla \mathbf{U}^{j+1} \|_{L^p}^p &\le |\mathbf{V}^j|_h^2 + (|\nabla \mathbf{U}^{j+1}|^{p-2}\nabla \mathbf{U}^{j+1}, \nabla(d_t\mathbf{U}^{j+1})) \\ &= k^{-1}\big(|\,\nabla \mathbf{U}^j\,|^{p-2}\nabla \mathbf{U}^j, \nabla \mathbf{R}^j\big) + \big(|\,\nabla \mathbf{U}^{j+1}\,|^{p-2}\nabla \mathbf{U}^{j+1} \\ &\qquad - |\,\nabla \mathbf{U}^j\,|^{p-2}\nabla \mathbf{U}^j, \nabla(d_t\mathbf{U}^{j+1})\big) \\ &\le k^{-1}\| \nabla \mathbf{U}^j \|_{L^\infty}^{p-1}\| \nabla \mathbf{R}^j \|_{L^1} + Ck^{1-\delta}[\,\|\nabla \mathbf{U}^{j+1}\|_{L^\infty} \\ &\quad + \|\nabla \mathbf{U}^j\|_{L^\infty}\,]^{p-2+\delta}\|\nabla(d_t\mathbf{U}^{j+1})\|_{L^2}^{2-\delta} \\ &\le Ck^{-1}h^{-p}\| \mathbf{R}^j \|_{L^1} + Ck^{1-\delta}h^{-p}\|d_t\mathbf{U}^{j+1}\|_{L^2}^{2-\delta} \\ &\le Ckh^{-p}\| \mathbf{V}^j \|_{L^2}^2 + Ck^{1-\delta}h^{-p}\|d_t\mathbf{U}^{j+1}\|_{L^2}^{2-\delta}. \end{aligned} \tag{3.8}$$

We first consider the simpler case, $p \in [2, \infty)$. It follows from our assumptions on $\mathbf{U}^\mathbf{0}$ that there exists a constant $C_1 > 0$ such that $\| \nabla \mathbf{U}^0 \|_{L^p} \le C_1$ for all $h > 0$. Assuming that $\| \nabla \mathbf{U}^j \|_{L^p} \le C_1$ and $k = O(h^{1+n(1-\frac{1}{p})})$, it then follows from (3.3) that

there exists a constant $C_2 > 0$ such that $kh^{-\frac{n}{2}}\|\mathbf{V}^j\|_{L^2} \leq C_2$. Therefore, combining (3.7) and (3.8) yields in the case $p \in [2,\infty)$ that there exists a constant $C_3 > 0$ such that

$$\left(1 - C_3\frac{k}{h^p}\right) k\| \mathbf{V}^j \|_{L^2}^2 + \frac{1}{p}\| \nabla\mathbf{U}^{j+1} \|_{L^p}^p \leq \frac{1}{p}\| \nabla\mathbf{U}^j \|_{L^p}^p. \tag{3.9}$$

If the time step k satisfies $C_3 k \leq h^p$, it follows from the above inequality that $\| \nabla\mathbf{U}^{j+1} \|_{L^p} \leq C_1$. Hence, by induction, (3.9) holds for $j = 0 \to J-1$ under the above two restrictions on k. On recalling our assumptions on k, (1.12), the desired stability result (3.1) for $p \in [2,\infty)$, with no c_2 term on the right-hand side, follows from summing (3.9) and noting from (3.7) that $\|d_t\mathbf{U}^{j+1}\|_{L^2}^2 \leq (1 + o(1))\|\mathbf{V}^j\|_{L^2}^2$.

We now consider the case $p \in (1,2)$. First, there exists a constant $C_4(p) > 0$ such that

$$\|d_t\mathbf{U}^{j+1}\|_{L^2}^p \leq \|d_t\mathbf{U}^{j+1}\|_{L^2}^2 + C_4. \tag{3.10}$$

Assuming $k = O(h^{p+\frac{n}{2}})$, it then follows from (3.2) that there exists a constant $C_5 > 0$ such that $kh^{-\frac{n}{2}}\|\mathbf{V}^j\|_{L^2} \leq C_5$. Therefore combining (3.7), (3.8), and (3.10) yields in the case $p \in (1,2)$ that there exists a constant $C_6 > 0$ such that

$$\left(1 - C_6\frac{k^{p-1}}{h^p}\right) k\| \mathbf{V}^j \|_{L^2}^2 + \frac{1}{p}\| \nabla\mathbf{U}^{j+1} \|_{L^p}^p \leq \frac{1}{p}\| \nabla\mathbf{U}^j \|_{L^p}^p + C_5\frac{k^p}{h^p}. \tag{3.11}$$

On recalling our assumptions on k, (1.12), the desired stability result (3.1) for $p \in (1,2)$ then follows from summing (3.11) and noting (3.7). ∎

4. Convergence. The following lemma, where we adopt the notation (1.11), will be needed for showing the convergence of our scheme.

LEMMA 4.1. *Let the assumptions of Lemma* 3.1 *hold. Then for all* $\mathbf{W} \in L^2\big((0,T);\boldsymbol{\mathcal{F}}_h(\underline{\mathbf{U}})\big)$ *it follows that*

$$\begin{aligned}\left|\int_0^T \left[(\mathbf{U}_t,\mathbf{W}) + (|\nabla\underline{\mathbf{U}}|^{p-2}\nabla\underline{\mathbf{U}},\nabla\mathbf{W})\right]\mathrm{d}t\right| \\ \leq C\left[kh^{-(\frac{n}{2}+1+\sigma)}\|\mathbf{W}\|_{L^2(\Omega_T)} + h\|\nabla\mathbf{W}\|_{L^2(\Omega_T)}\right],\end{aligned} \tag{4.1}$$

where $\sigma = 0$ *if* $p \in (1,2)$ *and* $\sigma = n(\frac{1}{2} - \frac{1}{p})$ *if* $p \in [2,\infty)$.

Proof. Write $\underline{\mathbf{V}} = \mathbf{U}_t - k^{-1}\underline{\mathbf{R}}$ in Step 2 of our scheme to obtain for any $\mathbf{W} \in L^2\big((0,T);\boldsymbol{\mathcal{F}}_h(\underline{\mathbf{U}})\big)$ that

$$\begin{aligned}\int_0^T \left[(\mathbf{U}_t,\mathbf{W}) + (|\nabla\underline{\mathbf{U}}|^{p-2}\nabla\underline{\mathbf{U}},\nabla\mathbf{W})\right]\mathrm{d}t \\ = k^{-1}\int_0^T (\underline{\mathbf{R}},\mathbf{W})\,\mathrm{d}t + \int_0^T \left[(\mathbf{U}_t,\mathbf{W}) - (\mathbf{U}_t,\mathbf{W})_h\right]\mathrm{d}t.\end{aligned} \tag{4.2}$$

From (3.6), (3.3), and (3.1) we have that

$$\begin{aligned}\int_0^T \| \underline{\mathbf{R}} \|_{L^2}^2\,\mathrm{d}t &\leq Ck^4h^{-n}\int_0^T \| \underline{\mathbf{V}} \|_{L^2}^4\,\mathrm{d}t \\ &\leq Ck^4h^{-(n+2+2\sigma)}\int_0^T \| \underline{\mathbf{V}} \|_{L^2}^2\,\mathrm{d}t \leq Ck^4h^{-(n+2+2\sigma)}.\end{aligned} \tag{4.3}$$

Hence the desired result (4.1) follows from (4.2), (4.3), (2.5), and (3.1). ∎

It follows from (3.1), our assumptions on $\mathbf{U}^0$, and as $\mathbf{U} \in \boldsymbol{\mathcal{M}}_h$ that there exists a function $\mathbf{u} \in H^1((0,T);\, L^2(\Omega,\mathbb{R}^m)) \cap L^\infty((0,T);W^{1,p}(\Omega,\mathbb{R}^m))$ and a subsequence of $\{\mathbf{U}\}_h$ such that as $h \to 0$

$$\begin{aligned} &\mathbf{U},\, \underline{\mathbf{U}} \rightharpoonup \mathbf{u} \text{ weakly* in } L^\infty\big(0,T;W^{1,p}(\Omega,\mathbb{R}^m)\big), \\ &\mathbf{U},\, \underline{\mathbf{U}} \to \mathbf{u} \text{ strongly in } L^q\big(\Omega_T,\mathbb{R}^m\big), \quad \mathbf{U}_t \rightharpoonup \mathbf{u}_t \text{ weakly in } L^2(\Omega_T,\mathbb{R}^m), \end{aligned} \tag{4.4}$$

where $q < \infty$ if $p \le n$ and $q = \infty$ if $p > n$. Furthermore, we have that (1.5) holds.

As $\mathbf{U} \in \boldsymbol{\mathcal{M}}_h$, it follows that $I_h[|\mathbf{U}|] \equiv 1$, and hence for every $K \in \mathcal{T}_h$ that

$$\begin{aligned} \big\| |\,\mathbf{U}\,|^2 - 1\big\|_{L^p(K)} &\le Ch^2 \|D^2(|\,\mathbf{U}\,|^2)\|_{L^p(K)} \\ &\le Ch^2 \big\|\nabla \mathbf{U}\big\|^2_{L^{2p}(K)} \le Ch \|\nabla \mathbf{U}\big\|_{L^p(K)}. \end{aligned} \tag{4.5}$$

Therefore, we deduce that

$$|\,\mathbf{u}\,| = 1 \text{ a.e. in } \Omega_T. \tag{4.6}$$

Next, in order to identify the limit of the p-Laplacian term in (4.1), we need to establish that

$$|\,\nabla \underline{\mathbf{U}}\,|^{p-2} \nabla \underline{\mathbf{U}} \rightharpoonup |\,\nabla \mathbf{u}\,|^{p-2} \nabla \mathbf{u} \quad \text{weakly in } L^{\frac{p}{p-1}}(\Omega_T,\mathbb{R}^{m\times n}) \quad \text{as } h \to 0. \tag{4.7}$$

The standard employment of Minty's lemma for monotone operators (see [37], "the decisive monotonicity trick") is not so straightforward, as (4.1) is only valid for $\mathbf{W} \in L^2((0,T);\boldsymbol{\mathcal{F}}_h(\underline{\mathbf{U}}))$ and not for all $\mathbf{W} \in L^2((0,T);\boldsymbol{\mathcal{V}}_h)$. Obviously, if $p = 2$, then (4.7) follows immediately from (4.4). The lemma below establishes a stronger version of (4.7) in the easier case when $p \in (n,\infty)$.

LEMMA 4.2. *In addition to the assumptions of Lemma* 3.1 *holding, let* $p \in (n,\infty)$. *Then we have for the subsequence* $\{\mathbf{U}\}_\mathbf{h}$ *of* (4.4) *that*

$$|\,\nabla \underline{\mathbf{U}}\,|^{p-2} \nabla \underline{\mathbf{U}} \to |\,\nabla \mathbf{u}\,|^{p-2} \nabla \mathbf{u} \quad \textit{strongly in } L^{\frac{p}{p-1}}(\Omega_T,\mathbb{R}^{m\times n}) \qquad \textit{as } h \to 0. \tag{4.8}$$

Proof. As $p \in (n,\infty)$, it follows that $\mathcal{I}_h \mathbf{u}$ is well-defined and

$$\mathcal{I}_h \mathbf{u} \to \mathbf{u} \quad \text{strongly in } L^\infty\big((0,T);W^{1,p}(\Omega,\mathbb{R}^m)\big) \text{ and hence in } L^\infty\big(\Omega_T,\mathbb{R}^m\big). \tag{4.9}$$

We deduce from (2.3)(ii) with $\delta = p-2$ that

$$\begin{aligned} \int_{\Omega_T} |\nabla(\mathbf{u} - \underline{\mathbf{U}})|^p \,\mathrm{d}\mathbf{x}\mathrm{d}t \le{}& \int_{\Omega_T} |\nabla \mathbf{u}|^{p-2} \langle \nabla \mathbf{u}, \nabla(\mathbf{u} - \underline{\mathbf{U}})\rangle_{\mathbb{R}^{m\times n}} \,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &- \int_{\Omega_T} |\nabla \underline{\mathbf{U}}|^{p-2} \langle \nabla \underline{\mathbf{U}}, \nabla(\mathbf{u} - \mathcal{I}_h \mathbf{u})\rangle_{\mathbb{R}^{m\times n}} \,\mathrm{d}\mathbf{x}\mathrm{d}t \\ &- \int_{\Omega_T} |\nabla \underline{\mathbf{U}}|^{p-2} \langle \nabla \underline{\mathbf{U}}, \nabla(\mathcal{I}_h \mathbf{u} - \underline{\mathbf{U}})\rangle_{\mathbb{R}^{m\times n}} \,\mathrm{d}\mathbf{x}\mathrm{d}t \\ =:{}& T_1 + T_2 + T_3. \end{aligned}$$

It follows from (4.4), (3.1), and (4.9) that $T_1, T_2 \to 0$ as $h \to 0$. As $\mathcal{I}_h \underline{\mathbf{U}} \equiv \underline{\mathbf{U}}$ and $\underline{\mathbf{U}}, \mathcal{I}_h \mathbf{u} \in \boldsymbol{\mathcal{M}}_h$ (recall (4.6)), we have that $\mathcal{I}_h \mathbf{u} - \underline{\mathbf{U}} \equiv \mathbf{W} + \mathbf{Z}$, where

$$\begin{aligned} &\mathbf{W} = \mathcal{I}_h[\mathbf{u} - \langle \mathbf{u}, \underline{\mathbf{U}}\rangle_{\mathbb{R}^m} \underline{\mathbf{U}}] \in \boldsymbol{\mathcal{F}}_h(\underline{\mathbf{U}}), \\ &\mathbf{Z} = \mathcal{I}_h[(\langle \mathbf{u}, \underline{\mathbf{U}}\rangle_{\mathbb{R}^m} - 1)\underline{\mathbf{U}}] = -\frac{1}{2}\mathcal{I}_h[|\mathbf{u} - \underline{\mathbf{U}}|^2 \underline{\mathbf{U}}]. \end{aligned} \tag{4.10}$$

It follows from (4.10), (4.1), (1.12), an inverse inequality, and (3.1) that

$$
\begin{aligned}
|T_3| &\le C\,\big[1+\|\mathbf{U}_t\|_{L^2(\Omega_T)}\big]\,\|\mathcal{I}_h[\mathbf{u}-\langle\mathbf{u},\underline{\mathbf{U}}\rangle_{\mathbb{R}^m}\underline{\mathbf{U}}]\|_{L^2(\Omega_T)}\\
&\quad + C\|\underline{\mathbf{U}}\|^{p-1}_{L^\infty(0,T;W^{1,p}(\Omega))}\,\|\mathcal{I}_h[(\langle\mathbf{u},\underline{\mathbf{U}}\rangle_{\mathbb{R}^m}-1)\underline{\mathbf{U}}]\|_{L^1(0,T;W^{1,p}(\Omega))}\\
&\le C\,\big[\|\mathbf{u}-\langle\mathbf{u},\underline{\mathbf{U}}\rangle_{\mathbb{R}^m}\underline{\mathbf{U}}\|_{L^\infty(\Omega_T)}+\|\mathcal{I}_h[|\mathbf{u}-\underline{\mathbf{U}}|^2\,\underline{\mathbf{U}}]\|_{L^1(0,T;W^{1,p}(\Omega))}\big]\\
&\le C\,\big[\|\mathbf{u}-\underline{\mathbf{U}}\|_{L^\infty(\Omega_T)}+\||\mathbf{u}-\underline{\mathbf{U}}|^2\,\underline{\mathbf{U}}\|_{L^1(0,T;W^{1,p}(\Omega))}\big]\\
&\le C\|\mathbf{u}-\underline{\mathbf{U}}\|_{L^\infty(\Omega_T)}.
\end{aligned}
\tag{4.11}
$$

On noting (4.4), as $p>n$, we have that $T_3\to 0$ as $h\to 0$; and hence we have that the subsequence of $\{\mathbf{U}\}_h$ in (4.4) is such that

$$
\underline{\mathbf{U}}\to\mathbf{u}\quad\text{strongly in } L^p\big(0,T;W^{1,p}(\Omega,\mathbb{R}^m)\big)\qquad\text{as } h\to 0.
\tag{4.12}
$$

The above and (2.3)(i) with $\delta=0$ immediately yields the desired result (4.8). ∎

Unfortunately, if $p\in(1,n]$ and $p\neq 2$, the proof of the desired result (4.7) is far more complicated. One difficulty occurs as $\mathcal{I}_h$ is not well-defined on $\mathbf{u}$. If one replaces $\mathcal{I}_h$ by a generalized interpolation operator, $\mathcal{I}^g_h$, then $\mathcal{I}^g_h\underline{\mathbf{U}}\not\equiv\underline{\mathbf{U}}$, $\mathcal{I}^g_h\mathbf{u}\notin\boldsymbol{\mathcal{M}}_{\mathbf{h}}$, and, moreover, a generalization of (4.10) with the second crucial identity for $\mathbf{Z}$, exploited in (4.11) above, does not hold. To overcome this difficulty, we employ a density argument by smoothing $\mathbf{u}$ and continue to work with $\mathcal{I}_h$. However, to obtain a generalization of the second identity for $\mathbf{Z}$ in (4.10) we require this smoothed $\mathbf{u}(\cdot,t)$ to belong to S^{m-1} and not just $\mathbb{R}^m$. This requires the denseness of $C^\infty(\overline{\Omega},S^{m-1})$ in $W^{1,p}(\Omega,S^{m-1})$, which imposes the restrictions of either $p=n$ or $p<m-1$; see [10]. Another difficulty occurs if $p\in(1,n]$ and $p\neq 2$ as $\underline{\mathbf{U}}\to\mathbf{u}$ in $L^q(\Omega_T,\mathbb{R}^m)$ only for $q<\infty$ and not for $q=\infty$; recall (4.4). To overcome this we require a discrete version of Theorem 2.1 in [15], which exploits a monotonicity argument to deduce that the term II' in the proof there is nonpositive. To obtain a discrete analogue of this, we require the right angle constraint, (A2), on our partitioning, which we now discuss in more detail.

Let $\{\mathbf{e}_i\}_{i=1}^n$ be the standard orthonormal vectors in $\mathbb{R}^n$, such that the jth component of $\mathbf{e}_i$ is δ_{ij}, $i,j=1\to n$. Given nonzero constants ρ_i, $i=1\to n$, let $\widehat{K}(\{\rho_i\}_{i=1}^n)$ be a reference simplex in $\mathbb{R}^n$ with vertices $\{\widehat{\mathbf{q}}_i\}_{i=0}^n$, where $\widehat{\mathbf{q}}_0$ is the origin and $\widehat{\mathbf{q}}_i=\widehat{\mathbf{q}}_{i-1}+\rho_i\mathbf{e}_i$, $i=1\to n$. Then under assumption (A2), given a $K\in\mathcal{T}_h$ with vertices $\{\mathbf{q}_{j_i}\}_{i=0}^n$ such that $\mathbf{q}_{i_0}$ is not a right-angled vertex, there exists a rotation/reflection matrix $B_K\in\mathbb{R}^{n\times n}$ such that the mapping $\mathcal{F}_K:\widehat{\mathbf{x}}\in\mathbb{R}^n\to\mathbf{q}_{j_0}+B_K\widehat{\mathbf{x}}\in\mathbb{R}^n$ maps the vertex $\widehat{\mathbf{q}}_i$ to $\mathbf{q}_{j_i}$, $i=0\to n$, and hence $\widehat{K}(\{\rho_i\}_{i=1}^n)$ to K. Then for all $K\in\mathcal{T}_h$, $\phi\in C(\overline{K},\mathbb{R})$, and $\boldsymbol{\phi}\in C(\overline{K},\mathbb{R}^m)$, we set for all $\widehat{\mathbf{x}}\in\widehat{K}(\{\rho_i\}_{i=1}^n)$

$$
\begin{aligned}
&\widehat{\phi}(\widehat{\mathbf{x}})\equiv\phi(\mathcal{F}_K\widehat{\mathbf{x}}), && (\widehat{I\phi})(\widehat{\mathbf{x}})\equiv(I_h\phi)(\mathcal{F}_K\widehat{\mathbf{x}});\\
&\widehat{\boldsymbol{\phi}}(\widehat{\mathbf{x}})\equiv\boldsymbol{\phi}(\mathcal{F}_K\widehat{\mathbf{x}}), && (\widehat{\mathcal{I}\boldsymbol{\phi}})(\widehat{\mathbf{x}})\equiv(\mathcal{I}_h\boldsymbol{\phi})(\mathcal{F}_K\widehat{\mathbf{x}}).
\end{aligned}
\tag{4.13}
$$

We have for any $\mathbf{Z}\in\boldsymbol{\mathcal{V}}_h$ and $K\in\mathcal{T}_h$ that

$$
\nabla\mathbf{Z}\equiv(\widehat{\nabla}\widehat{\mathbf{Z}})\,B_K^{-1}\qquad\text{on } K;
\tag{4.14}
$$

here $\mathbf{x}\equiv(x_1,\ldots,x_n)^T$, $\nabla\equiv(\frac{\partial}{\partial x_1},\ldots,\frac{\partial}{\partial x_n})$, $\widehat{\mathbf{x}}\equiv(\widehat{x}_1,\ldots,\widehat{x}_n)^T$, and $\widehat{\nabla}\equiv(\frac{\partial}{\partial\widehat{x}_1},\ldots,\frac{\partial}{\partial\widehat{x}_n})$. It is easily deduced (see, e.g., [4] for details) that for any z_1, $z_2\in C(\overline{\Omega},\mathbb{R})$

$$
\nabla(I_h[z_1z_2])=\nabla(I_hz_2)D(I_hz_1)+\nabla(I_hz_1)D(I_hz_2),
\tag{4.15}
$$

where for any $Z \in V^h$,

$$D(Z)\,|_K := B_K\,\widehat{D}(\widehat{Z})B_K^{-1} \qquad \forall\ K \in \mathcal{T}_h, \tag{4.16}$$

and $\widehat{D}(\widehat{Z})$ is the $n \times n$ diagonal matrix with diagonal entries

$$[\widehat{D}(\widehat{Z})]_{ii} := \frac{1}{2}\left[\widehat{Z}(\widehat{\mathbf{q}}_i) + \widehat{Z}(\widehat{\mathbf{q}}_{i-1})\right], \qquad i = 1 \to n. \tag{4.17}$$

LEMMA 4.3. *In addition to the assumptions of Lemma* 3.1 *holding, let either* $p = n$ *or* $p < m-1$, *and let the assumption* (A2) *hold. Then we have for the subsequence* $\{\mathbf{U}\}_{\mathbf{h}}$ *of* (4.4) *and for any* $s \in [1, p)$ *that*

$$\nabla \underline{\mathbf{U}} \to \nabla \mathbf{u} \qquad \textit{strongly in } L^s(\Omega_T, \mathbb{R}^{m\times n}) \textit{ as } h \to 0. \tag{4.18}$$

Hence the desired result (4.7) *holds.*

Proof. As either $p = n$ or $p < m-1$, it follows that $C^\infty(\overline{\Omega}, S^{m-1})$ is a dense subset of $W^{1,p}(\Omega, S^{m-1})$; see [10]. Hence for any fixed $\alpha \in (0,1)$ there exists $\mathbf{u}_\alpha \in L^\infty(0,T; C^\infty(\overline{\Omega}, S^{m-1}))$ such that

$$\|\mathbf{u} - \mathbf{u}_\alpha\|_{L^\infty(0,T;W^{1,p}(\Omega))} \le \alpha^2. \tag{4.19}$$

Therefore $\mathcal{I}_h\mathbf{u}_\alpha$ is well-defined and

$$\mathcal{I}_h\mathbf{u}_\alpha \to \mathbf{u}_\alpha \quad \text{strongly in } L^\infty\big(0,T; W^{1,p}(\Omega, \mathbb{R}^m\big). \tag{4.20}$$

In addition, we introduce $\boldsymbol{\eta}_\alpha : \mathbb{R}^m \to \mathbb{R}^m$ and $\eta_\alpha : \mathbb{R} \to \mathbb{R}$ such that

$$\boldsymbol{\eta}_\alpha(\mathbf{y}) := \eta_\alpha(|\mathbf{y}|)\mathbf{y} := \begin{cases} \mathbf{y} & \text{if } |\mathbf{y}| \le \alpha, \\ \frac{\alpha}{|\mathbf{y}|}\mathbf{y} & \text{if } |\mathbf{y}| \ge \alpha. \end{cases} \tag{4.21}$$

On adopting the notation in (4.13) and (4.14), we have for all $\mathbf{Z} \in \boldsymbol{\mathcal{V}}_h$ and $K \in \mathcal{T}_h$ that

$$\frac{\partial}{\partial \widehat{x}_k}\widehat{\mathcal{I}}[\boldsymbol{\eta}_\alpha(\widehat{\mathbf{Z}})] \equiv A_\alpha^{(k)}(\widehat{\mathbf{Z}})\frac{\partial \widehat{\mathbf{Z}}}{\partial \widehat{x}_k} \qquad \text{on } \widehat{K}, \qquad k = 1 \to n, \tag{4.22}$$

where $A_\alpha^{(k)}(\widehat{\mathbf{Z}}) \in \mathbb{R}^{m\times m}$ is such that for $i,\, j = 1 \to m$

$$\begin{aligned}[A_\alpha^{(k)}(\widehat{\mathbf{Z}})]_{ij} &= \frac{1}{2}[\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|) + \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)]\delta_{ij} \\ &+ \frac{1}{2}\frac{[\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|) - \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)]}{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| - |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|}\frac{([\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)]_i + [\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})]_i)([\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)]_j + [\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})]_j)}{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| + |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|}. \end{aligned}$$

For any $\mathbf{y} \in \mathbb{R}^m$, we deduce from the monotonicity of η_α that

$$\begin{aligned}\mathbf{y}^T A_\alpha^{(k)}(\widehat{\mathbf{Z}})\mathbf{y} &\ge \frac{1}{2}[\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|) + \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)]|\mathbf{y}|^2 \\ &\quad + \frac{1}{2}\frac{[\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|) - \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)]}{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| - |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|}\frac{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k) + \widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|^2}{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| + |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|}|\mathbf{y}|^2 \\ &\ge \frac{1}{2}\Bigg([\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|) + \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)] + \frac{[\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|) - \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)]}{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| - |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|} \\ &\qquad \times (|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| + |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)\Bigg)|\mathbf{y}|^2 \\ &\ge \frac{[\eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)|)|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| - \eta_\alpha(|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|)|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|]}{|\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_k)| - |\widehat{\mathbf{Z}}(\widehat{\mathbf{q}}_{k-1})|}|\mathbf{y}|^2 \ge 0. \end{aligned} \tag{4.23}$$

Therefore $A_\alpha^{(k)}(\widehat{\mathbf{Z}})$ is symmetric positive semidefinite for any $\mathbf{Z} \in \boldsymbol{\mathcal{V}}_h$. Similar to (4.23), we have for all $\mathbf{Z} \in \boldsymbol{\mathcal{V}}_h$ and on any $K \in \mathcal{T}^h$ that

$$\begin{aligned}\mathbf{y}^T A_\alpha^{(k)}(\widehat{\mathbf{Z}})\mathbf{y} &\leq \frac{\eta_\alpha(|\widehat{\mathbf{Z}}(\mathbf{q}_{k-1})|)|\widehat{\mathbf{Z}}(\mathbf{q}_k)| - \eta_\alpha(|\widehat{\mathbf{Z}}(\mathbf{q}_k)|)|\widehat{\mathbf{Z}}(\mathbf{q}_{k-1})|}{|\widehat{\mathbf{Z}}(\mathbf{q}_k)| - |\widehat{\mathbf{Z}}(\mathbf{q}_{k-1})|}|\mathbf{y}|^2 \\ &\leq [\eta_\alpha(|\widehat{\mathbf{Z}}(\mathbf{q}_k)|) + \eta_\alpha(|\widehat{\mathbf{Z}}(\mathbf{q}_{k-1})|)]|\mathbf{y}|^2 \leq 2|\mathbf{y}|^2 \qquad \forall \mathbf{y} \in \mathbb{R}^m. \end{aligned} \tag{4.24}$$

It follows from (4.14), $B_K^{-1} \equiv B_K^T$, (4.22), (4.23), and (4.24) that for all $\mathbf{Z}$, $\mathbf{Y} \in \boldsymbol{\mathcal{V}}_h$, and on any $K \in \mathcal{T}_h$

$$\begin{aligned}\langle \nabla \mathbf{Z}, \nabla(\mathcal{I}_{\mathbf{h}}[\eta_\alpha(\mathbf{Y}-\mathbf{Z})])\rangle_{\mathbb{R}^{\mathrm{m}\times\mathrm{n}}} &= \langle (\widehat{\nabla}\widehat{\mathbf{Z}})B_K^{-1}, (\widehat{\nabla}(\widehat{\mathcal{I}}[\eta_\alpha(\widehat{\mathbf{Y}}-\widehat{\mathbf{Z}})]))B_K^{-1}\rangle_{\mathbb{R}^{m\times n}} \\ &= \langle \widehat{\nabla}\widehat{\mathbf{Z}}, \widehat{\nabla}(\widehat{\mathcal{I}}[\eta_\alpha(\widehat{\mathbf{Y}}-\widehat{\mathbf{Z}})])\rangle_{\mathbb{R}^{m\times n}} \\ &= \sum_{k=1}^n \left\langle \frac{\partial \widehat{\mathbf{Z}}}{\partial \widehat{x}_k}, A_\alpha^{(k)}(\widehat{\mathbf{Y}}-\widehat{\mathbf{Z}})\frac{\partial(\widehat{\mathbf{Y}}-\widehat{\mathbf{Z}})}{\partial \widehat{x}_k}\right\rangle_{\mathbb{R}^m} \\ &\leq C|\widehat{\nabla}\widehat{\mathbf{Z}}||\widehat{\nabla}\widehat{\mathbf{Y}}| \leq C|\nabla \mathbf{Z}||\nabla \mathbf{Y}|. \end{aligned} \tag{4.25}$$

Hence we deduce from (4.25) that for all $\mathbf{Z}, \mathbf{Y} \in \boldsymbol{\mathcal{V}}_h$, and $K \in \mathcal{T}_h$

$$\int_K |\nabla \mathbf{Z}|^{p-2}\langle \nabla \mathbf{Z}, \nabla(\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{Y}-\mathbf{Z})])\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x} \leq C\|\nabla \mathbf{Z}\|_{L^p(K)}^{p-1}\|\nabla \mathbf{Y}\|_{L^p(K)}. \tag{4.26}$$

It is this bound, which we use in bounding T_3 below (containing the analogue of the term II' in the proof of Theorem 2.1 in [15]), that exploits the right-angle constraint, (A2), on the partitioning.

As $|\mathbf{u}_\alpha| = |\underline{\mathbf{U}}| = 1$ in Ω_T, we have from (4.21) that

$$\langle \boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha - \underline{\mathbf{U}}), \underline{\mathbf{U}}\rangle_{\mathbb{R}^m} = -\frac{1}{2}\langle \boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha - \underline{\mathbf{U}}), \mathbf{u}_\alpha - \underline{\mathbf{U}}\rangle_{\mathbb{R}^m}. \tag{4.27}$$

It follows from (4.27) and (4.21) that

$$\|I_h[\langle \boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha - \underline{\mathbf{U}}), \underline{\mathbf{U}}\rangle_{\mathbb{R}^m}]\|_{L^\infty(\Omega_T)} \leq \frac{1}{2}\alpha\|\mathbf{u}_\alpha - \underline{\mathbf{U}}\|_{L^\infty(\Omega_T)} \leq \alpha. \tag{4.28}$$

It follows from (4.14), (4.13), (4.27), and (4.21) that for all $K \in \mathcal{T}_h$

$$\begin{aligned}\|\nabla(I_h[\langle \boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha - \underline{\mathbf{U}}), \underline{\mathbf{U}}\rangle_{\mathbb{R}^m}])\|_{L^p(K)} &\leq C\|\widehat{\nabla}(\widehat{I}[\langle \boldsymbol{\eta}_\alpha(\widehat{\mathbf{u}}_\alpha - \widehat{\underline{\mathbf{U}}}), \widehat{\underline{\mathbf{U}}}\rangle_{\mathbb{R}^m}])\|_{L^p(\widehat{K})} \\ &\leq C\alpha\|\widehat{\nabla}|\widehat{I}(\widehat{\mathbf{u}}_\alpha) - \widehat{\underline{\mathbf{U}}}|\|_{L^p(\widehat{K})} \\ &\leq C\alpha\|\widehat{\nabla}[\widehat{I}(\widehat{\mathbf{u}}_\alpha) - \widehat{\underline{\mathbf{U}}}]\|_{L^p(\widehat{K})} \\ &\leq C\alpha\|\nabla[I_h(\mathbf{u}_\alpha) - \underline{\mathbf{U}}]\|_{L^p(K)}. \end{aligned} \tag{4.29}$$

For a.e. $t \in (0,T)$ let

$$\begin{aligned}\mathcal{J}_{h,\alpha}(t) &:= \{\text{nodes } \mathbf{q}_i \text{ of } \mathcal{T}_h : |(\mathcal{I}_h\mathbf{u}_\alpha)(t,\mathbf{q}_i) - \underline{\mathbf{U}}(t,\mathbf{q}_i)| \geq \alpha\}, \\ \mathcal{T}_{h,\alpha}(t) &:= \{K \in \mathcal{T}_h : K \text{ has a vertex } \mathbf{q}_i \in \mathcal{J}^{h,\alpha}(t)\}, \\ \mathcal{R}_{h,\alpha}(t) &:= \bigcup_{K \in \mathcal{T}_{h,\alpha}(t)} \overline{K}. \end{aligned} \tag{4.30}$$

It follows from (2.4), (1.9), and (4.30) that

$$\begin{aligned}\frac{\alpha^2}{n+1}\int_0^T |\mathcal{R}_{h,\alpha}(t)|\,\mathrm{d}t &\le \int_0^T |\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}}|_h^2\,\mathrm{d}t \le (n+2)\|\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}}\|^2_{L^2(\Omega_T)}\\ &\le 2(n+2)\left[\|\mathbf{u}_\alpha - \mathcal{I}_h\mathbf{u}_\alpha\|^2_{L^2(\Omega_T)} + \|\mathbf{u}_\alpha - \underline{\mathbf{U}}\|^2_{L^2(\Omega_T)}\right].\end{aligned} \tag{4.31}$$

Hence we deduce from (4.31), (4.20), (4.4), and (4.19) that

$$\lim_{h\to 0}\int_0^T |\mathcal{R}_{h,\alpha}(t)|\,\mathrm{d}t \le C\alpha^2. \tag{4.32}$$

In addition, it follows from (4.20) and (4.19) that

$$\lim_{h\to 0}\int_0^T\int_{\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^p\,\mathrm{d}\mathbf{x}\mathrm{d}t \le \lim_{h\to 0}\int_0^T\int_{\mathcal{R}_{h,\alpha}(t)} |\nabla\mathbf{u}|^p\,\mathrm{d}\mathbf{x}\mathrm{d}t + C\alpha^2. \tag{4.33}$$

For any $s\in[1,p)$, we have that

$$\begin{aligned}&\int_0^T\left(\int_{\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})|^s\mathrm{d}\mathbf{x}\right)\mathrm{d}t\\ &\quad\le \left(\int_0^T |\mathcal{R}_{h,\alpha}(t)|\,\mathrm{d}t\right)^{\frac{p-s}{p}}\left(\int_0^T\left(\int_{\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})|^p\mathrm{d}\mathbf{x}\right)\mathrm{d}t\right)^{\frac{s}{p}}.\end{aligned} \tag{4.34}$$

Let $p^\star := \max\{2,p\}$. Then on applying a Hölder inequality, noting (4.19), (4.20), (4.4), and (2.3)(ii) with $\delta = p^\star - 2$, we have that

$$\begin{aligned}&\left(\int_0^T\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})|^p\mathrm{d}\mathbf{x}\mathrm{d}t\right)^{\frac{p^\star}{p}}\\ &\quad\le C\int_0^T\left(\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)} [\,|\nabla(\mathcal{I}_h\mathbf{u}_\alpha)| + |\nabla\underline{\mathbf{U}}|\,]^p\,\mathrm{d}\mathbf{x}\right)^{-\frac{(p^\star-p)}{p}}\\ &\qquad\times\left(\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})|^p\mathrm{d}\mathbf{x}\right)^{\frac{p^\star}{p}}\mathrm{d}t\\ &\quad\le C\int_0^T\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)} [\,|\nabla(\mathcal{I}_h\mathbf{u}_\alpha)| + |\nabla\underline{\mathbf{U}}|\,]^{p-p^\star}\,|\nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})|^{p^\star}\mathrm{d}\mathbf{x}\mathrm{d}t\\ &\quad\le C\int_0^T\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^{p-2}\langle\nabla(\mathcal{I}_h\mathbf{u}_\alpha), \nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &\qquad - C\int_0^T\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)} |\nabla\underline{\mathbf{U}}|^{p-2}\langle\nabla\underline{\mathbf{U}}, \nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t =: T_1 + T_2.\end{aligned} \tag{4.35}$$

It follows from (3.1), (4.20), and (4.19) that

$$\begin{aligned}|T_1| \le &\left|\int_{\Omega_T} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^{p-2}\langle\nabla(\mathcal{I}_h\mathbf{u}_\alpha), \nabla(\mathcal{I}_h\mathbf{u}_\alpha - \underline{\mathbf{U}})\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t\right|\\ &+ C\left(\int_0^T\int_{\mathcal{R}_{h,\alpha}(t)} |\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^p\,\mathrm{d}\mathbf{x}\mathrm{d}t\right)^{\frac{p-1}{p}}.\end{aligned} \tag{4.36}$$

Hence we deduce from (4.36), (4.20), (4.4), and (4.19) that

$$\lim_{h\to 0}|T_1| \le C\,\alpha^2 + C\lim_{h\to 0}\left(\int_0^T\int_{\mathcal{R}_{h,\alpha}(t)}|\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^p\,\mathrm{d}\mathbf{x}\mathrm{d}t\right)^{\frac{p-1}{p}}. \tag{4.37}$$

Next we note from (4.30) and (4.21) that

$$\begin{aligned} T_2 &= -\int_0^T\int_{\Omega\setminus\mathcal{R}_{h,\alpha}(t)}|\nabla\underline{\mathbf{U}}|^{p-2}\langle\nabla\underline{\mathbf{U}},\nabla(\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}})])\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &= \int_0^T\int_{\mathcal{R}_{h,\alpha}(t)}|\nabla\underline{\mathbf{U}}|^{p-2}\langle\nabla\underline{\mathbf{U}},\nabla(\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}})])\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &\quad-\int_{\Omega_T}|\nabla\underline{\mathbf{U}}|^{p-2}\langle\nabla\underline{\mathbf{U}},\nabla(\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}})])\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t =: T_3 - T_4. \end{aligned} \tag{4.38}$$

It follows from (4.26) and (3.1) that

$$\begin{aligned} T_3 &\le C\|\nabla\underline{\mathbf{U}}\|^{p-1}_{L^p(\Omega_T)}\left(\int_0^T\int_{\mathcal{R}_{h,\alpha}(t)}|\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^p\,\mathrm{d}\mathbf{x}\mathrm{d}t\right)^{\frac{1}{p}}\\ &\le C\left(\int_0^T\int_{\mathcal{R}_{h,\alpha}(t)}|\nabla(\mathcal{I}_h\mathbf{u}_\alpha)|^p\,\mathrm{d}\mathbf{x}\mathrm{d}t\right)^{\frac{1}{p}}. \end{aligned} \tag{4.39}$$

Noting that $\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}})-\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}\underline{\mathbf{U}}]\in\boldsymbol{\mathcal{F}}_h(\underline{\mathbf{U}})$, we have that

$$\begin{aligned} T_4 &= \int_{\Omega_T}|\nabla\underline{\mathbf{U}}|^{p-2}\langle\nabla\underline{\mathbf{U}},\nabla(\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}})-\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}\underline{\mathbf{U}}])\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t\\ &\quad-\int_{\Omega_T}|\nabla\underline{\mathbf{U}}|^{p-2}\langle\nabla\underline{\mathbf{U}},\nabla(\mathcal{I}_h[\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}\underline{\mathbf{U}}])\rangle_{\mathbb{R}^{m\times n}}\,\mathrm{d}\mathbf{x}\mathrm{d}t =: T_5 + T_6. \end{aligned} \tag{4.40}$$

It then follows from (4.1), (1.12), an inverse inequality, (3.1), (2.4), (4.21), and (4.19) that

$$\begin{aligned} |T_5| &\le C\left[1+\|\mathbf{U}_t\|_{L^2(\Omega_T)}\right]\left[\int_0^T\|\mathcal{I}_h[\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}})-\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}\underline{\mathbf{U}}]\|^2_{L^2}\,\mathrm{d}t\right]^{\frac{1}{2}}\\ &\le C\|\mathcal{I}_h[\mathbf{u}_\alpha-\underline{\mathbf{U}}]\|_{L^2(\Omega_T)} \le C\,[\,\|\mathbf{u}-\underline{\mathbf{U}}\|_{L^2(\Omega_T)}+\|\mathbf{u}_\alpha-\mathcal{I}_h\mathbf{u}_\alpha\|_{L^2(\Omega_T)}+\alpha^2\,]. \end{aligned} \tag{4.41}$$

We note from (3.1), (4.15), (4.16), (4.17), (4.28), and (4.29) that

$$\begin{aligned} |T_6| &\le \|\nabla\underline{\mathbf{U}}\|^{p-1}_{L^p(\Omega_T)}\,\|\nabla(\mathcal{I}_h[\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}\underline{\mathbf{U}}])\|_{L^p(\Omega_T)}\\ &\le C\,\|\nabla(\mathcal{I}_h[\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}\underline{\mathbf{U}}])\|_{L^p(\Omega_T)}\\ &\le \|\underline{\mathbf{U}}\|_{L^\infty(\Omega_T)}\|\nabla(\mathcal{I}_h[\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}])\|_{L^p(\Omega_T)}\\ &\quad+\|\mathcal{I}_h[\langle\boldsymbol{\eta}_\alpha(\mathbf{u}_\alpha-\underline{\mathbf{U}}),\underline{\mathbf{U}}\rangle_{\mathbf{R}^m}]\|_{L^\infty(\Omega_T)}\|\nabla\underline{\mathbf{U}}\|_{L^p(\Omega_T)}\\ &\le C\alpha\left[\|\nabla\mathbf{u}_\alpha\|_{L^p(\Omega_T)}+\|\nabla\underline{\mathbf{U}}\|_{L^p(\Omega_T)}\right]. \end{aligned} \tag{4.42}$$

On combining (4.34)–(4.42), (3.1), (4.19), (4.20), (4.32), (4.33), and (4.4) we have that given any $\epsilon > 0$, there exist an $\alpha(\epsilon)$ and an $h_0(\alpha)$ such that for the subsequence $\{\mathbf{U}\}_{\mathbf{h}}$ of (4.4)

$$\|\nabla(\mathbf{u}_\alpha - \underline{\mathbf{U}})\|_{L^s(\Omega_T)} \le \epsilon \qquad \forall h \le h_0. \tag{4.43}$$

The desired result (4.18) then follows immediately from (4.43), (4.19), and (4.20). Finally, the desired result (4.7) follows immediately from (4.18) and (3.1); cf. [31, Lemma 6]. ∎

We now are ready to give a proof for Theorem 1.1.

Proof of Theorem 1.1. Given any $\boldsymbol{\phi} \in C^\infty(\overline{\Omega_T}, \mathbb{R}^m)$, let $\mathbf{w} = \mathbf{u} \wedge \boldsymbol{\phi}$ and $\mathbf{W} = \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})$. Here $\wedge$ is the wedge (exterior) product, which is the extension of the cross (vector) product on vectors in $\mathbb{R}^3$ to $\mathbb{R}^n$. Interpolation theory yields that

$$\begin{aligned} \| \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi}) - \underline{\mathbf{U}} \wedge \boldsymbol{\phi} \|_{L^2}^2 &\le Ch^4 \sum_{K \in \mathcal{T}_h} \| D^2\big(\underline{\mathbf{U}} \wedge \boldsymbol{\phi}\big) \|_{L^2(K)}^2 \\ &\le Ch^4 \left[\| \, |\underline{\mathbf{U}}| \, |D^2 \boldsymbol{\phi}| \, \|_{L^2}^2 + \| \, |\nabla \underline{\mathbf{U}}| \, |\nabla \boldsymbol{\phi}| \, \|_{L^2}^2 \right] \\ &\le Ch^4 \| \boldsymbol{\phi} \|_{H^2}^2 + Ch^{4-\gamma} \|\nabla \boldsymbol{\phi}\|_{L^\infty}^2 \| \nabla \underline{\mathbf{U}} \|_{L^p}^2, \end{aligned} \tag{4.44}$$

where $\gamma = n(2-p)/p$ if $p \in (1,2]$ and $\gamma = 0$ if $p \in (2, \infty)$. Therefore (4.44) and (4.4) yield that $\mathbf{W} \to \mathbf{w}$ strongly in $L^2(\Omega_T, \mathbb{R}^m)$, which in turn implies that

$$\int_{\Omega_T} \langle \mathbf{U}_t, \mathbf{W} \rangle_{\mathbb{R}^m} \, \mathrm{d}\mathbf{x}\mathrm{d}t \to \int_{\Omega_T} \langle \mathbf{u}_t, \mathbf{w} \rangle_{\mathbb{R}^m} \, \mathrm{d}\mathbf{x}\mathrm{d}t \qquad \text{as } h \to 0. \tag{4.45}$$

We now consider the p-Laplacian term. Similarly to (4.44), we have that

$$\| \nabla(\mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi}) - \underline{\mathbf{U}} \wedge \boldsymbol{\phi}) \|_{L^p}^2 \le Ch^2 \left[\| \boldsymbol{\phi} \|_{W^{2,p}}^2 + \|\nabla \boldsymbol{\phi}\|_{L^\infty}^2 \| \nabla \underline{\mathbf{U}} \|_{L^p}^2 \right]. \tag{4.46}$$

On noting the vector identity $\langle \nabla \mathbf{z}, \nabla(\mathbf{z} \wedge \boldsymbol{\phi}) \rangle_{\mathbb{R}^{m \times n}} = \langle \nabla \mathbf{z}, \mathbf{z} \wedge \nabla \boldsymbol{\phi} \rangle_{\mathbb{R}^{m \times n}}$, (4.4), and (4.7), it follows that as $h \to 0$

$$\begin{aligned} &\int_{\Omega_T} |\nabla \underline{\mathbf{U}}|^{p-2} \langle \nabla \underline{\mathbf{U}}, \nabla(\underline{\mathbf{U}} \wedge \boldsymbol{\phi}) \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}\mathrm{d}t = \int_{\Omega_T} |\nabla \underline{\mathbf{U}}|^{p-2} \langle \nabla \underline{\mathbf{U}}, \underline{\mathbf{U}} \wedge \nabla \boldsymbol{\phi} \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}\mathrm{d}t \\ &\to \int_{\Omega_T} |\nabla \mathbf{u}|^{p-2} \langle \nabla \mathbf{u}, \mathbf{u} \wedge \nabla \boldsymbol{\phi} \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}\mathrm{d}t = \int_{\Omega_T} |\nabla \mathbf{u}|^{p-2} \langle \nabla \mathbf{u}, \nabla(\mathbf{u} \wedge \boldsymbol{\phi}) \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}\mathrm{d}t. \end{aligned} \tag{4.47}$$

Noting (4.46), (4.47), and (4.7), we have that

$$\int_{\Omega_T} |\nabla \underline{\mathbf{U}}|^{p-2} \langle \nabla \underline{\mathbf{U}}, \nabla \mathbf{W} \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}\mathrm{d}t \longrightarrow \int_{\Omega_T} |\nabla \mathbf{u}|^{p-2} \langle \nabla \mathbf{u}, \nabla \mathbf{w} \rangle_{\mathbb{R}^{m \times n}} \, \mathrm{d}\mathbf{x}\mathrm{d}t \tag{4.48}$$

as $h \to 0$. Finally if $p \in (1,2]$, we deduce from an inverse inequality that

$$\begin{aligned} h^2 \|\nabla \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})\|_{L^2}^2 &\le h^2 \|\nabla \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})\|_{L^\infty}^{2-p} \|\nabla \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})\|_{L^p}^p \\ &\le Ch^p \|\mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})\|_{L^\infty}^{2-p} \|\nabla \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})\|_{L^p}^p \\ &\le Ch^p \|\nabla \mathcal{I}_h(\underline{\mathbf{U}} \wedge \boldsymbol{\phi})\|_{L^p}^p. \end{aligned} \tag{4.49}$$

It follows from (4.45), (4.48), (4.1), (4.44), (4.4), our constraints on the time step k, (1.12), (4.49), (4.46), and (3.1) that we now can pass to the limit $h \to 0$ in (4.1) to obtain that for all $\boldsymbol{\phi} \in C^\infty(\overline{\Omega_T}, \mathbb{R}^m)$

$$\int_0^T \left[(\mathbf{u}_t, \mathbf{u} \wedge \boldsymbol{\phi}) + (|\nabla \mathbf{u}|^{\mathrm{p}-2} \nabla \mathbf{u}, \nabla(\mathbf{u} \wedge \boldsymbol{\phi})) \right] \, \mathrm{dt} = 0. \tag{4.50}$$

However, as (4.6) holds, the above equation implies that $\mathbf{u} : \Omega_T \to S^{m-1}$ satisfies (1.3)–(1.4) in the weak sense; see Lemma 1.8 in [32] or the proof of Theorem 2.2 in [15]. Hence, we have proved Theorem 1.1. ∎

5. Numerical experiments: Finite-time blow-up and geometric changes. The global existence and the nonuniqueness of weak solutions to (1.3)–(1.4) for $p > 1$, and the local existence of smooth solutions motivate finite-time blow-up studies. We say that the numerical solution $\mathbf{U}$, for fixed mesh parameters, blows up at $t^\star$ if

$$\|\nabla \mathbf{U}(t^\star, \cdot)\|_{L^\infty} = \max_{\mathbf{V} \in \boldsymbol{\mathcal{M}}_h} \|\nabla \mathbf{V}\|_{L^\infty}.$$

We remark that this discrete blow-up behavior may disappear as the mesh is refined, and may be different if we changed from Neumann to Dirichlet-type boundary conditions. We employ our convergent numerical scheme to compute such phenomena. Throughout these numerical experiments, we set $\Omega := (-1, 1)^2 \subset \mathbb{R}^2$, i.e., $n = 2$, and $m = 3$; recall (1.13). We choose a uniform right-angled triangulation of Ω with $h = \sqrt{2}/2^3$ and set $\mathbf{U}^0 \equiv \mathcal{I}_h \mathbf{u_0}$. Unless otherwise stated, we choose $k = h^{s+1/2}/10$ for $s = \max\{p/(p-1), p\}$. In all of the experiments reported below we observed that $E_p(\mathbf{U}^{j+1}) \leq E_p(\mathbf{U}^j)$ for all $j \geq 0$ for this choice of k; recall the stability requirements of Theorem 1.1 and that for $p \in (1, 2)$ we computed with $p = 3/2$ and $5/4 \Rightarrow p/(p-1) \geq p + 1 \equiv p + \frac{n}{2}$. Finally, as $m = 3$, below we plot at each node $\mathbf{q}_i$ of $\mathcal{T}_h$ a vector based on the first two components of $\mathbf{U}^j(\mathbf{x}_i)$.

EXAMPLE 5.1. *Let $b > 0$, and define $\mathbf{u}_0 : \overline{\Omega} \to S^2$ by*

$$\mathbf{u}_0(\mathbf{x}) := \left(\frac{\mathbf{x}}{|\mathbf{x}|} \sin \phi(|\mathbf{x}|), \cos \phi(|\mathbf{x}|) \right), \quad \textit{where} \quad \phi(r) := \begin{cases} br^2 & \textit{for } r \leq 1, \\ b & \textit{for } r \geq 1. \end{cases}$$

According to the results in [13, 32] *we expect finite-time blow-up for $p = 2$ if $b > \pi$. We choose*
(ai) *$p = 2$ and $b = \pi/2$ and* (aii) *$p = 2$ and $b = 3\pi/2$,*
(bi) *$p = 3/2$ and $b = \pi/2$ and* (bii) *$p = 3/2$ and $b = 3\pi/2$,*
(ci) *$p = 5/2$ and $b = \pi/2$ and* (cii) *$p = 5/2$ and $b = 3\pi/2$.*

Figure 5.1 displays the numerical solution in Example 5.1(ai) at various times. As expected, we do not observe finite-time blow-up; at $t = 0.9090$ all vectors point in the same direction. We observe a similar behavior in (bi) and (ci).

In Figure 5.2 we plot the numerical solution in Example 5.1(aii) at various times. Blow-up occurs at $t \approx 0.4$ when the vector at the origin changes its direction from $(0, 0, 1)$ to $-(0, 0, 1)$. A zoom at the values of the nodes in a neighborhood of the origin at some times is displayed in Figure 5.5 and magnifies the change of direction at the origin.

The blow-up happens differently for (bii). Some snapshots of its dynamics are displayed in Figure 5.3. In the time interval $0.5 \leq t \leq 0.8$ all vectors apart from the one at $\mathbf{x} = \mathbf{0}$ approximately point out of the plane. Then, at time $t \approx 0.93$ the vector at the origin changes direction so that a uniform state is achieved.

The behavior in (cii) is different from that in (aii) and (bii). No blow-up occurs; cf. Figure 5.6. The vector field $\mathbf{U}$ obtained in (cii) is shown for various times in Figure 5.4.

The lower right plot in Figure 5.6 displays the energy $E_2\big(\mathbf{U}(t, \cdot)\big)$ in Example 5.1 (ai) obtained with $k = \frac{1}{2}h^2$ using the numerical integration rule (1.9) as stated in Step 2 of our scheme, and for comparison exact integration. The results clearly indicate

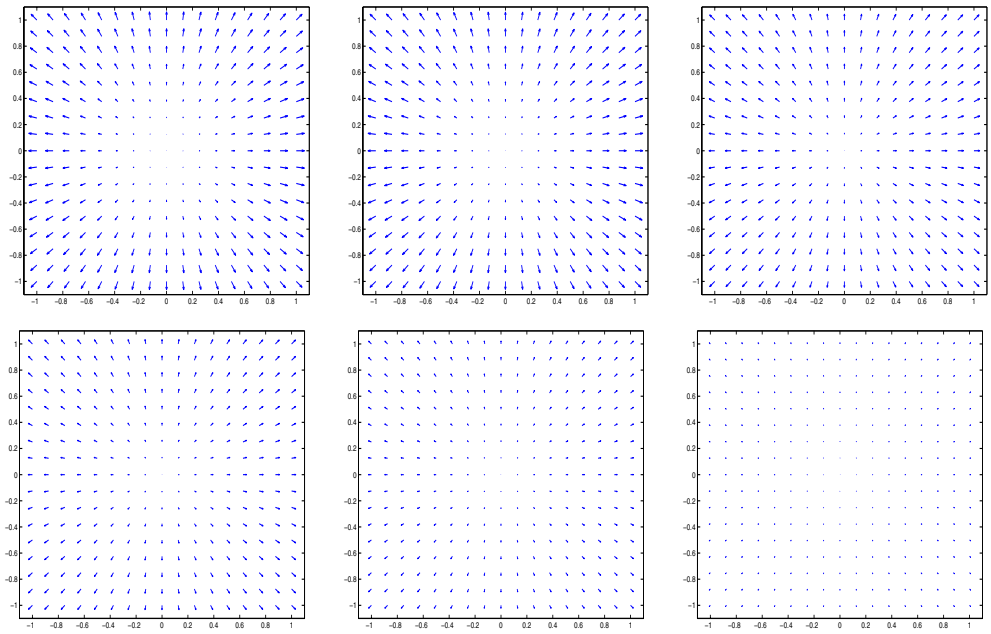


FIG. 5.1. $\mathbf{U}(t,\cdot)$ *in Example* 5.1(ai) *for* $t = 0$, 0.0102, 0.1625, 0.3301, 0.5078, 0.9090.

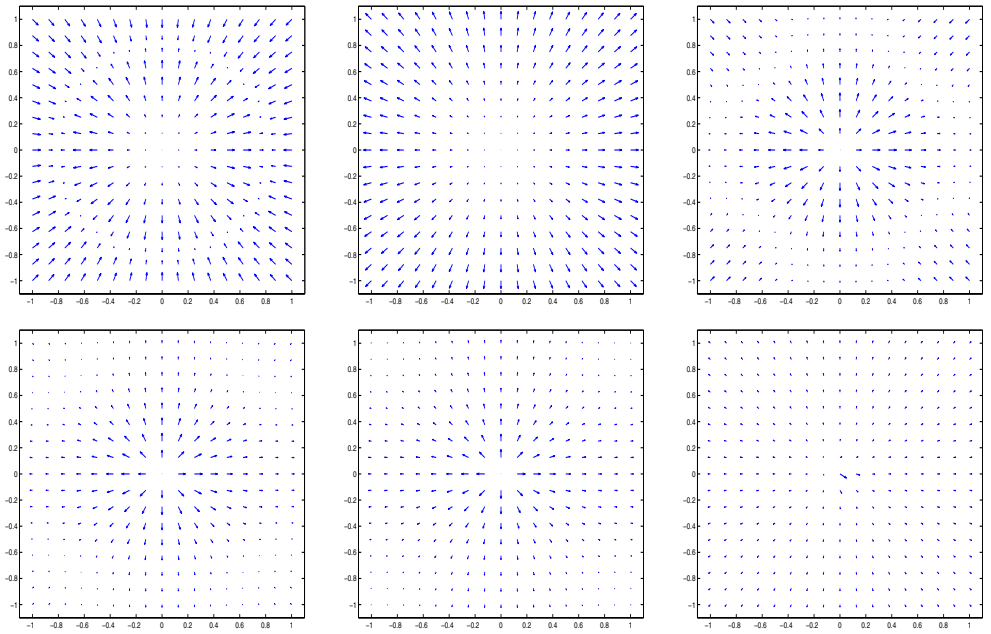


FIG. 5.2. $\mathbf{U}(t,\cdot)$ *in Example* 5.1(aii) *for* $t = 0$, 0.0102, 0.1016, 0.1828, 0.2539, 0.4467.

that $k = \frac{1}{2}h^2$ is not small enough for $p = 2$ and $n = 2$ in this experiment with exact integration, and reveal a stabilizing effect of numerical integration.

Analytical studies [3] of the scalar-valued total variation (TV) flow ($p = 1$) $-u_t \in \partial J(u)$, $u(0) = u_0 \in L^2(\Omega)$, for $J(u) = |Du|(\Omega)$ show interesting characterizations

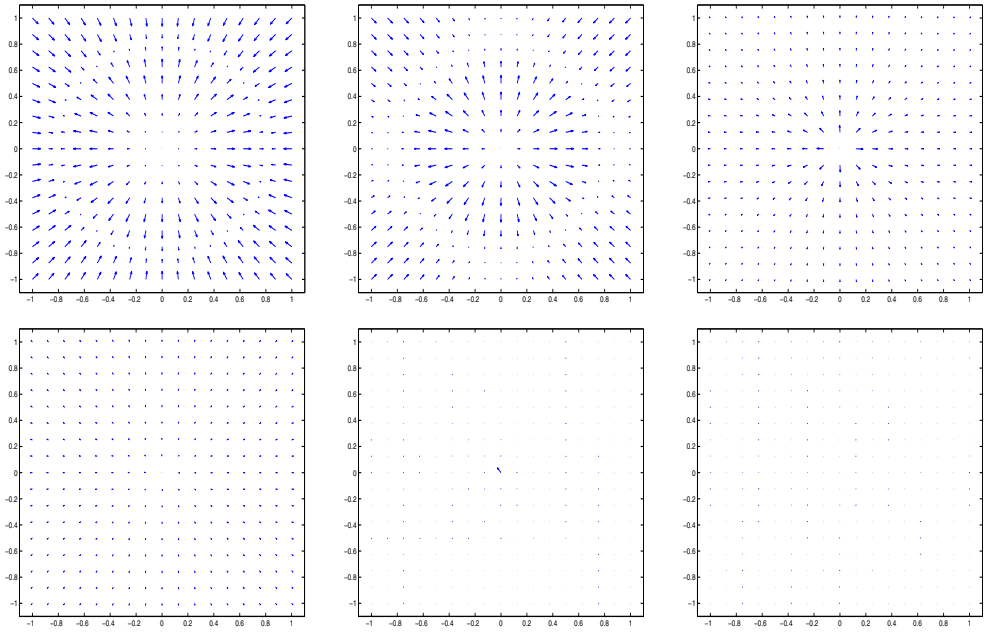


FIG. 5.3. $\mathbf{U}(t,\cdot)$ *in Example* 5.1(bii) *for* $t = 0$, 0.1051, 0.4054, 0.5105, 0.9259, 0.9910.

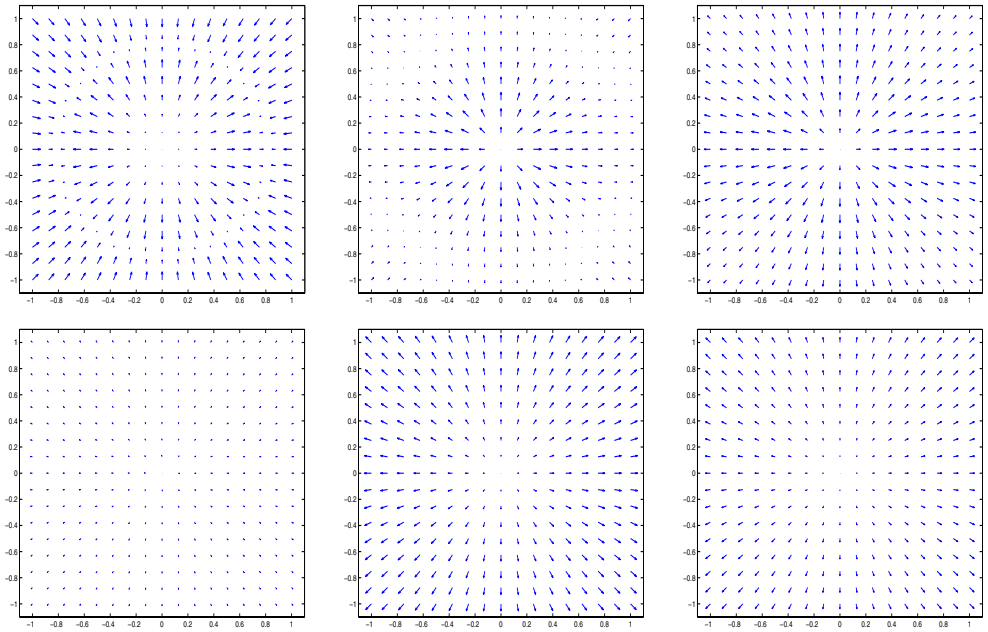


FIG. 5.4. $\mathbf{U}(t,\cdot)$ *in Example* 5.1(cii) *for* $t = 0$, 0.1001, 0.3053, 0.5105, 0.7157, 0.9209.

of the strong solution in the sense of semigroup theory: (i) finite extinction time ($n = 2$), (ii) $u(t,\cdot) \in L^\infty(\Omega)$, $t > 0$, if $u_0 \in L^n(\Omega)$, and no $L^1 - L^2$-regularizing effect for $L^1(\Omega)$-initial data, in general, (iii) $C^{1,\alpha}$-regularity of level sets $\partial^*\big[u(t) > \lambda\big]$ for $u_0 \in L^n(\Omega)$ of decreasing size, i.e., $\frac{d}{dt}\mathcal{H}^{n-1}\big(\partial^*\big[u(t) > \lambda\big]\big) \le 0$, and (iv) invariance

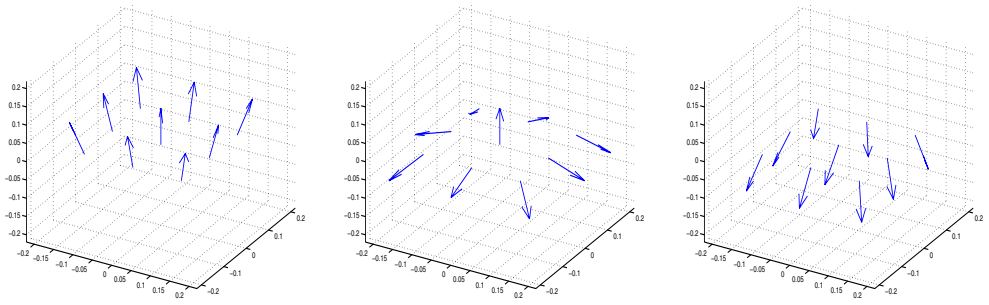


FIG. 5.5. *Nodal values* $\mathbf{U}(t, \mathbf{q}_i)$ *for nodes* $\mathbf{q}_i$ *close to the origin in Example* 5.1(aii) *for* $t = 0.0195, 0.2539, 0.3516$.

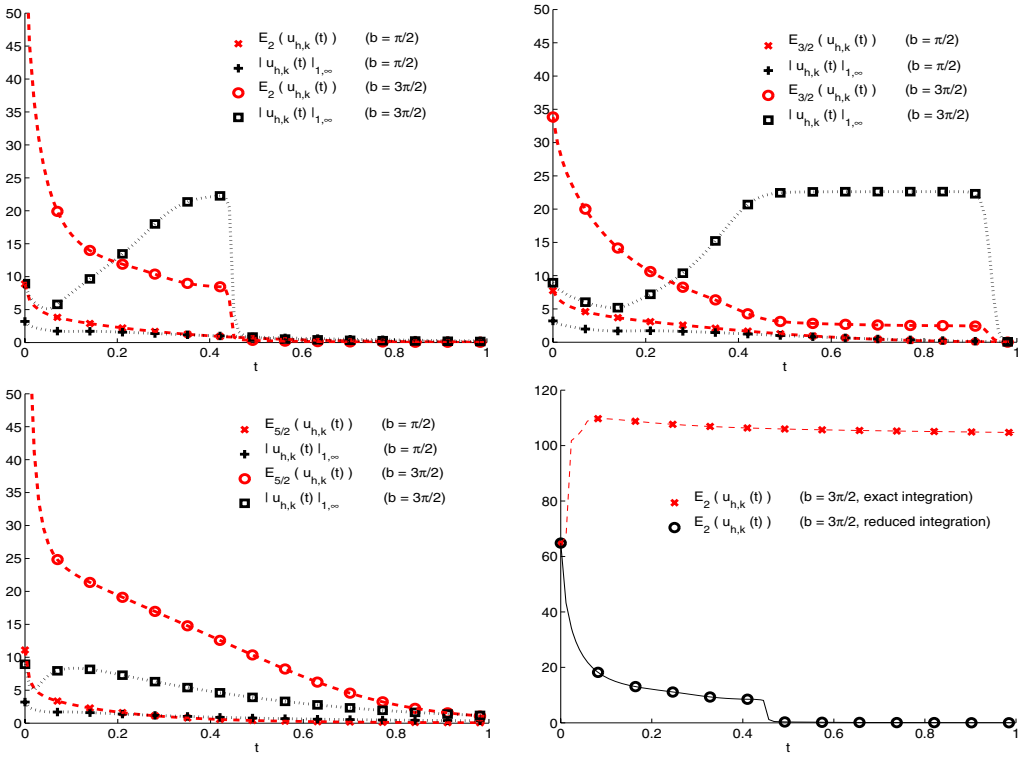


FIG. 5.6. *Energy decay and* $W^{1,\infty}$ *seminorm in Example* 5.1(a), (b), *and* (c) *and instability in Example* 5.1(ai) *for* $k = \frac{1}{2}h^2$.

of supports, provided, e.g., that the curvature of the smooth boundary of the simply connected convex staring support is not too large; cf. [21] for a convergence analysis of a regularized, fully discrete scheme and corresponding computational studies. We next discuss the latter issue in the present vectorial case.

EXAMPLE 5.2. *We define* $\mathbf{u}_0 : \overline{\Omega} \to S^2$ *by*

$$\mathbf{u}_0(\mathbf{x}) := \begin{cases} (1,0,0) & \text{for } |\mathbf{x}| < 0.5, \\ (0,1,0) & \text{for } |\mathbf{x}| \geq 0.5, \end{cases}$$

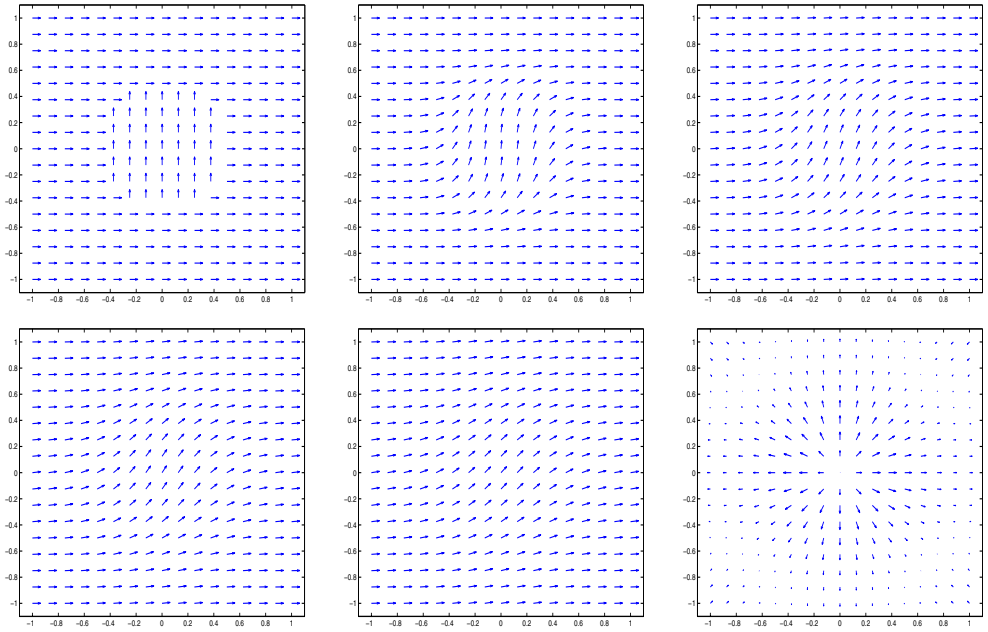


FIG. 5.7. $\mathbf{U}(t,\cdot)$ *in Example* 5.2(i) *for* $t = 0$, 0.02, 0.04, 0.06, 0.08, 0.10.

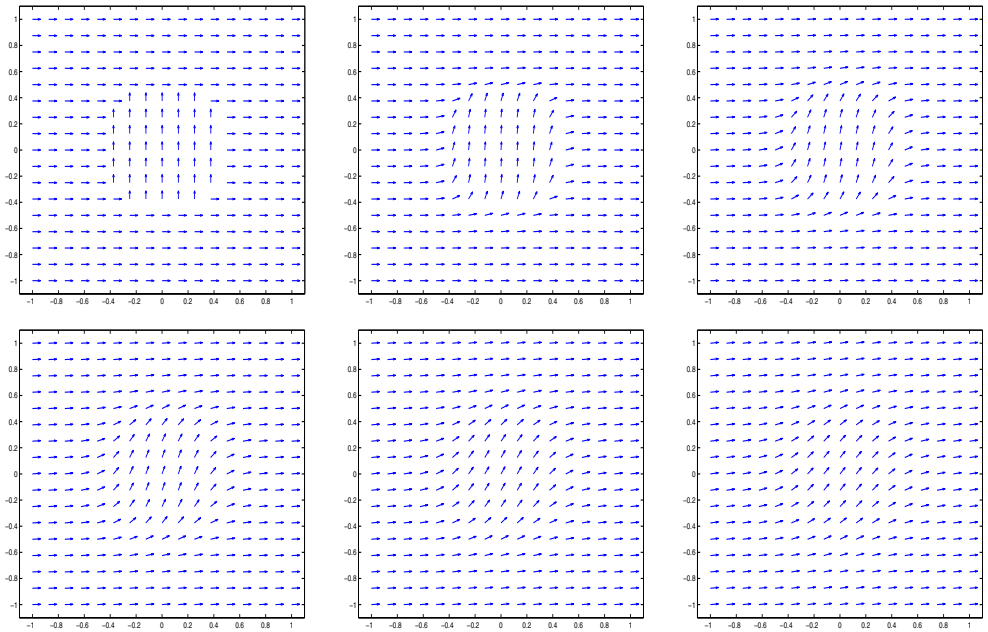


FIG. 5.8. $\mathbf{U}(t,\cdot)$ *in Example* 5.2(ii) *for* $t = 0$, 0.02, 0.04, 0.06, 0.08, 0.10.

and set (i) $p = 2$, (ii) $p = 3/2$, *and* (iii) $p = 5/4$.

Figures 5.7, 5.8, and 5.9 display snapshots of the numerical solutions in Example 5.2(i), (ii), and (iii), respectively. For $p = 2$ in (i) we observe that the solution is rather smooth for positive times and that at $t \approx 0.1$ an almost uniform (constant)

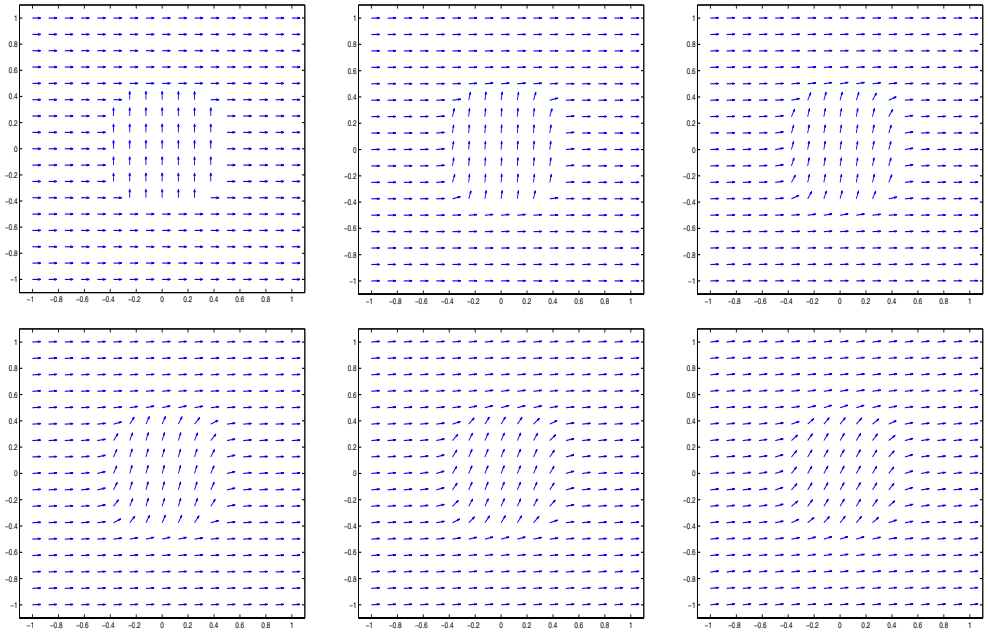


FIG. 5.9. $\mathbf{U}(t,\cdot)$ *in Example* 5.2(iii) *for* $t = 0,\ 0.02,\ 0.04,\ 0.06,\ 0.08,\ 0.10$.

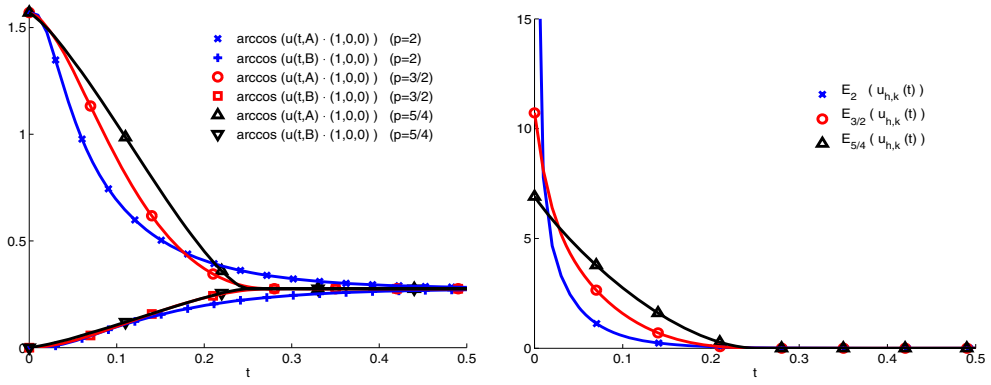


FIG. 5.10. *Angles and energy decay in Example* 5.2.

state is obtained. As opposed to the results in (i) for $p = 2$, the discontinuity along the circle $|\mathbf{x}| = 0.5$ is preserved for $p = 3/2$ in (ii) until $t \approx 0.04$. For $p = 5/4$ the discontinuity is preserved for a significantly longer time; cf. Figure 5.9. In the left plot of Figure 5.10 we display the angle between the vectors $\mathbf{U}(t,\mathbf{x})$ and $(1,0,0)$ for $t \in (0,1)$ and $\mathbf{x} \in \{\mathbf{A},\mathbf{B}\}$, where $\mathbf{A} = (0,0)$ and $\mathbf{B} = (3/4,3/4)$, and for $p = 2$, $p = 3/2$, and $p = 5/4$. We observe that the angle at the origin changes almost linearly in case $p = 3/2$. In the right plot of Figure 5.10 we display the energies $E_2(\mathbf{U}(t,\cdot))$, $E_{3/2}(\mathbf{U}(t,\cdot))$, and $E_{5/4}(\mathbf{U}(t,\cdot))$ as a function of t for the solutions in Example 5.2(i), (ii), and (iii), respectively. Of course, even though $\mathbf{u}_0$ is discontinuous, $\mathbf{U}^0 \equiv \mathcal{I}_h\mathbf{u}_0 \in W^{1,p}(\Omega,\mathbb{R}^3)$ with a mesh dependent norm, and so we still expect energy decay. We observe that this energy decay is slower for smaller exponents p.

Acknowledgment. J. W. Barrett, X. Feng, and A. Prohl would like to thank the Mathematisches Forschungsinstitut Oberwolfach for their kind hospitality and the opportunity of its "Research in Pairs" program in May, 2004.

REFERENCES

[1] F. Alouges, *A new algorithm for computing liquid crystal stable configurations: The harmonic mapping case*, SIAM J. Numer. Anal., 34 (1997), pp. 1708–1726.
[2] F. Alouges and P. Jaisson, *Convergence of a finite elements discretization for the Landau-Lifshitz equations*, Math. Models Methods Appl. Sci., 16 (2006), pp. 299–316.
[3] F. Andreu-Vaillo, V. Caselles, and J. M. Mazón, *Parabolic Quasilinear Equations Minimizing Linear Growth Functionals*, Birkhäuser-Verlag, Basel, Switzerland, 2004.
[4] J. W. Barrett and R. Nürnberg, *Finite element approximation of a Stefan problem with degenerate Joule heating*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 633–652.
[5] J. W. Barrett and W. B. Liu, *Quasi-norm error bounds for the finite element approximation of a non-Newtonian flow*, Numer. Math., 68 (1994), pp. 437–456.
[6] S. Bartels, *Stability and convergence of finite-element approximation schemes for harmonic maps*, SIAM J. Numer. Anal., 43 (2005), pp. 220–238.
[7] R. Becker, X. Feng, and A. Prohl, *Finite Element Approximations of the Ericksen-Leslie Model for Nematic Liquid Crystal Flow*, manuscript, 2007; available online from http://na.uni-tuebingen.de/preprints.shtml.
[8] M. Bertsch, R. Dal Passo, and A. Pisante, *Point singularities and nonuniqueness for the heat flow for harmonic maps*, Comm. Partial Differential Equations, 28 (2003), pp. 1135–1160.
[9] M. Bertsch, R. Dal Passo, and R. van der Hout, *Nonuniqueness for the heat flow of harmonic maps on the disk*, Arch. Ration. Mech. Anal., 161 (2002), pp. 93–112.
[10] F. Bethuel and X. Zheng, *Density of smooth functions between two manifolds in Sobolev spaces*, J. Funct. Anal., 80 (1988), pp. 60–75.
[11] P. V. Blomgren, *Total variation methods for restoration of vector valued images*, Ph.D. thesis, University of California, Los Angeles, Los Angeles, CA, 1998.
[12] W. F. Brown, *Micromagnetics*, Interscience, New York, 1963.
[13] K.-C. Chang, W.-Y. Ding, and R. Ye, *Finite-time blow-up of the heat flow of harmonic maps from surfaces*, J. Differential Geom., 36 (1992), pp. 507–515.
[14] Y. Chen, *The weak solutions to the evolution problems of harmonic maps*, Math. Z., 201 (1989), pp. 69–74.
[15] Y. Chen, M.-C. Hong, and N. Hungerbühler, *Heat flow of p-harmonic maps with values into spheres*, Math. Z., 215 (1994), pp. 25–35.
[16] Y. Chen and M. Struwe, *Existence and partial regularity results for the heat flow for harmonic maps*, Math. Z., 201 (1999), pp. 83–103.
[17] R. Cohen, S.-Y. Lin, and M. Luskin, *Relaxation and gradient methods for molecular orientation in liquid crystals*, Comput. Phys. Comm., 53 (1989), pp. 455–465.
[18] R. Cohen, R. Hardt, D. Kinderlehrer, S.-Y. Lin, and M. Luskin, *Minimum energy configurations for liquid crystals: Computational results*, in Theory and Applications of Liquid Crystals, IMA Vol. Math. Appl. 5, Springer, New York, 1987, pp. 99–121.
[19] P. Courilleau and F. Demengel, *Heat flow for p-harmonic maps with values in the circle*, Nonlinear. Anal., 41 (2000), pp. 689–700.
[20] A. Fardoun and R. Regbaoui, *Heat flow for p-harmonic maps with small initial data*, Calc. Var. Partial Differential Equations, 16 (2003), pp. 1–16.
[21] X. Feng, M. Von Oehsen, and A. Prohl, *Rate of convergence of regularization procedures and finite element approximations for the total variation flow*, Numer. Math., 100 (2005), pp. 441–456.
[22] Y. Giga, Y. Kashima, and N. Yamazaki, *Local solvability of a constrained gradient system of total variation*, Abstr. Appl. Anal., 8 (2004), pp. 651–682.
[23] R. Hardt, D. Kinderlehrer, and M. Luskin, *Remarks about the mathematical theory of liquid crystals*, in Calculus of Variations and Partial Differential Equations, Lecture Notes in Math. 1340, Springer, New York, 1988, pp. 123–138.
[24] N. Hungerbühler, *Heat flow into spheres for a class of energies*, in Progr. Nonlinear Differential Equations Appl. 59, Birkhäuser, Basel, Switzerland, 2004, pp. 45–65.
[25] N. Hungerbühler, *Non-uniqueness for the p-harmonic flow*, Canad. Math. Bull., 40 (1997), pp. 174–182.
[26] N. Hungerbühler, *p-harmonic Flow*, Diss. Math. Wiss. ETH Zürich, 10740, 1994.

[27] N. Hungerbühler, *Global weak solutions of the p-harmonic flow into homogeneous spaces*, Indiana Univ. Math. J., 45 (1996), pp. 275–288.
[28] M. Kružík and A. Prohl, *Recent developments in modeling, analysis, and numerics of ferromagnetism*, SIAM Rev., 48 (2006), pp. 439–483.
[29] S. Y. Lin and M. Luskin, *Relaxation methods for liquid crystal problems*, SIAM J. Numer. Anal., 26 (1989), pp. 1310–1324.
[30] M. Misawa, *Approximation of p-harmonic maps by the penalized equation*, Nonlinear Anal., 47 (2001), pp. 1069–1080.
[31] M. Misawa, *On the p-harmonic flow into spheres in the singular case*, Nonlinear Anal., 50 (2002), pp. 485–494.
[32] M. Struwe, *Geometric evolution problems*, IAS/Park City Math. Ser., 2 (1996), pp. 259–339.
[33] B. Tang, G. Sapiro, and V. Caselles, *Diffusion of generated data on non-flat manifolds via harmonic maps theory: The direction diffusion case*, Internat J. Comput. Vision, 36 (2000), pp. 149–161.
[34] B. Tang, G. Sapiro, and V. Caselles, *Color image enhancement via chromaticity diffusion*, IEEE Trans. Image Process., 10 (2001), pp. 701–707.
[35] P. Topping, *Reverse bubbling and nonuniqueness in the harmonic map flow*, Int. Math. Res. Notices, 10 (2002), pp. 505–520.
[36] L. A. Vese and S. J. Osher, *Numerical methods for p-harmonic flows and applications to image processing*, SIAM J. Numer. Anal., 40 (2002), pp. 2085–2104.
[37] E. Zeidler, *Nonlinear Functional Analysis and Its Applications,* II/B*: Nonlinear Monotone Operators*, Springer-Verlag, New York, 1990.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 928–945

TWO-SIDED A POSTERIORI ERROR ESTIMATES FOR MIXED FORMULATIONS OF ELLIPTIC PROBLEMS*

SERGEY REPIN†, STEFAN SAUTER‡, AND ANTON SMOLIANSKI‡

Abstract. The present work is devoted to the a posteriori error estimation for mixed approximations of linear self-adjoint elliptic problems. New guaranteed upper and lower bounds for the error measured in the natural product norm are derived, and individual sharp upper bounds are obtained for approximation errors in each of the physical variables. All estimates are reliable and valid for any approximate solution from the class of admissible functions. The estimates contain only global constants depending solely on the domain geometry and the given operators. Moreover, it is shown that, after an appropriate scaling of the coordinates and the equation, the ratio of the upper and lower bounds for the error in the product norm never exceeds 3. The possible methods of finding the approximate mixed solution in the class of admissible functions are discussed. The estimates are computationally very cheap and can also be used for the indication of the local error distribution. As applications, the diffusion problem as well as the problem of linear elasticity are considered.

Key words. a posteriori estimate, two-sided bounds, mixed approximation, elliptic problem

AMS subject classifications. 35J20, 65N15, 65N30

DOI. 10.1137/050641533

1. Introduction.

Most of the existing elliptic problems of continuum mechanics are originally derived in the mixed form; i.e., they contain *two* physical variables that are often equally important in the applications. For example, the stationary heat conduction (resp., diffusion) equation

$$-\Delta u + f = 0$$

comes from the energy (resp., mass) balance equation

$$-\operatorname{div} \mathbf{p} + f = 0 \tag{1}$$

and the empirical Fourier (resp., Fick) law for the heat (resp., mass) flux

$$\mathbf{p} = \nabla u. \tag{2}$$

Here we have set, for simplicity, the conduction (diffusion) coefficient equal to 1 and changed the sign in the flux relation (2). Both the temperature (molecular concentration) u and the flux $\mathbf{p}$ may be needed for understanding the real physical process, and this requirement becomes of utmost importance in the problems of the flows in porous media and in the elasticity problems, where the *complete* solution of the problem is the *pair* of the pressure and velocity for flows, or respectively the displacement and stress in elasticity.

These considerations served as a motivation for the extensive research in the field of so-called mixed methods, that is, the methods allowing one to obtain the

*Received by the editors September 29, 2005; accepted for publication (in revised form) October 31, 2006; published electronically May 4, 2007.
http://www.siam.org/journals/sinum/45-3/64153.html
†V.A. Steklov Institute of Mathematics, Laboratory of Mathematical Physics, Fontanka 27, 191 011 St. Petersburg, Russia (repin@pdmi.ras.ru).
‡Institute of Mathematics, Zurich University, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland (stas@math.unizh.ch, antsmol@amath.unizh.ch).

approximations to both physical variables of the problem. As general references on the subject, the books [9], [30], and [17] can be recommended. Although the computing methods are very important, reliable modeling also requires an explicit error control for the obtained approximations. This issue, namely, the a posteriori error estimation for the mixed formulations of elliptic problems, constitutes the primary goal of the present work.

There have been quite a few papers on the a posteriori error estimation for the mixed finite element methods (FEM). The residual-based estimates were developed in [2], [6], [12], [1], [16] for the diffusion-type equation and extended in [14] and [20] to equations of linear elasticity. The superconvergence-based (averaging-type) error estimators were proposed in [7] and [13] to control the L_2-error of the flux variable. Further, the estimators based on the solution of local problems were presented in [2], [16], and [20], and the hierarchical estimator can be found in [32]. Finally, a comparison of these four types of error estimators for mixed finite element discretizations by Raviart–Thomas elements (cf. [24]) was presented in [32]. We also refer to a recent paper [21], where further references on the subject can be found.

In this paper, we derive a posteriori error estimates of another type, the so-called *functional-type estimates* (see also [25], [26], [27], [28]). For the example of problem (1)–(2) equipped with the zero Dirichlet boundary condition for u and under the assumption that $f \in L_2(\Omega)$ (where Ω is the physical domain), the main estimates look as follows:

$$\begin{aligned} \|(u-v, \mathbf{p}-\mathbf{y})\|_{1\times\mathrm{div}} &\le \|\nabla v - \mathbf{y}\| + (1+2C_\Omega^2)^{1/2}\|\mathrm{div}\,\mathbf{y} - f\|, \\ \|(u-v, \mathbf{p}-\mathbf{y})\|_{1\times\mathrm{div}} &\ge \frac{1}{\sqrt{3}}\left(\|\nabla v - \mathbf{y}\| + \|\mathrm{div}\,\mathbf{y} - f\|\right). \end{aligned}$$

Here $\|\cdot\|$ is the norm in $L_2(\Omega)$, the *full* norm $\|(\cdot,\cdot)\|_{1\times\mathrm{div}}$ is defined as

$$\|(u-v, \mathbf{p}-\mathbf{y})\|_{1\times\mathrm{div}} := (\|\nabla(u-v)\|^2 + \|\mathbf{p}-\mathbf{y}\|^2 + \|\mathrm{div}(\mathbf{p}-\mathbf{y})\|^2)^{1/2},$$

and the pair $(v, \mathbf{y})$ from the product space $H_0^1(\Omega) \times H(\Omega;\mathrm{div})$ is *any* approximate solution to the mixed problem. The constant C_Ω is the global constant from the Friedrichs inequality and depends only on the domain geometry.

We see that, while these estimates provide *guaranteed* upper and lower bounds for the error of the mixed solution in the *full* norm, the estimates are also very flexible in the sense that they can be applied to a variety of different approximations, not being restricted to a particular discretization method. This fact makes the functional-type estimates especially attractive for the control of the modelling errors, like those arising in dimension reduction methods of continuum mechanics (see [29]). The sharpness of the estimates and the ability to indicate the local error distribution required for the mesh adaptation will also be shown. Last but not least, we remark that, once the approximate solution has been found in the product space, the estimates cost very little: their computation amounts to the calculation of the corresponding norms.

It is worth noting that, if the given data $f \in L_2(\Omega)$, the exact mixed solution $(u, \mathbf{p})$ belongs to the product space $H^1(\Omega) \times H(\Omega;\mathrm{div})$; thus, in this case, i.e., if the right-hand side f belongs to $L_2(\Omega)$, the latter product space seems to be a natural space for the approximation of the mixed solution. While the standard primal and dual mixed FEM approximate the mixed pair in $H^1(\Omega) \times L_2(\Omega;\mathbb{R}^n)$, resp., $L_2(\Omega) \times H(\Omega;\mathrm{div})$ (hence, not using the full regularity of the exact solution), there are alternative methods that allow us to construct the approximate solution directly

in $H^1(\Omega) \times H(\Omega;\mathrm{div})$. Some of these methods seem to be very promising and competitive, also in the case when one wants to find an approximation of the flux (stress) variable only. Although the comparative analysis of these methods is a subject of the next paper, we briefly review here four of them, since it is important for the application of our error estimates.

The rest of the paper is organized as follows. In section 2, we introduce the notation for the mixed formulation of a general linear self-adjoint elliptic problem. In section 3, the two-sided sharp a posteriori error estimate is derived for an arbitrary approximate solution from the natural class of admissible functions. Next, the individual a posteriori estimates for each of the two variables are derived and shown to be sharp as well. Section 4 is devoted to the applications of the developed theory. First, we consider the diffusion problem and obtain the explicit error bounds for its approximate mixed solution; then, we discuss possible methods of constructing the solution in the natural product space. Finally, the a posteriori error estimates are derived for both displacement and stress approximations in the problem of linear elasticity.

2. Preliminaries. Let V be a reflexive Banach space with the norm $\|\cdot\|_V$, Y a Hilbert space equipped with the inner product $(\cdot,\cdot)_Y$ and the norm $\|\cdot\|_Y$, and V_0 a linear subspace of V. By $\mathcal{B}$ we denote a linear bounded operator acting from V into Y, and by $\mathcal{B}^* : Y \to V_0^*$ the dual operator to $\mathcal{B}\big|_{V_0}$ (the restriction of $\mathcal{B}$ to V_0) in the sense that, for any $y \in Y$,

$$(y, \mathcal{B}w)_Y = \langle \mathcal{B}^* y, w\rangle \quad \forall w \in V_0.$$

Here $\langle w^*, w\rangle$ denotes the value of the functional $w^* \in V_0^*$ on the element $w \in V_0$.

Next, let us introduce a self-adjoint operator $\mathcal{A} \in \mathcal{L}(Y, Y)$ such that

$$\lambda_A \|y\|_Y^2 \le (\mathcal{A}y, y)_Y \le \Lambda_A \|y\|_Y^2 \quad \forall y \in Y, \tag{3}$$

where λ_A and Λ_A are positive constants independent of y. Such an operator defines the equivalent norm on Y:

$$|\!|\!| y |\!|\!| := (\mathcal{A}y, y)_Y^{1/2}.$$

The inverse operator $\mathcal{A}^{-1}$ satisfies an inequality of type (3) with constants Λ_A^{-1} and λ_A^{-1} and defines another equivalent norm on Y:

$$|\!|\!| y |\!|\!|_* := (\mathcal{A}^{-1} y, y)_Y^{1/2}.$$

Assume also that the operator $\mathcal{B}$ satisfies the coercivity inequality on V_0,

$$\|w\|_V \le C_{\mathcal{B}} \|\mathcal{B}w\|_Y \quad \forall w \in V_0, \tag{4}$$

where $C_{\mathcal{B}}$ is some positive constant independent of w. Using (3) and (4), one can define an equivalent norm $|\!|\!| \mathcal{B}\cdot |\!|\!|$ on V_0 as well as the following norm on the dual space V_0^*:

$$[\![w^*]\!] := \sup_{w \in V_0 \setminus \{0\}} \frac{\langle w^*, w\rangle}{|\!|\!| \mathcal{B}w |\!|\!|}.$$

Now let u_0 be some given function from V and

$$V_0 + u_0 := \{v \in V \mid v = w + u_0,\ w \in V_0\}.$$

Let, in addition, l be some given functional from V_0^*. Then, the problem

($\underline{\mathcal{P}}$): Find $u \in V_0 + u_0$ such that

$$(\mathcal{A}\mathcal{B}u, \mathcal{B}w)_Y + \langle l, w\rangle = 0 \quad \forall w \in V_0$$

has the unique solution (this follows from (3) and (4)).

The problem can be rewritten in the operator form as follows:

$$\mathcal{B}^*\mathcal{A}\mathcal{B}u + l = 0 \quad \text{in } V_0^*.$$

The mixed formulation of the problem can be immediately obtained by the introduction of the new unknown function

$$p = \mathcal{A}\mathcal{B}u,$$

which leads to the problem

($\underline{\mathcal{M}}$): Find $(u, p) \in (V_0 + u_0) \times Y$ such that

$$\begin{aligned} p &= \mathcal{A}\mathcal{B}u \quad \text{in } Y,\\ \mathcal{B}^*p + l &= 0 \quad \text{in } V_0^*. \end{aligned}$$

It is clear that problem $(\mathcal{M})$ has the well-known saddle-point structure; its unique solvability is a direct consequence of conditions (3) and (4) (see, e.g., [9]).

In what follows, we will adopt the terminology used in the duality theory of convex analysis (see, e.g., [15]) and call the solution of problem $(\mathcal{P})$ the *primal variable* (*primal solution*) and the new unknown p the *dual variable* (*dual solution*). Accordingly, the letters u, v, w will be reserved for the functions related to the primal variable, i.e., belonging to the space V, and the letters p, q, y for those related to the dual variable, i.e., in the space Y.

In view of the second equation of problem $(\mathcal{M})$, the dual variable p belongs to the set

$$Q_l := \{q \in Y \mid \mathcal{B}^*q = -l \text{ in } V_0^*\}.$$

Thus, for the full control of the dual variable one needs an extended norm on Y, which we define as

$$\|y\|_{\mathcal{B}^*} := \left(\| y \|_*^2 + [\![\mathcal{B}^*y]\!]^2\right)^{1/2} \quad \forall y \in Y. \tag{5}$$

Although this is an equivalent norm on Y (since $[\![\mathcal{B}^*y]\!] = \sup_{w\in V_0} \frac{\langle \mathcal{B}^*y, w\rangle}{\| \mathcal{B}w \|} = \sup_{w\in V_0} \frac{(y, \mathcal{B}w)_Y}{\| \mathcal{B}w \|} \le \sqrt{\frac{\Lambda_A}{\lambda_A}}\, \| y \|_*$), we need it to explicitly control the error in the "equilibrium equation," i.e., in the second equation of $(\mathcal{M})$.

Finally, we define the *full norm* on the product space $V_0 \times Y$:

$$\|(w, y)\|_{V_0\times Y} := \left(\| \mathcal{B}w \|^2 + \|y\|_{\mathcal{B}^*}^2\right)^{1/2} \quad \forall (w, y) \in V_0 \times Y. \tag{6}$$

3. General estimates.

3.1. Estimate in the full norm. Let $(v, q) \in (V_0 + u_0) \times Q_l$ be an arbitrary approximation to the exact solution (u, p) of problem $(\mathcal{M})$. Then, with the help of the relation $p = \mathcal{A}\mathcal{B}u$ and the fact that $\mathcal{A}$ is a self-adjoint linear operator, it is easy to show (see also [26]) that

$$\begin{aligned} \| \mathcal{B}(u - v) \|^2 + \| p - q \|_*^2 &= (\mathcal{A}\mathcal{B}(u - v), \mathcal{B}(u - v))_Y + (\mathcal{A}^{-1}(p - q), p - q)_Y\\ &= (\mathcal{A}\mathcal{B}v - q, \mathcal{B}v - \mathcal{A}^{-1}q)_Y + 2(p - q, \mathcal{B}(u - v))_Y. \end{aligned}$$

Since both p and q belong to the set Q_l, $\mathcal{B}^*(p-q) = -l + l = 0$ in V_0^* and $(p-q, \mathcal{B}(u-v))_Y = \langle \mathcal{B}^*(p-q), u-v\rangle = 0$; this implies

$$\| \mathcal{B}(u-v) \|^2 + \| p-q \|_*^2 = (\mathcal{A}\mathcal{B}v - q, \mathcal{B}v - \mathcal{A}^{-1}q)_Y = \| \mathcal{A}\mathcal{B}v - q \|_*^2. \tag{7}$$

This equality can be referred to as the generalized Prager–Synge hypercircle identity (see [23]).

Relation (7) may already be viewed as an a posteriori error estimate, since the right-hand side does not depend on the exact solution (u,p). However, the estimate holds only for $q \in Q_l$, which seriously restricts the field of its practical application. In fact, the constraint of the set Q_l is virtually impossible to satisfy exactly (this would be nearly equivalent to finding the exact dual solution p), and that is why it is desirable to obtain an estimate allowing the approximate dual solution to be in some large *unconstrained* space.

If we waive the constraint $\mathcal{B}^*q = -l$ in V_0^* for the approximate dual variable, the latter remains to be considered in the whole space Y. Let $y \in Y$ be an arbitrary approximation to p, and $v \in V_0 + u_0$ as before. Then, using (7), one can derive

$$\begin{aligned} &\| \mathcal{B}(u-v) \|^2 + \| p-y \|_*^2 \\ &\qquad = \| \mathcal{B}(u-v) \|^2 + \| p-q \|_*^2 + 2(\mathcal{A}^{-1}(p-q), q-y)_Y + \| q-y \|_*^2 \\ &\qquad = \| \mathcal{A}\mathcal{B}v - q \|_*^2 + 2(\mathcal{A}^{-1}(p-q), q-y)_Y + \| q-y \|_*^2 \quad \forall q \in Q_l. \end{aligned} \tag{8}$$

Now we would like to estimate the right-hand side of (8) from above, so as to eliminate $q \in Q_l$. It is clear that, in order to obtain an explicitly computable and efficient upper bound, one has to carefully choose some special q in Q_l.

First, define the auxiliary function $w_y \in V_0$ such that

$$\mathcal{B}^*\mathcal{A}\mathcal{B}w_y = l + \mathcal{B}^*y \quad \text{in } V_0^*.$$

Due to assumptions (3) and (4), this problem has a unique solution.

Now set $q := y - \mathcal{A}\mathcal{B}w_y$. It is evident that such a function q belongs to Y and $\mathcal{B}^*q = \mathcal{B}^*y - \mathcal{B}^*\mathcal{A}\mathcal{B}w_y = \mathcal{B}^*y - l - \mathcal{B}^*y = -l$ in V_0^*, that is, $q \in Q_l$. It may be noticed that, with this specific choice of q, the sum $q + \mathcal{A}\mathcal{B}w_y$ obviously becomes a nonorthogonal variant of the Helmholtz decomposition for the function $y \in Y$.

Now we can plug the constructed q into the right-hand side of (8). For the first term we have

$$\| \mathcal{A}\mathcal{B}v - q \|_* \le \| \mathcal{A}\mathcal{B}v - y \|_* + \| \mathcal{A}\mathcal{B}w_y \|_*. \tag{9}$$

Here $\| \mathcal{A}\mathcal{B}w_y \|_* = \big(\mathcal{A}^{-1}\mathcal{A}\mathcal{B}w_y, \mathcal{A}\mathcal{B}w_y\big)_Y^{1/2} = \| \mathcal{B}w_y \|$. The latter norm can be estimated by

$$\| \mathcal{B}w_y \|^2 = \langle \mathcal{B}^*\mathcal{A}\mathcal{B}w_y, w_y\rangle \le [\![\mathcal{B}^*\mathcal{A}\mathcal{B}w_y]\!] \, \| \mathcal{B}w_y \|,$$

which implies $\| \mathcal{B}w_y \| \le [\![\mathcal{B}^*\mathcal{A}\mathcal{B}w_y]\!]$. We notice now that, by the definition of w_y, $\mathcal{B}^*\mathcal{A}\mathcal{B}w_y = l + \mathcal{B}^*y$ in V_0^*, which ultimately leads to the estimate

$$\| \mathcal{A}\mathcal{B}w_y \|_* \le [\![l + \mathcal{B}^*y]\!]. \tag{10}$$

Inserting this into (9), one obtains

$$\| \mathcal{A}\mathcal{B}v - q \|_* \le \| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^*y]\!]. \tag{11}$$

The second term on the right-hand side of (8) can be rewritten as

$$2(\mathcal{A}^{-1}(p-q),q-y)_Y = 2(\mathcal{A}^{-1}(p-q),-\mathcal{A}\mathcal{B}w_y)_Y = -2(p-q,\mathcal{B}w_y)_Y \\ = -2\langle \mathcal{B}^*(p-q),w_y\rangle = 0, \tag{12}$$

since $\mathcal{B}^*(p-q) = -l + l = 0$ in V_0^*.

The third term on the right-hand side of (8) equals $\| \mathcal{A}\mathcal{B}w_y \|_*^2$, which has been estimated from above by $[\![l + \mathcal{B}^*y]\!]^2$ (see (10)). Hence, combining this result with (11) and (12), we obtain from (8)

$$\| \mathcal{B}(u-v) \|^2 + \| p-y \|_*^2 \le (\| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^*y]\!])^2 + [\![l + \mathcal{B}^*y]\!]^2, \tag{13}$$

where v is an arbitrary function from $V_0 + u_0$ and y is any function from Y.

From (13) one immediately derives the estimate

$$\left(\| \mathcal{B}(u-v) \|^2 + \| p-y \|_*^2\right)^{1/2} \le \| \mathcal{A}\mathcal{B}v - y \|_* + \sqrt{2}\, [\![l + \mathcal{B}^*y]\!] \tag{14}$$

and the following theorem.

THEOREM 3.1. *Let $(u,p) \in (V_0+u_0)\times Y$ be the solution of problem $(\mathcal{M})$, and let $v \in V_0+u_0$ and $y \in Y$ be arbitrary approximations to u and p.*

Then, the following estimates hold true:

$$\|(u-v,p-y)\|_{V_0\times Y} \le \| \mathcal{A}\mathcal{B}v - y \|_* + \sqrt{3}\, [\![l + \mathcal{B}^*y]\!], \tag{15}$$

$$\|(u-v,p-y)\|_{V_0\times Y} \ge \frac{1}{\sqrt{3}} \left(\| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^*y]\!]\right). \tag{16}$$

Proof. The upper bound (15) immediately follows from estimate (13) and the definition of the full norm $\|(\cdot,\cdot)\|_{V_0\times Y}$, since $[\![\mathcal{B}^*(p-y)]\!] = [\![l + \mathcal{B}^*y]\!]$.

To obtain the lower bound (16) we use first the triangle inequality to derive

$$\| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^*y]\!] \le \| \mathcal{A}\mathcal{B}v - \mathcal{A}\mathcal{B}u \|_* + \| p-y \|_* + [\![l + \mathcal{B}^*y]\!] \\ = \| \mathcal{B}(u-v) \| + \| p-y \|_* + [\![\mathcal{B}^*(p-y)]\!],$$

and then the inequality $a+b+c \le \sqrt{3}\sqrt{a^2+b^2+c^2}\ \ \forall a,b,c \ge 0$ to obtain the estimate

$$\| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^*y]\!] \le \sqrt{3}\, \|(u-v,p-y)\|_{V_0\times Y}.$$

This implies the lower bound (16). ∎

Let

$$M_\oplus := \| \mathcal{A}\mathcal{B}v - y \|_* + \sqrt{3}\, [\![l + \mathcal{B}^*y]\!] \tag{17}$$

denote the upper bound (15) for the error in the full norm.

Remarks. 1. If $y \to p$ in Y and $v \to u$ in V, the estimates (15) and (16) tend to zero, precisely as the exact error in the full norm $\|(u-v,p-y)\|_{V_0\times Y}$ does.

2. The error majorant $M_\oplus$ is *sharp*. Indeed, if one takes $y = p$ (i.e., $l + \mathcal{B}^*y \equiv 0$ in V_0^*), estimate (15) becomes

$$\| \mathcal{B}(u-v) \| \le \| \mathcal{A}\mathcal{B}v - p \|_* = \| \mathcal{B}(u-v) \|,$$

which shows that the constant "1" in front of the first term of $M_\oplus$ cannot be improved in general. On the other hand, if, in the case $u_0 = 0$, we set $v = 0$ and $y = 0$, then estimate (15) takes the form

$$\left(2 \| \mathcal{B}u \|^2 + [\![l]\!]^2\right)^{1/2} \le \sqrt{3}[\![l]\!],$$

which is a sharp estimate since $\| \mathcal{B}u \| \le [\![l]\!]$ (set $w = u$ in problem $(\mathcal{P})$), and this estimate, evidently, cannot be improved. Thus, the factor "$\sqrt{3}$" multiplying the second term in $M_\oplus$ cannot be taken smaller in a general case.

The sharpness of the lower bound (16) in a general case is obvious from the estimate's derivation.

3. The efficiency of the estimator $M_\oplus$ can be easily evaluated using the lower bound (16). Namely, for the effectivity index of $M_\oplus$ we have

$$\mathrm{i}_{\mathrm{eff}} := \frac{M_\oplus}{\|(u-v, p-y)\|_{V_0 \times Y}} \le \sqrt{3} \frac{\| \mathcal{A}\mathcal{B}v - y \|_* + \sqrt{3}\, [\![l + \mathcal{B}^* y]\!]}{\| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^* y]\!]} \le 3. \tag{18}$$

Estimate (18) provides a rough upper bound for the effectivity index that in most of the cases will be strictly less than 3. Indeed, if the second term $[\![l + \mathcal{B}^* y]\!]$ is essentially smaller than the first one, then $\mathrm{i}_{\mathrm{eff}}$ is close to $\sqrt{3}$. On the other hand, since (15) is a guaranteed upper bound of the error, we always have $\mathrm{i}_{\mathrm{eff}} \ge 1$.

The two-sided estimate (15)–(16) is important, because it provides a control over the error in the full norm, i.e., with respect to both primal and dual variables. However, the individual errors in primal and dual variables may also be of interest; in the next two sections we derive sharp upper bounds for the corresponding norms of these errors. It is worth noticing that the individual estimates which immediately follow from (15) are not sharp and hence may lead to a certain overestimation.

3.2. Estimate for error in the primal variable.

THEOREM 3.2. *Let $u \in V_0 + u_0$ be the solution of problem $(\mathcal{P})$, and $v \in V_0 + u_0$ an arbitrary approximate solution to $(\mathcal{P})$. Then*

$$\| \mathcal{B}(u-v) \| \le \| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^* y]\!] \quad \forall y \in Y. \tag{19}$$

Proof. It immediately follows from (7) that $\| \mathcal{B}(u-v) \|^2 = \inf_{q \in Q_l} \| \mathcal{A}\mathcal{B}v - q \|_*^2$; i.e.,

$$\| \mathcal{B}(u-v) \| \le \| \mathcal{A}\mathcal{B}v - q \|_* \quad \forall q \in Q_l. \tag{20}$$

The right-hand side of (20) has been already estimated for the proof of Theorem 3.1, where the function $q \in Q_l$ was constructed such that $q = y - \mathcal{A}\mathcal{B}w_y$ with y being any function from Y and w_y being the solution to the problem $\mathcal{B}^* \mathcal{A}\mathcal{B} w_y = l + \mathcal{B}^* y$ in V_0^*. Then, estimate (19) follows directly from (11). ∎

Remarks. 1. Estimate (19) is *sharp*. Indeed, if we set $y = p = \mathcal{A}\mathcal{B}u$, the estimate will be

$$\| \mathcal{B}(u-v) \| \le \| \mathcal{A}\mathcal{B}v - \mathcal{A}\mathcal{B}u \|_* = \| \mathcal{B}(u-v) \|.$$

On the other hand, in the case $u_0 = 0$, setting $v = 0$ and $y = 0$, we obtain from (19)

$$\| \mathcal{B}u \| \le [\![l]\!],$$

which is the sharp energy estimate for the solution u of problem $(\mathcal{P})$. Thus, the weights equal to 1 on the right-hand side of estimate (19) are optimal, in a general case.

2. Estimate (19) is *asymptotically exact* in the sense that, if $y \to p$ in Y, then the upper bound (19) tends to the norm $\| \mathcal{A}\mathcal{B}v - \mathcal{A}\mathcal{B}u \|_* = \| \mathcal{B}(u-v) \|$ of the error in primal variable.

3. The estimate remains *efficient* if y is close to p in Y, since

$$\begin{aligned}\| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^* y]\!] &\le \| \mathcal{A}\mathcal{B}v - \mathcal{A}\mathcal{B}u \|_* + \| p - y \|_* + [\![\mathcal{B}^*(p-y)]\!] \\ &\le \| \mathcal{B}(u-v) \| + \sqrt{2}\|p-y\|_{\mathcal{B}^*}.\end{aligned}$$

Here the last, presumably small, term measures the level of the overestimation due to estimate (19).

4. If one considers only $y \in Q_l$ in (19), one arrives at estimate (20), which is the "constitutive relation-based" estimate (see [19]). On the other hand, if one takes $y = \mathcal{A}\mathcal{B}v$ in (19), one obtains the estimate $\| \mathcal{B}(u-v) \| \le [\![l + \mathcal{B}^* \mathcal{A}\mathcal{B}v]\!]$, which is the "residual-based" estimate for problem $(\mathcal{P})$ (see [4]). Thus, estimate (19) includes these two estimates as particular cases, combining their advantages and providing a greater flexibility. More on the links between the error majorant and other estimates can be found in [25].

3.3. Estimate for error in the dual variable.

THEOREM 3.3. *Let $(u,p) \in (V_0 + u_0) \times Y$ be the solution to problem $(\mathcal{M})$, and let $y \in Y$ be any approximation of p. Then*

$$\| p - y \|_* \le \| \mathcal{A}\mathcal{B}v - y \|_* + [\![l + \mathcal{B}^* y]\!] \quad \forall v \in V_0 + u_0, \tag{21}$$

$$\|p - y\|_{\mathcal{B}^*} \le \| \mathcal{A}\mathcal{B}v - y \|_* + \sqrt{2}[\![l + \mathcal{B}^* y]\!] \quad \forall v \in V_0 + u_0. \tag{22}$$

Proof. We have for any $v \in V_0 + u_0$

$$\begin{aligned}\| \mathcal{B}(u-v) \|^2 + \| p - y \|_*^2 &= (\mathcal{A}\mathcal{B}(u-v), \mathcal{B}(u-v))_Y + (\mathcal{A}^{-1}(p-y), p-y)_Y \\ &= (\mathcal{A}\mathcal{B}v - y, \mathcal{B}v - \mathcal{A}^{-1}y)_Y + 2(p-y, \mathcal{B}(u-v))_Y,\end{aligned} \tag{23}$$

where the self-adjointness of $\mathcal{A}$ and $\mathcal{A}^{-1}$ as well as the relation $p = \mathcal{A}\mathcal{B}u$ have been used. For the second term on the right-hand side of (23) we have

$$(p-y, \mathcal{B}(u-v))_Y = \langle \mathcal{B}^*(p-y), u-v \rangle = \langle -l - \mathcal{B}^* y, u - v \rangle \quad \forall v \in V_0 + u_0,$$

which implies the estimate

$$| (p-y, \mathcal{B}(u-v))_Y | \le [\![l + \mathcal{B}^* y]\!] \| \mathcal{B}(u-v) \| \le \frac{1}{2} \left([\![l + \mathcal{B}^* y]\!]^2 + \| \mathcal{B}(u-v) \|^2 \right).$$

Using this estimate and noticing that the first term on the right-hand side of (23) equals $\| \mathcal{A}\mathcal{B}v - y \|_*^2$, we derive from (23)

$$\| \mathcal{B}(u-v) \|^2 + \| p - y \|_*^2 \le \| \mathcal{A}\mathcal{B}v - y \|_*^2 + [\![l + \mathcal{B}^* y]\!]^2 + \| \mathcal{B}(u-v) \|^2,$$

that is,

$$\| p - y \|_*^2 \le \| \mathcal{A}\mathcal{B}v - y \|_*^2 + [\![l + \mathcal{B}^* y]\!]^2 \quad \forall v \in V_0 + u_0.$$

This immediately yields estimate (21). Then, (22) is obvious. □

Remark. Estimate (21) is *sharp* (hence, estimate (22) is sharp too). Indeed, if our approximation y belongs to Q_l and we set $v = u$, we obtain from (21)

$$\| p - y \|_* \le \| \mathcal{A}\mathcal{B}u - y \|_* = \| p - y \|_*.$$

On the other hand, if, in the case $u_0 = 0$, we set $v = 0$ and $y = 0$, we have from (21)

$$\| p \|_* \le [\![l]\!], \quad \text{i.e.,} \quad \| \mathcal{B}u \| \le [\![l]\!],$$

which is the sharp energy estimate for the solution of problem $(\mathcal{P})$. Thus, the weights of both terms on the right-hand side of (21) are, in general, optimal.

3.4. Important special case. The estimates obtained above provide reliable measures of the errors in a very general situation when the exact solution to problem $(\mathcal{P})$ is sought in an arbitrary reflexive Banach space V and the given functional l belongs to V_0^*. As a result, the norm in V_0^* enters the estimates, making them less convenient for computational purposes. However, in most practically interesting cases one can significantly simplify the estimates. Indeed, usually one has

$$\text{the } \textit{continuous embedding } V \subset U \tag{24}$$

for some Hilbert space U with the inner product $(\cdot,\cdot)_U$ and the norm $\|\cdot\|_U$. This means that $\|\cdot\|_U \le C\|\cdot\|_V$ with some constant C; however, in what follows we mostly deal with cases like $U = L_2(\Omega)$, $V = H^1(\Omega)$ and hence make a stronger assumption,

$$\|\cdot\|_U \le \|\cdot\|_V. \tag{25}$$

One may notice that $U \subset V_0^*$. (It immediately follows from (24).) We also assume the given data

$$l \in U. \tag{26}$$

First, one can notice that, if assumption (26) holds true, the exact dual solution p satisfies the equation $\mathcal{B}^*p + l = 0$ in U and hence belongs to the space

$$Y_{\mathcal{B}^*} := \{y \in Y \mid \mathcal{B}^*y \in U\},$$

which is the Banach space with respect to the norm

$$|\!|\!| y |\!|\!|_{\mathcal{B}^*} := \big(\|\!|\!| y |\!|\!|_*^2 + \|\mathcal{B}^*y\|_U^2\big)^{1/2} \quad \forall y \in Y_{\mathcal{B}^*}. \tag{27}$$

As compared to the definition of the norm $\|\cdot\|_{\mathcal{B}^*}$ (see (5)), the newly defined norm is stronger, which reflects the fact that $Y_{\mathcal{B}^*}$ is a subspace of Y.

It is natural now to consider the approximation y of the exact dual solution in $Y_{\mathcal{B}^*}$ rather than in Y; this is still much less restrictive than an approximation in the set Q_l, whose definition contains the complicated constraint $\mathcal{B}^*y = -l$.

Then, we can estimate the term $[\![\, l + \mathcal{B}^*y \,]\!]$ as follows:

$$\begin{aligned}
[\![\, l + \mathcal{B}^*y \,]\!] &= \sup_{w\in V_0\setminus\{0\}} \frac{\langle l + \mathcal{B}^*y, w\rangle}{|\!|\!| \mathcal{B}w |\!|\!|} = \sup_{w\in V_0\setminus\{0\}} \frac{(l + \mathcal{B}^*y, w)_U}{|\!|\!| \mathcal{B}w |\!|\!|} \\
&\le \sup_{w\in V_0\setminus\{0\}} \frac{\|l + \mathcal{B}^*y\|_U \, \|w\|_U}{|\!|\!| \mathcal{B}w |\!|\!|} \le \sup_{w\in V_0\setminus\{0\}} \frac{\|l + \mathcal{B}^*y\|_U \, \|w\|_V}{|\!|\!| \mathcal{B}w |\!|\!|} \\
&\le \frac{C_{\mathcal{B}}}{\lambda_A^{1/2}} \, \|l + \mathcal{B}^*y\|_U \quad \forall y \in Y_{\mathcal{B}^*},
\end{aligned} \tag{28}$$

where inequalities (3), (4), and (25) have been used. It is important to notice that one often has the inequality

$$\|w\|_U \le \widetilde{C}_{\mathcal{B}} \, \|\mathcal{B}w\|_Y \quad \forall w \in V_0 \tag{29}$$

in addition to (4); in such a case, the constant $C_{\mathcal{B}}$ in (28) is to be replaced by $\widetilde{C}_{\mathcal{B}}$ from (29).

With the definition of the $Y_{\mathcal{B}^*}$-norm (see (27)), the *full norm* (6) should be understood on the product space $V_0 \times Y_{\mathcal{B}^*}$ in the following sense:

$$\|(v,y)\|_{V_0\times Y_{\mathcal{B}^*}} := \left(\| \mathcal{B}v \|^2 + \| y \|_*^2 + \|\mathcal{B}^* y\|_U^2\right)^{1/2} \quad \forall (v,y) \in V_0 \times Y_{\mathcal{B}^*}. \tag{30}$$

THEOREM 3.4. *Let V be continuously embedded into some Hilbert space U and $l \in U$. Suppose in addition that $\|\cdot\|_U \le \|\cdot\|_V$. Let $(u,p) \in (V_0+u_0)\times Y_{\mathcal{B}^*}$ be the solution to problem $(\mathcal{M})$, and let $(v,y) \in (V_0+u_0)\times Y_{\mathcal{B}^*}$ be any approximate solution to $(\mathcal{M})$.*

Then, the following a posteriori error estimates hold true:

$$\|(u-v,p-y)\|_{V_0\times Y_{\mathcal{B}^*}} \le \| \mathcal{A}\mathcal{B}v - y \|_* + \left(1 + 2\frac{C_{\mathcal{B}}^2}{\lambda_A}\right)^{1/2} \|l + \mathcal{B}^* y\|_U, \tag{31}$$

$$\|(u-v,p-y)\|_{V_0\times Y_{\mathcal{B}^*}} \ge \frac{1}{\sqrt{3}} \left(\| \mathcal{A}\mathcal{B}v - y \|_* + \|l + \mathcal{B}^* y\|_U\right), \tag{32}$$

$$\| \mathcal{B}(u-v) \| \le \| \mathcal{A}\mathcal{B}v - y \|_* + \frac{C_{\mathcal{B}}}{\lambda_A^{1/2}} \|l + \mathcal{B}^* y\|_U, \tag{33}$$

$$\| p - y \|_* \le \| \mathcal{A}\mathcal{B}v - y \|_* + \frac{C_{\mathcal{B}}}{\lambda_A^{1/2}} \|l + \mathcal{B}^* y\|_U, \tag{34}$$

$$\|p-y\|_{\mathcal{B}^*} \le \| \mathcal{A}\mathcal{B}v - y \|_* + \left(1 + \frac{C_{\mathcal{B}}^2}{\lambda_A}\right)^{1/2} \|l + \mathcal{B}^* y\|_U. \tag{35}$$

Proof. The upper bound (31) immediately follows from estimates (13) and (28); the lower bound (32) is a simple consequence of the triangle inequality, like the lower bound (16) in Theorem 3.1.

Estimate (33) can easily be derived from (19) and (28), while estimates (34) and (35) follow from (21) and (28). ∎

Remarks. 1. Estimates (31)–(35) are *sharp*, which follows from the sharpness of the estimates of Theorems 3.1–3.3 and of inequality (28).

2. Estimates (31) and (32) imply that the effectivity index of the error majorant (31) is always between 1 and $\sqrt{3}\left(1+2\frac{C_{\mathcal{B}}^2}{\lambda_A}\right)^{1/2}$. It is worth noting that the constant λ_A can be made equal to 1 if one performs the corresponding rescaling of the operator $\mathcal{A}$ and of the functional l (i.e., multiplication of the linear problem $(\mathcal{P})$ by $1/\lambda_A$). The constant $C_{\mathcal{B}}$ depends only on the operator $\mathcal{B}$ and can easily be evaluated a priori. (We discuss this issue in the next section.) We will also show that, after an appropriate scaling of the geometric coordinates, one can make the constant $C_{\mathcal{B}} \le 1$, which means that the effectivity index of the upper bound (31) for the new "rescaled" problem will always be between 1 and 3.

3. It is worthwhile to notice a remarkable symmetry of estimates (33) and (34) for the primal and dual variables.

4. Applications.

4.1. Diffusion problem.

4.1.1. Error estimates. Let $V = H^1(\Omega)$, where Ω is a bounded domain in $\mathbb{R}^n$ with Lipschitz boundary $\partial\Omega$, $V_0 = H_0^1(\Omega)$, $Y = L_2(\Omega;\mathbb{R}^n)$. Consider the case

$$\mathcal{B} = \nabla := \left(\frac{\partial}{\partial x_1}, \dots, \frac{\partial}{\partial x_n}\right).$$

Then, $\mathcal{B}^*\mathbf{y} = -\operatorname{div}\mathbf{y} \in H^{-1}(\Omega) = V_0^*$ for any $\mathbf{y} \in Y$, and

$$\langle \mathcal{B}^*\mathbf{y}, w\rangle = \int_\Omega \mathbf{y}\cdot\nabla w\,dx \quad \forall w \in V_0,$$

where the dot denotes the scalar product of vectors in $\mathbb{R}^n$. The operator $\mathcal{A}$ is defined by a symmetric uniformly positive definite matrix $\mathbf{A} = \{a_{ij}(x)\}_{i,j=\overline{1,n}}$ with coefficients from $L_\infty(\Omega)$. Then, the norms $\|\!|\cdot|\!\|$ and $\|\!|\cdot|\!\|_*$ are defined as

$$\|\!|\mathbf{y}|\!\|^2 = \int_\Omega \mathbf{A}\mathbf{y}\cdot\mathbf{y}\,dx, \qquad \|\!|\mathbf{y}|\!\|_*^2 = \int_\Omega \mathbf{A}^{-1}\mathbf{y}\cdot\mathbf{y}\,dx.$$

Inequality (3) is obviously satisfied, and (4) follows from the Friedrichs inequality.

Assume now that u_0 is some given function from $H^1(\Omega)$ and that l is some given functional from $H^{-1}(\Omega)$. Then, problem $(\mathcal{P})$ defines the weak solution of the boundary-value problem

$$-\operatorname{div}(\mathbf{A}\nabla u) + l = 0 \quad \text{in } \Omega, \tag{36}$$

$$u = u_0 \quad \text{on } \partial\Omega. \tag{37}$$

We can also write all the estimates of Theorems 3.1–3.3, where $[\![\cdot]\!]$ is equivalent to the $H^{-1}(\Omega)$-norm.

We see, however, that V is continuously embedded into the Hilbert space $U = L_2(\Omega)$ and that inequality (25) is valid; hence, if we suppose that the data $l \in L_2(\Omega)$, we can use the results of Theorem 3.4.

First, we note that the space $Y_{\mathcal{B}^*}$ is, in fact, the space $H(\Omega;\operatorname{div}) := \{\mathbf{y} \in L_2(\Omega;\mathbb{R}^n) \mid \operatorname{div}\mathbf{y} \in L_2(\Omega)\}$ with the norm

$$\|\mathbf{y}\|_{\operatorname{div}} := \left(\|\!|\mathbf{y}|\!\|_*^2 + \|\operatorname{div}\mathbf{y}\|^2\right)^{1/2},$$

where $\|\cdot\|$ denotes the norm in $L_2(\Omega)$.

The *full norm* then takes the form

$$\|(v,\mathbf{y})\|_{1\times\operatorname{div}} := \left(\|\!|\nabla v|\!\|^2 + \|\!|\mathbf{y}|\!\|_*^2 + \|\operatorname{div}\mathbf{y}\|^2\right)^{1/2} \quad \forall (v,\mathbf{y}) \in H_0^1(\Omega)\times H(\Omega;\operatorname{div}).$$

It is important to notice that, in the considered case, we have an inequality of type (29) that is exactly the Friedrichs inequality

$$\|w\| \le C_\Omega \|\!|\nabla w|\!\| \quad \forall w \in H_0^1(\Omega).$$

Thus, the constant $C_{\mathcal{B}}$ in (28) and in Theorem 3.4 is, in fact, the constant C_Ω from the Friedrichs inequality.

Hence, if $(u,\mathbf{p}) \in (H_0^1(\Omega)+u_0)\times H(\Omega;\operatorname{div})$ is the exact solution to the mixed problem

$$\mathbf{p} = \mathbf{A}\nabla u \quad \text{in } \Omega, \tag{38}$$

$$-\operatorname{div}\mathbf{p} + l = 0 \quad \text{in } \Omega, \tag{39}$$

and $(v,\mathbf{y}) \in (H_0^1(\Omega)+u_0)\times H(\Omega;\operatorname{div})$ is any approximate solution to the problem,

then the following a posteriori error estimates follow directly from Theorem 3.4:

$$\|(u-v,\mathbf{p}-\mathbf{y})\|_{1\times\mathrm{div}} \le \| \mathbf{A}\nabla v-\mathbf{y} \|_{*} + \left(1+2\frac{C_\Omega^2}{\lambda_A}\right)^{1/2} \|\mathrm{div}\,\mathbf{y}-l\|, \tag{40}$$

$$\|(u-v,\mathbf{p}-\mathbf{y})\|_{1\times\mathrm{div}} \ge \frac{1}{\sqrt{3}} \left(\| \mathbf{A}\nabla v-\mathbf{y} \|_{*} + \|\mathrm{div}\,\mathbf{y}-l\|\right), \tag{41}$$

$$\| \nabla(u-v) \| \le \| \mathbf{A}\nabla v-\mathbf{y} \|_{*} + \frac{C_\Omega}{\lambda_A^{1/2}}\|\mathrm{div}\,\mathbf{y}-l\|, \tag{42}$$

$$\| \mathbf{p}-\mathbf{y} \|_{*} \le \| \mathbf{A}\nabla v-\mathbf{y} \|_{*} + \frac{C_\Omega}{\lambda_A^{1/2}}\|\mathrm{div}\,\mathbf{y}-l\|, \tag{43}$$

$$\|\mathbf{p}-\mathbf{y}\|_{\mathrm{div}} \le \| \mathbf{A}\nabla v-\mathbf{y} \|_{*} + \left(1+\frac{C_\Omega^2}{\lambda_A}\right)^{1/2} \|\mathrm{div}\,\mathbf{y}-l\|. \tag{44}$$

Estimates (40)–(44) provide sharp error bounds that are explicitly computable, if one has the approximate solution to (38)–(39) in the product space $H^1(\Omega)\times H(\Omega;\mathrm{div})$. It is, of course, clear that, having found the approximate mixed solution $(v,\mathbf{y})$ by primal or dual mixed FEM, one can use some local averaging (projection) to recover the needed $H^1(\Omega)$ (respectively, $H(\Omega;\mathrm{div})$) regularity for the approximate primal (respectively, dual) variable. There exist, however, several methods allowing one to approximate the mixed solution $(u,\mathbf{p})$ in the space $H^1(\Omega)\times H(\Omega;\mathrm{div})$ directly. Below, we briefly review four of them.

4.1.2. Approximation of the mixed solution in $H^1(\Omega)\times H(\Omega;\mathrm{div})$.

(a) *Least-squares mixed method.* This method was analyzed in [22] and, under the name first-order-system least-squares (FOSLS), in [10], [11] (see also the references therein). In this method, the saddle-point (min-max) problem (38)–(39) is reformulated as a quadratic minimization (min-min) problem

$$\inf_{v\in H_0^1(\Omega)+u_0}\ \inf_{\mathbf{y}\in H(\Omega;\mathrm{div})} \left(\| \mathbf{A}\nabla v-\mathbf{y} \|_{*}^2 + \|\mathrm{div}\,\mathbf{y}-l\|^2\right), \tag{45}$$

which leads to the solution of the "stabilized" saddle-point problem, given next.

Find $(u,\mathbf{p})\in (H_0^1(\Omega)+u_0)\times H(\Omega;\mathrm{div})$ such that

$$\begin{aligned}\int_\Omega \mathbf{A}^{-1}\mathbf{p}\cdot\mathbf{q}\,dx + \int_\Omega (\mathrm{div}\,\mathbf{p})(\mathrm{div}\,\mathbf{q})\,dx - \int_\Omega \nabla u\cdot\mathbf{q}\,dx \\ = \int_\Omega l\,(\mathrm{div}\,\mathbf{q})\,dx \quad \forall \mathbf{q}\in H(\Omega;\mathrm{div}),\end{aligned} \tag{46}$$

$$\int_\Omega (\mathrm{div}\,\mathbf{p})\,v\,dx + \int_\Omega \mathbf{A}\nabla u\cdot\nabla v\,dx = 0 \quad \forall v\in H_0^1(\Omega). \tag{47}$$

We have to note that, in the original version of the method, the squared L_2-norm was used in the first term of the functional (45) instead of the squared $\| \cdot \|_*$-norm, which somehow changes the system of the functional's optimality conditions (46)–(47).

System (46)–(47), unlike (38)–(39), leads to a symmetric *positive definite* discrete problem, and the discrete *inf-sup* condition is always satisfied owing to the least-squares stabilization. The latter fact allows one to choose the approximation spaces for u and $\mathbf{p}$ independently of each other.

However, in (46)–(47) the primal and the dual variables are *strongly coupled*. The following method yields only a weak coupling of the variables.

(b) *Method of minimizing the squared majorant.* From estimate (42) for the error in the primal variable one can easily derive the estimate for the squared energy norm of the error:

$$\| \nabla(u-v) \|^2 \le (1+\beta) \| \mathbf{A}\nabla v - \mathbf{y} \|_*^2 + \left(1+\frac{1}{\beta}\right) \frac{C_\Omega^2}{\lambda_A} \|\operatorname{div} \mathbf{y} - l\|^2, \tag{48}$$

where $\beta > 0$ is an arbitrary number and $\mathbf{y}$ is any function from $H(\Omega;\operatorname{div})$. Denote the right-hand side of (48) by $M^2(v;\mathbf{y},\beta)$ ("the squared error majorant"). It is evident that $M^2(v;\mathbf{y},\beta)$ is, in fact, the least-squares functional (45) with differently weighted terms. However, instead of minimizing the functional with respect to both v and $\mathbf{y}$ simultaneously as in the least-squares mixed method, the following simple algorithm was proposed in [27]:

1. Find the approximate solution $v \in V_0 + u_0$ to the problem (36)–(37).
2. Set $\beta = 1$, and find $\mathbf{y}$ by minimizing $M^2(v;\mathbf{y},\beta)$ with respect to $\mathbf{y}$.

The algorithm was initially motivated by the goal of finding a best possible upper bound for the energy error in the primal variable; however, it also provides a computationally efficient way of computing approximate primal and dual solutions in $H^1(\Omega) \times H(\Omega;\operatorname{div})$ in a *weakly coupled* manner. Indeed, the problems for v and $\mathbf{y}$ now have to be solved successively.

While the problem of finding an approximate solution to (36)–(37) in step 1 is quite standard, the computation of $\mathbf{y}$ in step 2 also does not present serious difficulty. Since $M^2(v;\mathbf{y},\beta)$ is a quadratic functional with respect to the dual variable $\mathbf{y}$ for any fixed v and β, the minimization of the functional on any finite-dimensional subspace Y_h of $H(\Omega;\operatorname{div})$ leads to the solution of a linear system with symmetric positive definite matrix.

This algorithm was independently proposed in [8] as an alternative to the least-squares mixed method and considered as a single Picard–Uzawa-type iteration for the solution of the coupled system (46)–(47). (In [8], the least-squares functional (45) was used, not $M^2(v;\mathbf{y},\beta)$.) It has been shown in [8] that, with v found by a conforming FEM for the problem (36)–(37), the minimizer $\mathbf{y}_h$ of the functional on the subspace Y_h has the optimal order of the $H(\Omega;\operatorname{div})$-error with respect to the mesh size (i.e., the order of the interpolation error for Y_h), provided that the $H^1(\Omega)$-error of the approximation v is not of lower order. The advantage over the dual mixed FEM as well as the least-squares mixed method is obvious: the computation of the primal variable is completely independent of the calculation of the dual one (this reduces the total computational cost), and the discrete problem for each of the variables is moderately sized, symmetric, and positive definite (i.e., one does not have to deal with an indefinite saddle-point problem as in the case of the dual mixed FEM).

As follows from the numerical studies of [27], using the parameter β, one can gain a further improvement in the approximation of the dual solution. Namely, for the unique minimizer $\mathbf{y}_\beta \in H(\Omega;\operatorname{div})$ of $M^2(v;\mathbf{y},\beta)$ for any fixed $v \in V_0 + u_0$ and $\beta > 0$, it was proved in [27] that $\mathbf{y}_\beta$ converges to the exact dual solution p in $H(\Omega;\operatorname{div})$ as $\beta \to 0$, and, moreover,

$$\begin{aligned} \| \mathbf{p} - \mathbf{y} \|_* &\le C\,\beta^{1/2}, \\ \|\operatorname{div}(\mathbf{p} - \mathbf{y})\| &\le C\,\beta, \end{aligned}$$

with some constant C independent of $\mathbf{y}$ and β. Thus, the one-stroke minimization of the functional $M^2(v;\mathbf{y},\beta)$ with respect to $\mathbf{y}$ and with some moderately small β may yield even better accuracy of the dual-solution approximation than the minimization

with $\beta = 1$. Numerical experiments (see [27]) show that, for example, if one uses linear finite elements for both primal and dual variables, the value $\beta = 1/10$ is a good choice. Taking β moderately small allows one to circumvent the difficulties with the condition number of the resulting discrete system and with the locking phenomenon, typical for penalty methods.

A possible way of finding the concrete value of β is to minimize the functional $M^2(v; \mathbf{y}, \beta)$ with respect to β having fixed v and $\mathbf{y}$. This immediately implies the explicit formula

$$\beta = \frac{C_\Omega \|\operatorname{div} \mathbf{y} - l\|}{\lambda_A^{1/2} \, \| \mathbf{A}\nabla v - \mathbf{y} \|_*}, \tag{49}$$

and the modified algorithm for the approximation of the primal and dual solutions reads as follows:

1. Find the approximate solution $v \in V_0 + u_0$ to the problem (36)–(37).
2. Set $\beta^{(1)} = 1$ and find $\mathbf{y}^{(1)}$ by minimizing $M^2(v; \mathbf{y}, \beta^{(1)})$ with respect to $\mathbf{y}$.
3. Compute $\beta^{(2)}$ using $\mathbf{y}^{(1)}$ in formula (49); find $\mathbf{y}^{(2)}$ by minimizing $M^2(v; \mathbf{y}, \beta^{(2)})$ with respect to $\mathbf{y}$.

A further iteration of the process of minimizing the squared majorant with respect to $\mathbf{y}$ and β does not bring any essential benefits, as shown in the detailed study of [27].

To summarize, the minimization of the squared majorant either at a one-stroke (only steps 1 and 2 in the algorithm above) or by two iterations provides a competitive approach to the approximation of the dual solution. The whole method of finding the primal variable in $H^1(\Omega)$ and the dual variable in $H(\Omega; \operatorname{div})$ amounts to the successive solution of two elliptic problems. It is worth noting that the approximation spaces for v and $\mathbf{y}$ can be chosen independently of each other, as in the least-squares mixed method.

(c) *Dual penalty method.* This method can be viewed as a limiting case of the previous method, i.e., the case when the parameter β in the squared majorant is considered as a very small penalty parameter. The classical dual penalty method has, however, a slightly different formulation. Namely, after finding $v \in V_0 + u_0$ as an approximate solution to (36)–(37), one has to minimize the quadratic "penalized functional"

$$I(\mathbf{y}) := \| \mathbf{y} \|_*^2 + \frac{1}{\varepsilon} \|\operatorname{div} \mathbf{y} - l\|^2$$

over $H(\Omega; \operatorname{div})$ for some small $\varepsilon > 0$. The main difference with the method of minimizing the squared majorant is that the approximation of $\mathbf{y}$ is now *fully decoupled* from the approximation of v. As immediate drawbacks, one has the deterioration of the condition number of the resulting discrete problem and a possible locking phenomenon.

(d) *Method of local projections.* In this method, the dual variable is found by some local projections of the approximate flux $\mathbf{A}\nabla v$ into the space $H(\Omega; \operatorname{div})$. The approximate flux is derived from the approximate primal solution $v \in V_0 + u_0$ previously found by solving (36)–(37). Thus, we have here again a *weakly coupled* approach. The method is usually referred to as the "gradient recovery" or "gradient averaging," and its diverse variants have been considered by many researchers (see, e.g., [18], [34], [35], [33] and the references therein). In particular, the so-called "equilibrium-enhanced" gradient recovery methods (see [5], [31]) seem to be especially advantageous for computing an accurate approximation to the dual variable in the $H(\Omega; \operatorname{div})$-norm.

Remarks. 1. It is clear that each of the four methods addressed above has both advantages and drawbacks. A thorough comparison of the methods still remains to be done.

2. If the approximation to $(u, \mathbf{p})$ has been found in $(H^1_0(\Omega) + u_0) \times H(\Omega; \mathrm{div})$ by one of the above considered methods, it can be inserted into estimates (40)–(44) to yield the explicit a posteriori control of the errors in both variables. Since the norms in the estimates can be computed by summation of the local contributions from subdomains of Ω (given some finite subdivision of Ω), they may be used also for an *adaptive* improvement of the approximation. In particular, it is obvious that, if $\mathbf{y}$ is close to $\mathbf{p}$ in $H(\Omega; \mathrm{div})$, the term $\| \mathbf{A}\nabla v - \mathbf{y} \|_*$ computed over any subdomain $\omega \subset \Omega$ is close to $\| \nabla(u - v) \|$ considered on ω. More on the use of the error majorants for the indication of the local error distribution can be found in [27], [28].

3. The constant C_Ω stemming from Friedrichs' inequality is equal to $1/\sqrt{\lambda_\Omega}$, where λ_Ω is the minimal eigenvalue of the Laplace operator equipped with the homogeneous Dirichlet boundary condition on $\partial\Omega$. It is, however, clear that C_Ω can always be estimated from above by $C_\mathcal{D}$, where $\mathcal{D} \supset \Omega$ is some domain of a simple shape (e.g., a rectangle in two dimensions). Then, $C_\mathcal{D}$ can be computed analytically.

4. Since the evaluation of C_Ω is fairly easy, the total computational cost of estimates (40)–(44) is very small (only the computation of norms), provided that the pair $(v, \mathbf{y})$ is found in $(H^1_0(\Omega) + u_0) \times H(\Omega; \mathrm{div})$, for instance, by one of the methods discussed above.

5. Using translation and rescaling of the geometric coordinates (which amounts to a linear coordinate transformation), one can make it so that the rescaled physical domain $\widetilde{\Omega}$ would be completely inside of a unit cube (square in two dimensions). After having rewritten the original elliptic problem in the new coordinates and subsequently rescaling the equation so that $\lambda_A = 1$ (see Remark 2 at the end of section 3.4), we can write down all the estimates (40)–(44) for the approximation error of the solution to the rescaled problem on the new domain $\widetilde{\Omega}$. The most important fact here is that all the estimates for the new problem will contain only numerical constants (like $\sqrt{3}$, $\sqrt{2}$), since $\lambda_A = 1$ and the Friedrichs constant $C_{\widetilde{\Omega}}$ may be estimated from above by 1 (see Remark 3 above). As an immediate consequence, one infers that the effectivity index of the upper bound (40) for the error in the full norm will be between 1 and 3.

4.2. Linear elasticity. Although the application of the theory to the problem of linear elasticity is similar to the case of the diffusion problem, it is, however, interesting to consider the elasticity problem in detail.

Let $V = H^1(\Omega; \mathbb{R}^n)$, where Ω is a bounded domain in $\mathbb{R}^n$ with Lipschitz boundary $\partial\Omega$, $V_0 = \{\mathbf{v} \in V \mid \mathbf{v} = 0 \text{ on } \partial\Omega\}$, and $\mathbf{Y} = L_2(\Omega; \mathbb{M}^{n\times n}_s)$, where $\mathbb{M}^{n\times n}_s$ is the space of symmetric $n \times n$-matrices. Now define the operator $\mathcal{B}$ as follows:

$$\mathcal{B}\mathbf{v} := \underline{\underline{\mathbf{e}}}(\mathbf{v}) = \frac{1}{2}\left(\nabla\mathbf{v} + (\nabla\mathbf{v})^T\right).$$

Here $\nabla\mathbf{v} = \{v_{i,j}\}$ is a tensor (the gradient of the vector $\mathbf{v}$), and the symbol ^{T} means the transposition. Then, $\mathcal{B}^*\underline{\underline{\mathbf{y}}} = -\mathrm{div}\,\underline{\underline{\mathbf{y}}} \in H^{-1}(\Omega; \mathbb{R}^n) = V_0^*$ for any $\underline{\underline{\mathbf{y}}} \in \mathbf{Y}$, and

$$\langle \mathcal{B}^*\underline{\underline{\mathbf{y}}}, \mathbf{w} \rangle = \int_\Omega \underline{\underline{\mathbf{y}}} : \underline{\underline{\mathbf{e}}}(\mathbf{w})\, dx \quad \forall \mathbf{w} \in V_0,$$

where the colon denotes the inner product in $\mathbb{M}^{n\times n}_s$ ($a : b = \sum a_{ij}b_{ij}$ $\forall a, b \in \mathbb{M}^{n\times n}_s$).

The operator $\mathcal{A}$ is defined by the so-called tensor of elastic moduli $\mathbb{L} = \{\mathbb{L}_{ijkl}\}$, which satisfies the double inequality

$$\lambda_{\mathbb{L}} \,|\, \underline{\underline{\mathbf{e}}} \,|^2 \,\le\, \mathbb{L}\underline{\underline{\mathbf{e}}} \colon \underline{\underline{\mathbf{e}}} \,\le\, \Lambda_{\mathbb{L}} \,|\, \underline{\underline{\mathbf{e}}} \,|^2 \quad \forall \underline{\underline{\mathbf{e}}} \in \mathbb{M}^{n\times n}_s \tag{50}$$

and the symmetry and boundedness conditions

$$\mathbb{L}_{ijkl} = \mathbb{L}_{jikl} = \mathbb{L}_{klij}, \quad \mathbb{L}_{ijkl} \in L_\infty(\Omega). \tag{51}$$

Then, the norms $\|\!|\!| \cdot |\!|\!|$ and $\|\!|\!| \cdot |\!|\!|_*$ are defined as

$$\|\!|\!| \,\underline{\underline{\mathbf{y}}} \,|\!|\!|^2 = \int_\Omega \mathbb{L}\underline{\underline{\mathbf{y}}} \colon \underline{\underline{\mathbf{y}}}\, dx, \quad \|\!|\!| \,\underline{\underline{\mathbf{y}}} \,|\!|\!|^2_* = \int_\Omega \mathbb{L}^{-1}\underline{\underline{\mathbf{y}}} \colon \underline{\underline{\mathbf{y}}}\, dx \quad \forall \underline{\underline{\mathbf{y}}} \in \mathbf{Y}.$$

Inequality (3) is obviously satisfied, and (4) follows from the Korn inequality.

Assume now that $\mathbf{u}_0$ is some given function from $H^1(\Omega;\mathbb{R}^n)$ and that $\mathbf{f}$ is some given functional from $H^{-1}(\Omega;\mathbb{R}^n)$. Then, problem $(\mathcal{P})$ can be formulated as follows.

Find $\mathbf{u} \in V_0 + \mathbf{u}_0$ such that

$$\int_\Omega \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{u}) \colon \underline{\underline{\mathbf{e}}}(\mathbf{w})\, dx + \langle \mathbf{f}, \mathbf{w}\rangle = 0 \quad \forall \mathbf{w} \in V_0, \tag{52}$$

where $\langle \mathbf{f}, \mathbf{w}\rangle = \int_\Omega \mathbf{f}\cdot\mathbf{w}\, dx$. The corresponding mixed formulation of (52) defines the weak solution of the boundary-value problem of linear elasticity:

$$\underline{\underline{\mathbf{p}}} = \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{u}) \quad \text{in } \Omega, \tag{53}$$

$$\operatorname{div} \underline{\underline{\mathbf{p}}} = \mathbf{f} \quad \text{in } \Omega, \tag{54}$$

$$\mathbf{u} = \mathbf{u}_0 \quad \text{on } \partial\Omega. \tag{55}$$

We see that V is continuously embedded into the Hilbert space $U = L_2(\Omega;\mathbb{R}^n)$ and that inequality (25) holds true; hence, if we suppose the given body force $\mathbf{f} \in L_2(\Omega;\mathbb{R}^n)$, we can use the results of Theorem 3.4.

First, we note that the space $Y_{\mathcal{B}^*}$ is, in fact, the space $\mathbf{H}(\Omega;\operatorname{div}) := \{\underline{\underline{\mathbf{y}}} \in \mathbf{Y} \mid \operatorname{div} \underline{\underline{\mathbf{y}}} \in L_2(\Omega;\mathbb{R}^n)\}$ with the norm

$$\|\underline{\underline{\mathbf{y}}}\|_{\operatorname{div}} := \left(\|\!|\!| \,\underline{\underline{\mathbf{y}}} \,|\!|\!|^2_* + \|\operatorname{div} \underline{\underline{\mathbf{y}}}\|^2\right)^{1/2},$$

where $\|\cdot\|$ denotes the norm in $L_2(\Omega;\mathbb{R}^n)$.

The *full norm* then takes the form

$$\|(\mathbf{v}, \underline{\underline{\mathbf{y}}})\|_{1\times\operatorname{div}} := \left(\|\!|\!| \,\underline{\underline{\mathbf{e}}}(\mathbf{v}) \,|\!|\!|^2 + \|\!|\!| \,\underline{\underline{\mathbf{y}}} \,|\!|\!|^2_* + \|\operatorname{div} \underline{\underline{\mathbf{y}}}\|^2\right)^{1/2} \quad \forall (\mathbf{v}, \underline{\underline{\mathbf{y}}}) \in V_0 \times \mathbf{H}(\Omega;\operatorname{div}).$$

It is important to notice that, in the considered case, we have the inequality of type (29) that is a vector variant of the Friedrichs inequality, namely,

$$\|\mathbf{w}\| \,\le\, C_\Omega \,\|\underline{\underline{\mathbf{e}}}(\mathbf{w})\| \quad \forall \mathbf{w} \in V_0. \tag{56}$$

Thus, the constant $C_{\mathcal{B}}$ in (28) and in Theorem 3.4 is, in fact, the constant C_Ω from (56).

Remark. The constant C_Ω from (56) equals $1/\sqrt{\lambda_\Omega}$, where λ_Ω is the minimal eigenvalue of the vector-valued elliptic operator $\mathcal{L} : V_0 \to H^{-1}(\Omega;\mathbb{R}^n)$, $\mathcal{L}\mathbf{w} =$

$-\frac{1}{2}\left(\operatorname{div}(\nabla \mathbf{w}) + \nabla(\operatorname{div} \mathbf{w})\right)$ for any $\mathbf{w} \in V_0$, equipped with the zero Dirichlet boundary condition on $\partial\Omega$. The minimal eigenvalue λ_Ω can be estimated from below by one half of the sum of the minimal eigenvalues of the operators $\mathcal{L}_1 : V_0 \to H^{-1}(\Omega;\mathbb{R}^n)$, $\mathcal{L}_1\mathbf{w} = -\operatorname{div}(\nabla\mathbf{w}) = -\Delta\mathbf{w}$, and $\mathcal{L}_2 : V_0 \to H^{-1}(\Omega;\mathbb{R}^n)$, $\mathcal{L}_2\mathbf{w} = -\nabla(\operatorname{div}\mathbf{w})$. It is clear that the smallest eigenvalue of the second operator is zero, while the minimal eigenvalue of the first one equals the minimal eigenvalue of the scalar Laplace operator in Ω; the latter depends only on the geometry of the domain Ω and can be estimated from below by embedding Ω into a larger domain of a simpler shape, as discussed in Remark 3 at the end of section 4.1. This ultimately leads to an easily computable upper bound for the constant C_Ω from (56).

Hence, if $(\mathbf{u}, \underline{\underline{\mathbf{p}}}) \in (V_0 + \mathbf{u}_0) \times \mathbf{H}(\Omega;\operatorname{div})$ is the exact solution to the mixed problem (53)–(55) and $(\mathbf{v}, \underline{\underline{\mathbf{y}}}) \in (V_0 + \mathbf{u}_0) \times \mathbf{H}(\Omega;\operatorname{div})$ is any approximate solution to the problem, then the following a posteriori error estimates follow directly from Theorem 3.4:

$$\|(\mathbf{u}-\mathbf{v}, \underline{\underline{\mathbf{p}}} - \underline{\underline{\mathbf{y}}})\|_{1\times\operatorname{div}} \le \| \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{v}) - \underline{\underline{\mathbf{y}}} \|_* + \left(1 + 2\frac{C_\Omega^2}{\lambda_{\mathbb{L}}}\right)^{1/2} \|\operatorname{div}\underline{\underline{\mathbf{y}}} - \mathbf{f}\|, \tag{57}$$

$$\|(\mathbf{u}-\mathbf{v}, \underline{\underline{\mathbf{p}}} - \underline{\underline{\mathbf{y}}})\|_{1\times\operatorname{div}} \ge \frac{1}{\sqrt{3}}\left(\| \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{v}) - \underline{\underline{\mathbf{y}}} \|_* + \|\operatorname{div}\underline{\underline{\mathbf{y}}} - \mathbf{f}\|\right), \tag{58}$$

$$\| \underline{\underline{\mathbf{e}}}(\mathbf{u}-\mathbf{v}) \| \le \| \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{v}) - \underline{\underline{\mathbf{y}}} \|_* + \frac{C_\Omega}{\lambda_{\mathbb{L}}^{1/2}} \|\operatorname{div}\underline{\underline{\mathbf{y}}} - \mathbf{f}\|, \tag{59}$$

$$\| \underline{\underline{\mathbf{p}}} - \underline{\underline{\mathbf{y}}} \|_* \le \| \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{v}) - \underline{\underline{\mathbf{y}}} \|_* + \frac{C_\Omega}{\lambda_{\mathbb{L}}^{1/2}} \|\operatorname{div}\underline{\underline{\mathbf{y}}} - \mathbf{f}\|, \tag{60}$$

$$\|\underline{\underline{\mathbf{p}}} - \underline{\underline{\mathbf{y}}}\|_{\operatorname{div}} \le \| \mathbb{L}\underline{\underline{\mathbf{e}}}(\mathbf{v}) - \underline{\underline{\mathbf{y}}} \|_* + \left(1 + \frac{C_\Omega^2}{\lambda_{\mathbb{L}}}\right)^{1/2} \|\operatorname{div}\underline{\underline{\mathbf{y}}} - \mathbf{f}\|. \tag{61}$$

Estimates (57)–(61) provide sharp error bounds that are explicitly computable, if one has the approximate solution to problem (53)–(55) in the product space $(V_0 + \mathbf{u}_0) \times \mathbf{H}(\Omega;\operatorname{div})$. The construction of the approximation in this space can be done along the lines presented in the previous section for the case of a scalar elliptic problem.

REFERENCES

[1] Y. Achdou, C. Bernardi, and F. Coquel, *A priori and a posteriori analysis of finite volume discretizations of Darcy's equations*, Numer. Math., 96 (2003), pp. 17–42.

[2] A. Alonso, *Error estimators for a mixed method*, Numer. Math., 74 (1996), pp. 385–395.

[3] I. Babuška and G. N. Gatica, *On the mixed finite element method with Lagrange multipliers*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 192–210.

[4] I. Babuška and W. C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[5] T. Belytschko and T. Blacker, *Enhanced derivative recovery through least square residual penalty*, Appl. Numer. Math., 14 (1994), pp. 55–68.

[6] D. Braess and R. Verfürth, *A posteriori error estimators for the Raviart–Thomas element*, SIAM J. Numer. Anal., 33 (1996), pp. 2431–2444.

[7] J. H. Brandts, *Superconvergence and a posteriori error estimation for triangular mixed finite elements*, Numer. Math., 68 (1994), pp. 311–324.

[8] J. H. Brandts and Y. Chen, *An alternative to the least-squares mixed finite element method for elliptic problems*, in Numerical Mathematics and Advanced Applications, M. Feistauer, V. Dolejši, P. Knobloch, and K. Najzar, eds., Springer, Berlin, 2004, pp. 169–175.

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

[10] Z. Cai, R. Lazarov, T. A. Manteuffel, and S. F. McCormick, *First-order system least squares for second-order partial differential equations: Part* I, SIAM J. Numer. Anal., 31 (1994), pp. 1785–1799.

[11] Z. Cai, T. A. Manteuffel, S. F. McCormick, and S. V. Parter, *First-order system least squares (FOSLS) for planar linear elasticity: Pure traction problem*, SIAM J. Numer. Anal., 35 (1998), pp. 320–335.

[12] C. Carstensen, *A posteriori error estimate for the mixed finite element method*, Math. Comp., 66 (1997), pp. 465–476.

[13] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids.* I. *Low order conforming, non-conforming and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.

[14] C. Carstensen and G. Dolzmann, *A posteriori error estimates for mixed FEM in elasticity*, Numer. Math., 81 (1998), pp. 187–209.

[15] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North–Holland, New York, 1976.

[16] G. Gatica and M. Maischak, *A posteriori error estimates for the mixed finite element method with Lagrange multipliers*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 421–450.

[17] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations. Theory and Algorithms*, Springer-Verlag, New York, 1986.

[18] M. Křižek and P. Neittaanmäki, *Superconvergence phenomenon in the finite element method arising from averaging of gradients*, Numer. Math., 45 (1984), pp. 105–116.

[19] P. Ladeveze and D. Leguillon, *Error estimate procedure in the finite element method and applications*, SIAM J. Numer. Anal., 20 (1983), pp. 485–509.

[20] M. Lonsing and R. Verfürth, *A posteriori error estimators for mixed finite element methods in linear elasticity*, Numer. Math., 97 (2004), pp. 757–778.

[21] C. Lovadina and R. Stenberg, *Energy norm a posteriori error estimates for mixed finite element methods*, Math. Comp., 75 (2006), pp. 1659–1674.

[22] A. I. Pehlivanov, G. F. Carey, and R. D. Lazarov, *Least-squares mixed finite elements for second-order elliptic problems*, SIAM J. Numer. Anal., 31 (1994), pp. 1368–1377.

[23] W. Prager and J. L. Synge, *Approximations in elasticity based on the concept of functions space*, Quart. Appl. Math., 5 (1947), pp. 241–269.

[24] P.-A. Raviart and J.-M. Thomas, *A mixed finite element for second order elliptic problems*, in Mathematical Aspects of Finite Element Methods, I. Galligani and E. Magenes, eds., Springer, Berlin, 1977, pp. 292–315.

[25] S. Repin, *A posteriori error estimation for variational problems with uniformly convex functionals*, Math. Comp., 69 (2000), pp. 481–500.

[26] S. Repin, *Two-sided estimates of deviation from exact solutions of uniformly elliptic equations*, in Proceedings of the St. Petersburg Mathematical Society, Vol. 9, Amer. Math. Soc. Transl. (Ser. 2) 209, AMS, Providence, RI, 2003, pp. 143–171.

[27] S. Repin, S. Sauter, and A. Smolianski, *A posteriori error estimation for the Dirichlet problem with account of the error in approximation of boundary conditions*, Computing, 70 (2003), pp. 205–233.

[28] S. Repin, S. Sauter, and A. Smolianski, *A posteriori error estimation for the Poisson equation with mixed Dirichlet/Neumann boundary conditions*, J. Comput. Appl. Math., 164/165 (2004), pp. 601–612.

[29] S. Repin, S. Sauter, and A. Smolianski, *A posteriori estimation of dimension reduction errors for elliptic problems on thin domains*, SIAM J. Numer. Anal., 42 (2004), pp. 1435–1451.

[30] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, II, P. G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 523–639.

[31] N.-E. Wiberg, F. Abdulwahab, and S. Ziukas, *Enhanced superconvergent patch recovery incorporating equilibrium and boundary conditions*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 3417–3440.

[32] B. I. Wohlmuth and R. H. W. Hoppe, *A comparison of a posteriori error estimators for mixed finite element discretizations by Raviart–Thomas elements*, Math. Comp., 68 (1999), pp. 1347–1378.

[33] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.

[34] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Methods Engrg., 24 (1987), pp. 337–357.

[35] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates.* I. *The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 946–968

NONCONFORMING BOX-SCHEMES FOR ELLIPTIC PROBLEMS ON RECTANGULAR GRIDS*

ISABELLE GREFF†

Abstract. Recently, Courbet and Croisille [*RAIRO Modél. Math. Anal. Numér.*, 32 (1998), pp. 631–649] introduced the finite volume box-scheme for the two-dimensional (2D) Poisson problem in the case of triangular meshes. Generalizations to higher degree box-schemes have been published by Croisille and Greff [*Numer. Methods Partial Differential Equations*, 18 (2002), pp. 355–373]. These box-schemes are based on the principle of the finite volume method in that they take the average of the equations on each cell of the grid. This gives rise to a natural choice of unknowns located at the interface of the mesh. These box-schemes are conservative and use only one mesh. They can be interpreted as a discrete mixed Petrov–Galerkin formulation of the Poisson problem. In this paper we focus our interest on box-schemes for the Poisson problem in two dimensions on rectangular grids. We discuss the basic finite volume box-scheme and analyze and interpret it as three different box-schemes. The method is demonstrated by numerical examples.

Key words. Petrov–Galerkin method, finite volume, mixed finite elements, elliptic problems, nonconforming spaces

AMS subject classifications. 35J15, 65N15, 65N30

DOI. 10.1137/050647578

1. Introduction.

The aim of this paper is to introduce several box-schemes for elliptic problems on rectangular grids based on the model of [6, 7, 8, 13]. The principle of the box-scheme we intend to discuss here in the case of rectangular grids goes back to H. B. Keller [17], where a box-scheme for the one-dimensional (1D) heat equation is introduced. In the case of an elliptic system, the principle of the box-schemes consists of discretizing the mixed form of the equation, by taking the average of the conservation and the constitutive laws on the same grid, without any integration by part. In a sense, it is the most natural way to discretize a system in mixed form, very close to the physical setting of the equation. The finite element counterpart of that point of view is a Petrov–Galerkin formulation with two trial spaces and two test spaces. A comprehensive understanding of this kind of scheme has been introduced in [6, 7, 8] in the case of a triangular mesh. We also refer the reader to [21] for finite volume methods and their relations with Petrov–Galerkin formulation. It is the main purpose of this paper to understand the nature of the coupling between all the spaces involved, with a particular emphasis on the identification of the spurious modes typical of rectangular grids. It is hoped that this can be useful also for more standard schemes.

Here, we consider a rectangular domain $\Omega \subset \mathbb{R}^2$ covered by a regular grid $\mathcal{T}_h$ of rectangles with edges parallel to those of the domain. For the simplicity of the presentation and since we focus on the design principles of different box-schemes, we restrict ourselves to the Poisson problem $-\Delta u = f$ for $f \in L^2(\Omega)$ with homogeneous boundary conditions. The mixed form we consider is as follows: Find $(u, p) \in H_0^1(\Omega) \times$

*Received by the editors December 14, 2005; accepted for publication (in revised form) November 17, 2006; published electronically May 4, 2007.
http://www.siam.org/journals/sinum/45-3/64757.html
†Laboratoire de Mathématiques Appliquées, Université de Pau et des Pays de l'Adour, Avenue de l'Université, Bâtiment I.P.R.A. - BP 1155, F-64013 Pau Cedex (isabelle.greff@univ-pau.fr).

$H_{\mathrm{div}}(\Omega)$ such that

$$\begin{cases} (\operatorname{div} p + f, v)_{0,\Omega} = 0 \quad \text{for all } v \in L^2(\Omega), \\ (p - \nabla u, q)_{0,\Omega} = 0 \quad \text{for all } q \in (L^2(\Omega))^2, \end{cases} \tag{1}$$

where $H_{\mathrm{div}}(\Omega) = \{p \in (L^2(\Omega))^2 \,;\, \operatorname{div} p \in L^2(\Omega)\}$. As in [6, 8], the discretization of (1) is performed by a mixed Petrov–Galerkin scheme called a box-scheme. It involves four discrete spaces: $M_{1,h}, X_{1,h}$ as trial spaces and $M_{2,h}, X_{2,h}$ as test spaces. The box-scheme reads as follows: Find $(u_h, p_h) \in M_{1,h} \times X_{1,h}$ such that

$$\begin{cases} \displaystyle\sum_{K \in \mathcal{T}_h} (\operatorname{div} p_h + f, v_h)_{0,K} = 0 \quad \text{for all } v_h \in M_{2,h}, \\ \displaystyle\sum_{K \in \mathcal{T}_h} (p_h - \nabla u_h, q_h)_{0,K} = 0 \quad \text{for all } q_h \in X_{2,h}. \end{cases} \tag{2}$$

The uniqueness of the solution of (2) implies in particular the identity of the dimensions

$$\dim M_{1,h} + \dim X_{1,h} = \dim M_{2,h} + \dim X_{2,h}. \tag{3}$$

The starting point of this article is the paper [5] by Courbet, where an original algebraic box-scheme on quadrangles is introduced for the time dependent diffusive problem. We give a finite element interpretation of that scheme with three different box-schemes; this allows us to state its stability and accuracy properties. As is the case on a triangular mesh, a natural choice for the approximation of the flux p_h is the lowest order Raviart–Thomas space. For the unknown u_h three possible choices are the standard Q^1-Lagrange space, its nonconforming analogue, Q^1_{nc}, or the so-called P^1-nonconforming quadrilateral finite element, introduced by Park and Sheen [19]. Due to the properties of the different trial spaces, we can make the link between these three box-schemes explicit. An important characteristic of these box-schemes is their equivalence with a decoupled formulation in the unknowns u_h and p_h. This allows the computation of the discrete flux p_h in an inexpensive way, since it is given as a function of ∇u_h and the right-hand side f. This local reconstruction of the flux p_h in each cell is of particular interest for porous media problems, e.g., contaminant transport, where the velocity is computed by the Darcy law and introduced in a convection-diffusion equation for the computation of the concentration. This decoupled feature of the box-scheme extends the observation by Marini [18], that the flux in the mixed finite element method can be recovered in an inexpensive way. Concerning the a posteriori error estimates of the box-scheme, we refer the reader to the recent works by El Alaoui and Ern [10, 11]. Finally, let us mention that an increasing interest in box-schemes has recently appeared [3, 4]. Note that a different possibility for extending the box-scheme of [6] on rectangles, using the Rannacher–Turek nonconforming finite element space, has been studied in [3, 14, 15].

Let us give now some standard notation. We introduce the mesh dependent norms defined, respectively, on the mesh dependent spaces $H^1_0(\Omega)+M_{1,h}$ and $H_{\mathrm{div}}(\Omega)+X_{1,h}$:

$$|u|_{1,h} = \left(\sum_K |\nabla u|^2_{0,K} \right)^{1/2},$$

$$\|u\|_{1,h} = (|u|^2_{0,\Omega} + |u|^2_{1,h})^{1/2} \quad \text{for all } u \in H^1_0(\Omega) + M_{1,h},$$

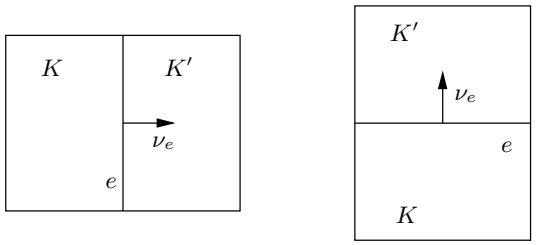


FIG. 1. *Vertical and horizontal edges.*

$$|p|_{\mathrm{div},h} = \left(\sum_K |\mathrm{div}\, p|^2_{0,K} \right)^{1/2},$$
$$\|p\|_{\mathrm{div},h} = (|p|^2_{0,\Omega} + |p|^2_{\mathrm{div},h})^{1/2} \quad \text{for all } p \in H_{\mathrm{div}}(\Omega) + X_{1,h}.$$

The geometrical notation is as follows. The rectangles are denoted by K with center $G_K(x_K, y_K)$, area $|K|$, and diameter h_K. We denote by h the maximum of the diameters of the elements of the mesh. The sizes of the sides of the rectangle K are $|e_{x,K}|$ and $|e_{y,K}|$. We will write ∂K for the set of edges of K. Let ν be the unit outward normal vector along the boundary $\partial\Omega$ and ν_K the one along the boundary ∂K of the rectangle K. The unitary normal vector to an edge e on the boundary $\partial\Omega$ is simply $\nu_e = \nu$. To each interior edge $e = \partial K \cap \partial K'$, we also associate a unit normal vector ν_e, which is arbitrarily defined as $\nu_e = \nu_K$ and $\nu_e = -\nu_{K'}$ in accordance with Figure 1. For an interior edge $e = \partial K \cap \partial K'$, $[u]_e = u_{|K',e} - u_{|K,e}$ denotes the jump of u along e with respect to the normal along the edge e. The midpoint of an edge e is x_e. The sets $\mathcal{A}_i$ and $\mathcal{A}_b$ denote the internal and boundary edges, respectively. We define $\mathcal{A} = \mathcal{A}_i \cup \mathcal{A}_b$ to be the set of all edges with global numbering. The number of rectangles is *NE*. The number of edges (respectively, internal, boundary edges) is *NA* (respectively, NA_i, NA_b). The number of vertices (respectively, internal, boundary vertices) is *NV* (respectively, NV_i, NV_b). The Euler relations are

$$4NE = NA_i + NA \quad \text{and} \quad NE - NA + NV = 1. \tag{4}$$

The gradient of f is $\nabla f = (\partial_x f, \partial_y f)^T$ and the two-dimensional (2D) rotational is $\mathrm{curl}\, f = (\partial_y f, -\partial_x f)^T$. Let P^0 be the space of piecewise constant functions, P^1 be the space of piecewise affine functions, and Q^1 be the space of bilinear functions. We define Π^0 to be the L^2-projection operator on the piecewise constant functions. Let us recall the definition of RT^0, the lowest order space of Raviart and Thomas [20], useful to discretize the flux $p = \nabla u$:

$$RT^0 = \{q_h \in H_{\mathrm{div}}(\Omega)\,;\, q_h \in RT^0(K) \quad \text{for all } K \in \mathcal{T}_h\},$$

where the local space $RT^0(K)$ is

$$RT^0(K) = P^0(K)^2 + P^0(K) \begin{pmatrix} x \\ 0 \end{pmatrix} + P^0(K) \begin{pmatrix} 0 \\ y \end{pmatrix}.$$

The space RT^0 is of dimension *NA*, the degrees of freedom being given by the linear forms

$$L_a(q_h) = \frac{1}{|a|} \int_a q_h \cdot \nu_a \, d\sigma \quad \text{for all } a \in \mathcal{A}.$$

Note that the normal component $p_h \cdot \nu_a$ of p along each interior edge is constant.

The outline of the paper is as follows. In section 2, we recall briefly the design principles of Courbet's scheme. To interpret it as a finite element method, we introduce in section 3 a finite element box-scheme based on the space used by Courbet to approximate the unknown u and the standard Q^1-Lagrange finite element space. The approximation of the flux $p = \nabla u$ is done using the Raviart–Thomas space. However, this box-scheme seems to be unstable. In section 4 we build a new box-scheme generalizing the previous one and based on the inclusion of the space Q^1-Lagrange into the nonconforming Q^1 space, Q^1_{nc}. Both unknowns u and p are discretized in nonconforming spaces with respect to $H^1_0(\Omega)$ and $H^1_{\mathrm{div}}(\Omega)$. We perform the numerical analysis of the scheme and its equivalence to a decoupled problem in u_h and p_h: a nonconforming scheme in u_h and a local reconstruction formula of p_h. Consequently, we can make explicit the link to the box-scheme of section 3. It turns out that the solution u_h of the box-scheme is only affine (and not bilinear) per rectangle. Section 5 is devoted to the development and the analysis of a reduced box-scheme. We conclude this work with numerical results in section 6. Note that this paper has been presented partly in [13, 14].

2. Courbet's box-scheme.

2.1. Introduction. In [5], Courbet introduced a box-scheme for the time dependent mixed formulation of the compressible Navier–Stokes equations. The scheme intended to extend to a rectangular grid the well-known box-scheme of H. B. Keller for the heat equation [17]. In the case of the Poisson problem, the box-scheme of Courbet is a derivation of the mixed form of the problem taken as mean value on each rectangle K:

$$\begin{cases} \displaystyle\int_K \operatorname{div} p \, dx + \int_K f \, dx = 0, \\ \displaystyle\int_K p \, dx - \int_K \nabla u \, dx = 0, \\ u = 0 \text{ on } \partial\Omega. \end{cases} \tag{5}$$

We refer to the Courbet box-scheme later as (BS1): Find $u = (u_a)_{a\in\mathcal{A}}$ and $p = (p_a)_{a\in\mathcal{A}}$ such that for all rectangles K of the grid,

$$\text{(BS1)} \quad \begin{cases} \displaystyle\sum_{a\in\partial K} |a|\, p_{a,K} + \int_K f \, dx = 0, \\ \displaystyle\frac{(p_{a_{1,K}} - p_{a_{3,K}})}{2} - \frac{(|a_{1,K}| u_{a_{1,K}} - |a_{3,K}| u_{a_{3,K}})}{|K|} = 0, \\ \displaystyle\frac{(p_{a_{2,K}} - p_{a_{4,K}})}{2} - \frac{(|a_{2,K}| u_{a_{2,K}} - |a_{4,K}| u_{a_{4,K}})}{|K|} = 0, \\ u_a = 0 \quad \text{for all } a \in \mathcal{A}_b, \end{cases} \tag{6}$$

where the subscripts $a_{1,K}$, $a_{2,K}$, $a_{3,K}$, and $a_{4,K}$ are related to the edges $a_{1,K}$, $a_{2,K}$, $a_{3,K}$, and $a_{4,K}$ of each rectangle K (see Figure 2). The unknowns u_a and p_a denote, respectively, the average of u and the normal component of the flux $p = \nabla u$ along an edge a and are located at the interface of the mesh. This gives $4NE$ unknowns and $3NE$ equations. In contrast to the analogous scheme on triangles introduced in [6], here is a lack of NE equations. Courbet suggests adding the constraint on each

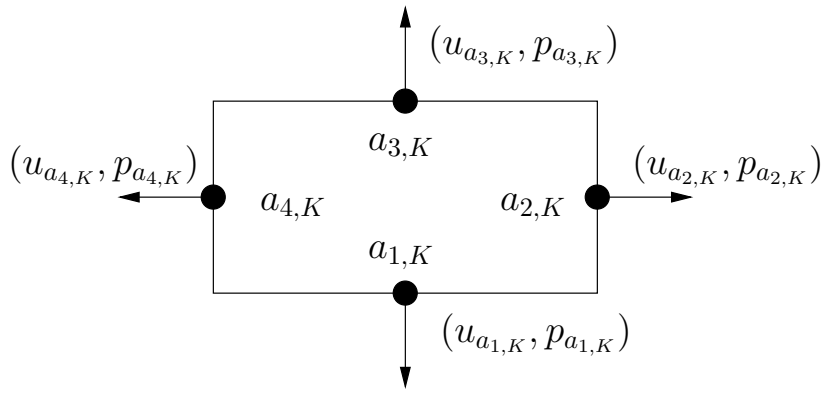


FIG. 2. *Rectangle K, with edges $a_{i,K}$ and unknowns $(u_{a_{i,K}}, p_{a_{i,K}})$ for $i = 1, \ldots, 4$.*

rectangle K as a discrete equation:

$$u_{a_{1,K}} + u_{a_{3,K}} = u_{a_{2,K}} + u_{a_{4,K}}. \tag{7}$$

In particular, the mean value of the solution u in each box coincides with its horizontal and vertical averages. Let us denote by C_0 the space introduced by Courbet to discretize the unknown u. It is generated by vectors of size NA, vanishing at the boundary, and satisfying the additional condition (7) on each rectangle of the grid. The space C_0 is defined by

$$\begin{aligned} C_0 = \big\{ (u_a)_{a\in\mathcal{A}} \in \mathbb{R}^{NA} \text{ such that } u_a = 0 \text{ for all } a \in \mathcal{A}_b \text{ and} \\ u_{a_{1,K}} + u_{a_{3,K}} = u_{a_{2,K}} + u_{a_{4,K}} \text{ on each rectangle } K \big\} \\ = \big\{ (u_a)_{a\in\mathcal{A}_i} \in \mathbb{R}^{NA_i},\ u_{a_{1,K}} + u_{a_{3,K}} = u_{a_{2,K}} + u_{a_{4,K}} \text{ on each rectangle } K \big\}. \end{aligned}$$

However, the dimension of the space C_0 is $\dim C_0 = NV_i = NA_i - NE + 1 > NA_i - NE$. Indeed, the boundary degrees of freedom of the space C_0 are not independent. In fact, if $u_a = 0$ for $NA_b - 1$ boundary edges, then $u_a = 0$ holds on the last one. This implies that the box-scheme (BS1) does not define a well-posed problem in the sense that the number of unknowns is larger than the number of equations. Actually, due to the dimension of the space C_0, the number of unknowns is: Number of unknowns $(u_a, p_a) = NV_i + NA = 3NE + 1$, whereas there are only $3NE$ equations. Despite this dimension inconsistency, very good numerical results are reported in [5] for the time dependent heat equation.

The observation that $\dim C_0 = NV_i$ suggests that the space C_0 is identical to the Q^1-Lagrange space with homogeneous boundary conditions.

LEMMA 2.1. *The mapping L defines a bijection between $Q^1_{c,0}$ and the Courbet space*

$$\begin{aligned} L :\ Q^1_{c,0} &\longrightarrow C_0, \\ u &\longmapsto \big(u(x_a)\big)_{a\in\mathcal{A}}, \end{aligned}$$

where x_a denotes the midedge of a and Q^1_c is the standard Q^1-Lagrange finite element space

$$Q^1_c = \{u \in \mathcal{C}^0(\Omega)\,;\, u \in Q^1(K) \text{ for all } K \in \mathcal{T}_h\}, \quad Q^1(K) = \mathrm{Span}\{1, x, y, xy\},$$

and $Q^1_{c,0}$ is its restriction to functions vanishing on $\partial\Omega$.

The proof of the lemma follows from the linearity, injectivity (see Proposition 2.1 hereafter) of the mapping L, and the equality of the dimensions of the spaces C_0 and $Q^1_{c,0}$. Before going further with the stabilization of the box-scheme (BS1), we recall some useful properties of the nonconforming Q^1 finite element space and its relation to Q^1_c.

2.2. Some properties of the Q^1 nonconforming space. The nonconforming Q^1 finite element space denoted by Q^1_{nc} is defined by

$$Q^1_{nc} = \Big\{ u_h \in L^2(\Omega)\,;\, u_h \in Q^1(K) \text{ for all } K \in \mathcal{T}_h; \\ \int_a u_{h|K_1}\, d\sigma = \int_a u_{h|K_2}\, d\sigma \text{ for all } a = \partial K_1 \cap \partial K_2 \in \mathcal{A}_i \Big\}.$$

The space $Q^1_{nc,0}$ is the zero boundary space:

$$Q^1_{nc,0} = \Big\{ u_h \in Q^1_{nc}\,;\, \int_a u_h\, d\sigma = 0 \text{ for all } a \in \mathcal{A}_b \Big\}.$$

Since the edges of the grid are parallel to the axis of the domain, the mean value of a function in Q^1 along an edge is the value at the midpoint of the edge. We recall that for all $v_h \in Q^1_{nc}$, the set of values

$$p_a(v_h) = v_h(x_a) \text{ for all } a \in \mathcal{A} \text{ with the associated midpoint } x_a$$

does not form a unisolvent set of degrees of freedom [1, 2, 12]. Indeed, let η be the function defined on $Q^1(K)$ by

$$\eta :\, Q^1(K) \longrightarrow \mathbb{R}^4, \\ p \longmapsto \big(p(x_a)\big)_{a\in\partial K}.$$

It is well known that the kernel of η is of dimension 1, generated by the nonconforming bubble b_K,

$$\operatorname{Ker}\eta = \operatorname{Span}\{b_K\}, \quad b_K(x,y) = \frac{4}{|K|}(x - x_K)(y - y_K),$$

where (x_K, y_K) is the center of the rectangle K. It is easy to check that for any $K \in \mathcal{T}_h$ and any $v \in Q^1(K)$, the function b_K has the following properties:

$$\int_{\partial K} b_K\, d\sigma = 0, \qquad \int_K \operatorname{curl} b_K \cdot \nabla v\, dx = 0, \\ \int_{\partial K} (\operatorname{curl} b_K \cdot \nu_K)\, v\, d\sigma = 0. \tag{8}$$

Let Ψ be the vector space generated by the local bubbles:

$$\Psi = \{\psi\,;\, \psi_{|K} = \alpha_K b_K,\, \alpha_K \in \mathbb{R} \text{ for all } K \in \mathcal{T}_h\}.$$

Then, $\dim \Psi = NE$ and $\Psi \subseteq Q^1_{nc,0}$.

DEFINITION 2.1. *We define $\mathcal{B} \in Q^1_c$ by $\mathcal{B} = \sum_{K\in\mathcal{T}_h} \operatorname{sgn}(K) b_K$, where $\operatorname{sgn}(K)$ takes alternately the values -1, $+1$ as displayed in Figure 3.*

−	+	−	+	−
+	−	K	−	+
−	+	−	+	−

FIG. 3. *Sign of K.*

The function $\mathcal{B}$ is the so-called *hourglass mode* introduced by Hansbo in [16], which gives rise to some instability. By using this definition and the properties of the previous spaces, we prove the following proposition.

PROPOSITION 2.1. *The spaces Q_c^1, Q_{nc}^1, and Ψ satisfy*

$$\text{(i)}\quad Q_c^1 \cap \Psi = \mathrm{Span}\{\mathcal{B}\}, \quad \text{(ii)}\quad Q_{nc}^1 = Q_c^1 + \Psi, \quad \text{(iii)}\quad Q_{nc,0}^1 = Q_{c,0}^1 \oplus \Psi.$$

In particular, $\dim Q_{nc}^1 = NA$ *and* $\dim Q_{nc,0}^1 = NA_i + 1$.

Proof. (i) Let $\psi = \sum_{K\in\mathcal{T}_h} \alpha_K b_K \in Q_c^1 \cap \Psi$. For an internal edge $a = \partial K_1 \cap \partial K_2$ the bubble satisfies $b_{K_1}|_a = -b_{K_2}|_a$. Then, the continuity of ψ along each internal edge implies $\alpha_K = \mathrm{sgn}(K)\,\alpha$ for all $K \in \mathcal{T}_h$, $\alpha \in \mathbb{R}$. So, $\mathrm{Span}\{\mathcal{B}\} \subset Q_c^1 \cap \Psi$. The reverse inclusion is clear.

(ii) Let us define the space $M = Q_c^1 + \Psi$. Then, $M \subseteq Q_{nc}^1$. To prove $M = Q_{nc}^1$, we will prove that $\dim M = \dim Q_{nc}^1$. Let i be the linear map

$$\begin{aligned} i :\, Q_{nc}^1 &\longrightarrow \mathbb{R}^{NA}, \\ u &\longmapsto \big(u(x_a)\big)_{a\in\mathcal{A}}, \end{aligned}$$

where x_a is the midpoint of the edge $a \in \mathcal{A}$. Using the definitions of Ψ and b_K, we prove that

$$\mathrm{Ker}\, i = \Big\{ u \in Q_{nc}^1 ;\, u(x_a) = 0 \text{ for all } a \in \mathcal{A} \Big\} \subseteq \Psi,$$

and according to the notation of Figure 2, where $x_{a_{i,K}}$ is the midpoint of the edge $a_{i,K} \in \partial K$, $i = 1, \ldots, 4$,

$$\mathrm{Im}\, i = \Big\{ \big(u(x_a)\big)_{a\in\mathcal{A}} \in \mathbb{R}^{NA} ;\, u(x_{a_{1,K}}) + u(x_{a_{3,K}}) = u(x_{a_{2,K}}) + u(x_{a_{4,K}}) \text{ for all } K \in \mathcal{T}_h \Big\}.$$

This in turn gives that $\dim Q_{nc}^1 = \dim(\mathrm{Ker}\, i) + \dim(\mathrm{Im}\, i) \le NA$. Moreover, $\dim M = \dim Q_c^1 + \dim \Psi - \dim(Q_c^1 \cap \Psi) = NV + NE - 1$. The Euler relation gives $\dim M = NA$. $M \subseteq Q_{nc}^1$, so $\dim M \le \dim Q_{nc}^1$. We deduce that $\dim Q_{nc}^1 = NA$, which concludes the proof of (ii).

The statement (iii) is directly implied by (i) and (ii). ∎

Using the property of $\mathcal{B}$ and the continuity of the normal component of the element in RT^0, we deduce the following lemma.

LEMMA 2.2. *Let Φ be the vector space generated by the curl of the nonconforming bubble*

$$\Phi = \mathrm{curl}\, \Psi = \left\{ \phi = \sum_{K\in\mathcal{T}_h} \beta_K\, \mathrm{curl}\, b_K, \quad \beta_K \in \mathbb{R} \right\}.$$

Then, $\Phi \cap RT^0 = \mathrm{Span}\{\mathrm{curl}(\mathcal{B})\}$.

Note that the box-scheme (BS1) is a derivation of the mixed formulation of the Laplace equation on each rectangle K given by the system (5). Since $\int_K \mathrm{curl}\, b_K\, dx = 0$ and $\mathrm{div}(\mathrm{curl}\, b_K) = 0$ we get from the mixed formulation (5) that for any β_K, we can superpose to p_h any function $\sum_K \beta_K \,\mathrm{curl}\, b_K$, which is a parasitic mode. Therefore a stabilization of the scheme has to eliminate that mode.

3. A first stabilization of Courbet's box-scheme. We will now give a first stabilization of the box-scheme (BS1) using the finite element interpretation of the space C_0 coupled with the Raviart–Thomas space RT^0 in order to discretize the unknowns (u, p). We also need to add one additional test function in order to have the right number of equations. The element $\mathcal{B}$ is the simplest choice according to results of the previous section.

PROPOSITION 3.1. *Let us call* (BS2) *the box-scheme: Find the solution* $(u_h, p_h) \in Q^1_{c,0} \times RT^0$ *of*

$$\text{(BS2)} \quad \begin{cases} (\mathrm{div}\, p_h + f, v_h)_{0,\Omega} = 0 & \text{for all } v_h \in P^0, \\ (p_h - \nabla u_h, q_h)_{0,\Omega} = 0 & \text{for all } q_h \in X_{2,h} = (P^0)^2 + \mathrm{Span}\{\mathrm{curl}(\mathcal{B})\}. \end{cases} \tag{9}$$

(i) *The box-scheme* (BS2) *has* $3NE + 1$ *unknowns.*

(ii) *The box-scheme* (BS2) *has a unique solution given by*

(a) $u_h \in Q^1_{c,0}$ *is the solution of*

$$\sum_{K \in \mathcal{T}_h} (\Pi^0 \nabla u_h, \Pi^0 \nabla v_h)_{0,K} = (\Pi^0 f, v_h)_{0,\Omega} \quad \text{for all } v_h \in Q^1_{c,0}. \tag{10}$$

(b) p_h *is given by*

$$p_{h|K} = (\Pi^0 \nabla u_h)_K - \frac{\Pi^0 f_{|K}}{2} \begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix} + \gamma_K \begin{pmatrix} x - x_K \\ -(y - y_K) \end{pmatrix}, \tag{11}$$

where γ_K *is the solution of a certain sparse linear system.*

Proof. (i) Using the Euler relation, we get the following identity between the number of unknowns and the number of equations:

$$\dim Q^1_{c,0} + \dim RT^0 = NV_i + NA = 3NE + 1 = \dim P^0 + \dim (X_{2,h}).$$

(ii) Let $(u_h, p_h) \in Q^1_{c,0} \times RT^0$ be a solution of (BS2). We prove that (u_h, p_h) satisfies the system ((a), (b)).

(a) Suppose we are given $v_h \in Q^1_{c,0}$; then $q_h = \Pi^0(\nabla v_h) \in X_{2,h}$. Introducing this value of q_h in the second equation of (9) and afterward using the decomposition $\nabla v_{h|K} = \Pi^0 \nabla v_{h|K} + \delta_K \nabla b_K$ for any $\delta_K \in \mathbb{R}$ and Green's formula, we get

$$\begin{aligned} \sum_K (\nabla u_h, \Pi^0 \nabla v_h)_{0,K} &= \sum_K (p_h, \Pi^0 \nabla v_h)_{0,K} \\ &= \sum_K (p_h, \nabla v_h - \delta_K \nabla b_K)_{0,K} \\ &= -\sum_K \int_K \mathrm{div}\, p_h\, v_h\, dx + \sum_K \int_{\partial K} v_h\, p_h \cdot \nu_K\, d\sigma \\ &\quad + \sum_K \int_K \delta_K \,\mathrm{div}\, p_h\, b_K\, dx - \sum_K \int_{\partial K} p_h \cdot \nu_K\, \delta_K\, b_K\, d\sigma. \end{aligned} \tag{12}$$

Since the mean value of the bubble function b_K vanishes and $p_h \in RT^0$, we have that $\int_K \delta_K \operatorname{div} p_h \, b_K \, dx = 0$ and $\int_{\partial K} p_h \cdot \nu_K \, \delta_K \, b_K \, d\sigma = 0$. On the other hand, the first equation of (9) gives $\operatorname{div} p_{h|K} = -\Pi^0 f_{|K}$ for all $K \in \mathcal{T}_h$. Therefore, the equality (12) becomes

$$\begin{aligned}\sum_K (\nabla u_h, \Pi^0 \nabla v_h)_{0,K} = \sum_K \int_K \Pi^0 f v_h \, dx - \sum_{a \in \mathcal{A}_i} \int_a p_h \cdot \nu_a \, [v_h]_a \, d\sigma \\ + \sum_{a \in \mathcal{A}_b} \int_a p_h \cdot \nu_a \, v_h \, d\sigma.\end{aligned} \tag{13}$$

Since $v_h \in Q^1_{c,0}$, $[v_h]_a = 0$ for all $a \in \mathcal{A}_i$ and $v_{h|a} = 0$ for all $a \in \mathcal{A}_b$, the relation (13) becomes

$$\sum_K (\nabla u_h, \Pi^0 \nabla v_h)_{0,K} = \sum_K (\Pi^0 f, v_h)_{0,K},$$

which concludes (a).

(b) Any function p_h in $RT^0(K)$ can be decomposed as

$$p_{h|K} = (\Pi^0 p_h)_{|K} + \frac{\operatorname{div} p_{h|K}}{2} \begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix} + \gamma_K \begin{pmatrix} x - x_K \\ -(y - y_K) \end{pmatrix}, \quad \gamma_K \in \mathbb{R}.$$

Using, respectively, the first and second equations of (9), we get $\operatorname{div} p_{h|K} = -\Pi^0 f_{|K}$ and $(\Pi^0 p_h)_{|K} = (\Pi^0 \nabla u_h)_{|K}$. Then

$$p_{h|K} = (\Pi^0 \nabla u_h)_{|K} - \frac{\Pi^0 f_{|K}}{2} \begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix} + \gamma_K \begin{pmatrix} x - x_K \\ -(y - y_K) \end{pmatrix}. \tag{14}$$

The computation of the coefficient γ_K is done using (a), the second equation of (9) with $q_h = \sum_K \operatorname{sgn}(K) \operatorname{curl} b_K$, and the continuity of the normal component of p_h. More details can be found at the end of this paper in the appendix.

This implies that any solution of the box-scheme (BS2) is a solution of the system ((a), (b)), which is unique. Indeed, $f = 0$ in (a) implies $\Pi^0 \nabla u_h = 0$. Lemma A.3 (see the appendix) and the zero values of u_h on $\partial\Omega$ permit us to conclude that $u_h = 0$. Replacing $f = 0$ and $u_h = 0$ in (11) implies $p_{h|K} = \gamma_K$. Proposition A.1 gives $\big(\gamma_K\big)_K$ as the solution of a linear system with a vanishing right-hand side. This means that $\gamma_K = 0$ for all K, and therefore $p_h = 0$. This concludes that $f = 0$ implies $u_h = 0$, $p_h = 0$. The existence of solutions of (BS2) is deduced from the uniqueness of the solution, the linearity of the problem, and the equality between the number of unknowns and equations. □

Remarks. (i) As proved by Hansbo [16], the one-point integration of the gradient of u_h (Proposition 3.1(ii)) is not sufficient to obtain the stability of the scheme.

(ii) The parasitic perturbation $\sum_K \beta_K \operatorname{curl} b_K \in \Phi$ seems to be controlled globally by the box-scheme but not locally. As a consequence, we do not get a local reconstruction of the flux p_h in each rectangle K.

4. A second stabilization of Courbet's box-scheme. Due to its possible instability, the box-scheme (BS2) is not totally satisfying. So, we want to build a box-scheme using larger spaces for both unknowns u and p. The basic idea is to use the nonconforming space $Q^1_{nc,0}$ containing the Q^1-Lagrange space $Q^1_{c,0}$ (used in (BS2)) for the approximation of u. For the flux, we consider the space RT^0 of Raviart

and Thomas, supplemented with the space Φ of the rotational of the bubble. Note that those spaces are both nonconforming, respectively, in $H_0^1(\Omega)$ and $H_{\mathrm{div}}(\Omega)$. Also this choice of spaces gives the advantage of getting the number of unknowns to be proportional to the number of rectangles; i.e., the trial spaces in (2) can be piecewise polynomial spaces.

4.1. Definition of the box-scheme.

PROPOSITION 4.1. *Let (BS_{nc}) be the following box-scheme: Find $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$, being the solution of*

(15)
$$(BS_{nc}) \quad \begin{cases} \displaystyle\sum_{K\in\mathcal{T}_h} (\operatorname{div} p_h + f, v_h)_{0,K} = 0 \text{ for all } v_h \in M_{2,h} = P^0, \\ \displaystyle\sum_{K\in\mathcal{T}_h} (p_h - \nabla u_h, q_h)_{0,K} = 0 \text{ for all } q_h \in X_{2,h} = (P^0)^2 + P^0 \begin{pmatrix} y \\ x \end{pmatrix} + P^0 \begin{pmatrix} x \\ -y \end{pmatrix}. \end{cases}$$

(i) *The box-scheme (BS_{nc}) has $5NE$ degrees of freedom.*

(ii) *The box-scheme (BS_{nc}) has a unique solution $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ given by the following:*

(a) *u_h is the solution of the following variational problem: Find $u_h \in Q^1_{nc,0}$ such that*
$$\sum_{K\in\mathcal{T}_h} (\nabla u_h, \nabla v_h)_{0,K} = (\Pi^0 f, v_h)_{0,\Omega} \quad \text{for all } v_h \in Q^1_{nc,0}.$$

(b) *p_h is locally given by*
$$p_{h|K} = (\nabla u_h)_{|K} - \frac{\Pi^0 f_{|K}}{|e_{x,K}|^2 + |e_{y,K}|^2} \begin{pmatrix} |e_{y,K}|^2 (x - x_K) \\ |e_{x,K}|^2 (y - y_K) \end{pmatrix}.$$

Note that this box-scheme is nonconforming for both unknowns u_h and p_h. The test spaces $M_{2,h}$ and $X_{2,h}$ (in the system (2)) are piecewise polynomial functions. We remark that $X_{2,h}$ is also $X_{2,h} = (P^0)^2 + P^0(\nabla b_K) + P^0(\operatorname{curl} b_K)$; in particular, $\nabla(M_{1,h}) \subseteq X_{2,h}$.

Proof. (i) By the Euler relations, we prove that $\dim Q^1_{nc,0} + \dim (RT^0 + \Phi) = (NA_i + 1) + (NA + NE - 1) = 5\,NE = \dim X_{2,h} + \dim P^0$.

(ii) Let us prove that any $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ satisfying the equations (BS_{nc}) fulfills the system ((a), (b)).

(a) Let $v_h \in Q^1_{nc,0}$. Let $q_h = \nabla v_h \in X_{2,h}$ in the second equation of (15). By integration by parts,
$$\sum_K (\nabla u_h, \nabla v_h)_{0,K} = -\sum_K \int_K \operatorname{div} p_h \, v_h \, dx + \sum_K \int_{\partial K} (p_h \cdot \nu_K) \, v_h \, d\sigma.$$
Moreover, since $\operatorname{div} p_{h|K} \in P^0(K)$, the first equation of (15) gives $\operatorname{div} p_{h|K} = -\Pi^0 f_{|K}$. Hence, $p_h \in RT^0 + \Phi$ can be written as $p_h = \bar{p}_h + \sum_K \beta_K \operatorname{curl} b_K$, with $\bar{p}_h \in RT^0$. This implies that
$$\begin{aligned} \sum_K (\nabla u_h, \nabla v_h)_{0,K} &= \sum_K (\Pi^0 f, v_h)_{0,K} + \sum_K \int_{\partial K} (\bar{p}_h + \beta_K \operatorname{curl} b_K) \cdot \nu_K \, v_h \, d\sigma \\ &= (\Pi^0 f, v_h)_{0,\Omega} + \sum_K \int_{\partial K} \bar{p}_h \cdot \nu_K \, v_h \, d\sigma \\ &\quad + \sum_K \int_{\partial K} \beta_K \operatorname{curl} b_K \cdot \nu_K \, v_h \, d\sigma. \end{aligned}$$

Using the properties (8) of the bubble b_K and the continuity of the normal component of elements $p_h \in RT^0 \subset H_{\mathrm{div}}(\Omega)$, we obtain

$$\sum_K (\nabla u_h, \nabla v_h)_{0,K} = (\Pi^0 f, v_h)_{0,\Omega} + \sum_{a\in\mathcal{A}_b} \int_a \bar{p}_h \cdot \nu_a \, v_h \, d\sigma - \sum_{a\in\mathcal{A}_i} \int_a \bar{p}_h \cdot \nu_a \, [v_h]_a \, d\sigma.$$

Since $v_h \in Q^1_{nc,0}$ and $p_h \cdot \nu_a \in P^0(a)$,

$$\int_a \bar{p}_h \cdot \nu_a \, v_h \, d\sigma = 0 \quad \text{for all } a \in \mathcal{A}_b \quad \text{and} \quad \int_a \bar{p}_h \cdot \nu_a \, [v_h]_a \, d\sigma = 0 \quad \text{for all } a \in \mathcal{A}_i,$$

which concludes (a). In particular, for $v_h = b_K \in Q^1_{nc,0}$,

$$(\nabla u_h, \nabla b_K)_{0,K} = (\Pi^0 f, b_K)_{0,K}.$$

The mean value of b_K equals zero on each rectangle; hence $(\Pi^0 f, b_K)_{0,K} = 0$. Also, ∇u_h is locally written as $(\nabla u_h)_{|K} = (\Pi^0 \nabla u_h)_{|K} + d_K \nabla b_K$, where d_K is given by u_h. We deduce that

$$\begin{aligned} 0 = (\nabla u_h, \nabla b_K)_{0,K} &= \big((\Pi^0 \nabla u_h)_{|K} + d_K \nabla b_K, \nabla b_K\big)_{0,K} \\ &= \underbrace{\big((\Pi^0 \nabla u_h)_{|K}, \nabla b_K\big)_{0,K}}_{=0} + d_K |\nabla b_K|^2_{0,K}. \end{aligned}$$

This means that $d_K = 0$ or, equivalently, that the bubble component of the solution u_h vanishes. In particular, $(\nabla u_h)_{|K} = (\Pi^0 \nabla u_h)_{|K}$.

(b) Since $p_h \in RT^0 + \Phi$, we have $p_h = \bar{p}_h + \sum_K \beta_K \operatorname{curl} b_K$ with $\bar{p}_h \in H_{\mathrm{div}}(\Omega)$. Again for each rectangle K, $\operatorname{div} p_{h|K} = -\Pi^0 f_{|K}$ and $\operatorname{div}(\operatorname{curl} b_K) = 0$, so that

$$\operatorname{div} p_h = \operatorname{div} \bar{p}_h = -\Pi^0 f.$$

The second equation of (15) implies that $(\Pi^0 p_h)_{|K} = (\Pi^0 \nabla u_h)_{|K} = (\nabla u_h)_{|K}$. On the other hand,

$$p_{h|K} = (\Pi^0 p_h)_{|K} + \frac{\operatorname{div} p_{h|K}}{2} \begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix} + \tilde{\beta}_K \begin{pmatrix} x - x_K \\ -(y - y_K) \end{pmatrix}, \quad \tilde{\beta}_K \in \mathbb{R}.$$

This is equivalent to

$$p_{h|K} = \underbrace{(\nabla u_h)_{|K} - \frac{\Pi^0 f_{|K}}{2} \begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix}}_{p_{h,1}} + \underbrace{\tilde{\beta}_K \begin{pmatrix} x - x_K \\ -(y - y_K) \end{pmatrix}}_{p_{h,2}}.$$

For evaluating the coefficient $\tilde{\beta}_K$, we use the second equation of (15) with $q_h = \operatorname{curl} b_K$:

$$\int_K (p_{h,1} + p_{h,2} - \nabla u_h) \cdot \operatorname{curl} b_K \, dx = 0. \tag{16}$$

We know by (8) that $\int_K \nabla u_h \cdot \operatorname{curl} b_K \, dx = 0$. Inserting the value of $p_{h,1}$, (16) becomes

$$\int_K p_{h,2} \cdot \operatorname{curl} b_K \, dx = \frac{2\Pi^0 f_{|K}}{|K|} \int_K \Big((x - x_K)^2 - (y - y_K)^2 \Big) \, dx.$$

Using the identities

$$\int_K (x-x_K)^2 = \tfrac{|K|}{12}|e_{x,K}|^2, \quad \int_K (y-y_K)^2 = \tfrac{|K|}{12}|e_{y,K}|^2 \tag{17}$$

and the definition of $p_{h,2}$ gives

$$\tilde{\beta}_K = \frac{\Pi^0 f_{|K}}{2}\frac{|e_{x,K}|^2 - |e_{y,K}|^2}{|e_{x,K}|^2 + |e_{y,K}|^2}.$$

We have proved that a solution (u_h, p_h) of the box-scheme (BS_{nc}) is also a solution of the problem ((a), (b)), which is unique. This proves the uniqueness of the solution of the box-scheme (BS_{nc}). The linearity and the equality between the number of unknowns and the number of equations permit us to conclude existence and uniqueness of the solution of the box-scheme (BS_{nc}) and its equivalence with the formulation ((a), (b)). This concludes (ii). ∎

The previous result states that the box-scheme (BS_{nc}) is well-posed and equivalent to a single scheme in u_h alone and an explicit reconstruction formula for p_h. More precisely, u_h is the solution of the nonconforming variational formulation for the problem $-\Delta u = \Pi^0 f$. It also generalizes the previous box-scheme (BS2) and addresses the above instability problem. This box-scheme seems to be a generalization on rectangles of the box-scheme $((u_h, p_h) \in P^1_{nc,0} \times RT^0)$ of Courbet and Croisille [6]. Contrary to the triangles case, here the unknowns are not located at the interface of the mesh. Nevertheless, in the particular case of a uniform grid consisting of squares, $\tilde{\beta}_K = 0$ on each K, p_h can be written in the square K as

$$p_{h|K} = (\nabla u_h)_{|K} - \frac{\Pi^0 f_{|K}}{2}\begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix},$$

which is the formulation of p_h in the box-scheme of Courbet and Croisille on triangles.

4.2. Numerical analysis. In this section, we provide the stability and the optimal a priori error estimates for the box-scheme (BS_{nc}). In the rest of the paper, C stands for a constant independent of the mesh.

LEMMA 4.1 (discrete Poincaré lemma). *There exists a constant $C > 0$ dependent only on Ω such that for all $u \in Q^1_{nc,0} + H^1_0(\Omega)$,*

$$|u|_{0,\Omega} \le C|u|_{1,h}.$$

Proof (see [13]). Let $u \in Q^1_{nc,0} + H^1_0(\Omega)$. Then

$$|u|_{0,\Omega} = \sup_{g \in L^2(\Omega)} \frac{|(u,g)_{0,\Omega}|}{|g|_{0,\Omega}}. \tag{18}$$

For $g \in L^2(\Omega)$, there exists $p \in H^1(\Omega)^2$ such that $\operatorname{div} p = g$ and $\|p\|_{1,\Omega} \le C|g|_{0,\Omega}$ [2]. By replacing g by this value in (18) and using Green's formula, we get

$$(u,g)_{0,\Omega} = (u, \operatorname{div} p)_{0,\Omega} = \underbrace{-\sum_K \int_K \nabla u \cdot p \, dx}_{(A)} + \underbrace{\sum_K \int_{\partial K} p \cdot \nu_K \, u \, d\sigma}_{(B)}. \tag{19}$$

First, we obtain $|(A)| = |\sum_K \int_K \nabla u \cdot p \, dx| \le |u|_{1,h} \, |p|_{1,\Omega}$. Let us estimate $|(B)|$. Since $p \in (H^1(\Omega))^2 \cap H_{\mathrm{div}}(\Omega)$,

$$(B) = \sum_K \int_{\partial K} p \cdot \nu_K \, u \, d\sigma = \sum_{a \in \mathcal{A}_b} \int_a p \cdot \nu_a \, u \, d\sigma - \sum_{a \in A_i} \int_a p \cdot \nu_a \, [u]_a \, d\sigma. \tag{20}$$

Let $\overline{p\cdot\nu_a} = \frac{1}{|a|}\int_a p\cdot\nu_a\,d\sigma$ be the mean value of $p\cdot\nu_a$ along the edge a. Since $u\in H_0^1(\Omega)+Q_{nc,0}^1$, by the property of $Q_{nc,0}^1$ to satisfy the *patch-test*, we have

$$\int_a \overline{p\cdot\nu_a}\,u\,d\sigma = 0 \quad \text{for all } a\in\mathcal{A}_b \quad \text{and} \quad \int_a \overline{p\cdot\nu_a}\,[u]_a\,d\sigma = 0 \quad \text{for all } a\in\mathcal{A}_i.$$

Therefore, equality (20) becomes

$$\begin{aligned}\sum_K \int_{\partial K} p\cdot\nu_K\,u\,d\sigma &= \sum_{a\in\mathcal{A}_b}\int_a (p\cdot\nu_a - \overline{p\cdot\nu_a})\,u\,d\sigma - \sum_{a\in\mathcal{A}_i}\int_a (p\cdot\nu_a - \overline{p\cdot\nu_a})\,[u]_a\,d\sigma\\ &= \sum_K\sum_{e\in\partial K}\int_e (p\cdot\nu_e - \overline{p\cdot\nu_e})\,u\,d\sigma.\end{aligned}$$

The lemma of Crouzeix and Raviart [9] gives

$$\left|\int_e (p\cdot\nu_e - \overline{p\cdot\nu_e})\,u\,d\sigma\right| \le C\,h_K|u|_{1,K}\,|p|_{1,K}.$$

Then,

$$|(B)| = \left|\sum_K \int_{\partial K} p\cdot\nu_K\,u\,d\sigma\right| \le 4\,Ch|u|_{1,h}\,|p|_{1,\Omega}.$$

Finally,

$$|(u,g)_{0,\Omega}| \le (4Ch+1)|u|_{1,h}\,|p|_{1,\Omega} \le (4\,Ch+1)|u|_{1,h}\underbrace{\|p\|_{1,\Omega}}_{\le C(\Omega)|g|_{0,\Omega}}. \qquad \square$$

PROPOSITION 4.2 (stability). *The solution $(u_h,p_h)\in Q_{nc,0}^1\times(RT^0+\Phi)$ of the problem (BS_{nc}) satisfies the stability estimate*

$$\|u_h\|_{1,h} + \|p_h\|_{\mathrm{div},h} \le C|f|_{0,\Omega}.$$

Proof. Using the formulation of Proposition 4.1 with $v_h = u_h$ and applying the Cauchy–Schwarz inequality and the Poincaré inequality give

$$\|u_h\|_{1,h} \le C(\Omega)\,|f|_{0,\Omega}.$$

On the other hand, the local formula (b) from Proposition 4.1 for p_h and the identity $\operatorname{div} p_h = -\Pi^0 f$ imply $\|p_h\|_{\mathrm{div},h}\le C|f|_{0,\Omega}$. This concludes the proof. $\square$

PROPOSITION 4.3 (a priori error estimates). *Let $(u,p)\in H_0^1(\Omega)\times H_{\mathrm{div}}(\Omega)$ be the solution of the continuous problem* (1) *and $(u_h,p_h)\in Q_{nc,0}^1\times(RT^0+\Phi)$ be the solution of the box-scheme (BS_{nc}). If $f\in H^1(\Omega)$, we have*

$$\begin{array}{ll}\text{(i)}\quad |u-u_h|_{1,h}\le Ch|f|_{0,\Omega}, & \text{(ii)}\quad |u-u_h|_{0,\Omega}\le Ch^2(|f|_{0,\Omega}+|f|_{1,\Omega}),\\ \text{(iii)}\quad |p-p_h|_{0,\Omega}\le Ch|f|_{0,\Omega}, & \text{(iv)}\quad |p-p_h|_{\mathrm{div},h}\le Ch|f|_{1,\Omega}.\end{array} \tag{21}$$

Proof. (i) Let us introduce the bilinear form a_h defined for all $u,v\in H_0^1(\Omega)+Q_{nc,0}^1$ by

$$a_h(u,v) = \sum_{K\in\mathcal{T}_h}(\nabla u,\nabla v)_{0,K}.$$

Then we obtain the classical inequality

$$|u-u_h|_{1,h} \le 2 \inf_{w_h \in Q^1_{nc,0}} |u-w_h|_{1,h} + \sup_{w_h \in Q^1_{nc,0}} \frac{|a_h(u_h-u,w_h)|}{|w_h|_{1,h}}.$$

The estimation of the consistency error is deduced from the variational formulation from Proposition 4.1:

$$\sup_{w_h \in Q^1_{nc,0}} \frac{|a_h(u_h-u,w_h)|}{|w_h|_{1,h}} \le C\, h\, |f|_{0,\Omega}. \tag{22}$$

By using the Q^1-Lagrange interpolation, we get

$$\inf_{w_h \in Q^1_{nc,0}} |u-w_h|_{1,h} \le Ch|u|_{2,\Omega}.$$

This concludes (i).

(ii) Part (ii) is proved by using the Aubin–Nitsche argument and the result (i).

(iii) Part (iii) is a deduction of the local formula p_h given by Proposition 4.1(ii).

(iv) Part (iv) results from $\operatorname{div} p = -f$ and $\operatorname{div} p_h = -\Pi^0 f$. ∎

4.3. Link to the box-scheme (BS2). We already mentioned that the one-point integration of the gradient of u_h is not sufficient to obtain the stability of the scheme (see section 2.1). Nevertheless, the addition of the local bubble in both trial and test spaces permits us to overcome the previous difficulty, as we have just observed. In this sense, the nonconforming bubble is a stabilization parameter. Moreover, from the decomposition of the space $Q^1_{nc,0}$ given in Proposition 2.1, we deduce the following result.

LEMMA 4.2 (link to the box-scheme (BS2)). *The solution* $(u_h,p_h) \in Q^1_{nc,0} \times (RT^0+\Phi)$ *of the box-scheme* (BS_{nc}) *is given as a function of the solution* $(\bar u_h,\bar p_h) \in Q^1_{c,0} \times RT^0$ *of the box-scheme* (BS2) *by*

$$u_h = \bar u_h + \sum_{K\in\mathcal{T}_h} \alpha_K\, b_K \quad \textit{and} \quad p_h = \bar p_h + \sum_{K\in\mathcal{T}_h} \beta_K \operatorname{curl} b_K,$$

where

$$\alpha_K = \frac{3|K|}{4} \frac{1}{|e_{x,K}|^2+|e_{y,K}|^2} (\bar p_h - \nabla \bar u_h, \nabla b_K)_{0,K},$$
$$\beta_K = -\frac{3|K|}{4} \frac{1}{|e_{x,K}|^2+|e_{y,K}|^2} (\bar p_h - \nabla \bar u_h, \operatorname{curl} b_K)_{0,K}.$$

Proof. Let $(\bar u_h,\bar p_h) \in Q^1_{c,0} \times RT^0$ be the solution of the box-scheme (BS2). We are looking for $(\alpha_K,\beta_K)_{K\in\mathcal{T}_h}$ such that

$$u_h = \bar u_h + \sum_K \alpha_K\, b_K, \qquad p_h = \bar p_h + \sum_K \beta_K \operatorname{curl} b_K$$

define the solution $(u_h,p_h) \in Q^1_{nc,0} \times (RT^0+\Phi)$ of the box-scheme (BS_{nc}). Due to $\operatorname{div} p_h = \operatorname{div} \bar p_h$, the first equation of (BS_{nc}) is valid for p_h. Let us suppose that (u_h,p_h) satisfies the second equation of (BS_{nc}). By the definition of (u_h,p_h) and since $(\bar u_h,\bar p_h)$ satisfies the second equation of (BS2),

$$\sum_{K\in\mathcal{T}_h} (p_h - \nabla u_h, q_h)_{0,K} = 0 \quad \text{for all } q_h \in P^0(\nabla b_K) + P^0(\operatorname{curl} b_K). \tag{23}$$

By taking $q_h = \nabla b_K$ in (23), we get

$$\begin{aligned} 0 &= (p_h - \nabla u_h, q_h)_{0,K} \\ &= (\bar{p}_h - \nabla \bar{u}_h, \nabla b_K)_{0,K} + (\beta_K \operatorname{curl} b_K, \nabla b_K)_{0,K} - (\alpha_K \nabla b_K, \nabla b_K)_{0,K}. \end{aligned}$$

Since $(\operatorname{curl} b_K, \nabla b_K)_{0,K} = 0$, we deduce the formula of α_K on each rectangle K. Then for each K, α_K is uniquely determined by the unique solution $(\bar{u}_h, \bar{p}_h)$ of the box-scheme (BS2). In the same way, by taking $q_h = \operatorname{curl} b_K$ in (23), we get

$$0 = (p_h - \nabla u_h, \operatorname{curl} b_K)_{0,K} = (\bar{p}_h - \nabla \bar{u}_h, \operatorname{curl} b_K)_{0,K} + (\beta_K \operatorname{curl} b_K, \operatorname{curl} b_K)_{0,K}$$

and deduce the formula for β_K. Then with this definition of the coefficients α_K, β_K, we prove that $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ is the unique solution of the box-scheme (BS_{nc}). □

5. A simplified stabilized box-scheme. In this section, we investigate a new way to stabilize the box-scheme (BS1). In fact it seems that the solution of the previous box-scheme is locally in $P^1(K) \times (RT^0(K) + \Phi)$ (see the proof of Proposition 4.1). We are looking for a space locally in $P^1(K)$ instead of $Q^1(K)$ with the same continuity properties as $Q^1_{nc,0}$. The space $\widetilde{M}_{1,h}$ recently introduced by Park and Sheen [19] fulfills those conditions:

$$\widetilde{M}_{1,h} = \left\{ v \in L^2(\Omega);\ v_{|K} \in P^1(K) \text{ for all } K \in \mathcal{T}_h; \right.$$
$$\left. \int_a v_{|K_1}\, dx = \int_a v_{|K_2} dx \text{ for all } a = \partial K_1 \cap \partial K_2 \in \mathcal{A}_i \right\}.$$

Its dimension is $\dim \widetilde{M}_{1,h} = 3NE - NA_i = NV - 1$, since there are three unknowns for each rectangle subject to NA_i independent continuity relations. The corresponding space with homogeneous boundary is

$$\widetilde{M}_{1,h,0} = \left\{ v \in \widetilde{M}_{1,h};\ \int_a v\, dx = 0 \text{ for all } a \in \mathcal{A}_b \right\}.$$

Its dimension is also $\dim \widetilde{M}_{1,h,0} = NA - NE - (NA_b - 1) = NV_i$. Note that this space satisfies the additional condition (7) of Courbet. However, in contrast to the space Q^1_c, it does not contain the nonconforming bubble. The space $\widetilde{M}_{1,h,0}$ is by definition included in $Q^1_{nc,0}$. Similarly to Lemma 2.1, we deduce from the linearity and the injectivity of L and the equality $\dim \widetilde{M}_{1,h,0} = \dim C_0$ the following lemma.

LEMMA 5.1. *The mapping L defines a bijection between $\widetilde{M}_{1,h,0}$ and the Courbet space C_0:*

$$\begin{aligned} L :\ \widetilde{M}_{1,h,0} &\longrightarrow C_0, \\ u &\longmapsto \big(u(x_a)\big)_{a \in \mathcal{A}}. \end{aligned}$$

DEFINITION 5.1. *Let* (BS3) *be the box-scheme: Find $(u_h, p_h) \in \widetilde{M}_{1,h,0} \times (RT^0 + \Phi)$ such that*

$$\text{(BS3)} \quad \begin{cases} \displaystyle\sum_{K \in \mathcal{T}_h} (\operatorname{div} p_h + f, v_h)_{0,K} = 0 & \text{for all } v_h \in P^0, \\ \displaystyle\sum_{K \in \mathcal{T}_h} (p_h - \nabla u_h, q_h)_{0,K} = 0 & \text{for all } q_h \in X_{2,h} = (P^0)^2 + \Phi. \end{cases} \tag{24}$$

The box-scheme has $4NE$ *unknowns.*

Indeed $\dim \widetilde{M}_{1,h,0} + \dim(RT^0 + \Phi) = NV_i + NA + NE - 1 = 4NE = \dim P^0 + \dim X_{2,h}$.

LEMMA 5.2 (link to the box-scheme (BS_{nc})). *The solution* $(\tilde{u}_h, \tilde{p}_h) \in \widetilde{M}_{1,h,0} \times (RT^0 + \Phi)$ *of the box-scheme* (BS3) *is unique and is given as a function of the solution* $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ *of the box-scheme* (BS_{nc}) *by*

$$\tilde{u}_h = u_h \quad \textit{and} \quad \tilde{p}_h = p_h.$$

Proof. Any solution $(\tilde{u}_h, \tilde{p}_h) \in \widetilde{M}_{1,h,0} \times (RT^0 + \Phi)$ of the box-scheme (BS3) is included in $Q^1_{nc,0} \times (RT^0 + \Phi)$ and satisfies (15). By uniqueness of the solution of the box-scheme (BS_{nc}) and the linearity of the scheme (BS3), we deduce the result. ∎

Note that we rediscover that $u_h \in Q^1_{nc,0}$ in the scheme (BS_{nc}) is locally in $P^1(K)$ (see the proof of Proposition 4.1). In particular, this means that the bilinear term "$x\,y$" is not needed. In fact the solution of the box-scheme (BS_{nc}) is already the solution of the box-scheme (BS3).

COROLLARY 5.1. *The box-scheme* (BS3) *has a unique solution* $(u_h, p_h) \in \widetilde{M}_{1,h,0} \times (RT^0 + \Phi)$ *such that*

(a) $u_h \in \widetilde{M}_{1,h,0}$ *is the solution of*

$$\sum_{K \in \mathcal{T}_h} (\nabla u_h, \nabla v_h)_{0,K} = (\Pi^0 f, v_h)_{0,\Omega} \quad \textit{for all } v_h \in \widetilde{M}_{1,h,0};$$

(b) p_h *is locally given by*

$$p_{h|K} = (\nabla u_h)_{|K} - \frac{\Pi^0 f_{|K}}{|e_{x,K}|^2 + |e_{y,K}|^2} \begin{pmatrix} |e_{y,K}|^2 (x - x_K) \\ |e_{x,K}|^2 (y - y_K) \end{pmatrix}.$$

Proof. This result is deduced from Lemma 5.2 and Proposition 4.1, since $\widetilde{M}_{1,h,0} \subset Q^1_{nc,0}$. ∎

COROLLARY 5.2 (a priori error estimates). *Let* $(u, p) \in H^1_0(\Omega) \times H_{\mathrm{div}}(\Omega)$ *be the solution of the continuous problem* (1) *and* $(u_h, p_h) \in \widetilde{M}_{1,h,0} \times (RT^0 + \Phi)$ *be the solution of the box-scheme* (BS3). *If* $f \in H^1(\Omega)$, *we have*

(i) $|u - u_h|_{1,h} \le Ch|f|_{0,\Omega}$, (ii) $|u - u_h|_{0,\Omega} \le Ch^2(|f|_{0,\Omega} + |f|_{1,\Omega})$,

(iii) $|p - p_h|_{0,\Omega} \le Ch|f|_{0,\Omega}$, (iv) $|p - p_h|_{\mathrm{div},h} \le Ch|f|_{1,\Omega}$.

6. Numerical results. In this section we present several numerical results which demonstrate the theoretical convergence rates obtained for the box-scheme of section 4. We compute the error estimates for the unknown u and the flux p of the box-scheme (BS_{nc}) on two different domains Ω meshed by rectangles. The solution of the box-scheme (BS_{nc}) is computed according to the decoupled formulation given in Proposition 4.1. The unknown u is the solution of the variational formulation, whereas p is deduced from the local reconstruction on each rectangle. From the computed error, we deduce the numerical convergence rate for each solution of each test. The results for each test case are reported in Tables 1–4.

Test cases 1 and 2 of section 6.1 are given on the unit square domain $\Omega = [0,1]^2$ meshed by squares, whereas section 6.2 is devoted to the computation of the error estimates of the box-scheme (BS_{nc}) on $\Omega = [0,1]^2$ meshed by rectangles. Finally, in section 6.3, we present one test case on the L-shaped domain $\Omega = [0,2] \times [0,1] \cup [1,2] \times [1,2]$, meshed by squares. All the computed convergence rates are in agreement with the theoretical ones given in Proposition 4.3.

TABLE 1
Box-scheme (BS_{nc}): $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ *for Test* 1.

Nb rect.	$\lvert u - u_h\rvert_{0,\Omega}$	$\lvert u - u_h\rvert_{1,h}$	$\lvert p - p_h\rvert_{0,\Omega}$	Space step h
100	2.261×10^{-3}	7.567×10^{-2}	7.976×10^{-2}	0.1414
225	1.008×10^{-3}	5.053×10^{-2}	5.326×10^{-2}	0.09428
400	5.677×10^{-4}	3.792×10^{-2}	3.997×10^{-2}	0.07071
900	2.525×10^{-4}	2.529×10^{-2}	2.665×10^{-2}	0.04714
Conv. rate	1.996	0.9977	0.9979	

TABLE 2
Box-scheme (BS_{nc}): $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ *for Test* 2.

Nb rect.	$\lvert u - u_h\rvert_{0,\Omega}$	$\lvert u - u_h\rvert_{1,h}$	$\lvert p - p_h\rvert_{0,\Omega}$	Space step h
100	3.927×10^{-2}	0.9945	1.035	0.1414
225	1.990×10^{-2}	0.6885	0.7112	0.09428
400	1.174×10^{-2}	0.5148	0.5333	0.07071
900	5.401×10^{-3}	0.3422	0.3553	0.04714
Conv. rate	1.808	0.9737	0.9751	

6.1. Square domain meshed by squares. The domain Ω is meshed by four different regular grids made of 100, 225, 400, and 900 squares.

1. *Test case* 1: In this first example, the source term f and the Dirichlet data g are chosen such that $u(x, y) = x\,(1 - x)\sin(\pi\, y)$ is the exact solution of the Poisson problem

$$\begin{cases} -\Delta u = f \quad \text{on} \quad \Omega, \\ u = g \quad \text{on} \quad \partial\Omega. \end{cases} \tag{25}$$

   The results for the box-scheme $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ defined by (BS_{nc}) are given in Table 1. The error for the unknown u is of order 1 in the seminorm $|\cdot|_{1,h}$ and of order 2 for the L^2-norm. For p we get also order 1 in the L^2-norm. The numerical results are of order of those computed theoretically in Proposition 4.3.
2. *Test case* 2: Our second example is a test case proposed by Arnold, Boffi, and Falk [1]. The source term and the boundary conditions are chosen such that $u(x, y) = \exp\big(-100\,((x - 1/4)^2 + (y - 1/3)^2)\big)$ is the exact solution of problem (25). It concerns a Gaussian pulse centered at the point $(x_0, y_0) = (\frac{1}{4}, \frac{1}{3})$. The error estimates for both unknowns u and $p = \nabla u$ are given in Table 2 for the box-scheme $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$. The convergence rates are a little lower than expected (1.8 instead of 2 for u in the L^2-norm and 0.97 instead of 1 for p in the L^2-norm) but still close to the a priori error estimates of Proposition 4.3. This is due to the high gradient of the exact solution at the point (x_0, y_0).

6.2. Test case 3: Square domain meshed by rectangles. In this example, we consider the domain $\Omega = [0, 1]^2$ meshed by rectangles and the solution $u(x, y) = x\,(1 - x)\,y\,(1 - y)\exp(5\,x)$ of the problem (25), where the right-hand side and the Dirichlet conditions are computed using the exact solution u. The grid is made of $n_x \times n_y$ rectangles, where n_x and n_y are the numbers of subdivisions of the segment $[0, 1]$ in each direction (O_x) and (O_y). We compute the solution (u_h, p_h) for (n_x, n_y)

TABLE 3
Box-scheme (BS_{nc}): $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ *for Test* 3.

Nb rect.	$\|u - u_h\|_{0,\Omega}$	$\|u - u_h\|_{1,h}$	$\|p - p_h\|_{0,\Omega}$	Space step h
20×5	3.386×10^{-2}	1.979	1.586	0.2061
40×10	8.488×10^{-3}	1.001	0.7991	0.1031
80×20	2.124×10^{-3}	0.5020	0.4003	0.05154
100×25	1.359×10^{-3}	0.4017	0.3203	0.04123
Conv. rate	1.998	0.9911	0.9942	

TABLE 4
Box-scheme (BS_{nc}): $(u_h, p_h) \in Q^1_{nc,0} \times (RT^0 + \Phi)$ *for Test* 4.

Nb rect.	$\|u - u_h\|_{0,\Omega}$	$\|u - u_h\|_{1,h}$	$\|p - p_h\|_{0,\Omega}$	Space step h
75	3.368×10^{-2}	0.5089	0.5215	0.2828
300	8.613×10^{-3}	0.2543	0.2616	0.1414
675	3.813×10^{-3}	0.1688	0.1739	0.09428
Conv. rate	1.981	1.004	0.9993	

taking the values $(20, 5)$, $(40, 10)$, $(80, 20)$, and $(100, 25)$, i.e., 100, 400, 1600, and 2500 rectangles. The exact solution presents a boundary layer at $x = 1$. Nevertheless, the computed solution u_h and the discrete flux p_h of (BS_{nc}) seem to take it into account. The convergence rates between the exact and the discrete solutions for both unknowns u and $p = \nabla u$ are assembled in Table 3. The numerical results really satisfy the theoretical estimates of Proposition 4.3.

6.3. Test case 4: Test case on an L-shaped domain. In this case we consider a different domain Ω_L, given by the square $[0, 2] \times [0, 2]$ without the part $[0, 1] \times [1, 2]$. We obtain an L-shaped domain. We compute the solution (u_h, p_h) of the box-scheme (BS_{nc}) associated with the Poisson problem (25). The data f and g are chosen such that $u(x, y) = x\,(2 - x)\,y^2\,(2 - y)\sin(x + 2y)$ is the exact solution of (25). The convergence rate and error for both unknowns u and $p = \nabla u$ are given in Table 4. The computed results conform to the theoretical ones.

7. Conclusion. We have presented three different box-schemes which are in fact strongly connected to each other through the initial box-scheme introduced by Courbet. The box-scheme (BS3) presents the advantage of giving a completely local formulation of the flux, which is not the case of the box-scheme (BS2). The box-scheme (BS_{nc}) gives the same solution as the box-scheme (BS3) but seems to be the most stable of all the schemes of the paper. The choice of the trial and test spaces erases the effect of the hourglass mode of the initial box-scheme. In particular, the box-scheme (BS_{nc}) is probably the one to implement for generalized quadrangles in two and three dimensions. The box-schemes can be generalized to the Poisson problem with a diffusion tensor $\mathcal{K}$. In this case, we would consider for the flux $p = \mathcal{K} \cdot \nabla u$. Particularly, the box-scheme method is adapted to the direct computation of the speed of the flow in the Darcy law.

Appendix. In this section, we are going to complete the proof of Proposition 3.1. Actually, we determine the coefficient γ_K of formula (11). For this purpose, we consider the H_{div}-property of elements of RT^0 to have their normal component continuous along the internal edges of the mesh. Indeed, for any internal edge $a = \partial K \cap \partial K'$,

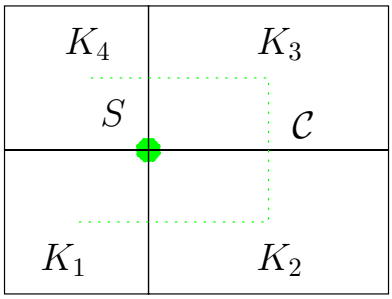


FIG. 4. *Path* $\mathcal{C}$.

$p_h \in RT^0$ satisfies

$$p_{h_{|K}} \cdot \nu_a + p_{h_{|K'}} \cdot \nu'_a = 0 \qquad \text{along the edge } a.$$

At first, we consider a grid made of four elements K_1, K_2, K_3, and K_4 and a path $\mathcal{C}$ crossing each element K_i for $i = 1, \ldots, 4$ once and only once as pictured in Figure 4.

Let a_1, a_2, a_3, and a_4 be the four internal edges $a_1 = \partial K_1 \cap \partial K_2$, $a_2 = \partial K_2 \cap \partial K_3$, $a_3 = \partial K_3 \cap \partial K_4$, and $a_4 = \partial K_1 \cap \partial K_4$. Let S be the common vertex of K_1, K_2, K_3, and K_4.

LEMMA A.1. *Let* u_h *be the solution of* (10) *and* p_h *be given on each rectangle* K *by the relation* (14), *where* γ_K *has to be defined. The continuity of* $p_h \cdot \nu_a$ *along the edges* $a = a_1, a_2, a_3$ *crossing the path* $\mathcal{C}$ *implies the continuity of* $p_h \cdot \nu_{a_4}$ *along the edge* a_4.

Proof. Considering the edges a_1, a_2, a_3 and the continuity of $p_h \cdot \nu_a$ along these edges, we get the following system:

$$\begin{cases} |K_2|\gamma_{K_2} = -|K_1|\gamma_{K_1} + F(u_h, f, a, K_1, K_2), \\ -|K_3|\gamma_{K_3} = -|K_1|\gamma_{K_1} + F(u_h, f, a_1, K_1, K_2) - G(u_h, f, a_2, K_2, K_3), \\ |K_4|\gamma_{K_4} = -|K_1|\gamma_{K_1} + F(u_h, f, a_3, K_4, K_3) \\ \qquad + F(u_h, f, a_1, K_1, K_2) - G(u_h, f, a_2, K_2, K_3), \end{cases} \tag{26}$$

where F is defined for the rectangles K_1 and K_2 by

$$F(u_h, f, a_1, K_1, K_2) = -2|a_1|\Big(\Pi^0(\nabla u_h)_{|K_1} - \Pi^0(\nabla u_h)_{|K_2}\Big) \cdot \begin{pmatrix} 1 \\ 0 \end{pmatrix} + \frac{|K_1|}{2}\Pi^0 f_{|K_1} + \frac{|K_2|}{2}\Pi^0 f_{|K_2}.$$

The same formula holds for the rectangles K_4 and K_3. The function G is defined by

$$G(u_h, f, a_2, K_2, K_3) = 2|a_2|\Big((\Pi^0\nabla u_h)_{|K_2} - (\Pi^0\nabla u_h)_{|K_3}\Big) \cdot \begin{pmatrix} 0 \\ 1 \end{pmatrix} - \frac{|K_2|}{2}\Pi^0 f_{|K_2} - \frac{|K_3|}{2}\Pi^0 f_{|K_3}.$$

The system (26) leads to the following relation between the rectangles K_1 and K_4:

$$
\begin{aligned}
|K_4|\gamma_{K_4} = &-|K_1|\gamma_{K_1} - 2\Big[|a_3|\Big(\Pi^0(\nabla u_h)_{|K_4} - \Pi^0(\nabla u_h)_{|K_3}\Big) \\
&+ |a_1|\Big(\Pi^0(\nabla u_h)_{|K_1} - \Pi^0(\nabla u_h)_{|K_2}\Big)\Big] \cdot \begin{pmatrix}1\\0\end{pmatrix} \\
&- 2|a_2|\Big((\Pi^0\nabla u_h)_{|K_2} - (\Pi^0\nabla u_h)_{|K_3}\Big) \cdot \begin{pmatrix}0\\1\end{pmatrix} \\
&+ \sum_K \frac{|K|}{2}\Pi^0 f_{|K} - \frac{|K_4|}{2}\Pi^0 f_{|K_4} - \frac{|K_1|}{2}\Pi^0 f_{|K_1}.
\end{aligned} \tag{27}
$$

Using the variational form (10) whose solution is u_h and taking v_h to be the basis function from $Q^1_{c,0}$ associated to the vertex S (Figure 4), we get

(28)

$$
\begin{aligned}
&\Big[|a_1|\Big(\Pi^0(\nabla u_h)_{K_1} - \Pi^0(\nabla u_h)_{K_2}\Big) + |a_3|\Big(\Pi^0(\nabla u_h)_{K_4} - \Pi^0(\nabla u_h)_{K_3}\Big)\Big] \cdot \begin{pmatrix}1\\0\end{pmatrix} \\
&\quad + \Big[|a_2|\Big(\Pi^0(\nabla u_h)_{K_2} - \Pi^0(\nabla u_h)_{K_3}\Big) + |a_4|\Big(\Pi^0(\nabla u_h)_{K_1} - \Pi^0(\nabla u_h)_{K_4}\Big)\Big] \cdot \begin{pmatrix}0\\1\end{pmatrix} \\
&= \sum_{i=1}^{4} \frac{|K_i|}{2}(\Pi^0 f)_{K_i}.
\end{aligned}
$$

After substituting equality (28) into (27), we get

$$
\begin{aligned}
|K_4|\gamma_{K_4} = &-|K_1|\gamma_{K_1} + 2|a_4|\Big(\Pi^0(\nabla u_h)_{K_1} - \Pi^0(\nabla u_h)_{K_4}\Big) \cdot \begin{pmatrix}0\\1\end{pmatrix} \\
&- \frac{|K_4|}{2}\Pi^0 f_{|K_4} - \frac{|K_1|}{2}\Pi^0 f_{|K_1},
\end{aligned}
$$

which is exactly the continuity of $p_h \cdot \nu_{a_4}$ along the edge a_4. This concludes the proof of the lemma. □

The result for four rectangles is extended recursively to a rectangle domain Ω with a grid of rectangles (see Figure 5) as follows.

LEMMA A.2. *Let u_h be the solution of* (10) *and p_h be given on each rectangle K by the relation* (14)*, where γ_K has to be defined. Let $\mathcal{C}$ be a path covering all the rectangles of the domain Ω. The continuity of $p_h \cdot \nu_a$ along each edge a crossing the path $\mathcal{C}$ (Figure* 5*) is equivalent to the continuity of $p_h \cdot \nu_a$ along each internal edge of the domain Ω.*

On the other hand, we consider the choice $q_h = \sum_K \operatorname{sgn}(K) \operatorname{curl} b_K$ in (9). Since the scalar product between ∇u_h and $\operatorname{curl} b_K$ (from (8)) is null on each rectangle K, we get

$$
0 = \sum_K (p_h, q_h)_{0,K} = \sum_K \int_K p_h \cdot \operatorname{sgn}(K) \operatorname{curl} b_K \, dx.
$$

Using the formula (14) for p_h,

$$
\sum_K \int_K \left[(\Pi^0\nabla u_h)_{|K} - \frac{\Pi^0 f_{|K}}{2}\begin{pmatrix}x - x_K\\ y - y_K\end{pmatrix} + \gamma_K \begin{pmatrix}x - x_K\\ -(y - y_K)\end{pmatrix}\right] \cdot \operatorname{sgn}(K) \operatorname{curl} b_K \, dx = 0,
$$

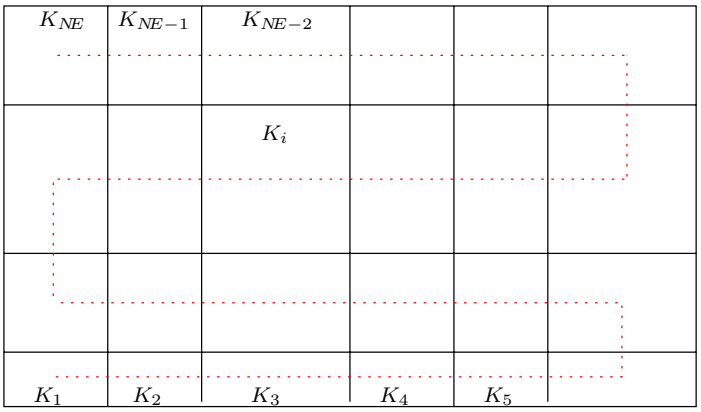


FIG. 5. *Path $\mathcal{C}$ for the whole grid $\mathcal{T}_h$ of the domain Ω.*

and $\int_K \Pi^0(\nabla u_h)\cdot \operatorname{curl} b_K\, dx = 0$ (property of the bubble), and the identities (17), we deduce

$$\sum_K \gamma_K \operatorname{sgn}(K)(|e_{x,K}|^2 + |e_{y,K}|^2) = \sum_K \operatorname{sgn}(K)\frac{\Pi^0 f_{|K}}{2}(|e_{x,K}|^2 - |e_{y,K}|^2). \tag{29}$$

Finally, we can deduce a linear system with NE equations and NE unknowns $(\gamma_1, \gamma_2, \ldots, \gamma_{NE})$. It is given by the $(NE-1)$ continuity conditions of $p_h\cdot\nu_a$ along each internal edge a, crossing the path $\mathcal{C}$ and (29).

PROPOSITION A.1. *Let u_h be the solution of* (10) *and p_h be given on each rectangle K by the relation* (14)*, where γ_K has to be defined on each rectangle. The vector $\gamma = (\gamma_1, \gamma_2, \ldots, \gamma_{NE})$ is a solution of the system $A\gamma = b$, where N denotes NE and*

$$A = \begin{pmatrix} |K_1| & |K_2| & 0 & \cdots & 0 \\ 0 & |K_2| & 0 & \cdots & 0 \\ \vdots & & \ddots & \ddots & \\ & 0 & & & 0 \\ & & \ddots & \ddots & \\ 0 & \cdots & & & 0 \\ 0 & \cdots & 0 & |K_{N-1}| & |K_N| \\ |e_{x,K_1}|^2 + |e_{y,K_1}|^2 & -(|e_{x,K_2}|^2 + |e_{y,K_2}|^2) & \cdots & \cdots & \operatorname{sgn}(K_N)(|e_{x,K_N}|^2 + |e_{y,K_N}|^2) \end{pmatrix},$$

$$b = \begin{pmatrix} H(u_h, f, |a_1|, K_1, K_2) \\ H(u_h, f, |a_2|K_2, K_3) \\ \vdots \\ \vdots \\ H(u_h, f, |a_{N-1}|, K_{N-2}, K_{N-1}) \\ \sum_K \operatorname{sgn}(K)\frac{\Pi^0 f_{|K}}{2}(|e_{x,K}|^2 - |e_{y,K}|^2) \end{pmatrix},$$

H is either F or G according to the cases, and a_i is the common edge to the rectangles K_i and K_{i+1} (see Figure 5*).*

Proof. We prove this result in the case of a uniform grid, i.e., where the rectangles have the same size, $e_x := e_{x,K}$, and $e_y := e_{y,K}$ for all rectangles K. Then, the determinant of the matrix A is

$$\det A = |K_2|\cdots|K_{NE-1}|\Big(|e_x|^2 + |e_y|^2\Big)\Big(\operatorname{sgn}(K_{NE})|K_1| + (-1)^{NE+1}\operatorname{sgn}(K_1)|K_{NE}|\Big).$$

Also the sign of the last rectangle seen by the path $\mathcal{C}$ is $\operatorname{sgn}(K_{NE}) = (-1)^{NE+1}$ and $\operatorname{sgn}(K_1) = 1$. Therefore, $|\det A| = |K_2| \cdots |K_{NE-1}| \big(|e_x|^2 + |e_y|^2\big)\big(|K_1| + |K_{NE}|\big)$ is different from zero.

Then, γ_K is given in each rectangle by the resolution of the system $A\gamma = b$ and p_h can be written in each rectangle as a function of γ_K (which depends on the neighboring rectangles):

$$p_{h|K} = \Pi^0(\nabla u_h) - \frac{\Pi^0 f_{|K}}{2} \begin{pmatrix} x - x_K \\ y - y_K \end{pmatrix} + \gamma_K \begin{pmatrix} x - x_K \\ -(y - y_K) \end{pmatrix}. \qquad \square$$

Resulting from this work on the path we can deduce the following characterization of the space Q_c^1.

LEMMA A.3 (characterization of Q_c^1). *Let u_h be such that $u_{h|K} \in Q^1(K)$ for all rectangles K of the grid $\mathcal{T}_h$. Then $u_h \in Q_c^1$ if and only if u_h is continuous at the midpoint of the internal edges and $\nabla u_h \cdot \tau$ is continuous along the path $\mathcal{C}$, where the path is crossing once and only once all the rectangles of the grid. This means*

$$Q_c^1 = \Big\{ u_h / u_{h|K} \in Q^1(K) \text{ for all } K \in \mathcal{T}_h,\ \int_a [u_h]\, d\sigma = 0 \text{ for all } a \in \mathcal{A}_i, \\ \nabla u_h \cdot \tau_a \text{ is continuous along the path } \mathcal{C} \Big\}.$$

Its dimension is $\dim Q_c^1 = 4NE - NA_i - (NE - 1) = NP$.

Proof. Let Q be defined by

$$Q = \Big\{ u_h \in Q^1(K) \text{ for all } K \in \mathcal{T}_h,\ \int_a [u_h]\, d\sigma = 0 \text{ for all } a \in \mathcal{A}_i, \\ \nabla u_h \cdot \tau_a \text{ is continuous along the path } \mathcal{C} \Big\}.$$

The inclusion $Q_c^1 \subseteq Q$ is clear. Let us prove that $Q \subseteq Q_c^1$. Again, we can restrict the proof to the case of four rectangles (see Figure 4). Let us suppose that u_h is continuous at the midpoint of the four edges a_1, a_2, a_3, and a_4 and that $\nabla u_h \cdot \tau_{a_i}$ is continuous along the edges a_i for $i = 1, 2, 3$. We deduce algebraically that $\nabla u_h \cdot \tau_{a_4}$ is continuous along the edge a_4. On the other hand, the continuity of u_h at the midpoint of the internal edges and the continuity of $\nabla u_h \cdot \tau_a$ at the internal edges imply the continuity of u_h at the interfaces of the mesh, which means the continuity of u_h on the whole domain. Then we get $Q \subseteq Q_c^1$, which concludes the lemma. $\square$

Acknowledgments. I am grateful to Professor J.-P. Croisille for fruitful discussions and suggestions concerning this work. I would like to thank the Max Planck Institute for Mathematics in the Sciences in Leipzig for providing the opportunity of finishing the present work.

REFERENCES

[1] D. N. ARNOLD, D. BOFFI, AND R. S. FALK, *Approximation by quadrilateral finite elements*, Math. Comp., 71 (2002), pp. 909–922.

[2] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts Appl. Math. 15, Springer-Verlag, New York, 2002.
[3] S.-H. Chou, D. Y. Kwak, and K. Y. Kim, *Mixed finite volume methods on nonstaggered quadrilateral grids for elliptic problems*, Math. Comp., 72 (2003), pp. 525–539.
[4] S.-H. Chou and S. Tang, *Comparing two approaches of analyzing mixed finite volume methods*, J. Korea Society for Industrial and Applied Mathematics, 5 (2001), pp. 55–78.
[5] B. Courbet, *Two-point schemes for computational fluid dynamics*, Rech. Aérospat., 5 (1990), pp. 21–46.
[6] B. Courbet and J.-P. Croisille, *Finite volume box schemes on triangular meshes*, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 631–649.
[7] J.-P. Croisille, *Finite volume box-schemes and mixed methods*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1087–1106.
[8] J.-P. Croisille and I. Greff, *Some nonconforming mixed box schemes for elliptic problems*, Numer. Methods Partial Differential Equations, 18 (2002), pp. 355–373.
[9] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations.* I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–75.
[10] L. El Alaoui, *Analyse d'erreur a priori et a posteriori pour des méthodes d'éléments finis mixtes non-conforme*, Ph.D. thesis, Ecole Nationale des Ponts et Chaussées, Paris, France, 2005.
[11] L. El Alaoui and A. Ern, *Residual and hierarchical a posteriori error estimates for nonconforming mixed finite element methods*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 903–929.
[12] V. Girault, *Problèmes variationnels et méthodes d'élements finis*, Notes de cours de DEA, Université Paris 6, Paris, France, 1995.
[13] I. Greff, *Schémas boîte: Etude théorique et numérique*, Ph.D. thesis, Université de Metz, Metz, France, 2003, http://www.univ-pau.fr/~igreff.
[14] I. Greff, *Some box-schemes for elliptic problems on rectangular meshes*, in Proceedings of LUXFEM03, Centre de Recherche Public, Henri Tudor, Luxembourg, 2003.
[15] I. Greff, *An Interpretation of the Raviart-Thomas Flux on Rectangular Grids via Box-Scheme*, in preparation, 2007.
[16] P. Hansbo, *A new approach to quadrature for finite elements incorporating hourglass control as a special case*, Comput. Methods Appl. Mech. Engrg., 158 (1998), pp. 301–309.
[17] H. B. Keller, *A new difference scheme for parabolic problems*, in Numerical Solution of Partial Differential Equations, II (SYNSPADE 1970), Academic Press, New York, 1971, pp. 327–350.
[18] L. D. Marini, *An inexpensive method for the evaluation of the solution of the lowest order Raviart–Thomas mixed method*, SIAM J. Numer. Anal., 22 (1985), pp. 493–496.
[19] C. Park and D. Sheen, P_1*-nonconforming quadrilateral finite element methods for second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 624–640.
[20] P.-A. Raviart and J. M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods, Lecture Notes in Math. 606, Springer-Verlag, Berlin, 1977, pp. 292–315.
[21] J.-M. Thomas and D. Trujillo, *Mixed finite volume methods*, Internat. J. Numer. Methods Engrg., 46 (1999), pp. 1351–1366.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 969–985

ON THE CONTRACTIVITY AND CONVERGENCE OF GENERAL LINEAR METHODS ON SEMI-INFINITE INTERVALS*

S. GONZÁLEZ-PINTO† AND D. HERNÁNDEZ-ABREU†

Abstract. The strict-contractivity and the convergence of General Linear Methods on the classes of strictly dissipative and dissipative differential systems regarding some inner product are analyzed. New convergence and contractivity results of the methods on semi-infinite intervals are provided for the case of strictly dissipative problems. Some applications of the main results to the class of Runge–Kutta multistep methods are supplied.

Key words. General Linear Methods, contractivity, convergence, dissipative differential systems

AMS subject classifications. 65L20, 65L06, 65L05

DOI. 10.1137/060657467

1. Introduction.

Let us consider initial value problems

$$y' = f(t,y), \quad y(0) = y_0, \qquad f : [0,\infty) \times \mathbb{C}^m \to \mathbb{C}^m, \tag{1.1}$$

where the function f is supposed to be continuously differentiable on an appropriate domain and it is also assumed to satisfy the one-sided Lipschitz condition

$$\mathrm{Re}\langle f(t,y) - f(t,z), y - z\rangle_X \le \nu \, \|y - z\|_X^2 \quad \forall t \ge 0 \; \forall y, z, \tag{1.2}$$

regarding some inner product in $\mathbb{C}^m$, where ν is some constant. It would suffice that (1.2) were satisfied in some cylinder around the exact solution $y(t)$ which will be assumed to exist for $t \ge 0$. The inner product is defined by

$$\langle u, v\rangle_X := \sum_{i,j=1}^{m} x_{ij} \overline{v}_j u_i; \quad u = (u_j)_{j=1}^m, \; v = (v_j)_{j=1}^m, \tag{1.3}$$

where the matrix $X = (x_{ij})_{i,j=1}^m \in \mathbb{R}^{m,m}$ is symmetric and positive definite, and $\bar{z}$ denotes the conjugate of a complex number z. The notation $\|\cdot\|_X$ will be used throughout this paper for the norm associated to that inner product. It is well known that under the previous assumptions, the difference between two solutions with initial values y_0 and z_0, respectively, satisfies [8, Ch. I]

$$\|y(t;0,y_0) - y(t;0,z_0)\|_X \le \exp(\nu t) \, \|y_0 - z_0\|_X \quad \forall t \ge 0.$$

Thus, when the constant ν is negative (strictly dissipative case), besides strict-contractivity for neighboring solutions, we have asymptotic stability (in the Lyapunov sense) for any solution defined on the interval $[0, +\infty)$. It would be desirable that in such a situation, the numerical methods applied to (1.1) also preserve the asymptotic stability in the Lyapunov sense for their numerical solutions.

*Received by the editors April 18, 2006; accepted for publication (in revised form) November 29, 2006; published electronically May 4, 2007.
http://www.siam.org/journals/sinum/45-3/65746.html

†Departamento de Análisis Matemático, Universidad de La Laguna, 38208, La Laguna, Tenerife, Canary Islands, Spain (spinto@ull.es, dhabreu@ull.es). The work of the second author was supported by project MTM2004-06466-C02-02 and grant AP2002-2761.

The goal of this paper is to study the strict-contractivity and the convergence (and B-convergence) of General Linear Methods on semi-infinite intervals for strictly dissipative differential systems, i.e., $\nu < 0$ in (1.2). Some of the B-convergence results presented here are similar to those by Huang, Chang, and Xiao [12] for the case of finite intervals and dissipative problems ($\nu = 0$ in (1.2)). However, our results also apply to semi-infinite intervals, whereas those results given in [12] do not. The strict-contractivity analysis extends the ideas given by Hairer and Zennaro [11] for implicit Runge–Kutta methods, and the superexponential character of the linear stability function associated to a General Linear Method is exploited in the same way as in [11]. The convergence results, which make use of the previous contractivity analysis and are partly inspired by the work of Hundsdorfer [14], are postponed to the last part of the paper. Applications of our results to some classes of General Linear Methods appearing in the literature are also given.

The paper is organized as follows. In section 2, some preliminaries about General Linear Methods and the norms to be used are given, and some tools to prove the main results are provided. In section 3, the strict-contractivity analysis is carried out. Section 4 is devoted to the existence and uniqueness of the stage solutions. The convergence studies are developed in section 5, and some applications of the main results are postponed to section 6.

2. Preliminary results. A k-step s-stage General Linear Method (see, e.g., [3], [10, p. 356]) applied to (1.1) is given by

$$\begin{array}{rcl} Y^{(n+1)} & = & h(A \otimes I_m)F(t_n, h, V^{(n)}) + (B \otimes I_m)Y^{(n)}, \quad n \geq 0, \\ V^{(n)} & = & h(\tilde{A} \otimes I_m)F(t_n, h, V^{(n)}) + (\tilde{B} \otimes I_m)Y^{(n)}, \end{array} \tag{2.1}$$

where $h > 0$ denotes the step-size, $V^{(n)} := ({v_1^{(n)}}^T, {v_2^{(n)}}^T, \ldots, {v_s^{(n)}}^T)^T \in \mathbb{C}^{ms}$ stands for the stages of the method, $Y^{(n+1)} := ({y_1^{(n+1)}}^T, {y_2^{(n+1)}}^T, \ldots, {y_k^{(n+1)}}^T)^T \in \mathbb{C}^{mk}$ is the advancing solution, $F(t_n, h, V^{(n)}) := (f(t_n + c_1 h, v_1^{(n)})^T, f(t_n + c_2 h, v_2^{(n)})^T, \ldots, f(t_n + c_s h, v_s^{(n)})^T)^T$, and $Y^{(n)}$ denotes a piece of information about the true solution known from the previous step. The method possesses as free parameters the matrices $\tilde{A} \in \mathbb{R}^{s,s}$, $\tilde{B} \in \mathbb{R}^{s,k}$, $A \in \mathbb{R}^{k,s}$, and $B \in \mathbb{R}^{k,k}$ and the vector $c := (c_1, \ldots, c_s)^T \in \mathbb{R}^s$, and they can be chosen in order to gain both stability and global stage order. Above $\otimes$ stands for the usual Kronecker product of matrices, $A \otimes B = (a_{ij}B)$, and I_m for the identity matrix of dimension m.

The difference $\Delta Y^{(n+1)} := \hat{Y}^{(n+1)} - Y^{(n+1)}$ between two numerical solutions provided by the method (2.1) when applied to (1.1) in a fixed step-size setting fulfills the recurrent relation

$$\Delta Y^{(n+1)} = M(Z^{(n)})\Delta Y^{(n)}, \quad n \geq 0, \quad Z^{(n)} = hJ^{(n)}, \tag{2.2}$$

with

$$J^{(n)} := \text{BlockDiag}(J_1^{(n)}, \ldots, J_s^{(n)}) \in \mathbb{C}^{ms,ms}, \tag{2.3}$$

and where $J_i^{(n)} \in \mathbb{C}^{m,m}$, $1 \leq i \leq s$, are matrices given by

$$J_i^{(n)} := \int_0^1 \frac{\partial f}{\partial y}(t_n + c_i h, v_i^{(n)} + \theta(\hat{v}_i^{(n)} - v_i^{(n)}))d\theta. \tag{2.4}$$

Besides, $M(Z) \in \mathbb{C}^{mk,mk}$ is the supermatrix given by

$$M(Z) = B \otimes I_m + (A \otimes I_m)Z(I_{sm} - (\tilde{A} \otimes I_m)Z)^{-1}(\tilde{B} \otimes I_m). \tag{2.5}$$

The linear stability analysis of General Linear Methods is carried out by considering the basic linear complex test problem $y' = \lambda y$, with $\lambda \in \mathbb{C}$. In such a case, the advancing solution satisfies $Y^{(n+1)} = R(z)Y^{(n)}$, $z = \lambda h$, $n \geq 0$, where $R(z)$ is the so-called *stability matrix*

$$R(z) := B + zA(I - z\tilde{A})^{-1}\tilde{B}. \tag{2.6}$$

Then, (2.1) is said to be A-stable if the eigenvalues of $R(z)$ satisfy the *root condition*, i.e., they have either modulus smaller than one or modulus one and they are simple, provided that $z \in \mathbb{C}^- := \{z \in \mathbb{C}, \operatorname{Re} z \leq 0\}$.

On the other hand, the nonlinear stability properties of the methods are studied on dissipative nonlinear systems ($\nu = 0$ in (1.2)). Thus, a General Linear Method is said to be G-stable if there exists a symmetric and positive definite matrix $G = (g_{ij})_{i,j=1}^k \in \mathbb{R}^{k,k}$ such that for two arbitrary advancing solutions we have that

$$\left\|\hat{Y}^{(n+1)} - Y^{(n+1)}\right\|_{G\otimes X} \leq \left\|\hat{Y}^{(n)} - Y^{(n)}\right\|_{G\otimes X},$$

where $\|Y^{(n)}\|_{G\otimes X} := \sum_{i,j=1}^k g_{ij}\langle Y_i^{(n)}, Y_j^{(n)}\rangle_X$. Burrage and Butcher proved in [3] that *algebraic stability* is enough for G-stability. We recall that a General Linear Method is said to be *algebraically stable* if there exist a nonnegative definite matrix $D = \operatorname{Diag}(d_1, \ldots, d_s)$ and a positive definite matrix $G \in \mathbb{R}^{k,k}$ such that

$$N = \begin{pmatrix} G - B^T G B & \tilde{B}^T D - B^T G A \\ D\tilde{B} - A^T G B & D\tilde{A} + \tilde{A}^T D - A^T G A \end{pmatrix} \in \mathbb{R}^{k+s,k+s} \tag{2.7}$$

is nonnegative definite. Moreover, both concepts G-stability and algebraic stability turn out to be equivalent for most of the General Linear Methods of interest. Thus, for a nonconfluent General Linear Method ($c_i \neq c_j \ \forall i \neq j$) which is preconsistent, i.e., there exists a vector $\xi_0 \in \mathbb{R}^k$ such that

$$B\xi_0 = \xi_0, \qquad \tilde{B}\xi_0 = e := (1, \ldots, 1)^T \in \mathbb{R}^s, \tag{2.8}$$

G-stability turns out to be equivalent to algebraic stability [4, 5] (see also [10, Ch. V.9]).

Next, we give some basic general results which exploit the superexponential character of certain complex valued matrix mappings. This will be used to prove new contractivity results in the forthcoming section. Concerning the matrix norms to be used, we recall that for a given vector norm $\|\cdot\|$ in $\mathbb{C}^m$, the operator norm will be $\|J\| := \max_{\|v\|=1} \|Jv\|$, $J \in \mathbb{C}^{m,m}$, and its associated logarithmic norm is defined by (see, e.g., [8, p. 27]) $\mu[J] := \lim_{\epsilon\to 0^+}(\|I_m + \epsilon J\| - 1)/\epsilon$. It must be noted that for the matrices $J_i^{(n)}$ given in (2.4), $\mu_X[J_i^{(n)}] \leq \nu$, $1 \leq i \leq s$, holds by virtue of (1.2). We will often consider norms associated to the inner product in (1.3). For this case, by splitting $X = Y^T Y$, where $Y \in \mathbb{R}^{m,m}$ is a nonsingular matrix (the decomposition is not unique), we have that

$$\|J\|_X := \left\|YJY^{-1}\right\|_2, \quad \mu_X[J] = \mu_2[YJY^{-1}] \quad \forall J \in \mathbb{C}^{m,m}. \tag{2.9}$$

We will also make use of an operator norm in the superspace $\mathbb{C}^{mk}$ for some $k \in \mathbb{N}$. Thus, for a given symmetric positive definite matrix $G = (g_{ij})_{i,j=1}^{k} \in \mathbb{R}^{k,k}$, the inner product in (1.3) can be extended to $\mathbb{C}^{mk}$ in the usual way:

$$\langle u, v\rangle_{G\otimes X} := \sum_{i,j=1}^{k} g_{ij}\langle u_i, v_j\rangle_X, \tag{2.10}$$
$$u = (u_j)_{j=1}^{k},\ v = (v_j)_{j=1}^{k},\ u_j, v_j \in \mathbb{C}^m,\ 1 \le j \le k.$$

By splitting the matrix $G = L^T L$, with $L \in \mathbb{R}^{k,k}$ nonsingular (the decomposition is not unique), it is not difficult to show that

$$\langle u, v\rangle_{G\otimes X} = \langle (L \otimes Y)u, (L \otimes Y)v\rangle_2. \tag{2.11}$$

Then, the operator norm associated to the inner product (2.10) satisfies for any $K \in \mathbb{C}^{mk,mk}$ that

$$\|K\|_{G\otimes X} := \max_{\|v\|_{G\otimes X}=1} \|Kv\|_{G\otimes X} = \max_{\substack{\|u\|_{G\otimes X}=1\\ \|v\|_{G\otimes X}=1}} |u^*(G \otimes X)Kv|. \tag{2.12}$$

It must be taken into account that for every l_p-norm with $p \ge 1$ and $q^{-1} + p^{-1} = 1$, it holds that

$$\|v\|_p = \max_{\|u\|_q=1} |u^* v|\ \ \forall v \in \mathbb{C}^n, \quad \|J\|_p = \max_{\substack{\|u\|_q=1\\ \|v\|_p=1}} |u^* J v|\ \ \forall J \in \mathbb{C}^{n,n},\ n \in \mathbb{N}. \tag{2.13}$$

This latter property (2.13) follows from the duality between the spaces l_p and l_q, with $p \ge 1$ and $q^{-1} + p^{-1} = 1$; see, e.g., [15, sect. 19]. Thus, by taking $p = 2$ in (2.13), the second equality in (2.12) is deduced from (2.11). Here, u^* stands for the transpose conjugate of the vector u. With these preliminaries and denoting by $\mu_X[\cdot]$ and $\mu_p[\cdot]$ the logarithmic norms associated to the norms $\|\cdot\|_X$ and $\|\cdot\|_p$, respectively, we have the following theorem.

THEOREM 1. *Let $M(Z_1, \dots, Z_s) := (m_{ij}(Z_1, \dots, Z_s))_{i,j=1}^{k} \in \mathbb{C}^{mk,mk}$ be a matrix mapping acting on the matrix-set $\{(Z_1, \dots, Z_s),\ Z_j \in \mathbb{C}^{m,m},\ j = 1, \dots, s\}$. Assume that $m_{ij}(Z_1 + zI_m, \dots, Z_s + zI_m)$, $1 \le i, j \le k$, is an (m, m)-matrix having analytic components on* $\operatorname{Re} z \le 0$ *whenever the matrices Z_j satisfy either*

(a) $\mu_X[Z_j] \le 0$ $(j = 1, \dots, s)$ *or*

(b) $\mu_p[Z_j] \le 0$ $(j = 1, \dots, s)$ *(for some l_p-norm, $p \ge 1$).*

Then, respectively, for each case we have that (a) *the real function*

$$\varphi_{G,M}(x) := \sup_{\substack{\mu_X[Z_j]\le x\\ 1\le j\le s}} \|M(Z_1, \dots, Z_s)\|_{G\otimes X} \tag{2.14}$$

is nondecreasing for $x \le 0$ and fulfills the superexponential property

$$\varphi(x)\varphi(y) \le \varphi(0)\varphi(x + y) \quad \forall\, x, y \le 0, \tag{2.15}$$

provided that $G \in \mathbb{R}^{k,k}$ is symmetric and positive definite; and (b) *the real function*

$$\varphi_{p,M}(x) := \sup_{\substack{\mu_p[Z_j]\le x\\ 1\le j\le s}} \|M(Z_1, \dots, Z_s)\|_p$$

is nondecreasing for $x \le 0$ and satisfies (2.15).

Proof. The proof closely follows the ideas of the proof in [11, Theorem 4.3]. However, we point out some minor modifications related to the matrix case.

(a) Take $x < 0$ and $y < 0$ and assume $\varphi_{G,M}(0) < \infty$. Let us consider vectors $u_i, v_i \in \mathbb{C}^{mk}$, with $\|u_i\|_{G\otimes X} = \|v_i\|_{G\otimes X} = 1$ $(i = 1, 2)$, and matrices $A_1, \ldots, A_s$, $B_1, \ldots, B_s \in \mathbb{C}^{m,m}$ satisfying $\mu_X[A_i] \le x + y$, $\mu_X[B_i] \le 0$, $1 \le i \le s$. Let us define, for $z \in \mathbb{C}$,

$$S(z) = (u_1^*(G \otimes X)M(A_1 - zI_m, \ldots, A_s - zI_m)v_1) \cdot (u_2^*(G \otimes X)M(B_1 + zI_m, \ldots, B_s + zI_m)v_2).$$

Thus, taking into account that in general the logarithmic norms satisfy [8, p. 31] $\mu[J + zI_m] = \mu[J] + \operatorname{Re} z$ for any $J \in \mathbb{C}^{m,m}$, we deduce that $S(z)$ is an analytic function on the strip $\mathcal{S}(0, x + y) := \{z \in \mathbb{C} : \ x + y \le \operatorname{Re} z \le 0\}$. Then from the maximum modulus principle and from (2.12), it follows that

$$|S(z)| \le \sup_{\substack{Re\ z=0,\\ Re\ z=x+y}} |S(z)| \le \varphi_{G,M}(0)\varphi_{G,M}(x + y) \quad \forall z \in \mathcal{S}(0, x + y). \tag{2.16}$$

Now, by taking $z = y$ and considering the maximum over $\|u_2\|_{G\otimes X} = \|v_2\|_{G\otimes X} = 1$ and the supremum over $\mu_X[B_j] \le 0$ $(j = 1, \ldots, s)$, it follows from (2.16) that

$$|u_1^*(G \otimes X)M(A_1 - yI_m, \ldots, A_s - yI_m)v_1|\, \varphi_{G,M}(y) \le \varphi_{G,M}(0)\varphi_{G,M}(x + y).$$

The proof is concluded by taking the maximum over $\|u_1\|_{G\otimes X} = \|v_1\|_{G\otimes X} = 1$ and the supremum over $\mu_X[A_i] \le x + y$ $(i = 1, \ldots, s)$.

(b) This part of the proof is along the lines of that of part (a) but this time defines

$$S(z) = \left(u_1^* M(A_1 - zI_m, \ldots, A_s - zI_m)v_1\right)\left(u_2^* M(B_1 + zI_m, \ldots, B_s + zI_m)v_2\right)$$

and takes $\|u_i\|_q = \|v_i\|_p = 1$ $(i = 1, 2)$, with $q^{-1} + p^{-1} = 1$. ∎

COROLLARY 1. *Let $M(z_1, \ldots, z_s) := (m_{ij}(z_1, \ldots, z_s))_{i,j=1}^k \in \mathbb{C}^{k,k}$, where $m_{ij}(\cdot)$ are analytic functions on each complex variable z_j for* $\operatorname{Re} z_j \le 0$ $(j = 1, \ldots, s)$. *Then the increasing functions*

$$\varphi_{G,M}(x) = \sup_{\substack{Re\ z_j \le x\\ 1\le j\le s}} \|M(z_1, \ldots, z_s)\|_G \quad \textit{and} \quad \varphi_{p,M}(x) = \sup_{\substack{Re\ z_j \le x\\ 1\le j\le s}} \|M(z_1, \ldots, z_s)\|_p$$

fulfill (2.15) *for any symmetric positive definite matrix $G \in \mathbb{R}^{k,k}$ and any l_p-norm $(p \ge 1)$.*

Remark 1. The statements in Corollary 1 remain true when x and y have the same sign and satisfy $x + y \le x_0$ for some $x_0 \ge 0$, and the functions $m_{ij}(z_1, \ldots, z_s)$, $1 \le i, j \le s$, are analytic on $\operatorname{Re} z_j \le x_0$ $(j = 1, \ldots, s)$. The statements in Theorem 1 also hold under similar assumptions.

Although the results presented in this section are related to both Euclidean and l_p-norms, the main applications for General Linear Methods arise when considering Euclidean norms.

3. Strict-contractivity analysis for General Linear Methods. We start with a simple lemma that will be used to prove the main results of strict-contractivity in this section.

LEMMA 1. *Let $\varphi : (-\infty, 0] \to [0, +\infty)$ be a function such that*

$$\text{(i)} \lim_{x\to-\infty} \varphi(x) = \zeta < 1, \quad \text{(ii)} \ \varphi(x)\varphi(y) \le \varphi(0)\varphi(x+y) \quad \forall x, y \in (-\infty, 0].$$

Then, there exists a constant $\sigma > 0$ such that $\varphi(x) \le \max\{\frac{1+\zeta}{2}, Ke^{\sigma x}\}$ $\forall x \le 0$, where $K = \max\{1, \varphi(0)\}$.

Proof. By virtue of (i), there exists $x_0 < 0$ such that $\varphi(x) \le \delta := \frac{1+\zeta}{2}$ $\forall x \le x_0$. Let us take $\sigma := \frac{\ln \delta}{2x_0} > 0$. Thus, if $x \in [2x_0, x_0]$, it follows that $\varphi(x) \le \delta = e^{\sigma(2x_0)} \le e^{\sigma x}$. By virtue of (ii), this implies $\varphi(\frac{x}{l}) \le Ke^{\sigma(\frac{x}{l})}$, $K = \max\{1, \varphi(0)\}$ $\forall x \in [2x_0, x_0]$ and $l \in \mathbb{N}$. On the other hand, for each $x \in (x_0, 0)$ there exists $n \in \mathbb{N}$ such that $x \in [\frac{x_0}{n}, \frac{x_0}{n+1})$. Therefore, $y := (n+1)x \in [2x_0, x_0)$, and $\varphi(x) = \varphi(\frac{y}{n+1}) \le Ke^{\sigma(\frac{y}{n+1})} = Ke^{\sigma x}$. This concludes the proof. ∎

The following theorem gives an estimate about the behavior at infinity of the function $\varphi_{G,M}(x)$ in (2.14).

THEOREM 2. *Let $M(Z)$, with $Z = \text{BlockDiag}(Z_1, \ldots, Z_s)$, be given by* (2.5). *Assume that $\tilde{A}$ is nonsingular. Then, for any symmetric positive definite matrix $G = L^T L \in \mathbb{R}^{k,k}$, we have for $\varphi_{G,M}$, given by* (2.14), *that*

$$\varphi_{G,M}(x) \le \left\|B - A\tilde{A}^{-1}\tilde{B}\right\|_G + \left\|LA\tilde{A}^{-1}\right\|_2 \left\|\tilde{B}L^{-1}\right\|_2 \frac{\left\|\tilde{A}^{-1}\right\|_2}{|x| - \left\|\tilde{A}^{-1}\right\|_2} \quad \forall x < -\left\|\tilde{A}^{-1}\right\|_2.$$

It must be observed that the upper bound in the theorem does not depend on the factorization of G.

Proof. Since $\tilde{A}$ is a nonsingular matrix, from (2.5) we can write

$$M(Z) = (B - A\tilde{A}^{-1}\tilde{B}) \otimes I_m + (A\tilde{A}^{-1} \otimes I_m)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}(\tilde{B} \otimes I_m).$$

By taking into account that $X = Y^T Y$, $Y \in \mathbb{R}^{m,m}$, from (2.9) and (2.11) it follows that

$$\|M(Z)\|_{G\otimes X} \le \left\|B - A\tilde{A}^{-1}\tilde{B}\right\|_G + \left\|LA\tilde{A}^{-1}\right\|_2 \left\|\tilde{B}L^{-1}\right\|_2 \left\|(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}\right\|_{I_s\otimes X}. \tag{3.1}$$

Take $u \in \mathbb{C}^{ms}$, $\|u\|_2 = 1$, and define $v = (I_s \otimes Y)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}(I_s \otimes Y^{-1})u$. We are addressed to provide an upper bound for $\|v\|_2$. From the definition of v

$$(\tilde{A}^{-1} \otimes I_m)u = (\tilde{A}^{-1} \otimes I_m)v - (I_s \otimes Y)Z(I_s \otimes Y^{-1})v, \tag{3.2}$$

and from here

$$\text{Re}\langle v, (\tilde{A}^{-1} \otimes I_m)u\rangle_2 = \text{Re}\langle v, (\tilde{A}^{-1} \otimes I_m)v\rangle_2 - \text{Re}\langle v, (I_s \otimes Y)Z(I_s \otimes Y^{-1})v\rangle_2. \tag{3.3}$$

Then, by denoting $v = (v_1^T, \ldots, v_s^T)^T$, with $v_i \in \mathbb{C}^m$, $1 \le i \le s$, we deduce that

$$\text{Re}\langle v, (I_s \otimes Y)Z(I_s \otimes Y^{-1})v\rangle_2 = \sum_{i=1}^{s} \text{Re}\langle v_i, YZ_iY^{-1}v_i\rangle_2 \le \sum_{i=1}^{s} \mu_X[Z_i]\,\|v_i\|_2^2 \le x\,\|v\|_2^2,$$

with $x = \max_{1\le i\le s} \mu_X[Z_i]$. Moreover, from elementary properties of the inner products we deduce that $\mathrm{Re}\langle v, (\tilde{A}^{-1}\otimes I_m)v\rangle_2 \ge -\|\tilde{A}^{-1}\|_2 \|v\|_2^2$, and $\mathrm{Re}\langle v, (\tilde{A}^{-1}\otimes I_m)u\rangle_2 \le \|\tilde{A}^{-1}\|_2 \|v\|_2$. From (3.3) we then conclude that

$$\left(-x - \left\|\tilde{A}^{-1}\right\|_2\right) \|v\|_2^2 \le \left\|\tilde{A}^{-1}\right\|_2 \|v\|_2 .$$

The proof is completed taking (3.1) into account. □

Next, we deduce, as a consequence of those previous results, the following contractivity result for General Linear Methods. Observe that $R(\infty) = B - A\tilde{A}^{-1}\tilde{B}$ holds for the stability function $R(z)$ in (2.6) associated to a General Linear Method with nonsingular $\tilde{A}$.

THEOREM 3. *Consider a G-stable General Linear Method* (2.1) *with a nonsingular matrix* $\tilde{A}$. *Assume that its linear stability matrix* (2.6) *satisfies* $\|R(\infty)\|_G < 1$. *Then, the method is strictly contractive on the class of strictly dissipative differential systems* ($\nu < 0$ *in* (1.2)); *i.e., there exist two positive constants* σ *and* $\gamma < 1$ (σ *and* γ *depending only on the coefficients of the method) such that*

$$\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \le \max\{\gamma, e^{h\sigma\nu}\} \left\|\Delta Y^{(n)}\right\|_{G\otimes X} \qquad \forall h > 0,\ n \ge 0.$$

Proof. By virtue of Theorem 1 and the G-stability of the method, we have that the function $\varphi_{G,M}(x)$ in (2.14) fulfills (2.15) with $\varphi_{G,M}(0) \le 1$. On the other hand, since $\|R(\infty)\|_G < 1$ we deduce from Theorem 2 and Lemma 1 that there exist two positive constants σ and $\gamma < 1$ such that $\varphi_{G,M}(h\nu) \le \max\{\gamma, e^{h\sigma\nu}\}$. This concludes the proof by taking (2.2) into account. □

Remark 2. In general, a G-stable General Linear Method (2.1) does not necessarily satisfy $\varphi_{G,M}(0) \le 1$. However, that statement is true when the linear system associated to the stages of the method

$$\begin{aligned} V &= (\widetilde{B}\otimes I_m)Y + (\widetilde{A}\otimes I_m)ZV, \\ Z &= \mathrm{BlockDiag}(Z_1, \ldots, Z_s) \in \mathbb{C}^{ms,ms} \end{aligned}$$

admits a solution $V \in \mathbb{C}^{ms}$ for every $Y \in \mathbb{C}^{mk,mk}$, provided that $\mu_X[Z_j] \le 0$, $j = 1, \ldots, s$ (see the proof of [10, Theorem 12.23, p. 193]). Some results on the existence of solution for G-stable General Linear Methods are presented in section 4.

The main drawback when applying the previous theorem to Runge–Kutta multistep methods comes from the fact that many of these methods fulfill $\|R(\infty)\|_G = 1$, where G denotes the G-stability matrix. For instance, this is the case of the two-step backward differentiation formula (BDF) method and of the multistep Runge–Kutta Gauss methods introduced by Burrage [1, 2], as it will be shown later in section 6. Thus, to reach strict-contractivity for some G-stable methods it seems necessary to compose the method on several consecutive steps, namely, l. For multistep Runge–Kutta methods l will sometimes coincide with the number of steps (k) on which the multistep method is based; see (2.1). This motivates the study of the *l-step composed method* related to a given General Linear Method. □

A straightforward calculation shows that the numerical solution provided by (2.1) after l consecutive steps, with fixed step-size $h > 0$, satisfies

$$Y^{(n+l)} = h\sum_{i=0}^{l-1} (B^{l-1-i}A\otimes I_m)F(t_{n+i}, h, V^{(n+i)}) + (B^l\otimes I_m)Y^{(n)}, \quad l \ge 1,$$

where the stages $V^{(n+j)}$ are computed from

$$\begin{aligned} V^{(n+j)} &= h\sum_{i=0}^{j-1}(\tilde{B}B^{j-1-i}A\otimes I_m)F(t_{n+i},h,V^{(n+i)}) \\ &\quad + h(\tilde{A}\otimes I_m)F(t_{n+j},h,V^{(n+j)}) + (\tilde{B}B^j\otimes I_m)Y^{(n)}, \quad 0\le j\le l-1. \end{aligned}$$

Thus, the l-step composed method can be seen as a new k-step ls-stage General Linear Method having the form

$$\begin{aligned} Y^{(n+l)} &= h(\alpha\otimes I_m)\overline{F}(t_n,h,\overline{V}^{(n+l-1)}) + (\beta\otimes I_m)Y^{(n)}, \\ \overline{V}^{(n+l-1)} &= h(\tilde{\alpha}\otimes I_m)\overline{F}(t_n,h,\overline{V}^{(n+l-1)}) + (\tilde{\beta}\otimes I_m)Y^{(n)}, \end{aligned} \tag{3.4}$$

where

$$\begin{aligned} \overline{V}^{(n+l-1)} &:= ({V^{(n)}}^T,{V^{(n+1)}}^T,\dots,{V^{(n+l-1)}}^T)^T\in\mathbb{C}^{mls}, \\ \overline{F}(t_n,h,\overline{V}^{(n+l-1)}) &:= (F(t_{n+j},h,{V^{(n+j)}}^T)^T)_{0\le j\le l-1}^T\in\mathbb{C}^{mls} \end{aligned}$$

and

$$\tilde{c}=\begin{pmatrix} c \\ c+e \\ \vdots \\ c+(l-1)e \end{pmatrix}\in\mathbb{R}^{ls},\quad \alpha=[B^{l-1}A,\ B^{l-2}A,\ \dots\ BA,\ A]\in\mathbb{R}^{k,ls},\quad \beta=B^l\in\mathbb{R}^{k,k},$$

$$\tilde{\alpha}=\begin{pmatrix} \tilde{A} & O & O & \dots & O & O \\ \tilde{B}A & \tilde{A} & O & \dots & O & O \\ \tilde{B}BA & \tilde{B}A & \tilde{A} & \dots & O & O \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \tilde{B}B^{l-2}A, & \tilde{B}B^{l-3}A, & \dots & \dots, & \tilde{B}A, & \tilde{A} \end{pmatrix}\in\mathbb{R}^{ls,ls},\quad \tilde{\beta}=\begin{pmatrix} \tilde{B} \\ \tilde{B}B \\ \vdots \\ \tilde{B}B^{l-1} \end{pmatrix}\in\mathbb{R}^{ls,k}.$$

To study the A-stability of the method (3.4), one has to consider the matrix $R_l(z)=\beta+z\alpha(I-z\tilde{\alpha})^{-1}\tilde{\beta}$. Since, for linear problems $y'=\lambda y$, $z=\lambda h$, the l-step composed method gives $Y^{(n+l)}=R_l(z)Y^{(n)}$ and also $Y^{(n+l)}=R(z)^lY^{(n)}$, where $R(z)$ is the stability matrix (2.6) of the original method (2.1), then it follows that $R_l(z)=R(z)^l$ $\forall z\in\mathbb{C}$. It also becomes evident that if $\tilde{A}$ is nonsingular, then so is $\tilde{\alpha}$. To study the strict-contractivity of the l-step composed method on nonlinear problems in general, proceeding the same way as in (2.2), one has that $\Delta Y^{(n+l)}=M_l(h\overline{J})\Delta Y^{(n)}$, $n\ge 0$, where $\overline{J}=\text{BlockDiag}(J^{(n+1)},\dots,J^{(n+l)})\in\mathbb{C}^{mls,mls}$, with the matrices $J^{(n+j)}$, $1\le j\le l$, given as in (2.3), and the supermatrix $M_l(\cdot)$ given by

$$M_l(\overline{Z}):=\beta\otimes I_m+(\alpha\otimes I_m)\overline{Z}(I_{mls}-(\tilde{\alpha}\otimes I_m)\overline{Z})^{-1}(\tilde{\beta}\otimes I_m)\in\mathbb{C}^{mk,mk}.$$

Of course, if the original method (2.1) is G-stable, then so is the l-step composed method (3.4). Thus, we can apply Theorem 3 to the resulting l-step composed method to get the following theorem.

THEOREM 4. *Assume that the coefficient matrix $\tilde{A}$ of a G-stable General Linear Method* (2.1) *is nonsingular and that $\rho(R(\infty))<1$. Then, there exist a first positive integer l and two positive constants σ and $\gamma<1$, depending only on the coefficients of the method, such that*

$$\left\|\Delta Y^{(n+l)}\right\|_{G\otimes X}\le\max\{\gamma,e^{h\sigma\nu}\}\left\|\Delta Y^{(n)}\right\|_{G\otimes X}\quad\forall h>0,\quad n\ge 0,$$

holds on the class of strictly dissipative differential systems ($\nu < 0$ in (1.2)). Hence, the l-step composed method (3.4) is strictly contractive.

Proof. The proof is an immediate consequence of Theorem 3, just by taking into account that $\rho(R(\infty)) = \lim_{l\to\infty} \|R(\infty)^l\|_G^{1/l} < 1$. ∎

Remark 3. It is still possible to deduce contractivity results when the matrix $\tilde{A}$ of the original General Linear Method is singular and the method is *algebraically stable*, preconsistent (see (2.8)), and irreducible (there are nonredundant stages; see, e.g., [12, p. 24] for a precise definition of irreducibility). On the other hand, it must be recalled that the matrix $D = \mathrm{Diag}(d_1, \ldots, d_s)$, appearing in the algebraic stability matrix N (2.7), has positive diagonal entries ($D > 0$) due to the algebraic stability; see [14, Lemma 4.1] or [12, Lemma 3.2].

THEOREM 5. *For a preconsistent, algebraically stable General Linear Method with $D > 0$ (see (2.7)), the difference $\Delta Y^{(n)} = \hat{Y}^{(n)} - Y^{(n)}$ between two numerical solutions for dissipative differential systems ($\nu \le 0$ in (1.2)) satisfies*

$$\left\|\Delta Y^{(n+1)}\right\|_{G\otimes X} \le \eta(h\nu) \left\|\Delta Y^{(n)}\right\|_{G\otimes X}, \ n \ge 0,$$

where

$$\eta(x) := \frac{\sqrt{1 - 2x\gamma(1-\zeta^2)} - 2x\gamma\zeta}{1 - 2x\gamma}, \tag{3.5}$$

$$\zeta = \|R(\infty)\|_G, \qquad R(\infty) := \lim_{\varepsilon\to 0}(B - A(\tilde{A} + \varepsilon I)^{-1}\tilde{B}),$$

and

$$\gamma = \left(\lim_{\varepsilon\to 0} \left\|LA(\tilde{A} + \varepsilon I)^{-1}D^{-1/2}\right\|_2\right)^{-2}, \quad \text{with } G = L^T L. \tag{3.6}$$

Remark 4. This result can be seen as a generalization of [11, Theorem 3.1] to General Linear Methods. For the sake of brevity, we do not reproduce its proof, which closely follows the ideas in the proof of the above-mentioned theorem.

The existence of $\lim_{\varepsilon\to 0} A(\tilde{A} + \varepsilon I)^{-1}$ can be ensured by virtue of [13, Lemma 3.8, p. 388].

According to [11, p. 214], the function $\eta(x)$ in (3.5) turns out to be superexponential ($\eta(0) = 1$ and $\eta(x)\eta(y) \le \eta(x + y)$, $x \le 0$, $y \le 0$). Moreover, it can be easily shown that $\zeta < \eta(x) < 1$ $\forall x < 0$ whenever $\zeta < 1$; whereas $\eta(x) = 1$ $\forall x \le 0$ in case $\zeta = 1$.

Analogous qualitative results of strict-contractivity for General Linear Methods on the class of strictly dissipative problems are deduced from Theorems 3 and 5. However, in order to simplify the presentation, our convergence results appearing in section 5 will be obtained from Theorems 2, 3, and 4.

4. Existence and uniqueness of solution for the stage equations. The existence of solution for the stage equations of General Linear Methods is derived along the same lines as in the case of implicit Runge–Kutta methods; see, e.g., [6], [8, Ch. 5], and [10, Ch. IV.14]. Below, we collect a few results whose proofs can be carried out by closely following the ideas in the above-mentioned works.

DEFINITION 1. *For any matrix $C \in \mathbb{R}^{s,s}$ and for any positive definite diagonal matrix $\tilde{D} \in \mathbb{R}^{s,s}$*

$$\alpha_{\tilde{D}}(C) := \inf_{u\neq 0} \frac{\langle Cu, u\rangle_{\tilde{D}}}{\langle u, u\rangle_{\tilde{D}}} = \frac{1}{2}\lambda_{min}(\tilde{D}^{1/2}C\tilde{D}^{-1/2} + \tilde{D}^{-1/2}C^T\tilde{D}^{1/2}),$$

and $\alpha_0(C) := \sup_{\tilde{D}>0} \alpha_{\tilde{D}}(C)$*, where* $\langle u, v\rangle_{\tilde{D}} := v^T \tilde{D} u$.

By following, for instance, the proofs of Theorems 14.2 and 14.3 in [10, Ch. IV.14] we have the following theorem.

THEOREM 6. *For the class of problems* (1.1)–(1.2)*, a General Linear Method having a nonsingular matrix* $\tilde{A}$ *has a unique solution for its stage equations whenever* $\nu h < \alpha_0(\tilde{A}^{-1})$ *is fulfilled.*

The condition $\nu h < \alpha_0(\tilde{A}^{-1})$ is *essentially optimal* as has been remarked by Kraaijevanger and Schneid [16, Theorem 2.12] for the more particular case of implicit Runge–Kutta methods. For the case of dissipative differential systems we also have the following theorem.

THEOREM 7. *An irreducible, preconsistent, and algebraically stable General Linear Method having a nonsingular matrix* $\tilde{A}$ *has a unique solution for its stage equations on the class of strictly dissipative problems* ($\nu < 0$ *in* (1.2))*.*

Proof. Due to the algebraic stability, the matrix N in (2.7) is nonnegative definite ($N \geq 0$). Moreover, the diagonal matrix D appearing in N is positive definite [14, Lemma 4.1]. Then, the matrix $D\tilde{A} + \tilde{A}^T D \geq 0$, and this implies $D\tilde{A}^{-1} + \tilde{A}^{-T} D \geq 0$. Consequently $\alpha_D(\tilde{A}^{-1}) \geq 0$. The proof is completed from Theorem 6. ∎

In case the matrix $\tilde{A}$ is singular, a more careful and particular study must be carried out. Thus, in situations where some stages are explicit, the above results can still be applied after deleting the corresponding rows (and columns) of the matrix $\tilde{A}$. This is, e.g., the case of the Runge–Kutta Lobatto IIIA methods.

5. Convergence results. We start by recalling that the discretization local errors associated to a General Linear Method (2.1) are given by the pair of vectors $(\xi(t_n), \eta(t_n))$ by means of [14, p. 366]

$$\begin{array}{rcl} \xi(t_n) & := & Y(t_n + h) - (B \otimes I_m) Y(t_n) - h(A \otimes I_m) V'(t_n), \\ \eta(t_n) & := & V(t_n) - (\tilde{B} \otimes I_m) Y(t_n) - h(\tilde{A} \otimes I_m) V'(t_n), \end{array} \tag{5.1}$$

where $V(t_n) = (y(t_n + c_j h))_{j=1}^s$, $V'(t_n) = (y'(t_n + c_j h))_{j=1}^s \in \mathbb{C}^{ms}$, and $Y(t_n) = (Y_j(t_n))_{j=1}^k \in \mathbb{C}^{mk}$. Here, $Y(t_n)$ denotes the exact advancing solution to be approximated for the method, and $y(t)$ stands for the exact solution of the initial value problem (1.1). It is straightforward to see that if the method has stage order q, then

$$\begin{array}{l} \xi(t_n) = h^{q+1}(d_1^{(q+1)} \otimes I_m) y^{(q+1)}(t_n) + \mathcal{O}(h^{q+2}), \\ \eta(t_n) = h^{q+1}(d_2^{(q+1)} \otimes I_m) y^{(q+1)}(t_n) + \mathcal{O}(h^{q+2}), \end{array} \tag{5.2}$$

where the vectors $d_1^{(q+1)} \in \mathbb{R}^k$ and $d_2^{(q+1)} \in \mathbb{R}^s$ depend only on the coefficients of the method; see, e.g., [14, p. 366]. From a straightforward calculation we deduce the next two upper bounds for the discretization local errors (5.1)

$$\begin{array}{rcl} \|\xi(t_n)\|_{I_k \otimes X} & \leq & h^q C_1 \int_{t_n}^{t_{n+1}} \left\| y^{(q+1)}(t) \right\|_X dt, \\ \|\eta(t_n)\|_{I_s \otimes X} & \leq & h^q C_2 \int_{t_n}^{t_{n+1}} \left\| y^{(q+1)}(t) \right\|_X dt, \end{array} \tag{5.3}$$

(5.4)

$$\begin{array}{rcl} \left\| \xi(t_n) - h^{q+1}(d_1^{(q+1)} \otimes I_m) y^{(q+1)}(t_n) \right\|_{I_k \otimes X} & \leq & h^{q+1} C_1' \int_{t_n}^{t_{n+1}} \left\| y^{(q+2)}(t) \right\|_X dt, \\ \left\| \eta(t_n) - h^{q+1}(d_2^{(q+1)} \otimes I_m) y^{(q+1)}(t_n) \right\|_{I_s \otimes X} & \leq & h^{q+1} C_2' \int_{t_n}^{t_{n+1}} \left\| y^{(q+2)}(t) \right\|_X dt, \end{array}$$

where the constants C_j, C_j', $j = 1, 2$, depend only on the coefficients of the method.

The study of the global errors $\epsilon_n := Y(t_n) - Y^{(n)}$ $(n = 1, 2, \dots)$, can be carried out, for instance, as it is done in [14, sect. 2]. Thus, it is immediate to show that they satisfy the recurrence

$$\begin{array}{c} \epsilon_{n+1} = M(Z^{(n)})\epsilon_n + \tau_n, \quad n = 0, 1, \dots, \\ \tau_n := \xi(t_n) + \omega(Z^{(n)})\eta(t_n), \\ \omega(Z) := (A \otimes I_m)Z(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}, \end{array} \tag{5.5}$$

where $M(Z)$ is the stability function given in (2.5), with $Z^{(n)} = hJ^{(n)}$ and $J^{(n)}$ given by (2.3) and (2.4). Thus, we are involved with getting upper bounds for $\|\prod_{j=0}^{q} M(Z^{(n-j)})\|_{G\otimes X}$ (for arbitrary $q \le n$) and for $\max_{\|u\|_{I_s\otimes X}=1} \|\omega(Z^{(n)})u\|_{G\otimes X}$. In order to bound the latter norm we have the following theorem.

THEOREM 8. *Assume that $\tilde{A}$ is nonsingular and that there exists a positive definite diagonal matrix $\tilde{D} = \mathrm{Diag}(\tilde{d}_1, \dots, \tilde{d}_s)$ such that $\tilde{\alpha} := \frac{1}{2}\lambda_{min}(\tilde{D}\tilde{A}^{-1} + \tilde{A}^{-T}\tilde{D}) \ge 0$. Then, for any symmetric positive definite matrix G and for any $Z = \mathrm{BlockDiag}(Z_1, \dots, Z_s)$, with $Z_j \in \mathbb{C}^m$, $1 \le j \le s$, it holds that*

$$\sup_{\substack{\|u\|_{I_s\otimes X}=1 \\ \mu_X[Z_j]\le x,\ 1\le j\le s}} \|\omega(Z)u\|_{G\otimes X} \le \kappa \left(1 + \frac{\left\|\tilde{D}\tilde{A}^{-1}\right\|_2}{\tilde{\alpha} - x\delta}\right) \quad \forall x < 0,$$

where

$$\kappa = \sqrt{\lambda_{max}((A\tilde{A}^{-1})^T G A\tilde{A}^{-1})} \quad \textit{and} \quad \delta = \min_{1\le j\le s} \tilde{d}_j. \tag{5.6}$$

Moreover, if $\tilde{\alpha} > 0$ the statement also holds for $x = 0$.

Proof. Let us consider $G = L^T L$ and $X = Y^T Y$ and take an arbitrary $u \in \mathbb{C}^{ms}$ satisfying $\|u\|_{I_s\otimes X} = \|(I_s \otimes Y)u\|_2 = 1$. By defining $y = \omega(Z)u$, it follows that

$$\begin{aligned} \|y\|_{G\otimes X} &= \left\|((A\tilde{A}^{-1}) \otimes I_m)((I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}u - u)\right\|_{G\otimes X} \\ &\le \left\|LA\tilde{A}^{-1}\right\|_2 \left(\left\|(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}\right\|_{I_s\otimes X} + 1\right). \end{aligned}$$

In order to bound $\|(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}\|_{I_s\otimes X}$, take an arbitrary $w \in \mathbb{C}^{ms}$, with $\|w\|_2 = 1$, and set $v = (I_s \otimes Y)(I_{ms} - (\tilde{A} \otimes I_m)Z)^{-1}(I_s \otimes Y^{-1})w$. Now, proceeding as in the proof of Theorem 2 and premultiplying by $v^*(\tilde{D} \otimes I_m)$ in both sides of the formula (3.2), we get that

$$v^*(\tilde{D}\tilde{A}^{-1} \otimes I_m)w = v^*(\tilde{D}\tilde{A}^{-1} \otimes I_m)v - v^*(I_s \otimes Y)(\tilde{D} \otimes I_m)Z(I_s \otimes Y^{-1})v. \tag{5.7}$$

By taking real part in the above equation and making use of the inequalities

$$\mathrm{Re}(v^*(I_s \otimes Y)(\tilde{D} \otimes I_m)Z(I_s \otimes Y^{-1})v) \le \delta x \|v\|_2^2 \quad \forall x < 0,$$

and

$$\mathrm{Re}(v^*((\tilde{D}\tilde{A}^{-1}) \otimes I_m)v) = 2^{-1}v^*((\tilde{D}\tilde{A}^{-1} + \tilde{A}^{-T}\tilde{D}) \otimes I_m)v \ge \tilde{\alpha}\|v\|_2^2,$$

it follows from (5.7) that

$$(\tilde{\alpha} - \delta x)\|v\|_2^2 \le \mathrm{Re}(v^*(\tilde{D}\tilde{A}^{-1} \otimes I_m)w) \le \left\|\tilde{D}\tilde{A}^{-1}\right\|_2 \|v\|_2.$$

Then, it yields that $\|(I_{ms}-(\tilde{A}\otimes I_m)Z)^{-1}\|_{I_s\otimes X}\leq \frac{\|\tilde{D}\tilde{A}^{-1}\|_2}{\tilde{\alpha}-\delta x}$. This completes the proof. ∎

Next, we give the main convergence results.

THEOREM 9. *Consider an irreducible and algebraically stable (for the matrices G and D) General Linear Method* (2.1)*, with stage order q and a nonsingular matrix $\tilde{A}$. Assume that $\rho(R(\infty))<1$. Then, by taking any positive definite diagonal matrix $\tilde{D}$ such that $\tilde{\alpha}=\alpha_{\tilde{D}}(\tilde{A}^{-1})\geq 0$, we have that the global errors $\epsilon_n=Y(t_n)-Y^{(n)}$ on strictly dissipative problems ($\nu<0$) satisfy for $n\geq 1$ and $h>0$ that*

$$\|\epsilon_n\|_{G\otimes X}\leq \varrho^{[l^{-1}n]}\,\|\epsilon_0\|_{G\otimes X}+h^{q+1}C_3 l\frac{1-\varrho^{1+l^{-1}h^{-1}t_{n-1}}}{1-\varrho}\mathcal{M}_{q+1}(t_n),$$

where $[x]$ denotes the integer part of the real number x and

$$C_3=C_1\sqrt{\rho(G)}+\kappa C_2\left(1+\frac{\left\|\tilde{D}\tilde{A}^{-1}\right\|_2}{\tilde{\alpha}+|\nu|\delta h}\right). \tag{5.8}$$

Here, l is the first positive integer verifying $\left\|R^l(\infty)\right\|_G<1$, $\varrho=\max\{\gamma,e^{h\sigma\nu}\}$, ($0<\gamma<1$ and $\sigma>0$ are those two constants given in Theorem 4), C_1 and C_2 are those two constants given in (5.3)*, κ and δ are both given in* (5.6)*, and*

$$\mathcal{M}_j(t_n):=\max_{t\in[0,t_n]}\left\|y^{(j)}(t)\right\|_X. \tag{5.9}$$

All the constants (with the exception of ϱ) depend only on the coefficients of the method. Recall that the matrix $\tilde{D}$ also depends exclusively on the matrix $\widetilde{A}$.

Proof. Let us take $\varrho=\max\{\gamma,e^{h\sigma\nu}\}<1$, where $0<\gamma<1$ and $\sigma>0$ are the constants given in Theorem 4. Then, by virtue of Theorem 4 it follows that (below $\|\cdot\|\equiv\|\cdot\|_{G\otimes X}$)

$$\left\|\prod_{k=1}^{j}M(Z^{(n-k)})\right\|\leq\varrho^{[j/l]}\quad\forall\, j\leq n. \tag{5.10}$$

From (5.5) we get for the global errors that

$$\|\epsilon_n\|\leq\|\tau_{n-1}\|+\sum_{j=1}^{n-1}\left\|\prod_{k=1}^{j}M(Z^{(n-k)})\right\|\|\tau_{n-j-1}\|+\left\|\prod_{k=1}^{n}M(Z^{(n-k)})\right\|\|\epsilon_0\|. \tag{5.11}$$

In order to bound $\|\tau_j\|$, from (5.3) and from Theorem 8 we deduce that

$$\|\tau_j\|\leq C_3h^{q+1}\mathcal{M}_{q+1}(t_{j+1}),\quad h>0,\ j\geq 0, \tag{5.12}$$

where C_3 is defined by (5.8). By inserting (5.10) and (5.12) into (5.11), it follows that

$$\|\epsilon_n\|\leq h^{q+1}C_3\mathcal{M}_{q+1}(t_n)\sum_{j=0}^{n-1}\varrho^{[j/l]}+\varrho^{[n/l]}\,\|\epsilon_0\|,\quad h>0,\ n\geq 1.$$

By setting $n-1=pl+q$, with p and q integers, $0\leq q<l$, we have that

$$\sum_{j=0}^{n-1}\varrho^{[j/l]}=l\sum_{k=0}^{p-1}\varrho^k+(q+1)\varrho^p\leq l\sum_{k=0}^{p}\varrho^k=l\frac{1-\varrho^{p+1}}{1-\varrho}=l\frac{1-\varrho^{1+[l^{-1}(n-1)]}}{1-\varrho}.$$

This concludes the proof since $\varrho^{1+[l^{-1}(n-1)]} \geq \varrho^{1+l^{-1}(n-1)}$. ∎

As a consequence of the previous theorem, we deduce the orders of convergence on semi-infinite intervals (and B-convergence on finite intervals) of algebraically stable General Linear Methods. It will be seen that the convergence results extend the pioneer results of B-convergence given by Frank, Schneid, and Ueberhuber [9] (see also [10, Theorem 15.3]) to the case of General Linear Methods. Our results for finite intervals are similar to those presented in [12]. However, the results on semi-infinite intervals, as far as we are aware, are completely new.

THEOREM 10. *Under the assumptions made in Theorem* 9 *for a General Linear Method, we have for dissipative problems* ($\nu \leq 0$) *that its global errors* $\epsilon_n := Y(t_n) - Y^{(n)}$ *satisfy for* $h > 0$, $n \geq 1$

1. *if* $\tilde{\alpha} > 0$ *and* $\nu < 0$, *then* $\|\epsilon_n\|_{G\otimes X} \leq h^q \frac{K}{|\nu|} \mathcal{M}_{q+1}(t_n) + \varrho^{[l^{-1}n]} \|\epsilon_0\|_{G\otimes X}$,
2. *if* $\tilde{\alpha} > 0$ *and* $\nu \leq 0$, *then* $\|\epsilon_n\|_{G\otimes X} \leq h^q K t_n \mathcal{M}_{q+1}(t_n) + \varrho^{[l^{-1}n]} \|\epsilon_0\|_{G\otimes X}$, *and*
3. *if* $\tilde{\alpha} = 0$ *and* $\nu < 0$, *then* $\|\epsilon_n\|_{G\otimes X} \leq h^{q-1} \frac{K}{|\nu|} \min\{\frac{1}{|\nu|}, t_n\} \mathcal{M}_{q+1}(t_n) + \varrho^{[l^{-1}n]} \|\epsilon_0\|_{G\otimes X}$,

where $\mathcal{M}_{q+1}(t_n)$ *is given in* (5.9), ϱ *and* l *are supplied by Theorem* 9, *and the constant* K *depends only on the coefficients of the method.*

Proof. The proof follows in a straightforward way from Theorem 9. It must be taken into account that if $\nu < 0$ then

$$\chi(\nu, h) := l \frac{1 - e^{\sigma\nu h(1+\frac{n-1}{l})}}{1 - e^{\sigma\nu h}} \leq \frac{l}{1 - e^{\sigma\nu h}} = \frac{l}{\sigma|\nu|h} + \mathcal{O}(1),\ h \to 0.$$

Moreover, for $\nu < 0$, $\chi(\nu, h) \leq \lim_{\nu\to 0^-} \chi(\nu, h) = h^{-1}t_n + l - 1$, since the function $(1 - e^x)^{-1}(1 - e^{ax})$ is increasing on x whenever the constant $a > 1$. ∎

It is still possible to gain one more order of convergence for a General Linear Method when

$$d_1^{(q+1)} = (I_k - B)x, \quad d_2^{(q+1)} = -\tilde{B}x \tag{5.13}$$

hold for some vector $x \in \mathbb{R}^k$. Here, $d_1^{(q+1)}$ and $d_2^{(q+1)}$ are given by (5.2).

THEOREM 11. *If a General Linear Method satisfies* (5.13) *and the assumptions in Theorem* 9, *then, for dissipative problems* ($\nu \leq 0$), *its global errors fulfill the statements in Theorem* 10 *with* q *replaced by* $q + 1$.

Proof. The proof is along the lines of that of Theorem 9, but with the following minor modifications, which are inspired by the proof of Theorem 4.2 in [14]. We define the modified global errors

$$\bar{\epsilon}_n := \epsilon_n - (x \otimes I_m) h^{q+1} y^{(q+1)}(t_n),$$

which, by virtue of (5.4) and (5.13), fulfill the recurrent relation

$$\bar{\epsilon}_n = M(Z^{(n)})\bar{\epsilon}_{n-1} + \bar{\tau}_{n-1},\ n \geq 1,\ \text{with}\ \|\bar{\tau}_{n-1}\|_{G\otimes X} \leq h^{q+2} C_3' \mathcal{M}_{q+2}(t_n).$$

Here C_3' has the same expression as C_3 in (5.8) but just replacing the constants C_i by C_i' ($i = 1, 2$), with C_i' given in (5.4). ∎

From the previous theorem it seems that one more order than the stage order can be achieved, but as it has been remarked by Hundsdorfer [14, p. 378], very few interesting methods satisfy (5.13), and this assumption seems necessary also to gain one order with regard to the stage order [14, Theorem 4.2]. We discuss this possibility more in the next section.

6. Some applications of the main results. Below, we consider a few methods in the Runge–Kutta multistep family which take the form of a General Linear Method (2.1) with

$$A = \begin{pmatrix} \gamma_1 & \cdots & \gamma_s \\ O & \dots & O \end{pmatrix}, \qquad B = \left(\begin{array}{ccc|c} \alpha_1 & \dots & \alpha_{k-1} & \alpha_k \\ \hline & I_{k-1} & & O \end{array}\right),$$

where $\gamma_1, \dots, \gamma_s, \alpha_1, \dots, \alpha_k \in \mathbb{R}$, I_{k-1} denotes the identity matrix of dimension $k-1$ and $O \in \mathbb{R}^{k-1}$ stands for the null vector. Moreover, due to preconsistency, they always satisfy $\alpha_1 + \cdots + \alpha_k = 1$.

In the case of $k = 1$, we get the classical Runge–Kutta methods. Thus, for the s-stage Runge–Kutta Radau IA, Radau IIA, and Lobatto IIIC methods (see, e.g., [10, Ch. IV.5]), we have that $(\gamma_k)_{k=1,s}$ is the weight vector and $B = (1)$. Then, since $R(\infty) = 0$, $G = (1)$, and $D = \mathrm{Diag}(\gamma_k)_{k=1,s} > 0$, by virtue of Theorem 4 these methods turn out to be strictly contractive for the strictly dissipative class ($\nu < 0$). Moreover, from Theorem 10 (see also [10, p. 220]) we deduce that for semi-infinite intervals, the Radau IIA methods have order of convergence $q = s$, whereas Radau IA methods reach order of convergence $q = s-1$. In the same way, the two-stage Lobatto IIIC methods reach order $q = 1$. However, for Lobatto IIIC methods with $s \geq 3$, only order of convergence $q = s - 2$ can be derived from Theorem 10.

It should be mentioned that one more order than the stage order cannot be guaranteed since these families do not fulfill the condition (5.13).

Next, we consider the classical two-step backward differentiation formula $BDF2$, given by $\frac{3}{2}y_{n+1} - 2y_n + \frac{1}{2}y_{n-1} = hf(t_{n+1}, y_{n+1})$, $n \geq 1$. This method is a two-step one-stage G-stable General Linear Method for the symmetric positive definite matrix $G := \begin{pmatrix} \frac{5}{4} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{1}{4} \end{pmatrix}$ (see, e.g., [7], [10, Example 6.5, pp. 308–309]). Its stability matrix satisfies $\|R(\infty)\|_G = 1$ and $\left\|R(\infty)^2\right\|_G = 0$. Then, according to Theorem 4, the composed method defined by two consecutive steps of the $BDF2$ formula defines a strictly contractive method on the class of strictly dissipative differential systems. By taking $D = \tilde{A} = (2/3)$, it is readily seen that $\alpha_D(\tilde{A}^{-1}) = 3/2$. Then, from Theorem 10, this method achieves order of convergence two on semi-infinite intervals. It can be easily checked that this method does not fulfill the condition (5.13), and hence Theorem 11 cannot be applied. It is well known that this method has exactly order of convergence 2 on the class $\nu < 0$.

On the other hand, Burrage [1, 2] introduced the following set of simplifying conditions in order to construct high order algebraically stable multistep Runge–Kutta methods:

$$\begin{aligned}
B(p): \quad \bar{b}_p &:= q\sum_{j=1}^{s} \gamma_j c_j^{q-1} + \sum_{j=1}^{k} \alpha_j (1-j)^q - 1 = 0, && 1 \leq q \leq p, \\
C(p): \quad \bar{c}_p &:= q\sum_{j=1}^{s} \tilde{a}_{ij} c_j^{q-1} + \sum_{j=1}^{k} \tilde{b}_{ij} (1-j)^q - c_i^q = 0, && 1 \leq q \leq p,\ \forall i, \\
D(p): \quad \bar{d}_p &:= q\sum_{i=1}^{s} \gamma_i c_i^{q-1} \tilde{b}_{ij} - \alpha_j (1-(1-j)^q) = 0, && 1 \leq q \leq p,\ \forall j, \\
E(p): \quad \bar{e}_p &:= q\sum_{i=1}^{s} \gamma_i c_i^{q-1} \tilde{a}_{ij} - \gamma_j (1 - c_j^q) = 0, && 1 \leq q \leq p,\ \forall j.
\end{aligned}$$

Thus, the family of k-step s-stage Runge–Kutta Gauss methods with stage order $q = s$ and order of consistency $2s$ has been derived by Burrage [1, 2] by imposing

$B(s)$, $C(s)$, $D(s)$, and $E(s)$. This family of methods, with parameters $\alpha_j \geq 0$ ($j = 1, \ldots, k-1$), $\alpha_k > 0$, and $\alpha_1 + \ldots \alpha_k = 1$ provides G-stable methods, where $G = \mathrm{Diag}(1, \alpha_2 + \ldots \alpha_k, \ldots, \alpha_{k-1} + \alpha_k, \alpha_k)$; see Theorem 9.15 in [10, p. 367]. The nodes $c_1, \ldots, c_s$ depend on $k-1$ parameters α_j ($j = 2, \ldots, k$), they are placed on the real interval $(1-k, 1)$, and they are uniquely determined in terms of the $k-1$ parameters by the condition $B(2s)$ (see, e.g., [10, Lemma 9.11]); that is,

$$\sum_{j=1}^{k} \alpha_j \int_{1-j}^{1} \pi(x) x^{q-1} dx = 0, \quad 1 \leq q \leq s, \tag{6.1}$$

where $\pi(x) = (x - c_1) \cdot \cdots \cdot (x - c_s)$ stands for the nodal polynomial. In particular, for the uniparametric family of two-step two-stage Runge–Kutta Gauss methods, which will be here denoted by $Gauss(\alpha)$, with free parameter $\alpha \equiv \alpha_2 \in (0, 1]$ ($\alpha_1 = 1 - \alpha$), the nodes c_1 and c_2 can be computed in terms of α from (6.1). The weights γ_1, γ_2 and the coefficient matrices $\widetilde{B}$ and $\widetilde{A}$ are uniquely determined in terms of α by the conditions $B(2)$, $D(2)$, and $C(2)$, respectively. For every $\alpha \in (0, 1]$, such methods are G-stable and algebraically stable with matrix $G = \mathrm{Diag}(1, \alpha)$ and diagonal matrix $D = \mathrm{Diag}(\gamma_1, \gamma_2)$. For $\alpha \to 0^+$ we get the classical one-step Gauss method with step-size $h > 0$ to advance from t_n to t_{n+1}; whereas for $\alpha \to 1^-$ the classical one-step Gauss method with step-size $2h$ to advance from t_{n-1} to t_{n+1} is deduced. In Table 1 the main features of this family of methods are displayed. In particular, from

TABLE 1

Some features of the 2-step 2-stage algebraically stable Runge–Kutta multistep methods, denoted by $Gauss(\alpha)$, $RadauIIA(\alpha)$, $RadauIA(\alpha)$, and $LobattoIIIC(\alpha)$, $0 < \alpha \leq 1$.

	$Gauss(\alpha)$
$\|R(\infty)\|_G < 1$	Never
$\left\|R(\infty)^2\right\|_G < 1$	$0 < \alpha < 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha < 1$
$\rho(R(\infty)) = 1,\ \alpha_D(\widetilde{A}^{-1}) = 0$	$\alpha = 1$
Simplifying conditions	$B(2),\ C(2),\ D(2),\ E(2)$
Convergence on $[0, \infty)$	$q = 2$ (only for $0 < \alpha < 1$)
	$RadauIIA(\alpha)$
$\|R(\infty)\|_G < 1$	$0 < \alpha < 1$
$\left\|R(\infty)^2\right\|_G < 1$	$0 < \alpha \leq 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha \leq 1$
Simplifying conditions	$B(2),\ C(2),\ D(2),\ E(1),\ c_2 = 1$
Convergence on $[0, \infty)$	$q = 2$ ($0 < \alpha \leq 1$)
	$RadauIA(\alpha)$
$\|R(\infty)\|_G < 1$	$0 < \alpha < 1$
$\left\|R(\infty)^2\right\|_G < 1$	$0 < \alpha \leq 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha \leq 1$
Simplifying conditions	$B(2),\ C(1),\ D(2),\ E(2),\ c_1 = -1$
Convergence on $[0, \infty)$	$q = 1$ ($0 < \alpha \leq 1$)
	$LobattoIIIC(\alpha)$
$\|R(\infty)\|_G < 1$	$0 < \alpha < 1$
$\left\|R(\infty)^2\right\|_G < 1$	$0 < \alpha \leq 1$
$\alpha_D(\widetilde{A}^{-1}) > 0$	$0 < \alpha \leq 1$
Simplifying conditions	$B(2),\ C(1),\ D(2),\ E(1),\ c_{1,2} = -1, 1,\ \tilde{a}_{22} = 1$
Convergence on $[0, \infty)$	$q = 1$ ($0 < \alpha \leq 1$)

Theorems 4 and 10 it can be deduced that they are strictly contractive and convergent of order two on semi-infinite intervals whenever $0 < \alpha < 1$.

Under the same simplifying conditions above, Li [17] derived six classes of high order algebraically stable and B-convergent Runge–Kutta multistep methods. In particular, those methods belonging to the classes 2–4 (see [17, p. 1491]) can be respectively regarded as generalizations of the one-step Runge–Kutta RadauIIA, RadauIA, and LobattoIIIC methods. Here, we have studied in detail the cases in [17, p. 1491] for which $k = s = 2$. In Table 1, we have considered particular uniparametric families belonging to the classes 2, 3, and 4 in [17], denoted, respectively, by $RadauIIA(\alpha)$, $RadauIA(\alpha)$, and $LobattoIIIC(\alpha)$, where $0 < \alpha_2 = \alpha \leq 1$, $\alpha_1 = 1 - \alpha$, with α as a free parameter. The main features concerning the derivation of the methods, the strict-contractivity and its convergence order on semi-infinite intervals, are displayed in Table 1. All these methods are algebraically stable whenever $0 < \alpha \leq 1$, with matrices $G = (1, \alpha)$ and $D = (\gamma_1, \gamma_2)$. The weights $\{\gamma_j,\, j = 1, 2\}$ are readily obtained from the simplifying conditions in Table 1. The order of convergence is deduced from Theorem 10. It should be mentioned that one more order than the stage order cannot be guaranteed since the condition (5.13) is never satisfied.

7. Concluding remarks. New results related to the strict-contractivity and the convergence of General Linear Methods on the class of strictly dissipative problems have been derived. A previous result of strict-contractivity on Runge–Kutta methods by Hairer and Zennaro [11] has been generalized to the class of General Linear Methods. On the other hand, the convergence results on Runge–Kutta methods by Frank, Schneid, and Ueberhuber [9] and on General Linear Methods by Huang, Chang, and Xiao [12] are extended to semi-infinite intervals. The new convergence results meet applications on interesting methods appearing in the literature.

REFERENCES

[1] K. Burrage, *High order algebraically stable multistep Runge–Kutta methods*, SIAM J. Numer. Anal., 24 (1987), pp. 106–115.
[2] K. Burrage, *Order properties of implicit multivalue methods for ordinary differential equations*, IMA J. Numer. Anal., 8 (1988), pp. 43–69.
[3] K. Burrage and J. C. Butcher, *Non-linear stability of a general class of differential equation methods*, BIT, 20 (1980), pp. 185–203.
[4] J. C. Butcher, *Linear and non-linear stability for General Linear Methods*, BIT, 27 (1987), pp. 182–189.
[5] J. C. Butcher, *The equivalence of algebraic stability and AN-stability*, BIT, 27 (1987), pp. 510–533.
[6] M. Crouzeix, W. H. Hundsdorfer, and M. N. Spijker, *On the existence of solutions to the algebraic equations in implicit Runge-Kutta methods*, BIT, 23 (1983), pp. 84–91.
[7] G. Dahlquist, *Error analysis for a class of methods for stiff nonlinear initial value problems*, in Numerical Analysis (Dundee, 1975), Lecture Notes in Math. 506, Springer-Verlag, Berlin, 1976, pp. 60–74.
[8] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, CWI Monograph, North–Holland, Amsterdam, 1984.
[9] R. Frank, J. Schneid, and C. W. Ueberhuber, *Order results for implicit Runge–Kutta methods applied to stiff systems*, SIAM J. Numer. Anal., 22 (1985), pp. 515–534.
[10] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, New York, 1996.
[11] E. Hairer and M. Zennaro, *On error growth functions of Runge-Kutta methods*, Appl. Numer. Math., 22 (1996), pp. 205–216.
[12] C. Huang, Q. Chang, and A. Xiao, *B-convergence of General Linear Methods for stiff problems*, Appl. Numer. Math., 47 (2003), pp. 31–44.

[13] C. Huang, S. Li, H. Fu, and G. Chen, *Nonlinear stability of General Linear Methods for delay differential equations*, BIT, 42 (2002), pp. 380–392.
[14] W. Hundsdorfer, *On the error of General Linear Methods for stiff dissipative differential equations*, IMA J. Numer. Anal., 14 (1994), pp. 363–379.
[15] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis*, revised English ed., Dover Publications, New York, 1975.
[16] J. F. B. M. Kraaijevanger and J. Schneid, *On the unique solvability of the Runge-Kutta equations*, Numer. Math., 59 (1991), pp. 129–157.
[17] S. Li, *Stability and B-convergence properties of multistep Runge-Kutta methods*, Math. Comp., 69 (2000), pp. 1481–1504.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 986–1004

OPTIMAL SUPERCONVERGENCE RESULTS FOR DELAY INTEGRO-DIFFERENTIAL EQUATIONS OF PANTOGRAPH TYPE*

HERMANN BRUNNER† AND QIYA HU‡

Abstract. We analyze the optimal (global and local) orders of superconvergence of collocation solutions u_h on uniform meshes I_h for delay Volterra integro-differential equations with proportional delay functions given by $\theta(t) = qt$ $(0 < q < 1,\ t \in [0,T])$. In particular, we show that if u_h is a continuous piecewise polynomial of degree $m \geq 2$, and if collocation is at the Gauss (–Legendre) points, then the (optimal) order of local superconvergence on I_h is $p^* = m+2$. It turns out that the same order p^* holds for nonlinear (strictly increasing) delay functions vanishing at $t = 0$. However, on judiciously chosen geometric meshes, collocation at the Gauss points yields the order $2m - \varepsilon_N$, where $\varepsilon_N \to 0$ as the number N of mesh points tends to infinity. Optimal local superconvergence results for the pantograph delay differential equation are obtained as special cases of our general analysis.

Key words. Volterra integro-differential equation, vanishing delays, proportional delays, pantograph equation, collocation solutions, optimal order of superconvergence

AMS subject classifications. 65R20, 34K06, 34K28

DOI. 10.1137/060660357

1. Introduction.

It is well known that collocation in the space of continuous piecewise polynomials of degree $m \geq 1$ for first-order delay differential and integro-differential equations with *nonvanishing delays* leads to (optimal) $\mathcal{O}(h^{2m})$-superconvergence at the mesh points of a suitably chosen (constrained) mesh I_h if the collocation points are the Gauss (–Legendre) points (see, e.g., Bellen [3], Bellen and Zennaro [7], Brunner [8, 11], and Chapter 4 of the monograph [12]). For the prototype of a functional differential equation with *vanishing delay*, the pantograph equation,

$$y'(t) = a(t)y(t) + b(t)y(qt), \quad t \in I := [0,T] \quad (0 < q < 1) \tag{1.1}$$

(first analyzed by Fox et al. [18] and Kato and McLeod [29]; see also the survey paper [23] by Iserles); this high order of local superconvergence can be attained on special *geometric meshes* (Bellen [4]). The same is true for its generalization, the pantograph Volterra integro-differential equation

$$y'(t) = a(t)y(t) + b(t)y(qt) + g(t) + \int_0^t K_0(t,s)y(s)ds + \int_0^{qt} K_1(t,s)y(s)ds, \quad t \in I \tag{1.2}$$

*Received by the editors May 19, 2006; accepted for publication (in revised form) December 1, 2006; published electronically May 4, 2007.
http://www.siam.org/journals/sinum/45-3/66035.html

†Department of Mathematics and Statistics, Memorial University of Newfoundland, St. John's, NL, A1C 5S7 Canada (hermann@math.mun.ca). This author's research was supported by the Natural Sciences and Engineering Research Council of Canada (Discovery grant 9406).

‡LSEC and Institute of Computational Mathematics and Scientific/Engineering Computing, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100080, China (hqy@lsec. cc.ac.cn). This author's research was supported by the Natural Science Foundation of China G10371129, the Key Project of the Natural Science Foundation of China G10531080, and the National Basic Research Program of China 2005CB321702.

(Bellen et al. [5]).

On *uniform meshes* the problem regarding optimal (local) superconvergence of collocation solutions for pantograph-type functional differential equations, and in particular for the pantograph equation itself, has remained open (compare the survey paper [10]). An indication that the "classical" superconvergence results might not remain valid for pantograph-type differential (and integral) equations was given in Brunner [9], Takama, Muroya, and Ishiwata [33], Ishiwata [28], and Muroya, Ishiwata, and Brunner [32]: it was shown in these papers that the collocation solution corresponding to the m Gauss points does not exhibit the (local) order $2m+1$ at $t = t_1 = h$, in contrast to ordinary differential equations or delay differential equations with constant delay.

It is the aim of the present paper to show, employing techniques rather different from those in [13], that for $m > 2$ the attainable order of superconvergence at the points of a uniform mesh $I_h = \{t_n = nh : \ 0 \le n \le N \ (t_N = T)\}$ (with $N \ge 2$) cannot exceed $p^* := m+2$, and that this optimal value p^* can be attained for *any* $q \in (0,1)$ and all $m \ge 2$. This is in sharp contrast to the optimal local superconvergence result in (iterated) collocation solutions for pantograph-type Volterra *integral* equations: Brunner and Hu [13] have shown that the optimal local order $p^* = m+2$ can be attained *only* when $q = 1/2$ and m is even.

2. Volterra integro-differential equations of pantograph type.

2.1. Regularity and representation of solutions. Consider the general linear pantograph Volterra integro-differential equation

$$y'(t) = a(t)y(t) + b(t)y(\theta(t)) + g(t) + (\mathcal{V}y)(t) + (\mathcal{V}_\theta y)(t), \quad t \in I := [0,T], \tag{2.1}$$

with initial condition $y(0) = y_0$ and delay function $\theta(t) := t - \tau(t)$ satisfying $\theta(0) = 0$ and $\theta(t) < t$ when $t \in (0,T]$ (see also (D1)–(D3) below). The Volterra integral operators $\mathcal{V}$ and $\mathcal{V}_\theta$ are defined by

$$(\mathcal{V}y)(t) := \int_0^t K_0(t,s)y(s)ds, \quad K_0 \in C(D), \tag{2.2}$$

$$(\mathcal{V}_\theta y)(t) := \int_0^{\theta(t)} K_1(t,s)y(s)ds, \quad K_1 \in C(D_\theta), \tag{2.3}$$

where $D := \{(t,s) : \ 0 \le s \le t \le T\}$. We set

$$D_\theta^{(k)} := \{(t,s) : \ 0 \le s \le \theta^k(t), \ t \in I\} \quad (k \ge 1) \quad \text{and} \quad D_\theta := D_\theta^{(1)}.$$

$\theta(t) = qt$ $(0 < q < 1)$, and we will usually write $D_\theta^{(k)} = D_q^{(k)}$.

The delay function $\theta(t) = t - \tau(t)$ is subject to the following conditions:

(D1) $\theta \in C^d(I)$, with $d \ge 1$, and $\theta(0) = 0$;
(D2) $\theta(t) \le q_1 t$ $(t \in I)$, with $q_1 < 1$; and
(D3) $\min_{t\in I} \theta'(t) =: q_0 > 0$, with $q_0 \le q_1$.

Equation (2.1) includes two important special cases: the *pantograph equation* with variable coefficients, (1.1) (which corresponds to $\theta(t) = qt = t - (1-q)t$), and the delay Volterra integro-differential equation with "pure delay,"

$$y'(t) = b(t)y(\theta(t)) + g(t) + (\mathcal{V}_\theta y)(t), \quad t \in I, \tag{2.4}$$

which exhibits all the essential (quantitative) properties of (2.1) on which the subsequent analysis will focus.

It is readily verified that the initial-value problem for (2.1) is equivalent to the delay integral equation

$$y(t) = g_0(t) + \int_0^t H_0(t,s)y(s)ds + \int_0^{\theta(t)} H_1(t,s)y(s)ds, \quad t \in I, \tag{2.5}$$

where we have set

$$g_0(t) := y_0 + \int_0^t g(s)ds, \tag{2.6}$$

$$H_0(t,s) := a(s) + \int_s^t K_0(v,s)dv, \tag{2.7}$$

$$H_1(t,s) := b(\theta^{-1}(s))\theta'(\theta^{-1}(s)) + \int_{\theta^{-1}(s)}^t K_1(v,s)dv. \tag{2.8}$$

In complete analogy with the pantograph equation (1.1), smooth data in (2.1) lead to smooth solutions on I. This is made precise in the following theorem.

THEOREM 2.1. *Assume that, for some $d \geq 0$, the given functions in* (2.1) *satisfy*
(i) $a,\ b,\ g \in C^d(I)$;
(ii) $K_0 \in C^d(D),\ K_1 \in C^d(D_\theta)$; *and*
(iii) (D1)–(D3), *with* $\theta \in C^{d+1}(I)$.
Then for each $y_0 \in \mathbb{R}$ the initial-value problem for (2.1) *has a unique solution $y \in C^{d+1}(I)$.*

The proof of this existence and regularity result (which generalizes Theorems 5.16 and 5.18 in Brunner [12]) follows readily from the Picard iteration process applied to the delay integral equation (2.5). For the delay Volterra integro-differential equation (2.4)—on which our subsequent analysis will focus—we obtain the following theorem on the representation of solutions. This result is an obvious generalization of the one for $\theta(t) = qt$ $(0 < q < 1)$ derived by Chambers [17]; in fact, this result is already implicitly contained in the 1914 paper [1] by Andreoli.

THEOREM 2.2. *Under the assumptions of Theorem* 2.1, *the unique solution of the initial-value problem for the delay Volterra integro-differential equation* (2.4) *can be represented in the form*

$$y(t) = g_0(t) + \sum_{k=1}^{\infty} \int_0^{\theta^k(t)} H_k(t,s)g_0(s)ds, \quad t \in I, \tag{2.9}$$

where $\theta^k := \underbrace{\theta \circ \cdots \circ \theta}_{k}$. The iterated kernels $H_k(t,s)$ of the kernel $H_1(t,s)$ defined in (2.8) *are determined recursively by*

$$H_{k+1}(t,s) := \int_{\theta^{-k}(s)}^{\theta(t)} H_1(t,v)H_k(v,s)dv \tag{2.10}$$

$$= \int_{\theta^{-1}(s)}^{\theta^k(t)} H_k(t,v)H_1(v,s)dv, \quad (t,s) \in D_\theta^{(k+1)}\ (k \geq 1). \tag{2.11}$$

An alternative form of the solution representation (2.9) *is*

$$y(t) = \left(1 + \sum_{k=1}^{\infty} \tilde{H}_k(t,0)\right) y_0 + \int_0^t g(s)ds + \sum_{k=1}^{\infty} \int_0^{\theta^k(t)} \tilde{H}_k(t,s)g(s)ds, \tag{2.12}$$

with

$$\tilde{H}_k(t,s) := \int_s^{\theta^k(t)} H_k(t,v)dv, \quad (t,s) \in D_\theta^{(k)}. \tag{2.13}$$

Proof. The only new ingredient in Theorem 2.2 is the alternative expression (2.11) for the iterated kernel $H_{k+1}(t,s)$. It is readily verified by induction, using an obvious change in the order of integration (which is based on assumption (D3) for θ). We will illustrate this by considering $k = 2$. By (2.10),

$$H_3(t,s) = \int_{\theta^{-2}(s)}^{\theta(t)} H_1(t,v)H_2(v,s)dv,$$

and hence,

$$\begin{aligned} H_3(t,s) &= \int_{\theta^{-2}(s)}^{\theta(t)} H_1(t,v)\left(\int_{\theta^{-1}(s)}^{\theta(v)} H_1(v,z)H_1(z,s)dz\right)dv \\ &= \int_{\theta^{-1}(s)}^{\theta^2(t)} \left(\int_{\theta^{-1}(z)}^{\theta(t)} H_1(t,v)H_1(v,z)dv\right) H_1(z,s)dz \\ &= \int_{\theta^{-1}(s)}^{\theta^2(t)} H_2(t,z)H_1(z,s)dz, \quad (t,s) \in D_\theta^{(3)}. \end{aligned}$$

The completion of the induction argument is now clear. The representation (2.12) follows from (2.9) and the definition (2.6) of $g_0(t)$. □

COROLLARY 2.3. *Let* $\theta(t) = qt,\ 0 < q < 1$, *and define*

$$\beta := q^{-1}\max\{|b(t)| :\ t \in I\}, \quad \bar{K}_1 := \max\{|K_1(t,s)| :\ (t,s) \in D_q\}.$$

Then the iterated kernels $\{H_k(t,s)\}$ *corresponding to the kernel* $K_1(t,s)$ *in* (2.4), (2.8) *are bounded by*

$$|H_k(t,s)| \le \frac{(\beta + \bar{K}_1 T)^k}{(k-1)!} q^{(k-1)(k-2)/2}[qt - q^{-(k-1)}s]^{k-1}, \tag{2.14}$$
$$(t,s) \in D_q^{(k)} \quad (k \ge 1).$$

Moreover, the kernels $\{\tilde{H}_k(t,s)\}$ *defined in* (2.13) *satisfy*

$$|\tilde{H}_k(t,s)| \le \frac{(\beta + \bar{K}_1 T)^k}{k!} q^{k(k-1)/2}[qt - q^{-(k-1)}s]^k, \quad (t,s) \in D_q^{(k)} \quad (k \ge 1). \tag{2.15}$$

The estimates (2.14) follow directly from the expression (2.11) for the iterated kernels, using an inductive argument and integration by parts. Hence, the estimates for the integrated kernels $\tilde{H}_k(t,s)$ are an obvious consequence of (2.13).

COROLLARY 2.4. *If* $K_1(t,s) \equiv 0$, *then the iterated kernels associated with the delay integral equation* (2.5) *corresponding to the pantograph equation* (1.1) *satisfy*

$$|H_k(t,s)| \le \frac{\beta^k q^{(k-1)(k-2)/2}}{(k-1)!}[qt - q^{-(k-1)}s]^{k-1}, \quad (t,s) \in D_q^{(k)} \quad (k \ge 1). \tag{2.16}$$

For the sake of completeness we conclude this section by observing that the results of Theorem 2.2 on the representation of the solution to (2.4) are easily extended to

the general pantograph equation (2.1), by applying Picard iteration to its equivalent delay integral equation (2.5). The precise result is given in the following theorem.

THEOREM 2.5. *If the given functions in the general pantograph Volterra integro-differential equation* (2.1) *are continuous on their respective domains, and if the delay function* θ *is subject to* (D1)–(D3)*, then its unique solution* $y \in C^1(I)$ *has the representation*

$$\begin{aligned} y(t) &= g_0(t) + \int_0^t \sum_{k=1}^{\infty} H_{0,k}(t,s) g_0(s) ds \\ &\quad + \sum_{k=1}^{\infty} \int_0^{\theta^k(t)} H_{1,k}(t,s) g_0(s) ds + \mathcal{M}(t), \quad t \in I, \end{aligned} \tag{2.17}$$

where

$$\mathcal{M}(t) := \sum_{k=1}^{\infty} \int_0^{\theta^k(t)} H_k^{(0,1)}(t,s) g_0(s) ds, \tag{2.18}$$

and $\theta^0(t) := t$*. The kernels* $H_{0,k}(t,s)$ *and* $H_{1,k}(t,s)$ *are the iterated kernels associated with the kernels* $H_0(t,s)$ *and* $H_1(t,s)$ *defined in* (2.7) *and* (2.8)*, and* g_0 *is given by* (2.6)*. The infinite series in* (2.17) *and* (2.18) *converge absolutely and uniformly on* I*, and we have* $H_k^{(0,1)}(t,s) \equiv 0$ $(k \geq 1)$ *when* $H_0(t,s) \equiv 0$ *or* $H_1(t,s) \equiv 0$*.*

Proof. As indicated before, the representation (2.17) follows readily from Picard iteration applied to the integral equation (2.5); here, the following lemma (an obvious extension of Dirichlet's formula regarding the change in the order of integration in integrals with variable limits of integration) plays a key role. ∎

LEMMA 2.6. *Let* k *and* ℓ *be given nonnegative integers, and assume that the delay function* θ *is subject to the hypotheses* (D1)–(D3)*. Then for any function* $\phi \in C(D_\theta^{(k+\ell)})$*,*

$$\int_0^{\theta^k(t)} \left(\int_0^{\theta^\ell(s)} \phi(s,v) dv \right) ds = \int_0^{\theta^{k+\ell}(t)} \left(\int_{\theta^{-\ell}(v)}^{\theta^k(t)} \phi(s,v) ds \right) dv, \quad (t,s) \in D_\theta^{(k+\ell)}. \tag{2.19}$$

We leave the details of the proof of Theorem 2.5 to the reader but illustrate the basic result underlying its induction argument, thus revealing the structure of the kernels $H_k^{(0,1)}(t,s)$ in (2.18). Setting $y_0(t) := g_0(t)$ and $H_{0,1}(t,s) := H_0(t,s)$, $H_{1,1}(t,s) := H_1(t,s)$, the first two iterates obtained by Picard iteration for (2.5) are, respectively,

$$y_1(t) = g_0(t) + \int_0^t H_{0,1}(t,s) g_0(s) ds + \int_0^{\theta(t)} H_{1,1}(t,s) g_0(s) ds,$$

and hence, by Lemma 2.6,

$$\begin{aligned} y_2(t) &= g_0(t) + \int_0^t H_{0,1}(t,s) y_1(s) ds + \int_0^{\theta(t)} H_{1,1}(t,s) y_1(s) ds \\ &= g_0(t) + \int_0^t H_{0,1}(t,s) \Bigg(g_0(s) + \int_0^s H_{0,1}(s,v) g_0(v) dv \\ &\quad + \int_0^{\theta(s)} H_{1,1}(s,v) g_0(v) dv \Bigg) ds \end{aligned}$$

$$+\int_0^{\theta(t)} H_{1,1}(t,s)\left(g_0(s)+\int_0^s H_{0,1}(s,v)g_0(v)dv+\int_0^{\theta(s)} H_{1,1}(s,v)g_0(v)dv\right)ds$$
$$=g_0(t)+\int_0^t H_{0,1}(t,s)g_0(s)ds+\int_0^t H_{0,2}(t,s)g_0(s)ds$$
$$+\int_0^{\theta(t)} H_{1,1}(t,s)g_0(s)ds+\int_0^{\theta^2(t)} H_{1,2}(t,s)g_0(s)ds+\int_0^{\theta(t)} H_2^{(0,1)}(t,s)g_0(s)ds.$$

Here, $H_{0,2}(t,s)$ and $H_{1,2}(t,s)$ are the first nontrivial iterated kernels for the given kernels $H_0(t,s)$ and $H_1(t,s)$:

$$H_{0,2}(t,s):=\int_s^t H_{0,1}(t,v)H_{0,k-1}(v,s)dv,$$
$$H_{1,2}(t,s):=\int_{\theta^{-(k-1)}(s)}^{\theta(t)} H_{1,1}(t,v)H_{1,k-1}(v,s)dv$$

(cf. (2.10)), and the "mixed" kernel $H_2^{(0,1)}(t,s)$ has the form

$$H_2^{(0,1)}(t,s):=\int_{\theta^{-1}(s)}^t H_{0,1}(t,v)H_{1,1}(v,s)dv+\int_v^{\theta(t)} H_{1,1}(t,v)H_{0,1}(v,s)dv.$$

2.2. Collocation methods in piecewise polynomial spaces. As we have seen in section 2.1 (Theorem 2.1), the analytical solution to the pantograph integro-differential equation (2.1) with smooth data is smooth on the entire interval $I:=[0,T]$, in contrast to solutions to such functional equations with nonvanishing delays. Thus, the mesh underlying the piecewise polynomial space in which the approximation u_h to the exact solution will be sought will be chosen as

$$I_h:=\{t_n:\ 0=t_0<t_1<\cdots<t_N=T\}\quad(\text{with }\ t_n=t_n^{(N)}),\tag{2.20}$$

and we set $h_n:=t_{n+1}-t_n,\ h:=\max\{h_n:\ 0\le n\le N-1\}$. For given I_h and $m\ge 1$, the collocation solution u_h to (2.1) will be an element of the space

$$S_m^{(0)}(I_h):=\{v\in C^0(I):\ v|_{[t_n,t_{n+1}]}\in\pi_m\ (0\le n\le N-1)\}\tag{2.21}$$

of continuous (real) piecewise polynomials of degree not exceeding m; u_h is determined by the collocation equation

$$u_h'(t)=a(t)u_h(t)+b(t)u_h(\theta(t))+g(t)+(\mathcal{V}u_h)(t)+(\mathcal{V}_\theta u_h)(t),\quad t\in X_h,\ \ u_h(0)=y_0.\tag{2.22}$$

Here,

$$X_h:=\{t_n+c_ih_n:\ 0<c_1<\cdots<c_m\le 1\ (0\le n\le N-1)\}\tag{2.23}$$

is the set of collocation points corresponding to given collocation parameters $\{c_i\}$. For continuous data there exists $\bar h>0$ so that (2.19) has a unique solution $u_h\in S_m^{(0)}(I_h)$ for all meshes I_h with $h\in(0,\bar h)$ (see, e.g., [12, section 5.5.1]).

It is also known [12, section 5.5.2] that, for arbitrarily chosen collocation parameters $\{c_i\}$, the collocation error tends to zero uniformly, as $h\to 0$. More precisely, we

have the following global convergence result (which we state for further reference); its proof can be found in [12, Ch. 5].

THEOREM 2.7. *Assume that* a, b, g, K_0, K_1 *are in* C^m *on their respective domains, and let* θ *be subject to the conditions* (D1)–(D3), *with* $d \geq m+1$ *(cf. section* 2.1*). Then for* $h \in (0, \bar{h})$ *the collocation solution* $u_h \in S_m^{(0)}(I_h)$ *for* (2.1) *satisfies*

$$\|y^{(\nu)} - u_h^{(\nu)}\|_\infty \leq C_\nu h^m \quad (\nu = 0, 1), \tag{2.24}$$

where the constants C_ν *depend on* $\{c_i\}$ *and on* q *but not on* h.

The collocation solution $u_h \in S_m^{(0)}(I_h)$ for (2.1) induces the *defect* (or residual) δ_h defined by

$$\delta_h(t) := -u_h'(t) + a(t)u_h(t) + b(t)u_h(\theta(t)) + g(t) + (\mathcal{V}u_h)(t) + (\mathcal{V}_\theta u_h)(t), \quad t \in I, \tag{2.25}$$

with $\delta_h(t) = 0$ for all $t \in X_h$. It inherits (piecewise, on the subintervals $[t_n, t_{n+1}]$) the regularity of the given functions in (2.1). Since $\delta_h(t)$ can also be written as

$$\delta_h(t) = e_h'(t) - a(t)e_h(t) - b(t)e_h(\theta(t)) - (\mathcal{V}e_h)(t) - (\mathcal{V}_\theta e_h)(t), \quad t \in I, \tag{2.26}$$

where $e_h := y - u_h$, the following result is an immediate consequence of Theorem 2.7.

COROLLARY 2.8. *Under the assumptions of Theorem* 2.7 *we have*

$$\|\delta_h\|_\infty \leq D_0 h^m$$

for all $h \in (0, \bar{h})$, *where the constant* D_0 *does not depend on* h.

3. Optimal global superconvergence on I.

THEOREM 3.1. *Assume the following:*

(i) *The given functions in the pantograph Volterra integro-differential equation* (2.1) *satisfy, for* κ *specified in* (3.1), a, b, $g \in C^{m+\kappa}(I)$, $K_0 \in C^{m+\kappa}(D)$, $K_1 \in C^{m+\kappa}(D_\theta)$, *and* $\theta(t) = qt$ $(0 < q < 1)$.

(ii) $u_h \in S_m^{(0)}(I_h)$ *is the collocation solution, with respect to the uniform mesh* I_h *and collocation parameters* $\{c_i\}$, *to* (2.1).

(iii) *The collocation parameters are such that, for some* κ *with* $1 \leq \kappa \leq m$,

$$J_\nu := \int_0^1 s^\nu \prod_{i=1}^m (s - c_i)ds = 0, \quad \nu = 0, \ldots, \kappa - 1; \tag{3.1}$$

that is, the interpolatory m*-point quadrature based on the abscissas given by the collocation parameters* $\{c_i\}$ *has degree of precision* $m + \kappa - 1$.

Then u_h *is globally superconvergent:*

$$\|y - u_h\|_\infty \leq Ch^{m+1},$$

where in general the order $m+1$ *cannot be replaced by* $m+2$. *The constant* C *depends on* q *and on* $\{c_i\}$ *but not on* h.

Proof. For ease of exposition we will give the proof for (2.4); it is readily adapted (using the representation (2.17) in Theorem 2.5) to the general pantograph integro-differential equation (2.1). It follows from (2.4) and (2.25) (with $a = 0$, $\mathcal{V} = 0$) that the collocation error $e_h := y - u_h$ solves the initial-value problem

$$e_h'(t) = b(t)e_h(qt) + \delta_h(t) + (\mathcal{V}_\theta e_h)(t), \quad t \in I, \quad e_h(0) = 0. \tag{3.2}$$

Hence, by (2.12) of Theorem 2.2 (where the roles of y and g are now assumed by e_h and δ_h, respectively),

$$e_h(t) = \int_0^t \delta_h(s)ds + \sum_{k=1}^{\infty} \int_0^{q^k t} \tilde{H}_k(t,s)\delta_h(s)ds, \quad t \in I. \tag{3.3}$$

Let now $t = t_n + vh$ ($v \in [0,1]$), and define

$$\begin{aligned} q_{k,n}(v) &:= \lfloor q^k(n+v) \rfloor, \quad \gamma_{k,n} = \gamma_{k,n}(v) := q^k n - q_{k,n}, \\ k_n^*(q) &:= \max\{k : \ q_{k,n}(v) \geq 1\}. \end{aligned} \tag{3.4}$$

Thus, (3.3) can be written in the form

$$e_h(t) = h\sum_{\ell=0}^{n-1} \int_0^1 \delta_h(t_\ell + sh)ds + h\int_0^v \delta_h(t_n + sh)ds + S_n^I(v) + S_n^{II}(v), \tag{3.5}$$

where we have set

$$S_n^I(v) := \sum_{k=1}^{k_n^*(q)} \int_0^{t_{q_{k,n}}} \tilde{H}_k(t,s)\delta_h(s)ds = h\sum_{k=1}^{k_n^*(q)} \left(\sum_{\ell=0}^{q_{k,n}-1} \int_0^1 \tilde{H}_k(t, t_\ell + sh)\delta_h(t_\ell + sh)ds \right) \tag{3.6}$$

and

$$S_n^{II}(v) := h\sum_{k=1}^{\infty} \int_0^{\gamma_{n,k}} \tilde{H}_k(t, t_{q_{k,n}} + sh)\delta_h(t_{q_{k,n}} + sh)ds. \tag{3.7}$$

In order to derive the (optimal) order estimate for $S_n^I(v)$, we first observe that, using m-point interpolatory quadrature formulas, with abscissas based on the m collocation parameters $\{c_i\}$ and with $E_{n,\ell}(v)$ and $\tilde{E}_{n,\ell}^{(k)}(v)$ denoting, respectively, the resulting quadrature errors for the integrals with integrands $\delta_h(t_\ell + sh)$ and those with integrands $\tilde{H}_k(t, t_\ell + sh)\delta_h(t_\ell + sh)$, and observing that on each subinterval $[t_n, t_{n+1}]$ the defect δ_h is in $C^{m+\kappa}$, we may write

$$|S_n^I(v)| \leq h\sum_{k=1}^{k_n^*(q)} \left(\sum_{\ell=0}^{q_{k,n}-1} |\tilde{E}_{n,\ell}^{(k)}(v)| \right). \tag{3.8}$$

By assumption (3.1) on the collocation parameters $\{c_i\}$, these quadrature formulas have degree of precision $m + \kappa - 1$, and hence, by the regularity of the integrands on the subintervals $[t_n, t_{n+1}]$,

$$|E_{n,\ell}(v)| \leq Q_m h^{m+\kappa}, \quad |\tilde{E}_{n,\ell}^{(k)}(v)| \leq \tilde{Q}_m h^{m+\kappa}.$$

We therefore obtain

$$|S_n^I(v)| \leq h \cdot \tilde{Q}_m h^{m+\kappa} \sum_{k=1}^{k_n^*(q)} q_{k,n}(v) \leq Nh \cdot \tilde{Q}_m h^{m+\kappa} \frac{q}{1-q} = T\tilde{Q}_m h^{m+\kappa} \frac{q}{1-q} \tag{3.9}$$

($0 \leq n \leq N-1$), since, by (3.4), $q_{k,n} \leq q^k n \leq q^k N$. ∎

Consider now $S_n^{II}(v)$: it follows from (3.7), Corollary 2.8, and (2.15) that

$$|S_n^{II}(v)| \le h\|\delta_h\|_\infty \sum_{k=1}^{\infty} \frac{(\beta + \bar{K}_1 T)^k}{k!} q^{k(k-1)/2} \int_0^{\gamma_{k,n}} [qt - q^{-(k-1)}s]^k ds$$

$$\le h\|\delta_h\|_\infty \sum_{k=1}^{\infty} \frac{(\beta + \bar{K}_1 T)^k T^{k+1}}{(k+1)!} q^{k(k+3)/2} =: D_0 \tilde{B}(q) h^{m+1}, \quad v \in [0,1] \tag{3.10}$$

$(0 \le n \le N-1)$. Thus, (3.5) together with (3.9) and (3.10) lead to the (optimal) global superconvergence estimate

$$|e_h(t)| \le Ch^{m+1}, \quad \text{with} \quad C := (TQ_m + D_0) + (T\tilde{Q}_m q/(1-q) + D_0\tilde{B}(q)),$$

which holds uniformly for $t \in I$, and for any κ between 1 and m in (3.1).

Owing to the smoothness of the exact solution of the pantograph integro-differential equation (2.1), the optimal global convergence estimate in Theorem 3.1 is not surprising: it agrees with the one for classical and constant-delay Volterra integral equations. However, the picture changes completely for the optimal order of the collocation solution for (2.1) at the mesh points of a uniform mesh, as shown in the following section.

4. Optimal local superconvergence on uniform meshes. It was shown in Brunner and Hu [13] that for Volterra *integral* equations of pantograph type, the optimal order of local superconvergence at the points of a uniform mesh cannot exceed $p^* = m+2$, and that the optimal value is attained only when $q = 1/2$ and m is even. For Volterra *integro-differential* equations of pantograph type, we obtain the same value of p^* but it is now attained for all $q \in (0,1)$ and all $m \ge 2$. This result, stated in the following theorem, provides the (affirmative) answer to a conjecture in [11, section 4] and [12, section 5.5.2]. We will comment on the reason for this difference in the optimal local superconvergence orders following the proof of Theorem 4.1.

THEOREM 4.1. *Assume the following:*

(i) *The given functions in* (2.1) *satisfy* a, b, $g \in C^{m+\kappa}(I)$, $K_0 \in C^{m+\kappa}(D)$, *and* $K_1 \in C^{m+\kappa}(D_\theta)$ *for some* κ *with* $1 \le \kappa \le m$.

(ii) $\theta(t) = qt$ $(0 < q < 1)$.

(iii) $u_h \in S_m^{(0)}(I_h)$ *is the collocation solution to* (2.1) *on a uniform mesh* I_h *and corresponding to collocation parameters* $\{c_i\}$ *satisfying the orthogonality conditions* (3.1) *with* $1 \le \kappa \le m$.

Then for any $q \in (0,1)$ *and any* $m \ge 2$,

$$\max\{|y(t) - u_h(t)| : \ t \in I_h\} \le C^* h^{m+2}, \tag{4.1}$$

and the exponent $m+2$ *cannot, in general, be replaced by* $m+3$.

Proof. The starting point is again the error equation (3.2): for the representation of its solution we now use (2.9), where the roles of g in (2.6) and those of y are assumed by δ_h and e_h, respectively. For a given mesh point $t = t_n$ $(n = 1, \ldots, N)$ we set, in analogy to (3.4),

$$q_{k,n} := \lfloor q^k n \rfloor, \quad \gamma_{k,n} := q^k n - q_{k,n}, \quad k_n^*(q) := \max\{k : \ q_{k,n} \ge 1\}. \tag{4.2}$$

Thus, the counterpart of (3.5) is given by

$$e_h(t_n) = \int_0^{t_n} \delta_h(s) ds + S_n^I + S_n^{II}, \tag{4.3}$$

where

$$S_n^I := \sum_{k=1}^{k_n^*(q)} \int_0^{t_{q_{k,n}}} H_k(t_n, s) \left(\int_0^s \delta_h(v)dv \right) ds \tag{4.4}$$
$$= h \sum_{k=1}^{k_n^*(q)} \sum_{\ell=0}^{q_{k,n}-1} H_k(t_n, t_\ell + sh) \left(h \sum_{\nu=0}^{\ell-1} \int_0^1 \delta_h(t_\nu + vh)dv + h \int_0^s \delta_h(t_\ell + vh)dv \right) ds$$

and

$$S_n^{II} := h \sum_{k=1}^{\infty} \int_0^{\gamma_{k,n}} H_k(t_n, t_{q_{k,n}} + sh) \left(\int_0^{t_{q_{k,n}}+sh} \delta_h(v)dv \right) ds \tag{4.5}$$
$$= h \sum_{k=1}^{\infty} \int_0^{\gamma_{k,n}} H_k(t_n, t_{q_{k,n}} + sh) \left(h \sum_{\ell=0}^{q_{k,n}-1} \int_0^1 \delta_h(t_\ell + vh)dv + h \int_0^s \delta_h(t_{q_{k,n}} + vh)dv \right) ds.$$

The techniques for estimating $|S_n^I|$ and $|S_n^{II}|$ closely parallel the ones used in section 3 (cf. (3.8) and (3.9)). Observe first that in the upper bound for $|S_n^I|$ we have

$$h \sum_{\nu=0}^{\ell-1} \left| \int_0^1 \delta_h(t_\nu + vh)dv \right| + h \left| \int_0^s \delta_h(t_\ell + vh)dv \right| \le TQ_m h^{m+\kappa} + h \cdot D_0 h^m =: \tilde{D}_0 h^{m+1}.$$

Hence, by (2.14) of Corollary 2.3 and by observing the factor h in front of the first summation sign in the second line of (4.4) we are led to

$$|S_n^I| \le h \cdot \tilde{D}_0 h^{m+1} \cdot \sum_{k=1}^{k_n^*(q)} \sum_{\ell=0}^{q_{k,n}-1} \int_0^1 |H_k(t_n, t_\ell + sh)|ds$$
$$\le \tilde{D}_0 h^{m+2} \sum_{k=1}^{k_n^*(q)} \frac{(\beta + \bar{K}_1 T)^k T^{k-1}}{(k-1)!} q^{(k-1)(3k-4)/2} =: C_0 h^{m+2}. \tag{4.6}$$

Similarly, (4.5) together with (2.14) and Corollary 2.8 allow us to obtain the estimate

$$|S_n^{II}| \le C_1 h^{m+2} \quad (1 \le n \le N). \tag{4.7}$$

Hence, (4.3) and the two estimates (4.6) and (4.7) yield the desired optimal superconvergence result on I_h,

$$|e_h(t_n)| \le (C_0 + C_1) h^{m+2} =: C^* h^{m+2}, \quad n = 1, \ldots, N;$$

as the above analysis shows, the power h^{m+2} cannot be replaced by h^{m+3}, except in trivial cases. □

The above result answers the question regarding the optimal order of local superconvergence for the pantograph delay differential equation (1.1).

COROLLARY 4.2. *The collocation solution* $u_h \in S_m^{(0)}(I_h)$ $(m \geq 2)$ *for the pantograph delay differential equation* (1.1)*, with uniform mesh* I_h *and collocation points based on the Gauss points* $\{c_i\}$*, has the optimal local superconvergence order* $p^* = m+2$ *on* I_h*:*

$$\max\{|y(t) - u_h(t)| : \ t \in I_h\} \leq Ch^{m+2}.$$

Remark. As we have indicated before, the above optimal local superconvergence result differs sharply from the one corresponding to the collocation solution $u_h \in S_{m-1}^{(-1)}(I_h)$ (the space of piecewise polynomials of degree $m-1 \geq 0$ that are allowed to possess finite jumps at the mesh points) and its iterate $u_h^{it|}$ for the pantograph integral equation

$$u(t) = g(t) + \int_0^{qt} K_1(t,s)u(s)ds, \quad t \in [0,T]. \tag{4.8}$$

Here, the iterated collocation error e_h^{it} $(= e_h - \delta_h)$ at $t = t_n$ has the representation

$$e_h^{it}(t_n) = \sum_{k=1}^{\infty} \int_0^{q^k t_n} H_k(t_n, s)\delta_h(s)ds, \quad n = 1, \ldots, N,$$

which, recalling (4.2), can be written in a form analogous to (4.3), namely, $e_h^{it}(t_n) = \hat{S}_n^I + \hat{S}_n^{II}$. However, the terms $\hat{S}_n^I$ and $\hat{S}_n^{II}$ corresponding to S_n^I and S_n^{II} (cf. (4.4) and (4.5)) no longer contain integrals of the defect function δ_h. In particular, we now have

$$\hat{S}_n^{II} = h\sum_{k=1}^{\infty} \int_0^{\gamma_{k,n}} H_k(t_n, t_{q_{k,n}} + sh)\delta_h(s)ds,$$

where, as before, $\|\delta_h\|_\infty = \mathcal{O}(h^m)$. This implies that if collocation is at the Gauss points, then $\hat{S}_n^{II} = \mathcal{O}(h^{m+2})$ $(n = 1, \ldots, N)$ can now be attained only for special values of q and m, namely, when $q = 1/2$ and m is even. Details can be found in Brunner and Hu [13, p. 1940].

5. Equations with nonlinear vanishing delays vanishing at $t = 0$. We now turn to the superconvergence analysis of collocation solutions for pantograph-type Volterra integro-differential equations (2.1) where delay function given by $\theta(t) = t - \tau(t)$ is nonlinear and satisfies the hypotheses (D1)–(D3). Since (D2) implies the inequalities

$$\theta^k(t) \leq q_1^k t \quad (k \geq 1) \quad \text{and} \quad \theta^{-1}(s) \geq q_1^{-1} s \tag{5.1}$$

(cf. (2.11)), they suggest that our global and local superconvergence results of Theorems 3.1 and 4.1 will remain valid for such nonlinear vanishing delays. The basis for the proofs is given by the following generalization of Corollary 2.3.

LEMMA 5.1. *Let* θ *be subject to the hypotheses* (D1)–(D3) *of section* 2*, and define*

$$\beta := \max\{|b(\theta^{-1}(t))|\theta'(\theta^{-1}(t)) : \ t \in I\}, \quad \bar{K}_1 := \max\{|K_1(t,s)| : \ (t,s) \in D_\theta\}.$$

The iterated kernels associated with the delay integral equation (2.5) *corresponding to the pantograph integro-differential equation* (2.4) *satisfy*

$$|H_k(t,s)| \leq \frac{(\beta + \bar{K}_1 T)^k}{(k-1)!} q_1^{(k-1)(k-2)/2}[q_1 t - q_1^{-(k-1)}s]^{k-1}, \tag{5.2}$$
$$(t,s) \in D_\theta^{(k)} \ \ (k \geq 1).$$

Moreover, the kernels $\{\tilde{H}_k(t,s)\}$ *defined in* (2.13) *satisfy*

$$|\tilde{H}_k(t,s)| \le \frac{(\beta+\bar{K}_1T)^k}{k!} q_1^{k(k-1)/2}[q_1t - q_1^{-(k-1)}s]^k, \quad (t,s)\in D_\theta^{(k)} \ \ (k\ge 1). \tag{5.3}$$

The proof is an immediate consequence of (5.1) (which is based on the condition (D2) for the delay function θ)) and the result of Corollary 2.3, where the role of q is now assumed by q_1.

THEOREM 5.2. *Assume the following:*

(i) *The given functions* a, b, g, K_0, *and* K_1 *in the general Volterra pantograph integro-differential equation* (2.1) *possess continuous derivatives of order* $m+\kappa$ *for some* κ *with* $1\le\kappa\le m$ *on their respective domains.*

(ii) *The delay function* θ *is subject to the hypotheses* (D1)–(D3) *of section* 2, *with* $d\ge m+\kappa+1$.

(iii) $u_h\in S_m^{(0)}(I_h)$ *is the collocation solution, with respect to a uniform mesh* I_h *with sufficiently small mesh diameter* $h>0$, *to the initial-value problem for* (2.1).

(iv) *The collocation parameters* $\{c_i\}$ *in* (2.21) *satisfy* (3.1) *for some* κ *with* $1\le\kappa\le m$.

Then the following hold:

(a) u_h *is globally superconvergent on* I, *with optimal order described by*

$$\|y-u_h\|_\infty \le Ch^{m+1}. \tag{5.4}$$

(b) *For any* $q_1\in(0,1)$ *(cf. hypothesis* (D3)*) and any* $m\ge 2$, u_h *is locally superconvergent at the mesh points*

$$\max\{|y(t)-u_h(t)|:\ t\in I_h\}\le C^*h^{m+2}, \tag{5.5}$$

where in general the order $p^*:=m+2$ *cannot be replaced by* $m+3$.

Proof. We will prove the local superconvergence estimate (5.5); the global estimate (5.4) can be established along very similar lines, as is already suggested by the proof of Theorem 3.1.

By (2.1) and (2.20) the collocation error satisfies the initial-value problem

$$\begin{aligned} &e_h'(t) = a(t)e_h(t)+b(t)e_h(\theta(t))+\delta_h(t)+(\mathcal{V}e_h)(t)+(\mathcal{V}_\theta e_h)(t), \quad t\in I,\\ &e_h(0)=0, \end{aligned}$$

where $\delta_h(t)=0$ on the set X_h of collocation points. Hence, it follows from the solution representation (2.9) in Theorem 2.5 (with e_h and δ_h replacing y and g, respectively) that, at $t=t_n$ $(n=1,\dots,N)$,

$$\begin{aligned} e_h(t_n) &= \int_0^{t_n}\delta_h(s)ds+\sum_{k=0}^{\infty}\int_0^{\theta^k(t_n)}H_k(t_n,s)\left(\int_0^s\delta_h(v)dv\right)ds\\ &= \int_0^{t_n}\delta_h(s)ds+\int_0^{t_n}H_0(t,s)(t_n,s)\left(\int_0^s\delta_h(v)dv\right)ds \qquad (5.6)\\ &\quad +\sum_{k=1}^{\infty}\int_0^{\theta^k(t_n)}H_k(t_n,s)\left(\int_0^s\delta_h(v)dv\right)ds. \qquad (5.7) \end{aligned}$$

For given n and $k\ge 1$ let, in analogy to (4.2), the (unique) $q_{k,n}\in\mathbb{N}$ be such that

$$\theta^k(t_n)\in[t_{q_{k,n}},t_{q_{k,n}+1}), \tag{5.8}$$

and define

$$\gamma_{k,n} := (\theta^k(t_n) - t_{q_{k,n}})/h, \quad k_n^* := \max\{k : \ q_{k,n} \geq 1\}. \tag{5.9}$$

We thus may write, in analogy to (4.3)–(4.5),

$$e_h(t_n) = h\sum_{\ell=0}^{n-1}\int_0^1 \delta_h(t_\ell + sh)ds + h\sum_{\ell=0}^{n-1}\int_0^1 H_0(t_n, t_\ell + sh)\left(\int_0^{t_\ell+sh}\delta_h(v)dv\right)ds$$
$$+ S_n^I + S_n^{II} \quad (n = 1, \ldots, N), \tag{5.10}$$

where

$$S_n^I := \sum_{k=1}^{k_n^*}\int_0^{t_{q_{k,n}}} H_k(t_n, s)\left(\int_0^s \delta_h(v)dv\right)ds \tag{5.11}$$

and

$$S_n^{II} := h\sum_{k=1}^{\infty}\int_0^{\gamma_{k,n}} H_k(t_n, t_{q_{k,n}} + sh)\left(\int_0^{t_{q_{k,n}}+sh}\delta_h(v)dv\right)ds. \tag{5.12}$$

(Recall that by Lemma 5.1 the above infinite series converge uniformly for all $n = 1, \ldots, N$.)

A glimpse at (4.4) and (4.5) now reveals that the estimates (4.6) and (4.7) carry over to S_n^I and S_n^{II} defined in (5.11) and (5.12), except that now, by (5.2) in Lemma 5.1, q has to be replaced by $q_1 \in (0,1)$, corresponding to the hypothesis (D2) for the nonlinear delay function θ. Thus, employing the arguments used in the proof of Theorem 4.1 allows us to derive the optimal estimate (5.5) of Theorem 5.2.

COROLLARY 5.3. *Assume that the delay function θ satisfies the conditions* (D1)–(D3) *of section 2, with $d \geq 2m+1$, and let I_h be a uniform mesh with sufficiently small $h > 0$. If a, b, $g \in C^{2m}(I)$, then the collocation solution $u_h \in S_m^{(0)}(I_h)$ $(m \geq 2)$ for the generalized pantograph delay differential equation*

$$y'(t) = a(t)y(t) + b(t)y(\theta(t)) + g(t), \quad t \in I, \quad y(0) = y_0, \tag{5.13}$$

with collocation points (2.21) *based on the Gauss points $\{c_i\}$, has the optimal local superconvergence order $p^* = m+2$ for all $q \in (0,1)$:*

$$\max\{|y(t) - u_h(t)| : \ t \in I_h\} \leq C^* h^{m+2}.$$

Here, the constant C^* depends on the $\{c_i\}$ and on q_1 but not on h.

Remarks. 1. Recall that the hypothesis (D2) (section 2) requires that $\theta(t) \leq q_1 t$, $t \in I$, for some $q_1 \in (0,1)$. Do the optimal superconvergence estimates (5.4) and (5.5) remain true if the nonlinear delay function θ satisfies only (D1) and (D3) and is such that $\theta'(0) = 1$? Examples of such delay functions are $\theta(t) = q_1\log(1+t)$, $0 < q_1 \leq 1$, for which we have $\theta'(t) = q_1\frac{1}{1+t}$, $t \in [0,T]$, and thus $\theta'(0) = 1$ when $q_1 = 1$; and $\theta(t) = t - t^r$, $r \in \mathbb{N}$ $(r \geq 2)$. In this case, $\theta'(0) = 1$, $\theta^{(\nu)}(0) = 0$ $(\nu = 2, \ldots, r-1)$ (see also [5, section 4]).

The analysis of optimal superconvergence of collocation solutions for (2.1) with delay functions of the above type is yet to be carried out. It appears that the approach in Brunner and Maset [15] will yield the tools to extend the analysis in the present

paper to functional differential and integro-differential equations (2.1) containing these more general θ.

2. A related question concerns delay functions θ that satisfy (D1) and (D3) on $[0,T]$ but have the properties that (i) $\theta(0) = 0$ and $\theta(t) < t$ for $t \in (0, t^*)$, (ii) $\theta(t^*) = t^*$ (i.e., $\tau(t^*) = 0$), and (iii) e.g., (D1)–(D3) hold on $[t^*, T]$. The optimal convergence analysis of (2.1) with such "doubly vanishing" delay functions is studied in [15].

6. Optimal local superconvergence on geometric meshes. We shall now show that, in analogy to pantograph-type Volterra integral equations (Brunner and Hu [13]), $\mathcal{O}(h^{2m})$-superconvergence at the mesh points can (almost) be restored if we replace the uniform mesh I_h by a judiciously chosen *geometric mesh* (see also [21] and [14]). To be precise, we shall seek the collocation solution u_h to (2.1) in $S_m^{(0)}(I_h)$, where the underlying mesh I_h is defined by the mesh points

$$t_0 = 0, \;\; t_n = t_n^{(N)} = d^{N-n}T \quad (n = 1, \ldots, N) \tag{6.1}$$

for some $d = d(q; N) \in (0, 1)$.

THEOREM 6.1. *Let the assumptions of Theorem* 3.1 *hold with* $\kappa = m$ *(implying, by* (3.1)*, that the* $\{c_i\}$ *are the* m *Gauss points). If the mesh* I_h *corresponds to the points defined by* (6.1)*, with*

$$d = q^{1/r}, \quad r := \left\lfloor \frac{\ln(q)}{\ln[1 - \frac{2m\ln(N)}{(m+2)N}]} \right\rfloor, \tag{6.2}$$

then the estimate

$$\max\{|y(t) - u_h(t)| : \; t \in I_h\} \le CN^{-(2m-\varepsilon_N)} \quad \text{as} \;\; N \to \infty \tag{6.3}$$

for the collocation solution $u_h \in S_m^{(0)}(I_h)$ $(m \ge 2)$ *to* (2.1) *holds for any* $q \in (0, 1)$. *Here,* $C = C(q)$*, and* ε_N *is defined by*

$$\varepsilon_N := \log_N \left(\frac{(2m \cdot \ln N)^{2m}}{(2m+1)(m+2)^{2m}} \right) \tag{6.4}$$

and has the property $\varepsilon_N \to 0^+$ *as* $N \to \infty$.

Proof. For ease of exposition we describe the proof of (2.4); the analysis is readily extended to the general pantograph integro-differential equation (2.1) by employing the error representation based on the result (2.17) in Theorem 2.5. By (2.9) of Theorem 2.2 we obtain the error representation

$$e_h(t_n) = \int_0^{t_n} \delta_h(s)ds + \sum_{k=1}^{\infty} \int_0^{\theta^k(t_n)} H_k(t_n, s) \left(\int_0^s \delta_h(\tau)d\tau \right) ds, \quad n = 1, \ldots, N$$

(recall also (3.3)). Using integration by parts, this expression can be rewritten as

$$e_h(t_n) = \int_0^{t_n} \delta_h(s)ds + \sum_{k=1}^{\infty} \int_0^{\theta^k(t_n)} \tilde{H}_k(t_n, \tau)\delta_h(\tau)d\tau, \quad n = 1, \ldots, N, \tag{6.5}$$

with

$$\tilde{H}_k(t_n, \tau) := \int_\tau^{\theta^k(t_n)} H_k(t_n, s)ds.$$

Resorting to the quadrature argument used in the proof of Theorem 4.1, we find that

$$\left|\int_0^{t_n} \delta_h(s)ds\right| \le CN^{-2m} \quad (n=1,\dots,N). \tag{6.6}$$

Thus, the proof reduces to estimating the infinite series in (6.5). It is easy to see that

$$\theta^k(t_n) = q^k t_n = d^{kr}\cdot t_n = d^{N-n+kr}T,$$

with r as defined in (6.2). This, together with the definition (6.1) of t_n, leads to the following results:

(i) For $kr-n \ge -1$, we have $\theta^k(t_n) \le t_1 \le CN^{-\frac{2m}{m+2}}$ (as $N\to+\infty$).

(ii) For $kr-n \le -1$, there holds $\theta^k(t_n) = t_{n-kr} \in I_h$.

Note that (i) follows from Lemma 3.1(i) of [14].

We now assume, without loss of generality, that $n \ge r+1$. Otherwise, the case (i) always occurs for each k, and the corresponding analysis is obvious. We decompose the infinite series in (6.5) into two parts:

$$\begin{aligned}\sum_{k=1}^{\infty}\int_0^{\theta^k(t_n)} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau &= \sum_{k=[\frac{n-1}{r}]+1}^{\infty}\int_0^{\theta^k(t_n)} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau \\ &\quad + \sum_{k=1}^{[\frac{n-1}{r}]}\int_0^{t_{n-kr}} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau =: I_1 + I_2.\end{aligned} \tag{6.7}$$

Since

$$|\tilde{H}_k(t_n,\tau)| \le C\theta^k(t_n), \quad \tau\in[0,\ \theta^k(t_n)],$$

a standard argument leads to

$$\left|\int_0^{\theta^k(t_n)} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau\right| \le C(\theta^k(t_n))^{m+2} \le C(d^{N-n+kr})^{m+2}.$$

Furthermore, it is easy to verify that

$$|I_1| \le Cd^{(m+2)(N-1)} \le Ct_1^{m+2} \le CN^{-2m}. \tag{6.8}$$

Since the term I_2 in (6.7) can be written in the form

$$I_2 = \sum_{k=1}^{[\frac{n-1}{r}]}\sum_{i=1}^{n-kr}\int_{t_{i-1}}^{t_i} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau,$$

we obtain, observing Corollary 2.3,

$$\left|\int_{t_{i-1}}^{t_i} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau\right| \le C(t_i-t_{i-1})^{2m+1}.$$

It then follows from part (ii) of Lemma 3.1 in [14] that

$$|I_2| \le C(1-d)^{2m+1}\sum_{k=1}^{[\frac{n-1}{r}]}\sum_{i=1}^{n-kr} d^{(N-i)(2m+1)}$$

$$\leq C(1-d)^{2m} \sum_{k=1}^{[\frac{n-1}{r}]} d^{(2m+1)(N-n+kr)}$$
$$\leq C\frac{q^{2m+1}(1-d)^{2m}}{1-q^{2m+1}} \leq CN^{-(2m-\varepsilon_N)},$$

with ε_N given by (6.4). We now substitute (6.8) and the above inequality into (6.7); this yields

$$\left|\sum_{k=1}^{\infty} \int_0^{\theta^k(t_n)} \tilde{H}_k(t_n,\tau)\delta_h(\tau)d\tau\right| \leq CN^{-(2m-\varepsilon_N)}.$$

The result (6.3) in Theorem 6.1 now follows from (6.5), (6.6), and the above estimate. ∎

COROLLARY 6.2. *Let $a, b, g \in C^{2m}(I)$. If the solution y of the pantograph delay differential equation* (5.13) *is approximated by the collocation solution $u_h \in S_m^{(0)}(I_h)$ $(m \geq 2)$, with geometric mesh I_h given by* (6.1), (6.2)*, and with collocation points corresponding to the Gauss points $\{c_i\}$, then*

$$\max\{|y(t)-u_h(t)| : t \in I_h\} \leq CN^{-(2m-\varepsilon_N)} \quad \text{as } N \to \infty,$$

with ε_N defined in (6.4).

Remark. It was shown in Bellen [4] (see also [7] and [5]) that the optimal local superconvergence order $p^* = 2m$ can be restored if a different type of (quasi-) geometric mesh is employed (see also [31] and [6], where such meshes were introduced for the stability analysis of one-point collocation methods for the pantograph equation). This approach relies, however, on the assumption that a sufficiently accurate initial approximation is known on a "small" initial interval $[0, t_0]$; the quasi-geometric mesh is then generated on $[t_0, T]$. This contrasts the quasi-optimal local superconvergence results of Theorem 6.1 and Corollary 6.2 which do not require such an initial approximation.

7. Concluding remarks. We conclude our presentation by pointing to a number of open problems in the numerical analysis of pantograph-type functional differential equations.

(i) *Higher-order pantograph-type integro-differential equations.* Special cases of the initial-value problem

$$y''(t) = a(t)y(t) + b(t)y(\theta) + (\mathcal{V}y)(t) + (\mathcal{V}_\theta y)(t), \quad t \in I := [0,T], \tag{7.1}$$

with $\theta(t) = qt$ $(0 < q < 1)$, were studied, both theoretically and numerically, by Bélair [2] and by Zhang and Brunner [34]. If (7.1) is rewritten as a system of first-order integro-differential equations and solved by collocation in the piecewise polynomial space $S_m^{(0)}(I_h)$, then the superconvergence analysis of the previous sections can be readily extended to this system. However, if the initial-value problem (7.1) is solved directly, by using the collocation space $S_{m+1}^{(1)}(I_h)$, then the analysis of the optimal order of superconvergence on uniform meshes remains to be carried out.

(ii) *Pantograph equations of neutral type.* The convergence analysis in $S_m^{(0)}(I_h)$ for the neutral-type analogue of (2.4),

$$u'(t) = b(t)u(\theta(t)) + c(t)u'(\theta(t)) + g(t) + \int_0^{\theta(t)} (K_1(t,s)u(s) + K_2(t,s)u'(s))\,ds, \quad t \in [0,T], \tag{7.2}$$

with $\theta(t) = qt$ $(0 < q < 1)$, is much more complex than the one for (2.4) and is the subject of ongoing work. The key difficulty lies in the fact that (7.2) is in essence equivalent to a nonstandard pantograph-type Volterra integral equation of the form

$$u(t) = g_0(t) + \hat{b}(t)u(qt) + \int_0^{qt} \hat{K}(t,s)u(s)ds, \quad t \in I.$$

We note that even for the initial-value problem for the neutral pantograph equation

$$u'(t) = a(t)u(t) + b(t)u(qt) + c(t)u'(qt) + g(t), \quad t \in [0,T],$$

the existence of a unique (exact or collocation) solution is a nontrivial problem (it depends on the "size" of $c(t)$); compare, e.g., [27, 26, 16].

(iii) *Pantograph equations with multiple delays.* The paper [35] by Zhao, Xu, and Qiao contains an analysis of the optimal order of piecewise polynomial collocation solutions at $t = t_1 = h$ for the double pantograph equation (that is, for (1.1) with two proportional delays). Their result generalizes the analogous ones in [9, 33, 28, 32] for (1.1) with constant a and b. It is not yet known if the optimal superconvergence results of Corollaries 4.2 and 5.3 remain valid for delay differential equations with several proportional delays.

(iv) *Asymptotic stability of collocation solutions.* The problem of the asymptotic behavior (asymptotic stability; contractivity) of collocation solutions on *uniform meshes* to pantograph-type (integro-) differential equations remains essentially open. While there is such a result for the special case $u_h \in S_1^{(0)}(I_h)$ $(m = 1)$ and $q = 1/2$ (see Buhmann, Iserles, and Nørsett [16]; compare also Iserles [23, 24], Iserles and Liu [25]), there has been extensive work on asymptotic stability for (1.1) when I_h is a *geometric mesh*. We refer the reader to the papers by Liu [30, 31], Bellen, Guglielmi, and Torelli [6], Guglielmi and Zennaro [20], Huang and Vandewalle [22], and Guglielmi [19] and to the monograph [7] by Bellen and Zennaro.

Acknowledgments. Part of the work by the first author was carried out while he was a University Fellow at Hong Kong Baptist University (January–May 2006). He gratefully acknowledges the financial support and the hospitality extended to him by HKBU's Department of Mathematics.

The authors wish to thank Wei Gong (Academy of Mathematics and Systems Science, Chinese Academy of Sciences) and the anonymous referee for their careful reading of a previous version of this paper and for suggestions that led to a better presentation of the material.

REFERENCES

[1] G. Andreoli, *Sulle equazioni integrali*, Rend. Circ. Mat. Palermo, 37 (1914), pp. 76–112.
[2] J. Bélair, *Sur une équation différentielle fonctionnelle analytique*, Canad. Math. Bull., 24 (1981), pp. 43–46.
[3] A. Bellen, *One-step collocation for delay differential equations*, J. Comput. Appl. Math., 10 (1984), pp. 275–283.
[4] A. Bellen, *Preservation of superconvergence in the numerical integration of delay differential equations with proportional delay*, IMA J. Numer. Anal., 22 (2002), pp. 529–536.
[5] A. Bellen, H. Brunner, S. Maset, and L. Torelli, *Superconvergence in collocation methods on quasi-graded meshes for functional differential equations with vanishing delays*, BIT, 46 (2006), pp. 229–247.

[6] A. Bellen, N. Guglielmi, and L. Torelli, *Asymptotic stability properties of θ-methods for the pantograph equation*, Appl. Numer. Math., 24 (1997), pp. 275–293.
[7] A. Bellen and M. Zennaro, *Numerical Methods for Delay Differential Equations*, Oxford University Press, Oxford, UK, 2003.
[8] H. Brunner, *The numerical solution of neutral Volterra integro-differential equations with delay arguments,* Ann. Numer. Math., 1 (1994), pp. 309–322.
[9] H. Brunner, *On the discretization of differential and Volterra integral equations with variable delay*, BIT, 37 (1997), pp. 1–12.
[10] H. Brunner, *The discretization of Volterra functional integral equations with proportional delays*, in Difference and Differential Equations (Changsha, 2002), S. Elaydi, G. Ladas, J. Wu, and X. Zou, eds., Fields Inst. Commun. 42, AMS, Providence, RI, 2004, pp. 3–27.
[11] H. Brunner, *The numerical analysis of functional integral and integro-differential equations of Volterra type*, Acta Numer., 13 (2004), pp. 55–145.
[12] H. Brunner, *Collocation Methods for Volterra Integral and Related Functional Equations*, Cambridge University Press, Cambridge, UK, 2004.
[13] H. Brunner and Q.-Y. Hu, *Optimal superconvergence orders of iterated collocation solutions for Volterra integral equations with vanishing delays*, SIAM J. Numer. Anal., 43 (2005), pp. 1934–1949.
[14] H. Brunner, Q.-Y. Hu, and Q. Lin, *Geometric meshes in collocation methods for Volterra integral equations with proportional delays*, IMA J. Numer. Anal., 21 (2001), pp. 783–798.
[15] H. Brunner and S. Maset, *Time Transformations for Delay Differential Equations*, preprint, Department of Mathematics and Computer Science, University of Trieste, Trieste, Italy, 2006.
[16] M. Buhmann, A. Iserles, and S. P. Nørsett, *Runge–Kutta methods for neutral differential equations*, in Contributions in Numerical Mathematics (Singapore, 1993), R. P. Agarwal, ed., World Scientific, River Edge, NJ, 1993, pp. 85–98.
[17] Ll. G. Chambers, *Some properties of the functional equation* $\phi(x) = f(x) + \int_0^{\lambda x} g(x, y, \phi(y))dy$, Internat. J. Math. Math. Sci., 14 (1990), pp. 27–44.
[18] L. Fox, D. F. Mayers, J. R. Ockendon, and A. B. Tayler, *On a functional differential equation*, J. Inst. Math. Appl., 8 (1971), pp. 271–307.
[19] N. Guglielmi, *Short proofs and a counterexample for analytical and numerical stability of delay equations with infinite memory*, IMA J. Numer. Anal., 26 (2006), pp. 60–77.
[20] N. Guglielmi and M. Zennaro, *Stability of one-leg Θ-methods for the variable coefficient pantograph equation on the quasi-geometric mesh*, IMA J. Numer. Anal., 23 (2003), pp. 421–438.
[21] Q.-Y. Hu, *Geometric meshes and their application to Volterra integro-differential equations with singularities*, IMA J. Numer. Anal., 18 (1998), pp. 151–164.
[22] C. Huang and S. Vandewalle, *Discretized stability and error growth of the nonautonomous pantograph equation*, SIAM J. Numer. Anal., 42 (2005), pp. 2020–2042.
[23] A. Iserles, *On the generalized pantograph functional differential equation*, European J. Appl. Math., 4 (1993), pp. 1–38.
[24] A. Iserles, *Numerical analysis of delay differential equations with variable delays*, Ann. Numer. Math., 1 (1994), pp. 133–152.
[25] A. Iserles and Y. Liu, *On pantograph integro-differential equations*, J. Integral Equations Appl., 6 (1994), pp. 213–237.
[26] A. Iserles and Y. Liu, *On neutral functional-differential equations with proportional delays*, J. Math. Anal. Appl., 207 (1997), pp. 73–95.
[27] A. Iserles and J. Terjéki, *Stability and asymptotic stability of functional-differential equations*, J. London Math. Soc. Ser. II, 51 (1995), pp. 559–572.
[28] E. Ishiwata, *On the attainable order of collocation methods for the neutral functional-differential equations with proportional delays*, Computing, 64 (2000), pp. 207–222.
[29] T. Kato and J. B. McLeod, *The functional-differential equation* $y'(x) = ay(\lambda x) + by(x)$, Bull. Amer. Math. Soc., 77 (1971), pp. 891–937.
[30] Y. Liu, *Stability analysis of θ-methods for neutral functional-differential equations*, Numer. Math., 70 (1995), pp. 473–485.
[31] Y. Liu, *On θ-methods for delay differential equations with infinite lag*, J. Comput. Appl. Math., 71 (1996), pp. 177–190.
[32] Y. Muroya, E. Ishiwata, and H. Brunner, *On the attainable order of collocation methods for pantograph integro-differential equations*, J. Comput. Appl. Math., 152 (2003), pp. 347–366.

[33] N. Takama, Y. Muroya, and E. Ishiwata, *On the attainable order of collocation methods for the delay differential equations with proportional delay*, BIT, 40 (2000), pp. 374–394.
[34] W. Zhang and H. Brunner, *Collocation approximations for second-order differential equations and Volterra integro-differential equations with variable delays*, Canad. Appl. Math. Quart., 6 (1998), pp. 269–285.
[35] J. Zhao, Y. Xu, and Y. Qiao, *The attainable order of a collocation method for a double-pantograph delay differential equation*, Numer. Math. J. Chinese Univ., 27 (2005), pp. 297–308 (in Chinese).

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1005–1034

A STOCHASTIC COLLOCATION METHOD FOR ELLIPTIC PARTIAL DIFFERENTIAL EQUATIONS WITH RANDOM INPUT DATA*

IVO BABUŠKA†, FABIO NOBILE‡, AND RAÚL TEMPONE§

Abstract. In this paper we propose and analyze a stochastic collocation method to solve elliptic partial differential equations with random coefficients and forcing terms (input data of the model). The input data are assumed to depend on a finite number of random variables. The method consists in a Galerkin approximation in space and a collocation in the zeros of suitable tensor product orthogonal polynomials (Gauss points) in the probability space and naturally leads to the solution of uncoupled deterministic problems as in the Monte Carlo approach. It can be seen as a generalization of the stochastic Galerkin method proposed in [I. Babuška, R. Tempone, and G. E. Zouraris, *SIAM J. Numer. Anal.,* 42 (2004), pp. 800–825] and allows one to treat easily a wider range of situations, such as input data that depend nonlinearly on the random variables, diffusivity coefficients with unbounded second moments, and random variables that are correlated or even unbounded. We provide a rigorous convergence analysis and demonstrate exponential convergence of the "probability error" with respect to the number of Gauss points in each direction in the probability space, under some regularity assumptions on the random input data. Numerical examples show the effectiveness of the method.

Key words. collocation method, stochastic partial differential equations, finite elements, uncertainty quantification, exponential convergence

AMS subject classifications. 65N35, 65N15, 65C20, 65N12, 65N30

DOI. 10.1137/050645142

Introduction. Thanks to the fast growing power of computers, numerical simulations are increasingly used to produce predictions of the behavior of complex physical or engineering systems. Some sources of errors arising in computer simulations now can be controlled and reduced by using sophisticated techniques such as a posteriori error estimation [1, 3, 37], mesh adaptivity, and the more recent *modeling error* analysis [31, 32, 10]. All this has increased the accuracy of numerical predictions as well as our confidence in them.

Yet, many engineering applications are affected by a relatively large amount of uncertainty in the input data such as model coefficients, forcing terms, boundary conditions, and geometry. In this case, to obtain a reliable numerical prediction, one has to include uncertainty quantification due to the uncertainty in the input data.

*Received by the editors November 14, 2005; accepted for publication (in revised form) October 16, 2006; published electronically May 7, 2007.
http://www.siam.org/journals/sinum/45-3/64514.html

†ICES, The University of Texas at Austin, Austin, TX 78712 (babuska@ices.utexas.edu). This author's work was partially supported by Sandia National Lab contract 268687 and Office of Naval Research grant N00014-99-1-0724.

‡MOX, Dipartimento di Matematica, Politecnico di Milano, 20133 Milano, Italy (fabio.nobile@polimi.it). This author's work was partially supported by M.U.R.S.T. Cofin 2004 grant "Metodi Numerici Avanzati per Equazioni alle Derivate Parziali di Interesse Applicativo" and a J. T. Oden Visiting Faculty Fellowship.

§School of Computational Sciences and Department of Mathematics, Florida State University at Tallahassee, Tallahassee, FL 32306 (rtempone@scs.fsu.edu). This author's work was partially supported by the ICES Postdoctoral Fellowship Program, UdelaR in Uruguay, and the European Network HYKE (HYperbolic and Kinetic Equations: Asymptotics, Numerics, Analysis), funded by EC contract HPRN-CT-2002-00282.

Uncertainty can be described in several ways, depending on the amount of information available; among others we mention the worst case scenario analysis and fuzzy set theory, evidence theory, probabilistic setting, etc. (see [6, 24] and the references therein). In this paper we focus on elliptic partial differential equations with a probabilistic description of the uncertainty in the input data. The model problem has the form

$$\mathcal{L}(a)u = f \quad \text{in } D, \tag{0.1}$$

where $\mathcal{L}$ is an elliptic operator in a domain $D \subset \mathbb{R}^d$, which depends on some coefficients $a(x, \omega)$, with $x \in D$, $\omega \in \Omega$, and Ω indicating the set of possible outcomes. Similarly, the forcing term $f = f(x, \omega)$ can be assumed to be random as well.

We will focus on the case where the probability space has a low dimensionality, which means that the stochastic problem depends only on a relatively small number of random variables.

This can be the case if, for instance, the mathematical model depends on few parameters, which can be taken as random variables with a given joint probability distribution. For example, we might think of the deformation of an elastic homogeneous material in which Young's modulus and Poisson's ratio (parameters that characterize the material properties) are random variables, either independent or not.

In other situations, the input data may vary randomly from one point of the physical domain D to another, and their uncertainty should rather be described in terms of random fields with a given covariance structure (i.e., each point of the domain is a random variable, and the correlation between two distinct points in the domain is known and nonzero, in general; this case is sometimes referred to as *colored noise*).

Examples of this situation are the deformation of inhomogeneous materials such as wood, foams, or biomaterials such as arteries and bones; groundwater flow problems where the permeability in each layer of sediments (rocks, sand, etc.) should not be assumed constant; and the action of wind (direction and point intensity) on structures.

A possible way to describe such random fields consists in using a Karhunen–Loève [27, 28] or a polynomial chaos (PC) expansion [38, 42]. The former represents the random field as a linear combination of an infinite number of uncorrelated random variables, while the latter uses polynomial expansions in terms of independent random variables. Both expansions exist provided that the random field $a : \Omega \to V$, as a mapping from the probability space into a functional space V, has bounded second moments. Other nonlinear expansions can be considered as well (see, e.g., [22] for a technique to express a stationary random field with given covariance structure and marginal distribution as a function of (infinite) independent random variables; nonlinear transformations also have been used in [30, 39]). The use of nonpolynomial expansions may be advantageous in some situations: for instance, in groundwater flow problems, the permeability coefficient within each layer of sediments can feature huge variability, which is often expressed in a logarithm scale. In this case, one might want to use a Karhunen–Loève (or PC) expansion for the logarithm of the permeability, instead of the permeability field itself. This leads to an exponential dependence of the permeability on the random variables, and the resulting random field might even have unbounded second moments. An advantage of such a nonlinear expansion is that it guarantees a positive permeability almost surely (a condition which is difficult to enforce with a standard truncated Karhunen–Loève or PC expansion).

Although such random fields are properly described only by means of an infinite number of random variables, whenever the input data vary slowly in space, with a correlation length comparable to the size of the domain, only a few terms in the above

mentioned expansions are typically enough to describe the random field with sufficient accuracy. Therefore, for this type of application it is reasonable to limit the analysis to just a few random variables in the expansion (see, e.g., [2]).

This argument is also strengthened by the fact that the amount of measured data one has at hand to identify the input data as random fields is in general very limited and barely sufficient to identify the first few random variables in the expansion.

Conversely, situations in which the random fields are highly oscillatory with a short correlation length, as in the case of materials with a random microstructure, do not fall into this category and will not be considered in the present work. The interested reader should refer, instead, to the wide literature in homogenization and multiscale analysis (see, e.g., [16] and references therein).

To solve numerically the stochastic PDE (0.1), a relatively new numerical technique, which has gained much attention in the last few years, is the so-called *spectral Galerkin* approximation (see, e.g., [21]). It employs standard approximations in space (finite elements, finite volumes, spectral or h-p finite elements, etc.) and polynomial approximation in the probability domain, either by full polynomial spaces [41, 30, 20, 34], tensor product polynomial spaces [4, 18], or piecewise polynomial spaces [4, 26].

The use of tensor product spaces is particularly attractive in the case of a small number of random variables, since it allows naturally the use of anisotropic spaces where the polynomial degree is chosen differently with respect to each random variable. Moreover, whenever the random fields are expanded in a truncated Karhunen–Loève expansion and the underlying random variables are assumed independent, a particular choice of the basis for the tensor product space (as proposed in [4, 5]), leads to the solution of uncoupled deterministic problems as in a Monte Carlo simulation. In this case, exponential convergence of the "probability error" has been proved in [4].

On the other hand, tensor product spaces suffer from the so-called *curse of dimensionality* since the dimension of the approximating space grows exponentially fast in the number of random variables. If the number of random variables is moderate or large, one should rather consider full polynomial spaces or sparse tensor product spaces [7, 18, 40]. We will not address this issue in this paper.

The extension of the spectral Galerkin method to cases in which the input data depend nonlinearly on the random variables and possibly have unbounded second moments is not straightforward and, in any case, would lead to fully coupled systems of equations, which call for highly efficient parallel solvers.

In this work we propose a collocation method which consists in collocating problem (0.1) in the zeros of tensor product orthogonal polynomials with respect to the joint probability density ρ of the random variables, should they be independent, or any other auxiliary density $\hat{\rho}$ corresponding to independent random variables, as long as the ratio $\rho/\hat{\rho}$ is bounded. Stochastic collocation has already been applied in a variety of problems and is the subject of ongoing research; see among others [35, 29] and the recent work [40], which the authors became aware of upon completion of this work.

As will be pointed out in the paper, this method offers several advantages:

- It naturally leads to uncoupled deterministic problems also in the case of input data which depend nonlinearly on the random variables.
- It treats efficiently the case of nonindependent random variables with the introduction of the auxiliary density $\hat{\rho}$.
- It can easily deal with unbounded random variables, such as Gaussian or exponential ones.

- It deals without difficulty with a diffusivity coefficient a with unbounded second moment.

The main result of the paper is given in Theorem 4.1 in section 4, where we prove that the collocation method preserves the same accuracy as the spectral Galerkin approach and achieves exponential convergence in all the above mentioned cases, provided that the input data are infinitely differentiable with respect to the random variables, under very mild assumptions on the growth of such derivatives, as is the case for standard expansions of random fields.

The collocation method can also be seen as a pseudospectral method (see, e.g., [33] and [19] for unbounded domains), i.e., a spectral Galerkin approximation with the use of suitable Gaussian quadrature formulas. We will also show that in some particular cases, where such Gaussian quadratures are exact, it actually coincides with the spectral Galerkin method based on tensor product spaces.

The outline of the paper is as follows: in section 1 we introduce the mathematical problem and the main notation used throughout. In section 2 we describe the collocation method. In section 3 we provide some regularity results on the solution of the stochastic PDE. In particular, we show that the solution is analytic with respect to the random variables, provided that the input data, as functions of the random variables, have infinite derivatives which do not grow too fast. In section 4 we give a complete convergence result for the collocation method and prove exponential convergence. Finally, in section 5 we present some numerical results showing the effectiveness of the proposed method.

1. Problem setting. Let D be a convex bounded polygonal domain in $\mathbb{R}^d$ and let $(\Omega, \mathcal{F}, P)$ be a complete probability space. Here Ω is the set of outcomes, $\mathcal{F} \subset 2^\Omega$ is the σ-algebra of events, and $P : \mathcal{F} \to [0,1]$ is a probability measure. Consider the stochastic linear elliptic boundary value problem: find a random function, $u : \Omega \times \overline{D} \to \mathbb{R}$, such that P-almost everywhere (a.e.) in Ω, or in other words, almost surely (a.s.) the following equation holds:

$$\begin{aligned} -\nabla \cdot (a(\omega,\cdot)\nabla u(\omega,\cdot)) &= f(\omega,\cdot) \quad \text{on} \quad D, \\ u(\omega,\cdot) &= 0 \quad \text{on} \quad \partial D. \end{aligned} \tag{1.1}$$

We will make the following assumptions:

(A_1) $a(\omega,\cdot)$ is uniformly bounded from below; i.e.,

$$\text{there exist } a_{min} > 0 \text{ such that} \quad P\big(\, \omega \in \Omega : \; a(\omega,x) > a_{min} \; \forall x \in \overline{D} \,\big) = 1.$$

(A_2) $f(\omega,\cdot)$ is square integrable with respect to P; i.e., $\int_D E[f^2]\,dx < \infty$.

Moreover, we introduce the following Hilbert spaces:

- $V_P = L^2_P(\Omega) \otimes H^1_0(D)$, equipped with the norm $\|v\|_P^2 = \int_D E\big[|\nabla v|^2\big]\,dx$.
- $V_{P,a} \equiv \{v \in V_P : \; \int_D E\big[a|\nabla v|^2\big]\,dx < \infty\}$, equipped with the norm $\|v\|_{P,a} = \sqrt{\int_D E\big[a|\nabla v|^2\big]\,dx}$.

Observe that under the above assumptions, the space $V_{P,a}$ is continuously embedded in V_P and

$$\|v\|_P \le \frac{1}{\sqrt{a_{min}}} \|v\|_{P,a}.$$

Problem (1.1) can be written in a weak form as

$$\text{find } u \in V_{P,a} \text{ such that} \quad \int_D E[a\nabla u \cdot \nabla v]\,dx = \int_D E[fv]\,dx \quad \forall\, v \in V_{P,a}. \tag{1.2}$$

A straightforward application of the Lax–Milgram theorem allows one to state the well posedness of problem (1.2) in the following lemma.

LEMMA 1.1. *Under assumptions* (A_1) *and* (A_2), *problem* (1.2) *admits a unique solution* $u \in V_{P,a}$, *which satisfies the estimate*

$$\|u\|_P \le \frac{C_P}{a_{min}} \left(\int_D E[f^2]\, dx \right)^{\frac{1}{2}}. \tag{1.3}$$

In the previous lemma we have used the Poincaré inequality

$$\|w\|_{L^2(D)} \le C_P \|\nabla w\|_{L^2(D)} \qquad \forall w \in H_0^1(D).$$

Weaker assumptions on the random coefficients. It is possible to relax the assumptions (A_1) and (A_2) substantially and still guarantee the existence and uniqueness of the solution u to problem (1.2). In particular, if the lower bound for the coefficient a is no longer a constant but a random variable, i.e., $a(x,\omega) \ge a_{min}(\omega) > 0$ a.s. a.e. on D, we have the following estimate for the moments of the solution.

LEMMA 1.2 (moments estimates). *Let* $p, q \ge 0$ *with* $1/p + 1/q = 1$. *Take a positive integer* k. *Then if* $f \in L_P^{kp}(\Omega; L^2(D))$ *and* $1/a_{min} \in L_P^{kq}(\Omega)$, *we have that* $u \in L_P^k(\Omega; H_0^1(D))$.

Proof. Since

$$\|u(\cdot,\omega)\|_{H_0^1(D)} \le C_P \frac{\|f(\cdot,\omega)\|_{L^2(D)}}{a_{min}(\omega)} \quad a.s.,$$

the result is a direct application of Hölder's inequality:

$$\begin{aligned}\int_\Omega \|u(\cdot,\omega)\|_{H_0^1(D)}^k dP(\omega) &\le C_P^k \int_\Omega \left(\frac{\|f(\cdot,\omega)\|_{L^2(D)}}{a_{min}(\omega)} \right)^k dP(\omega) \\ &\le C_P^k \left(\int_\Omega \|f(\cdot,\omega)\|_{L^2(D)}^{kp} dP(\omega) \right)^{1/p} \left(\int_\Omega \left(\frac{1}{a_{min}(\omega)} \right)^{qk} dP(\omega) \right)^{1/q}. \qquad \square\end{aligned}$$

Example 1 (lognormal diffusion coefficient). As an application of the previous lemma, we can conclude the well posedness of (1.2) with a lognormal diffusion coefficient. For instance, let

$$a(x,\omega) = \exp\left(\sum_{n=1}^N b_n(x) Y_n(\omega) \right), \quad Y_n \sim N(0,1) \quad \text{independent and identically distributed.}$$

Use the lower bound

$$a_{min}(\omega) = \exp\left(-\sum_{n=1}^N \|b_n\|_{L^\infty(D)} |Y_n(\omega)| \right)$$

and then for $k, q < \infty$

$$\begin{aligned}\|1/a_{min}\|_{L_P^{kq}(\Omega)}^{kq} &= \int_\Omega \left(\frac{1}{a_{min}(\omega)} \right)^{qk} dP(\omega) \\ &= \int_{\mathbb{R}^N} \exp\left(kq \sum_{n=1}^N \|b_n\|_{L^\infty(D)} |z_n| \right) \exp\left(-\frac{1}{2} \sum_{n=1}^N z_n^2 \right) dz_1 \cdots dz_N < \infty.\end{aligned} \tag{1.4}$$

Now let $\epsilon > 0$. Then by Lemma 1.2 the assumption $f \in L_P^{k(1+\epsilon)}(\Omega; L^2(D))$ together with (1.4) implies $u \in L_P^k(\Omega; H_0^1(D))$.

1.1. Finite dimensional noise assumption. In many problems the source of randomness can be approximated using just a small number of uncorrelated, sometimes independent, random variables; take, for example, the case of a truncated Karhunen–Loève expansion [4]. This motivates us to make the following assumption.

ASSUMPTION 1 (finite dimensional noise). *The coefficients used in the computations have the form*

$$a(\omega,x)=a(Y_1(\omega),\dots,Y_N(\omega),x) \quad \text{and} \quad f(\omega,x)=f(Y_1(\omega),\dots,Y_N(\omega),x) \quad \text{on} \quad \Omega\times\overline{D},$$

where $N\in\mathbb{N}_+$ and $\{Y_n\}_{n=1}^N$ are real valued random variables with mean value zero and unit variance.

We will denote with $\Gamma_n\equiv Y_n(\Omega)$ the image of Y_n, $\Gamma=\prod_{n=1}^N\Gamma_n$ and we will assume that the random variables $[Y_1,Y_2,\dots,Y_N]$ have a joint probability *density function* $\rho:\Gamma\to\mathbb{R}_+$, with $\rho\in L^\infty(\Gamma)$.

Example 2. The following standard transformation guarantees that the diffusivity coefficient is bounded away from zero a.s.:

$$\log(a-a_{min})(\omega,x)=b_0(x)+\sum_{1\le n\le N}\sqrt{\lambda_n}b_n(x)Y_n(\omega); \tag{1.5}$$

i.e., one performs a Karhunen–Loève expansion for $\log(a-a_{min})$, assuming that $a>a_{min}$ a.s. On the other hand, the right-hand side of (1.1) can be represented as a truncated Karhunen–Loève expansion:

$$f(\omega,x)=c_0(x)+\sum_{1\le n\le N}\sqrt{\mu_n}c_n(x)Y_n(\omega).$$

Remark 1. It is usual to have f and a independent, because the loads and the material properties are seldom related. In such a situation we have $a(Y(\omega),x)=a(Y_a(\omega),x)$ and $f(Y(\omega),x)=f(Y_f(\omega),x)$, with $Y=[Y_a,Y_f]$ and the vectors Y_a, Y_f independent.

After making Assumption 1, the solution u of the stochastic elliptic boundary value problem (1.2) can be described by just a finite number of random variables, i.e., $u(\omega,x)=u(Y_1(\omega),\dots,Y_N(\omega),x)$. Thus, the goal is to approximate the function $u=u(y,x)$, where $y\in\Gamma$ and $x\in\overline{D}$. Observe that the stochastic variational formulation (1.2) has a "deterministic" equivalent which is the following: find $u\in V_{\rho,a}$ such that

$$\int_\Gamma \rho\,(a\nabla u,\nabla v)_{L^2(D)}\,dy=\int_\Gamma\rho\,(f,v)_{L^2(D)}\,dy\quad\forall\,v\in V_{\rho,a}, \tag{1.6}$$

noting that here and later in this work the gradient notation, ∇, always means differentiation with respect to $x\in D$ only, unless otherwise stated. The space $V_{\rho,a}$ is the analogue of $V_{P,a}$ with $(\Omega,\mathcal{F},P)$ replaced by $(\Gamma,\mathcal{B}^N,\rho\,dy)$.

Since the solution to (1.6) is unique and is also a solution to (1.2), it follows that the solution has necessarily the form $u(\omega,x)=u(Y_1(\omega),\dots,Y_N(\omega),x)$. The stochastic boundary value problem (1.1) now becomes a deterministic Dirichlet boundary value problem for an elliptic PDE with an N-dimensional parameter. For convenience, we consider the solution u as a function $u:\Gamma\to H_0^1(D)$ and we use the notation $u(y)$ whenever we want to highlight the dependence on the parameter y. We use similar notation for the coefficient a and the forcing term f. Then it can be shown that problem (1.1) is equivalent to

$$\int_D a(y)\nabla u(y)\cdot\nabla\phi\,dx=\int_D f(y)\phi\,dx\quad\forall\phi\in H_0^1(D),\ \ \rho\text{-a.e. in }\Gamma. \tag{1.7}$$

For our convenience, we will suppose that the coefficient a and the forcing term f admit a smooth extension on the $\rho\,dy$-zero measure sets. Then (1.7) can be extended a.e. in Γ with respect to the Lebesgue measure (instead of the measure $\rho\,dy$).

Remark 2. Strictly speaking, (1.7) will hold only for those values of $y \in \Gamma$ for which the coefficient $a(y)$ is finite. In this paper we will assume that $a(y)$ may go to infinity only at the boundary of the parameter domain Γ.

Making Assumption 1 is a crucial step, turning the original stochastic elliptic equation into a deterministic parametric elliptic equation and allowing the use of finite element and finite difference techniques to approximate the solution of the resulting deterministic problem (cf. [25, 13]).

Observe that the knowledge of $u = u(y, x)$ fully determines the law of the random field $u(\omega, x)$. Yet, the computation of some quantities of interest such as failure probabilities might pose extra challenges from the numerical point of view. On the other hand, computation of moments of the solution or functionals of the solution is direct (see sections 2 and 4.1).

2. Collocation method. We seek a numerical approximation to the exact solution of (1.6) in a finite dimensional subspace $V_{p,h}$ based on a tensor product, $V_{p,h} = \mathcal{P}_p(\Gamma) \otimes H_h(D)$, where the following hold.

- $H_h(D) \subset H_0^1(D)$ is a standard finite element space of dimension N_h, which contains continuous piecewise polynomials defined on regular triangulations $\mathcal{T}_h$ that have a maximum mesh spacing parameter $h > 0$.
- $\mathcal{P}_p(\Gamma) \subset L^2_\rho(\Gamma)$ is the span of tensor product polynomials with degree at most $p = (p_1, \dots, p_N)$; i.e., $\mathcal{P}_p(\Gamma) = \bigotimes_{n=1}^N \mathcal{P}_{p_n}(\Gamma_n)$, with

$$\mathcal{P}_{p_n}(\Gamma_n) = \operatorname{span}(y_n^m,\ m = 0, \dots, p_n), \quad n = 1, \dots, N.$$

  Hence the dimension of $\mathcal{P}_p$ is $N_p = \prod_{n=1}^N (p_n + 1)$.

We first introduce the semidiscrete approximation $u_h : \Gamma \to H_h(D)$, obtained by projecting (1.7) onto the subspace $H_h(D)$, for each $y \in \Gamma$, i.e.,

$$\int_D a(y)\nabla u_h(y) \cdot \nabla \phi_h \, dx = \int_D f(y)\phi_h \, dx \quad \forall \phi_h \in H_h(D), \ \text{ for a.e. } y \in \Gamma. \tag{2.1}$$

The next step consists in collocating (2.1) on the zeros of orthogonal polynomials and building the discrete solution $u_{h,p} \in \mathcal{P}_p(\Gamma) \otimes H_h(D)$ by interpolating in y the collocated solutions.

To this end, we first introduce an auxiliary probability density function $\hat\rho : \Gamma \to \mathbb{R}^+$ that can be seen as the joint probability of N independent random variables; i.e., it factorizes as

$$\hat\rho(y) = \prod_{n=1}^N \hat\rho_n(y_n) \ \ \forall y \in \Gamma, \qquad \text{and is such that} \quad \left\| \frac{\rho}{\hat\rho} \right\|_{L^\infty(\Gamma)} < \infty. \tag{2.2}$$

For each dimension $n = 1, \dots, N$, let y_{n,k_n}, $1 \le k_n \le p_n + 1$, be the $p_n + 1$ roots of the orthogonal polynomial q_{p_n+1} with respect to the weight $\hat\rho_n$, which then satisfies $\int_{\Gamma_n} q_{p_n+1}(y) v(y) \hat\rho_n(y) dy = 0$ for all $v \in \mathcal{P}_{p_n}(\Gamma_n)$.

Standard choices for $\hat\rho$, such as constant, Gaussian, etc., lead to well-known roots of the polynomial q_{p_n+1}, which are tabulated to full accuracy and do not need to be computed.

To any vector of indexes $[k_1, \dots, k_N]$ we associate the global index

$$k = k_1 + p_1(k_2 - 1) + p_1 p_2 (k_3 - 1) + \cdots$$

and we denote by y_k the point $y_k = [y_{1,k_1}, y_{2,k_2}, \ldots, y_{N,k_N}] \in \Gamma$. We also introduce, for each $n = 1, 2, \ldots, N$, the Lagrange basis $\{l_{n,j}\}_{j=1}^{p_n+1}$ of the space $\mathcal{P}_{p_n}$,

$$l_{n,j} \in \mathcal{P}_{p_n}(\Gamma_n), \qquad l_{n,j}(y_{n,k}) = \delta_{jk}, \quad j,k = 1, \ldots, p_n + 1,$$

where δ_{jk} is the Kronecker symbol, and we set $l_k(y) = \prod_{n=1}^{N} l_{n,k_n}(y_n)$. Hence, the final approximation is given by

$$u_{h,p}(y,x) = \sum_{k=1}^{N_p} u_h(y_k, x) l_k(y),$$

where $u_h(y_k, x)$ is the solution of problem (2.1) for $y = y_k$.

Equivalently, if we introduce the Lagrange interpolant operator $\mathcal{I}_p : C^0(\Gamma; H_0^1(D)) \to \mathcal{P}_p(\Gamma) \otimes H_0^1(D)$, such that

$$\mathcal{I}_p v(y) = \sum_{n=1}^{N} v(y_k) l_k(y) \qquad \forall v \in C^0(\Gamma; H_0^1(D)),$$

then we have simply $u_{h,p} = \mathcal{I}_p u_h$.

Finally, for any continuous function $g : \Gamma \to \mathbb{R}$ we introduce the Gauss quadrature formula $E_{\hat\rho}^p[g]$ approximating the integral $\int_\Gamma g(y)\hat\rho(y)\,dy$ as

$$E_{\hat\rho}^p[g] = \sum_{k=1}^{N_p} \omega_k g(y_k), \quad \omega_k = \prod_{n=1}^{N} \omega_{k_n}, \quad \omega_{k_n} = \int_{\Gamma_n} l_{k_n}^2(y)\hat\rho_n(y)\,dy. \tag{2.3}$$

This can be used to approximate the mean value or the variance of u as

$$\bar u_h \in H_h(D), \qquad \bar u_h(x) = E_{\hat\rho}^p\left[\frac{\rho}{\hat\rho} u_h(x)\right],$$
$$\mathrm{var}_h(u_h) \in L^1(D), \qquad \mathrm{var}_h(u_h)(x) = E_{\hat\rho}^p\left[\frac{\rho}{\hat\rho}\left(u_h(x) - \bar u_h(x)\right)^2\right]$$

as long as $\rho/\hat\rho$ is a smooth function. Otherwise, $\bar u_h$ and $\mathrm{var}_h(u_h)$ should be computed with a suitable quadrature formula which takes into account eventual discontinuities or singularities of $\rho/\hat\rho$.

2.1. Collocation versus spectral Galerkin approximation. An alternative approach to the collocation method introduced thus far consists in approximating problem (1.6) with a spectral Galerkin method: find $u_{h,p}^G \in \mathcal{P}_p(\Gamma) \otimes H_h(D)$ such that

$$\int_\Gamma \rho\,(a\nabla u_{h,p}^G, \nabla v)_{L^2(D)}\,dy = \int_\Gamma \rho\,(f, v)_{L^2(D)}\,dy \quad \forall\, v \in \mathcal{P}_p(\Gamma) \otimes H_h(D). \tag{2.4}$$

This approach has been considered by several authors [4, 13, 18, 41, 21, 30]. Observe that, in general, problem (2.4) leads to a fully coupled system of linear equations, whose dimension is $N_h \times N_p$ and that demands highly efficient strategies and parallel computations for its numerical solution [15]. Conversely, the collocation method requires only the solutions of N_p uncoupled linear systems of dimension N_h and is fully parallelizable.

In [4, 5] a particular choice of basis functions (named *double orthogonal polynomials*) for the space $P_p(\Gamma)$ is proposed. This choice allows us to decouple the system in

the special case where the diffusivity coefficient and the forcing term are multilinear combinations of the random variables $Y_n(\omega)$ (as is the case if one performs a truncated linear Karhunen–Loève expansion) and the random variables are independent, i.e., $\rho(y) = \prod_{n=1}^{N} \rho_n(y_n)$. The proposed basis is then obtained by solving the following eigenvalue problems, for each $n = 1, \dots, N$:

$$\int_{\Gamma_n} z\psi_{kn}(z)v(z)\rho_n(z)\,dz = c_{kn}\int_{\Gamma_n} \psi_{kn}(z)v(z)\rho_n(z)\,dz, \qquad k = 1, \dots, p_n + 1.$$

The eigenvectors ψ_{kn} are normalized so as to satisfy the property

$$\int_{\Gamma_n} \psi_{kn}(z)\psi_{jn}(z)\rho_n(z)\,dz = \delta_{kj}, \qquad \int_{\Gamma_n} z\psi_{kn}(z)\psi_{jn}(z)\rho_n(z)\,dz = c_{kn}\delta_{kj}.$$

See [4, 5] for further details on the double orthogonal basis.

We aim at analyzing, now, the analogies between the collocation and the spectral Galerkin methods. The collocation method can be seen as a *pseudospectral* Galerkin method (see, e.g., [33]), where the integrals over Γ in (2.4) are replaced by the quadrature formula (2.3): *find* $u_{h,p} \in \mathcal{P}_p(\Gamma) \otimes H_h(D)$ *such that*

$$E_{\hat{\rho}}^p\left[\frac{\rho}{\hat{\rho}}\,(a\nabla u_{h,p}, \nabla v)_{L^2(D)}\right] = E_{\hat{\rho}}^p\left[\frac{\rho}{\hat{\rho}}\,(f, v)_{L^2(D)}\right] \quad \forall\, v \in \mathcal{P}_p(\Gamma) \otimes H_h(D). \tag{2.5}$$

Indeed, by choosing in (2.5) the test functions of the form $v(y, x) = l_k(y)\phi(x)$, where $\phi(x) \in H_h(D)$ and $l_k(y)$ is the Lagrange basis function associated to the knot y_k, $k = 1, \dots, N_p$, one is led to solve a sequence of uncoupled problems of the form (2.1) collocated in the points y_k, which, ultimately, gives the same solution as the collocation method.

In the particular case where the diffusivity coefficient and the forcing term are multilinear combinations of the random variables $Y_n(\omega)$ and the random variables are independent, it turns out that the quadrature formula is exact if one chooses $\hat{\rho} = \rho$. In this case, the solution obtained by the collocation method actually coincides with the spectral Galerkin one. This can be seen easily by observing that, with the above assumptions, the integrand in (2.4), i.e., $(a\nabla u_{h,p} \cdot \nabla v)$, is a polynomial at most of degree $2p_n + 1$ in the variable y_n and the Gauss quadrature formula is exact for polynomials up to degree $2p_n + 1$ integrated against the weight ρ.

The collocation method is a natural generalization of the spectral Galerkin approach and has the following advantages:

- It decouples the system of linear equations in Y also in the case where the diffusivity coefficient a and the forcing term f are nonlinear functions of the random variables Y_n.
- It treats efficiently the case of nonindependent random variables with the introduction of the auxiliary measure $\hat{\rho}$.
- It can easily deal with unbounded random variables (see Theorem 4.1 in section 4).

As will be shown in section 4, the collocation method preserves the same accuracy as the spectral Galerkin approach and achieves exponential convergence if the coefficient a and forcing term f are infinitely differentiable with respect to the random variables Y_n, under very mild requirements on the growth of their derivatives in Y.

As a final remark, we show that the double orthogonal polynomials proposed in [4] coincide with the Lagrange basis $l_k(y)$ and the eigenvalues c_{kn} are nothing but the Gauss knots of integration.

LEMMA 2.1. *Let $\Gamma \subset \mathbb{R}$, $\rho : \Gamma \to \mathbb{R}$ be a positive weight, and $\{\psi_k\}_{k=1}^{p+1}$ be the set of double orthogonal polynomials of degree p satisfying*

$$\int_\Gamma \psi_k(y)\psi_j(y)\rho(y)\,dy = \delta_{kj}, \qquad \int_\Gamma y\psi_k(y)\psi_j(y)\rho(y)\,dy = c_k\delta_{kj}.$$

Then the eigenvalues c_k are the nodes of the Gaussian quadrature formula associated to the weight ρ, and the eigenfunctions ψ_k are, up to multiplicative factors, the corresponding Lagrange polynomials built on the nodes c_k.

Proof. We have, for $k = 1, \dots, p+1$,

$$\int_\Gamma (y - c_k)\psi_k(y)v(y)\rho(y)dy = 0 \quad \forall v \in P_p(\Gamma).$$

Take $v = \prod_{\substack{j=1\\ j\neq k}}^{p+1}(y - c_j) \in P_p(\Gamma)$ in the above and let $w = \prod_{j=1}^{p+1}(y - c_j)$. Then

$$\int_\Gamma w(y)\psi_k(y)\rho(y)dy = 0 \quad \forall k = 1, \dots, p+1.$$

Since $\{\psi_k\}_{k=1}^{p+1}$ defines a basis of the space $\mathcal{P}_p(\Gamma)$, the previous relation implies that w is ρ-orthogonal to $P_p(\Gamma)$. Besides, the functions $(y - c_k)\psi_k$ are also orthogonal to the same subspace: this yields, due to the one-dimensional nature of the orthogonal complement of $P_p(\Gamma)$ over $P_{p+1}(\Gamma)$,

$$(y - c_k)\psi_k = \alpha_k w = \alpha_k \prod_{j=1}^{p+1}(y - c_j), \quad k = 1, \dots, p+1,$$

which gives

$$\psi_k = \alpha_k \prod_{\substack{j=1\\ j\neq k}}^{p+1}(y - c_j), \quad k = 1, \dots, p+1;$$

i.e., the double orthogonal polynomials ψ_k are collinear to Lagrange interpolants at the nodes c_j. Moreover, the eigenvalues c_j are the roots of the polynomial $w \in P_{p+1}(\Gamma)$, which is ρ-orthogonal to $P_p(\Gamma)$, and therefore they coincide with the nodes of the Gaussian quadrature formula associated with the weight ρ. ∎

3. Regularity results. Before going through the convergence analysis of the method, we need to state some regularity assumptions on the data of the problem and consequent regularity results for the exact solution u and the semidiscrete solution u_h.

In what follows we will need some restrictive assumptions on f and ρ. In particular, we will assume f to be a continuous function in y, whose growth at infinity, whenever the domain Γ is unbounded, is at most exponential. At the same time we will assume that ρ behaves as a Gaussian weight at infinity, as does the auxiliary density $\hat{\rho}$, in light of assumption (2.2).

Other types of growth of f at infinity and corresponding decay of the probability density ρ, for instance, of exponential type, could be considered as well. Yet we will limit the analysis to the aforementioned case.

To make these assumptions precise, we introduce a weight $\sigma(y) = \prod_{n=1}^{N} \sigma_n(y_n) \leq 1$, where

$$\sigma_n(y_n) = \begin{cases} 1 & \text{if } \Gamma_n \text{ is bounded,} \\ e^{-\alpha_n |y_n|} \text{ for some } \alpha_n > 0 & \text{if } \Gamma_n \text{ is unbounded,} \end{cases} \tag{3.1}$$

and the functional space

$$C^0_\sigma(\Gamma; V) \equiv \left\{ v : \Gamma \to V, \;\; v \text{ continuous in } y, \;\; \max_{y \in \Gamma} \|\sigma(y) v(y)\|_V < +\infty \right\},$$

where V is a Banach space of functions defined in D.

ASSUMPTION 2 (growth at infinity). *In what follows we will assume that*

(a) $f \in C^0_\sigma(\Gamma; L^2(D))$, *and*

(b) *the joint probability density* ρ *satisfies*

$$\rho(y) \leq C_\rho \, e^{-\sum_{n=1}^{N} (\delta_n y_n)^2} \qquad \forall y \in \Gamma, \tag{3.2}$$

for some constant $C_\rho > 0$ *and* δ_n *strictly positive if* Γ_n *is unbounded and zero otherwise.*

The parameter δ_n in (3.2) gives a scale for the decay of ρ at infinity and provides an estimate of the dispersion of the random variable Y_n. On the other hand, the parameter α_n in (3.1) controls the growth of the forcing term f at infinity.

Remark 3 (growth of f). The convergence result given in Theorem 4.1 in section 4 extends to a wider class of functions f. For instance, we could take $f \in C^0_\sigma(\Gamma; L^2(D))$ with $\sigma = e^{-\sum_{n=1}^{N} (\delta_n y_n)^2/8}$. Yet the class given in (3.1) is already large enough for most practical applications (see Example 2).

We can now choose any suitable auxiliary density $\hat{\rho}(y) = \prod_{n=1}^{N} \hat{\rho}_n(y_n)$ that satisfies, for each $n = 1, \ldots, N$,

$$C^n_{min} e^{-(\delta_n y_n)^2} \leq \hat{\rho}_n(y_n) < C^n_{max} e^{-(\delta_n y_n)^2} \qquad \forall y_n \in \Gamma_n \tag{3.3}$$

for some positive constants C^n_{min} and C^n_{max} that do not depend on y_n.

Observe that this choice satisfies the requirement given in (2.2), i.e., $\|\rho/\hat{\rho}\|_{L^\infty(\Gamma)} \leq C_\rho / C_{min}$ with $C_{min} = \prod_{n=1}^{N} C^n_{min}$.

Under the above assumptions, the following inclusions hold true:

$$C^0_\sigma(\Gamma; V) \subset L^2_{\hat{\rho}}(\Gamma; V) \subset L^2_\rho(\Gamma; V)$$

with continuous embedding. Indeed, on one hand we have

$$\|v\|_{L^2_\rho(\Gamma;V)} \leq \left\| \frac{\rho}{\hat{\rho}} \right\|^{\frac{1}{2}}_{L^\infty(\Gamma)} \|v\|_{L^2_{\hat{\rho}}(\Gamma;V)} \leq \sqrt{\frac{C_\rho}{C_{min}}} \|v\|_{L^2_{\hat{\rho}}(\Gamma;V)}.$$

On the other hand,

$$\|v\|^2_{L^2_{\hat{\rho}}(\Gamma;V)} = \int_\Gamma \hat{\rho}(y) \|v(y)\|^2_V \, dy \leq \|v\|^2_{C^0_\sigma(\Gamma;V)} \int_\Gamma \frac{\hat{\rho}(y)}{\sigma^2(y)} \, dy \leq \prod_{n=1}^{N} M_n \|v\|^2_{C^0_\sigma(\Gamma;V)}$$

with $M_n = \int_{\Gamma_n} \hat{\rho}_n / \sigma_n^2$. Now, for Γ_n bounded, $M_n \leq C^n_{max} |\Gamma_n|$, whereas if Γ_n is unbounded,

$$M_n = \int_{\Gamma_n} \left(e^{-\frac{(\delta_n y)^2}{2} + 2\alpha_n |y|} \right) e^{\frac{(\delta_n y)^2}{2}} \hat{\rho}_n(y) \, dy \leq C^n_{max} \sqrt{\frac{2\pi}{\delta_n}} e^{2(\alpha_n/\delta_n)^2}.$$

The first result we need is the following lemma.

LEMMA 3.1. *If $f \in C^0_\sigma(\Gamma; L^2(D))$ and $a \in C^0_{loc}(\Gamma; L^\infty(D))$, uniformly bounded away from zero, then the solution to problem* (1.7) *satisfies $u \in C^0_\sigma(\Gamma; H^1_0(D))$.*

The proof of this lemma is immediate. The next result concerns the analyticity of the solution u whenever the diffusivity coefficient a and the forcing term f are infinitely differentiable with respect to y, under mild assumptions on the growth of their derivatives in y. We will perform a one-dimensional analysis in each direction y_n, $n = 1, \dots, N$. For this, we introduce the following notation:

$$\Gamma_n^* = \prod_{\substack{j=1 \\ j\neq n}}^{N} \Gamma_j,$$

with y_n^* denoting an arbitrary element of Γ_n^*. Similarly, we set

$$\hat{\rho}_n^* = \prod_{\substack{j=1 \\ j\neq n}}^{N} \hat{\rho}_j$$

and

$$\sigma_n^* = \prod_{\substack{j=1 \\ j\neq n}}^{N} \sigma_j.$$

LEMMA 3.2. *Under the assumption that, for every $y = (y_n, y_n^*) \in \Gamma$, there exists $\gamma_n < +\infty$ such that*

$$\left\| \frac{\partial_{y_n}^k a(y)}{a(y)} \right\|_{L^\infty(D)} \leq \gamma_n^k k! \qquad \text{and} \qquad \frac{\|\partial_{y_n}^k f(y)\|_{L^2(D)}}{1 + \|f(y)\|_{L^2(D)}} \leq \gamma_n^k k!, \tag{3.4}$$

the solution $u(y_n, y_n^*, x)$ as a function of y_n, $u : \Gamma_n \to C^0_{\sigma_n^*}(\Gamma_n^*; H^1_0(D))$ admits an analytic extension $u(z, y_n^*, x)$, $z \in \mathbb{C}$, in the region of the complex plane

$$\Sigma(\Gamma_n; \tau_n) \equiv \{z \in \mathbb{C}, \ \operatorname{dist}(z, \Gamma_n) \leq \tau_n\} \tag{3.5}$$

with $0 < \tau_n < 1/(2\gamma_n)$. Moreover, for all $z \in \Sigma(\Gamma_n; \tau_n)$,

$$\|\sigma_n(\operatorname{Re} z)\, u(z)\|_{C^0_{\sigma_n^*}(\Gamma_n^*; H^1_0(D))} \leq \frac{C_P e^{\alpha_n \tau_n}}{a_{min}(1 - 2\tau_n\gamma_n)} (2\|f\|_{C^0_\sigma(\Gamma; H^1_0(D))} + 1) \tag{3.6}$$

with the constant C_p as in (1.3).

Proof. In every point $y \in \Gamma$, the kth derivative of u with respect to y_n satisfies the problem

$$B(y; \partial_{y_n}^k u, v) = -\sum_{l=1}^{k} \binom{k}{l} \partial_{y_n}^l B(y; \partial_{y_n}^{k-l} u, v) + (\partial_{y_n}^k f, v) \quad \forall v \in H^1_0(D),$$

where B is the parametric bilinear form $B(y; u, v) = \int_D a(y) \nabla u \cdot \nabla v \, dx$. Hence

$$\begin{aligned} \|\sqrt{a(y)} \nabla \partial_{y_n}^k u\|_{L^2(D)} \leq & \sum_{l=1}^{k} \binom{k}{l} \left\| \frac{\partial_{y_n}^l a(y)}{a(y)} \right\|_{L^\infty(D)} \|\sqrt{a(y)} \nabla \partial_{y_n}^{k-l} u\|_{L^2(D)} \\ & + \frac{C_p}{\sqrt{a_{min}}} \|\partial_{y_n}^k f\|_{L^2(D)}. \end{aligned}$$

Setting $R_k = \|\sqrt{a(y)}\nabla\partial_{y_n}^k u\|_{L^2(D)}/k!$ and using the bounds on the derivatives of a and f, we obtain the recursive inequality

$$R_k \le \sum_{l=1}^{k} \gamma_n^l R_{k-l} + \frac{C_p}{\sqrt{a_{min}}}\gamma_n^k(1+\|f\|_{L^2(D)}).$$

The generic term R_k admits the bound

$$R_k \le (2\gamma_n)^k R_0 + \frac{C_p}{\sqrt{a_{min}}}(1+\|f\|_{L^2(D)})\gamma_n^k \sum_{l=0}^{k-1} 2^l.$$

Observing that $R_0 = \|\sqrt{a(y)}\nabla u(y)\|_{L^2(D)} \le \frac{C_p}{\sqrt{a_{min}}}\|f(y)\|_{L^2(D)}$ and

$$\frac{\|\nabla\partial_{y_n}^k u\|_{L^2(D)}}{k!} \le \frac{R_k}{\sqrt{a_{min}}},$$

we get the final estimate on the growth of the derivatives of u,

$$\frac{\|\nabla\partial_{y_n}^k u(y)\|_{L^2(D)}}{k!} \le \frac{C_p}{a_{min}}(2\|f(y)\|_{L^2(D)}+1)(2\gamma_n)^k.$$

We now define for every $y_n \in \Gamma_n$ the power series $u : \mathbb{C} \to C^0_{\sigma_n^*}(\Gamma_n^*, H_0^1(D))$ as

$$u(z, y_n^*, x) = \sum_{k=0}^{\infty} \frac{(z-y_n)^k}{k!}\partial_{y_n}^k u(y_n, y_n^*, x).$$

Hence,

$$\begin{aligned}
\sigma_n(y_n)\|u(z)\|_{C^0_{\sigma_n^*}(\Gamma_n^*,H_0^1(D))} &\le \sum_{k=0}^{\infty} \frac{|z-y_n|^k}{k!}\sigma_n(y_n)\|\partial_{y_n}^k u(y_n)\|_{C^0_{\sigma_n^*}(\Gamma_n^*;H_0^1(D))} \\
&\le \frac{C_P}{a_{min}} \max_{y_n\in\Gamma_n}\left\{\sigma_n(y_n)\left(2\|f(y_n)\|_{C^0_{\sigma_n^*}(\Gamma_n^*;L^2(D))}+1\right)\right\}\sum_{k=0}^{\infty}\left(|z-y_n|2\gamma_n\right)^k \\
&\le \frac{C_P}{a_{min}}(2\|f\|_{C^0_\sigma(\Gamma;L^2(D))}+1)\sum_{k=0}^{\infty}\left(|z-y_n|2\gamma_n\right)^k,
\end{aligned}$$

where we have exploited the fact that $\sigma_n(y_n) \le 1$ for all $y_n \in \Gamma_n$; the series converges for all $z \in \mathbb{C}$ such that $|z-y_n| \le \tau_n < 1/(2\gamma_n)$. Moreover, in the ball $|z-y_n| \le \tau_n$, we have, by virtue of (3.1), $\sigma_n(\mathrm{Re}\, z) \le e^{\alpha_n\tau_n}\sigma_n(y_n)$, and then

$$\sigma_n(\mathrm{Re}\, z)\|u(z)\|_{C^0_{\sigma_n^*}(\Gamma_n^*,H_0^1(D))} \le \frac{C_P e^{\alpha_n\tau_n}}{a_{min}(1-2\tau_n\gamma_n)}(2\|f\|_{C^0_\sigma(\Gamma;L^2(D))}+1).$$

The power series converges for every $y_n \in \Gamma_n$; hence, by a continuation argument, the function u can be extended analytically on the whole region $\Sigma(\Gamma_n;\tau_n)$ and estimate (3.6) follows. □

Example 3. Let us consider the case where the diffusivity coefficient a is expanded in a linear truncated Karhunen–Loève expansion

$$a(\omega, x) = b_0(x) + \sum_{n=1}^{N}\sqrt{\lambda_n}b_n(x)Y_n(\omega),$$

provided that such expansion guarantees $a(\omega, x) \geq a_{min}$ for almost every $\omega \in \Omega$ and $x \in D$ [18]. In this case we have

$$\left\| \frac{\partial_{y_n}^k a}{a} \right\|_{L^\infty(\Gamma \times D)} \leq \begin{cases} \sqrt{\lambda_n} \|b_n\|_{L^\infty(D)} / a_{min} & \text{for } k = 1, \\ 0 & \text{for } k > 1 \end{cases}$$

and we can safely take $\gamma_n = \sqrt{\lambda_n} \|b_n\|_{L^\infty(D)} / a_{min}$ in (3.4).

If we consider, instead, a truncated *exponential* expansion

$$a(\omega, x) = a_{min} + e^{b_0(x) + \sum_{n=1}^N \sqrt{\lambda_n} b_n(x) Y_n(\omega)},$$

we have

$$\left\| \frac{\partial_{y_n}^k a}{a} \right\|_{L^\infty(\Gamma \times D)} \leq \left(\sqrt{\lambda_n} \|b_n\|_{L^\infty(D)} \right)^k$$

and we can take $\gamma_n = \sqrt{\lambda_n} \|b_n\|_{L^\infty(D)}$. Hence, both choices fulfill the assumption in Lemma 3.2.

Example 4. Similarly to the previous case, let us consider a forcing term f of the form

$$f(\omega, x) = c_0(x) + \sum_{n=1}^N c_n(x) Y_n(\omega),$$

where the random variables Y_n are Gaussian (either independent or not) and the functions $c_n(x)$ are square integrable for any $n = 1, \ldots, N$. Then, the function f belongs to the space $C^0_\sigma(\Gamma; L^2(D))$, with weight σ defined in (3.1), for any choice of the exponent coefficients $\alpha_n > 0$.

Moreover,

$$\frac{\|\partial_{y_n}^k f(y)\|_{L^2(D)}}{1 + \|f(y)\|_{L^2(D)}} \leq \begin{cases} \|c_n\|_{L^2(D)} & \text{for } k = 1, \\ 0 & \text{for } k > 1, \end{cases}$$

and we can safely take $\gamma_n = \|c_n\|_{L^2(D)}$ in (3.4). Hence, such a forcing term satisfies the assumptions of Lemma 3.2. In this case, though, the solution u is linear with respect to the random variables Y_n (hence, clearly analytic), and our theory is not needed.

Observe that the regularity results are valid also for the semidiscrete solution u_h.

4. Convergence analysis. Our aim is to give a priori estimates for the total error $\epsilon = u - u_{h,p}$ in the natural norm $L^2_\rho(\Gamma) \otimes H^1_0(D)$. The next theorem states the convergence result we are seeking, and the rest of the section will be devoted to its proof. In particular, we will prove that the error decays (sub)exponentially fast with respect to p under the regularity assumptions made in section 3. The convergence with respect to h will be dictated by standard approximability properties of the finite element space $H_h(D)$ and the regularity in space of the solution u (see, e.g., [12, 11]).

THEOREM 4.1. *Under the assumptions of Lemmas* 3.1 *and* 3.2*, there exist positive constants r_n, $n = 1, \ldots, N$, and C, independent of h and p, such that*

$$\begin{aligned} \|u - u_{h,p}\|_{L^2_\rho \otimes H^1_0} \leq & \frac{1}{\sqrt{a_{min}}} \inf_{v \in L^2_\rho \otimes H_h} \left(\int_{\Gamma \times D} \rho a |\nabla(u - v)|^2 \right)^{\frac{1}{2}} \\ & + C \sum_{n=1}^N \beta_n(p_n) \exp\{-r_n p_n^{\theta_n}\}, \end{aligned} \tag{4.1}$$

where

- *if* Γ_n *is bounded,* $\begin{cases}\theta_n = \beta_n = 1,\\ r_n = \log\left[\frac{2\tau_n}{|\Gamma_n|}\left(1+\sqrt{1+\frac{|\Gamma_n|^2}{4\tau_n^2}}\right)\right],\end{cases}$
- *if* Γ_n *is unbounded,* $\begin{cases}\theta_n = 1/2, \quad \beta_n = O(\sqrt{p_n}),\\ r_n = \tau_n\delta_n,\end{cases}$

τ_n *is smaller than the distance between* Γ_n *and the nearest singularity in the complex plane, as defined in Lemma* 3.2*, and* δ_n *is defined as in* (3.2).

The first term on the right-hand side of (4.1) concerns the space approximability of u in the subspace $H_h(D)$ and is controlled by the mesh size h. The actual rate of convergence will depend on the regularity in space of $a(y)$ and $f(y)$ for each $y \in \Gamma$ as well as on the smoothness on the domain D. Observe that an h or $h-p$ adaptive strategy to reduce the error in space is not precluded by this approach.

The exponential rate of convergence in the Y direction depends on the constants r_n, which in turn are related to the distances from the sets Γ_n to their nearest singularities in the complex plane. In Examples 3 and 4 we have estimated these constants in the case where the random fields a and f are represented by either a linear or exponential truncated Karhunen–Loève expansion. Hence, a full characterization of the convergence rate is available in these cases.

Observe that in Theorem 4.1 it is not necessary to assume the finiteness of the second moment of the coefficient a.

Before proving the theorem, we recall some known results of approximation theory for a function f defined on a one-dimensional domain (bounded or unbounded) with values in a Banach space V, $f : \Gamma \subset \mathbb{R} \to V$. As in section 2, let $\rho : \Gamma \to \mathbb{R}^+$ be a positive weight which satisfies, for all $y \in \Gamma$, $\rho(y) \le C_M e^{-(\delta y)^2}$ for some $C_M > 0$ and δ strictly positive if Γ is unbounded and zero otherwise; let $y_k \in \Gamma$, $k = 1, \dots, p+1$ be the set of zeros of the polynomial of degree p orthogonal to the space $\mathcal{P}_{p-1}$ with respect to the weight ρ; and let σ be an extra positive weight such that $\sigma(y) \ge C_m e^{-(\delta y)^2/4}$ for some $C_m > 0$. With this choice, the embedding $C^0_\sigma(\Gamma; V) \subset L^2_\rho(\Gamma; V)$ is continuous. Observe that the condition on σ is satisfied both by a Gaussian weight $\sigma = e^{-(\mu y)^2}$ with $\mu \le \delta/2$ and by an exponential weight $\sigma = e^{-\alpha|y|}$ for any $\alpha \ge 0$. Finally, we denote by $\mathcal{I}_p$ the Lagrange interpolant operator, $\mathcal{I}_p v(y) = \sum_{k=1}^{p+1} v(y_k) l_k(y)$ for every continuous function v, and by $\omega_k = \int_\Gamma l_k^2(y)\rho(y)\,dy$ the weights of the Gaussian quadrature formula built upon $\mathcal{I}_p$.

The following two lemmas are a slight generalization of a classical result by Erdös and Turán [17].

LEMMA 4.2. *The operator* $\mathcal{I}_p : C^0_\sigma(\Gamma; V) \to L^2_\rho(\Gamma; V)$ *is continuous.*

Proof. We have, indeed, that for any $v \in C^0_\sigma(\Gamma; V)$

$$\|\mathcal{I}_p v\|^2_{L^2_\rho(\Gamma;V)} = \int_\Gamma \left\| \sum_{k=1}^{p+1} v(y_k) l_k(y) \right\|^2_V \rho(y)\,dy \le \int_\Gamma \left(\sum_{k=1}^{p+1} \|v(y_k)\|_V l_k(y) \right)^2 \rho(y)\,dy.$$

Thanks to the orthogonality property $\int_\Gamma l_j(y) l_k(y) \rho(y)\, dy = \delta_{jk}$, we have

$$\begin{aligned}
\|\mathcal{I}_p v\|^2_{L^2_\rho(\Gamma;V)} &\le \int_\Gamma \sum_{k=1}^{p+1} \|v(y_k)\|^2_V \, l_k^2(y) \rho(y)\, dy \\
&\le \max_{k=1,\dots,p+1} \|v(y_k)\|^2_V \sigma^2(y_k) \sum_{k=1}^{p+1} \int_\Gamma \frac{l_k^2(y)\rho(y)}{\sigma^2(y_k)}\, dy \\
&\le \|v\|^2_{C^0_\sigma(\Gamma;V)} \sum_{k=1}^{p+1} \frac{\omega_k}{\sigma^2(y_k)}.
\end{aligned}$$

In the case of Γ bounded, we have $\sigma \ge C_m$ and $\sum_{k=1}^{p+1} \omega_k = 1$ for any p, and the result follows immediately. For Γ unbounded, since $\rho(y) \le C_M e^{-(\delta y)^2}$, all the even moments $c_{2m} = \int_\Gamma y^{2m} \rho(y)\, dy$ are bounded, up to a constant, by the moments of the Gaussian density $e^{-(\delta y)^2}$. Therefore, using a result from Uspensky in 1928 [36], it follows that

$$\sum_{k=1}^{p+1} \frac{\omega_k}{\sigma^2(y_k)} \overset{p\to\infty}{\longrightarrow} \int_\Gamma \frac{\rho(y)}{\sigma^2(y)}\, dy \le \frac{C_M}{C_m^2} \sqrt{\frac{2\pi}{\delta}},$$

and we conclude that

$$\|\mathcal{I}_p v\|_{L^2_\rho(\Gamma;V)} \le C_1 \|v\|_{C^0_\sigma(\Gamma;V)}. \qquad \square$$

LEMMA 4.3. *For every function $v \in C^0_\sigma(\Gamma;V)$ the interpolation error satisfies*

$$\|v - \mathcal{I}_p v\|_{L^2_\rho(\Gamma;V)} \le C_2 \inf_{w \in \mathcal{P}_p(\Gamma)\otimes V} \|v - w\|_{C^0_\sigma(\Gamma;V)}$$

with a constant C_2 independent of p.

Proof. Let us observe that for all $w \in \mathcal{P}_p(\Gamma) \otimes V$, it holds that $\mathcal{I}_p w = w$. Then,

$$\begin{aligned}
\|v - \mathcal{I}_p v\|_{L^2_\rho(\Gamma;V)} &\le \|v - w\|_{L^2_\rho(\Gamma;V)} + \|\mathcal{I}_p(w - v)\|_{L^2_\rho(\Gamma;V)} \\
&\le C_2 \|v - w\|_{C^0_\sigma(\Gamma;V)}.
\end{aligned}$$

Since w is arbitrary in the right-hand side, the result follows. $\square$

Lemma 4.3 relates the approximation error $(v - \mathcal{I}_p v)$ in the L^2_ρ-norm with the *best approximation* error in the weighted C^0_σ-norm for any weight $\sigma(y) \ge C_m e^{-(\delta y)^2/4}$. We now analyze the best approximation error for a function $v : \Gamma \to V$ which admits an analytic extension in the complex plane, in the region $\Sigma(\Gamma;\tau) = \{z \in \mathbb{C},\ \operatorname{dist}(z,\Gamma) < \tau\}$ for some $\tau > 0$. We will still denote the extension by v; in this case, τ represents the distance between $\Gamma \subset \mathbb{R}$ and the nearest singularity of $v(z)$ in the complex plane.

We study separately the two cases of Γ bounded and unbounded. We start with the bounded case, in which the extra weight σ is set equal to 1. The following result is an immediate extension of the result given in [14, Chapter 7, section 8]

LEMMA 4.4. *Given a function $v \in C^0(\Gamma;V)$ which admits an analytic extension in the region of the complex plane $\Sigma(\Gamma;\tau) = \{z \in \mathbb{C},\ \operatorname{dist}(z,\Gamma) \le \tau\}$ for some $\tau > 0$, it holds that*

$$\min_{w \in \mathcal{P}_p \otimes V} \|v - w\|_{C^0(\Gamma;V)} \le \frac{2}{\varrho - 1} e^{-p \log(\varrho)} \max_{z \in \Sigma(\Gamma;\tau)} \|v(z)\|_V,$$

where

$$1 < \varrho = \frac{2\tau}{|\Gamma|} + \sqrt{1 + \frac{4\tau^2}{|\Gamma|^2}}.$$

Proof. We sketch the proof for completeness. We first make a change of variables, $y(t) = y_0 + \frac{|\Gamma|}{2}t$, where y_0 is the midpoint of Γ. Hence, $y([-1,1]) = \Gamma$. We set $\tilde{v}(t) = v(y(t))$. Clearly, $\tilde{v}$ can be extended analytically in the region of the complex plane $\Sigma([-1,1]; 2\tau/|\Gamma|) \equiv \{z \in \mathbb{C}, \mathrm{dist}(z,[-1,1]) \le 2\tau/|\Gamma|\}$.

We then introduce the Chebyshev polynomials $C_k(t)$ on $[-1,1]$ and the expansion of $\tilde{v} : [-1,1] \to V$ as

$$\tilde{v}(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} a_k C_k(t), \tag{4.2}$$

where the Fourier coefficients $a_k \in V$, $k = 0, 1, \ldots$, are defined as

$$a_k = \frac{1}{\pi}\int_{-\pi}^{\pi} \tilde{v}(\cos(t))\,\cos(kt)\,dt.$$

It is well known (see, e.g., [14, 9]) that the series (4.2) converges in any elliptic disc $D_\varrho \subset \mathbb{C}$, with $\varrho > 1$, delimited by the ellipse

$$E_\varrho = \left\{ z = t + is \in \mathbb{C},\ \ t = \frac{\varrho + \varrho^{-1}}{2}\cos\phi,\ \ s = \frac{\varrho - \varrho^{-1}}{2}\sin(\phi),\ \ \phi \in [0, 2\pi) \right\}$$

in which the function $\tilde{v}$ is analytic. Moreover (see [14] for details), we have

$$\|a_k\|_V \le 2\varrho^{-k} \max_{z \in D_\varrho} \|\tilde{v}(z)\|_V.$$

If we denote by $\Pi_p v \in \mathcal{P}_p(\Gamma) \otimes V$ the truncated Chebyshev expansion up to the polynomial degree p and observe that $|C_k(t)| \le 1$ for all $t \in [-1,1]$, we have

$$\begin{aligned}\min_{w \in \mathcal{P}_p \otimes V} \|v - w\|_{C^0(\Gamma;V)} &\le \|\tilde{v} - \Pi_p \tilde{v}\|_{C^0([-1,1];V)} \\ &\le \sum_{k=p+1}^{\infty} \|a_k\|_V \le \frac{2}{\varrho - 1}\varrho^{-p} \max_{z \in D_\varrho} \|\tilde{v}(z)\|_V.\end{aligned}$$

Finally, we have to link ϱ to the size of the analyticity region of $\tilde{v}$. It is easy to verify that the ellipse given by

$$\varrho = \frac{2\tau}{|\Gamma|}\left(1 + \sqrt{1 + \frac{|\Gamma|^2}{4\tau^2}}\right)$$

is the largest ellipse that can be drawn inside $\Sigma([-1,1]; 2\tau/|\Gamma|)$, and this proves the stated result. ∎

For the case of unbounded Γ we first recall a result given in [23] and then we state in Lemma 4.6 a result tuned to our situation.

We denote by $H_n(y) \in P_n(\mathbb{R})$ the normalized Hermite polynomials

$$H_n(y) = \sqrt{\pi^{\frac{1}{2}} 2^n n!}\,(-1)^n e^{y^2}\frac{\partial^n}{\partial y^n}\left(e^{-y^2}\right)$$

and by $h_n(y) = e^{-y^2/2}H_n(y)$ the Hermite functions. We recall that the Hermite polynomials form a complete orthonormal basis of the $L^2(\mathbb{R})$ space with respect to the weight e^{-y^2}, i.e.,

$$\int_{\mathbb{R}} H_k(y)H_l(y)e^{-y^2}\,dy = \delta_{kl}.$$

LEMMA 4.5 (Hille [23]). *Let $f(z)$ be an analytic function in the strip of the complex plane $\Sigma(\mathbb{R};\tau) \equiv \{z = (y+iw) \in \mathbb{C}, -\tau \le w \le \tau\}$. A necessary and sufficient condition in order that the Fourier–Hermite series*

$$\sum_{k=0}^{\infty} f_k h_k(z), \qquad f_k = \int_{\mathbb{R}} f(y)h_k(y)\,dy, \tag{4.3}$$

shall exist and converge to the sum $f(z)$ in $\Sigma(\mathbb{R};\tau)$ is that for every β, $0 \le \beta < \tau$, there exists a finite positive $C(\beta)$ such that

$$|f(y+iw)| \le C(\beta)e^{-|y|\sqrt{\beta^2-w^2}}, \quad -\infty < y < \infty, \ -\beta \le w \le \beta. \tag{4.4}$$

Moreover, the following bound for the Fourier coefficients holds:

$$|f_n| \le Ce^{-\tau\sqrt{2n+1}}. \tag{4.5}$$

In particular, the previous result tells us that, in order to have exponential decay of the Fourier coefficients f_n, we not only need $f(z)$ to be analytic in $\Sigma(\mathbb{R};\tau)$ but also must require that it decays on the real line, for $y \to \infty$, at least as $e^{-\tau|y|}$.

We now introduce two weights: the exponential $\sigma = e^{-\alpha|y|}$, for some $\alpha > 0$, and the Gaussian $G = e^{-(\delta y)^2/4}$. We recall that Lemma 4.3 holds for both. We will assume that the function v is in the space $C^0_\sigma(\Gamma;V)$, but we will measure the best approximation error in the weaker norm $C^0_G(\Gamma;V)$, with Gaussian weight, so that we can use the result from Hille given in Lemma 4.5. The following lemma holds.

LEMMA 4.6. *Let v be a function in $C^0_\sigma(\mathbb{R};V)$. We suppose that v admits an analytic extension in the strip of the complex plane $\Sigma(\mathbb{R};\tau) = \{z \in \mathbb{C}, \ \operatorname{dist}(z,\mathbb{R}) \le \tau\}$ for some $\tau > 0$, and that*

$$\forall z = (y+iw) \in \Sigma(\mathbb{R};\tau), \quad \sigma(y)\|v(z)\|_V \le C_v(\tau).$$

Then, for any $\delta > 0$, there exist a constant C, independent of p, and a function $\Theta(p) = O(\sqrt{p})$ such that

$$\min_{w\in\mathcal{P}_p\otimes V} \max_{y\in\mathbb{R}} \left| \|v(y) - w(y)\|_V e^{-\frac{(\delta y)^2}{4}} \right| \le C\Theta(p)e^{-\tau\delta\sqrt{p}}.$$

Proof. We introduce the change of variable $t = \delta\, y/\sqrt{2}$ and we denote $\tilde{v}(t) = v(y(t))$. Observe that $\tilde{v} \in C^0_{\tilde{\sigma}}(\mathbb{R};V)$ with weight $\tilde{\sigma} = e^{-\sqrt{2}\frac{\alpha}{\delta}|t|}$. We consider the expansion of $\tilde{v}$ in Hermite polynomials

$$\tilde{v}(t) = \sum_{k=0}^{\infty} v_k H_k(t), \quad \text{where } v_k \in V, \quad v_k = \int_{\mathbb{R}} \tilde{v}(t)H_k(t)e^{-t^2}\,dt. \tag{4.6}$$

We now set $f(z) = \tilde{v}(z)e^{-\frac{z^2}{2}}$. Observe that the Hermite expansion of f as defined in (4.3) has the same Fourier coefficients as the expansion of $\tilde{v}$ defined in (4.6). Indeed

$$f_k = \int_{\mathbb{R}} f(t)h_k(t)\,dt = \int_{\mathbb{R}} \tilde{v}(t)H_k(t)e^{-t^2}\,dt = v_k.$$

Clearly, $f(z)$ is analytic in the strip $\Sigma(\mathbb{R}; \frac{\tau\delta}{\sqrt{2}})$, being the product of analytic functions. Moreover,

$$\|f(y+iw)\|_V = |e^{-\frac{(y+iw)^2}{2}}| \|\tilde{v}(z)\|_V \leq e^{-\frac{y^2-w^2}{2}} e^{\sqrt{2}\frac{\alpha}{\delta}|y|} C_v(\tau).$$

Setting

$$C(\beta) = \max_{\substack{-\infty<y<\infty \\ -\beta\leq w\leq\beta}} \exp\left\{-\frac{y^2-w^2}{2} + \sqrt{2}\frac{\alpha}{\delta}|y| + |y|\sqrt{\beta^2-w^2}\right\},$$

which is bounded for all $-\frac{\tau\delta}{\sqrt{2}} \leq \beta \leq \frac{\tau\delta}{\sqrt{2}}$, the function $f(z)$ satisfies the hypotheses of Lemma 4.5. Hence the Hermite series converges in $\Sigma(\mathbb{R}; \frac{\tau\delta}{\sqrt{2}})$ and the Fourier coefficients v_k behave as in (4.5). We chose $w \in \mathcal{P}_p \otimes V$ as the truncated Hermite expansion of v, up to degree p: $\tilde{w}(t) = \Pi_p \tilde{v}(t) = \sum_{k=0}^{p} v_k H_k(t)$. We have

$$\begin{aligned} E_p(v) &= \min_{w\in\mathcal{P}_p\otimes V} \max_{y\in\mathbb{R}} \left| \|v(y) - w(y)\|_V e^{-\frac{(\delta y)^2}{4}} \right| \\ &\leq \max_{t\in\mathbb{R}} \left| \|\tilde{v}(t) - \Pi_p\tilde{v}(t)\|_V e^{-\frac{t^2}{2}} \right| \leq \max_{t\in\mathbb{R}} \| \sum_{k=p+1}^{\infty} v_k h_k(t) \|_V. \end{aligned}$$

It is well known (see, e.g., [8]) that the Hermite functions $h_k(t)$ satisfy $|h_k(t)| < 1$ for all $t \in \mathbb{R}$ and all $k = 0, 1, \ldots$. Hence, the previous series can be bound as

$$E_p(v) \leq \sum_{k=p+1}^{\infty} \|v_k\|_V \leq C \sum_{k=p+1}^{\infty} e^{-\frac{\tau\delta}{\sqrt{2}}\sqrt{2k+1}}.$$

Lemma A.2 in the appendix provides a bound for such a series, and this concludes the proof. ∎

We are now ready to prove Theorem 4.1.

Proof of Theorem 4.1. The error naturally splits into $\epsilon = (u-u_h)+(u_h-u_{h,p})$. The first term depends on the space discretization only and can be estimated easily; indeed, the function u_h is the orthogonal projection of u onto the subspace $L^2_\rho(\Gamma) \otimes H^1_0(D)$ with respect to the inner product $\int_{\Gamma\times D} \rho a |\nabla \cdot|^2$. Hence

$$\begin{aligned} \|u - u_h\|_{L^2_\rho(\Gamma)\otimes H^1_0(D)} &\leq \frac{1}{\sqrt{a_{min}}} \left(\int_{\Gamma\times D} \rho a |\nabla(u-u_h)|^2 \right)^{\frac{1}{2}} \\ &\leq \frac{1}{\sqrt{a_{min}}} \inf_{v\in L^2_\rho(\Gamma)\otimes H_h(D)} \left(\int_{\Gamma\times D} \rho a |\nabla(u-v)|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

The second term $u_h - u_{h,p}$ is an interpolation error. We recall, indeed, that $u_{h,p} = \mathcal{I}_p u_h$. To lighten the notation, we will drop the subscript h, being understood that we work on the semidiscrete solution. We recall, moreover, that u_h has the same regularity as the exact solution u with respect to y.

To analyze this term we employ a one-dimensional argument. We first pass from the norm L^2_ρ to $L^2_{\hat{\rho}}$:

$$\|u - \mathcal{I}_p u\|_{L^2_\rho\otimes H^1_0} \leq \left\| \frac{\rho}{\hat{\rho}} \right\|^{\frac{1}{2}}_{L^\infty(\Gamma)} \|u - \mathcal{I}_p u\|_{L^2_{\hat{\rho}}\otimes H^1_0}.$$

Here we adopt the same notation as in section 3; namely, we indicate with $\bullet_n$ a quantity relative to the direction y_n and with $\bullet_n^*$ the analogous quantity relative to all other directions y_j, $j \neq n$. We focus on the first direction y_1 and define an interpolation operator $\mathcal{I}_1 : C^0_{\sigma_1}(\Gamma_1; L^2_{\hat{\rho}_1^*} \otimes H^1_0) \to L^2_{\hat{\rho}_1}(\Gamma_1; L^2_{\hat{\rho}_1^*} \otimes H^1_0)$,

$$\mathcal{I}_{p_1} v(y_1, y_1^*, x) = \sum_{k=1}^{p_1+1} v(y_{1,k}, y_1^*, x) l_{1,k}(y_1).$$

Then, the global interpolant $\mathcal{I}_p$ can be written as the composition of two interpolation operators $\mathcal{I}_p = \mathcal{I}_1 \circ \mathcal{I}_p^{(1)}$, where $\mathcal{I}_p^{(1)}$ is the interpolation operator in all directions $y_2, y_3, \ldots, y_N$ except y_1: $\mathcal{I}_p^{(1)} : C^0_{\sigma_1^*}(\Gamma_1^*; H^1_0) \to L^2_{\hat{\rho}_1^*}(\Gamma_1^*; H^1_0)$. We then have

$$\|u - \mathcal{I}_p u\|_{L^2_{\hat{\rho}} \times H^1_0} \leq \underbrace{\|u - \mathcal{I}_1 u\|_{L^2_{\hat{\rho}} \times H^1_0}}_{\mathrm{I}} + \underbrace{\|\mathcal{I}_1(u - \mathcal{I}_p^{(1)} u)\|_{L^2_{\hat{\rho}} \times H^1_0}}_{\mathrm{II}}.$$

Let us bound the first term. We think of u as a function of y_1 with values in a Banach space V, $u \in L^2_{\hat{\rho}_1}(\Gamma_1; V)$, where $V = L^2_{\hat{\rho}_1^*}(\Gamma_1^*) \otimes H^1_0(D)$. Under Assumption 2 in section 3 and the choice of $\hat{\rho}$ given in (3.3), the following inclusions hold true:

$$C^0_{\sigma_1}(\Gamma_1; V) \subset C^0_{G_1}(\Gamma_1; V) \subset L^2_{\hat{\rho}_1}(\Gamma_1; V)$$

with $\sigma_1 = G_1 = 1$ if Γ_1 is bounded and $\sigma_1 = e^{-\alpha_1 |y_1|}$, $G_1 = e^{-\frac{(\delta_1 y_1)^2}{4}}$ if Γ_1 is unbounded. We know also from Lemma 4.2 that the interpolation operator $\mathcal{I}_1$ is continuous both as an operator from $C^0_{\sigma_1}(\Gamma_1; V)$ with values in $L^2_{\hat{\rho}_1}(\Gamma_1; V)$ and from $C^0_{G_1}(\Gamma_1; V)$ in $L^2_{\hat{\rho}_1}(\Gamma_1; V)$. In particular, we can estimate

$$\mathrm{I} = \|u - \mathcal{I}_1 u\|_{L^2_{\hat{\rho}_1}(\Gamma_1; V)} \leq C_2 \inf_{w \in \mathcal{P}_{p_1} \otimes V} \|u - w\|_{C^0_{G_1}(\Gamma; V)}.$$

To bound the best approximation error in $C^0_{G_1}(\Gamma; V)$, in the case of Γ_1 bounded we use Lemma 4.4, whereas if Γ_1 is unbounded, we employ Lemma 4.6 and the fact that $u \in C^0_{\sigma_1}(\Gamma_1; V)$ (see Lemma 3.1). In both cases, we need the analyticity result, for the solution u, stated in Lemma 3.2. Putting everything together, we can say that

$$\mathrm{I} \leq \begin{cases} C e^{-r_1 p_1}, & \Gamma_1 \text{ bounded}, \\ C \Theta(p_1) e^{-r_1 \sqrt{p_1}}, & \Gamma_1 \text{ unbounded}, \end{cases}$$

the value of r_1 being specified in Lemmas 4.4 and 4.6. To bound the term II, we use Lemma 4.2:

$$\mathrm{II} \leq C_1 \|u - \mathcal{I}_p^{(1)} u\|_{C^0_{\sigma_1}(\Gamma_1; V)}.$$

The term on the right-hand side is again an interpolation error. Thus we have to bound the interpolation error in all the other $N-1$ directions, uniformly with respect to y_1 (in the weighted norm $C^0_{\sigma_1}$). We can proceed iteratively, defining an interpolation $\mathcal{I}_2$, bounding the resulting error in the direction y_2, and so on. ∎

4.1. Convergence of moments. In some cases one might be interested only in computing the first few moments of the solution, namely $E[u^m]$, $m = 1, 2, \ldots$. We show in the next two lemmas that the error in the first two moments, measured in a

suitable spatial norm, is bounded by the *mean square error* $\|u - u_{h,p}\|_{L^2_\rho \otimes H^1_0}$, which, due to Theorem 4.1, is exponentially convergent with respect to the polynomial degree p employed in the y directions. In particular, without extra regularity assumptions on the solution u of the problem, we have optimal convergence for the error in the mean value (first moment) measured in $L^2(D)$ or $H^1(D)$ and for the error in the second moment measured in $L^1(D)$.

LEMMA 4.7 (approximation of mean value).

$$\|E[u - u_{h,p}]\|_{V(D)} \le \|u - u_{h,p}\|_{L^2_\rho(\Gamma)\otimes V(D)}, \text{ with } V(D) = L^2(D) \text{ or } H^1(D).$$

The proof is immediate and omitted. Although the previous estimate implies exponential convergence with respect to p, under the assumptions of Theorem 4.1, the above estimate is suboptimal and can be improved by a duality argument (see [4] and Remark 5.2 in [5]).

LEMMA 4.8 (approximation of the second moment).

$$\|E[u^2 - u^2_{h,p}]\|_{L^1(D)} \le C\|u - u_{h,p}\|_{L^2_\rho(\Gamma)\otimes L^2(D)}$$

with C independent of the discretization parameters h and p.

Proof. We have

$$\begin{aligned}\|E[u^2 - u^2_{h,p}]\|_{L^1(D)} &\le \|E[(u - u_{h,p})(u + u_{h,p})]\|_{L^1(D)} \\ &\le \|u - u_{h,p}\|_{L^2_\rho(\Gamma)\otimes L^2(D)}\|u + u_{h,p}\|_{L^2_\rho(\Gamma)\otimes L^2(D)} \\ &\le \|u - u_{h,p}\|_{L^2_\rho(\Gamma)\otimes L^2(D)} \left(\|u\|_{L^2_\rho(\Gamma)\otimes L^2(D)} + \|u_{h,p}\|_{L^2_\rho(\Gamma)\otimes L^2(D)}\right).\end{aligned}$$

The term $\|u_{h,p}\|_{L^2_\rho \otimes L^2}$ can be bounded as

$$\|u_{h,p}\|_{L^2_\rho(\Gamma)\otimes L^2(D)} = \|\mathcal{I}_p u_h\|_{L^2_\rho(\Gamma)\otimes L^2(D)} \le C_1\|u_h\|_{C^0_\sigma(\Gamma;L^2(D))} \le C(f, a_{\min}),$$

where we have used the boundedness of the interpolation operator $\mathcal{I}_p$ stated in Lemma 4.2. The last inequality follows from the fact that the semidiscrete solution u_h is the orthogonal projection of the exact solution u onto the subspace H_h with respect to the *energy* inner product; hence

$$\|\sqrt{a(y)}\nabla u_h(y)\|_{L^2(D)} \le \|\sqrt{a(y)}\nabla u(y)\|_{L^2(D)} \quad \forall y \in \Gamma,$$

and the last term can be controlled in terms of $a_{\min}$ and the forcing term, f. ∎

Similarly, it is possible to estimate the approximation error in the covariance function of the solution u.

On the other hand, to estimate the convergence rate of the error in higher order moments, or of the second moment in higher norms, we need extra regularity assumptions on the solution to ensure proper integrability and then be able to use analyticity.

5. Numerical examples. This section illustrates the convergence of the collocation method for a stochastic elliptic problem in two dimensions. The computational results are in accordance with the convergence rate predicted by the theory.

The problem to solve is

$$\begin{aligned} -\nabla\cdot(a\nabla u) &= 0 \quad \text{on } \Omega \times D, \\ u &= 0 \quad \text{on } \Omega \times \partial D_D, \\ -a\partial_n u &= 1 \quad \text{on } \Omega \times \partial D_N, \\ \partial_n u &= 0 \quad \text{on } \Omega \times (\partial D - (\partial D_D \cup \partial D_N)), \end{aligned}$$

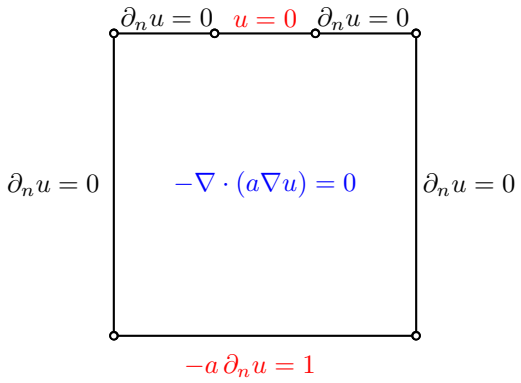


FIG. 1. *Geometry and boundary conditions for the numerical example.*

with

$$D = \{(x,z) \in \mathbb{R}^2 : -1.5 \le x \le 0,\ -0.4 \le z \le 0.8\},$$
$$\partial D_D = \{(x,z) \in \mathbb{R}^2 : -1 \le x \le -0.5,\ z = 0.8\},$$
$$\partial D_N = \{(x,z) \in \mathbb{R}^2 :\ -1.5 \le x \le 0,\ z = -0.4\};$$

cf. Figure 1.

The random diffusivity coefficient is a nonlinear function of the random vector Y, namely,

$$a(\omega, x) = a_{min} + \exp\Big\{ [Y_1(\omega)\cos(\pi z) + Y_3(\omega)\sin(\pi z)]\, e^{-\frac{1}{8}} + [Y_2(\omega)\cos(\pi x) + Y_4(\omega)\sin(\pi x)]\, e^{-\frac{1}{8}} \Big\}. \tag{5.1}$$

Here $a_{min} = 1/100$, and the real random variables Y_n, $n = 1, \ldots, 4$, are independent and identically distributed with mean value zero and unit variance. To illustrate the behavior of the collocation method with either unbounded or bounded random variables Y_n, this section presents two different cases, corresponding to either Gaussian or uniform densities. The corresponding collocation points are then Cartesian products determined by the roots of either Hermite or Legendre polynomials.

Observe that the collocation method requires only the solution of uncoupled deterministic problems in the collocation points, even in the presence of a diffusivity coefficient which depends nonlinearly on the random variables as in (5.1). This is a great advantage with respect to the classical stochastic Galerkin finite element method as considered in [4] or [30] (see also the considerations given in section 2.1). Observe, moreover, how easily the collocation method can deal with unbounded random variables.

Figure 2 shows some realizations of the logarithm of the diffusivity coefficient, while Figures 3 and 4 show the mean and variance of the corresponding solutions. The finite element space for spatial discretization is the span of continuous functions that are piecewise polynomials with degree five over a triangulation with 1178 triangles and 642 vertices; see Figure 5. This triangulation has been adaptively graded to control

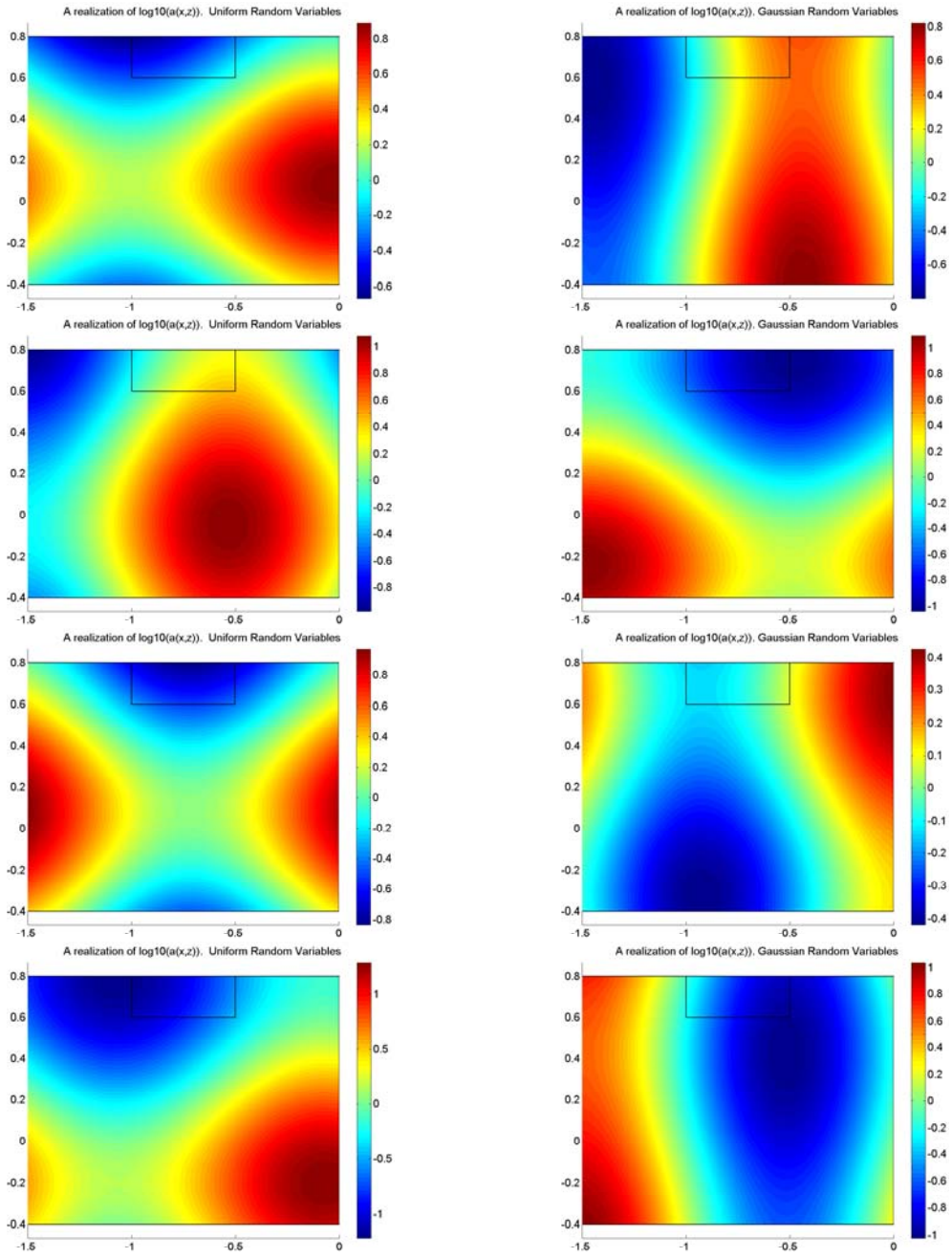


FIG. 2. *Some realizations of* $\log(a)$.

the singularities at the boundary points $(-1, 0.8)$ and $(-0.5, 0.8)$. These singularities occur where the Dirichlet and Neumann boundaries meet, and they essentially behave like $\sqrt{r}$, with r being the distance to the closest singularity point.

To study the convergence of the tensor product collocation method, we increase the order p for the approximating polynomial spaces, $\mathcal{P}_p(\Gamma)$, following the adaptive

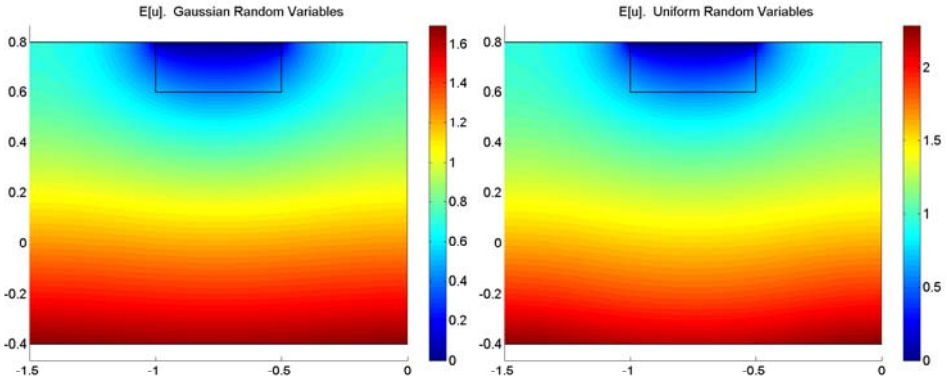


FIG. 3. *Results for the computation of the expected value for the solution, $E[u]$.*

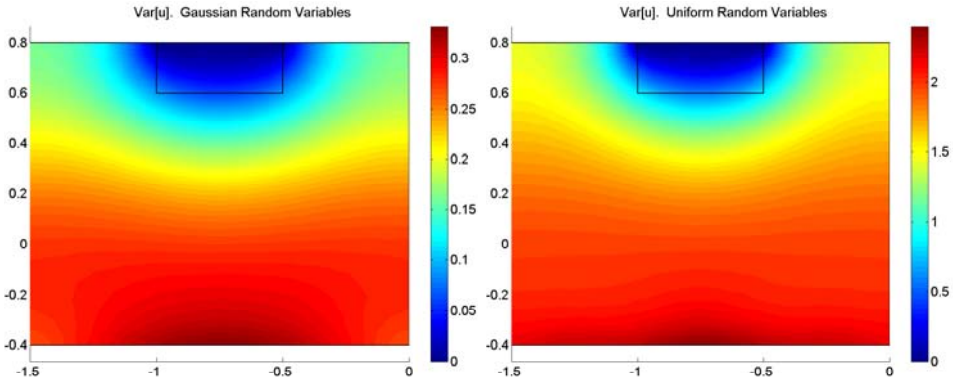


FIG. 4. *Results for the computation of the variance of the solution, $Var[u]$.*

algorithm described on page 1287 of [5]. This adaptive algorithm increases the tensor polynomial degree with an anisotropic strategy: it increases the order of approximation in one direction as much as possible before considering the next direction.

The computational results for the $H_0^1(D)$ approximation error in the expected value, $E[u]$, are shown on Figure 6, while those corresponding to the approximation of the second moment, $E[u^2]$, are shown on Figure 7. To estimate the computational error in the ith direction, corresponding to a multi-index $p = (p_1, \ldots, p_i, \ldots, p_N)$, we approximate it by $E[e] \approx E[u_{h,p} - u_{h,\tilde{p}}]$, with $\tilde{p} = (p_1, \ldots, p_i + 1, \ldots, p_N)$. We proceed similarly for the error in the approximation of the second moment.

As expected, the estimated approximation error decreases exponentially fast as the polynomial order increases, for both the computation of $E[u]$ and $E[u^2]$, with either Gaussian or uniform probability densities.

6. Conclusions. In this work we have proposed a collocation method for the solution of elliptic partial differential equations with random coefficients and forcing terms. This method has the advantages of leading to uncoupled deterministic problems also in the case of input data which depend nonlinearly on the random variables; treating efficiently the case of nonindependent random variables with the introduction of an auxiliary density $\hat{\rho}$; dealing easily with unbounded random variables, such as Gaussian or exponential ones; and dealing with no difficulty with a diffusivity

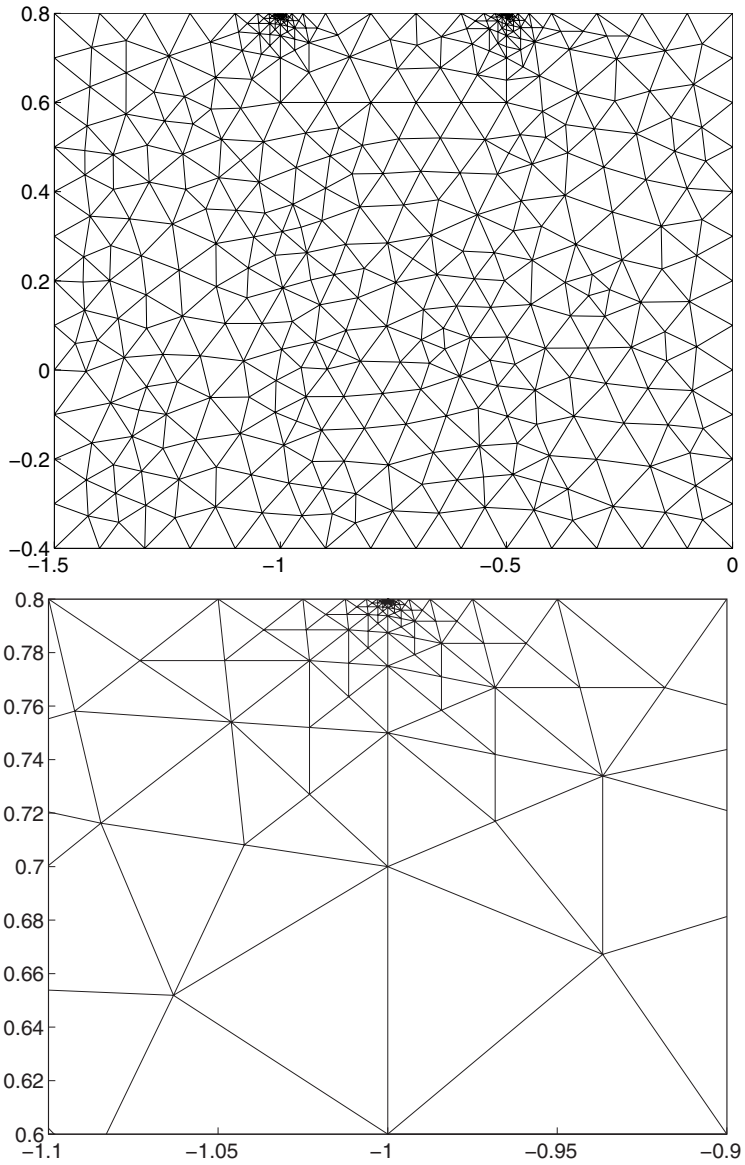


FIG. 5. *Top: Unstructured grid for the spatial discretization. The corresponding finite element spaces are the span of continuous functions that are piecewise polynomials with degree five. Bottom: Detail of the mesh refinement near the left singularity.*

coefficient a with unbounded second moment.

We have provided a full convergence analysis and proved exponential convergence "in probability" for a broad range of situations. The theoretical result is given in Theorem 4.1 and confirmed numerically by the tests presented in section 5.

The method is very versatile and very accurate for the class of problems considered (as accurate as the stochastic Galerkin approach). It leads to the solution of uncoupled deterministic problems and, as such, is fully parallelizable like a Monte Carlo method. The extension of the analysis to other classes of linear and nonlinear problems is the subject of ongoing research.

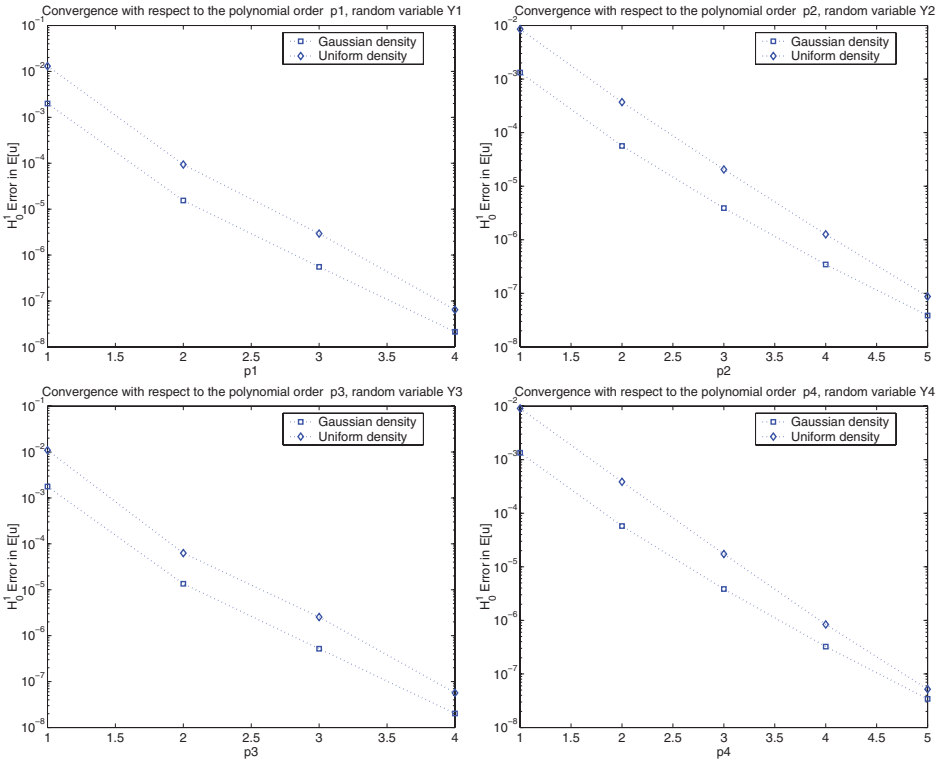


FIG. 6. *Convergence results for the approximation of the expected value, $E[u]$.*

The use of tensor product polynomials suffers from the *curse of dimensionality*. Hence, this method is efficient only for a small number of random variables. For a moderate or large dimensionality of the probability space, one should rather turn to *sparse tensor product spaces*. This aspect will be investigated in a future work.

Appendix.

LEMMA A.1. *Let $r \in \mathbb{R}^+$, $r < 1$. Then*

- $$\sum_{k=0}^{n}(2k+1)r^k = \frac{1}{(1-r)^2}\left\{1 + r - r^{n+1}\left[(2n+1)(1-r)+2\right]\right\}.$$
- $$\sum_{k=n+1}^{\infty}(2k+1)r^k = r^{n+1}\frac{(2n+1)(1-r)+2}{(1-r)^2}.$$

Proof. We use the summation-by-parts formula

$$\sum_{k=0}^{n} f_k g_k = f_n G_n - \sum_{k=0}^{n-1} G_k(f_{k+1} - f_k), \qquad G_k = \sum_{j=0}^{k} g_j$$

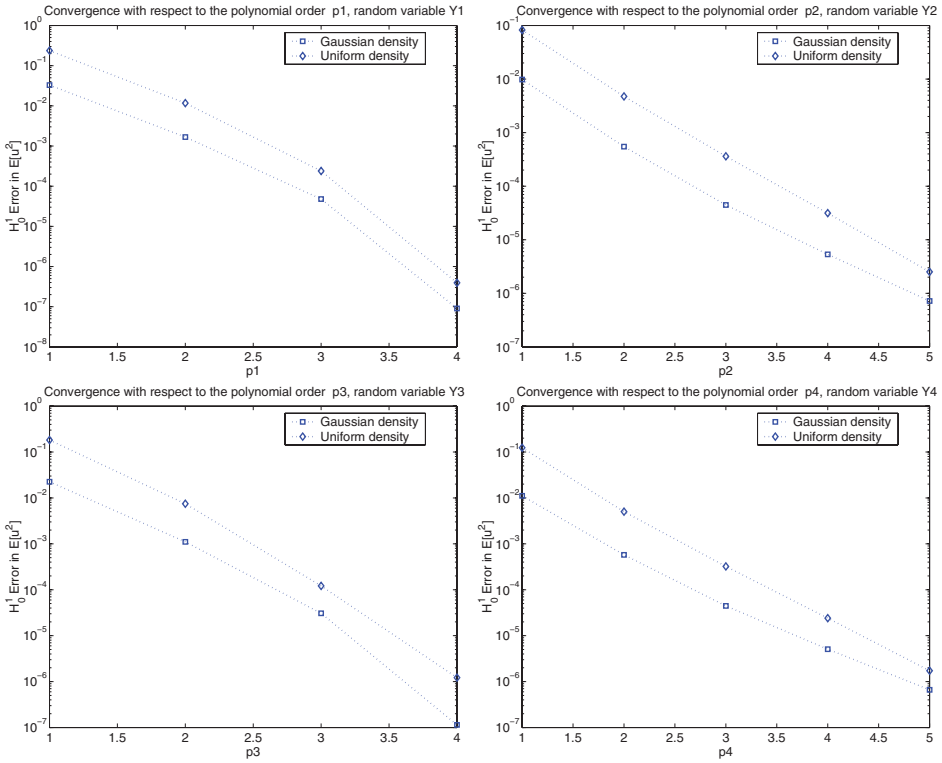


FIG. 7. *Convergence results for the approximation of the second moment,* $E[u^2]$.

with $f_k = (2k+1)$, $g_k = r^k$, and $G_k = (1-r^{k+1})/(1-r)$. Then

$$\begin{aligned}\sum_{k=0}^{n}(2k+1)r^k &= (2n+1)\frac{1-r^{n+1}}{1-r} - \sum_{k=0}^{n-1} 2\frac{1-r^{k+1}}{1-r} \\ &= (2n+1)\frac{1-r^{n+1}}{1-r} - \frac{2}{1-r}\left[n - r\frac{1-r^n}{1-r}\right] \\ &= \frac{1}{1-r}\left[(2n+1) - (2n+1)r^{n+1} - 2n + 2r\frac{1-r^n}{1-r}\right] \\ &= \frac{1}{1-r}\left\{1 + \frac{2r}{1-r} - r^{n+1}\left[(2n+1) + \frac{2}{1-r}\right]\right\},\end{aligned}$$

which gives the first result. Clearly,

$$\sum_{k=0}^{\infty}(2k+1)r^k = \frac{1+r}{(1-r)^2}.$$

Then, computing the tail series as

$$\sum_{k=n+1}^{\infty}(2k+1)r^k = \sum_{k=0}^{\infty}(2k+1)r^k - \sum_{k=0}^{n}(2k+1)r^k,$$

we easily obtain the second result as well. □

LEMMA A.2. *Let $r \in \mathbb{R}^+$, $r < 1$. Then*

$$\sum_{k=n+1}^{\infty} r^{\sqrt{2k+1}} \leq \left[\frac{2\sqrt{n+1}}{a(1-a)} + O(1)\right] a^{\sqrt{n}}, \qquad a = r^{\sqrt{2}}.$$

Proof. We start bounding

$$\sum_{k=n+1}^{\infty} r^{\sqrt{2k+1}} \leq \sum_{k=n+1}^{\infty} r^{\sqrt{2k}} = \sum_{k=n+1}^{\infty} a^{\sqrt{k}}.$$

Let us observe, now, that

$$\sum_{k=n+1}^{\infty} a^{\sqrt{k}} \leq \sum_{k=[\sqrt{n+1}]}^{\infty} (2k+1)a^k,$$

where we have denoted by $[v]$ the integer part of a real number v. Then, using the result from Lemma A.1, we have

$$\sum_{k=[\sqrt{n+1}]}^{\infty} (2k+1)a^k \leq a^{[\sqrt{n+1}]} \frac{(2[\sqrt{n+1}]-1)(1-a)+2}{(1-a)^2}.$$

Observing now that $\sqrt{n+1} - 1 \leq [\sqrt{n+1}] \leq \sqrt{n+1} + 1$, we obtain

$$\sum_{k=n+1}^{\infty} a^{\sqrt{k}} \leq a^{\sqrt{n+1}} \frac{(2\sqrt{n+1}+1)(1-a)+2}{a(1-a)^2},$$

which leads immediately to the final result. ∎

REFERENCES

[1] M. AINSWORTH AND J.-T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley, New York, 2000.

[2] I. BABUŠKA, K-M. LIU, AND R. TEMPONE, *Solving stochastic partial differential equations based on the experimental data*, Math. Models Methods Appl. Sci., 13 (2003), pp. 415–444.

[3] I. BABUŠKA AND T. STROUBOULIS, *The Finite Element Method and Its Reliability*, Numerical Mathematics and Scientific Computation, The Clarendon Press, Oxford University Press, New York, 2001.

[4] I. BABUŠKA, R. TEMPONE, AND G. E. ZOURARIS, *Galerkin finite element approximations of stochastic elliptic partial differential equations*, SIAM J. Numer. Anal., 42 (2004), pp. 800–825.

[5] I. BABUŠKA, R. TEMPONE, AND G. E. ZOURARIS, *Solving elliptic boundary value problems with uncertain coefficients by the finite element method: The stochastic formulation*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1251–1294.

[6] I. BABUŠKA, F. NOBILE, AND R. TEMPONE, *Worst-case scenario analysis for elliptic problems with uncertainty*, Numer. Math., 101 (2005), pp. 185–219.

[7] V. BARTHELMANN, E. NOVAK, AND K. RITTER, *High dimensional polynomial interpolation on sparse grids*, Adv. Comput. Math., 12 (2000), pp. 273–288.

[8] J. P. BOYD, *Asymptotic coefficients of Hermite function series*, J. Comput. Phys., 54 (1984), pp. 382–410.

[9] J. P. BOYD, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, Mineola, NY, 2001.

[10] M. BRAACK AND A. ERN, *A posteriori control of modeling errors and discretization errors*, Multiscale Model. Simul., 1 (2003), pp. 221–238.

[11] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[12] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, New York, 1978.
[13] M. K. Deb, I. Babuška, and J. T. Oden, *Solution of stochastic partial differential equations using Galerkin finite element techniques*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 6359–6372.
[14] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. 303, Springer-Verlag, Berlin, 1993.
[15] H. C. Elman, O. G. Ernst, D. P. O'Leary, and M. Stewart, *Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1037–1055.
[16] B. Engquist, P. Löstedt, and O. Runborg, eds., *Multiscale Methods in Science and Engineering*, Lect. Notes Comput. Sci. Eng. 44, Springer-Verlag, Berlin, 2005.
[17] P. Erdös and P. Turán, *On interpolation.* I. *Quadrature- and mean-convergence in the Lagrange-interpolation*, Ann. of Math. (2), 38 (1937), pp. 142–155.
[18] P. Frauenfelder, C. Schwab, and R. A. Todor, *Finite elements for elliptic problems with stochastic coefficients*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 205–228.
[19] D. Funaro and O. Kavian, *Approximation of some diffusion evolution equations in unbounded domains by Hermite functions*, Math. Comp., 57 (1991), pp. 597–619.
[20] R. Ghanem, *Ingredients for a general purpose stochastic finite elements implementation*, Comput. Methods Appl. Mech. Engrg., 168 (1999), pp. 19–34.
[21] R. G. Ghanem and P. D. Spanos, *Stochastic Finite Elements: A Spectral Approach*, Springer-Verlag, New York, 1991.
[22] M. Grigoriu, *Stochastic Calculus: Applications in Science and Engineering*, Birkhäuser Boston, Boston, 2002.
[23] E. Hille, *Contributions to the theory of Hermitian series.* II. *The representation problem*, Trans. Amer. Math. Soc., 47 (1940), pp. 80–94.
[24] J. Hlaváček, I. Chleboun, and I. Babuška, *Uncertain Input Data Problems and the Worst Scenario Method*, Elsevier, Amsterdam, 2004.
[25] S. Larsen, *Numerical Analysis of Elliptic Partial Differential Equations with Stochastic Input Data*, Ph.D. thesis, University of Maryland, College Park, MD, 1986.
[26] O. P. Le Maître, O. M. Knio, H. N. Najm, and R. G. Ghanem, *Uncertainty propagation using Wiener–Haar expansions*, J. Comput. Phys., 197 (2004), pp. 28–57.
[27] M. Loève, *Probability Theory.* I, 4th ed., Grad. Texts in Math. 45, Springer-Verlag, New York, 1977.
[28] M. Loève, *Probability Theory.* II, 4th ed., Grad. Texts in Math. 46, Springer-Verlag, New York, 1978.
[29] L. Mathelin, M. Y. Hussaini, and T. A. Zang, *Stochastic approaches to uncertainty quantification in CFD simulations*, Numer. Algorithms, 38 (2005), pp. 209–236.
[30] H. G. Matthies and A. Keese, *Galerkin methods for linear and nonlinear elliptic stochastic partial differential equations*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1295–1331.
[31] J. T. Oden and S. Prudhomme, *Estimation of modeling error in computational mechanics*, J. Comput. Phys., 182 (2002), pp. 496–515.
[32] J. T. Oden and K. S. Vemaganti, *Estimation of local modeling error and goal-oriented adaptive modeling of heterogeneous materials.* I. *Error estimates and adaptive algorithms*, J. Comput. Phys., 164 (2000), pp. 22–47.
[33] A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations*, Springer Ser. Comput. Math. 23, Springer-Verlag, Berlin, 1994.
[34] L. J. Roman and M. Sarkis, *Stochastic Galerkin method for elliptic SPDEs: A white noise approach*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 941–955.
[35] M. A. Tatang, *Direct Incorporation of Uncertainty in Chemical and Environmental Engineering Systems*, Ph.D. thesis, MIT, Cambridge, MA, 1995.
[36] J. V. Uspensky, *On the convergence of quadrature formulas related to an infinite interval*, Trans. Amer. Math. Soc., 30 (1928), pp. 542–559.
[37] R. Verfürth, *A Review of A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner, Wiley, Stuttgart, 1996.
[38] N. Wiener, *The homogeneous chaos*, Amer. J. Math., 60 (1938), pp. 897–936.
[39] C. L. Winter and D. M. Tartakovsky, *Groundwater flow in heterogeneous composite aquifers*, Water Resour. Res., 38 (2002), p. 23.1 (doi:10.1029/2001WR000450).

[40] D. Xiu and J. S. Hesthaven, *High-order collocation methods for differential equations with random inputs*, SIAM J. Sci. Comput., 27 (2005), pp. 1118–1139.

[41] D. Xiu and G. E. Karniadakis, *Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 4927–4948.

[42] D. Xiu and G. E. Karniadakis, *The Wiener–Askey polynomial chaos for stochastic differential equations*, SIAM J. Sci. Comput., 24 (2002), pp. 619–644.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1035–1044

CONVERGENCE ANALYSIS OF PERTURBED TWO-GRID AND MULTIGRID METHODS*

YVAN NOTAY†

Abstract. We consider multigrid methods for symmetric positive definite linear systems. We present a new algebraic convergence analysis of two-grid schemes with inexact solution of the coarse grid system. This analysis allows us to bound the convergence factor of such perturbed two-grid schemes, assuming only a certain bound on the convergence factor for the unperturbed scheme (with exact solution of the coarse grid system). Applied to multigrid methods with the standard W-cycle, this analysis shows that if the convergence factor of the (unperturbed) two-grid method is uniformly bounded by $\sigma < 1/2$, then the convergence factor of the multigrid method is uniformly bounded by $\sigma/(1-\sigma)$. The analysis is purely algebraic and requires only that pre- and postsmoothing are applied in a symmetric way. It covers both geometric and algebraic multigrid methods, and the coarse grid matrix may be of any type (not necessarily Galerkin).

Key words. multigrid, convergence analysis, linear systems, W-cycle, condition number

AMS subject classifications. 65F10, 65N55

DOI. 10.1137/060652312

1. Introduction.

We consider multigrid methods for solving symmetric positive definite (SPD) $n \times n$ linear systems

$$A\,\mathbf{u} \;=\; \mathbf{b}. \tag{1.1}$$

We focus on symmetric multigrid schemes, more precisely on methods for which the basic two-grid cycle is defined as follows:

- Relax ν times on $A\,\mathbf{u} = \mathbf{b}$ using a smoother R;
  we assume that R is an $n \times n$ matrix such that $\rho(I - R\,A) < 1$,
  where $\rho(\cdot)$ stands for the spectral radius;
  ν (the number of pre- and postsmoothing steps) is a given positive integer.
- Perform the coarse grid correction: $\mathbf{u} \leftarrow \mathbf{u} + p\,A_C^{-1} p^T(\mathbf{b} - A\,\mathbf{u})$;
  we assume that A_C (the coarse grid matrix) is an $n_c \times n_c$ SPD matrix,
  where $n_c \le n$ is the number of coarse variables;
  p (the prolongation) is an $n \times n_c$ matrix.
- Relax ν times on $A\,\mathbf{u} = \mathbf{b}$ using R^T.

Note that we do not assume any specific form for the coarse grid matrix.

In this paper, we analyze the influence of the perturbations that arise when the coarse grid systems are solved approximately. More precisely, we consider schemes as above in which A_C^{-1} is exchanged for some $n_c \times n_c$ SPD matrix K_C that approximates it. Our main result relates the convergence of this perturbed two-grid method with that of the "ideal" two-grid method that requires the inversion of A_C.

To express this relation, consider the iteration matrices that govern the convergence of the schemes described above:

$$T_{\mathrm{TG}} \;=\; (I - R^T A)^\nu\,(I - p\,A_C^{-1}\,p^T\,A)\,(I - R\,A)^\nu \tag{1.2}$$

*Received by the editors February 17, 2006; accepted for publication (in revised form) December 19, 2006; published electronically May 7, 2007. This work was supported by the Belgian FNRS (Maître de Recherches).

http://www.siam.org/journals/sinum/45-3/65231.html

†Service de Métrologie Nucléaire (C.P. 165/84), Université Libre de Bruxelles, 50 Av. F.D. Roosevelt, B-1050 Brussels, Belgium (ynotay@ulb.ac.be).

for the unperturbed two-grid cycle and

$$T_{\mathrm{PTG}} \;=\; (I - R^T A)^\nu \,(I - p\, K_C\, p^T\, A)\,(I - R\, A)^\nu \tag{1.3}$$

for the perturbed one; see, e.g., [19, p. 40]. As is well known, performing m iterations implies that

$$\hat{\mathbf{u}} - \mathbf{u}_{\mathrm{TG}}^{(m)} = \big(T_{\mathrm{TG}}\big)^m \big(\hat{\mathbf{u}} - \mathbf{u}_{\mathrm{TG}}^{(0)}\big) \qquad \text{or} \qquad \hat{\mathbf{u}} - \mathbf{u}_{\mathrm{PTG}}^{(m)} = \big(T_{\mathrm{PTG}}\big)^m \big(\hat{\mathbf{u}} - \mathbf{u}_{\mathrm{PTG}}^{(0)}\big)$$

(respectively), where $\hat{\mathbf{u}} = A^{-1}\mathbf{b}$ is the exact solution to (1.1). Hence, the asymptotic convergence rate is equal to the spectral radius of the iteration matrix, which is referred to as the *convergence factor*. On the other hand, both the perturbed and the unperturbed two-grid cycle implicitly define a preconditioner, which we denote B_{TG} and B_{PTG}, respectively. They are related to the iteration matrices by

$$I - B_{\mathrm{TG}}^{-1}\, A \;=\; T_{\mathrm{TG}} \qquad \text{and} \qquad I - B_{\mathrm{PTG}}^{-1}\, A \;=\; T_{\mathrm{PTG}}. \tag{1.4}$$

Note that because A, B_{TG}, and B_{PTG} are SPD (see below), the eigenvalues of $B_{\mathrm{TG}}^{-1}A$ and $B_{\mathrm{PTG}}^{-1}A$ are real and the convergence factors satisfy

$$\begin{aligned}
\rho\,(T_{\mathrm{TG}}) &= \max\big(\lambda_{\max}(B_{\mathrm{TG}}^{-1}A) - 1,\; 1 - \lambda_{\min}(B_{\mathrm{TG}}^{-1}A)\big),\\
\rho\,(T_{\mathrm{PTG}}) &= \max\big(\lambda_{\max}(B_{\mathrm{PTG}}^{-1}A) - 1,\; 1 - \lambda_{\min}(B_{\mathrm{PTG}}^{-1}A)\big),
\end{aligned}$$

where $\lambda_{\max}(\cdot)$ and $\lambda_{\min}(\cdot)$ stand for the largest and the smallest eigenvalue, respectively. In section 2, we prove

$$\lambda_{\max}(B_{\mathrm{PTG}}^{-1}\, A) \le \lambda_{\max}(B_{\mathrm{TG}}^{-1}\, A) \;\cdot\; \max(\lambda_{\max}(K_C\, A_C),\, 1), \tag{1.5}$$

$$\lambda_{\min}(B_{\mathrm{PTG}}^{-1}\, A) \ge \lambda_{\min}(B_{\mathrm{TG}}^{-1}\, A) \;\cdot\; \min(\lambda_{\min}(K_C\, A_C),\, 1). \tag{1.6}$$

Multigrid cycles are obtained when a two-grid method is used recursively, exchanging the solution of the coarse grid system for a given number γ of two-grid cycles on the coarser level, and so on, until the coarsest level on which an exact solve is performed. Multigrid cycles are thus particular cases of perturbed two-grid cycles, and the above results enable us to analyze them. For the so-called W-cycle (which corresponds to $\gamma = 2$), we show in section 3 that if $\sigma < 1/2$ is a uniform (i.e., holding at every level) bound on the convergence factor of the unperturbed two-grid method, then the convergence factor of the multigrid method is bounded by $\sigma/(1-\sigma)$.

This improves the state of the art in multigrid convergence theory. Analyzing the two-grid convergence factor is often sufficient to assess the convergence of a multigrid scheme; see, e.g., [19, p. 77]. However, this is not yet completely supported by theoretical results. The standard algebraic analysis considers the multigrid iteration matrix as a two-grid iteration matrix plus some perturbation term; see [11, sect. 4.2] or [19, Thm. 3.2.1]. It allows us to obtain a useful bound on the convergence factor of the multigrid method with the W-cycle if σ satisfies

$$\sigma \;\le\; \frac{1}{4\,C}, \tag{1.7}$$

where C is a constant whose exact value is difficult to predict, except that it is in general not smaller than 1 (see section 3 below for details). Because of this condition, this result on multigrid convergence is sometimes stated as follows: "if the two-grid method converges *sufficiently well*, then the multigrid method with W-cycle will have

similar convergence properties" [19, p. 77]. However, condition (1.7) may be violated when textbook multigrid efficiency is difficult to achieve. It may also be difficult to check when using an algebraic multigrid (AMG) method. Below we also show that, when both our new analysis and the standard algebraic analysis apply, our bound on the convergence factor is generally sharper.

For the case of Galerkin coarse grid matrices (i.e., assuming $A_C = p^T A\, p$), an interesting analysis of the W-cycle multigrid has been developed by Braess in [6, pp. 226–228]. This analysis is based on two-grid schemes without postsmoothing, which also gives a worst case estimate for the general case. As will be seen in section 3, our bound on the convergence factor for the W-cycle is always equal to the square of Braess's bound. This suggests that, as proved for the two-grid case in [15, eq. (41)], the W-cycle multigrid scheme with symmetrized pre- and postsmoothing can converge twice as fast as the corresponding scheme without postsmoothing.

On the other hand, in the SPD case, it is also possible to prove optimal convergence properties (with respect to the number of levels) of multigrid methods via so-called smoothing and approximation properties or via the theory of subspace correction methods (using the multilevel splitting of finite element spaces); see, e.g., [5, 7, 11, 12, 13, 14, 16, 17, 22, 23]. However, bounds derived in this way do not, in general, give satisfactorily sharp predictions of actual multigrid convergence [19, p. 96]. Moreover, they require assumptions that are more restrictive than just the convergence of the two-grid method. To our knowledge, for instance, these assumptions have not yet been checked for AMG methods. These analyses, nevertheless, play an important role in the multigrid convergence theory, complementary to the algebraic approach developed here. Indeed, they cover the V-cycle, for which both the standard analysis mentioned above and our new analysis fail to deliver bounds independent of the number of levels.

Eventually, it should be noted that our bounds (1.5), (1.6) have a wider scope than just the analysis of standard multigrid cycles. First, these relations are similar to the ones holding for AMLI-type methods [2, 3]. Hence, when the two-grid method does not converge fast enough for the standard W-cycle, it is possible to use polynomially accelerated cycles based on Chebyshev polynomials, as considered in these references [21]. Another potential application lies in the simplification of coarse grid matrices: (1.5), (1.6) indeed show that one may replace a given coarse grid matrix by a spectrally equivalent approximation. For instance, the theory of algebraic two-grid methods heavily relies on the use of Galerkin coarse grid matrices, that is, on the assumption that $A_C = p^T A\, p$ [8, 9, 10, 15, 18]. However, in practice, such matrices may be costly to compute, and so it could be interesting to develop cheaper alternatives.

The remainder of this paper is organized as follows. In section 2, we prove the main inequalities (1.5), (1.6). Their application to multigrid cycles is discussed in section 3.

2. Perturbed two-grid methods. We first show that B_{TG} is the Schur complement of an extended matrix $\widehat{B}_{\mathrm{TG}}$ given in factored form. Note that this holds for any SPD coarse grid matrix A_C; hence, similarly, B_{PTG} is the Schur complement of an extended matrix $\widehat{B}_{\mathrm{PTG}}$. This factored form has been inspired by the factored form existing for the preconditioner defined by the so-called hierarchical basis multigrid method[1] [4]. Our derivation is also related to equation (15) in [10].

[1] This latter method does not fit into our framework because only fine grid unknowns are relaxed during smoothing steps, and hence R is singular.

Let us introduce some notation. Let M be the matrix such that

$$I - M^{-1}A \;=\; (I - R\,A)^{\nu}; \tag{2.1}$$

from the assumption, $\rho(I - R\,A) < 1$, and M exists, is invertible, and is such that $\rho(I - M^{-1}A) < 1$. This latter relation implies that

$$Q \;= M^{-1} + M^{-T} - M^{-T}\,A\,M^{-1} \;=\; M^{-T}\left(M + M^T - A\right)\,M^{-1} \tag{2.2}$$

is positive definite; see [9, 15].

Define

$$\widehat{B}_{\mathrm{TG}} \;=\; \begin{pmatrix} I_{n\times n} & 0 \\ -p^T(I - A\,M^{-1}) & I_{n_c\times n_c}\end{pmatrix}\begin{pmatrix} Q^{-1} & 0 \\ 0 & A_C\end{pmatrix}\begin{pmatrix} I_{n\times n} & -(I - M^{-T}A)p \\ 0 & I_{n_c\times n_c}\end{pmatrix}. \tag{2.3}$$

Straightforward calculation shows that

$$\begin{aligned}\widehat{B}_{\mathrm{TG}}{}^{-1} &= \begin{pmatrix} I_{n\times n} & (I - M^{-T}A)p \\ 0 & I_{n_c\times n_c}\end{pmatrix}\begin{pmatrix} Q & 0 \\ 0 & A_C^{-1}\end{pmatrix}\begin{pmatrix} I_{n\times n} & 0 \\ p^T(I - A\,M^{-1}) & I_{n_c\times n_c}\end{pmatrix} \\ &= \begin{pmatrix} Q + (I - M^{-T}A)p\,A_C^{-1}\,p^T(I - A\,M^{-1}) & (I - M^{-T}A)p\,A_C^{-1} \\ A_C^{-1}\,p^T(I - A\,M^{-1}) & A_C^{-1}\end{pmatrix} \\ &= \begin{pmatrix} B_{\mathrm{TG}}^{-1} & (I - M^{-T}A)p\,A_C^{-1} \\ A_C^{-1}\,p^T(I - A\,M^{-1}) & A_C^{-1}\end{pmatrix}.\end{aligned}$$

Because $\widehat{B}_{\mathrm{TG}}$ is SPD, this first shows that B_{TG} is SPD too. Further, the inverse of B_{TG} is a principal submatrix of the inverse of $\widehat{B}_{\mathrm{TG}}$ if and only if B_{TG} is equal to the corresponding Schur complement in $\widehat{B}_{\mathrm{TG}}$; see, e.g., [1, eq. (3.4), p. 93]. That is, considering the 2×2 block form

$$\widehat{B}_{\mathrm{TG}} \;=\; \begin{pmatrix} (\widehat{B}_{\mathrm{TG}})_{FF} & (\widehat{B}_{\mathrm{TG}})_{FC} \\ (\widehat{B}_{\mathrm{TG}})_{CF} & (\widehat{B}_{\mathrm{TG}})_{CC}\end{pmatrix}$$

(where $(\widehat{B}_{\mathrm{TG}})_{FF}$ is $n\times n$ and $(\widehat{B}_{\mathrm{TG}})_{CC}$ is $n_c\times n_c$), B_{TG} is the Schur complement of $\widehat{B}_{\mathrm{TG}}$ with respect to the bottom right block:

$$B_{\mathrm{TG}} \;=\; (\widehat{B}_{\mathrm{TG}})_{FF} - (\widehat{B}_{\mathrm{TG}})_{FC}\,(\widehat{B}_{\mathrm{TG}})_{CC}^{-1}\,(\widehat{B}_{\mathrm{TG}})_{CF}.$$

This, together with Theorem 3.8 in [1], proves the following lemma.

LEMMA 2.1. *Let B_{TG} be defined by* (1.2), (1.4) *with A, R, p, A_C satisfying the assumptions stated in section* 1. *Let $\widehat{B}_{TG}$ be defined by* (2.3) *with M defined by* (2.1) *and Q defined by* (2.2). *B_{TG} is SPD, and one has, for all $\mathbf{z}\in\Re^n$,*

$$\mathbf{z}^T\,B_{TG}\,\mathbf{z} \;=\; \min_{\mathbf{w}_C\in\Re^{n_c}}\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{TG}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}. \tag{2.4}$$

Moreover,

$$\lambda_{\max}(B_{TG}^{-1}A) = \max_{\mathbf{z}\in\Re^n\setminus\{0\}}\;\max_{\mathbf{w}_C\in\Re^{n_c}}\;\frac{\mathbf{z}^T\,A\,\mathbf{z}}{\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{TG}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}}, \tag{2.5}$$

$$\lambda_{\min}(B_{TG}^{-1}A) = \min_{\mathbf{z}\in\Re^n\setminus\{0\}}\;\max_{\mathbf{w}_C\in\Re^{n_c}}\;\frac{\mathbf{z}^T\,A\,\mathbf{z}}{\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{TG}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}}. \tag{2.6}$$

We now prove (1.5), (1.6). Note that these inequalities may also be proved from the factored form of two-grid preconditioners as obtained by Vassilevski in [20, 21]. Their use is also implicit in the building of AMLI-cycle multigrid developed independently by the same author [21].

THEOREM 2.2. *Let B_{TG}, B_{PTG} be defined by* (1.2), (1.3), (1.4) *with A, R, p, A_C, K_C satisfying the assumptions stated in section* 1. *Inequalities* (1.5) *and* (1.6) *hold.*

Proof. Let M, Q, $\widehat{B}_{\mathrm{TG}}$ be defined by (2.1), (2.2), (2.3), and define $\widehat{B}_{\mathrm{PTG}}$ similarly to $\widehat{B}_{\mathrm{TG}}$, exchanging A_C for K_C^{-1} in (2.3). Lemma 2.1 yields

$$\begin{aligned}
\lambda_{\max}(B_{\mathrm{PTG}}^{-1}A) &= \max_{\mathbf{z}\in\Re^n\setminus\{0\}} \max_{\mathbf{w}_C\in\Re^{n_c}} \frac{\mathbf{z}^T A\,\mathbf{z}}{\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{\mathrm{PTG}}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}} \\
&\le \max_{\mathbf{z}\in\Re^n\setminus\{0\}} \max_{\mathbf{w}_C\in\Re^{n_c}} \frac{\mathbf{z}^T A\,\mathbf{z}}{\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{\mathrm{TG}}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}} \cdot \max_{\widehat{\mathbf{z}}\in\Re^{n+n_c}\setminus\{0\}} \frac{\widehat{\mathbf{z}}^T\widehat{B}_{\mathrm{TG}}\,\widehat{\mathbf{z}}}{\widehat{\mathbf{z}}^T\widehat{B}_{\mathrm{PTG}}\,\widehat{\mathbf{z}}} \\
&= \lambda_{\max}(B_{\mathrm{TG}}^{-1}A) \cdot \max_{\widehat{\mathbf{w}}\in\Re^{n+n_c}\setminus\{0\}} \frac{\widehat{\mathbf{w}}^T\begin{pmatrix}Q^{-1} & 0\\ 0 & A_C\end{pmatrix}\widehat{\mathbf{w}}}{\widehat{\mathbf{w}}^T\begin{pmatrix}Q^{-1} & 0\\ 0 & K_C^{-1}\end{pmatrix}\widehat{\mathbf{w}}} \\
&= \lambda_{\max}(B_{\mathrm{TG}}^{-1}A) \cdot \max(\lambda_{\max}(K_C\,A_C),\,1).
\end{aligned}$$

Similarly, one finds

$$\begin{aligned}
\lambda_{\min}(B_{\mathrm{PTG}}^{-1}A) &= \min_{\mathbf{z}\in\Re^n\setminus\{0\}} \max_{\mathbf{w}_C\in\Re^{n_c}} \frac{\mathbf{z}^T A\,\mathbf{z}}{\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{\mathrm{PTG}}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}} \\
&\ge \min_{\mathbf{z}\in\Re^n\setminus\{0\}} \max_{\mathbf{w}_C\in\Re^{n_c}} \frac{\mathbf{z}^T A\,\mathbf{z}}{\begin{pmatrix}\mathbf{z}^T & \mathbf{w}_C^T\end{pmatrix}\widehat{B}_{\mathrm{TG}}\begin{pmatrix}\mathbf{z}\\ \mathbf{w}_C\end{pmatrix}} \cdot \min_{\widehat{\mathbf{z}}\in\Re^{n+n_c}\setminus\{0\}} \frac{\widehat{\mathbf{z}}^T\widehat{B}_{\mathrm{TG}}\,\widehat{\mathbf{z}}}{\widehat{\mathbf{z}}^T\widehat{B}_{\mathrm{PTG}}\,\widehat{\mathbf{z}}} \\
&= \lambda_{\min}(B_{\mathrm{TG}}^{-1}A) \cdot \min_{\widehat{\mathbf{w}}\in\Re^{n+n_c}\setminus\{0\}} \frac{\widehat{\mathbf{w}}^T\begin{pmatrix}Q^{-1} & 0\\ 0 & A_C\end{pmatrix}\widehat{\mathbf{w}}}{\widehat{\mathbf{w}}^T\begin{pmatrix}Q^{-1} & 0\\ 0 & K_C^{-1}\end{pmatrix}\widehat{\mathbf{w}}} \\
&= \lambda_{\min}(B_{\mathrm{TG}}^{-1}A) \cdot \min(\lambda_{\min}(K_C\,A_C),\,1). \qquad \square
\end{aligned}$$

3. Multigrid cycles. Multigrid methods are recursively defined. In the SPD case considered here, the iteration matrix $T_{\mathrm{MG}}^{(\ell)}$ at level ℓ depends on the iteration matrix $T_{\mathrm{MG}}^{(\ell-1)}$ at level $\ell-1$ (the next coarser level) according to

$$T_{\mathrm{MG}}^{(\ell)} = \left(I - R_\ell^T A_\ell\right)^{\nu_\ell}\left(I - p_\ell\left(I - \left(T_{\mathrm{MG}}^{(\ell-1)}\right)^\gamma\right)A_{\ell-1}^{-1}\,p_\ell^T A_\ell\right)\left(I - R_\ell\,A_\ell\right)^{\nu_\ell} \tag{3.1}$$

(see, e.g., [19, pp. 48–49]). In this equation, γ is the *cycle index*; $\gamma = 1$ corresponds to the V-cycle and $\gamma = 2$ to the W-cycle; larger values of γ are seldom considered in practice.

Now $T_{\rm MG}^{(\ell)}$ is a perturbed two-grid iteration matrix (1.3) with $A = A_\ell$, $R = R_\ell$, $\nu = \nu_\ell$, $p = p_\ell$ and K_C given by

$$K_C \;=\; \left(I - \left(T_{\rm MG}^{(\ell-1)}\right)^\gamma\right) A_{\ell-1}^{-1}.$$

Defining the iteration matrix $T_{\rm TG}^{(\ell)}$ of the (unperturbed) two-grid method by (1.2) with $A_C = A_{\ell-1}$ and letting $B_{\rm TG}^{(\ell)}$, $B_{\rm MG}^{(\ell)}$ be such that

$$I - {B_{\rm TG}^{(\ell)}}^{-1} A_\ell \;=\; T_{\rm TG}^{(\ell)}, \qquad I - {B_{\rm MG}^{(\ell)}}^{-1} A_\ell \;=\; T_{\rm MG}^{(\ell)},$$

inequalities (1.5), (1.6) imply

$$\begin{aligned}
\lambda_{\max}\left({B_{\rm MG}^{(\ell)}}^{-1} A_\ell\right) &\le \lambda_{\max}\left({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\right) \;\cdot\; \max\left(\lambda_{\max}\big(I - \left(T_{\rm MG}^{(\ell-1)}\right)^\gamma\big),\, 1\right),\\
\lambda_{\min}\left({B_{\rm MG}^{(\ell)}}^{-1} A_\ell\right) &\ge \lambda_{\min}\left({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\right) \;\cdot\; \min\left(\lambda_{\min}\big(I - \left(T_{\rm MG}^{(\ell-1)}\right)^\gamma\big),\, 1\right).
\end{aligned}$$

Let

$$\sigma_{\rm MG}^{(\ell)} \;=\; \rho\left(T_{\rm MG}^{(\ell)}\right) \;=\; \max\left(\lambda_{\max}\big({B_{\rm MG}^{(\ell)}}^{-1} A_\ell\big) - 1,\; 1 - \lambda_{\min}\big({B_{\rm MG}^{(\ell)}}^{-1} A_\ell\big)\right)$$

be the convergence factor of the multigrid method at level ℓ. One has, assuming $\sigma_{\rm MG}^{(\ell-1)} \le 1$,

$$\lambda_{\max}\left({B_{\rm MG}^{(\ell)}}^{-1} A_\ell\right) - 1 \;\le\; \begin{cases} \lambda_{\max}\left({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\right)\left(1 + \left(\sigma_{\rm MG}^{(\ell-1)}\right)^\gamma\right) - 1 & \text{if } \gamma \text{ is odd and } T_{\rm MG}^{(\ell-1)} \text{ has some negative eig.,} \\ \lambda_{\max}\left({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\right) - 1 & \text{otherwise,} \end{cases}$$

whereas

$$1 - \lambda_{\min}\left({B_{\rm MG}^{(\ell)}}^{-1} A_\ell\right) \;\le\; 1 - \lambda_{\min}\left({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\right)\left(1 - \left(\sigma_{\rm MG}^{(\ell-1)}\right)^\gamma\right).$$

Then let

$$\sigma_{\rm TG}^{(\ell)} \;=\; \rho\left(T_{\rm TG}^{(\ell)}\right) \;=\; \max\left(\lambda_{\max}\big({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\big) - 1,\; 1 - \lambda_{\min}\big({B_{\rm TG}^{(\ell)}}^{-1} A_\ell\big)\right)$$

be the convergence factor of the (unperturbed) two-grid method at level ℓ. If $\sigma_{\rm TG}^{(\ell)} \le 1$ and if either $\lambda_{\max}({B_{\rm TG}^{(\ell)}}^{-1} A_\ell) \le 1$ for all ℓ (as occurs when using Galerkin coarse grid matrices) or γ is even (or both), there holds

$$\sigma_{\rm MG}^{(\ell)} \;\le\; 1 - \left(1 - \sigma_{\rm TG}^{(\ell)}\right)\left(1 - \left(\sigma_{\rm MG}^{(\ell-1)}\right)^\gamma\right) \;\le\; 1. \tag{3.2}$$

If the matrix A_0 on the coarsest level is inverted exactly, one has $\sigma_{\rm MG}^{(1)} = \sigma_{\rm TG}^{(1)}$, and (3.2) defines a recursion which may be followed from $\ell = 2, 3, \ldots$ until the finest level.

For $\gamma = 1$ (the V-cycle), this does not yield bounds independent of the number of levels, although the estimates may be practically relevant for few levels if the $\sigma_{\rm TG}^{(\ell)}$ are small. For instance, $\sigma_{\rm TG}^{(\ell)} \le 0.1$ yields $\sigma_{\rm MG}^{(\ell)} \le 0.41$ for $\ell = 5$, which is not that bad.

The most interesting application is $\gamma = 2$ (the W-cycle). No additional assumption on $\lambda_{\max}(B_{\mathrm{TG}}^{(\ell)\,-1} A)$ is needed (γ is even), and one may check that if

$$\sigma_{\mathrm{TG}}^{(\ell)} \;\leq\; \sigma$$

and

$$\sigma_{\mathrm{MG}}^{(\ell-1)} \;\leq\; \frac{\sigma}{1-\sigma}$$

hold for some $\sigma < 1/2$, then

$$\sigma_{\mathrm{MG}}^{(\ell)} \;\leq\; \frac{\sigma}{1-\sigma}.$$

This proves the following theorem.

THEOREM 3.1. *Consider a multigrid method recursively defined by the iteration matrix* (3.1) *with* $\gamma = 2$ *for* $\ell = 1, 2, \ldots$ *and* $T_{MG}^{(0)} = 0$ *(exact inversion on the coarsest level). Assume that* A_ℓ, $\ell = 0, 1, \ldots$, *is SPD and that* R_ℓ, $\ell = 1, 2, \ldots$, *is such that* $\rho(I - R_\ell A_\ell) < 1$. *If the spectral radius of the two-grid iteration matrix* (1.2) *with* $A = A_\ell$, $R = R_\ell$, $\nu = \nu_\ell$, $p = p_\ell$, *and* $A_C = A_{\ell-l}$ *is bounded by some* $\sigma < 1/2$ *independently of* ℓ, *then the spectral radius of the multigrid iteration matrix* (3.1) *is bounded by* $\sigma/(1-\sigma)$, *independently of* ℓ.

Comparison with the standard algebraic analysis. This analysis, see, e.g., [11, 19], is based on matrix norms, instead of spectral radii, and applies to the nonsymmetric case as well. In the framework considered here, this analysis proves that if

$$\|T_{\mathrm{TG}}^{(\ell)}\| \;\leq\; \sigma^*$$

and

$$\left\| \left(I - R_\ell^T A_\ell\right)^{\nu_\ell} p_\ell \right\| \;\cdot\; \left\| A_{\ell-1}^{-1} p_\ell^T A_\ell \left(I - R_\ell A_\ell\right)^{\nu_\ell} \right\| \;\leq\; C$$

hold for some σ^*, C such that

$$4\, C\, \sigma^* \;\leq\; 1, \tag{3.3}$$

then the iteration matrix for the W-cycle ($\gamma = 2$) satisfies

$$\|T_{\mathrm{TG}}^{(\ell)}\| \;\leq\; \frac{1 - \sqrt{1 - 4\, C\, \sigma^*}}{2\, C} \;\leq\; 2\, \sigma^*. \tag{3.4}$$

Note that this result holds for any matrix norm $\|\cdot\|$.

To comment on it, first observe that

$$\begin{aligned}
&\left\| \left(I - R_\ell^T A_\ell\right)^{\nu_\ell} p_\ell \right\| \;\cdot\; \left\| A_{\ell-1}^{-1} p_\ell^T A_\ell \left(I - R_\ell A_\ell\right)^{\nu_\ell} \right\| \\
&\qquad \geq \left\| \left(I - R_\ell^T A_\ell\right)^{\nu_\ell} p_\ell A_{\ell-1}^{-1} p_\ell^T A_\ell \left(I - R_\ell A_\ell\right)^{\nu_\ell} \right\| \qquad (3.5) \\
&\qquad = \left\| \left(I - R_\ell^T A_\ell\right)^{\nu_\ell} \left(I - R_\ell A_\ell\right)^{\nu_\ell} - T_{\mathrm{TG}}^{(\ell)} \right\| \\
&\qquad \geq \left\| \left(I - R_\ell^T A_\ell\right)^{\nu_\ell} \left(I - R_\ell A_\ell\right)^{\nu_\ell} \right\| - \left\| T_{\mathrm{TG}}^{(\ell)} \right\| \\
&\qquad \geq \left(\rho\left(I - R_\ell A_\ell\right)\right)^{2\,\nu_\ell} - \sigma^*. \qquad (3.6)
\end{aligned}$$

In practical situations, $\rho\,(I - R_\ell\, A_\ell) \approx 1$ (otherwise, the coarse grid correction would be unnecessary for fast convergence). Hence, $C \gtrsim 1-\sigma^*$. Moreover, if $A_{\ell-1} = p_\ell^T A_\ell\, p_\ell$ (i.e., with a Galerkin coarse grid matrix), the middle term in the right-hand side of (3.5) is a projector that leaves the vectors in the range of p_ℓ unchanged, leading to expect $C \gtrsim 1$ in such cases. Since $\sigma^* \geq \sigma$ (the spectral radius is a lower bound on the matrix norm for any norm), the condition (3.3) is thus generally more restrictive than our condition $\sigma < 1/2$, even if one uses the energy norm for which $\sigma^* = \sigma$. For instance, $C \gtrsim 1$ then means that (3.3) requires $\sigma \leq \sigma^* \lesssim 1/4$.

When both bounds apply, if, in addition,

$$C \;\geq\; 1 - \sigma^*$$

(as one expects according to (3.6) and the discussion above), then our bound $\sigma/(1-\sigma)$ is always better than the bound (3.4). Indeed, taking $\sigma^* = \sigma$ (which is the most favorable for (3.4)), one has (since $\sigma(1+2\,C) = \sigma + \frac{1}{2} 4\,C\,\sigma < 1$)

$$\begin{aligned}
\frac{\sigma}{1-\sigma} \;\leq\; \frac{1-\sqrt{1-4\,C\,\sigma}}{2\,C} \;&\Longleftrightarrow\; \sqrt{1-4\,C\,\sigma^*} \;\leq\; \frac{1-\sigma(1+2\,C)}{1-\sigma} \\
&\Longleftrightarrow\; (1-\sigma)^2(1-4\,C\,\sigma) - (1-\sigma(1+2\,C))^2 \;\leq\; 0 \\
&\Longleftrightarrow\; -4\,C\,\sigma^3 + (4\,C - 4\,C^2)\sigma^2 \;\leq\; 0 \\
&\Longleftrightarrow\; -\sigma + 1 - C \;\leq\; 0.
\end{aligned}$$

By way of illustration, consider the case $\sigma = \sigma^* = 1/4$ and $C = 1$. Then our bound for the W-cycle is $1/3$, whereas (3.4) gives $1/2$. Note, however, that, for σ going to 0, both bounds converge to σ.

Comparison with Braess's analysis. Braess's analysis [6, pp. 226–228] assumes Galerkin coarse grid matrices and is based on two-grid and multigrid schemes without postsmoothing. To make things clear, let

$$\begin{aligned}
\widetilde{T}_{\rm TG}^{(\ell)} &= \left(I - p_\ell\, A_{\ell-1}^{-1}\, p_\ell^T\, A_\ell\right)\, (I - R_\ell\, A_\ell)^{\nu_\ell}\,, \\
\widetilde{T}_{\rm MG}^{(\ell)} &= \left(I - p_\ell \left(I - \left(\widetilde{T}_{\rm MG}^{(\ell-1)}\right)^{\gamma}\right) A_{\ell-1}^{-1}\, p_\ell^T\, A_\ell\right)\, (I - R_\ell\, A_\ell)^{\nu_\ell}
\end{aligned}$$

be the corresponding iteration matrices, and denote by $\widetilde{\rho}_{\rm TG}^{(\ell)}$, $\widetilde{\rho}_{\rm MG}^{(\ell)}$ their energy norm:

$$\widetilde{\rho}_{\rm TG}^{(\ell)} \;=\; \|\widetilde{T}_{\rm TG}^{(\ell)}\|_{A_\ell}, \qquad \widetilde{\rho}_{\rm MG}^{(\ell)} \;=\; \|\widetilde{T}_{\rm MG}^{(\ell)}\|_{A_\ell}.$$

The main convergence result for multigrid cycles in [6] is inequality (3.9) from Chapter V. With the above notation, this inequality amounts to

$$\left(\widetilde{\rho}_{\rm MG}^{(\ell)}\right)^2 \;\leq\; 1 \;-\; \left(1 - \left(\widetilde{\rho}_{\rm TG}^{(\ell)}\right)^2\right) \left(1 - \left(\widetilde{\rho}_{\rm MG}^{(\ell-1)}\right)^{2\gamma}\right). \tag{3.7}$$

Comparing with (3.2), the requirement on $\widetilde{\rho}_{\rm TG}^{(\ell)}$ to have an optimal method is less restrictive than the requirement we have on $\sigma_{\rm TG}^{(\ell)}$. However, as seen in [15, eq. (41)], when $A_{\ell-1} = p_\ell^T A_\ell\, p_\ell$ (as needed to prove (3.7)), there holds

$$A_\ell\, T_{\rm TG}^{(\ell)} \;=\; \left(\widetilde{T}_{\rm TG}^{(\ell)}\right)^T\, A_\ell\, \widetilde{T}_{\rm TG}^{(\ell)}$$

entailing

$$\sigma_{\mathrm{TG}}^{(\ell)} \;=\; \left(\widetilde{\rho}_{\mathrm{TG}}^{(\ell)}\right)^2. \tag{3.8}$$

That is, if one uses the same smoother and prolongation, the two-grid method with symmetrized pre- and postsmoothing converges twice as fast as the method without postsmoothing.

Acknowledging this fact, one sees that (3.7) defines recursively a bound on $\widetilde{\rho}_{\mathrm{MG}}^{(\ell)}$ which is equal to the square root of the bound on $\sigma_{\mathrm{MG}}^{(\ell)}$ obtained from our result (3.2). This suggests that, as shown by (3.8) in the two-grid case, the W-cycle multigrid with symmetrized smoothing can converge twice as fast as the corresponding algorithm without postsmoothing.

Sometimes the bound for the scheme without postsmoothing is used as worst case estimate for the general case. From that point of view, our analysis gives sharper bounds. For instance, Theorem 3.4 in [6, Chapter V] states that the convergence factor for the W-cycle is not larger that $3/5$ when $\widetilde{\rho}_{\mathrm{TG}}^{(\ell)} \leq 1/2$ for all ℓ, whereas, with $\sigma_{\mathrm{TG}}^{(\ell)} \leq 1/4$, our theorem, Theorem 3.1, then proves that the convergence factor for the W-cycle does not exceed $1/3$.

Acknowledgments. I thank M. Hochstenbach for a careful reading of the manuscript. An anonymous referee suggested numerous improvements which are deeply appreciated. Another referee drew our attention to [6].

REFERENCES

[1] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.
[2] O. Axelsson and P. S. Vassilevski, *Algebraic multilevel preconditioning methods*, I, Numer. Math., 56 (1989), pp. 157–177.
[3] O. Axelsson and P. S. Vassilevski, *Algebraic multilevel preconditioning methods*, II, SIAM J. Numer. Anal., 27 (1990), pp. 1569–1590.
[4] R. E. Bank, T. F. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math., 52 (1988), pp. 427–458.
[5] D. Braess, *The convergence rate of a multigrid method with Gauss-Seidel relaxation for the Poisson equation*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., Lecture Notes in Math. 960, Springer-Verlag, Berlin, Heidelberg, New York, 1982, pp. 368–386.
[6] D. Braess, *Finite Elements*, Cambridge University Press, Cambridge, UK, 1997.
[7] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal., 20 (1983), pp. 967–975.
[8] A. Brandt, *Algebraic multigrid theory: The symmetric case*, Appl. Math. Comput., 19 (1986), pp. 23–56.
[9] R. D. Falgout and P. S. Vassilevski, *On generalizing the algebraic multigrid framework*, SIAM J. Numer. Anal., 42 (2004), pp. 1669–1693.
[10] R. D. Falgout, P. S. Vassilevski, and L. T. Zikatanov, *On two-grid convergence estimates*, Numer. Linear Algebra Appl., 12 (2005), pp. 471–494.
[11] W. Hackbusch, *Multi-grid convergence theory*, in Multigrid Methods, W. Hackbusch and U. Trottenberg, eds., Lecture Notes in Math. 960, Springer-Verlag, Berlin, Heidelberg, New York, 1982, pp. 177–219.
[12] W. Hackbusch, *Multi-grid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[13] J. Mandel, S. McCormick, and J. Ruge, *An algebraic theory for multigrid methods for variational problems*, SIAM J. Numer. Anal., 25 (1988), pp. 91–110.
[14] S. F. McCormick, *Multigrid methods for variational problems: General theory for the V-cycle*, SIAM J. Numer. Anal., 22 (1985), pp. 634–643.
[15] Y. Notay, *Algebraic multigrid and algebraic multilevel methods: A theoretical comparison*, Numer. Linear Algebra Appl., 12 (2005), pp. 419–451.
[16] P. Oswald, *Multilevel Finite Element Approximation: Theory and Applications*, Teubner Skr. Numer., Teubner, Stuttgart, 1994.

[17] P. Oswald, *Subspace correction methods and multigrid theory*, in Multigrid, Academic Press, London, 2001, pp. 533–572.
[18] K. Stüben, *An introduction to algebraic multigrid*, in Multigrid, Academic Press, London, 2001, pp. 413–532.
[19] U. Trottenberg, C. W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, London, 2001.
[20] P. S. Vassilevski, *A block–factorization (algebraic) formulation of multigrid and Schwarz methods*, East-West J. Numer. Math., 6 (1998), pp. 65–79.
[21] P. S. Vassilevski, *Multilevel Block Factorization Preconditioners*, Springer-Verlag, New York, to appear.
[22] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[23] H. Yserentant, *Old and new convergence proofs for multigrid methods*, Acta Numer., 2 (1993), pp. 285–326.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1045–1063

THE STABILITY OF FRONT-TRACKING SCHEME FOR TWO SPATIAL DIMENSIONAL MISCIBLE TWO PHASE FLOW*

KOU-KUNG ALEX CHANG†

Abstract. The front-tracking method of Glimm et al. is designed for physical applications which are governed by a hyperbolic system of equations in which codimension one (cd-1) jump discontinuities in the solution are important. The core of this method is the curve propagation algorithm. A theoretical analysis is presented here to show that this curve propagation algorithm is stable. Chang and Lindquist developed a new curve propagation algorithm which conserves mass for any grid of finite spacing. We also present a proof to show the new curve propagation algorithm is stable under some conditions.

Key words. front tracking, discontinuity propagation, two phase flow

AMS subject classifications. 76S05, 76E15, 82A42

DOI. 10.1137/040616693

1. Introduction.

In 1981, Glimm et al. presented a computational method called front tracking [5], [8], [9], which is designed for physical applications which are governed by a hyperbolic system of equations in which codimension one (cd-1) jump discontinuities in the solution are important. The front-tracking scheme has been developed into a computational code, FronTier, for dealing with two or three spatial dimensional problems [4], [11] and has been used in various applications [3], [6], [7], [12].

In this computational method, two important problems are of concern; one is if the mass balance is conserved when it is applied in practice; the other is if this numerical method is stable.

For the first problem, Glimm, Lindquist, and Zhang address five algorithmic areas that need to be corrected for maintaining mass conservation when applying this method to two phase flow in porous media [7]. In the same paper, Glimm et al. present algorithms to resolve four of them. The fifth algorithmic area of problem is the explicit movement of the fluid discontinuity curves (the curve propagation algorithm). Errors produced by the curve propagation algorithm for a nonlinear case can be absorbed by the numerical approximations. For the linear case, errors are not dissipated and increase with time. Chang and Lindquist analyzed the mass balance error produced by the curve propagation algorithm for miscible flow case (linear case) and showed that up to a finite time T, the mass balance error vanishes as the length of the line segments of the continuous piecewise linear discontinuity curve goes to zero when the numerical velocity field is piecewise constant or the numerical velocity field is Lipschitz continuous [1], [2]. Meanwhile, Chang and Lindquist developed a new curve propagation algorithm which is assured to conserve mass for any grid of finite spacing [1].

In this paper, we study the other problem: the stability property of the front-tracking scheme applied to miscible two phase incompressible flow in porous media.

*Received by the editors October 10, 2004; accepted for publication (in revised form) January 3, 2007; published electronically May 7, 2007.
http://www.siam.org/journals/sinum/45-3/61669.html
†Department of Applied Mathematics, National Pingtung University of Education, Pingtung 90003, Taiwan, Republic of China (chang@mail.npue.edu.tw).

Our analysis focuses on the stability of the curve propagation (both original and new) algorithms, which are the core of the front-tracking scheme.

The behavior of two phase incompressible flow in porous media is governed by a specific system of equations,

$$\phi(X)\frac{\partial s}{\partial t} + \nabla \cdot \overrightarrow{v}(X) f(s) = 0, \tag{1}$$

$$\nabla \cdot \overrightarrow{v}(X) = 0, \tag{2}$$

$$\overrightarrow{v}(X) = -\lambda(s)\kappa(X) \bigtriangledown P(X), \tag{3}$$

where (1) is a hyperbolic subsystem, and (2) and (3) are elliptic subsystems. In this system of equations, X is the space variable in R^2, and ϕ is the medium porosity (volume fraction of pore space); s and $1-s$ are the respective saturations (fractions of available pore volume) of the two flowing fluid phases; $\overrightarrow{v}$ is the total fluid velocity; $f\overrightarrow{v}$ and $(1-f)\overrightarrow{v}$ are the respective fractions of the total fluid velocity carried by each phase; P is the pressure field in the medium; κ is the medium permeability; and λ is the saturation-dependent total relative transmissibility. We shall also refer to the mobility ratio, $\mathrm{M} \equiv \lambda(1)/\lambda(0)$, which governs the linearized analysis for fingering instability in two phase flows; $\mathrm{M} \leq 1$ corresponds to stable flows and $\mathrm{M} > 1$ to unstable flows. We specify our problem on miscible two phase flow for more fundamental reasons, and (1) becomes

$$\phi(X)\frac{\partial s}{\partial t} + \overrightarrow{v}(X) \cdot \nabla s = 0. \tag{1'}$$

We outline the rest of this paper as follows: section 2 will describe the front-tracking method, curve propagation algorithms, and some relative assumptions and definitions. Section 3 contains our proof of the stability of the original curve propagation algorithm. Section 4 shows the stability of the mass conserved curve propagation algorithm.

2. The front-tracking method. The front-tracking method approach to solving (1)–(3) utilizes an implicit pressure explicit saturation scheme of sequential solution of the coupled hyperbolic-elliptic system (1)–(3). The elliptic subsystem is solved by mixed finite elements yielding simultaneous solutions of P and $\overrightarrow{v}$. Because our analysis focuses on the stability of the front tracking for solving the hyperbolic subsystem, we do not discuss this mixed finite element algorithm further. Besides, without loss of generality, we may assume that the velocity field is Lipschitz continuous.

The hyperbolic subsystem is solved by the front-tracking algorithms which employ a fixed, volume filling grid to resolve the smooth part of the solution and moving cd-1 grids to resolve jump discontinuities and their motion. The fixed grid is regular rectangular and of discretization spacing Δx and Δy; we shall refer to it as the hyperbolic grid. The moving cd-1 discontinuity grids in two dimensional space are unions of *curves*. For our specific miscible flow problem, the curves separate the domain geometrically to several disconnected subdomains. Each curve has an orientation and is replaced by its continuous piecewise linear approximant. We denote a linear segment from this approximant by $\overline{ab}$, where a is the start point and b is the end point of this segment. The start and end points of all segments in the piecewise linear approximant are called *points of the curve* or simply *interior points* or *points*. The first point on the piecewise linear approximant is called the *start node*, and the last is the *end node*. The points on a curve carry physical data in the left and right sides of the curve, which are the saturations in our case. For details, we refer the reader to [10].

The algorithmic procedure for updating the numerical solution to subsystem (1) consists of two major tasks: (I) achieve curve propagation, which generates the discontinuity curves, transmits waves across the front, and updates flow tangentially to the front; (II) update the solution on the fixed grid. We describe them as follows:

(I) Basically, curve propagation is achieved by changing the coordinates locally at each curve point in normal and tangential directions to the discontinuity curve. Therefore (1) becomes

$$\phi(X)s_t + (\hat{n}\cdot\nabla)\overrightarrow{v}f(s) + (\hat{t}\cdot\nabla)\overrightarrow{v}f(s) = 0 \tag{4}$$

at each point p, $\hat{n}$ is the normal vector to the discontinuity curve, and $\hat{t}$ is the tangent vector. The algorithm for solving (4) can be described in terms of series of splittings of hyperbolic operators:

(a) Solve the local Riemann problem at curve point p for equation

$$\phi(X)s_t + (\hat{n}\cdot\nabla)\overrightarrow{v}f(s) = 0, \tag{5}$$

with state values

$$s(p+\hat{n}) \equiv s_{left}, \qquad s(p-\hat{n}) \equiv s_{right},$$

to find propagation speed. This solution provides a new propagated position p' for the point. In the miscible case (i.e., the linear case), the propagation speed is $\overrightarrow{v}(x)\cdot\hat{n}|_p$.

(b) At each point on the propagated discontinuity grid, solve equation

$$\phi(X)s_t + (\hat{t}\cdot\nabla)\overrightarrow{v}f(s) = 0 \tag{6}$$

by using a one dimensional finite difference scheme to update the state value on both sides of the discontinuity curve. For the miscible flow case, the state value on each side should be constant.

(c) It may happen that an unphysical crossing of the discontinuity curve occurred during movement. We need to untangle this unphysical crossing to ensure physical correctness of the discontinuity grids. This untangling step is required between steps (a) and (b).

(d) Points on the discontinuity curve may become too far apart or too close together, and so a redistribution step is necessary to redistribute points on the propagated grids for ensuring adequate sampling of the discontinuity grid along its arc length. For more detail on these actions, we refer the reader to [1] and [13].

(II) Updating the solution on the fixed grid is done by solving an initial-boundary value problem on both sides of the front (discontinuity curve). We treat the front as a moving boundary and never use states (phase saturations s here) on the opposite side of the front. This can be done by using any stable, two spatial dimensional, numerical algorithm. For the miscible flow case studied here, the phase saturations remain piecewise constant in space, and implementation of (II) is trivial and stable. Therefore we concentrate on the curve propagation algorithms. We recall some fundamental definitions, underlying algorithmic procedures, and theorems in [1] and [2].

DEFINITION NBC. *The normal direction to the linear segment $\overline{p_1p_2}$ is the direction perpendicular to $\overline{p_1p_2}$, oriented from the right side to the left side.*

DEFINITION NC. *Let $\overline{p_1p_2}$ and $\overline{p_2p_3}$ be two consecutive segments from the piecewise linear approximant. Then the normal to the curve at p_2 is the direction perpendicular to the segment $\overline{p_1p_3}$, oriented from the right side to the left side of the curve.*

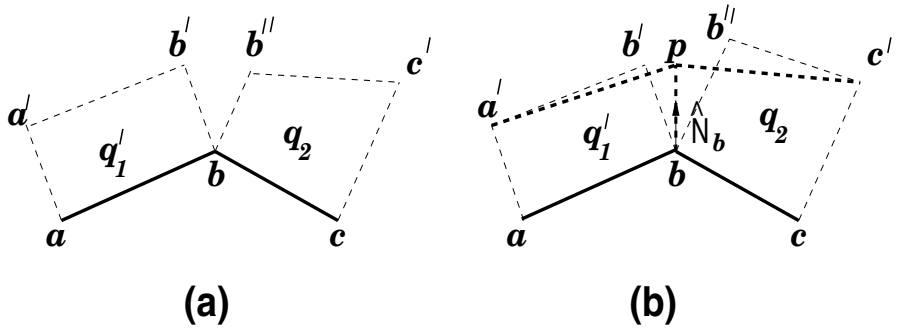


FIG. 2.1. (a) q_1' *(area* $aa'b'b$*) and* q_2 *(area* $bb''c'c$*) are the conserved mass changes by the movement of linear segments* $\overline{ab}$ *and* $\overline{bc}$*.* (b) $\hat{N}_b$ *is the curve normal at point* b*. Along* $\hat{N}_b$*, find a position* p *such that the mass change in the area* $aa'pc'cb$ *is equal to* $q_1' + q_2$*.*

ALGORITHM BP. *Given* p *of* $\overline{p_1 p}$ *(or* $\overline{pp_1}$*), then its propagated image,* p'*, is obtained by solving a one dimensional Riemann problem along the* normal to the linear segment $\overline{p_1 p}$ *(*$\overline{pp_1}$*).*

ALGORITHM PC. *Given* p_2 *of* $\overline{p_1 p_2}$ *and* $\overline{p_2 p_3}$*, then its propagated image,* p_2''*, is obtained by the solution of a one dimensional Riemann problem along the* normal to the curve *at* p_2*.*

ALGORITHM CP. *The movement of a tracked curve is achieved by propagating each interior point by Algorithm PC over a time step* Δt*.*

THEOREM 2.1. *Let* $E(T, \Delta t)$ *be the total mass error at time* T *produced by moving the interior of discontinuity curve under Algorithm PC. Assume the curvature of the curve is bounded by* $K(T) < \infty$*. Then* $E(T, \Delta t) = O(\Delta t)$ *as* $\Delta t \to 0$*.*

ALGORITHM NPC. *Consider the propagation of point* b*, where* $\overline{ab}$ *and* $\overline{bc}$ *are two consecutive segments from the piecewise linear approximant to the discontinuity curve. Assume point "a" has been previously propagated to* a' *in a mass conserving manner. With reference to Figure* 2.1(a)*, propagate* b *to* b' *by Algorithm BP (*b *is considered as a point from* $\overline{ab}$*); propagate* b *and* c*, respectively, to* b'' *and* c' *by Algorithm BP (*b *and* c *are considered as points from* $\overline{bc}$*). Consider the respective conservative mass changes* q_1' *and* q_2 *produced by the movements* $\overline{ab} \to a'b'$ *and* $\overline{bc} \to b''c'$*. Find a point* p *(illustrated in Figure* 2.1(b)*) lying along the curve normal direction at* b *such that the mass change in the area* $abpa' + bcc'p$ *is equal to* $q_1' + q_2$*.*

Remark 1. For the miscible flow case, the phase saturations are piecewise constant in space (we assume the saturations are 0 and 1 here). Therefore the mass change q_1' in the quadrilateral $abb'a'$ is equal to the area of $abb'a'$, i.e., $q_1' = \frac{1}{2}[\overrightarrow{ab} \times (\overrightarrow{aa'} + \overrightarrow{bb'}) + \overrightarrow{bb'} \times \overrightarrow{aa'}]$. Similarly, $q_2 = \frac{1}{2}[\overrightarrow{bc} \times (\overrightarrow{bb''} + \overrightarrow{cc'}) + \overrightarrow{bb''} \times \overrightarrow{cc'}]$. Since $\overrightarrow{bb''}$ is parallel to $\overrightarrow{cc'}$, q_2 becomes $\frac{1}{2}\overrightarrow{bc} \times (\overrightarrow{bb''} + \overrightarrow{cc'})$.

THEOREM 2.2. *In Algorithm NPC, there exists a unique point* p *lying along the curve normal direction at* b *such that the mass change in the area* $abpa' + bcc'p$ *is equal to* $q_1' + q_2$ *iff* $\hat{N}_b \times \overrightarrow{c'a'} \neq 0$*, where* $\hat{N}_b$ *is the unit curve normal vector at* b*.*

Proof. In order to find a unique point p lying along the curve normal direction at b such that the mass change in the area $abpa' + bcc'p$ is equal to $q_1' + q_2$, we need to

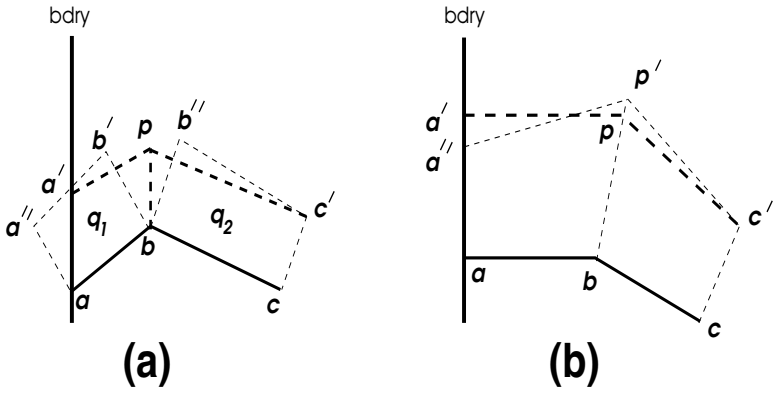


FIG. 2.2. (a) *Propagation of a point adjacent to a start node.* (b) *Propagation of a point adjacent to a start node with the first linear segment perpendicular to the boundary.*

find a scalar λ such that

$$[\lambda\hat{N}_b \times \overrightarrow{aa'} + \overrightarrow{ab} \times (\lambda\hat{N}_b + \overrightarrow{aa'})] + [\overrightarrow{cc'} \times \lambda\hat{N}_b + \overrightarrow{bc} \times (\overrightarrow{cc'} + \lambda\hat{N}_b)] = 2(q_1' + q_2). \tag{7}$$

After an easy calculation, we have

$$\lambda[\hat{N}_b \times \overrightarrow{c'a'}] = 2(q_1' + q_2) + \overrightarrow{aa'} \times \overrightarrow{ab} + \overrightarrow{cc'} \times \overrightarrow{bc}. \tag{8}$$

This equation has a unique solution λ iff $\hat{N}_b \times \overrightarrow{c'a'} \neq 0$. □

Remark 2. Equation (8) has no solution for λ when $\hat{N}_b \times \overrightarrow{c'a'} = 0$. For resolving this problem, we reduce the time step size to produce new propagated positions c' and a' for c and a. After reducing the time step, $\hat{N}_b \times \overrightarrow{c'a'}$ will still vanish only if the linear segments $\overline{ab}$ and $\overline{bc}$ are colinear. In this case, we propagate point b by Algorithm PC, which will be mass conserving.

Remark 3. In general, Algorithm NPC requires a minor modification when applied to the first interior point of a curve. Assume point a of Figure 2.2(a) is the start node of a curve and is propagated to point a' by an appropriate node propagation algorithm. Consider the point a'' obtained by propagating a by Algorithm BP. In propagating the first interior point, b, the mass change q_1 in the area $abb'a''$ (rather than $q_1{}'$ of $abb'a'$) should be used in applying Algorithm NPC. Analogous modification is required when NPC is applied to the last interior point of a curve.

Remark 4. In practice, for many fluid calculation cases, the contact discontinuity curve should be perpendicular to the boundary. Algorithm NPC propagates the first interior point of a curve in the following way: After every point of a curve has been propagated by Algorithm NPC (general case), adjust point a'' to a' as well as p' to p such that the mass change of the area $abcc'pa'$ equals that of the area $abcc'p'a''$ (see Figure 2.2(b)), and the linear approximation $\overline{ap}$ must be perpendicular to the boundary. Similar modification is required when NPC is applied to the last interior point of a curve.

3. Stability of front tracking. In this section, we will discuss the stability of the original curve propagation algorithm. Before continuing our discussion, we will present the definition of stability and some assumptions.

DEFINITION 3.1. *Let U be a numerical method and $U_{\Delta t}$ be the numerical solution with the length of time step Δt. Then U is stable if there are R, M, and N depending*

on the initial U_0 and the velocity field $\overrightarrow{v}(x,y)$ such that all approximations $U_{\Delta t}$ satisfy the following conditions:

(a) *Supp*$(U_{\Delta t}(\cdot, n\Delta t)) \subset [-M,M]\times[-N,N]$, *where* $n\Delta t \in [0,T]$.

(b) $\sum_{n=0}^{T/\Delta t}\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}(\Delta t\Delta x|U^n_{i,j+1}-U^n_{i,j}|+\Delta t\Delta y|U^n_{i+1,j}-U^n_{i,j}|+\Delta x\Delta y|U^{n+1}_{i,j}-U^n_{i,j}|) \le R$ *for all* $\Delta t < \Delta t_0$ *for some* Δt_0. *Here* Δx, Δy *are the block lengths in the direction of* x *and* y, *and* $U^n_{i,j}$ *is the value of* $U_{\Delta t}$ *at* $t = n\Delta t$, $x = i\Delta x$, *and* $y = j\Delta y$.

Our discussions require two assumptions: one is that the velocity field $\overrightarrow{v}(x)$ satisfies the Lipschitz condition; the other is the discontinuity curves at all time steps are C^2 curves, and the curvature on every point of the curve is bounded by κ. These two assumptions are explained as follows.

Assumption L1. Let $\{\Delta_i\}$ denote the set of the mesh elements of the velocity field. We assume the velocity $\overrightarrow{v_i}$ in a mesh element Δ_i is assigned by $\overrightarrow{v}(X)$ for some $X \in \Delta_i$. Furthermore, we assume the velocity field $\overrightarrow{v}(X)$ is Lipschitz continuous; i.e., $|\overrightarrow{v}(X)-\overrightarrow{v}(Y)| \le K||X-Y||$ for some constant K, where $||\cdot||$ is the space norm and X and Y are two dimensional space variables.

Assumption C1. Let σ^n be the discontinuity curve at time step n; we assume σ^n is a C^2 curve for all n, and the curvature $|\kappa(p)| \le \kappa < \infty$ for all points p of σ^n.

A curve is represented by its continuous piecewise linear approximant. If the length of any of the segments from this approximant is too big, then this is a poor approximation to the curve. If it is too small, it leads to numerical induced tangling during propagation. Therefore we need an assumption (Assumption R1) that restricts a piecewise linear approximation. In practice, The FronTier package uses algorithms described in section 2 to achieve this restriction.

Assumption R1. Let σ^n be the discontinuity curve at time step n and h be the hyperbolic grid size. We restrict the length b^n_i of a linear segment from the piecewise linear approximant to $c_1 h < |b^n_i| < c_2 h$ for all i, n, where c_1 and c_2 are constants.

Assumption R2. We assume the algorithm satisfies the CFL condition $\Delta t \le \alpha c_1 h/V$, where α is some constant $\alpha < 1$, c_1 is the same as in Assumption R1, and V is the largest magnitude velocity of discontinuity motion in the computational region during the entire flow process. Furthermore, we assume there is a positive integer N such that $T = \Delta t N$.

LEMMA 3.2. *Let b^n_i be a linear segment from the continuous piecewise linear approximation to the discontinuity curve σ^n at time step n and b^{n+1}_i be the line segment propagated from b^n_i. Then under Assumptions L1 and C1, the length* $|b^{n+1}_i| \le |b^n_i|(1+\Delta t(2V\kappa + K))$.

Proof. Let $\overrightarrow{pq} = b^n_i$, $\hat{n}_1$ be the normal to the curve at q and $\hat{n}_2$ be the normal to the curve at p (see Figure 3.1). Then

$$b^{n+1}_i = \overrightarrow{p'q'} = (\Delta t\,\overrightarrow{v}_1\cdot\hat{n}_1)\hat{n}_1 + b^n_i - (\Delta t\,\overrightarrow{v}_2\cdot\hat{n}_2)\hat{n}_2. \tag{9}$$

By triangle inequality,

$$|b^{n+1}_i| \le |b^n_i| + |(\Delta t\overrightarrow{v_1}\cdot\hat{n}_1)\cdot\hat{n}_1 - (\Delta t\overrightarrow{v_2}\cdot\hat{n}_2)\cdot\hat{n}_2|, \tag{10}$$

and by an intermediate calculation, we have

$$\begin{aligned}|b^{n+1}_i| \le |b^n_i| &+ \Delta t|(\overrightarrow{v_1}\cdot(\hat{n}_1-\hat{n}_2))\hat{n}_1| + \Delta t|(\overrightarrow{v_1}\cdot\hat{n}_2)(\hat{n}_1-\hat{n}_2)| \\ &+ \Delta t|((\overrightarrow{v_1}-\overrightarrow{v_2})\cdot\hat{n}_2)\hat{n}_2|.\end{aligned} \tag{11}$$

Since $\overrightarrow{v}(X)$ is continuous on a closed and bounded region, $[-M,M]\times[-N,N]$, there is a maximum velocity, V, such that $|\overrightarrow{v}(X)| \le V$ for all $X \in [-M,M]\times[-N,N]$.

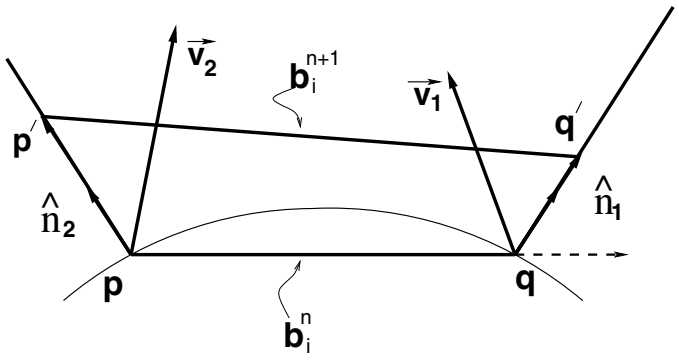


FIG. 3.1. *Piecewise linear approximation $\overline{pq}$ of discontinuity curve σ^n.*

And, because σ^n is a C^2 curve, if $|b_i^n|$ is small enough, then

$$|\hat{n}_1 - \hat{n}_2| \approx |\kappa(p)||b_i^n| \le \kappa|b_i^n|, \tag{12}$$

where $\kappa(p)$ is the curvature at p on σ^n. By Assumption L1, $\overrightarrow{v}(X)$ is Lipschitz continuous; then

$$|(\overrightarrow{v_1} - \overrightarrow{v_2})| \le K|b_i^n|. \tag{13}$$

Hence, from (11), we have

$$\begin{aligned} |b_i^{n+1}| &\le |b_i^n| + \Delta t V\kappa|b_i^n| + \Delta t V\kappa|b_i^n| + \Delta t K|b_i^n| \\ &\le |b_i^n|(1 + \Delta t(2V\kappa + K)). \end{aligned} \tag{14}$$

This completes our proof. ∎

LEMMA 3.3. *Let σ^n be the discontinuity curve at the nth time step. Then under Assumptions L1 and C1, $|\sigma^n| \le |\sigma^0|(1 + \Delta t(2V\kappa + K))^n$.*

Proof. Assume σ^n consists of m linear pieces b_i^n, $i = 1, 2, \cdot, m$. From Lemma 3.2,

$$|\sigma^{n+1}| = \sum_{i=1}^{m} |b_i^{n+1}| \le \sum_{i=1}^{m} |b_i^n|(1 + \Delta t(2V\kappa + K)) \le |\sigma^n|(1 + \Delta t(2V\kappa + K)). \tag{15}$$

By induction, we have

$$|\sigma^n| \le |\sigma^0|(1 + \Delta t(2V\kappa + K))^n. \quad \blacksquare \tag{16}$$

LEMMA 3.4. *Let $U_{i,j}^n$ be the numerical approximation solution of $s(x, y; t)$ of* (1), (2), and (3) *at $t = n\Delta t$, $x = i\Delta x$, and $y = j\Delta y$. Without loss of generality, we assume $\Delta x = \Delta y = h$; then under Assumptions L1, C1, R1, and R2,*

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta t h|U_{i,j+1}^n - U_{i,j}^n| \le R_1 \tag{17}$$

and

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta t h|U_{i+1,j}^n - U_{i,j}^n| \le R_2 \tag{18}$$

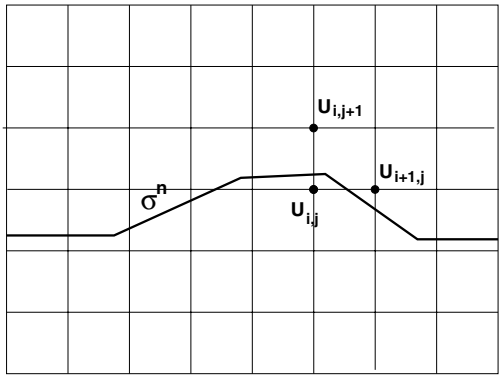


FIG. 3.2. *The discontinuity curve σ^n crosses the grid line between $(i, j+1)$ and (i, j).*

for some constants R_1 and R_2 and for all $\Delta t < \Delta t_0$.

Proof. We want to prove

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta t h |U_{i,j+1}^n - U_{i,j}^n| \le R_1$$

for some constant R_1. First, we see

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta t h |U_{i,j+1}^n - U_{i,j}^n| = \sum_{n=0}^{T/\Delta t} \Delta t \left(\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h |U_{i,j+1}^n - U_{i,j}^n| \right). \tag{19}$$

Since $U_{i,j}^n$ only has two values, say 1 or 0, then

$$|U_{i,j+1}^n - U_{i,j}^n| = \begin{cases} 1 & \text{if } \sigma^n \text{ crosses the grid line between } (i,\, j+1) \text{ and } (i,\, j), \\ 0 & \text{otherwise,} \end{cases} \tag{20}$$

where σ^n is the discontinuity curve at time step $n\Delta t$ (see Figure 3.2). Then the computation of

$$\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h |U_{i,j+1}^n - U_{i,j}^n|$$

is equivalent to counting how many vertical grid lines σ^n intersects. Note that we restricted that the length of each piece of the piecewise linear approximant is bounded by $c_1 h \le |b_i^n| \le c_2 h$. Then each piece may cross $r = [c_2 h/h] + 1 = [c_2] + 1$ vertical grid lines, where $[c_2]$ means the largest integer that is no greater than c_2. A discontinuity curve σ^n may have at most $|\sigma^n|/(c_1 h)$ linear pieces. Therefore $\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h |U_{i,j+1}^n - U_{i,j}^n| \le hr|\sigma^n|/(c_1 h)$. Hence

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta t h |U_{i,j+1}^n - U_{i,j}^n| \le \sum_{n=0}^{T/\Delta t} \Delta t h r \frac{|\sigma^n|}{c_1 h} = \frac{r}{c_1} \sum_{n=0}^{T/\Delta t} \Delta t |\sigma^n|. \tag{21}$$

By Lemma 3.3, $|\sigma^n| \le |\sigma^0|(1+\Delta t(2V\kappa+K))^n$; then

$$\begin{aligned}\frac{r}{c_1}\sum_{n=0}^{T/\Delta t}\Delta t|\sigma^n| &\le \frac{r}{c_1}\sum_{n=0}^{T/\Delta t}\Delta t|\sigma^0|(1+\Delta t(2V\kappa+K))^n \\ &\le \frac{r|\sigma^0|}{c_1}\sum_{n=0}^{T/\Delta t}\Delta t(1+\Delta t(2V\kappa+K))^n. \qquad (22)\end{aligned}$$

From our assumption, $T=\Delta tN$ implies $\Delta t=\frac{T}{N}$. Therefore

$$\begin{aligned}&\frac{r|\sigma^0|}{c_1}\sum_{n=0}^{T/\Delta t}\Delta t(1+\Delta t(2V\kappa+K))^n \\ &= \Delta t\frac{r|\sigma^0|}{c_1}+\frac{r|\sigma^0|}{c_1}\sum_{n=1}^{T/\Delta t}\Delta t\left(1+\frac{T(2V\kappa+K)}{N}\right)^n. \qquad (23)\end{aligned}$$

Since $1\le n\le N$,

$$1+\frac{T(2V\kappa+K)}{N} \le 1+\frac{T(2V\kappa+K)}{n}. \qquad (24)$$

Furthermore, $(1+T(2V\kappa+K)/n)^n$ is increasing to $e^{T(2V\kappa+K)}$ as n goes to ∞; then

$$\left(1+\frac{T(2V\kappa+K)}{n}\right)^n \le e^{T(2V\kappa+K)} \qquad (25)$$

for all n. Then

$$\begin{aligned}&\Delta t\frac{r|\sigma^0|}{c_1}+\frac{r|\sigma^0|}{c_1}\sum_{n=1}^{T/\Delta t}\Delta t\left(1+\frac{T(2V\kappa+K)}{N}\right)^n \\ &\le \Delta t\frac{r|\sigma^0|}{c_1}+\frac{r|\sigma^0|}{c_1}\sum_{n=1}^{T/\Delta t}\Delta t\left(1+\frac{T(2V\kappa+K)}{n}\right)^n \\ &\le \Delta t\frac{r|\sigma^0|}{c_1}+\frac{r|\sigma^0|}{c_1}\sum_{n=1}^{T/\Delta t}\Delta te^{T(2V\kappa+K)} = \frac{r|\sigma^0|}{c_1}Te^{T(2V\kappa+K)}. \qquad (26)\end{aligned}$$

Define $R_1=(r|\sigma^0|/c_1)Te^{T(2V\kappa+K)}$; then we have

$$\sum_{n=0}^{T/\Delta t}\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}\Delta th|U^n_{i,j+1}-U^n_{i,j}| \le R_1. \qquad (27)$$

Similarly, we may have

$$\sum_{n=0}^{T/\Delta t}\sum_{i=-\infty}^{\infty}\sum_{j=-\infty}^{\infty}\Delta th|U^n_{i+1,j}-U^n_{i,j}| \le R_2 \qquad (28)$$

for all $\Delta t<\Delta t_0$. □

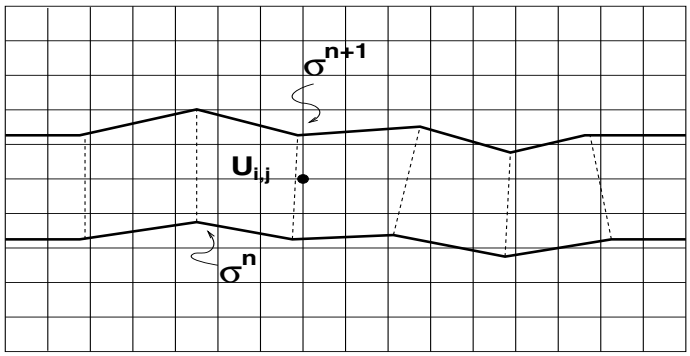


FIG. 3.3. $|U_{i,j}^{n+1} - U_{i,j}^n| \neq 0$ *only when* (i,j) *is in the region of* σ *swept from time step* $n\Delta t$ *to time step* $(n+1)\Delta t$.

LEMMA 3.5. *Let* $U_{i,j}^n$ *be the numerical approximation solution of* $s(x,y;t)$ *of* (1), (2), and (3) *at* $t = n\Delta t$, $x = i\Delta x$, *and* $y = j\Delta y$. *Then under Assumptions* $L1$, $C1$, $R1$, *and* $R2$,

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \Delta x \Delta y |U_{i,j}^{n+1} - U_{i,j}^n| \leq R_3 \tag{29}$$

for some constant R_3 *and for all* $\Delta t < \Delta t_0$.

Proof. Without loss of generality, one may assume $\Delta x = \Delta y = h$. Because we are considering miscible two phase flow, calculating

$$\sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h^2 |U_{i,j}^{n+1} - U_{i,j}^n|$$

is equivalent to calculating the area that the discontinuity curve sweeps from time step $n\Delta t$ to time step $(n+1)\Delta t$ (see Figure 3.3). Therefore evaluating

$$\sum_{n=0}^{T/\Delta t} \sum_{i=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} h^2 |U_{i,j}^{n+1} - U_{i,j}^n|$$

is equivalent to evaluating

$$\sum_{n=0}^{T/\Delta t} A(\sigma^n),$$

where $\mathrm{A}(\sigma^n)$ is the area that the discontinuity curve σ^n sweeps from time $n\Delta t$ to time $(n+1)\Delta t$. Since a discontinuity curve consists of linear pieces, we first need to compute the area swept by a linear piece $\overline{ab}$ from time $n\Delta t$ to time $(n+1)\Delta t$. Let us denote by $\overline{cd}$ the image of $\overline{ab}$ at time $(n+1)\Delta t$. Then the swept area is the area of the quadrilateral $abcd$, denoted by $A(abcd)$. Let $q_{\overline{ab}}$ be the conserved mass movement across the linear piece $\overline{ab}$ from time $n\Delta t$ to time $(n+1)\Delta t$. Then

$$q_{\overline{ab}} = \Delta t \left| \int_0^{l_{ab}} \overrightarrow{v_\perp}(x,y) d\sigma \right|, \tag{30}$$

where $\overrightarrow{v_\perp}$ is the normal component of the velocity field at the linear piece $\overline{ab}$ with length l_{ab}, and σ is the arc length along $\overline{ab}$. Because we are considering miscible flow, $\overrightarrow{v_\perp}$ becomes $\overrightarrow{v}(x,y)\cdot\hat{n}$, where $\overrightarrow{v}(x,y)$ is the velocity of the linear piece $\overline{ab}$ and $\hat{n}$ is its normal. If the numerical velocity field $\overrightarrow{v}(x,y)$ is bounded by a velocity V, i.e., $|\overrightarrow{v}(x,y)|\le V$ for all x and y, then

$$q_{\overline{ab}} = \int_0^{l_{ab}} \Delta t(\overrightarrow{v}\cdot\hat{n})d\sigma \le \Delta t V l_{ab}. \tag{31}$$

If we define $err(\overline{ab})$ to be $|A(abdc) - q_{\overline{ab}}|$, then

$$|A(abdc)| = |A(abdc) - q_{\overline{ab}} + q_{\overline{ab}}| \le \Delta t V|\overline{ab}| + err(\overline{ab}). \tag{32}$$

Now assume σ^n has m pieces b_i^n, $i=1,\dots,m$, at time $t=n\Delta t$. The area swept by σ^n from time $n\Delta t$ to $(n+1)\Delta t$ is

$$|A(\sigma^n)| \le \sum_{i=1}^m |A(b_i^n)| \le \sum_{i=1}^m (\Delta t V|b_i^n| + err(b_i^n)) \le \Delta t V|\sigma^n| + E(\sigma^n), \tag{33}$$

where $E(\sigma^n) = \sum_{i=1}^m err(b_i^n)$. Similarly to the proof of Lemma 3.4,

$$\sum_{n=0}^{T/\Delta t} \Delta t V|\sigma^n| \le V|\sigma^0|Te^{T(2V\kappa+K)}. \tag{34}$$

By Theorem 2.1,

$$\sum_{n=0}^{T/\Delta t} E(\sigma^n) \to 0$$

as $\Delta t\to 0$. Therefore, given any constant, say 1, there is Δt_0 such that $\sum_{n=0}^{T/\Delta t} E(\sigma^n) < 1$ whenever $\Delta t \le \Delta t_0$. Let $R_3 = TV|\sigma^0|\exp(T(2V\kappa+K))+1$. Then

$$\sum_{n=0}^{T/\Delta t} |A(\sigma^n)| \le \sum_{n=0}^{T/\Delta t} (\Delta t V|\sigma^n| + E(\sigma^n)) = \sum_{n=0}^{T/\Delta t} \Delta t V|\sigma^n| + \sum_{n=0}^{T/\Delta t} E(\sigma^n) \le R_3. \qquad \square \tag{35}$$

Now we have the following theorem.

THEOREM 3.6. *Under Assumptions L1, R1, R2, and C1, the curve propagation algorithm, Algorithm PC, of the front-tracking scheme for incompressible miscible two phase flow on porous media is stable.*

Proof. By Lemmas 3.4 and 3.5, the result holds. $\square$

4. Stability of mass conserved curve propagation algorithm. By Theorem 2.2, the new point propagation algorithm, Algorithm NPC, determines unique point p' iff $\hat{N}_b \times \overrightarrow{c'a'} \ne 0$. The only problem is that if $\hat{N}_b \times \overrightarrow{c'a'} = \epsilon \ne 0$ is very small, the new point p' will be far away from the old point p. This will change the shape of the discontinuity curve and make this algorithm unstable. Therefore we need to analyze in which condition this algorithm will be stable.

In Assumption C1, we assume that the curvature of the discontinuity curve is bounded by a constant κ. A natural consequential assumption is as follows.

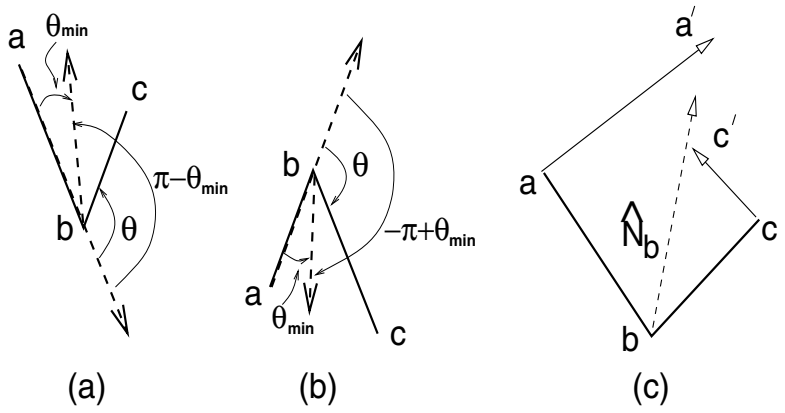


FIG. 4.1. (a) and (b) *the angle* θ *between* $\overrightarrow{ab}$ *and* $\overrightarrow{bc}$ *is bounded by* $-\pi + \theta_{min} \le \theta \le \pi - \theta_{min}$. (c) $\overrightarrow{c'a'}$ *is almost parallel to* $\hat{N}_b$.

Assumption C2. Let $\overline{ab}$ and $\overline{bc}$ be two adjoint linear segments. We assume there is a θ_{min} such that the angle θ between $\overrightarrow{ab}$ and $\overrightarrow{bc}$ is bounded by $-\pi+\theta_{min} \le \theta \le \pi-\theta_{min}$ (see Figure 4.1(a)–(b)).

Remark 5. After several propagation time steps, Assumptions R1, C1, or C2 may be violated; i.e., a linear segment may become excessively lengthened or shortened, or the angle between two adjacent linear segments may become smaller than θ_{min}. These problems can be resolved by redistributing points on the propagated grids for ensuring adequate sampling of the discontinuity grid along its arc length and eliminating any small angle. In practice, a redistribution routine [1], [13] is used to maintain these restrictions.

We want to investigate when $\hat{N}_b \times \overrightarrow{c'a'} = \epsilon$ small will happen. The case happened when the propagation time step Δt is large enough so that $\overrightarrow{c'a'}$ is almost parallel to the normal vector $\hat{N}_b$ (see Figure 4.1(c)). Therefore we need to give a restriction for Δt such that $\hat{N}_b \times \overrightarrow{c'a'}$ is larger than some constant. Lemmas 4.1 and 4.2 give the condition

$$\Delta t V \left(\frac{1}{2} + \frac{4c_2}{c_1 \theta_{min}} \right) \le \frac{1}{4} c_1 h \theta_{min}, \tag{36}$$

where c_1 and c_2 are the same as those in Assumption R1, so that the length between point b and propagation image p is less than $|\overrightarrow{ac}|/4$.

LEMMA 4.1. *Let* $\overline{ab}$ *and* $\overline{bc}$ *be two adjoint linear segments. Let* a' *be the point propagated from* a *by algorithm NPC and* c' *be the point propagated from* c *along normal to the linear segment* $\overline{bc}$ *by algorithm PB. If* $|\overline{aa'}| \le |\overline{ac}|/4$ *and* $|\overline{cc'}| \le |\overline{ac}|/4$, *then* $|\hat{N}_b \times \overrightarrow{a'c'}| \ge |\overline{ac}|/2$.

Proof. Let s_1 be the length of $|\overline{aa'}|$, s_2 the length of $|\overline{cc'}|$, θ_1 the angle between $\overrightarrow{ac}$ and $\overrightarrow{aa'}$, and θ_2 the angle between $\overrightarrow{ac}$ and $\overrightarrow{cc'}$. Since s_1, s_2 are less than or equal to $|\overline{ac}|/4$ and θ_1, θ_2 are arbitrary, this problem is equivalent to calculating that the minimum of $|\hat{N}_b \times \overrightarrow{a'c'}|$ restricted by a' is in the disk with center a and radius $|\overline{ac}|/4$ and c' is in the disk with center c and radius $|\overline{ac}|/4$.

Without loss of generality, we may assume that $a = (0, 0)$, $c = (\ell, 0)$, $\hat{N}_b = (0, 1)$, $\overrightarrow{aa'} = (s_1 \cos\theta_1, s_1 \sin\theta_1)$, and $\overrightarrow{cc'} = (s_2 \cos\theta_2 + \ell, s_2 \sin\theta_2)$. Then

$$|\hat{N}_b \times \overrightarrow{a'c'}| = |(s_2 \cos\theta_2 - s_1 \cos\theta_1 + \ell)|. \tag{37}$$

By the method of Lagrange multiplier, it is easy to find the minimum of $|\hat{N}_b \times \overrightarrow{a'c'}|$ is $\ell/2 = |\overrightarrow{ac}|/2$ when $\theta_2 = \pi$, $\theta_1 = 0$, and s_1 and s_2 are $|\overline{ac}|/4$. ∎

LEMMA 4.2. *Let $\overline{ab}$ and $\overline{bc}$ be two adjoint linear segments and p be the image of b by Algorithm NPC. If we choose Δt such that*

$$\Delta t V \left(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}\right) \le \frac{1}{4} c_1 h \theta_{min}, \tag{38}$$

then under Assumptions L1, R1, R2, C1, and C2, $|\overline{bp}| \le |\overline{ac}|/4$, where V is the maximum of the velocity field and c_1, c_2 are the same as those in Assumption R1.

Proof. Let a' be the previously propagated image of a by algorithm NPC and c' be the image of c (if c is considered as point from $\overline{bc}$; see Figure 2.1) by Algorithm BP. Let q_1' and q_2 be the mass changes produced by the movements $\overline{ab} \rightarrow a'b'$ and $\overline{bc} \rightarrow b''c'$. Then, by Remark 1,

$$q_1' = \frac{1}{2}[\overrightarrow{ab} \times (\overrightarrow{aa'} + \overrightarrow{bb'}) + \overrightarrow{bb'} \times \overrightarrow{aa'}], \quad q_2 = \frac{1}{2}[\overrightarrow{bc} \times (\overrightarrow{bb''} + \overrightarrow{cc'})]. \tag{39}$$

Similar to the proof of Theorem 2.2, λ is the unique scalar such that

$$\lambda[\hat{N}_b \times \overrightarrow{c'a'}] = 2(q_1' + q_2) + \overrightarrow{aa'} \times \overrightarrow{ab} + \overrightarrow{cc'} \times \overrightarrow{bc} = \overrightarrow{ab} \times \overrightarrow{bb'} + \overrightarrow{bb'} \times \overrightarrow{aa'} + \overrightarrow{bc} \times \overrightarrow{bb''}. \tag{40}$$

By Lemma 4.1, $|\hat{N}_b \times \overrightarrow{c'a'}| \ge |\overrightarrow{ac}|/2$; then

$$|\lambda| = \frac{|\overrightarrow{ab} \times \overrightarrow{bb'} + \overrightarrow{bb'} \times \overrightarrow{aa'} + \overrightarrow{bc} \times \overrightarrow{bb''}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \le \frac{2}{|\overrightarrow{ac}|}(|\overrightarrow{ab}||\overrightarrow{bb'}| + |\overrightarrow{bb'}||\overrightarrow{aa'}| + |\overrightarrow{bc}||\overrightarrow{bb''}|). \tag{41}$$

Since $|\overrightarrow{bb'}| \le \Delta t V$, $|\overrightarrow{bb''}| \le \Delta t V$, and by induction we may assume $|\overrightarrow{aa'}| \le |\overrightarrow{ac}|/4$, then

$$|\lambda| \le \frac{2}{|\overrightarrow{ac}|}\Delta t V(|\overrightarrow{ab}| + |\overrightarrow{aa'}| + |\overrightarrow{bc}|) \le \frac{1}{2}\Delta t V + \frac{2}{|\overrightarrow{ac}|}\Delta t V(|\overrightarrow{ab}| + |\overrightarrow{bc}|). \tag{42}$$

Since by Assumption R1, both $|\overrightarrow{ab}|$ and $|\overrightarrow{bc}|$ are greater than or equal to $c_1 h$, and by Assumption C2, the angle between $\overrightarrow{ba}$ and $\overrightarrow{bc}$ is greater than θ_{min} (see Figure 4.1(a)–(b)), we have $|\overrightarrow{ac}| \ge c_1 h \theta_{min}$. By our assumptions, $|\overrightarrow{ab}| \le c_2 h$, $|\overrightarrow{bc}| \le c_2 h$, and

$$|\lambda| \le \Delta t V \left(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}\right). \tag{43}$$

If we choose Δt such that

$$\Delta t V \left(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}\right) \le \frac{1}{4} c_1 h \theta_{min}, \tag{44}$$

we have

$$|\overrightarrow{bp}| = |\lambda \hat{N}_b| \le \Delta t V \left(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}\right) \le \frac{1}{4} c_1 h \theta_{min} \le \frac{1}{4}|\overrightarrow{ac}|. \tag{45}$$

∎

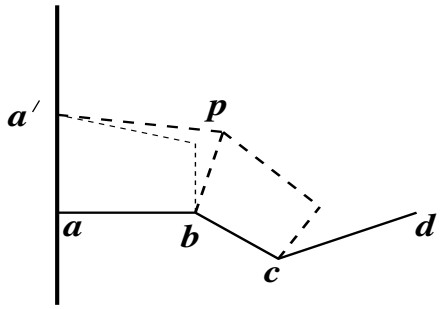


FIG. 4.2. *The propagation of the first interior point of a curve.*

Remark 6. In propagating the first interior point of a curve, we first propagate start node a to a' by a nonmass conserving propagating algorithm (see Figure 4.2). Because $|\overrightarrow{aa'}| \le \Delta tV$ and $(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}) > 1$, $|\overrightarrow{aa'}| \le \Delta tV(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}) \le \frac{1}{4}c_1h\theta_{min}$, and this implies $|\overrightarrow{bp}| \le \frac{1}{4}c_1h\theta_{min}$ (the proof of Lemma 4.2). Next, in propagating point c, we need $|\overrightarrow{bp}| \le \frac{1}{4}|\overrightarrow{bd}|$. But because $|\overrightarrow{bp}| \le \frac{1}{4}c_1h\theta_{min}$ and the angle between $\overrightarrow{cb}$ and $\overrightarrow{cd}$ is greater than θ_{min} as well as both $|\overrightarrow{bc}|$ and $|\overrightarrow{cd}|$ are greater than or equal to c_1h, we have $|\overrightarrow{bp}| \le \frac{1}{4}|\overrightarrow{bd}|$. Thus by induction it can be assumed $|\overrightarrow{aa'}| \le \frac{1}{4}c_1h\theta_{min} \le \frac{1}{4}|\overrightarrow{ac}|$ for each interior point propagation.

Next, we want to use a method similar to that in section 3 to analyze the stability of Algorithm NPC. Therefore we have to show that the length of the discontinuity curve σ^n is bounded by $|\sigma^0|(1 + \Delta t\widetilde{C})^n$ for Δt small and some constant $\widetilde{C}$ while using Algorithm NPC. Before doing this, we need to show that the difference between two propagation images by Algorithms PC and NPC is less than ΔthK for some constant K. This is done by Lemma 4.3.

LEMMA 4.3. *Let $\overline{ab}$ and $\overline{bc}$ be two adjoint linear pieces, p' be the image of b by algorithm PC, and p be the image of b by Algorithm NPC. Then under assumptions L1, R1, R2, C1, C2, and*

$$\Delta tV\left(\frac{1}{2} + \frac{4c_2}{c_1\theta_{min}}\right) \le \frac{1}{4}c_1h\theta_{min}, \tag{46}$$

$|\overline{bp} - \overline{bp'}| \le \Delta thK$ *for some constant* K.

Proof. Because propagating point b by Algorithm NPC is related to previous propagated point a', two cases for a' need to be considered. The first case is to consider a as a start node; line segment $\overline{ab}$ is always perpendicular to the boundary (see Figure 4.3(a)). This implies a can be propagated to a' by Algorithm BP and $\overrightarrow{aa'}$ is always perpendicular to $\overline{ab}$. The second case considers a as an interior point. In this case, a' is the image of a by Algorithm NPC, and $\overrightarrow{aa'}$ may not be perpendicular to $\overline{ab}$ (see Figure 4.3(b)).

Let a' be the previously propagated image of a by Algorithm NPC or Algorithm BP, let c' be the image of c by Algorithm BP, where c is considered as a point from $\overline{bc}$, and let b' and b'' be the images of b by Algorithm BP if b is considered as a point from $\overline{ab}$ and $\overline{bc}$, respectively. From (7), there exists a scalar λ such that $\overrightarrow{bp} = \lambda\hat{N}_b$

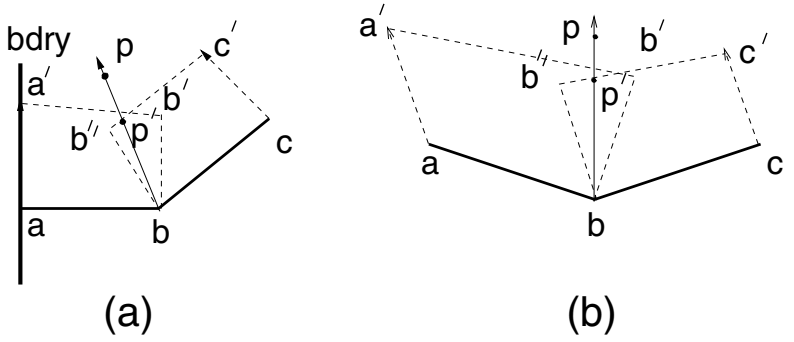


FIG. 4.3. (a) *a propagated to a' by Algorithm PC.* (b) *a propagated to a' by Algorithm NPC.*

and

$$\lambda[\hat{N}_b \times (\overrightarrow{aa'} - \overrightarrow{ab} - \overrightarrow{cc'} - \overrightarrow{bc})] = 2(q_1' + q_2) + \overrightarrow{aa'} \times \overrightarrow{ab} + \overrightarrow{cc'} \times \overrightarrow{bc}, \tag{47}$$

where q_1' and q_2 are as in Remark 1. Let $\gamma = (\Delta t \overrightarrow{v_b} \cdot \hat{N}_b)$ and $\overrightarrow{p'p} = \beta \hat{N}_b$, where β is a scalar. Since $\overrightarrow{bp} = \overrightarrow{bp'} + \overrightarrow{p'p}$ as well as $\overrightarrow{bp'} = \gamma \hat{N}_b$, $\overrightarrow{bp} = \gamma \hat{N}_b + \beta \hat{N}_b = (\gamma + \beta)\hat{N}_b$. Then (47) implies

$$(\gamma + \beta)[\hat{N}_b \times (\overrightarrow{aa'} - \overrightarrow{ab} - \overrightarrow{cc'} - \overrightarrow{bc})] = -\overrightarrow{bb'} \times \overrightarrow{ab} - \overrightarrow{bb''} \times \overrightarrow{bc}. \tag{48}$$

Case I. In this case, a' is the image of a by Algorithm BP; then $\overrightarrow{aa'} = (\Delta t \overrightarrow{v_a} \cdot \hat{N}_a)\hat{N}_a$. From (48),

$$\beta = \frac{1}{\hat{N}_b \times \overrightarrow{c'a'}}[\gamma \hat{N}_b \times \overrightarrow{cc'} - \gamma \hat{N}_b \times \overrightarrow{aa'} + (\gamma \hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab} + (\gamma \hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}]. \tag{49}$$

Hence

$$|\beta| \le \frac{|\gamma \hat{N}_b \times \overrightarrow{cc'}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} + \frac{|\gamma \hat{N}_b \times \overrightarrow{aa'}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} + \frac{|(\gamma \hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} + \frac{|(\gamma \hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}|}{|\hat{N}_b \times \overrightarrow{c'a'}|}. \tag{50}$$

We want to estimate the values $|\gamma \hat{N}_b \times \overrightarrow{cc'}|$, $|\gamma \hat{N}_b \times \overrightarrow{cc'}|$, $|(\gamma \hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab}|$, and $|(\gamma \hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}|$ separately. Now

$$|\gamma \hat{N}_b \times \overrightarrow{cc'}| = |\gamma||\overrightarrow{cc'}||\hat{N}_b \times \hat{n}_{bc}| = (\Delta t \overrightarrow{v_a} \cdot \hat{N}_a)(\Delta t \overrightarrow{v_a} \cdot \hat{N}_a) \sin\theta, \tag{51}$$

where θ is the angle between $\hat{N}_a$ and $\hat{n}_{bc}$. By Assumption R1, κ is the maximum curvature of all σ^n; therefore

$$\sin\theta \approx \theta \le \kappa |\overrightarrow{ac}|. \tag{52}$$

Then

$$|\gamma \hat{N}_b \times \overrightarrow{cc'}| \le (\Delta t \overrightarrow{v}_a)(\Delta t \overrightarrow{v}_c) \sin\theta \le (\Delta t V)^2 \kappa |\overrightarrow{ac}|, \tag{53}$$

where V is the maximum velocity on velocity field. By Assumption R2, $\Delta t \leq \alpha h$, and let $\alpha V^2 \kappa = K_1$. Then

$$|\gamma \hat{N}_b \times \overrightarrow{cc'}| \leq (\Delta t V)^2 \kappa |\overrightarrow{ac}| \leq \Delta t h K_1 |\overrightarrow{ac}|. \tag{54}$$

From Lemma 4.1, we have $|\hat{N}_b \times \overrightarrow{c'a'}| \geq \frac{1}{2}|\overrightarrow{ac}|$. Therefore

$$\frac{|\gamma \hat{N}_b \times \overrightarrow{cc'}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \leq \Delta t h K_1. \tag{55}$$

Similarly,

$$\frac{|\gamma \hat{N}_b \times \overrightarrow{aa'}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \leq \Delta t h K_2 \tag{56}$$

for some constant K_2. Next,

$$|(\gamma \hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab}| = |\gamma \hat{N}_b - \overrightarrow{bb'}||\overrightarrow{ab}||\sin\phi|, \tag{57}$$

where ϕ is the angle between $\gamma \hat{N}_b - \overrightarrow{bb'}$ and $\overrightarrow{ab}$. Since $|\sin\phi| \leq 1$ and $|\overrightarrow{ab}| \leq c_2 h$ by Assumption R1, we also have

$$\begin{aligned} |\gamma \hat{N}_b - \overrightarrow{bb'}| &= |(\Delta t \overrightarrow{v_b} \cdot \hat{N}_b)\hat{N}_b - (\Delta t \overrightarrow{v_b} \cdot \hat{n}_{ab})\hat{n}_{ab}| \\ &\leq |(\Delta t \overrightarrow{v_b} \cdot \hat{N}_b)\hat{N}_b - (\Delta t \overrightarrow{v_b} \cdot \hat{N}_b)\hat{n}_{ab}| + |(\Delta t \overrightarrow{v_b} \cdot \hat{N}_b)\hat{n}_{ab} - (\Delta t \overrightarrow{v_b} \cdot \hat{n}_{ab})\hat{n}_{ab}| \\ &\leq ||(\Delta t \overrightarrow{v_b} \cdot \hat{N}_b)||\hat{N}_b - \hat{n}_{ab}| + |(\Delta t \overrightarrow{v_b})||\hat{N}_b - \hat{n}_{ab}| \leq 2\Delta t V |\hat{N}_b - \hat{n}_{ab}|. \end{aligned} \tag{58}$$

Since $|\hat{N}_b - \hat{n}_{ab}| \approx \theta \leq \kappa c_2 h$, then we have

$$|\gamma \hat{N}_b - \overrightarrow{bb'}| \leq 2\Delta t V |\hat{N}_b - \hat{n}_{ab}| \leq 2\Delta t V \kappa c_2 h. \tag{59}$$

Because $|\hat{N}_b \times \overrightarrow{c'a'}| \geq \frac{1}{2}|\overrightarrow{ac}|$ and $|\overrightarrow{ac}| \geq c_1 h \theta_{min}$,

$$\frac{|(\gamma \hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \leq \frac{|\gamma \hat{N}_b - \overrightarrow{bb'}||\overrightarrow{ab}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \leq \frac{4\Delta t V \kappa c_2 h c_2 h}{c_1 h \theta_{min}} \leq \frac{4\Delta t V \kappa c_2 h c_2}{c_1 \theta_{min}}. \tag{60}$$

Let

$$K_3 = \frac{4V\kappa c_2 c_2}{c_1 \theta_{min}};$$

then

$$\frac{|(\gamma \hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \leq \Delta t h K_3. \tag{61}$$

With a similar method, we have

$$\frac{|(\gamma \hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} \leq \Delta t h K_4. \tag{62}$$

Therefore

(63)
$$|\overline{bp}-\overline{bp'}| = |\beta| \le \frac{|\gamma\hat{N}_b \times \overrightarrow{cc'}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} + \frac{|\gamma\hat{N}_b \times \overrightarrow{aa'}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} + \frac{|(\gamma\hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab}|}{|\hat{N}_b \times \overrightarrow{c'a'}|} + \frac{|(\gamma\hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}|}{|\hat{N}_b \times \overrightarrow{c'a'}|}$$
$$\le \Delta thK_1 + \Delta thK_2 + \Delta thK_3 + \Delta thK_4 = \Delta thK.$$

Case II. Consider a' to be the image of a by Algorithm NPC. Then $\overrightarrow{aa'} = (\gamma' + \beta')\hat{N}_a$, where $\gamma' = \Delta t\overrightarrow{v_a}\cdot\hat{N}_a$ and $|\beta'| \le \Delta thK$ for some constant K. Let $\xi = \Delta t\overrightarrow{v_c}\cdot\hat{N}_c$. Similarly, we have

$$\lambda[\hat{N}_b \times (\overrightarrow{aa'} - \overrightarrow{ab} - \overrightarrow{cc'} - \overrightarrow{bc})] = 2(q_1' + q_2) + \overrightarrow{aa'} \times \overrightarrow{ab} + \overrightarrow{cc'} \times \overrightarrow{bc}, \tag{64}$$

which leads to

$$\beta(\hat{N}_b \times \overrightarrow{c'a'}) = -\gamma(\gamma' + \beta')(\hat{N}_b \times \hat{N}_a) + \gamma\xi(\hat{N}_b \times \hat{n_{bc}}) + (\gamma\hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab} + (\gamma\hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}. \tag{65}$$

Therefore

(66)
$$|\beta| \le \frac{1}{|(\hat{N}_b \times \overrightarrow{c'a'})|}|\gamma\gamma'(\hat{N}_b \times \hat{N}_a) + \gamma\xi(\hat{N}_b \times \hat{n_{bc}}) + (\gamma\hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab} + (\gamma\hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}|$$
$$+ \frac{1}{|(\hat{N}_b \times \overrightarrow{c'a'})|}|\gamma\beta'(\hat{N}_b \times \hat{N}_a)|.$$

By Case I of this proof, we have

(67)
$$\frac{1}{|(\hat{N}_b \times \overrightarrow{c'a'})|}|\gamma\gamma'(\hat{N}_b \times \hat{N}_a) + \gamma\xi(\hat{N}_b \times \hat{n_{bc}}) + (\gamma\hat{N}_b - \overrightarrow{bb'}) \times \overrightarrow{ab} + (\gamma\hat{N}_b - \overrightarrow{bb''}) \times \overrightarrow{bc}|$$
$$\le \Delta thK$$

for some constant K. Since $\hat{N}_b \times \hat{N}_a = \sin\theta$, where θ is the angle between $\hat{N}_b$ and $\hat{N}_a$, then

$$\sin\theta \approx \theta \le \kappa c_2 h. \tag{68}$$

Hence

$$\frac{1}{|(\hat{N}_b \times \overrightarrow{c'a'})|}|\gamma\beta'(\hat{N}_b \times \hat{N}_a)| \le \Delta t^2 hK. \tag{69}$$

Choose $\Delta t \le 1$; then

$$\frac{1}{|(\hat{N}_b \times \overrightarrow{c'a'})|}|\gamma\beta'(\hat{N}_b \times \hat{N}_a)| \le \Delta t^2 hK \le \Delta thK. \tag{70}$$

Therefore

$$|\beta| \le \Delta thK. \qquad \square \tag{71}$$

LEMMA 4.4. *Let σ^n be the discontinuity curve at the nth time step by Algorithm NPC. Then under Assumptions L1, R1, R2, C1, C2, and condition*

$$\Delta t V\left(\frac{1}{2}+\frac{4c_2}{c_1\theta_{min}}\right)\le\frac{1}{4}c_1h\theta_{min},$$

$|\sigma^n|\le|\sigma^0|(1+\Delta t\widetilde{C})^n$ *for some constant* $\widetilde{C}$.

Proof. From Lemma 4.2, we choose Δt small such that

$$\Delta t V\left(\frac{1}{2}+\frac{4c_2}{c_1\theta_{min}}\right)\le\frac{1}{4}c_1h\theta_{min}$$

for every propagation time step, where c_1, c_2 are the same as in Assumption R1 and θ_{min} is the same as in Assumption C2.

Let $\overrightarrow{ab}=b_i^n$, $\hat{N}_a$ be the normal to the curve at a and $\widehat{N}_b$ be the normal to the curve at b. Then

$$b_i^{n+1}=-(\gamma_a'+\beta_a')\hat{N}_a+b_i^n+(\gamma_b'+\beta_b')\hat{N}_b, \tag{72}$$

where $\gamma_a'=(\Delta t\overrightarrow{v_a}\hat{N}_a)$, $\gamma_b'=(\Delta t\overrightarrow{v_b}\hat{N}_b)$, $|\beta_a'|\le\Delta thK$, and $|\beta_b'|\le\Delta thK$. By triangle inequality,

$$|b_i^{n+1}|\le|-\gamma_a'\hat{N}_a+\overrightarrow{b_i^n}+\gamma_b'\hat{N}_b|+|\beta_a'\hat{N}_a+\beta_b'\hat{N}_b|, \tag{73}$$

and by Lemma 3.2,

$$|-\gamma_a'\hat{N}_a+b_i^n+\gamma_b'\hat{N}_b|\le|b_i^n|(1+\Delta t(2V\kappa+C)), \tag{74}$$

where V is the maximum of the velocity field, κ is the maximum curvature of all discontinuity curves, and C is a constant. Hence

$$|b_i^{n+1}|\le|b_i^n|(1+\Delta t(2V\kappa+C))+2\Delta thK. \tag{75}$$

Now

$$\begin{aligned}|\sigma^{n+1}|=\sum_{i=1}^{m}|b_i^{n+1}|&\le\sum_{i=1}^{m}|b_i^n|(1+\Delta t(2V\kappa+C))+\sum_{i=1}^{m}2\Delta thK\\&\le|\sigma^n|(1+\Delta t(2V\kappa+C))+m2\Delta thK,\end{aligned} \tag{76}$$

where m is the number of pieces in σ^n. Since the length of every linear piece is no less than c_1h, then $m\le|\sigma^n|/c_1h$. Therefore

$$\begin{aligned}|\sigma^{n+1}|&\le|\sigma^n|(1+\Delta t(2V\kappa+C))+\frac{|\sigma^n|}{c_1h}2\Delta thK\le|\sigma^n|(1+\Delta t(2V\kappa+C))+|\sigma^n|\frac{2}{c_1}\Delta tK\\&=|\sigma^n|\left(1+\Delta t\left(2V\kappa+C+\frac{2}{c_1}K\right)\right).\end{aligned} \tag{77}$$

Let $\widetilde{C}=2V\kappa+C+\frac{2}{c_1}K$; then

$$|\sigma^{n+1}|\le|\sigma^n|(1+\Delta t\widetilde{C}). \tag{78}$$

By induction,

$$|\sigma^{n+1}| \le |\sigma^0|(1+\Delta t\widetilde{C})^n. \qquad \square \tag{79}$$

THEOREM 4.5. *Under Assumptions L1, R1, R2, C1, C2, and condition*

$$\Delta t V\left(\frac{1}{2}+\frac{4c_2}{c_1\theta_{min}}\right) \le \frac{1}{4}c_1 h\theta_{min},$$

the curve propagation algorithm, Algorithm NPC, of the front-tracking scheme for incompressible miscible two phase flow on porous media is stable.

Proof. The proof is similar to those of Lemmas 3.4 and 3.5 and Theorem 3.6. We omit it here. $\square$

Remark 7. A condition is required in Theorem 4.5 that $\Delta t V(\frac{1}{2}+\frac{4c_2}{c_1\theta_{min}}) \le \frac{1}{4}c_1 h\theta_{min}$. Because V, c_1, c_2, h, and θ_{min} are constants, $\Delta t V(\frac{1}{2}+\frac{4c_2}{c_1\theta_{min}}) \le \frac{1}{4}c_1 h\theta_{min}$ will be satisfied if the time step Δt is chosen to be small enough.

Acknowledgment. I wish to thank the anonymous referee for his/her due diligence in editing this paper. This resulted in a new understanding and clarified my thinking about this topic.

REFERENCES

[1] K. A. CHANG AND W. B. LINDQUIST, *Mass-conserving front tracking for miscible two-phase flow*, SIAM J. Sci. Comput., 18 (1997), pp. 1310–1327.

[2] K. A. CHANG, *An error analysis of front-tracking scheme for miscible two phase flow*, Int. Math. J., 4 (2003), pp. 247–263.

[3] X. GARAIZAR, J. GLIMM, AND W. GUO, *Elastic deformation and slug flow as applications of front tracking*, in Transactions of the Seventh Army Conference on Applied Mathematics and Computing, Army Research Office, Research Triangle Park, NC, 1990, pp. 705–717.

[4] J. GLIMM, J. W. GROVE, X. L. LI, AND N. ZHAO, *Simple front tracking*, in Nonlinear Partial Differential Equations, AMS, Providence, RI, 1999, pp. 133–149.

[5] J. GLIMM, E. ISAACSON, D. MARCHESIN, AND O. MCBRYAN, *Front tracking for hyperbolic systems*, Adv. in Appl. Math., 2 (1981), pp. 91–119.

[6] J. GLIMM, W. B. LINDQUIST, O. MCBRYAN, AND L. PADMANHABAN, *A front tracking reservoir simulator: Five-spot validation studies and the water coning problem*, in The Mathematics of Reservoir Simulation, Frontiers Appl. Math. 1, R. E. Ewing, ed., SIAM, Philadelphia, 1984, pp. 107–135.

[7] J. GLIMM, W. B. LINDQUIST, AND Q. ZHANG, *Front tracking, oil reservoirs, engineering problems, and mass conservation*, in Multidimensional Hyperbolic Problems and Computations, IMA Vol. Math. Appl. 29, Springer, New York, 1991, pp. 123–139.

[8] J. GLIMM, D. MARCHESIN, AND O. MCBRYAN, *Unstable fingers in two phase flow,* Comm. Pure Appl. Math., 34 (1981), pp. 53–75.

[9] J. GLIMM, D. MARCHESIN, AND O. MCBRYAN, *A numerical method for two phase flow with an unstable interface*, J. Comput. Phys., 39 (1981), pp. 179–200.

[10] J. GLIMM AND O. MCBRYAN, *A computational model for interfaces*, Adv. in Appl. Math., 6 (1985), pp. 422–435.

[11] J. GLIMM, X. LI, Y. LIU, Z. XU, AND N. ZHAO, *Conservative front tracking with improved accuracy*, SIAM J. Numer. Anal., 41 (2003), pp. 1926–1947.

[12] J. GROVE, D. H. SHARP, Y. YANG, AND Q. ZHANG, *Quantitative theory of Richtmyer–Meshkov instability,* Phys. Rev. Lett., 71 (1993), pp. 3473–3476.

[13] C. KLINGENBERG AND B. PLOHR, *An introduction to front tracking*, in Multidimensional Hyperbolic Problems and Computations, IMA Vol. Math. Appl. 29, Springer, New York, 1991, pp. 203–216.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1064–1080

FUNCTION, GRADIENT, AND HESSIAN RECOVERY USING QUADRATIC EDGE-BUMP FUNCTIONS*

JEFFREY S. OVALL†

Abstract. An approximate error function for the discretization error on a given mesh is obtained by projecting (via the energy inner product) the functional residual onto the space of continuous, piecewise quadratic functions which vanish on the vertices of the mesh. Conditions are given under which one can expect this hierarchical basis error estimator to give efficient and reliable function recovery, asymptotically exact gradient recovery, and convergent Hessian recovery in the square norms. One does not find similar function recovery results in the literature. The analysis given here is based on a certain superconvergence result which has been used elsewhere in the analysis of gradient recovery methods. Numerical experiments are provided which demonstrate the effectivity of the approximate error function in practice.

Key words. finite elements, a posteriori estimates, hierarchical bases, superconvergence, gradient recovery

AMS subject classifications. 65N15, 65N30, 65N50

DOI. 10.1137/060648908

1. Introduction.

Hierarchical basis a posteriori error estimators were introduced in the early 1980s [22], and a general framework for the analysis of their effectivity and computational cost has been given by Bank [5] and Bank and Smith [1]. The basic idea behind such methods is that the base space of functions V_h, in which we wish to find our finite element approximation u_h, is augmented by a complementary space $\tilde{V}_h$ such that the composite space $V_h \oplus \tilde{V}_h$ provides an improved finite element approximation $\bar{u}_h$. In symbols, we show this as $|||u - \bar{u}_h||| \leq \beta |||u - u_h|||$ for some $\beta \in [0, 1)$, where $|||\cdot|||$ is the energy norm associated with the underlying bilinear form. This improved approximation assumption is referred to as a saturation assumption. An approximate error function $\varepsilon_h \approx u - u_h$ is computed in the space $\tilde{V}_h$. Using the saturation assumption and strengthened Cauchy inequalities between the spaces V_h and $\tilde{V}_h$, effectivity estimates of the form

$$c_1 \leq \frac{|||\varepsilon_h|||}{|||u - u_h|||} \leq c_2 \tag{1}$$

are proved.

In this paper a different sort of analysis, which yields stronger assertions, is given for the case where V_h is the space of continuous, piecewise linear functions on a given mesh and $\bar{V}_h$ is the space of continuous, piecewise quadratic functions on that same mesh. The augmenting space $\tilde{V}_h$ consists of quadratic "bump" functions which vanish on the vertices of the mesh. In particular, we show that the approximate error function, $\varepsilon_h \approx u - u_h$, provides efficient and reliable function recovery, asymptotically exact gradient recovery, and convergent Hessian recovery:

$$c_1 \leq \frac{\|\varepsilon_h\|_{0,\Omega}}{\|u - u_h\|_{0,\Omega}} \leq c_2, \qquad \frac{\|\varepsilon_h\|_{1,\Omega}}{\|u - u_h\|_{1,\Omega}} \to 1, \qquad \sum_{\tau \in \mathcal{T}_h} |\varepsilon_h|^2_{2,\tau} \to |u|^2_{2,\Omega}. \tag{2}$$

*Received by the editors January 3, 2006; accepted for publication (in revised form) January 8, 2007; published electronically May 7, 2007.
http://www.siam.org/journals/sinum/45-3/64890.html
†Max Planck Institute for Mathematics in the Sciences, Inselstraße 22-26, D-04103 Leipzig, Germany (ovall@mis.mpg.de).

Our analysis is based on a superconvergence result of Bank and Xu [3, 4], which also appears in a slightly more general form in [20]. This result was used in these papers to explain the success of a number of popular gradient recovery methods, but we use it here in the context of hierarchical basis error estimation to establish our key approximation results (2).

The rest of this paper is organized as follows. In section 2 we lay out the basic notation and assumptions for this paper. Section 3 contains a statement of the superconvergence result of Bank and Xu, which we then use to prove the above mentioned gradient and Hessian recovery results. In section 4 we prove the function recovery result and show why we cannot generally hope for asymptotic exactness in this case. Section 5 comprises almost half of the paper and consists of four examples, which are used to verify the effectivity of our estimator in practice, and a brief subsection on computational cost.

2. Notation and basic assumptions. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$, and define

$$\mathcal{H} \equiv \left\{ v \in H^1(\Omega) : v|_{\partial\Omega_D} = 0 \text{ in the trace sense} \right\}. \tag{3}$$

The usual spaces $W_p^k(\Omega)$ and $H^k(\Omega) \equiv W_2^k(\Omega)$ are equipped with their standard norms $\|\cdot\|_{k,p,\Omega}$ and $\|\cdot\|_{k,\Omega} \equiv \|\cdot\|_{k,2,\Omega}$, and seminorms $|\cdot|_{k,p,\Omega}$ and $|\cdot|_{k,\Omega}$, respectively. For simplicity in exposition, we will assume that $\partial\Omega$ is a polygon. Let data functions $a : \bar{\Omega} \to \mathbb{R}^{2\times 2}$, $\mathbf{b} : \bar{\Omega} \to \mathbb{R}^2$, $c, f : \bar{\Omega} \to \mathbb{R}$, and $g : \partial\Omega_N \to \mathbb{R}$ be given. The problem is to find $u \in \mathcal{H}$ such that

$$B(u, v) = F(v) \text{ for all } v \in \mathcal{H}, \tag{4}$$

$$B(u, v) \equiv \int_\Omega a\nabla u \cdot \nabla v + (\mathbf{b} \cdot \nabla u + cu)v \, dx, \tag{5}$$

$$F(v) \equiv \int_\Omega fv \, dx + \int_{\partial\Omega_N} gv \, ds. \tag{6}$$

We will assume that the data functions are sufficiently smooth, and that the matrix a is positive definite, with the smallest eigenvalue bounded below on Ω by some constant $\gamma > 0$. We make the following standard assumptions concerning the bilinear form B and linear functional F: There exist constants $\alpha, \nu, \mu > 0$, such that, for all $v, w \in \mathcal{H}$,

$$\begin{aligned} |F(v)| &\le \alpha \|v\|_{1,\Omega}, \\ |B(v, w)| &\le \nu \|v\|_{1,\Omega} \|w\|_{1,\Omega}, \\ B(v, v) &\ge \mu \|v\|_{1,\Omega}^2. \end{aligned}$$

Let $\mathcal{T}_h$ denote a shape-regular triangulation of Ω with mesh size $h \in (0, 1)$. Let $V_h \subset \mathcal{H}$ denote the space of continuous, piecewise-linear polynomials defined on $\mathcal{T}_h$, and let $\bar{V}_h \subset \mathcal{H}$ denote the continuous, piecewise-quadratic polynomials. We will think of $\bar{V}_h$ hierarchically as

$$\bar{V}_h = V_h \oplus \tilde{V}_h, \tag{7}$$

where $\tilde{V}_h$ is the space of quadratic "bump" functions, i.e., continuous piecewise-quadratic polynomials which vanish at all of the vertices of the triangulation. In what follows, $u_h \in V_h$ and $\bar{u}_h \in \bar{V}_h$ denote, respectively, the piecewise-linear and

piecewise-quadratic approximate solutions of (4):

$$B(u_h, v) = F(v) \text{ for all } v \in V_h, \tag{8}$$

$$B(\bar{u}_h, v) = F(v) \text{ for all } v \in \bar{V}_h. \tag{9}$$

Let $u_\ell \in V_h$ and $u_q \in \bar{V}_h$ denote piecewise-linear and piecewise-quadratic interpolants of u on $\mathcal{T}_h$. We make the following standard assumptions about their asymptotic approximation quality:

$$\|u - u_\ell\|_{k,\Omega} \lesssim h^{2-k}\|u\|_{2,\Omega}, \tag{10}$$

$$\|u - u_q\|_{k,\Omega} \lesssim h^{3-k}\|u\|_{3,\Omega}, \tag{11}$$

for $0 \le k \le 1$.

3. Gradient and Hessian recovery. In this section we prove asymptotically exact gradient recovery and convergent Hessian recovery results,

$$\frac{\|\varepsilon_h\|_{1,\Omega}}{\|u - u_h\|_{1,\Omega}} \to 1, \qquad \sum_{\tau \in \mathcal{T}_h} |\varepsilon_h|^2_{2,\tau} \to |u|^2_{2,\Omega} \tag{12}$$

for the approximate error function $\varepsilon_h \approx u - u_h$ described below. We first describe the key assumption on the mesh that will play a role in these results. This mesh condition and a slight generalization of it can be found in [3, 20].

Let e denote an interior edge in $\mathcal{T}_h$ with adjacent triangles τ and τ'. We say that the quadrilateral formed by τ and τ' satisfies the *approximate $O(h^2)$-parallelogram property* provided that the lengths of opposite edges differ by only $O(h^2)$. The equivalent property at the boundary is as follows: Let e and e' denote adjacent boundary edges sharing the vertex x, and let τ and τ' be the triangles having the edges e and e', respectively. Let $\mathbf{t}$ and $\mathbf{t}'$ be the unit tangent vectors, corresponding to a counterclockwise orientation on τ and τ'. Starting with e for τ and e' for τ' we identify corresponding edges of τ and τ' by traversing their edges counterclockwise. We say that the triangles τ and τ' associated with the boundary vertex x satisfy the approximate $O(h^2)$-parallelogram property, provided that the lengths of corresponding edges in τ and τ' differ by only $O(h^2)$, and $|\mathbf{t} - \mathbf{t}'| = O(h)$. The key assumption on the triangulation follows.

ASSUMPTION 3.1 (an $O(h^{2\sigma})$-irregular triangulation).

1. *Let $\mathcal{E} = \mathcal{E}_1 \oplus \mathcal{E}_2$ denote the set of interior edges in $\mathcal{T}_h$. For each $e \in \mathcal{E}_1$, τ and τ' satisfy the approximate $\mathcal{O}(h^2)$-parallelogram property, while*

$$\sum_{e \in \mathcal{E}_2} |\tau| + |\tau'| = \mathcal{O}(h^{2\sigma}).$$

2. *Let $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{P}_2$ denote the set of boundary vertices. The elements associated with $x \in \mathcal{P}_1$ satisfy the approximate $\mathcal{O}(h^2)$-parallelogram property, and $|\mathcal{P}_2| = \kappa$, where κ is fixed independent of h.*

The second condition is necessary only in the case of Neumann boundary conditions, $\partial\Omega_N \neq \emptyset$. The following result, due to Bank and Xu [3], is the key lemma for the results in this paper.

LEMMA 3.2. *Under Assumption* 3.1*, we have*

$$\|u_h - u_\ell\|_{1,\Omega} \lesssim h^{1+\min(\sigma,1)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}. \tag{13}$$

We now present a new result based on Lemma 3.2 for computing a superconvergent approximation of the gradient. Suppose that we first solve for the linear finite element approximation, $u_h \in V_h$, and then augment this approximation by solving the residual equation on $\tilde{V}_h$, the space of quadratic bumps. In other words,

$$B(u_h, v) = F(v) \text{ for all } v \in V_h, \tag{14}$$

$$B(\varepsilon_h, v) = F(v) - B(u_h, v) \text{ for all } v \in \tilde{V}_h. \tag{15}$$

One can think of this as a projection of the residual error onto the space $\tilde{V}_h$. We have the following result.

THEOREM 3.3. *Under Assumption* 3.1*, we have*

$$\|u - (u_h + \varepsilon_h)\|_{1,\Omega} \lesssim h^{1+\min(\sigma,1)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}. \tag{16}$$

Proof. Using Galerkin orthogonality to replace $\varepsilon_h \in \tilde{V}_h$ with $u_b \in \tilde{V}_h$, the "bump" portion of the quadratic interpolant $u_q = u_\ell + u_b$, we get the following estimate:

$$\|u - (u_h + \varepsilon_h)\|_{1,\Omega}^2 \lesssim B(u - (u_h + \varepsilon_h), u - (u_h + \varepsilon_h)) \tag{17}$$

$$= B(u - (u_h + \varepsilon_h), u - (u_h + u_b)) \tag{18}$$

$$\lesssim \|u - (u_h + \varepsilon_h)\|_{1,\Omega} \|u - (u_h + u_b)\|_{1,\Omega}. \tag{19}$$

We bound the term $\|u - (u_h + u_b)\|_{1,\Omega}$ as follows:

$$\|u - (u_h + u_b)\|_{1,\Omega} \le \|u - u_q\|_{1,\Omega} + \|u_q - (u_h + u_b)\|_{1,\Omega} \tag{20}$$

$$= \|u - u_q\|_{1,\Omega} + \|u_\ell - u_h\|_{1,\Omega} \tag{21}$$

$$\lesssim h^2 \|u\|_{3,\Omega} + h^{1+\min(\sigma,1)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}. \tag{22}$$

This completes the proof. ∎

As an immediate corollary, we see conditions under which we can expect $\|\varepsilon_h\|_{1,\Omega}$ to be an asymptotically exact estimator of the true gradient error $\|u - u_h\|_{1,\Omega}$.

COROLLARY 3.4. *Suppose that there is some constant $c > 0$ such that $\|u - u_h\|_{1,\Omega} \ge ch$. Then under Assumption* 3.1*, we have*

$$\frac{\|\varepsilon_h\|_{1,\Omega}}{\|u - u_h\|_{1,\Omega}} \to 1. \tag{23}$$

Proof. It holds that

$$\left| \frac{\|\varepsilon_h\|_{1,\Omega}}{\|u - u_h\|_{1,\Omega}} - 1 \right| \le \frac{\|u - (u_h + \varepsilon_h)\|_{1,\Omega}}{\|u - u_h\|_{1,\Omega}}. \tag{24}$$

Combining this with the estimate from Theorem 3.3 completes the proof. ∎

Theorem 3.3 and Corollary 3.4 and their proofs have also appeared in [17, 18].

Recall that the quadratic interpolant $u_q \in \bar{V}_h$ of u is decomposed as the sum $u_q = u_\ell + u_b$ with $u_\ell \in V_h$ and $u_b \in \tilde{V}_h$. In the following lemma we compare the first and second derivatives of ε_h and u_b. The second of these results is used in the proof of Theorem 3.6 to establish the Hessian recovery result, and the first will play an important role in the next section, where we prove the function recovery result.

LEMMA 3.5. *Under Assumption* 3.1*, we have*

$$\|\varepsilon_h - u_b\|_{1,\Omega} \lesssim h^{1+\min(\sigma,1)} |\log h|^{1/2} \|u\|_{3,\infty,\Omega}, \tag{25}$$

$$\sum_{\tau \in \mathcal{T}_h} |\varepsilon_h - u_b|_{2,\tau}^2 \lesssim h^{2\min(\sigma,1)} |\log h| \|u\|_{3,\infty,\Omega}^2. \tag{26}$$

Proof. In the proof of Theorem 3.3, we saw that

$$\|u-(u_h+\varepsilon_h)\|_{1,\Omega},\ \|u-(u_h+u_b)\|_{1,\Omega} \lesssim h^{1+\min(\sigma,1)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}. \tag{27}$$

This gives us

$$\|\varepsilon_h-u_b\|_{1,\Omega} \le \|u-(u_h+\varepsilon_h)\|_{1,\Omega}+\|u-(u_h+u_b)\|_{1,\Omega} \tag{28}$$

$$\lesssim h^{1+\min(\sigma,1)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}. \tag{29}$$

Using a standard inverse estimate, we see that

$$\sum_{\tau\in\mathcal{T}_h}|\varepsilon_h-u_b|_{2,\tau}^2 \lesssim h^{-2}\|\varepsilon_h-u_b\|_{1,\Omega}^2 \lesssim h^{2\min(1,\sigma)}\,|\log h|\,\|u\|_{3,\infty,\Omega}^2, \tag{30}$$

so we have proven both results. ∎

The convergent Hessian recovery result follows.

THEOREM 3.6. *Under Assumption* 3.1*, we have*

$$\sum_{\tau\in\mathcal{T}_h}|u-\varepsilon_h|_{2,\tau}^2 \lesssim h^{2\min(\sigma,1)}|\log h|\|u\|_{3,\infty,\Omega}^2. \tag{31}$$

Proof. We have $|u-\varepsilon_h|_{2,\tau} \le |u-u_b|_{2,\tau}+|u_b-\varepsilon_h|_{2,\tau}$, so

$$\sum_{\tau\in\mathcal{T}_h}|u-\varepsilon_h|_{2,\tau}^2 \le 2\left(\sum_{\tau\in\mathcal{T}_h}|u-u_b|_{2,\tau}^2+\sum_{\tau\in\mathcal{T}_h}|u_b-\varepsilon_h|_{2,\tau}^2\right) \tag{32}$$

$$\lesssim h^2\|u\|_{3,\infty,\Omega}^2+\sum_{\tau\in\mathcal{T}_h}|u_b-\varepsilon_h|_{2,\tau}^2. \tag{33}$$

Combining this with the second estimate in Lemma 3.5 completes the proof. ∎

Provided that $\|u\|_{3,\infty,\Omega}<\infty$, the estimate in Theorem 3.5 is equivalent to

$$\sum_{\tau\in\mathcal{T}_h}|\varepsilon_h|_{2,\tau}^2 \to |u|_{2,\Omega}^2. \tag{34}$$

4. Function recovery. In this section we prove that the approximate error function ε_h provides efficient and reliable approximation of the true error $u-u_h$ in the L_2-norm,

$$c_1 \le \frac{\|\varepsilon_h\|_{0,\Omega}}{\|u-u_h\|_{0,\Omega}} \le c_2. \tag{35}$$

We also explain why we cannot generally expect the same sort of asymptotic exactness result which we saw for the gradient error. In other words, we *cannot* generally expect that

$$\frac{\|\varepsilon_h\|_{0,\Omega}}{\|u-u_h\|_{0,\Omega}} \to 1, \tag{36}$$

although the constants c_1, c_2 may be near 1 in practice.

This first lemma will allow us to convert the gradient approximation result from Lemma 3.5 into the function (L_2) approximation results that follow.

LEMMA 4.1. *Let $\mathcal{T}_h$ be a shape-regular quasi-uniform mesh. For any $b \in \tilde{V}_h$, we have*

$$\|b\|_{0,\Omega} \lesssim h\|\nabla b\|_{0,\Omega}. \tag{37}$$

Proof. Let $\tau \in \mathcal{T}_h$ be given, and write b in terms of its three bump basis functions on τ, $b = c_1 b_1 + c_2 b_2 + c_3 b_3$. We denote the length of the edge on which b_k does *not* vanish by L_k, and without loss of generality take $L_1 \le L_2 \le L_3$. It holds that

$$\|b\|_{0,\tau}^2 = \frac{8|\tau|}{45}(c_1^2 + c_2^2 + c_3^2 + c_1c_2 + c_1c_3 + c_2c_3), \tag{38}$$

$$\|\nabla b\|_{0,\tau}^2 = \frac{1}{3|\tau|}\left((c_1 - c_2 - c_3)^2 L_1^2 + (c_2 - c_1 - c_3)^2 L_2^2 + (c_3 - c_1 - c_2)^2 L_3^2\right). \tag{39}$$

We bound $\|\nabla b\|_{0,\tau}^2$ from below as follows:

$$\|\nabla b\|_{0,\tau}^2 \ge \frac{L_1^2}{3|\tau|}\left((c_1 - c_2 - c_3)^2 + (c_2 - c_1 - c_3)^2 + (c_3 - c_1 - c_2)^2\right) \tag{40}$$

$$= \frac{L_1^2}{3|\tau|}(3c_1^2 + 3c_2^2 + 3c_3^2 - 2c_1c_2 - 2c_1c_3 - 2c_2c_3) \tag{41}$$

$$\ge \frac{L_1^2}{3|\tau|}\frac{1}{2}(c_1^2 + c_2^2 + c_3^2 + c_1c_2 + c_1c_3 + c_2c_3). \tag{42}$$

This gives us

$$\|b\|_{0,\tau}^2 \le \frac{48}{45}\frac{|\tau|^2}{L_1^2}\|\nabla b\|_{0,\tau}^2 \lesssim h^2\|\nabla b\|_{0,\tau}^2. \tag{43}$$

Summing over triangles completes the proof. ∎

LEMMA 4.2. *Under Assumption* 3.1*, we have*

$$\|\varepsilon_h - u_b\|_{0,\Omega} \lesssim h^{2+\min(\sigma,1)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}, \tag{44}$$

$$\|u - (u_\ell + \varepsilon_h)\|_{0,\Omega} \lesssim h^{2+\min(\sigma,1)}|\log h|^{1/2}\|u\|_{3,\infty,\Omega}. \tag{45}$$

Proof. Combining the first estimate from Lemma 3.5 with the result of Lemma 4.1 proves the first of these two estimates. We also have

$$\|u - (u_\ell + \varepsilon_h)\|_{0,\Omega} \le \|u - u_q\|_{0,\Omega} + \|\varepsilon_h - u_b\|_{0,\Omega} \lesssim h^3\|u\|_{3,\Omega} + \|\varepsilon_h - u_b\|_{0,\Omega}. \tag{46}$$

Combining this second estimate with the first completes the proof. ∎

We see from the estimate $\|u-(u_\ell+\varepsilon_h)\|_{0,\Omega} = o(h^2)$ that $\|\varepsilon_h\|_{0,\Omega}$ is an asymptotically exact estimator of the interpolation error $\|u-u_\ell\|_{0,\Omega}$, provided that $\|u-u_\ell\|_{0,\Omega} > m_1h^2$ for some positive constant m_1. We are now ready to prove the main result of this section.

THEOREM 4.3. *Suppose that there are constants $m_1, m_2 > 0$, such that $\|u - u_\ell\|_{0,\Omega} \ge m_1h^2$ and $\|u - u_h\|_{0,\Omega} \ge m_2h^2$. Then, under Assumption* 3.1*, there are constants $c_1, c_2 > 0$, such that*

$$c_1 \le \frac{\|\varepsilon_h\|_{0,\Omega}}{\|u - u_h\|_{0,\Omega}} \le c_2. \tag{47}$$

Proof. It is certainly the case that there are constants $M_1, M_2 > 0$, such that $\|u - u_\ell\|_{0,\Omega} \leq M_1 h^2$ and $\|u - u_h\|_{0,\Omega} \leq M_2 h^2$. So we have

$$\frac{m_1}{M_2} \leq \frac{\|u - u_\ell\|_{0,\Omega}}{\|u - u_h\|_{0,\Omega}} \leq \frac{M_1}{m_2}. \tag{48}$$

The proof is completed by using the fact that $\|\varepsilon_h\|_{0,\Omega}$ is an asymptotically exact estimator of $\|u - u_\ell\|_{0,\Omega}$. ∎

Recall that the proof of the asymptotic exactness of $\|\varepsilon_h\|_{1,\Omega}$ as an estimator of $\|u-u_h\|_{1,\Omega}$ relied on the fact that $\|u_\ell-u_h\|_{1,\Omega} = o(h)$. We see in Lemma 4.4 below that we need $\|u_\ell - u_h\|_{0,\Omega} = o(h^2)$ to get asymptotic exactness of $\|\varepsilon_h\|_{0,\Omega}$ as an estimator of $\|u - u_h\|_{0,\Omega}$.

LEMMA 4.4. *Under Assumption* 3.1*, we have*

$$\|u - (u_h + \varepsilon_h)\|_{0,\Omega} = o(h^2) \iff \|u_h - u_\ell\|_{0,\Omega} = o(h^2). \tag{49}$$

Proof. We have the inequalities

$$\|u - (u_h + \varepsilon_h)\|_{0,\Omega} \leq \|u - (u_\ell + \varepsilon_h)\|_{0,\Omega} + \|u_h - u_\ell\|_{0,\Omega}, \tag{50}$$

$$\|u_h - u_\ell\|_{0,\Omega} \leq \|u - (u_\ell + \varepsilon_h)\|_{0,\Omega} + \|u - (u_h + \varepsilon_h)\|_{0,\Omega}. \tag{51}$$

Lemma 4.2 completes the proof. ∎

The rest of this section is devoted to demonstrating by example that we *cannot* generally expect $\|u_\ell - u_h\|_{0,\Omega} = o(h^2)$ even in an ideal situation for which we can prove $\|u_\ell - u_h\|_{1,\Omega} \lesssim h^2 |\log h|^{1/2} \|u\|_{3,\infty,\Omega}$. Thus, we cannot generally expect asymptotic exactness in the L_2-norm.

Consider the following simple problem on the unit square $\Omega = (0,1) \times (0,1)$:

$$\begin{aligned} -\Delta u &= 2x(1-x) + 2y(1-y) \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned}$$

The exact solution is $u = x(1-x)y(1-y)$. We take the family of uniform meshes having mesh size $h = \frac{1}{n+1}$ and n^2 degrees of freedom located at $(x_i, y_j) = (ih, jh)$; see Figure 1. We will show that $h^2 \lesssim \|u_\ell - u_h\|_{0,\Omega}$.

Let $T \in \mathbb{R}^{n \times n}$ be the tridiagonal matrix with stencil $(-1,\ 2,\ -1)$. Under the standard ordering of unknowns (left to right, bottom to top) the stiffness matrix for this problem is given by

$$A = T \otimes I + I \otimes T = (V \otimes V)(D \otimes I + I \otimes D)(V \otimes V), \tag{52}$$

$$V_{ij} = \sqrt{\frac{2}{n+1}} \sin \frac{ij\pi}{n+1}\ ,\ D_{ij} = \delta_{ij} \left(2 - 2\cos\frac{i\pi}{n+1}\right) = \delta_{ij} 4 \sin^2 \frac{i\pi}{2(n+1)}. \tag{53}$$

We note that $V = V^T = V^{-1}$. As a notational convenience, for $\mathbf{x} \in \mathbb{R}^{n^2}$ we use $\mathbf{x}_{(i,j)} \equiv \mathbf{x}_{(i-1)n+j}$. Similarly, we take $\phi_{(i,j)}$ to be the Lagrange nodal basis function associated with the grid point (x_i, y_j). We define $\mathbf{d}$ and $\mathbf{r}$ to be the error and residual, respectively, at the grid points

$$\mathbf{d}_{(i,j)} = u(x_i, y_j) - u_h(x_i, y_j) = u_\ell(x_i, y_j) - u_h(x_i, y_j), \tag{54}$$

$$\mathbf{r}_{(i,j)} = h^2 f(x_i, y_j) - \int_\Omega f \phi_{(i,j)}\, dxdy = \frac{2}{3} h^4. \tag{55}$$

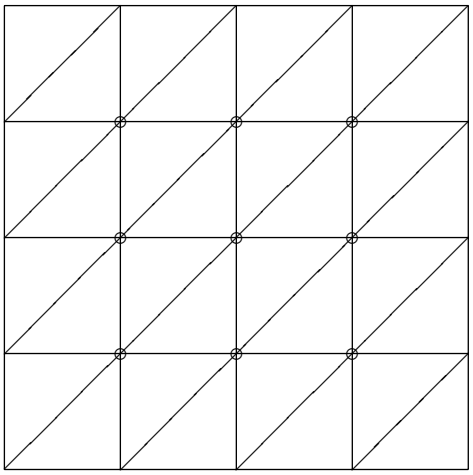


FIG. 1. *Uniform mesh with* $n = 3$.

We have $A\mathbf{d} = \mathbf{r}$. We first argue that $\|u_\ell - u_h\|_{0,\Omega} \geq \frac{h}{2}\|\mathbf{d}\|$, and then establish that $\|\mathbf{d}\| \geq Ch$, thereby proving that $h^2 \lesssim \|u_\ell - u_h\|_{0,\Omega}$. We begin by noting that for any linear function g on a triangle τ, given in terms of its three nodal basis functions, $g = c_1\ell_1 + c_2\ell_2 + c_3\ell_3$, we have

$$\|g\|_{0,\tau}^2 = \frac{|\tau|}{6}(c_1^2 + c_2^2 + c_3^2 + c_1c_2 + c_1c_3 + c_2c_3) \geq \frac{|\tau|}{12}(c_1^2 + c_2^2 + c_3^2). \tag{56}$$

Therefore, if g is continuous and piecewise-linear on $\mathcal{T}$, we have

$$\|g\|_{0,\Omega}^2 = \sum_{\tau\in\mathcal{T}_h} \|g\|_{0,\tau}^2 \geq \frac{|\tau|}{2}\|\mathbf{c}\|^2 = \frac{h^2}{4}\|\mathbf{c}\|^2, \tag{57}$$

where $\mathbf{c}$ is the vector of coefficients. The factor of 6 comes from the fact that each coefficient appears in 6 of the summands $\|g\|_{0,\tau}^2$. This proves that

$$\|u_\ell - u_h\|_{0,\Omega} \geq \frac{h}{2}\|\mathbf{d}\|. \tag{58}$$

We now consider $\|\mathbf{d}\| = \|A^{-1}\mathbf{r}\| = \frac{2}{3}h^4\|A^{-1}(\mathbf{e}\otimes\mathbf{e})\|$, where $\mathbf{e} \in \mathbb{R}^n$ is the vector of ones. It holds that $\|A^{-1}(\mathbf{e}\otimes\mathbf{e})\| = \|(D\otimes I + I\otimes D)^{-1}(V\mathbf{e}\otimes V\mathbf{e})\|$, and

$$(V\mathbf{e})_i = \sqrt{\frac{2}{n+1}}\sum_{j=1}^{n}\sin\frac{ij\pi}{n+1} = \sqrt{\frac{2}{n+1}}\cot\frac{i\pi}{2(n+1)}\left|\sin\frac{i\pi}{2}\right|. \tag{59}$$

This gives us

$$\|A^{-1}(\mathbf{e}\otimes\mathbf{e})\|^2 = \frac{h^2}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|\sin\frac{i\pi}{2}\sin\frac{j\pi}{2}\right|\left(\frac{\cot\frac{i\pi}{2(n+1)}\cot\frac{j\pi}{2(n+1)}}{\sin^2\frac{i\pi}{2(n+1)} + \sin^2\frac{j\pi}{2(n+1)}}\right)^2 \tag{60}$$

$$> \frac{h^2}{4}\left(\frac{\cot\frac{\pi}{2(n+1)}\cot\frac{\pi}{2(n+1)}}{\sin^2\frac{\pi}{2(n+1)} + \sin^2\frac{\pi}{2(n+1)}}\right)^2 \tag{61}$$

$$= \frac{h^2}{16}\frac{\cos^4\frac{\pi}{2(n+1)}}{\sin^8\frac{\pi}{2(n+1)}} > \frac{h^2}{16}\frac{(\frac{1}{\sqrt{2}})^4}{(\frac{\pi}{2(n+1)})^8} = \frac{4}{\pi^8}\,h^{-6}. \tag{62}$$

Combining these results we have $\|u_\ell - u_h\|_{0,\Omega} > \frac{h}{2}\frac{2h^4}{3}\frac{2h^{-3}}{\pi^4} = \frac{2h^2}{3\pi^4}$, which completes the argument.

5. Experiments. In this section we offer four examples which illustrate the effectivity of our estimator and provide some comments on its computational cost. In particular, we wish to verify (2), the key results of this paper, in practice. The exact error for each of the examples solution is known, so we can judge the quality of our estimator directly. Throughout this section we use $e_h \equiv u - u_h$ for the exact error and the abbreviation EFF for each of the effectivity ratios

$$\frac{\|\varepsilon_h\|_{0,\Omega}}{\|e_h\|_{0,\Omega}}, \qquad \frac{|\varepsilon_h|_{1,\Omega}}{|e_h|_{1,\Omega}}, \qquad \frac{|\varepsilon_h|_{2,\Omega}}{|u|_{2,\Omega}}. \tag{63}$$

For the sake of convenience we abuse notation slightly by taking

$$|\varepsilon_h|_{2,\Omega} \equiv \sqrt{\sum_{\tau\in\mathcal{T}} |\varepsilon_h|^2_{2,\tau}}. \tag{64}$$

This is an abuse because $|v|_{2,\Omega}$ is infinite by its standard definition for functions such as ε_h, which have a gradient jump between elements in a mesh. Additionally, we abbreviate the standard scientific notation by placing the base 10 exponent as a subscript, for example, $3.54_{-2} \equiv 3.54 \times 10^{-2}$.

The quantity N appearing in the tables is the number of triangles in the mesh. For the larger values of N, this is roughly twice the number of vertices in the mesh. In the first four examples, for which the exact error is known, we use the error model $E = CN^{-p}$, derived from standard a priori estimates and $Nh^2 \sim 1$, to give a sense of the rate of convergence of error. In particular, we give the least-squares best fit for each of the normed errors. We note that $p = 1$ (resp., $p = 1/2$) corresponds to what is generally called *quadratic* (resp., *linear*) convergence—in terms of the mesh parameter h—and we use this language in the explanations below. The code used for the numerical experiments is PLTMG [6], with modifications necessary to implement our error estimation technique.

5.1. The simple problem. For our first experiment, we revisit the example from section 4 which was used to demonstrate that one cannot generally expect asymptotic exactness from our estimator in L_2. We will see, however, that the function recovery is very nearly exact in this case. Recall that the problem is to find u such that

$$\begin{aligned} -\Delta u &= 2x(1-x) + 2y(1-y) \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega. \end{aligned}$$

Here Ω is the unit square, and the exact solution is $u = x(1-x)y(1-y)$. We provide the values of the various norms of u so that the relative errors can be readily assessed if desired:

$$\|u\|_{0,\Omega} = \sqrt{\frac{1}{900}} = 0.0\bar{3}, \qquad |u|_{1,\Omega} = \sqrt{\frac{1}{45}} \approx 0.149, \qquad |u|_{2,\Omega} = \sqrt{\frac{22}{45}} \approx 0.699.$$

This example is also used in the numerical experiments in [21, 23].

In Table 1 we see the predicted performance of the estimator in each of the square norms, with the L_2 error estimate having effectivity very near 1 on each mesh. Below, we give the approximate error models for the function and gradient errors:

$$\|e_h\|_{0,\Omega} \approx 0.159N^{-1.02}, \qquad |e_h|_{1,\Omega} \approx 0.307N^{-0.502}.$$

TABLE 1
Estimates, exact values, and effectivity for the simple problem.

N	88	441	1887	7765	31505	126919
$\|\varepsilon_h\|_{0,\Omega}$	1.71_{-3}	2.93_{-4}	6.98_{-5}	1.62_{-5}	3.90_{-6}	9.64_{-7}
$\|e_h\|_{0,\Omega}$	1.65_{-3}	3.09_{-4}	7.22_{-5}	1.63_{-5}	3.93_{-6}	9.76_{-7}
EFF	1.04	0.947	0.966	0.993	0.994	0.987
$\|\varepsilon_h\|_{1,\Omega}$	3.19_{-2}	1.36_{-2}	6.61_{-3}	3.14_{-3}	1.54_{-3}	7.67_{-4}
$\|e_h\|_{1,\Omega}$	3.14_{-2}	1.37_{-2}	6.61_{-3}	3.15_{-3}	1.55_{-3}	7.72_{-4}
EFF	1.01	0.998	1.00	0.997	0.996	0.996
$\|\varepsilon_h\|_{2,\Omega}$	0.726	0.713	0.709	0.705	0.703	0.703
$\|u\|_{2,\Omega}$	0.699	0.699	0.699	0.699	0.699	0.699
EFF	1.04	1.02	1.01	1.01	1.01	1.00

TABLE 2
Estimates, exact values, and effectivity for the oscillatory problem.

N	88	434	1888	7825	31679	127552
$\|\varepsilon_h\|_{0,\Omega}$	0.369	0.149	8.43_{-2}	1.50_{-2}	3.27_{-3}	7.99_{-4}
$\|e_h\|_{0,\Omega}$	0.499	0.172	9.60_{-2}	1.76_{-2}	3.89_{-3}	9.49_{-4}
EFF	0.738	0.865	0.878	0.853	0.846	0.842
$\|\varepsilon_h\|_{1,\Omega}$	15.1	8.43	6.56	3.04	1.46	0.716
$\|e_h\|_{1,\Omega}$	17.6	9.78	6.92	3.07	1.46	0.720
EFF	0.859	0.862	0.949	0.991	0.993	0.995
$\|\varepsilon_h\|_{2,\Omega}$	304	458	603	632	634	634
$\|u\|_{2,\Omega}$	632	632	632	632	632	632
EFF	0.481	0.693	0.954	1.00	1.00	1.00

We point out that we observe the predicted a priori quadratic convergence of $\|e_h\|_{0,\Omega}$ and linear convergence of $|e_h|_{1,\Omega}$.

5.2. The oscillatory problem. In this second example we consider the situation where the exact solution still possesses no singularities, but oscillates strongly. The problem is to find u such that

$$-\Delta u = 128\pi^2 \sin(8\pi x)\sin(8\pi y) \text{ in } \Omega,$$
$$u = 0 \text{ on } \partial\Omega.$$

Here Ω is again the unit square, and the exact solution is $u = \sin(8\pi x)\sin(8\pi y)$. The pertinent norms of u are given below:

$$\|u\|_{0,\Omega} = \sqrt{\frac{1}{4}} = 0.5, \qquad |u|_{1,\Omega} = \sqrt{32\pi^2} \approx 17.8, \qquad |u|_{2,\Omega} = \sqrt{4096\pi^4} \approx 632.$$

In Table 2 we again see effectivity approaching 1 for the gradient error and the Hessian in both norms. The effectivity of the function error estimate tends to stay in the 80–85% range. We see in the approximate error models below that the adaptive refinement seems to be producing suboptimal reduction of function and gradient error:

$$\|e_h\|_{0,\Omega} \approx 36.5N^{-0.873}, \qquad |e_h|_{1,\Omega} \approx 149N^{-0.443}.$$

This is due to the fact that the two coarsest meshes are just beginning to resolve the oscillatory behavior. When the error data from these two initial meshes is removed, we see the expected quadratic and linear convergence for the function and gradient errors, respectively. More precisely, the exponents for the L_2 and H^1 error models are $p = 1.09$ and 0.536, respectively.

TABLE 3
Estimates, exact values, and effectivity for the slit domain problem.

N	94	481	2031	8334	33704	135632
$\|\varepsilon_h\|_{0,\Omega}$	2.81_{-2}	3.20_{-3}	5.26_{-4}	1.39_{-4}	3.43_{-5}	8.51_{-6}
$\|e_h\|_{0,\Omega}$	0.122	3.92_{-2}	1.33_{-2}	3.57_{-3}	9.08_{-4}	1.78_{-4}
EFF	0.230	8.18_{-2}	3.96_{-2}	3.88_{-2}	3.78_{-2}	4.78_{-2}
$\|\varepsilon_h\|_{1,\Omega}$	0.419	0.231	0.132	6.93_{-2}	3.51_{-2}	1.62_{-2}
$\|e_h\|_{1,\Omega}$	0.590	0.331	0.189	9.91_{-2}	4.99_{-2}	2.25_{-2}
EFF	0.710	0.698	0.697	0.699	0.703	0.720
$\|\varepsilon_h\|_{2,\Omega_s}$	5.34	19.9	24.2	18.2	17.5	17.2
$\|u\|_{2,\Omega_s}$	17.2	17.2	17.2	17.2	17.2	17.2
EFF	0.310	1.16	1.40	1.06	1.02	1.00
N	94	481	2031	8334	33704	135632
$\|\varepsilon_h\|_{0,\Omega_s}$	2.81_{-2}	3.20_{-3}	5.04_{-4}	1.38_{-4}	3.43_{-5}	8.51_{-6}
$\|e_h\|_{0,\Omega_s}$	0.122	3.92_{-2}	1.32_{-2}	3.56_{-3}	9.04_{-4}	1.78_{-4}
EFF	0.230	8.18_{-2}	3.81_{-2}	3.89_{-2}	3.79_{-2}	4.80_{-2}
$\|\varepsilon_h\|_{1,\Omega_s}$	0.419	0.231	6.28_{-2}	2.69_{-2}	1.37_{-2}	6.89_{-3}
$\|e_h\|_{1,\Omega_s}$	0.590	0.331	0.119	3.47_{-2}	1.48_{-2}	6.99_{-3}
EFF	0.710	0.698	0.526	0.774	0.925	0.986

5.3. The slit domain problem. For our third example we consider a problem for which the boundary conditions force a singularity at the origin. Because of the infinite gradient at the origin, it is interesting to investigate the effectivity of the estimators. The problem is to find u such that

$$-\Delta u = 0 \text{ in } \Omega, \qquad u(r,0^+) = 0, \qquad \nabla u \cdot \mathbf{n}\,(r,2\pi^-) = 0, \qquad u(1,\theta) = \sin(\theta/4).$$

Here Ω is the unit disk with the positive x-axis removed, and the exact solution is $u = r^{1/4}\sin(\theta/4)$. Though the gradient of u is infinite at the origin, $|u|_{1,\Omega}$ is finite. However, this is not the case for $|u|_{2,\Omega}$—here we must avoid the origin to get a finite H^2 seminorm. Let Ω_s denote Ω with the disk of radius s about the origin removed. In the experiments, we take $s = 1/100$. The pertinent norms are given below:

$$\|u\|_{0,\Omega} = \sqrt{\frac{2\pi}{5}} \approx 1.12,\ |u|_{1,\Omega} = \sqrt{\frac{\pi}{4}} \approx 0.886,\ |u|_{2,\Omega_s} = \sqrt{\frac{3\pi}{32}(s^{-3/2}-1)} \approx 17.2.$$

We note that the global smoothness condition $u \in W^3_\infty(\Omega)$ is certainly not satisfied here.

In Table 3 we see the clear effects of this singularity on the performance of the function error estimates and the gradient error. Here the function error estimates underestimate the true function error by roughly a factor of 26.5 at worst and a factor of 5 at best, and the gradient error estimate underestimates the true gradient error by 28% at best, though it is slowly improving. We also point out that the second derivatives are recovered quite well. Concerning Table 3, we mention finally that the performance of the gradient error estimate improves markedly if we restrict our attention to the error on the subdomain Ω_s, as is seen at the bottom of that table, but the performance of the function error estimate does not improve appreciably. The approximate error models given below, though showing subquadratic convergence of the function error and sublinear convergence of the gradient error, are actually quite encouraging for a problem with this sort of singularity, where we would expect $p \approx 1/8$ asymptotically for the gradient error $|e_h|_{1,\Omega}$ under uniform refinement:

$$\|e_h\|_{0,\Omega} \approx 9.31N^{-0.894}, \qquad |e_h|_{1,\Omega} \approx 5.11N^{-0.447}.$$

TABLE 4
Estimates, exact values, and effectivity for the jumping coefficient problem.

N	66	353	1530	6337	25734	103617
$\|\varepsilon_h\|_{0,\Omega}$	6.67	0.396	8.04_{-2}	1.86_{-2}	4.33_{-3}	1.22_{-3}
$\|e_h\|_{0,\Omega}$	11.1	0.811	0.116	2.29_{-2}	5.21_{-3}	1.49_{-3}
EFF	0.603	0.488	0.691	0.814	0.831	0.820
$\|\varepsilon_h\|_{1,\Omega}$	96.0	33.3	13.6	6.06	2.90	1.42
$\|e_h\|_{1,\Omega}$	108	36.1	14.1	6.16	2.92	1.43
EFF	0.886	0.923	0.960	0.985	0.994	0.997
$\|\varepsilon_h\|_{2,\Omega_s}$	1.17_3	2.49_3	2.50_3	2.40_3	2.35_3	2.33_3
$\|u\|_{2,\Omega_s}$	2.32_3	2.32_3	2.32_3	2.32_3	2.32_3	2.32_3
EFF	0.504	1.07	1.07	1.03	1.01	1.00

5.4. The jumping coefficient problem. The problem is to find u such that

$$-a_k \Delta u = 0 \text{ in } \Omega, u(r,0) = 0,$$
$$\nabla u \cdot \mathbf{n}\,(r,\pi) = 0, u(1,\theta) = b_k \sin(\alpha\theta) + c_k \cos(\alpha\theta).$$

Here Ω is the upper half of the unit disk, which is divided into two regions having differing coefficients of diffusion. In the first region, $0 < \theta < \frac{\pi}{4}$, we have $a_1 = 10^3$. In the second region, $\frac{\pi}{4} < \theta < \pi$, we have $a_2 = 1$. The exact solution is $u = r^\alpha(b_k \sin(\alpha\theta) + c_k \cos(\alpha\theta))$, where the values α, b_k, c_k are determined by the boundary conditions at $\theta = 0, \pi$ and the continuity of u and $a_k \nabla u \cdot \mathbf{n}$ along the interface $\theta = \frac{\pi}{4}$ between the two regions. The boundary condition at $r = 1$ is chosen to match the solution in the interior. The boundary conditions on the positive and negative x-axes and the continuity conditions at the interface provide four equations which are linear in b_1, c_1, b_2, c_2 (and trigonometric in α). It is clear that $b_1 = c_1 = b_2 = c_2 = 0$ trivially satisfies all of the specified conditions, so we must select α so that the resulting linear system is singular—therefore admitting nontrivial solutions. If there are any such choices of α, then there are infinitely many. We selected the following solution, with $\alpha \approx 0.666422$:

$$b_1 = 1, \qquad c_1 = 0, \qquad b_2 \approx 750.416, \qquad c_2 \approx -432.484,$$
$$\|u\|_{0,\Omega} \approx 515, \qquad |u|_{1,\Omega} \approx 767, \qquad |u|_{2,\Omega_s} \approx 2.32_3.$$

Again we take Ω_s to be Ω with the disk of radius $s = 1/100$ removed. Although $u \notin H^2(\Omega_s)$ because of the jump discontinuity of ∇u at the interface between the two regions, we abuse notation by taking

$$|u|^2_{2,\Omega_s} \equiv \sum_{\tau \in \mathcal{T}_s} |u|^2_{2,\tau} \tag{65}$$

for Table 4. This sum is finite because the interface between the two regions does not pass through the interior of any of the triangles.

In Table 4, we see the data for this experiment. We point out that the performance of the various error estimates based on the approximate error function seem to be unaffected by the jump in the coefficient. In particular, we see effectivity ratios near or approaching 1 for the gradient error and the Hessian, and slightly better than 80% for the function values in each norm. The approximate error models given below show error convergence which is better than one would expect, with superquadratic convergence in function error and superlinear convergence in gradient error:

$$\|e_h\|_{0,\Omega} \approx 1.12_3 N^{-1.20}, \qquad |e_h|_{1,\Omega} \approx 1.58_3 N^{-0.589}.$$

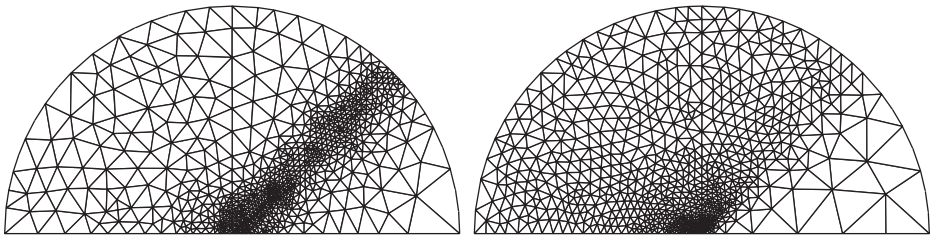


FIG. 2. *The meshes for the jumping coefficient problem after three stages of adaptive refinement, using Bank–Xu gradient recovery estimates (left) and bump function error estimates (right). The mesh on the left has* 804 *vertices and* 1534 *triangles, and the mesh on the right has* 808 *vertices and 1530 triangles.*

These convergence rates are elevated in the models because of the significant error reduction in the early stages of refinement. When we remove the error data from the first two meshes, the convergence rates drop to the more normal quadratic and linear levels.

In addition to having an r^α, $\alpha < 1$ singularity at the origin, the solution also has a jump discontinuity in its gradient at $\theta = \pi/4$. It is relevant at this point to consider which of these two types of singularities has the stronger (negative) effect on the performance of the estimator for problems of this sort. Considering that the slit domain problem possesses only an r^α singularity and that the α for that problem is smaller than the one for this problem, comparing the performance of the estimator in both cases suggests that r^α singularities are more influential than jump discontinuities in the gradient. In fact, a careful reading of either the Bank–Xu paper [3] or the Xu–Zhang paper [20] reveals that the key superconvergence result for this paper,

$$\|u_h - u_\ell\|_{1,\Omega} = o(h),$$

holds for u having a finite number of gradient jump discontinuities provided that u is sufficiently smooth in each of corresponding subdomains. So we see that, asymptotically, the effectivity of the estimator is affected by jumping coefficients only if they lead to singularities which are worse than gradient jump discontinuities.

We also mention that, for problems of the sort for which we can expect gradient jumps, a naive application of gradient recovery error estimators will lead to suboptimal and sometimes terrible performance. This is because of the fact that gradient recovery schemes involve some sort of local or global averaging. If care is not taken to avoid averaging across an interface where ∇u jumps, then the local error estimates near the interface will tend to overestimate the actual error there—particularly when u_h approximates u well. To illustrate this explicitly we give a brief summary of the result using the Bank–Xu recovery technique, which is a global recovery technique. In Figure 2, we see a clear qualitative difference between the sort of refinement produced by the bump estimator and the naive use of the Bank–Xu estimator—the sort of difference we might have guessed due to the overestimation of error near the interface for the latter. The error model for this refinement is $|e_h|_{1,\Omega} \approx 845\, N^{-0.487}$, with effectivity $EFF \approx 3$ as the mesh is refined. We are not trying to make the point that this sort of bad behavior is unavoidable for gradient recovery schemes—in practice it can be avoided by taking care to not average out a gradient jump where there should be one. Bank and Xu noted this in an example in [4], and performing their gradient recovery scheme for our problem on each subdomain separately restores the optimal performance. The point that we are trying to make with this discussion is that with

TABLE 5

Timing comparison: the ratio of the costs to compute ε_h and $\mathcal{R}\nabla u_h$. Ratios in the first three rows correspond to using SGS-CG to compute ε_h, and the bottom three rows correspond to using unpreconditioned CG.

Simple	3.17	3.37	3.53	3.25	2.66	2.24
Oscillatory	3.15	4.07	3.61	3.40	2.05	2.38
Slit domain	3.19	2.83	2.85	3.01	2.66	2.03
Simple	2.50	2.56	2.63	2.38	1.98	1.49
Oscillatory	2.45	2.53	2.74	2.59	1.44	1.53
Slit domain	2.55	2.09	2.15	2.12	1.94	1.40

the bump error estimator it is not necessary to treat subdomains differently. We think that this is an attractive feature of the estimator, particularly in cases where the number of jumps in the coefficient on the diffusion term (and hence the number of jumps in the gradient of the solution) is large, or where there are small or narrow regions in which the number of elements needed to get a good approximation of the true solution there is smaller than the number of elements needed to perform any of the standard gradient recovery techniques.

5.5. Computational cost. Although the linear system involved in the computation of ε_h can be expected to have roughly three times the number of unknowns as that for computing u_h, the system itself is readily solved because it is well-conditioned (see [5, p. 11], for example). But how does the cost compare with that of various gradient recovery schemes? We content ourselves with a direct comparison to the recovery scheme of Bank and Xu as it is currently implemented in PLTMG. In Table 5, we have the ratios of the times needed to compute ε_h and the recovered gradient $\mathcal{R}\nabla u_h$ for three of the four problems considered here—the jumping coefficient problem was omitted because it would have required a modification of the gradient recovery subroutines in PLTMG. We have used the symmetric Gauß–Seidel method as a preconditioner for CG in the computation of ε_h, as in all of the experiments above, and these data correspond to those experiments. For example, the ratio 3.17 for the simple problem corresponds to the coarsest mesh (88 triangles for both ε_h and $\mathcal{R}\nabla u_h$), and 2.24 corresponds to the finest mesh (126919 triangles for ε_h and 127020 for $\mathcal{R}\nabla u_h$).

For these three problems, unpreconditioned CG can be used instead with no loss in effectivity. When this is done, the timing ratios improve, as is shown in the bottom three rows of Table 5. We generally advocate using some sort of preconditioner for problems such as the jumping coefficient problem because otherwise one notices a drop in effectivity. We suggest that the greater computational cost, still quite small with respect to the total computational cost of the adaptive algorithm, may be worthwhile for this very robust and flexible error estimator. The robustness of the estimator is seen theoretically in that, even in situations where the assumptions taken here do not apply, we can fall back on the "old" analysis based on the milder saturation assumption and on the strengthened Cauchy inequality, which hold under quite general conditions (see [9] and [10, pp. 436–445]). The flexibility of the approximate error function $\varepsilon_h \approx u - u_h$ is clear in that it can be used to measure error in other norms or to approximate error in certain functionals of interest (see [18]), as well as for mesh smoothing procedures such as that proposed by Bank and Smith [2].

6. Final remarks. We have given proof and numerical evidence of the effectiveness of the hierarchical basis type bump function estimator $\varepsilon_h \approx u - u_h$ in recovering function values and first and second derivatives. The proofs offered here are based on

the superconvergence result $\|u_h - u_\ell\|_{1,\Omega} = o(h)$, which is usually used in the proofs of the effectiveness of gradient recovery methods. In our proofs, we replace the standard saturation assumption and strengthened Cauchy inequality used in the analysis commonly given for hierarchical basis methods with relatively mild mesh symmetry conditions and relatively strong smoothness assumptions, which are sufficient but often not seen to be necessary in practice. We thereby obtain stronger theoretical results than are generally given for such estimators, and these results are borne out in practice. The approximation $\varepsilon_h \approx u - u_h$ is provably quite robust and can be used for error estimation and adaptivity in a variety of norms and other measures.

In terms of the asymptotically exact recovery of gradient error, our estimator $\|\nabla\varepsilon_h\|_{0,\Omega}$ has a lot of very good competition in the many gradient recovery procedures proposed in the literature. In addition to the recovery procedure of Bank and Xu, which is mentioned several times above, we also cite the local least-squares fitting of Zienkiewicz and Zhu [23, 24] (perhaps the most popular), the polynomial preserving method of Zhang and Naga [16, 21], and the method proposed by Wiberg and Li [15, 19], which has the most in common with our own in that it can be used directly to produce a locally quadratic (though not globally continuous) approximation of the error $u-u_h$. These methods should also be suitable for recovering second derivatives—Bank and Xu argue as much for their estimator—but not much has been written in the gradient recovery literature about estimating the function error. The notable exception in this regard is in the aforementioned works of Wiberg and Li, where numerical evidence of efficiency and reliability of their estimator are given, but no analysis is provided.

We now briefly consider a few straightforward generalizations of what has been presented here. The $\mathcal{O}(h^{2\sigma})$-irregular triangulation assumption is generalized in [20], where Xu and Zhang call it *Condition*(α,σ). We note that the σ in the Xu–Zhang paper plays the role of the 2σ used in both the Bank–Xu paper and our own, and an $\mathcal{O}(h^{1+\alpha})$-parallelogram property is used instead of an $\mathcal{O}(h^2)$-parallelogram property. In their paper, Xu and Zhang also use the less stringent regularity condition $u \in H^3(\Omega) \cap W^2_\infty(\Omega)$. Under these assumptions and a few natural assumptions on the bilinear form for the problem, they prove that

$$\|u_h - u_\ell\|_{1,\Omega} \lesssim h^{1+\min(\alpha,1/2,\sigma/2)}(\|u\|_{3,\Omega} + |u|_{2,\infty,\Omega}). \tag{66}$$

The results in this paper can be modified in the obvious way to incorporate the Xu–Zhang version of the mesh symmetry conditions and the weaker regularity assumption, with no change in the proofs.

We will mention two other ways in which the arguments given here can be readily generalized. The first is to consider linear simplicial elements in $\mathbb{R}^n$, $n > 2$. Recall that the key result from which all of the other estimates were proved was of the form

$$\|u_h - u_\ell\|_{1,\Omega} = o(h), \tag{67}$$

where u_h is the linear finite element approximation and u_ℓ is the linear Lagrange interpolant. Brandts and Křížek [7, 12] show that

$$\|u_h - u_\ell\|_{1,\Omega} \lesssim h^2\|u\|_{3,\Omega} \tag{68}$$

on very regular meshes for $u \in H^1_0(\Omega) \cap H^s(\Omega)$ and $s = 3$ for $n \le 5$ and $s > n/2$ for $n \ge 6$. Any s greater than 3 is needed only to ensure that the nodal interpolant u_ℓ can be well-defined. Chen [8] generalizes the argument of [3] to mildly structured

tetrahedral meshes in $\mathbb{R}^3$ to obtain

$$\|u_h - u_\ell\|_{1,\Omega} \lesssim h^{1+\min(1,\sigma)}\|u\|_{3,\infty,\Omega}, \tag{69}$$

where $u \in H_0^1(\Omega) \cap W_\infty^3(\Omega)$ and σ measures the violation of an $\mathcal{O}(h^2)$-parallelepiped property. With such superconvergence results, the extension of our results proceeds in the obvious fashion.

Another generalization would be to consider hierarchical error estimators for higher order elements. For example, let $\hat{V}_h = \bar{V}_h \oplus (\hat{V}_h \setminus \bar{V}_h)$ be the piecewise cubic finite element space, which we think of hierarchically. If $\bar{u}_h \in \bar{V}_h$ is the finite element solution, we might want to estimate the error $u - \bar{u}_h$ using a function in $\hat{V}_h \setminus \bar{V}_h$; call it $\bar{\varepsilon}_h$. Li [13, 14] has shown that Lagrange interpolation does not generally give the analogous superconvergence results for elements of degree 3 or higher in $\mathbb{R}^2$, but we are free to use some other appropriate interpolation scheme. Let $\Pi_q : C(\bar{\Omega}) \to \bar{V}_h$ and $\Pi_c : C(\bar{\Omega}) \to \hat{V}_h$ be defined by

$$\begin{aligned}
&\Pi_q u(v_i) = \Pi_c u(v_i) = u(v_i) \text{ for vertices } v_i,\\
&\int_{e_j} u - \Pi_q u \, ds = \int_{e_j} (u - \Pi_c u) v \, ds = 0 \text{ for edges } e_j \text{ and linear functions } v,\\
&\int_\tau u - \Pi_c u \, dx = 0 \text{ for triangles } \tau.
\end{aligned}$$

Huang and Xu [11] argue that

$$\|\bar{u}_h - \Pi_q u\|_{1,\Omega} \lesssim h^{2+\min(1,\sigma)/2}(\|u\|_{4,\Omega} + |u|_{3,\infty,\Omega}),\ \Pi_c u - \Pi_q u \in \hat{V}_h \setminus \bar{V}_h. \tag{70}$$

One might correctly infer from the statement of the result that a similar argument to those found in [3, 20] is used. With an estimate like this, the analogue of Theorem 3.3 can be proved in the obvious way. Using arguments like those given in Lemma 3.5 and Theorem 3.6, we see that our approximate error function $\bar{\varepsilon}_h \approx u - \bar{u}_h$ provides superconvergent approximation of $\|u - \bar{u}_h\|_{2,\Omega}$ and convergent approximation of $\|u\|_{3,\Omega}$. Finally, arguing along the same lines as in section 4 we get even better results than in the case of piecewise linears, because it actually does hold that $\|\bar{u}_h - \Pi_q u\|_{0,\Omega} = o(h^3)$. Huang and Xu have plans to extend their results to higher order elements as well, and the analogous results should be able to be plugged into our framework with little difficulty.

Acknowledgments. The author thanks Wolfgang Hackbusch and other colleagues in the scientific computation group at the Max Planck Institute in Leipzig for many helpful discussions, as well as Randy Bank at University of California, San Diego for assistance with PLTMG, and for useful comments concerning this manuscript. Additionally I thank the referees for their helpful input.

REFERENCES

[1] R. E. Bank and R. K. Smith, *A posteriori error estimates based on hierarchical bases*, SIAM J. Numer. Anal., 30 (1993), pp. 921–935.

[2] R. E. Bank and R. K. Smith, *Mesh smoothing using a posteriori error estimates*, SIAM J. Numer. Anal., 34 (1997), pp. 979–997.

[3] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators, Part* I. *Grids with superconvergence*, SIAM J. Numer. Anal., 41 (2003), pp. 2294–2312.

[4] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators, Part* II. *General unstructured grids*, SIAM J. Numer. Anal., 41 (2003), pp. 2313–2332.

[5] R. E. Bank, *Hierarchical bases and the finite element method*, in Acta Numerica, Volume 5, Cambridge University Press, Cambridge, UK, 1996, pp. 1–43.
[6] R. E. Bank, *Pltmg: A Software Package for Solving Elliptic Partial Differential Equations, Users' Guide* 9.0, Technical report, University of California, San Diego, 2004.
[7] J. Brandts and M. Křížek, *Gradient superconvergence on uniform simplicial partitions of polytopes*, IMA J. Numer. Anal., 23 (2003), pp. 489–505.
[8] L. Chen, *Superconvergence of tetrahedral linear finite elements*, Int. J. Numer. Anal. Model., 3 (2006), pp. 273–282.
[9] W. Dörfler and R. H. Nochetto, *Small data oscillation implies the saturation assumption*, Numer. Math., 91 (2002), pp. 1–12.
[10] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Appl. Math. Sci. 159, Springer-Verlag, New York, 2004.
[11] Y. Huang and J. Xu, *Superconvergence of Quadratic Finite Elements on Mildly Structured Grids*, Math. Comput., to appear.
[12] M. Křížek, *Superconvergence phenomena on three-dimensional meshes*, Int. J. Numer. Anal. Model., 2 (2005), pp. 43–56.
[13] B. Li, *Superconvergence for higher-order triangular finite elements*, Chinese J. Numer. Math. Appl., 12 (1990), pp. 75–79.
[14] B. Li, *Lagrange interpolation and finite element superconvergence*, Numer. Methods Partial Differential Equations, 20 (2004), pp. 33–59.
[15] X. D. Li and N.-E. Wiberg, *A posteriori error estimate by element patch post-processing, adaptive analysis in energy and L_2 norms*, Comput. & Structures, 53 (1994), pp. 907–919.
[16] A. Naga and Z. Zhang, *A posteriori error estimates based on the polynomial preserving recovery*, SIAM J. Numer. Anal., 42 (2004), pp. 1780–1800.
[17] J. S. Ovall, *Duality-Based Adaptive Refinement for Elliptic PDEs*, Ph.D. thesis, Department of Mathematics. University of California at San Diego, 2004.
[18] J. S. Ovall, *Asymptotically exact functional error estimators based on superconvergent gradient recovery*, Numer. Math., 102 (2006), pp. 543–558.
[19] N.-E. Wiberg and X. D. Li, *Superconvergent patch recovery of finite-element solution and a posteriori L_2 norm error estimate*, Comm. Numer. Methods Engrg., 10 (1994), pp. 313–320.
[20] J. Xu and Z. Zhang, *Analysis of recovery type a posteriori error estimators for mildly structured grids*, Math. Comp., 73 (2004), pp. 1139–1152.
[21] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.
[22] O. C. Zienkiewicz, D. W. Kelly, J. Gago, and I. Babuška, *Hierarchical finite element approaches, error estimates and adaptive refinement*, in The Mathematics of Finite Elements and Applications, IV (Uxbridge, 1981), Academic Press, London, 1982, pp. 313–346.
[23] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates.* I. *The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.
[24] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergent patch recovery and a posteriori error estimates.* II. *Error estimates and adaptivity*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1365–1382.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1081–1106

NUMERICAL METHODS FOR QUASI-LINEAR ELLIPTIC EQUATIONS WITH NONLINEAR BOUNDARY CONDITIONS*

C. V. PAO†

Abstract. The purpose of this paper is to give a numerical treatment for a class of quasi-linear elliptic equations under nonlinear boundary conditions, including the three basic types of linear boundary conditions. The quasi-linear equation is discretized by the finite difference method, and the method of upper-lower solutions and its associated monotone iteration are used to compute the solutions of the finite difference system. This method leads to monotone iterative schemes for the computation of numerical solutions as well as some comparison results among the monotone iterative schemes. It also leads to the existence of a maximal and a minimal finite difference solution, including the uniqueness of the solution, and the convergence of the finite difference solution to the corresponding continuous solution. Applications are given to two physical problems in heat conduction and combustion theory, and numerical results for the heat-conduction problem are given, and are compared with the known true continuous solution.

Key words. density-dependent reaction diffusion, nonlinear boundary condition, monotone iterative schemes, upper and lower solutions, convergence of finite difference solution, heat conduction

AMS subject classifications. 65N22, 65N06, 65N12

DOI. 10.1137/060653640

1. Introduction.

Nonlinear elliptic boundary value problems arise from many fields of applied sciences and have been investigated extensively in the literature both analytically and numerically. The analytical consideration is mostly for the existence, uniqueness, multiplicity, and bifurcation of solutions, while the numerical investigation is often devoted to accurate and efficient computational algorithms, error estimates, and convergence of the discrete solution to the corresponding continuous solution of the original problem. In this paper, we investigate some of the numerical aspects for a class of quasi-linear elliptic equations under nonlinear boundary conditions, including the three basic types of linear boundary conditions: Dirichlet, Neumann, and Robin. The class of quasi-linear boundary problems under consideration is given in the form

$$\begin{aligned} -\nabla\cdot(D(u)\nabla u) \;+\; \boldsymbol{c}(x)\cdot(D(u)\nabla u) \;&=\; f(x,u) \quad (x\in\Omega), \\ D(u)\partial u/\partial\nu \;&=\; g(x,u) \qquad\qquad\quad (x\in\partial\Omega), \end{aligned} \tag{1.1}$$

where Ω is a bounded domain in $\mathbb{R}^p$ with boundary $\partial\Omega (p = 1, 2, \ldots)$, ∇ is the gradient operator in Ω, and $\partial u/\partial\nu$ denotes the outward normal derivative of u on $\partial\Omega$. The vector $\boldsymbol{c}(x) = (c^{(1)}(x), \ldots, c^{(p)}(x))$ and the functions $D(u)$, $f(x,u)$, and $g(x,u)$ (which, in general, are nonlinear in u) are prescribed continuous functions of their respective arguments. In terms of reaction diffusion problems, the three terms in the quasi-linear equation of (1.1) are referred to as diffusion, convection, and reaction, respectively. It is assumed that $D(u) > 0$ for u in a subset $\mathcal{S}_0$ of $\mathbb{R}^1$(see (2.12)), but we allow $D(u) = 0$ at $u = 0$ for homogeneous Dirichlet boundary condition in (1.1a) below. This assumption implies that problem (1.1a) may be degenerate at the boundary points when the boundary condition is of Dirichlet type. The consideration

*Received by the editors March 5, 2006; accepted for publication (in revised form) January 10, 2007; published electronically May 7, 2007.
http://www.siam.org/journals/sinum/45-3/65364.html
†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 (cvpao@math.ncsu.edu).

of the boundary condition in (1.1) includes the Neumann or Robin type

$$\partial u/\partial \nu + \beta(x)u = g^*(x,u) \quad (x \in \partial\Omega),$$

where $\beta(x) \geq 0$ on $\partial\Omega$. In fact, the above boundary condition, including the linear case $g^*(x,u) = g^*(x)$, can be written in the form of (1.1) with $g(x,u) = D(u)(g^*(x,u) - \beta(x)u)$. The above nonlinear boundary condition has been given considerable attention in the literature, and various special forms of the nonlinear functions were treated (cf. [13, 17, 21, 23, 25, 31]). In order to include all three basic types of linear boundary conditions, we also consider the Dirichlet boundary condition

$$u(x) = \xi(x) \qquad (x \in \partial\Omega) \tag{1.1a}$$

and refer to this problem as problem (1.1a). It is to be noted that if the diffusion coefficient $D(u)$ is replaced by the more general function $b(x)D(u)$ for a smooth function $b(x)$ in Ω, then problem (1.1) can be reduced to the same form except with the convection coefficient $\boldsymbol{c}$ replaced by $(\boldsymbol{c} + \nabla b)$. Hence our investigation in the following discussions is directly applicable to the above more general equation without any complication.

Numerical treatment of the quasi-linear boundary problem (1.1) is extensive, and various aspects of the problem, such as method of computation, error estimate, and convergence of discrete solution, have been discussed (cf. [2, 8, 11, 15, 16, 18, 19, 32, 35]). There is also extensive numerical treatment for the corresponding time-dependent problem (cf. [1, 6, 9, 26, 34]). However, most of the treatments in these works are devoted either to semilinear equations where the diffusion coefficient is density independent or to quasi-linear equations with linear boundary conditions. The papers in [23, 26, 29] are for semilinear equations with linear boundary conditions where monotone iterative methods are used to develop computational algorithms. The same method is used in [25] for semilinear equations with nonlinear boundary conditions and in [26, 34] for time-dependent problems. On the other hand, the treatment in [2, 8, 11, 15, 16, 18, 19, 20, 35] is for linear Dirichlet or Neumann boundary condition, and most of the discussions are concerned with error estimates of solutions using the finite element method. In the above works, a unique solution to the problem is often assumed to exist, although the existence of more than one solution for nonlinear elliptic boundary problems occurs often in many physical problems (cf. [23]).

In this paper, we use the method of upper and lower solutions and its associated monotone iteration for the computation of numerical solutions of problems (1.1) and (1.1a). Our approach to the problem is to formulate it as a coupled system of a semilinear elliptic equation and an algebraic equation and then to develop monotone iterative schemes for the corresponding discrete system by the finite difference method (see also Remark 3.1). This approach makes it possible to apply the ideas and techniques for semilinear equations to the present quasi-linear boundary problems. The purpose of this paper is (i) to present three monotone iterative schemes (called Picard, Gauss–Seidel, and Jacobi iterations, respectively) for the computation of solutions of the finite difference system, including the existence of maximal and minimal solutions, and some comparison theorems among the three monotone iterations, and (ii) to show the convergence of the finite difference solution to the corresponding continuous solution as the mesh size tends to zero. Applications are given to some physical problems in heat transfer and combustion theory, and numerical results are given to some heat-conduction problems.

The plan of the paper is as follows: In section 2, we formulate the quasi-linear boundary problem (1.1) (or (1.1a)) as a coupled system of semilinear boundary problem and an algebraic equation and then develop a Picard type of monotone iteration for the finite difference system of the coupled equations. The Gauss–Seidel and Jacobi iterations are given in section 3, and some comparison theorems among these monotone iterations are given in section 4. Section 5 is devoted to the convergence of the maximal and minimal solutions to the corresponding maximal and minimal solutions of the continuous system. In section 6, we give some applications of the monotone iterations to some heat-conduction problems. Some numerical results with (and without) a known continuous solution are given in section 7 to demonstrate the monotone property of the iterative schemes and the reliability of the numerical computations.

2. A Picard-type monotone iteration. In order to develop computational schemes for numerical solutions of (1.1) or (1.1a), we form the problem as a coupled system of a semilinear elliptic boundary value problem and an algebraic equation. Define

$$w(x) \quad = \quad \int_0^{u(x)} D(s)ds \quad (x \in \overline{\Omega}). \tag{2.1}$$

Then $\nabla w = D(u)\nabla u$ and $\partial w/\partial \nu = D(u)\partial u/\partial \nu$. Since $dw/du = D(u)$, we see that if $D(u) > 0$ for $u \in \mathcal{S}_0$, then the inverse function of (2.1), denoted by $u = q(w)$, exists and $dq/dw = 1/D(u)$. Hence problem (1.1) may be written in the equivalent form

$$\begin{aligned} &-\nabla^2 w + \boldsymbol{c}\cdot\nabla w = f(x,u), \quad u = q(w) \quad (x \in \Omega), \\ &\partial w/\partial \nu = g(x,u) \qquad\qquad\qquad\quad (x \in \partial\Omega). \end{aligned} \tag{2.2}$$

Let $\gamma^{(l)}(x)$, $l = 1, 2$, be some nonnegative functions to be chosen, and define

$$\begin{aligned} F(x,u) \quad &= \quad f(x,u) + \gamma^{(1)}(x)\int_0^u D(s)ds \quad (x \in \Omega), \\ G(x,u) \quad &= \quad g(x,u) + \gamma^{(2)}(x)\int_0^u D(s)ds \quad (x \in \partial\Omega). \end{aligned} \tag{2.3}$$

Then, by adding $\gamma^{(1)}w$ and $\gamma^{(2)}w$ on both sides of the respective equations in (2.2) and using the relation (2.1), we obtain the equivalent system

$$\begin{aligned} &-\nabla^2 w + \boldsymbol{c}\cdot\nabla w + \gamma^{(1)}w = F(x,u), \quad u = q(w) \quad (x \in \Omega), \\ &\partial w/\partial \nu + \gamma^{(2)}w = G(x,u) \qquad\qquad\qquad\quad (x \in \partial\Omega). \end{aligned} \tag{2.4}$$

For the Dirichlet problem (1.1a), we replace the boundary condition in (2.4) (or (2.2)) by

$$w(x) \quad = \quad \int_0^{\xi(x)} D(s)ds \equiv \xi^*(x) \quad (x \in \partial\Omega) \tag{2.4a}$$

and refer to this system as problem (2.4a). It is obvious that u is a solution of (1.1) (resp., (1.1a)) if and only if (u, w) is a solution of (2.4) (resp., (2.4a)). Although the above system can be written as an uncoupled boundary value problem in w, we find it more convenient to treat it as a coupled system. In fact, our discretized system for numerical solutions is based on the form (2.4) or (2.4a).

Let $i = (i_1, \dots, i_p)$ be a multiple index with $i_\nu = 1, \dots, M_\nu$, and let $x_i = (x_{i_1}, \dots, x_{i_p})$ be a mesh point on $\overline{\Omega} \equiv \Omega \cup \partial\Omega$, where $\nu = 1, \dots, p$ and M_ν is the total number of intervals in the x_ν-direction. Denote by Λ, $\partial\Lambda$, and $\overline{\Lambda}$ the sets of mesh points of Ω, $\partial\Omega$, and $\overline{\Omega}$, respectively, and when no confusion arises we write $i \in \Lambda'$ when $x_i \in \Lambda'$, where Λ' stands for Λ, $\partial\Lambda$, or $\overline{\Lambda}$. Let h_ν be the spatial increment in the x_ν-direction, and let $u_i = u(x_i)$, $w_i = w(x_i)$, and $\boldsymbol{c}_i = (c^{(1)}(x_i), \dots, c^{(p)}(x_i))$. Define

$$\begin{aligned} &D(u_i) = D(u(x_i)), && q(w_i) = q(w(x_i)), \\ &F_i(u_i) = F(x_i, u(x_i)), && G_i(u_i) = G(x_i, u(x_i)). \end{aligned} \tag{2.5}$$

Then, by the central difference approximations

$$\begin{aligned} \Delta_p[w_i] &\equiv \sum_{\nu=1}^{p} \Delta^{(\nu)} w_i \equiv \sum_{\nu=1}^{p} h_\nu^{-2}[w(x_i + h_\nu e_\nu) - 2w(x_i) + w(x_i - h_\nu e_\nu], \\ \boldsymbol{c}_i \cdot \delta_p[w_i] &= \sum_{\nu=1}^{p} (c^{(\nu)}(x_i)/2h_\nu)[w(x_i + h_\nu e_\nu) - w(x_i - h_\nu e_\nu)] \end{aligned} \tag{2.6}$$

and the boundary approximation

$$\hat{B}[w_0] \quad = \quad [w(x_0) - w(\hat{x})]/|x_0 - \hat{x}| \quad (x_0 \in \overline{\Lambda}), \tag{2.7}$$

where e_ν is the unit vector in $\mathbb{R}^p$ with the νth component one and zero elsewhere and $\hat{x}$ is a suitable neighboring point of x_0 in Λ, we approximate (2.4) by the finite difference system

$$\begin{aligned} &-\Delta_p[w_i] + \boldsymbol{c}_i \cdot \delta_p[w_i] + \gamma_i^{(1)} w_i = F_i(u_i), \quad u_i = q(w_i) \quad (i \in \Lambda), \\ &\hat{B}[w_i] + \gamma_i^{(2)} w_i = G_i(u_i) \qquad\qquad\qquad\qquad (i \in \partial\Lambda). \end{aligned} \tag{2.8}$$

For the Dirichlet problem (2.4a), the boundary condition in (2.8) is replaced by

$$w_i = \xi_i^* \qquad (i \in \partial\Lambda), \tag{2.8a}$$

and the corresponding finite difference system is referred to as problem (2.8a). Although the boundary approximation in (2.7) can be approximated by a suitable central differencing scheme, the present formulation is more convenient in the discussion for the general multidimensional domain Ω (cf. [3, 12, 27, 29]).

To develop monotone iterative schemes for the solution of (2.8) or (2.8a), we impose the following basic hypothesis:

(H_1) (i) $D(u) > 0$ for $u \in \mathcal{S}_0$ and $h_\nu < |c^{(\nu)}(x_i)|^{-1}$ for $x_i \in \Lambda$ and $\nu = 1, \dots, p$.

(ii) $f(\cdot, u)$ and $g(\cdot, u)$ are C^1-functions of u, and there exist nonnegative functions $\gamma^{(1)}(x)$, $\gamma^{(2)}(x)$, not both identically zero, such that

$$\begin{aligned} \gamma^{(1)}(x) D(u) + f_u(x, u) \quad &\geq \quad 0, \\ &\text{for} \quad u \in \mathcal{S}_0 \quad (x \in \Omega, x' \in \partial\Omega), \\ \gamma^{(2)}(x') D(u) + g_u(x', u) \quad &\geq \quad 0, \end{aligned} \tag{2.9}$$

where $\mathcal{S}_0$ is a sector in $\mathbb{R}^1$ given by (2.12) below. It is clear that condition (2.9) is trivially satisfied (with $\gamma^{(1)}(x) = \gamma^{(2)}(x) = 0$) if $f(\cdot, u)$ and $g(\cdot, u)$ are either independent of u or nondecreasing in u for $u \in \mathcal{S}_0$. It is also satisfied by any functions $\gamma^{(1)}(x)$, $\gamma^{(2)}(x)$ satisfying

$$\gamma^{(1)}(x) \geq -f_u(x, u)/d_0, \quad \gamma^{(2)}(x') \geq -g_u(x', u)/d_0 \quad (u \in \mathcal{S}_0)$$

if $D(u) \geq d_0 > 0$ for $u \in \mathcal{S}_0$. Hence condition (2.9) is needed only for the degenerate case $D(0) = 0$, and $f(\cdot, u)$ and $g(\cdot, u)$ are not nondecreasing in u. Since, by (2.3),

$$\begin{aligned} F_u(x,u) &= f_u(x,u) + \gamma^{(1)}(x)D(u), \\ G_u(x,u) &= g_u(x,u) + \gamma^{(2)}(x)D(u), \end{aligned} \tag{2.10}$$

condition (2.9) implies that $F(\cdot, u)$ and $G(\cdot, u)$ are nondecreasing in u for $u \in \mathcal{S}_0$. The subset $\mathcal{S}_0$ is the sector between a pair of upper and lower solutions which are defined by the following.

DEFINITION 2.1. *A function $(\tilde{u}_i, \tilde{w}_i)$ is called an upper solution of* (2.8) *if*

$$\begin{aligned} &-\Delta_p[\tilde{w}_i] + \boldsymbol{c}_i \cdot \delta_p[\tilde{w}_i] + \gamma_i^{(1)}\tilde{w}_i \geq F_i(\tilde{u}_i), \quad \tilde{u}_i \geq q(\tilde{w}_i) \quad (i \in A), \\ &\hat{B}[\tilde{w}_i] + \gamma_i^{(2)}\tilde{w}_i \geq G_i(\tilde{u}_i) \qquad\qquad\qquad\qquad\quad (i \in \partial\Lambda). \end{aligned} \tag{2.11}$$

Similarly, $(\hat{u}_i, \hat{w}_i)$ is called a lower solution if it satisfies (2.11) *with the inequalities reversed. The pair $(\tilde{u}_i, \tilde{w}_i)$, $(\hat{u}_i, \hat{w}_i)$ is said to be ordered if $(\tilde{u}_i, \tilde{w}_i) \geq (\hat{u}_i, \hat{w}_i)$ for every $i \in \overline{\Lambda}$.*

For the Dirichlet problem (2.8a), the inequalities for $\tilde{w}_i$ and $\hat{w}_i$ on the boundary in (2.11) are replaced by

$$\tilde{w}_i \geq \xi_i^* \geq \hat{w}_i \qquad (i \in \partial\Lambda). \tag{2.11a}$$

It is obvious from the above definition that a solution (u_i, w_i) of (2.8) (or (2.8a)) is an upper solution as well as a lower solution of the corresponding problem. For a given pair of ordered upper and lower solutions, we set

$$\begin{aligned} \mathcal{S}_0 &= \{u_i \in \mathbb{R}^1; \hat{u}_i \leq u_i \leq \tilde{u}_i\}, \\ \mathcal{S} &= \{(u_i, w_i) \in \mathbb{R}^2; (\hat{u}_i, \hat{w}_i) \leq (u_i, w_i) \leq (\tilde{u}_i, \tilde{w}_i)\}. \end{aligned} \tag{2.12}$$

Also, for notational convenience, we define the linear operators

$$\begin{aligned} &L[w_i] = -\Delta_p[w_i] + \boldsymbol{c}_i \cdot \delta_p[w_i] + \gamma_i^{(1)}w_i \quad (i \in \Lambda), \\ &B[w_i] = \hat{B}[w_i] + \gamma_i^{(2)}w_i \qquad\qquad\qquad\quad (i \in \partial\Lambda). \end{aligned} \tag{2.13}$$

Using $\tilde{u}_i$ or $\hat{u}_i$ as an initial iteration, we can construct a sequence $\{u_i^{(m)}, w_i^{(m)}\}$ from the Picard type of iteration process

$$\begin{aligned} &L[w_i^{(m)}] = F_i(u_i^{(m-1)}) \quad (i \in \Lambda), \\ &B[w_i^{(m)}] = G_i(u_i^{(m-1)}) \quad (i \in \partial\Lambda), \\ &u_i^{(m)} = q(w_i^{(m)}) \qquad\quad (i \in \overline{\Lambda}), \qquad m = 1, 2, \ldots. \end{aligned} \tag{2.14}$$

For the Dirichlet problem (2.8a), the boundary condition in (2.14) is replaced by

$$w_i^{(m)} = \xi_i^* \qquad (i \in \partial\Lambda), \qquad m = 1, 2, \ldots. \tag{2.14a}$$

It is clear from (2.14) (with $m = 1$) that starting from any $u^{(0)}$ we can compute $w_i^{(1)}$ from the first two equations in (2.14) because $F_i(u^{(0)})$ and $G_i(u^{(0)})$ are known. Using the value of $w^{(1)}$ in the third equation of (2.14) gives $u_i^{(1)}$. Continuation of this process shows that the sequence $\{u^{(m)}, w^{(m)}\}$ is well defined and can be computed

from (2.14) for every $m = 1, 2, \ldots$. The same is true for the Dirichlet problem (2.14a). Denote the sequence by $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$ if $u_i^{(0)} = \tilde{u}_i$ and by $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ if $u_i^{(0)} = \hat{u}_i$, and refer to them as maximal and minimal sequences, respectively. Our aim is to show that each of these sequences converges monotonically to a solution of (2.8) or (2.8a), respectively. For this purpose, we state the following well-known positivity lemma (e.g., see [24]).

LEMMA 2.1. *Let hypothesis* (H_1) *hold, and let* $\gamma_i^{(l)} \geq 0$, $l = 1, 2$. *If* z_i *satisfies*

$$L[z_i] \geq 0 \quad \text{in} \quad \Lambda, \quad B[z_i] \geq 0 \quad \text{on} \quad \partial\Lambda, \tag{2.15}$$

and $\gamma_i^{(1)} + \gamma_i^{(2)} > 0$ *for at least one* $i \in \overline{\Lambda}$, *then either* $z_i > 0$ *in* Λ *or* $z_i \equiv 0$ *on* $\overline{\Lambda}$. *The same conclusion holds if*

$$L[z_i] \geq 0 \quad \text{in} \quad \Lambda, \quad z_i \geq 0 \quad \text{on} \quad \partial\Lambda \tag{2.16}$$

without the requirement of $\gamma_i^{(1)} + \gamma_i^{(2)} > 0$ *for some* i.

Lemma 2.1 is a discrete version of the maximum principle and is useful for proving the following monotone property of the maximal and minimal sequences.

LEMMA 2.2. *Under the hypothesis* (H_1), *the sequences* $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ *governed by* (2.14) *or* (2.14a) *possess the monotone property*

$$(\underline{u}_i^{(m)}, \underline{w}_i^{(m)}) \leq (\underline{u}_i^{(m+1)}, \underline{w}_i^{(m+1)}) \leq (\overline{u}_i^{(m+1)}, \overline{w}_i^{(m+1)}) \leq (\overline{u}_i^{(m)}, \overline{w}_i^{(m)}) (i \in \overline{\Lambda}) \tag{2.17}$$

for every $m = 0, 1, 2, \ldots$.

Proof. Let $\overline{z}_i^{(0)} = \overline{w}_i^{(0)} - \overline{w}_i^{(1)} \equiv \tilde{w}_i - \overline{w}_i^{(1)}$. By (2.11), (2.13), and (2.14),

$$\begin{aligned} L[\overline{z}_i^{(0)}] &= (-\Delta_p[\tilde{w}_i] + \boldsymbol{c}_i \cdot \delta_p[\tilde{w}_i] + \gamma_i^{(1)}\tilde{w}_i) - F_i(\tilde{u}_i) \geq 0, \\ B[\overline{z}_i^{(0)}] &= (\hat{B}[\tilde{w}_i] + \gamma_i^{(2)}\tilde{w}_i) - G_i(\tilde{u}_i) \geq 0. \end{aligned}$$

In view of Lemma 2.1, we have $\overline{z}_i^{(0)} \geq 0$, which gives $\overline{w}_i^{(0)} \geq \overline{w}_i^{(1)}$. Since $q(w_i)$ is nondecreasing in w_i, the above result and (2.11) imply that

$$\overline{u}_i^{(0)} - \overline{u}_i^{(1)} = \tilde{u}_i - q(\overline{w}^{(1)}) \geq \tilde{u}_i - q(\tilde{w}_i) \geq 0.$$

This proves $(\overline{u}_i^{(0)}, \overline{w}_i^{(0)}) \geq (\overline{u}_i^{(1)}, \overline{w}_i^{(1)})$. A similar argument gives $(\underline{u}_i^{(1)}, \underline{w}_i^{(1)}) \geq (\underline{u}_i^{(0)}, \underline{w}_i^{(0)})$. Moreover, by (2.14) and the nondecreasing property of $F_i(u_i)$ and $G_i(u_i)$,

$$\begin{aligned} L[\overline{w}_i^{(1)} - \underline{w}_i^{(1)}] &= F_i(\overline{u}_i^{(0)}) - F_i(\underline{u}_i^{(0)}) \geq 0, \\ B[\overline{w}_i^{(1)} - \underline{w}_i^{(1)}] &= G_i(\overline{u}_i^{(0)}) - G_i(\underline{u}_i^{(0)}) \geq 0. \end{aligned}$$

This leads to $\overline{w}_i^{(1)} \geq \underline{w}_i^{(1)}$. Again the nondecreasing property of $q(w_i)$ gives $\overline{u}_i^{(1)} - \underline{u}_i^{(1)} = q(\overline{w}_i^{(1)}) - q(\underline{w}_i^{(1)}) \geq 0$, which proves $(\overline{u}_i^{(1)}, \overline{w}_i^{(1)}) \geq (\underline{u}_i^{(1)}, \underline{w}_i^{(1)})$. The above conclusions show that (2.17) holds for $m = 0$. Assume, by induction, that (2.17) holds when m is replaced by $(m-1)$. Then, by (2.14),

$$\begin{aligned} L[\overline{w}_i^{(m)} - \overline{w}_i^{(m+1)}] &= F_i(\overline{u}_i^{(m-1)}) - F_i(\overline{u}_i^{(m)}) \geq 0, \\ B[\overline{w}_i^{(m)} - \overline{w}_i^{(m+1)}] &= G_i(\overline{u}_i^{(m-1)}) - G_i(\overline{u}_i^{(m)}) \geq 0. \end{aligned}$$

It follows from Lemma 2.1 that $\overline{w}_i^{(m)} \geq \overline{w}_i^{(m+1)}$. This implies that

$$\overline{u}_i^{(m)} - \overline{u}_i^{(m+1)} = q(\overline{w}_i^{(m)}) - q(\overline{w}_i^{(m+1)}) \geq 0,$$

which proves $(\overline{u}_i^{(m+1)}, \overline{w}_i^{(m+1)}) \leq (\overline{u}_i^{(m)}, \overline{w}_i^{(m)})$. A similar argument yields $(\underline{u}_i^{(m+1)}, \underline{w}_i^{(m+1)}) \geq (\underline{u}_i^{(m)}, \underline{w}_i^{(m)})$ and $(\overline{u}_i^{(m+1)}, \overline{w}_i^{(m+1)}) \geq (\underline{u}_i^{(m+1)}, \underline{w}_i^{(m+1)})$. The conclusion of the lemma follows from the principle of induction. The proof for the Dirichlet problem (2.14a) is similar and is omitted. □

Based on the monotone property (2.17), we have the following conclusion for the nonlinear boundary problem (2.8).

THEOREM 2.1. *Let $(\tilde{u}_i, \tilde{w}_i)$, $(\hat{u}_i, \hat{w}_i)$ be ordered upper and lower solutions of* (2.8)*, and let hypothesis* (H_1) *hold. Then $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$ converges monotonically to a maximal solution $(\overline{u}_i, \overline{w}_i)$ of* (2.8) *in $\mathcal{S}$, while $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ converges monotonically to a minimal solution $(\underline{u}_i, \underline{w}_i)$. Moreover,*

$$\begin{aligned}(\hat{u}_i, \hat{w}_i) \leq (\underline{u}_i^{(m)}, \underline{w}_i^{(m)}) \quad &\leq (\underline{u}_i^{(m+1)}, \underline{w}_i^{(m+1)}) \leq (\underline{u}_i, \underline{w}_i) \leq (\overline{u}_i, \overline{w}_i) \\ \leq (\overline{u}_i^{(m+1)}, \overline{w}_i^{(m+1)}) &\leq (\overline{u}_i^{(m)}, \overline{w}_i^{(m)}) \quad \leq (\tilde{u}_i, \tilde{w}_i), \quad m = 1, 2, \ldots,\end{aligned} \tag{2.18}$$

and if $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$ $(\equiv (u_i^*, w_i^*))$, then (u_i^*, w_i^*) is the unique solution of (2.8) *in $\mathcal{S}$.*

Proof. By the monotone property (2.17), the limits

$$\lim_{m\to\infty} (\overline{u}_i^{(m)}, \overline{w}_i^{(m)}) = (\overline{u}_i, \overline{w}_i), \quad \lim_{m\to\infty} (\underline{u}_i^{(m)}, \underline{w}_i^{(m)}) = (\underline{u}_i, \underline{w}_i) \tag{2.19}$$

exist and satisfy relation (2.18). Letting $m \to \infty$ in (2.14) shows that both $(\overline{u}_i, \overline{w}_i)$ and $(\underline{u}_i, \underline{w}_i)$ are solutions of (2.8). We show that these solutions are the respective maximal and minimal solutions in the sense that if (u_i, w_i) is any other solution of (2.8) in $\mathcal{S}$, then $(\underline{u}_i, \underline{w}_i) \leq (u_i, w_i) \leq (\overline{u}_i, \overline{w}_i)$ on $\overline{\Lambda}$.

Let $\overline{z}_i^{(m)} = \overline{w}_i^{(m)} - w_i$. By (2.8) and (2.14),

$$\begin{aligned}L[\overline{z}_i^{(m)}] &= F_i(\overline{u}_i^{(m-1)}) - F_i(u_i), \\ B[\overline{z}_i^{(m)}] &= G_i(\overline{u}_i^{(m-1)}) - G_i(u_i), \quad m = 1, 2, \ldots.\end{aligned}$$

Since $\hat{u}_i \leq u_i \leq \tilde{u}_i$ and $F_i(u_i)$ and $G_i(u_i)$ are nondecreasing functions of u_i, the above relation for $m = 1$ gives

$$\begin{aligned}L[\overline{z}_i^{(1)}] &= F_i(\tilde{u}_i) - F_i(u_i) \geq 0, \\ B[\overline{z}_i^{(1)}] &= G_i(\tilde{u}_i) - G_i(u_i) \geq 0.\end{aligned}$$

This yields $\overline{w}_i^{(1)} \geq w_i$. Hence $\overline{u}_i^{(1)} - u_i = q(\overline{w}_i^{(1)}) - q(w_i) \geq 0$, which proves $(\overline{u}_i^{(1)}, \overline{w}_i^{(1)}) \geq (u_i, w_i)$. As in the proof of Lemma 2.2, an induction argument shows that $(\overline{u}_i^{(m)}, \overline{w}_i^{(m)}) \geq (u_i, w_i)$ for every m. Letting $m \to \infty$ and using the relation in (2.19) lead to $(\overline{u}_i, \overline{w}_i) \geq (u_i, w_i)$. A similar argument gives $(\overline{u}_i, \overline{w}_i) \leq (u_i, w_i)$, which proves the maximal and minimal property of $(\overline{u}_i, \overline{w}_i)$ and $(\underline{u}_i, \underline{w}_i)$. It is obvious from this property that (u_i^*, w_i^*) is the unique solution of (2.8) in $\mathcal{S}$ if $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$. This proves the theorem. □

To ensure that $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$, it is necessary to impose some additional conditions on $f(\cdot, u)$ and $g(\cdot, u)$. A sufficient condition is given in the following.

(H_2) $f_u(x,u) \le 0$, $g_u(x',u) \le 0$ for $x \in \Omega$, $x' \in \partial\Omega$, $u \in \mathcal{S}_0$, and either $f_u(x,u) < 0$ for some $x \in \Omega$ or $g_u(x',u) < 0$ for some $x' \in \partial\Omega$.

THEOREM 2.2. *Let $(\tilde{u}_i, \tilde{w}_i)$, $(\hat{u}_i, \hat{w}_i)$ be ordered upper and lower solutions of* (2.8)*, and let hypotheses* (H_1)*,* (H_2) *hold. Then* $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$ $(\equiv (u_i^*, w_i^*))$ *and* (u_i^*, w_i^*) *is the unique solution of* (2.8) *in* $\mathcal{S}$.

Proof. Since, by (2.1) and (2.3),

$$F_i(u_i) = f_i(u_i) + \gamma_i^{(1)} w_i, \quad G_i(u_i) = g_i(u_i) + \gamma_i^{(2)} w_i,$$

the maximal and minimal solutions $(\overline{u}_i, \overline{w}_i)$, $(\underline{u}_i, \underline{w}_i)$ of (2.8) satisfy the equations

$$\begin{aligned} -\Delta_p[\overline{w}_i] + \boldsymbol{c}_i \cdot \delta_p[\overline{w}_i] &= f_i(\overline{u}_i), \quad \hat{B}[\overline{w}_i] = g_i(\overline{u}_i), \quad \overline{u}_i = q(\overline{w}_i), \\ -\Delta_p[\underline{w}_i] + \boldsymbol{c}_i \cdot \delta_p[\underline{w}_i] &= f_i(\underline{u}_i), \quad \hat{B}[\underline{w}_i] = g_i(\underline{u}_i), \quad \underline{u}_i = q(\underline{w}_i). \end{aligned}$$

A subtraction of the corresponding equations in the above relation and using the nonincreasing property of $f_i(u_i)$, $g_i(u_i)$ in (H_2) lead to

$$\begin{aligned} &-\Delta_p[\underline{w}_i - \overline{w}_i] + \boldsymbol{c}_i \cdot \delta_p[\underline{w}_i - \overline{w}_i] = f_i(\underline{u}_i) - f_i(\overline{u}_i) \ge 0, \\ &\hat{B}[\underline{w}_i - \overline{w}_i] = g_i(\underline{u}_i) - g_i(\overline{u}_i) \ge 0, \\ &\underline{u}_i - \overline{u}_i = q(\underline{w}_i) - q(\overline{w}_i). \end{aligned}$$

Applying Lemma 2.1 to the first two of the above relations gives $\underline{w}_i - \overline{w}_i \ge 0$. By Theorem 2.1, we obtain $\underline{w}_i = \overline{w}_i$, which ensures that $\underline{u}_i = \overline{u}_i$. This proves $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$ and thus the uniqueness of the solution. □

When the nonlinear boundary condition in (2.8) is replaced by the Dirichlet boundary condition (2.8a), the equation for $w_i^{(m)}$ on $\partial\Lambda$ in the iteration process (2.14) is replaced by (2.14a). By using the same argument as that in the proof of Theorems 2.1 and 2.2, we have the following analogous conclusions.

THEOREM 2.3. *Let $(\tilde{u}_i, \tilde{w}_i)$, $(\hat{u}_i, \hat{w}_i)$ be ordered upper and lower solutions of the Dirichlet problem* (2.8a)*, and let hypothesis* (H_1) *hold. Then all the conclusions in Theorem* 2.1 *hold true for the corresponding sequences* $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ *from* (2.14), (2.14a)*. Moreover, the uniqueness result in Theorem* 2.1 *holds if* $f_u(x,u) \le 0$ *for* $x \in \Omega$, $u \in \mathcal{S}_0$.

Proof. The proof follows from the same argument as that in the proof for problem (2.8) and is omitted. □

Remark 2.1. (a) If $D(u) = d_0$ is a positive constant, then all the conclusions in Theorems 2.1–2.3 hold true for the semilinear system (2.8) with $w_i = d_0 u_i$ and $q(w_i) = w_i/d_0$. In this situation, the iteration process (2.14) (or (2.14a)) is reduced to that in [25, 27]. On the other hand, Theorems 2.1 and 2.3 hold true for any $D(u) > 0$ if either $f_i(u_i) = f_i$ or $g_i(u) = g_i$ is independent of u since the requirement in (H_1) is trivially satisfied.

(b) For the uniqueness result in Theorem 2.2, the strict inequality $f_u(x,u) < 0$ for some $x \in \Omega$ or $g_u(x',u) < 0$ for some $x' \in \partial\Omega$ in hypothesis (H_2) is needed. For example, if $f(x,u) = f(x)$, $g(x',u) = g(x')$, and $D(u) = d_0$ are all independent of u, then problem (2.2) has no solution or infinite number of solutions depending on whether $\int_\Omega f(x)dx + d_0 \int_{\partial\Omega} g(x')dx'$ is nonzero or zero. However, for the Dirichlet problem (2.8a), this requirement is not needed.

(c) For each m, $(\overline{u}_i^{(m)}, \overline{w}_i^{(m)})$ is an upper bound of the maximal solution $(\overline{u}_i, \overline{w}_i)$, while $(\underline{u}_i^{(m)}, \underline{w}_i^{(m)})$ is a lower bound of the minimal solution $(\underline{u}_i, \underline{w}_i)$, and if $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$ $(\equiv (u_i^*, w_i^*))$, then they become upper and lower bounds of (u_i^*, w_i^*). In the latter situation, the difference $(\overline{u}_i^{(m)} - \underline{u}_i^{(m)}, \overline{w}_i^{(m)} - \underline{w}_i^{(m)})$ gives a maximum possible error of the solution and is often used as a stopping criterion in practical computation.

(d) In the hypothesis $(H_1) - (i)$, it is assumed that $h_\nu < |c^{(\nu)}(x_i)|^{-1}$ for $\nu = 1, \ldots, p$. If the convection coefficient $|\boldsymbol{c}|$ is extremely large (that is, convection dominate diffusion), then the conclusions in Theorems 2.1–2.3 remain true, provided that an upwind differencing scheme is used for the convection term $\boldsymbol{c} \cdot \nabla w$ (cf. [3, 12]).

3. Gauss–Seidel and Jacobi monotone iterations. In the Picard-type iteration (2.14), it is necessary to solve a linear algebraic system for each iteration. Since the number of equations in this system may be very large when the spatial domain Ω is of multiple dimension, it may require another iterative scheme for the computation of the iteration $(u_i^{(m)}, w_i^{(m)})$. To avoid additional iterations while maintaining the monotone convergence of the sequence, we consider two additional iteration processes, called Gauss–Seidel and Jacobi iterations. To describe these iterations, it is more convenient to write the finite difference system (2.8) (or (2.8a)) in vector form.

Let $M = (M_1 - 1) \cdots (M_p - 1)$ be the total number of interior mesh points in Λ. Define (column) vectors

$$\begin{aligned} &U = (u_1, \ldots, u_M)^T, \quad W = (w_1, \ldots, w_M)^T, \quad Q(W) = (q(w_1), \ldots, q(w_M))^T, \\ &F(U) = (F_1(u_1), \ldots, F_M(u_M))^T, \quad G(U) = (G_1(u_1), \ldots, G_M(u_M))^T, \end{aligned} \tag{3.1}$$

where $(\cdot)^T$ denotes the transpose of a row vector. Then the finite difference system (2.8) may be written in the vector form

$$\mathcal{A}W = F(U) + G(U), \quad U = Q(W), \tag{3.2}$$

where $\mathcal{A} = A + \Gamma$. The matrix A is, in general, an M by M block matrix which is associated with the diffusion-convection operator in (2.6) and the boundary approximation in (2.7), and Γ is a nonnegative diagonal matrix with its diagonal elements determined by $\gamma_i^{(1)}$ and $\gamma_i^{(2)}$ (cf. [25, 26, 27]) for some details). For the Dirichlet problem (2.8a), the vector form is given by

$$\mathcal{A}W = F(U) + \boldsymbol{\xi}^*, \quad U = Q(W), \tag{3.2a}$$

where $\boldsymbol{\xi}^*$ is associated with $(\xi_1^*, \ldots, \xi_M^*)^T$. Since our concern here is the mathematical structure of the finite difference system, detailed formulation of the above vector form will not be given here. However, we make the following hypothesis on the matrix A.

(H_3) The matrix $A = (a_{jk})$ is irreducible, and $a_{jj} > 0$, $a_{jk} \le 0$ for $k \ne j$ and

$$\sum_{k=1}^{M} a_{jk} \ge 0 \quad \text{for} \quad \text{all} \quad j = 1, \ldots, M. \tag{3.3}$$

It is easy to show from (2.6) and $h_\nu < |c^{(\nu)}(x_i)|^{-1}$ that for both the Neumann boundary problem (2.8) and the Dirichlet problem (2.8a) the conditions in (H_3) are all satisfied. In fact, for the Dirichlet problem (2.8a), A is symmetric, and strict inequality in (3.3) holds for at least one j, while for the Neumann problem (2.8)

condition (3.3) becomes

$$\sum_{k=1}^{M} a_{jk} = 0 \quad \text{for} \quad \text{all} \quad j = 1, \ldots, M. \tag{3.4}$$

The connectedness of Ω ensures that A is irreducible. The condition in (H_3) implies that A is an M-matrix, and for any nontrivial nonnegative diagonal matrix Γ the inverse $\mathcal{A}^{-1} = (A + \Gamma)^{-1}$ exists and is a positive matrix (cf. [33, 36]). Moreover, the smallest eigenvalue μ_0 of A is nonnegative, and its corresponding eigenvector may be chosen positive. For the boundary problem (2.8a), $\mu_0 > 0$, and A^{-1} exists and is a positive matrix (cf. [33]). In the following discussion, we consider the Dirichlet problem (3.2a) as a special case of (3.2) with $G(U) = \boldsymbol{\xi}^*$ and $\Gamma \equiv \text{diag}(\gamma_1^{(1)}, \ldots, \gamma_M^{(1)})$.

In terms of the above vector form, the definition of upper and lower solutions in Definition 2.1 is reduced to the following.

DEFINITION 3.1. *A vector* $(\tilde{U}, \tilde{W}) \in \mathbb{R}^M \times \mathbb{R}^M$ *is called an upper solution of* (3.2) *if*

$$\mathcal{A}\tilde{W} \geq F(\tilde{U}) + G(\tilde{U}), \quad \tilde{U} \geq Q(\tilde{W}). \tag{3.5}$$

Similarly, $(\hat{U}, \hat{W})$ *is called a lower solution if it satisfies* (3.5) *with the inequalities reversed.*

In the above definition, inequalities between vectors are always in the componentwise sense. It is easy to check that if the components $(\tilde{u}_i, \tilde{w}_i)$, $(\hat{u}_i, \hat{w}_i)$ of $(\tilde{U}, \tilde{W})$ and $(\hat{U}, \hat{W})$ satisfy the requirements in Definition 2.1, then $(\tilde{U}, \tilde{W})$ and $(\hat{U}, \hat{W})$ satisfy the requirements in Definition 3.1. For a given pair of ordered upper and lower solutions $(\tilde{U}, \tilde{W})$, $(\hat{U}, \hat{W})$ (that is, $(\tilde{U}, \tilde{W}) \geq (\hat{U}, \hat{W})$), we again set

$$\begin{aligned} \mathcal{S}_0 &= \{U \in \mathbb{R}^M; \hat{U} \leq U \leq \tilde{U}\}, \\ S &= \{(U, W) \in \mathbb{R}^M \times \mathbb{R}^M; (\hat{U}, \hat{W}) \leq (U, W) \leq (\tilde{U}, \tilde{W})\}. \end{aligned} \tag{3.6}$$

To describe the Gauss–Seidel and Jacobi iterations, we write the matrix $\mathcal{A}$ in the split form $\mathcal{A} = \mathcal{D} - \mathcal{L} - \mathcal{U}$, where $\mathcal{D}$, $-\mathcal{L}$, and $-\mathcal{U}$ are the diagonal, lower-off-diagonal, and upper-off-diagonal submatrices of $\mathcal{A}$, respectively. It is clear from hypothesis (H_3) that all the diagonal elements of $\mathcal{D}$ are positive and all the elements of $\mathcal{L}$ and $\mathcal{U}$ are nonnegative. Define a triangular matrix $\mathcal{G}$ and a diagonal matrix $\mathcal{J}$ by

$$\mathcal{G} = \mathcal{D} + \Gamma - \mathcal{L}, \quad \mathcal{J} = \mathcal{D} + \Gamma. \tag{3.7}$$

Then we have the following three types of iterations:

Picard iteration

$$\mathcal{A}W^{(m)} = F(U^{(m-1)}) + G(U^{(m-1)}), \quad U^{(m)} = Q(W^{(m)}). \tag{3.8}$$

Gauss–Seidel iteration

$$\mathcal{G}W^{(m)} = \mathcal{U}W^{(m-1)} + F(U^{(m-1)}) + G(U^{(m-1)}), \quad U^{(m)} = Q(W^{(m)}). \tag{3.9}$$

Jacobi iteration

$$\mathcal{J}W^{(m)} = (\mathcal{U} + \mathcal{L})W^{(m-1)} + F(U^{(m-1)}) + G(U^{(m-1)}), \quad U^{(m)} = Q(W^{(m)}). \tag{3.10}$$

For the Dirichlet problem (3.2a), we replace $G(U^{(m-1)})$ in (3.8)–(3.10) by $\boldsymbol{\xi}^*$ for every m. It is clear that the Picard iteration (3.8) is simply a vector representation of the iteration process (2.14).

To obtain monotone convergent sequences, we observe from hypothesis (H_3) that the inverse matrices $\mathcal{A}^{-1}$, $\mathcal{G}^{-1}$, and $\mathcal{J}^{-1}$ all exist and are positive matrices (cf. [33, 36]). This implies that given any initial iteration $U^{(0)}$ the sequence $\{U^{(m)}, W^{(m)}\}$ governed by any one of the iteration processes in (3.8), (3.9), and (3.10) is well defined. In each case, we denote the sequences by $\{\overline{U}^{(m)}, \overline{W}^{(m)}\}$ if $U^{(0)} = \tilde{U}$, and by $\{\underline{U}^{(m)}, \underline{W}^{(m)}\}$ if $U^{(0)} = \hat{U}$, and refer to them as maximal and minimal sequences, respectively. The following theorem gives the monotone convergence of these sequences.

THEOREM 3.1. *Let $(\tilde{U}, \tilde{W})$, $(\hat{U}, \hat{W})$ be ordered upper and lower solutions of* (3.2)*, and let $\{\overline{U}^{(m)}, \overline{W}^{(m)}\}$, $\{\underline{U}^{(m)}, \underline{W}^{(m)}\}$ be the maximal and minimal sequences governed by any one of the iteration processes in* (3.8)*,* (3.9)*, and* (3.10)*. Assume that hypotheses (H_1) and (H_3) hold. Then $\{\overline{U}^{(m)}, \overline{W}^{(m)}\}$ converges monotonically to a maximal solution $(\overline{U}, \overline{W})$ of* (3.2)*, and $\{\underline{U}^{(m)}, \underline{W}^{(m)}\}$ converges monotonically to a minimal solution $(\underline{U}, \underline{W})$. Moreover,*

$$
\begin{aligned}
(\hat{U}, \hat{W}) \le (\underline{U}^{(m)}, \underline{W}^{(m)}) \quad &\le (\underline{U}^{(m+1)}, \underline{W}^{(m-1)}) \le (\underline{U}, \underline{W}) \le (\overline{U}, \overline{W}) \\
&\le (\overline{U}^{(m+1)}, \overline{W}^{(m+1)}) \le (\overline{U}^{(m)}, \overline{W}^{(m)}) \quad \le (\tilde{U}, \tilde{W}), \quad m = 1, 2, \ldots,
\end{aligned} \tag{3.11}
$$

and if hypothesis (H_2) holds, then $(\overline{U}, \overline{W}) = (\underline{U}, \underline{W})$ $(\equiv (U^*, W^*))$ and (U^*, W^*) is the unique solution of (3.2) *in $\mathcal{S}$.*

Proof. Since the Picard iteration (3.8) is a vector representation of the iteration process (2.14), the conclusion of the theorem for the Picard iteration follows from Theorems 2.1 and 2.2. We show the monotone property

$$
\begin{aligned}
(\underline{U}^{(m)}, \underline{W}^{(m)}) &\le (\underline{U}^{(m+1)}, \underline{W}^{(m+1)}) \\
&\le (\overline{U}^{(m+1)}, \overline{W}^{(m+1)}) \le (\overline{U}^{(m)}, \overline{W}^{(m)}), \quad m = 0, 1, 2, \ldots,
\end{aligned} \tag{3.12}
$$

for the Gauss–Seidel iteration (3.9). It is obvious from (3.5), (3.9) (with $m = 1$), and $\mathcal{A} = \mathcal{G} - \mathcal{U}$ that

$$
\begin{aligned}
\mathcal{G}(\overline{W}^{(0)} - \overline{W}^{(1)}) &= \mathcal{G}\overline{W}^{(0)} - [\mathcal{U}\overline{W}^{(0)} + F(\overline{U}^{(0)}) + G(\overline{U}^{(0)})] \\
&= A\tilde{W} - F(\tilde{U}) - G(\tilde{U}) \ge 0.
\end{aligned}
$$

The positivity of $\mathcal{G}^{-1}$ implies that $\overline{W}^{(0)} \ge \overline{W}^{(1)}$. Using this relation in the second equation of (3.9) and applying the nondecreasing property of $Q(W)$ yield

$$
\overline{U}^{(0)} - \overline{U}^{(1)} = \tilde{U} - Q(\overline{W}^{(1)}) \ge \tilde{U} - Q(\tilde{W}) \ge 0.
$$

This proves $(\overline{U}^{(1)}, \overline{W}^{(1)}) \le (\overline{U}^{(0)}, \overline{W}^{(0)})$. A similar argument gives $(\underline{U}^{(0)}, \underline{W}^{(0)}) \le (\underline{U}^{(1)}, \underline{W}^{(1)})$. Moreover, by the nonnegative property of $\mathcal{U}$ and the nondecreasing property of $F(U)$ and $G(U)$, we have

$$
\mathcal{G}(\overline{W}^{(1)} - \underline{W}^{(1)}) = \mathcal{U}(\overline{U}^{(0)} - \underline{U}^{(0)}) + F(\overline{U}^{(0)}) - F(\underline{U}^{(0)}) + G(\overline{U}^{(0)}) - G(\underline{U}^{(0)}) \ge 0.
$$

This leads to $\overline{W}^{(1)} \ge \underline{W}^{(1)}$, and therefore $\overline{U}^{(1)} - \underline{U}^{(1)} = Q(\overline{W}^{(1)}) - Q(\underline{W}^{(1)}) \ge 0$. The above conclusions show that (3.12) holds for $m = 0$. Assume, by induction, that

(3.12) is satisfied when m is replaced by $m-1$ for some $m > 1$. Then, again by $\mathcal{U} \geq 0$ and the nondecreasing property of $F(U)$ and $G(U)$,

$$\mathcal{G}(\overline{W}^{(m)} - \overline{W}^{(m+1)}) = \mathcal{U}(\overline{W}^{(m-1)} - \overline{W}^{(m)}) + F(\overline{U}^{(m-1)}) - F(\overline{U}^{(m)}) + G(\overline{U}^{(m-1)}) - G(\overline{U}^{(m)}) \geq 0.$$

This gives $\overline{W}^{(m)} \geq \overline{W}^{(m+1)}$, and hence $\overline{U}^{(m)} - \overline{U}^{(m+1)} = Q(\overline{W}^{(m)}) - Q(\overline{W}^{(m+1)}) \geq 0$, which proves $(\overline{U}^{(m+1)}, \overline{W}^{(m+1)}) \leq (\overline{U}^{(m)}, \overline{W}^{(m)})$. The same reasoning shows that $(\underline{U}^{(m)}, \underline{W}^{(m)}) \leq (\underline{U}^{(m+1)}, \underline{W}^{(m+1)}) \leq (\overline{U}^{(m+1)}, \overline{W}^{(m+1)})$. The monotone property (3.12) follows from the principle of induction.

In view of (3.12), the limits

$$\lim_{m\to\infty} (\overline{U}^{(m)}, \overline{W}^{(m)}) = (\overline{U}, \overline{W}), \quad \lim_{m\to\infty} (\underline{U}^{(m)}, \underline{W}^{(m)}) = (\underline{U}, \underline{W}) \tag{3.13}$$

exist and satisfy relation (3.11). Letting $m \to \infty$ in (3.9) and using the relation $\mathcal{A} = \mathcal{G} - \mathcal{U}$ show that both $(\overline{U}, \overline{W})$ and $(\underline{U}, \underline{W})$ are solutions of (3.2). The maximal property of $(\overline{U}, \overline{W})$ and the minimal property of $(\underline{U}, \underline{W})$ follow from the argument in the proof of Theorem 2.1. Finally, if (H_2) holds, then, by (3.2),

$$\mathcal{A}(\underline{W} - \overline{W}) = [F(\underline{U}) + G(\underline{U})] - [F(\overline{U}) + G(\overline{U})] \geq 0.$$

The positivity of $\mathcal{A}^{-1}$ ensures $\underline{W} \geq \overline{W}$. In view of (3.11), we have $\underline{W} = \overline{W}$. This implies that $\underline{U} - \overline{U} = Q(\underline{W}) - Q(\overline{W}) = 0$, which shows that $(\overline{U}, \overline{W}) = (\underline{U}, \underline{W})$. The uniqueness of the solution in $\mathcal{S}^*$ follows from the maximal and minimal property of $(\overline{U}, \overline{W})$ and $(\underline{U}, \underline{W})$. This proves the theorem for the Gauss–Seidel iteration. The proof for the Jacobi iteration is similar and is omitted. □

For the Dirichlet problem (3.2a), we have the following analogous results.

THEOREM 3.2. *Let $(\tilde{U}, \tilde{W})$, $(\hat{U}, \hat{W})$ be ordered upper and lower solutions of* (3.2a), *and let hypotheses (H_1) and (H_3) hold. Then the maximal and minimal sequences obtained from any one of the iterations in* (3.8), (3.9), *and* (3.10), *where $\mathcal{G}(U^{(m-1)})$ is replaced by $\boldsymbol{\xi}^*$, possess the convergence property in Theorem* 3.1. *Moreover, the solution is unique in $\mathcal{S}$ if $f_u(x,u) \leq 0$ for $x \in \Omega$, $u \in \mathcal{S}_0$.*

Proof. Since the proof of the theorem is similar to that for Theorem 3.1, we omit the details. □

Remark 3.1. Although the discrete system (3.2a) for the Dirichlet boundary problem is formulated by the finite difference method, the same vector form can also be obtained by the finite element method. It can be shown by a suitable choice of the basis in the finite element method that the matrix A and the function $F(U)$ possess the same property as that given in hypotheses (H_1) and (H_3). Hence all the conclusions in Theorems 2.3 and 3.2 are directly applicable to the corresponding finite element system of the Dirichlet problem (2.8a).

4. Comparison of monotone sequences. In this section, we present some comparison results for the maximal and minimal sequences of the three iterative schemes (3.8), (3.9), and (3.10). The following comparison results among the three monotone iterations are for both the nonlinear boundary condition and Dirichlet boundary condition.

THEOREM 4.1. *Let the conditions in Theorem* 3.1 *be satisfied, and let* $(\{\overline{U}_P^{(m)}, \overline{W}_P^{(m)}\}, \{\underline{U}_P^{(m)}, \underline{W}_P^{(m)}\})$, $(\{\overline{U}_G^{(m)}, \overline{W}_G^{(m)}\}, \{\underline{U}_G^{(m)}, \underline{W}_G^{(m)}\})$, $(\{\overline{U}_J^{(m)}, \overline{W}_J^{(m)}\}, \{\underline{U}_J^{(m)}, \underline{W}_J^{(m)}\})$ *be the maximal-minimal sequences given by* (3.8), (3.9), *and* (3.10), *respectively, where* $(\overline{U}_P^{(0)}, \overline{W}_P^{(0)}) = (\overline{U}_G^{(0)}, \overline{W}_G^{(0)}) = (\overline{U}_J^{(0)}, \overline{W}_J^{(0)}) = (\tilde{U}, \tilde{W})$ *and* $(\underline{U}_P^{(0)}, \underline{W}_P^{(0)}) = (\underline{U}_G^{(0)}, \underline{W}_G^{(0)}) = (\underline{U}_J^{(0)}, \underline{W}_J^{(0)}) = (\hat{U}, \hat{W})$. *Then*

$$\begin{aligned} (\overline{U}_P^{(m)}, \overline{W}_P^{(m)}) &\le (\overline{U}_G^{(m)}, \overline{W}_G^{(m)}) \le (\overline{U}_J^{(m)}, \overline{W}_J^{(m)}), \\ (\underline{U}_P^{(m)}, \underline{W}_P^{(m)}) &\ge (\underline{U}_G^{(m)}, \underline{W}_G^{(m)}) \ge (\underline{U}_J^{(m)}, \underline{W}_J^{(m)}), \quad m = 1, 2, \ldots. \end{aligned} \tag{4.1}$$

Proof. We first prove the theorem for the maximal sequences between Picard and Gauss–Seidel iterations. Let $Z^{(m)} = \overline{W}_G^{(m)} - \overline{W}_P^{(m)}$ for $m = 1, 2, \ldots$. By a subtraction of (3.8) from (3.9) and using the relation $\mathcal{A} = \mathcal{G} - \mathcal{U}$ and $\mathcal{U}(\overline{U}_G^{(m-1)} - \overline{U}_G^{(m)}) \ge 0$, we have

$$\begin{aligned} \mathcal{A}Z^{(m)} &= (\mathcal{G} - \mathcal{U})\overline{W}_G^{(m)} - \mathcal{A}\overline{W}_P^{(m)} \\ &= \mathcal{U}(\overline{W}_G^{(m-1)} - \overline{W}_G^{(m)}) + [F(\overline{U}_G^{(m-1)}) + G(\overline{U}_G^{(m-1)})] \\ &\qquad - [F(\overline{U}_P^{(m-1)}) + G(\overline{U}_P^{(m-1)})] \\ &\ge F(\overline{U}_G^{(m-1)}) - F(\overline{U}_P^{(m-1)}) + G(\overline{U}_G^{(m-1)}) - G(\overline{U}_P^{(m-1)}), \\ \overline{U}_G^{(m)} - \overline{U}_P^{(m)} &= Q(\overline{W}_G^{(m)}) - Q(\overline{W}_P^{(m)}), \quad m = 1, 2, \ldots. \end{aligned} \tag{4.2}$$

Since $\overline{U}_G^{(0)} = \overline{U}_P^{(0)}$, the above relation for $m = 1$ yields $\mathcal{A}Z^{(1)} \ge 0$. This gives $Z^{(1)} \ge 0$ or, equivalently, $\overline{W}_P^{(1)} \le \overline{W}_G^{(1)}$. Using this result in the last equation in (4.2) (with $m = 1$) gives $\overline{U}_G^{(1)} - \overline{U}_P^{(1)} = Q(W_G^{(1)}) - Q(W_P^{(1)}) \ge 0$. This proves $(\overline{U}_P^{(1)}, \overline{W}_P^{(1)}) \le (\overline{U}_G^{(1)}, \overline{W}_G^{(1)})$. Assume $(\overline{U}_P^{(m-1)}, \overline{W}_P^{(m-1)}) \le (\overline{U}_G^{(m-1)}, \overline{W}_G^{(m-1)})$ for some $m > 1$. Then, by (4.2) and the nondecreasing property of $F(U)$ and $G(U)$, we obtain $\mathcal{A}Z^{(m)} \ge 0$, which yields $\overline{W}_P^{(m)} \le \overline{W}_G^{(m)}$. This implies that $\overline{U}_G^{(m)} - \overline{U}_P^{(m)} = Q(\overline{W}_G^{(m)}) - Q(\overline{W}_P^{(m)}) \ge 0$, and therefore $(\underline{U}_P^{(m)}, \underline{W}_P^{(m)}) \le (\overline{U}_G^{(m)}, W_G^{(m)})$. The first inequality in (4.1) for the maximal sequences follows from the principle of induction.

To show the second inequality for the maximal sequences between Gauss–Seidel and Jacobi iterations, we consider $\overline{Z}^{(m)} = \overline{W}_J^{(m)} - \overline{W}_G^{(m)}$ for $m = 1, 2, \ldots$. By a subtraction of (3.9) from (3.10) and using the relation $\mathcal{G} = \mathcal{J} - \mathcal{L}$ and $\mathcal{L}(\overline{W}_J^{(m-1)} - \overline{W}_J^{(m)}) \ge 0$, we obtain

$$\begin{aligned} \mathcal{G}\overline{Z}^{(m)} &= (\mathcal{J} - \mathcal{L})\overline{W}_J^{(m)} - \mathcal{G}\overline{W}_G^{(m)} \\ &= [\mathcal{U}\overline{W}_J^{(m-1)} + \mathcal{L}(\overline{W}_J^{(m-1)} - \overline{W}_J^{(m)}) + F(\overline{U}_J^{(m-1)}) + G(\overline{U}_J^{(m-1)})] \\ &\quad - [\mathcal{U}\overline{W}_G^{(m-1)} + F(\overline{U}_G^{(m-1)}) + G(\overline{U}_G^{(m-1)})] \\ &\ge \mathcal{U}\overline{Z}^{(m-1)} + F(\overline{U}_J^{(m-1)}) - F(\overline{U}_G^{(m-1)}) \\ &\quad + G(\overline{U}_J^{(m-1)}) - G(\overline{U}_G^{(m-1)}), \quad m = 1, 2, \ldots. \end{aligned} \tag{4.3}$$

Since $\overline{U}_J^{(0)} = \overline{U}_G^{(0)}$ and $\overline{W}_J^{(0)} = \overline{W}_G^{(0)}$, the above relation for $m = 1$ gives $\mathcal{G}\overline{Z}^{(1)} \geq 0$. The positivity of $\mathcal{G}^{-1}$ yields $\overline{Z}^{(1)} \geq 0$, that is, $\overline{W}_G^{(1)} \leq \overline{W}_J^{(1)}$. It follows from $\overline{U}_J^{(1)} - \overline{U}_G^{(1)} = Q(\overline{W}_J^{(1)}) - Q(\overline{W}_G^{(1)}) \geq 0$ that $(\overline{U}_G^{(1)}, \overline{W}_G^{(1)}) \leq (\overline{U}_J^{(1)}, \overline{W}_J^{(1)})$. Using the relation (4.3), an induction argument shows that $(\overline{U}_G^{(m)}, \overline{W}_G^{(m)}) \leq (\overline{U}_J^{(m)}, \overline{W}_J^{(m)})$ for every m. This proves the theorem for the maximal sequences. The proof for the minimal sequences is similar and is omitted. □

Remark 4.1. The comparison result in Theorem 4.1 implies that with the same initial iteration, which is either an upper solution or a lower solution, the sequence given by the Picard iteration converges faster than the one given by the Gauss–Seidel iteration, which in turn converges faster than the one by Jacobi iteration. This is true for both the maximal sequence and the minimal sequence. However, the Jacobi iteration is the simplest to use in practical computation, while the Picard iteration may require additional iterations when the size of the system is large.

5. Convergence of finite difference solutions. To investigate the convergence of a finite difference solution of (2.8) to its corresponding continuous solution of (2.4), we make use of a similar monotone iteration process for the continuous problem by the method of upper and lower solutions. To ensure the existence of a classical solution of (2.4), we assume that $f(x,\cdot)$, $g(x,\cdot)$, $\boldsymbol{c}(x)$, and $\gamma^{(l)}(x)$, $l = 1, 2$, are all Hölder continuous in x and hypothesis (H_1) is satisfied. For the Dirichlet problem (2.4a), it is assumed that $\xi^* \in C^{2+\alpha}(\Omega)$, where $C^{m+\alpha}(\Omega)$ $(m = 0, 1, 2, \ldots)$ is the set of functions $C^{(m)}(\Omega)$ that are Hölder continuous in Ω with exponents $\alpha \in (0, 1)$. The product space $C^{(m)}(\Omega') \times C^{(m)}(\Omega')$ is denoted by $\mathcal{C}^{(m)}(\Omega')$, where Ω' stands for Ω, $\partial\Omega$, or $\overline{\Omega}$. Set $|h| = h_1 + \cdots + h_p$, and define

$$\mathcal{L}[w] = -\nabla^2 w + \boldsymbol{c} \cdot \nabla w + \gamma^{(1)} w.$$

Then we have the following similar definition of upper and lower solutions for the continuous problem.

DEFINITION 5.1. *A function* $(\tilde{u}, \tilde{w}) \in \mathcal{C}^{(2)}(\Omega) \cap \mathcal{C}^{(0)}(\overline{\Omega})$ *is called an upper solution of* (2.4) *if*

$$\begin{aligned} &\mathcal{L}[\tilde{w}] \geq F(x, \tilde{u}), \quad \tilde{u} \geq q(\tilde{w}) \quad \textit{in} \quad \Omega, \\ &\partial\tilde{w}/\partial\nu + \gamma^{(2)}\tilde{w} \geq G(x, \tilde{u}) \qquad \textit{on} \quad \partial\Omega. \end{aligned} \tag{5.1}$$

Similarly, $(\hat{u}, \hat{w})$ *is called a lower solution if it satisfies* (5.1) *with the inequalities reversed.*

For the Dirichlet boundary problem (2.4a), the boundary requirement in (5.1) is replaced by

$$\tilde{w}(x) \geq \xi^*(x) \geq \hat{w}(x) \qquad (x \in \partial\Omega). \tag{5.1a}$$

The pair $(\tilde{u}, \tilde{w})$, $(\hat{u}, \hat{w})$ is said to be ordered if $(\tilde{u}, \tilde{w}) \geq (\hat{u}, \hat{w})$. For a given pair of ordered upper and lower solutions, we set

$$\begin{aligned} \mathcal{S}_0^* &\equiv \{u \in C(\overline{\Omega}); \quad \hat{u} \leq u \leq \tilde{u}\}, \\ \mathcal{S}^* &\equiv \{(u, w) \in \mathcal{C}(\overline{\Omega}); \quad (\hat{u}, \hat{w}) \leq (u, w) \leq (\tilde{u}, \tilde{w})\}. \end{aligned} \tag{5.2}$$

Using either $\tilde{u}$ or $\hat{u}$ as an initial iteration, we construct a sequence from the linear iteration process

$$\begin{aligned} &\mathcal{L}[w^{(m)}] = F(x, u^{(m-1)}) && (x \in \Omega), \\ &\partial w^{(m)}/\partial \nu + \gamma^{(2)} w^{(m)} = G(x, u^{(m-1)}) && (x \in \partial\Omega), \\ &u^{(m)} = q(w^{(m)}) && (x \in \overline{\Omega}), \quad m = 1, 2, \ldots . \end{aligned} \tag{5.3}$$

For the Dirichlet problem (2.4a), we replace the boundary condition in (5.3) by

$$w^{(m)}(x) = \xi^*(x) \quad (x \in \partial\Omega), \quad m = 1, 2, \ldots . \tag{5.3a}$$

Denote the sequence by $\{\overline{u}^{(m)}, \overline{w}^{(m)}\}$ if $u^{(0)} = \tilde{u}$ and by $\{\underline{u}^{(m)}, \underline{w}^{(m)}\}$ if $u^{(0)} = \hat{u}$, and refer to them as maximal and minimal sequences, respectively. The following theorem from [28] is analogous to Theorem 2.1.

THEOREM 5.1. *Let $(\tilde{u}, \tilde{w})$, $(\hat{u}, \hat{w})$ be ordered upper and lower solutions of* (2.4) *(or* (2.4a)*, and let hypothesis (H_1) hold. Then $\{\overline{u}^{(m)}, \overline{w}^{(m)}\}$ converges monotonically to a maximal solution $(\overline{u}, \overline{w})$ of* (2.4) *in $\mathcal{S}^*$, while $\{\underline{u}^{(m)}, \underline{w}^{(m)}\}$ converges monotonically to a minimal solution $(\underline{u}, \underline{w})$. Moreover,*

$$\begin{aligned} (\hat{u}, \hat{w}) &\leq (\underline{u}^{(m)}, \underline{w}^{(m)}) \leq (\underline{u}^{(m+1)}, \underline{w}^{(m+1)}) \leq (\underline{u}, \underline{w}) \leq (\overline{u}, \overline{w}) \\ &\leq (\overline{u}^{(m+1)}, \overline{w}^{(m+1)}) \leq (\overline{u}^{(m)}, \overline{w}^{(m)}) \leq (\tilde{u}, \tilde{w}), \quad m = 1, 2, \ldots , \end{aligned} \tag{5.4}$$

and if either hypothesis (H_2) is satisfied or $(\overline{u}, \overline{w}) = (\underline{u}, \underline{w})$ $(\equiv (u^*, w^*))$, then (u^*, w^*) is the unique solution of (2.4)*. The same conclusions hold true for the Dirichlet boundary problem* (2.4a)*.*

Based on Theorems 2.1 and 5.1, we show the convergence of the maximal and minimal finite difference solutions to their respective maximal and minimal solutions of the continuous problem for $x_i \in \overline{\Lambda}^*$, where $\overline{\Lambda}^*$ is a fixed partition of $\overline{\Omega}$. It is assumed that every refinement of $\overline{\Lambda}^*$ contains $\overline{\Lambda}^*$ and there exist ordered upper and lower solutions $((\tilde{u}_i, \tilde{w}_i), (\hat{u}_i, \hat{w}_i))$ and $((\tilde{u}(x), \tilde{w}(x)), (\hat{u}(x), \hat{w}(x)))$ of (2.4) and (2.8), respectively (or (2.4a) and (2.8a), respectively). We also assume that given any $\epsilon > 0$ there exists $\delta > 0$ such that

$$|\tilde{u}(x_i) - \tilde{u}_i| + |\tilde{w}(x_i) - \hat{w}_i| < \epsilon, \ |\hat{u}(x_i) - \hat{u}_i| + |\hat{w}(x_i) - \hat{w}_i| < \epsilon \text{ whenever } |h| < \delta. \tag{5.5}$$

THEOREM 5.2. *Let hypothesis (H_1) and condition* (5.5) *be satisfied, and let $((\overline{u}_i, \overline{w}_i), (\underline{u}_i, \underline{w}_i))$ and $((\overline{u}(x), \overline{w}(x)), (\underline{u}(x), \underline{w}(x)))$ be the respective maximal-minimal solutions of* (2.4) *and* (2.8) *(or* (2.4a) *and* (2.8a)*). Then, as $|h| \to 0$,*

$$(\overline{u}_i, \overline{w}_i) \to (\overline{u}(x_i), \overline{w}(x_i)) \quad \text{and} \quad (\underline{u}_i, \underline{w}_i) \to (\underline{u}(x_i), \underline{w}(x_i)) \tag{5.6}$$

at every point $x_i \in \overline{\Lambda}^*$. The same convergence result holds true for the Dirichlet boundary problem (2.4a) *and* (2.8a)*.*

Proof. It suffices to show that given any $\epsilon' > 0$ there exists $\delta' > 0$ such that for every $x_i \in \overline{\Lambda}^*$

$$|\overline{u}(x_i) - \overline{u}_i| + |\overline{w}(x_i) - \overline{w}_i| < \epsilon', \ |\underline{u}(x_i) - \underline{u}_i| + |\underline{w}(x_i) - \underline{w}_i| < \epsilon' \text{ when } |h| < \delta'. \tag{5.7}$$

We prove (5.7) for the maximal solutions $(\overline{u}(x_i), \overline{w}(x_i))$ and $(\overline{u}_i, \overline{w}_i)$ because the proof for the minimal solutions is similar. Let $\{\overline{u}^{(m)}(x_i), \overline{w}^{(m)}(x_i)\}$, $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$ be the

respective maximal sequences of (2.4) and (2.8). By Theorems 2.1 and 5.1, there exists $m^* \geq 1$ such that for all $m \geq m^*$ and $i \in \overline{\Lambda}^*$

$$|\overline{u}^{(m)}(x_i) - \overline{u}(x_i)| + |\overline{u}_i^{(m)} - \overline{u}_i| < \epsilon'/3, \quad |\overline{w}^{(m)}(x_i) - \overline{w}(x_i)| + |\overline{w}_i^{(m)} - \overline{w}_i| < \epsilon'/3.$$

Since

$$|\overline{u}(x_i) - \overline{u}_i| \leq |\overline{u}(x_i) - \overline{u}^{(m)}(x_i)| + |\overline{u}^{(m)}(x_i) - \overline{u}_i^{(m)}| + |\overline{u}_i^{(m)} - \overline{u}_i|,$$
$$|\overline{w}(x_i) - \overline{w}_i| \leq |\overline{w}(x_i) - \overline{w}^{(m)}(x_i)| + |\overline{w}^{(m)}(x_i) - \overline{w}_i^{(m)}| + |\overline{w}_i^{(m)} - \overline{w}_i|,$$

condition (5.7) is fulfilled if, for some $m \geq m^*$,

$$|\overline{u}^{(m)}(x_i) - \overline{u}_i^{(m)}| + |\overline{w}^{(m)}(x_i) - \overline{w}_i^{(m)}| < \epsilon'/3 \quad \text{when} \quad |h| < \delta'. \tag{5.8}$$

It is easily seen from (5.3) and the central difference and boundary approximations in (2.6) and (2.7) that

$$\begin{aligned} -L[\overline{w}^{(m)}(x_i)] &= F_i(\overline{u}^{(m-1)}(x_i)) + o^{(m)}(|h|) \quad (x_i \in \Lambda^*), \\ B[\overline{w}^{(m)}(x_i')] &= G_i(\overline{u}^{(m-1)}(x_i')) + o^{(m)}(|h|) \quad (x_i' \in \partial\overline{\Lambda}^*), \\ \overline{u}^{(m)}(x_i) &= q(\overline{w}^{(m)}(x_i)) + o^{(m)}(|h|) \quad (x_i \in \Lambda^*), \end{aligned} \tag{5.9}$$

where $o^{(m)}(|h|) \to 0$ as $|h| \to 0$. Let $(v_i^{(m)}, z_i^{(m)}) = (\overline{u}^{(m)}(x_i) - \overline{u}_i^{(m)}, \overline{w}^{(m)}(x_i) - \overline{w}_i^{(m)})$. Then a subtraction of (2.14) from (5.9) gives

$$\begin{aligned} L[z_i^{(m)}] &= F_i(\overline{u}^{(m-1)}(x_i)) - F_i(\overline{u}_i^{(m-1)}) + o^{(m)}(|h|), \\ B[z_i^{(m)}] &= G_i(\overline{u}^{(m-1)}(x_i')) - G_i(\overline{u}_i^{(m-1)}) + o^{(m)}(|h|), \\ v_i^{(m)} &= q(\overline{w}^{(m)}(x_i)) - q(\overline{w}_i^{(m)}) + o^{(m)}(|h|). \end{aligned}$$

In vector form, the above system is equivalent to

$$\begin{aligned} \mathcal{A}Z^{(m)} &= [F(\overline{U}^{(m-1)}(x)) + G(\overline{U}^{(m-1)}(x))] - [F(\overline{U}^{(m-1)}) + G(\overline{U}^{(m-1)})] + O^{(m)}(|h|) \\ V^{(m)} &= Q(\overline{W}^{(m)}(x)) - Q(\overline{W}^{(m)}) + O^{(m)}(|h|), \end{aligned} \tag{5.10}$$

where $Z^{(m)} = (z_1^{(m)}, \ldots, z_M^{(m)})^T$, $\overline{U}^{(m)}(x) = (\overline{u}^{(m)}(x_1), \ldots, \overline{u}^{(m)}(x_M))^T$, etc., and for any convenient norm in $\mathbb{R}^M$, $||O^{(m)}(|h|)|| \to 0$ as $|h| \to 0$. By the positivity of $\mathcal{A}^{-1}$, the above relation implies that

$$\begin{aligned} |Z^{(m)}| &\leq \mathcal{A}^{-1}[(k_f + k_g)|V^{(m-1)}| + |O^{(m)}(|h|)|], \\ |V^{(m)}| &\leq k_q|Z^{(m)}| + |O^{(m)}(|h|)|, \end{aligned} \tag{5.11}$$

where k_f, k_g, and k_q are the Lipschitz constants of $F(U)$, $G(U)$, and $Q(W)$, respectively, and $|Y| = |y_1| + \cdots + |y_M|$ for any vector $Y = (y_1, \ldots, y_M)^T$.

It is well known that given any $\epsilon_1 > 0$ there exist a matrix norm and a vector norm in $\mathbb{R}^M$ such that

$$||\mathcal{A}^{-1}|| \leq (\mu_0 + \overline{\gamma} - \epsilon_1)^{-1} \equiv \sigma, \quad ||\mathcal{A}^{-1}Y|| \leq \sigma||Y||$$

for every $Y \in \mathbb{R}^M$, where μ_0 is the smallest eigenvalue of A and $\overline{\gamma} \equiv \max\{\gamma_i^{(l)}; i = 1, \ldots, M, l = 1, 2\}$ (cf. [36]). Since $\mu_0 \geq 0$, it suffices to choose $\sigma = (\overline{\gamma} - \epsilon)^{-1}$, which

is independent of A and therefore is independent of $|h|$. Using this relation in (5.11) leads to

$$||Z^{(m)}|| \le \sigma(k_f + k_g)||V^{(m-1)}|| + ||O^{(m)}(|h|)||,$$
$$||V^{(m)}|| \le k_q||Z^{(m)}|| + ||O^{(m)}(|h|)||,$$

where $||O^{(m)}(|h|)|| \to 0$ as $|h| \to 0$. For notational convenience, we define

$$r^{(m)} = ||V^{(m)}||, \quad s^{(m)} = ||Z^{(m)}||, \quad c = \max\{k_q, \sigma(k_f + k_g)\}.$$

Then the above inequalities yield

$$\begin{aligned} &s^{(m)} \le cr^{(m-1)} + o(|h|), \\ &r^{(m)} \le cs^{(m)} + o(|h|), \qquad m = 1, 2, \dots, \end{aligned} \tag{5.12}$$

where $o(|h|) = \sup\{||O^{(m)}(|h|)||; m \ge 1\}$.

Consider the case $m = 1$. Since $r^{(0)} = ||V^{(0)}|| = ||\overline{U}^{(0)}(x) - \overline{U}^{(0)}||$ and $\overline{U}^{(0)}(x) = (\tilde{u}(x_1), \dots, \tilde{u}(x_M))^T$, $\overline{U}^{(0)} = (\tilde{u}_1, \dots, \tilde{u}_M)^T$, we see from (5.5) that for any $\epsilon_0 > 0$ there exists $\delta_0 > 0$ such that $O(|h|) < \epsilon_0$ when $|h| < \delta_0$. In view of (5.12), we have

$$s^{(1)} \le cr^{(0)} + o(|h|) \le c\epsilon_0 + \epsilon_0 = (1 + c)\epsilon_0,$$
$$r^{(1)} \le cs^{(1)} + o(|h|) \le c(1 + c)\epsilon_0 + \epsilon_0 = (c^2 + c + 1)\epsilon_0.$$

An induction argument gives

$$s^{(2)} = (c^3 + c^2 + c + 1)\epsilon_0, \qquad r^{(2)} \le (c^4 + c^3 + c^2 + 1)\epsilon_0,$$
$$\vdots$$
$$s^{(m)} = (c^{2m-1} + c^{2m-2} + \cdots + 1)\epsilon_0, \quad r^{(m)} \le (c^{2m} + c^{2m-1} + \cdots + 1)\epsilon_0.$$

Let $m \ge m^*$ be fixed. Then by choosing ϵ_0 sufficiently small there exists $\delta' > 0$ such that $s^{(m)} + r^{(m)} < \epsilon'/3$ when $|h| < \delta'$. This is equivalent to

$$||V^{(m)}|| + ||Z^{(m)}|| < \epsilon'/3 \quad \text{when} \quad |h| < \delta'.$$

The above relation implies that (5.8) holds because the components of $V^{(m)}$ and $Z^{(m)}$ are $(\overline{u}^{(m)}(x_i) - \overline{u}_i^{(m)})$ and $(\overline{w}^{(m)}(x_i) - \overline{w}_i^{(m)})$, respectively. This proves (5.7) for the maximal solutions. Proofs for the minimal solutions and for the Dirichlet problem are similar and are omitted. □

An immediate consequence of Theorem 5.2 is the following result.

THEOREM 5.3. *Let the conditions in Theorem* 5.2 *be satisfied. If hypothesis* (H_2) *holds, then each of the problems* (2.4) *and* (2.8) *has a unique solution* $(u^*(x), w^*(x))$ *and* (u_i^*, w_i^*)*, and at every mesh point* $x_i \in \overline{\Lambda}^*$*,*

$$(u_i^*, w_i^*) \to (u^*(x_i), w^*(x_i)) \quad \textit{as} \quad |h| \to 0. \tag{5.13}$$

The same is true for the Dirichlet problems (2.4a) *and* (2.8a) *if* $f_u(x, u) \le 0$.

6. Application to model problems. In this section, we apply the results of the previous sections to some model problems in heat transfer and chemical engineering as illustrations. In the application of the iterative scheme (2.14) (or (2.14a), we make use of the relation

$$
\begin{aligned}
F_i(u_i^{(m-1)}) &= \gamma_i^{(1)} \int_0^{u_i^{(m-1)}} D(s)ds + f_i(u_i^{(m-1)}) = \gamma_i^{(1)} w_i^{(m-1)} + f_i(u_i^{(m-1)}), \\
G_i(u_i^{(m-1)}) &= \gamma_i^{(2)} \int_0^{u_i^{(m-1)}} D(s)ds + g_i(u_i^{(m-1)}), \\
&= \gamma_i^{(2)} w_i^{(m-1)} + g_i(u_i^{(m-1)}), \quad m = 1, 2, \ldots,
\end{aligned} \tag{6.1}
$$

to obtain an equivalent iterative scheme in the form

$$
\begin{aligned}
L[w_i^{(m)}] &= \gamma_i^{(1)} w_i^{(m-1)} + f_i(u_i^{(m-1)}), \\
B[w_i^{(m)}] &= \gamma_i^{(2)} w_i^{(m-1)} + g_i(u_i^{(m-1)}), \\
u_i^{(m)} &= q(w_i^{(m)}), m = 1, 2, \ldots.
\end{aligned} \tag{6.2}
$$

This form of iteration avoids the computation of the integral term $\int_0^u D(s)ds$ in each iteration, especially when it cannot be given in explicit form. The value of $u_i^{(m)}$ in (6.2) can be solved from the algebraic equation $w_i^{(m)} = D(u_i^{(m)})$ if the inverse function $q(w)$ cannot be explicitly given. For Dirichlet boundary problems, the boundary condition in (6.2) is replaced by

$$
w_i^{(m)} = \xi_i^*, \qquad m = 1, 2, \ldots, \tag{6.2a}
$$

where ξ_i^* is assumed nonnegative.

Some heat-conduction problems. In the heat-conduction problem, the thermal conductivity is often considered temperature dependent, and the boundary or internal source may also be temperature dependent. A frequently assumed form of the source function is based on the so-called Boltzmann fourth-power law, which is given in the form $\sigma(a^4 - u^4)$, where $u \equiv u(x)$ denotes the steady-state temperature distribution, $a \equiv a(x)$ is the surrounding temperature, and σ is a positive constant. This leads to the boundary value problem (1.1) with

$$
f(x, u) = 0, \quad g(x, u) = \sigma(a^4 - u^4) \tag{6.3}
$$

if the fourth-power law applies to the boundary surface and

$$
f(x, u) = \sigma(a^4 - u^4), \quad g(x, u) = 0 \tag{6.4}
$$

if the law applies to the conducting medium (cf. [2, 10, 13, 22, 23]), The above heat-conduction problem, including some numerical aspect and the corresponding time-dependent system, has been investigated by many workers (e.g., see [13, 14, 21, 22, 23, 24, 25]). Although various forms of the thermal conductivity $D(u)$ have been assumed, including the special cases of $D(u) = d_0 + d_1 u$ and $D(u) = e^{\alpha u}$, we need only the positivity of $D(u) \geq d_0 > 0$ in the construction of upper and lower solutions.

It is easy to verify that the functions $f(x, u)$, $g(x, u)$ in either (6.3) or (6.4) satisfy hypotheses (H_1) and (H_2) for $u \geq 0$ (that is, $\mathcal{S}_0 = R^+$). Hence, by Theorems 2.1

and 2.2, the finite difference system (2.8), (6.3) (that is, problem (2.8) with $f(x,u)$, $g(x,u)$ given by (6.3)) has a unique nonnegative solution (u_i, w_i) if there exists a pair of ordered nonnegative upper and lower solutions. The same is true for the system (2.8), (6.4) (or (2.8a), (6.4)). Moreover, by the positivity lemma, Lemma 2.1, for discrete boundary problems, the solution (u_i, w_i) is positive in Λ. Consider the case where $f(x,u)$ and $g(x,u)$ are given by (6.3). It is easy to verify from $f_i(u_i) = 0$, $g_i(0) = \sigma a_i^4 > 0$ and $g_i(\overline{a}) \le 0$ for any constant $\overline{a} \ge \max\{a_i, i \in \overline{\Lambda}\}$ that the pair

$$(\tilde{u}_i, \tilde{w}_i) = (\overline{a}, \overline{a}^*), \quad (\hat{u}_i, \hat{w}_i) = (0,0), \tag{6.5}$$

where $\overline{a}^* = \int_0^{\overline{a}} D(s)ds$, are ordered upper and lower solutions of (2.8), (6.3). By Theorems 2.1 and 2.2, problem (2.8), (6.3) has a unique positive solution (u_i^*, w_i^*) and $(u_i^*, w_i^*) \le (\overline{a}, \overline{a}^*)$.

In order to compute the solution (u_i^*, w_i^*) by any one of the iterative schemes in (3.8), (3.9), and (3.10), we need to find the functions $\gamma^{(1)}(x)$, $\gamma^{(2)}(x)$ that satisfy the condition (2.9) in (H_1). Since $f_u(x,u) = 0$, $g_u(x,u) = -4\sigma u^3$ and $\mathcal{S}_0 = [0, \overline{a}]$, it suffices to choose $\gamma^{(1)} = 0$, $\gamma^{(2)} = 4\sigma\overline{a}^3/d_0$ (or any $\gamma^{(2)} \ge 4\sigma\overline{a}^3/d_0$). Using the above values of $(\gamma^{(1)}, \gamma^{(2)})$ and the functions

$$f_i(u_i^{(m-1)}) = 0, \quad g_i(u_i^{(m-1)}) = \sigma\left[a_i^4 - (u_i^{(m-1)})^4\right] \tag{6.6}$$

in the iteration process (6.2), we can compute the maximal and minimal sequences $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$, where $(\overline{u}_i^{(0)}, \overline{w}_i^{(0)}) = (\overline{a}, \overline{a}^*)$ and $(\underline{u}_i^{(0)}, \underline{w}_i^{(0)}) = (0,0)$. The function $q(w_i)$ in (6.2) is determined from (2.1) and can sometimes be obtained explicitly. For example, if $D(u) = d_0 e^{\alpha u}$ or $D(u) = d_0 + d_1 u$, where d_0 and d_1 are positive constants and $\alpha \ne 0$, then

$$q(w) = \alpha^{-1}\ln[1 + (\alpha/d_0)w] \quad \text{and} \quad q(w) = d_1^{-1}[d_0^2 + 2d_1 w)^{1/2} - d_0], \tag{6.7}$$

respectively. In the special case of the linear function $D(u)$, which will be used in our numerical computations, the equation $u_i^{(m)} = q(w_i^{(m)})$ in (6.2) is given explicitly by

$$u_i^{(m)} \quad = \quad d_1^{-1}[(d_0^2 + 2d_1 w^{(m)})^{1/2} - d_0]. \tag{6.8}$$

By writing the finite difference problem (2.8), (6.3) in the vector form (3.2), we can also compute the maximal and minimal sequences $\{\overline{U}^{(m)}, \overline{W}^{(m)}\}$, $\{\underline{U}^{(m)}, \underline{W}^{(m)}\}$ from either the Gauss–Seidel iteration (3.9) or the Jacobi iteration (3.10), where $(\overline{U}^{(0)}, \overline{W}^{(0)}) = (\overline{a}E, \overline{a}^*E)$, $(\underline{U}^{(0)}, \underline{W}^{(0)}) = (0,0)$, and $E = (1, \ldots, 1)^T \in \mathbb{R}^M$. By Theorem 3.1, these sequences converge monotonically to the unique solution (U^*, W^*), where $U^* = (u_1^*, \ldots, u_M^*)^T$, and $W^* = (w_1^*, \ldots, w_M^*)^T$.

To guarantee the convergence of the finite difference solution of this model, we observe that the pair in (6.5) are also ordered upper and lower solutions of the continuous problem (2.4), (6.3). By Theorem 5.1, the continuous problem (2.4), (6.3) has a unique solution $(u^*(x), w^*(x))$. Since condition (5.5) is trivially satisfied, Theorem 5.2 implies that $(u_i^*, w_i^*) \to (u^*(x_i), w^*(x_i))$ as $|h| \to 0$. To summarize the above conclusions, we have the following.

THEOREM 6.1. *Let $f(x,u)$, $g(x,u)$ be given by* (6.3) *and $D(u) \ge d_0 > 0$, and let $(\overline{u}_i^{(0)}, \overline{w}^{(0)}) = (\overline{a}, \overline{a}^*)$, $(\underline{u}_i^{(0)}, \underline{w}_i^{(0)}) = (0,0)$, and $(\gamma_i^{(1)}, \gamma_i^{(2)}) = (0, 4\sigma\overline{a}^3/d_0)$, where $\overline{a} \ge a(x)$ on $\overline{\Omega}$. Then the following statements hold true for the problem* (2.8), (6.3)*:*

(i) *A unique positive solution (u_i^*, w_i^*) exists and is bounded by $(\overline{a}, \overline{a}^*)$.*

(ii) *The maximal sequence* $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$ *from* (6.2) *converges monotonically from above to* (u_i^*, w_i^*), *while the minimal sequence* $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ *converges monotonically from below to* (u_i^*, w_i^*).

(iii) *In vector form, the sequences* $\{\overline{U}^{(m)}, \overline{W}^{(m)}\}$, $\{\underline{U}^{(m)}, \underline{W}^{(m)}\}$ *given either by* (3.9) *or by* (3.10) *converge monotonically to the unique solution* (U^*, W^*).

(iv) *As* $|h| \to 0$, *the finite difference solution* (u^*, w^*) *converges to the continuous solution* $(u^*(x_i), w^*(x_i))$ *at every mesh point* $x_i \in \Lambda^*$.

We next consider the case where $f(x,u)$ and $g(x,u)$ are given by (6.4). It is easy to verify that the pair in (6.5) are also ordered upper and lower solutions of (2.8), (6.4). This leads to the choice of $(\gamma_i^{(1)}, \gamma_i^{(2)}) = (4\sigma\overline{a}^3/d_0, 0)$, which ensures that condition (2.9) is satisfied. With this value of $(\gamma_i^{(1)}, \gamma_i^{(2)})$ and the functions

$$f_i(u_i^{(m-1)}) = \sigma[a_i^4 - (u_i^{(m-1)})^4], \quad g_i(u_i^{(m-1)}) = 0 \tag{6.9}$$

in the iteration process (6.2), we compute the maximal and minimal sequences $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ as in the previous problem. In the special case of $D(u) = d_0 + d_1 u$, the equation $u_i^{(m)} = q(w_i^{(m)})$ is given by (6.8). As a consequence of the theorems in the previous sections, we have the following conclusion.

THEOREM 6.2. *Let* $f(x,u)$, $g(x,u)$ *be given by* (6.4) *and* $D(u) \geq d_0 > 0$, *and let* $(\overline{u}_i^{(0)}, \overline{w}_i^{(0)}) = (\overline{a}, \overline{a}^*)$, $(\underline{u}_i, \underline{w}_i) = (0,0)$, *and* $(\gamma_i^{(1)}, \gamma_i^{(2)}) = (4\sigma\overline{a}^3/d_0, 0)$. *Then all the conclusions in* (i)–(iv) *of Theorem* 6.1 *hold true for problem* (2.8), (6.4).

If the boundary condition in (2.8), is replaced by (2.8a), where $0 \leq \xi_i^* \leq \overline{a}^*$ (that is, $0 \leq \xi_i \leq \overline{a}$), then the pair in (6.4) are ordered upper and lower solutions of problem (2.8a), (6.4), and therefore the choice of $(\gamma_i^{(1)}, \gamma_i^{(2)})$ remains the same. The only difference in the iteration process (6.2) is that the boundary condition should be replaced by (6.2a). In vector form, the function $G(U^{(m-1)})$ in (3.9) and (3.10) should be replaced by $\boldsymbol{\xi}^* \equiv (\xi_1^*, \ldots, \xi_M^*)^T$. This observation leads to the following.

THEOREM 6.3. *Let the conditions in Theorem* 6.2 *be satisfied for the problem* (2.8a), (6.4), *where* $0 \leq \xi_i^* \leq \overline{a}^*$, *and let the boundary condition in* (6.2) *be replaced by* (6.2a). *In vector form, let* $G(U^{(m-1)})$ *in* (3.9) *and* (3.10) *be replaced by* $\boldsymbol{\xi}^*$. *Then all the conclusions in* (i)–(iv) *of Theorem* 6.1 *hold true for the Dirichlet problem* (2.8a), (6.4).

A chemical reactor problem. In a nonisothermal chemical reactor with first-order reaction, if the diffusion coefficient is density dependent, then the equation governing the steady-state chemical concentration $u(x)$ is given by (1.1) with

$$f(x,u) = \sigma(1-u)\exp[\gamma u/(1+u)], \quad g(x,u) = -\beta(x)u, \tag{6.10}$$

where σ and γ are positive constants and $\beta(x) \geq 0$ on $\partial\Omega$ (cf. [4, 5, 7, 10, 23, 27, 30]). This implies that the boundary condition of this model problem is given by the Robin type

$$\mathcal{D}(u)\partial u/\partial\nu + \beta(x)u = 0 \quad (x \in \partial\Omega).$$

Problem (1.1), (6.10) also describes the temperature distribution in a combustible material and has been investigated by many workers in combustion theory (see [10] and the references therein). It is obvious that for any $D(u) \geq d_0 > 0$, condition (2.9) in (H_1) is satisfied by some $(\gamma^{(1)}, \gamma^{(2)})$. Unlike the heat-conduction problem, this model may have multiple positive solutions depending on the parameters of σ and γ.

In fact, it is known that in the case of constant diffusion coefficient $D(u) = d_0$ this problem has a unique solution if σ and γ are either small or large, and it has two or more positive solutions if σ and γ have certain intermediate values (see [23, p. 124] and [27] for some numerical results). Since, by (6.10),

$$f(x,0) = \sigma > 0, \quad g(x,0) = 0 \quad \text{and} \quad f(x,1) = 0, \quad g(x,1) \le 0,$$

Definition 2.1 implies that for any constant $\overline{a} \ge 1$ the pair in (6.5) are ordered upper and lower solutions of problem (2.8), (6.10). To find $(\gamma_i^{(1)}, \gamma_i^{(2)})$, we observe from (6.10) that

$$\begin{aligned} f_u(x,u) &= -\sigma(1+u)^{-2}\exp[(\gamma u/(1+u))][u^2 + (2+\gamma)u + (u-\gamma)], \\ g_u(x,u) &= -\beta(x). \end{aligned} \tag{6.11}$$

This implies that $-f_u(x,u) \le 4\sigma e^{\gamma}$ for $0 \le u \le 1$. Hence, by choosing $\overline{a} = 1$, we see that condition (2.9) holds for any $\gamma_i^{(1)} \ge 4\sigma e^{\gamma}/d_0$, $\gamma_i^{(2)} \ge \beta_i/d_0$. Using this value of $(\gamma_i^{(1)}, \gamma_i^{(2)})$, we compute the maximal and minimal sequences from (6.2) with

$$\begin{aligned} f_i(u_i^{(m-1)}) &= \sigma(1 - u_i^{(m-1)})\exp[\gamma u_i^{(m-1)}/(1 + u_i^{(m-1)})], \\ g_i(u_i^{(m-1)}) &= -\beta_i u_i^{(m-1)}, \quad m = 1, 2, \ldots. \end{aligned} \tag{6.12}$$

It is easy to verify that the pair in (6.5) (with $\overline{a} = 1$) are also ordered upper and lower solutions of the continuous problem (2.4), (6.10). This implies that condition (5.5) is trivially satisfied, and, by Theorem 5.1, problem (2.4), (6.10) has a positive maximal solution $(\overline{u}(x), \overline{w}(x))$ and a positive minimal solution $(\underline{u}(x), \underline{w}(x))$. It follows from Theorems 2.1, 3.1, and 5.2 that we have the following conclusion.

THEOREM 6.4. *Let $D(u) \ge d_0 > 0$ and $f(x,u)$, $g(x,u)$ be given by* (6.10)*, and let $(\overline{u}_i^{(0)}, \overline{w}_i^{(0)}) = (\overline{a}, \overline{a}^*)$, $(\underline{u}_i^{(0)}, \underline{w}_i^{(0)}) = (0,0)$, and $(\gamma_i^{(1)}, \gamma_i^{(2)}) = (4\sigma e^{\gamma}/d_0, \beta_i/d_0)$, where $\overline{a} = 1$. Then the following statements hold true:*

(i) *$\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$ converges monotonically from above to a maximal solution $(\overline{u}_i, \overline{w}_i)$ of problem* (2.8)*,* (6.10)*, and $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ converges monotonically from below to a minimal solution $(\underline{u}_i, \underline{w}_i)$.*

(ii) *Every solution (u_i, w_i) of problem* (2.8)*,* (6.10) *in $\mathcal{S}$ satisfies $(\underline{u}_i, \underline{w}_i) \le (u_i, w_i) \le (\overline{u}_i, \overline{w}_i)$, and problem* (2.8)*,* (6.10) *has a unique solution in $\mathcal{S}$ if $(\overline{u}_i, \overline{w}_i) = (\underline{u}_i, \underline{w}_i)$.*

(iii) *In vector form, the sequences $\{\overline{U}^{(m)}, \overline{W}^{(m)}\}$, $\{\underline{U}^{(m)}, \underline{W}^{(m)}\}$ obtained from either* (3.9) *or* (3.10) *converge monotonically to $(\overline{U}, \overline{W})$ and $(\underline{U}, \underline{W})$, respectively.*

(iv) *As $|h| \to 0$, the finite difference solution $(\overline{u}_i, \overline{w}_i)$ converges to $(\overline{u}(x_i), \overline{w}(x_i))$, and $(\underline{u}_i, \underline{w}_i)$ converges to $(\underline{u}(x_i), \underline{w}(x_i))$ at every mesh point $x_i \in \overline{\Lambda}^*$.*

7. Numerical results. To compute numerical solutions of (2.8) or (2.8a) by the monotone iterative schemes, we consider some model problems from heat transfer, where the true continuous solution is explicitly given. This continuous solution is then used to compare with the computed solution to check the accuracy of the monotone iterative scheme. Numerical results for the model where the true solution is not known explicitly is also obtained.

Example 1. In the first example, we consider a heat-conduction problem in a one-dimensional domain $\Omega = (0,1)$, which is given by

$$\begin{aligned} &-[(1+u)u_x]_x + 2(1+u)u_x = a^4(x) - u^4 \quad (0 < x < 1), \\ &u(0) = 0, \quad u(1) = 1 - e^{-1}, \end{aligned} \tag{7.1}$$

TABLE 1
Numerical results of Example 1(a).

(a) Picard's method

Iteration n		$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1.0$
1	$\overline{u}$	0.547609	0.792582	0.893433	0.868080	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.092069	0.163246	0.243168	0.375888	0.632121
2	$\overline{u}$	0.329443	0.555804	0.682750	0.711287	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.137211	0.249549	0.357278	0.481156	0.632121
3	$\overline{u}$	0.246641	0.437489	0.566088	0.630337	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.159969	0.291740	0.408116	0.519978	0.632121
5	$\overline{u}$	0.195209	0.353711	0.477574	0.569074	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.176472	0.321280	0.441844	0.544120	0.632121
8	$\overline{u}$	0.182710	0.332185	0.453958	0.552606	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.180772	0.328812	0.450226	0.549997	0.632121
11	$\overline{u}$	0.181422	0.329944	0.451479	0.550874	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.181221	0.329594	0.451092	0.550603	0.632121
13	$\overline{u}$	0.181306	0.329742	0.451256	0.550718	0.632121
	u^*	0.181269	0.329680	0.451188	0.550671	0.632121
	$\underline{u}$	0.181262	0.329664	0.451170	0.550658	0.632121

where

$$a(x) = \left[6e^{-x} - 4e^{-2x} + (1 - e^{-x})^4\right]^{1/4}. \tag{7.2}$$

The above model is a special case of (1.1a), (6.4) with $D(u) = 1+u$, $c = 2$, $\sigma = 1$, and $\xi(0) = 0$, $\xi(1) = 1 - e^{-1}$. It is easy to check that $u = 1 - e^{-x}$ is the unique solution of (7.1). To compute numerical values of the solution by the monotone iterative scheme (6.2), (6.2a), we observe that

$$w_i = u_i + u_i^2/2, \quad L[w_i] = -\Delta_1[w_i] + 2\delta_i[w_i], \quad q(w_i) = (1 + 2w_i)^{1/2} - 1.$$

Since $a_i^4 \le 7$ and $0 \le \xi_i^* \le 3/2$, it suffices to choose $(\tilde{u}_i, \tilde{w}_i) = (7, 32)$, $(\hat{u}_i, \hat{w}_i) = (0, 0)$, and $(\gamma_i^{(1)}, \gamma_i^{(2)}) = (28, 0)$. With the above data in the iteration process (6.2)–(6.2a), we compute the maximal and minimal sequences $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$ for various values of $h \equiv \Delta x$. The stopping criterion in the iteration process is $||\overline{u}^{(m)} - \underline{u}^{(m)}|| + ||\overline{w}^{(m)} - \underline{w}^{(m)}|| < \epsilon$ for various $\epsilon > 0$, where $||u||$ is the maximal norm of u_i over $i = 1, \ldots, M$. Numerical values of the sequences $\{\overline{u}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}\}$ together with the values of the true analytical solution u_i^* at various mesh points in $(0, 1)$ for the case $h = 1/80, \epsilon = 10^{-4}$ are given in Table 1. It is seen from this table that the monotone property of both the maximal sequence and minimal sequence are observed at every mesh point, and after about 14 iterations the values of $\overline{u}_i^{(m)}$ and $\underline{u}_i^{(m)}$ differ from u_i^* by less than 0.01 percent. We also compute the solution for various values of $a(x)$ where the true solution is not explicitly known. Numerical values of $\{\overline{u}_i^{(m)}\}$ and $\{\underline{u}_i^{(m)}\}$ corresponding to the case $a(x) = 1$ are given in Table 2 and are sketched in Figure 1.

TABLE 2
Numerical results of Example 1(b).

(b) Picard's method

Iteration n		$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1.0$
1	$\overline{u}$	0.530901	0.781567	0.890490	0.869091	0.632121
	$\underline{u}$	0.060073	0.121327	0.208178	0.358766	0.632121
2	$\overline{u}$	0.271157	0.476435	0.605334	0.656579	0.632121
	$\underline{u}$	0.091499	0.185127	0.296503	0.442104	0.632121
3	$\overline{u}$	0.179227	0.338634	0.466401	0.560791	0.632121
	$\underline{u}$	0.106495	0.213778	0.331399	0.468508	0.632121
5	$\overline{u}$	0.128554	0.253374	0.375640	0.499062	0.632121
	$\underline{u}$	0.116027	0.231208	0.351283	0.482456	0.632121
7	$\overline{u}$	0.119933	0.238172	0.358983	0.487717	0.632121
	$\underline{u}$	0.117746	0.234284	0.354695	0.484792	0.632121
9	$\overline{u}$	0.118431	0.235503	0.356041	0.485711	0.632121
	$\underline{u}$	0.118049	0.234822	0.355290	0.485198	0.632121
11	$\overline{u}$	0.118169	0.235036	0.355526	0.485359	0.632121
	$\underline{u}$	0.118102	0.234917	0.355394	0.485269	0.632121
12	$\overline{u}$	0.118136	0.234978	0.355462	0.485316	0.632121
	$\underline{u}$	0.118108	0.234928	0.355407	0.485278	0.632121

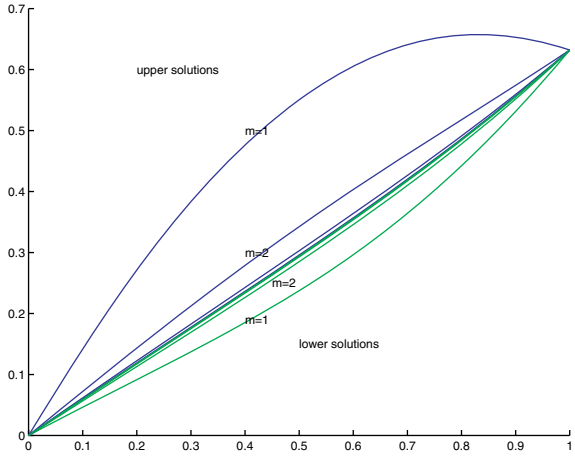


FIG. 1. *Upper and lower sequences for Example* 1(b).

Example 2. We next consider a quasi-linear equation with a nonlinear boundary condition in the form

$$\begin{aligned} &-[(1+u)u_x]_x + 2(1+u)u_x = 0 \quad (0 \le x < 1), \\ &u(0) = 0, \quad (1+u(1))u_x(1) = a^4 - u^4(1), \end{aligned} \tag{7.3}$$

where a is a constant given by

$$a^4 = 2e^2 + [(2e^2-1)^{1/2} - 1]^4. \tag{7.4}$$

This problem is a special case of (1.1a), (6.3) with $D(u) = 1 + u$, $c = 2$, and $\sigma = 1$, and the analytical solution is

$$u(x) = (2e^{2x} - 1)^{1/2} - 1. \tag{7.5}$$

TABLE 3
Numerical results of Example 2(a).

(a) Picard's method

Iteration n		$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1.0$
	$\overline{u}$	0.430854	0.899955	1.437382	2.068933	2.821820
1	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.118280	0.274495	0.477150	0.736053	1.062758
	$\overline{u}$	0.417727	0.875273	1.400922	2.019567	2.757727
2	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.217687	0.484242	0.810366	1.209152	1.696607
	$\overline{u}$	0.412393	0.865217	1.386046	1.999409	2.731543
3	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.294903	0.639031	1.047602	1.537672	2.129240
	$\overline{u}$	0.409151	0.859097	1.376986	1.987128	2.715586
5	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.378317	0.800575	1.290110	1.869162	2.562162
	$\overline{u}$	0.408459	0.857789	1.375050	1.984503	2.712175
8	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.405597	0.852380	1.367037	1.973637	2.698054
	$\overline{u}$	0.408401	0.857680	1.374889	1.984284	2.711891
11	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.408160	0.857223	1.374212	1.983366	2.710698
	$\overline{u}$	0.408397	0.857671	1.374875	1.984265	2.711867
14	u^*	0.408421	0.857709	1.374918	1.984303	2.711888
	$\underline{u}$	0.408376	0.857633	1.374819	1.984189	2.711767

TABLE 4
Numerical results of Example 2(b).

(b) Picard's method

Iteration n		$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1.0$
1	$\overline{u}$	0.070658	0.168152	0.300080	0.475127	0.703153
	$\underline{u}$	0.012119	0.029934	0.055952	0.093618	0.147514
2	$\overline{u}$	0.054097	0.129999	0.234591	0.375926	0.563204
	$\underline{u}$	0.019717	0.048443	0.089892	0.148951	0.231807
3	$\overline{u}$	0.045148	0.109093	0.198161	0.319914	0.483089
	$\underline{u}$	0.024473	0.059934	0.110733	0.182470	0.282051
5	$\overline{u}$	0.036911	0.089655	0.163907	0.266616	0.405965
	$\underline{u}$	0.029268	0.071447	0.131452	0.215478	0.331009
8	$\overline{u}$	0.033264	0.080987	0.148504	0.242428	0.370638
	$\underline{u}$	0.031521	0.076832	0.141091	0.230735	0.353479
11	$\overline{u}$	0.032440	0.079024	0.145004	0.236912	0.362550
	$\underline{u}$	0.032041	0.078073	0.143307	0.234234	0.358620
14	$\overline{u}$	0.032252	0.078576	0.144204	0.235649	0.360697
	$\underline{u}$	0.032161	0.078358	0.143815	0.235036	0.359797
17	$\overline{u}$	0.032209	0.078473	0.144021	0.235360	0.360273
	$\underline{u}$	0.032188	0.078423	0.143932	0.235220	0.360067
19	$\overline{u}$	0.032201	0.078454	0.143986	0.235306	0.360194
	$\underline{u}$	0.032193	0.078435	0.143953	0.235254	0.360117

By using $(\tilde{u}_i, \tilde{w}_i) = (3, 8)$, $(\hat{u}_i, \hat{w}_i) = (0, 0)$, and $(\gamma_i^{(1)}, \gamma_i^{(2)}) = (0, 108)$ in the iteration process (6.2), we compute the maximal and minimal sequences $\{\overline{u}_i^{(m)}, \overline{w}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}, \underline{w}_i^{(m)}\}$, using the same criteria as that in Example 1. Numerical values of $\{\overline{u}_i^{(m)}\}$, $\{\underline{u}_i^{(m)}\}$ and the true solution $u^*(x_i)$ are given in Table 3. Table 4 and Figure 2 give the values of $\{\overline{u}_i^{(m)}\}$ and $\{\underline{u}_i^{(m)}\}$ for the case $a = 1$.

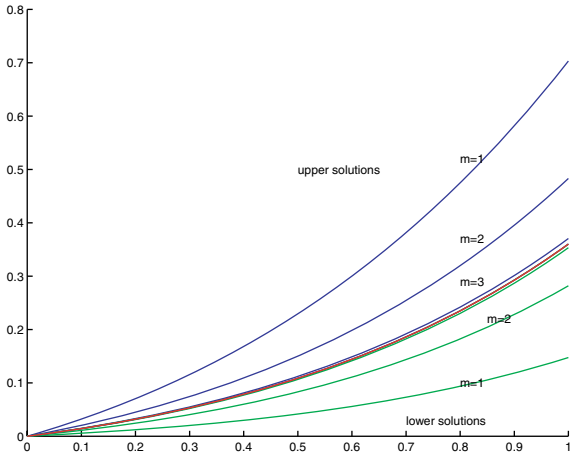


FIG. 2. *Upper and lower sequences for Example* 2(b).

Acknowledgment. The author is thankful to Taiping He for the computation of numerical results in section 7.

REFERENCES

[1] E. ADAMS AND W. F. AMES, *On contracting interval iteration for nonlinear problems in* $\mathbb{R}^n$. II. *Applications*, Nonlinear Anal., 5 (1981), pp. 525–542.

[2] M. AINSWORTH, D. W. KELLY, I. H. SLOAN, AND S. L. WANG, *Post-processing with computable error bounds for the finite element approximation of a nonlinear heat conduction problem*, IMA J. Numer. Anal., 17 (1997), pp. 547–561.

[3] W. F. AMES, *Numerical Methods for Partial Differential Equations*, 3rd ed., Academic Press, San Diego, 1992.

[4] R. ARIS, *The Mathematical Theory of Diffusion and Reaction in Permeable Catalysts*, Vols. I and II, 3rd ed., Oxford University Press, London, 1975.

[5] J. W. BEBERNES AND D. EBERLY, *Mathematical Problems from Combustion Theory*, Springer-Verlag, New York, 1989.

[6] Z. CHEN, R. E. EWING, Q. JIANG, AND A. M. SPAGNUOLO, *Error analysis for characteristics-based methods for degenerate parabolic problems*, SIAM J. Numer. Anal., 40 (2002), pp. 1491–1515.

[7] D. S. COHEN, *Multiple stable solutions of nonlinear boundary value problems arising in chemical reactor theory*, SIAM J. Appl. Math., 20 (1971), pp. 1–13.

[8] J. DOUGLAS, JR., AND T. DUPONT, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comp., 29 (1975), pp. 689–696.

[9] S. EVJE AND K. H. KARLSEN, *Monotone difference approximations of BV solutions to degenerate convection-diffusion equations*, SIAM J. Numer. Anal., 37 (2000), pp. 1838–1860.

[10] D. A. FRANK-KAMENETSKII, *Diffusion and Heat Transfer in Chemical Kinetics*, Plenum Press, New York, 1969.

[11] J. FREHSE AND R. RANNACHER, *Asymptotic* L^∞*-error estimates for linear finite element approximations of quasilinear boundary value problems*, SIAM J. Numer. Anal., 15 (1978), pp. 418–431.

[12] C. A. HALL AND T. A. PORCHING, *Numerical Analysis of Partial Differential Equations*, Prentice–Hall, Englewood Cliffs, NJ, 1990.

[13] H. B. KELLER, *Elliptic boundary value problems suggested by nonlinear diffusion processes*, Arch. Rational Mech. Anal., 35 (1969), pp. 363–381.

[14] J. B. KELLER AND W. E. OLMSTEAD, *Temperature of a nonlinear radiating semi-infinite solid*, Quart. Appl. Math., 29 (1972), pp. 559–566.

[15] D. Y. KWAK AND K. Y. KIM, *Mixed covolume methods for quasi-linear second-order elliptic problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1057–1072.

[16] T. Linss, H.-G. Roos, and R. Vulanović, *Uniform pointwise convergence on Shishkin-type meshes for quasi-linear convection-diffusion problems*, SIAM J. Numer. Anal., 38 (2000), pp. 897–912.
[17] M. Luskin, *A Galerkin method for nonlinear parabolic equations with nonlinear boundary conditions*, SIAM J. Numer. Anal., 16 (1979), pp. 284–299.
[18] S. Meddahi, *On a mixed finite element formulation of a second order quasilinear problem in the plane*, Numer. Methods Partial Differential Equations, 20 (2004), pp. 90–103.
[19] F. A. Milner, *Mixed finite element methods for quasilinear second-order elliptic problems*, Math. Comp., 44 (1985), pp. 303–320.
[20] T. Nakaki and K. Tomoeda, *A finite difference scheme for some nonlinear diffusion equations in an absorbing medium: Support splitting phenomena*, SIAM J. Numer. Anal., 40 (2002), pp. 945–964.
[21] W. E. Olmstead, *Temperature distribution in a convex solid with nonlinear radiation boundary condition*, J. Math. Mech., 15 (1966), pp. 899–908.
[22] M. N. Őzisik, *Boundary Value Problems of Heat Conduction*, Dover, New York, 1989.
[23] C. V. Pao, *Nonlinear Parabolic and Elliptic Equations*, Plenum Press, New York, 1992.
[24] C. V. Pao, *Numerical solutions for some coupled systems of nonlinear boundary value problems*, Numer. Math., 51 (1987), pp. 381–394.
[25] C. V. Pao, *Finite difference reaction diffusion equations with nonlinear boundary conditions*, Numer. Methods Partial Differential Equations, 11 (1995), pp. 355–374.
[26] C. V. Pao, *Numerical methods for semilinear parabolic equations*, SIAM J. Numer. Anal., 24 (1987), pp. 24–35.
[27] C. V. Pao, *Monotone iterative methods for finite difference system of reaction diffusion equations*, Numer. Math., 46 (1985), pp. 571–586.
[28] C. V. Pao, *Quasilinear parabolic and elliptic equations with nonlinear boundary conditions*, Nonlinear Anal., 66 (2007), pp. 639–662.
[29] S. V. Parter, *Mildly nonlinear elliptic partial differential equations and their numerical solution*, Numer. Math., 7 (1965), pp. 113–128.
[30] S. V. Parter, *Solutions of a differential arising in chemical reactor process*, SIAM J. Appl. Math., 26 (1974), pp. 687–716.
[31] L. W. Ross, *Perturbation analysis of diffusion-coupled biochemical reaction kinetics*, SIAM J. Appl. Math., 19 (1970), pp. 323–329.
[32] M.-N. Le Roux, *Numerical solution of nonlinear reaction diffusion processes*, SIAM J. Numer. Anal., 37 (2000), pp. 1644–1656.
[33] R. S. Varga, *Matrix Iterative Analysis*, Prentice–Hall, Englewood Cliffs, NJ, 1962.
[34] J. H. Wang and C. V. Pao, *Finite difference reaction-diffusion equations with nonlinear diffusion coefficients*, Numer. Math., 85 (2000), pp. 485–502.
[35] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs*, SIAM J. Numer. Anal., 33 (1996), pp. 1759–1777.
[36] D. M. Young, *Iterative Solution of Large Linear System*, Academic Press, New York, 1971.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1107–1108

STRICT DIAGONAL DOMINANCE AND OPTIMAL BOUNDS FOR THE SKEEL CONDITION NUMBER*

J.M. PEÑA†

Abstract. An optimal bound for the Skeel condition number of a triangular matrix strictly diagonally dominant by rows is obtained.

Key words. Skeel condition number, diagonal dominance, Gaussian elimination

AMS subject classifications. 65F05, 65G05, 65G99

DOI. 10.1137/060677562

We say that matrix $A = (a_{ij})_{1\le i,j\le n}$ is *strictly diagonally dominant by rows* if $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ for all $i = 1, \dots, n$. The *Skeel condition number* of a matrix A is defined as $\mathrm{Cond}(A) = \| \, |A^{-1}| \, |A| \, \|_\infty$, where $|M|$ denotes the matrix whose entries are the absolute values of the entries of the matrix M. Given a real matrix $A = (a_{ij})_{1\le i,j\le n}$ whose rows are not null, the *minimal scaled diagonal element* of A is the number p defined as: $p := \min_{i=1,\dots,n} |a_{ii}|/(\sum_{j=1}^n |a_{ij}|)$. Observe that, if U is the upper triangular matrix obtained after Gauss elimination of a matrix A, then the minimal scaled diagonal element of U coincides with the concept (introduced in [1]) of the minimal scaled pivot of A. The minimal scaled diagonal element p of matrix which is strictly diagonally dominant by rows satisfies $1/2 < p \le 1$. The following result uses p to obtain an optimal bound for the Skeel condition number of a triangular matrix strictly diagonally dominant by rows. Besides, it provides a proof for Theorem 3.1 of [1] when $p > 1/2$ and shows the optimality of the corresponding bound. The proof of Theorem 3.1 of [1] is correct only when $p < 1/2$ because only in this case $(1-p)/p > 1$ and the proof that $|(V^{-1})_{ij}| < ((1-p)/p)^{(j-i)}$ is correct.

THEOREM 0.1. *Let $U = (u_{ij})_{1\le i,j\le n}$ be an upper triangular matrix which is strictly diagonally dominant by rows, and let $p(> 1/2)$ be its minimal scaled diagonal element. Then*

$$\mathrm{Cond}(U) \le \frac{1 - \left(\frac{1-p}{p}\right)^{n-1}(2-2p)}{2p-1}. \tag{0.1}$$

Moreover, for any $p > 1/2$ and $n \ge 1$ there exist matrices U for which the previous inequality (0.1) *is an equality.*

Proof. Let $V := D^{-1}U$, where D is the diagonal matrix whose (i,i)-entry is u_{ii} for all i. Then $\mathrm{Cond}(U) = \mathrm{Cond}(V)$ and $V = (V_{ij})_{1\le i,j\le n}$ is upper triangular with $V_{ii} = 1$. Let $r := \max_{1\le i\le n}\{\sum_{j=i+1}^n |V_{ij}|\} (< 1)$. Then

$$r = \max_{1\le i\le n}\left\{\sum_{j=i+1}^{n} \frac{|u_{ij}|}{|u_{ii}|}\right\} = \max_{1\le i\le n}\left\{\frac{\sum_{j=i}^n |u_{ij}| - |u_{ii}|}{|u_{ii}|}\right\} = \frac{1}{p} - 1 = \frac{1-p}{p}. \tag{0.2}$$

Let us now prove by induction on n that, if V is an $n \times n$ upper triangular matrix with a unit diagonal which is strictly diagonally dominant by rows with a minimal

*Received by the editors December 12, 2006; accepted for publication (in revised form) January 31, 2007; published electronically May 7, 2007. This research was partially supported by the Spanish Research grant MTM2006-03388 and by Gobierno de Aragón and Fondo Social Europeo.
http://www.siam.org/journals/sinum/45-3/67756.html
†Departamento de Matemática Aplicada, Universidad de Zaragoza, 50009 Zaragoza, Spain (jmpena@unizar.es).

scaled diagonal element not less than p, then

$$\|V^{-1}\|_\infty \leq \frac{1-r^{n-1}(2-2p)}{(2p-1)(r+1)}. \tag{0.3}$$

The result is trivial for $n=1$. Let us assume that it holds for $n-1$, and let us prove it for n.

If we compute V^{-1} by Gauss–Jordan, starting from the last column, we can easily obtain the following bound for the absolute value of $(V^{-1})_{ij}$ for any $i \in \{1,\dots,n\}$ and $j > i$:

$$|(V^{-1})_{ij}| \leq |V_{ij}| + |V_{i,j-1}|\,|(V^{-1})_{j-1,j}| + |V_{i,j-2}|\,|(V^{-1})_{j-2,j}| + \cdots + |V_{i,i+1}|\,|(V^{-1})_{i+1,j}|.$$

Then, taking into account that $(V^{-1})_{ii} = 1$, we can derive

$$\sum_{j=i}^{n} |(V^{-1})_{ij}| \leq 1 + |V_{i,i+1}| \left(\sum_{j=i+1}^{n} |(V^{-1})_{i+1,j}| \right) + \cdots + |V_{i,n}|.$$

Let Z^{-1} be the submatrix of V^{-1} formed by rows and columns $2,\dots,n$. From the previous formula, we can deduce that

$$\sum_{j=i}^{n} |(V^{-1})_{ij}| \leq 1 + \|Z^{-1}\|_\infty \left(\sum_{j=i+1}^{n} |V_{ij}| \right). \tag{0.4}$$

Since Z is an $(n-1)\times(n-1)$ matrix satisfying the induction hypothesis, $\|Z^{-1}\|_\infty \leq (1-r^{n-2}(2-2p))/((2p-1)(r+1))$. Then, using the definition of r and taking the maximum in (0.4) among $i=1,\dots,n$, we can deduce from the previous formula that

$$\|V^{-1}\|_\infty \leq 1 + \|Z^{-1}\|_\infty r \leq \frac{2p(1+r)-1-r^{n-1}(2-2p)}{(2p-1)(r+1)}. \tag{0.5}$$

By (0.2), $2p(1+r)-1=1$, and so we deduce from (0.5) that (0.3) (and so, the induction) holds. Since $\|V\|_\infty \leq 1+r$, the bound (0.1) follows from (0.3) and (0.2).

Now let $0 \leq r < 1$, and let us consider the following matrix U and its inverse U^{-1}:

$$U = \begin{pmatrix} 1 & -r & 0 & \cdots & 0 \\ 0 & 1 & -r & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & 1 & -r \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}, \quad U^{-1} = \begin{pmatrix} 1 & r & r^2 & \cdots & r^{n-1} \\ 0 & 1 & r & \cdots & r^{n-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & 1 & r \\ 0 & \cdots & \cdots & 0 & 1 \end{pmatrix}.$$

Then the minimal scaled diagonal element of U is $p = 1/(1+r)$ and then

$$\mathrm{Cond}(U) = \frac{2r^n - r - 1}{r-1}. \tag{0.6}$$

Substituting $p = 1/(1+r)$ in (0.1), we deduce that $\mathrm{Cond}(U)$ is bounded above by the right-hand side of (0.6), and so this bound can be achieved for any $p > 1/2$ and positive integer $n \geq 1$. ∎

REFERENCE

[1] J. M. Peña, *Scaled pivots and scaled partial pivoting strategies*, SIAM J. Numer. Anal., 41 (2003), pp. 1022–1031.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1109–1121

ON CORNER AVOIDANCE PROPERTIES OF RANDOM-START HALTON SEQUENCES*

JÜRGEN HARTINGER† AND VOLKER ZIEGLER‡

Abstract. Recently, the analysis of quasi-Monte Carlo (QMC) sampling of integrands with singularities has gained considerable interest. In this setting error bounds for QMC integration, in addition to discrepancy, include a measure of how well the singularities are avoided by the utilized sequences. The article aims to generalize results for the corner avoidance of the classical Halton sequence to Halton sequences that start in an arbitrary point of the unit cube. In particular, it is shown that almost all (in Lebesgue sense) random-start Halton sequences exhibit the same corner avoidance property as the original Halton sequence.

Key words. quasi-Monte Carlo integration for singular integrands, random-start Halton sequence, corner avoidance

AMS subject classifications. Primary, 11K45; Secondary, 65C05, 11J25, 11D61

DOI. 10.1137/050645361

1. Introduction.

Quasi-Monte Carlo (QMC) methods are deterministic alternatives to classical Monte Carlo methods with asymptotically superior error bounds for integration problems when the integrand belongs to a suitable class of functions.

Let $\bar{U}^s = [0,1]^s$ be the s-dimensional unit cube, and consider functions $f : \bar{U}^s \to \mathbb{R}$ such that $I = \int_{\bar{U}^s} f(\mathbf{x})d\mathbf{x}$ exists. Furthermore, let $(\mathbf{x}_n)_{n>0}$ be a sequence with $\mathbf{x}_n = (x_n^{(1)}, \ldots, x_n^{(s)}) \in [0,1)^s$. For a subset $B \subseteq \bar{U}^s$, denote by $\lambda(B)$ its Lebesgue measure and by χ_B its characteristic function, i.e., $\chi_B(\mathbf{x})$ equals 1 for $\mathbf{x} \in B$ and 0 otherwise. The star discrepancy of the set $(\mathbf{x}_n)_{1\le n\le N}$, measuring its uniformness, is given by

$$D_N^* (\mathbf{x}_1, \ldots, \mathbf{x}_N) = \sup_{J\in\mathcal{J}^*} \left| \frac{1}{N} \sum_{n=1}^{N} \chi_J (\mathbf{x}_n) - \lambda(J) \right|,$$

where $\mathcal{J}^*$ is the set of all subintervals of $\bar{U}^s$ of the form $\prod_{i=1}^s [0, u_i)$. Sequences with best known order of discrepancy ($\mathcal{O}(N^{-1}(\log N)^s)$) are called low-discrepancy sequences (LDS). Different LDS constructions were proposed by Halton [3], Sobol' [15], Faure [2], and Niederreiter [7, 8]. Tezuka and Tokuyama [17] show that the last three approaches can be unified by a generalization of Niederreiter's principles.

We define the QMC estimator of I as $\hat{I}_N = \frac{1}{N} \sum_{n=1}^{N} f(\mathbf{x}_n)$. Hlawka's theorem [6] bounds the QMC integration error as follows:

$$\left| I - \hat{I}_N \right| \le V_{HK}(f) D_N^* (\mathbf{x}_1, \ldots, \mathbf{x}_N),$$

*Received by the editors November 16, 2005; accepted for publication (in revised form) November 9, 2006; published electronically May 10, 2007.

http://www.siam.org/journals/sinum/45-3/64536.html

†Johann Radon Institute for Computational and Applied Mathematics (RICAM), Austrian Academy of Sciences, Altenbergerstraße 69, A-4040 Linz, Austria (juergen.hartinger@oeaw.ac.at). The work of this author was supported by the Austrian Science Foundation, project S-8308-MAT.

‡Institut für Mathematik, Graz University of Technology, Steyrergasse 30, A-8010 Graz, Austria (ziegler@finanz.math.tugraz.at). The work of this author was supported by the Austrian Science Foundation, project S-8307-MAT.

where $V_{HK}(f)$ denotes the variation in the sense of Hardy and Krause (for a definition and a survey on concepts of multidimensional variation, see [11]). A thorough introduction to the field of QMC integration may be found in the monograph [9].

Mainly inspired by financial applications, the theory of QMC quadrature formulae for singular integrands (those that do not have bounded Hardy–Krause variation) was intensively studied recently. In a setting where the singularities lie in the boundary, Owen [12] shows that the QMC convergence order under suitable growth conditions on the integrand in addition to the discrepancy depends on the speed with which the utilized sequence approaches the boundary of the unit cube.

The quantity of main interest within this setting is the hyperbolic distance from a point $\mathbf{z} = (z_1, \dots, z_s) \in [0,1)^s$ to a corner of the unit cube $\mathbf{h} = (h_1, \dots, h_s) \in \{0,1\}^s$, which is defined by $\|\mathbf{z}\|_{\mathbf{h}} = \prod_{i=1}^{s} |z_i - h_i|$. In order to get (asymptotically) efficient QMC rules for these integrands one has to find point sequences satisfying the condition

$$\exists n_0 \in \mathbb{Z}^+ : \forall n \geq n_0 : \ \|\mathbf{x}_n\|_{\mathbf{h}} \geq c\, n^{-r}, \tag{1}$$

with a constant $c = c(r) > 0$ and small r. A sequence is said to avoid the corner $\mathbf{h}$ if (1) is fulfilled with $r \leq 1 + \varepsilon$ for all $\varepsilon > 0$.

Let us state Owen's result [12] in order to shed some light on the role of r. Let $f : (0,1]^s \to \mathbb{R}$ be a real-valued Lebesgue measurable function which is singular in the origin. Moreover, we assume some growth conditions. For a set $u \subset \{1, \dots, s\}$ of indices, the symbol $\partial^u f(\mathbf{x})$ represents $(\prod_{j \in u} \partial/\partial x_j) f(\mathbf{x})$, with the convention that $\partial^{\emptyset} f(\mathbf{x}) = f(\mathbf{x})$. Our growth condition is that

$$|\partial^u f(\mathbf{x})| \leq B \prod_{j=1}^{s} (x_j)^{-A_j - \mathbf{1}_{j \in u}} \tag{2}$$

holds for some $A_j > 0$, some $B < \infty$, and all $u \subseteq \{1, \dots, s\}$, where $\mathbf{1}_{j \in u} = 1$ if $j \in u$ and $\mathbf{1}_{j \in u} = 0$ otherwise. Now we can state Owen's result [12, Theorem 5.5].

THEOREM 1. *Let $f(\mathbf{x})$ be as above and suppose $\mathbf{x}_1, \dots, \mathbf{x}_N$ satisfy (1). Then for any $\eta > 0$,*

$$|\hat{I}_N - I| \leq C_1 D_N^*(\mathbf{x}_1, \dots \mathbf{x}_N) N^{\eta + r \max_j A_j} + C_2 N^{r(\max_j A_j - 1)}$$

holds for finite C_1 and C_2 that may depend on η.

A similar result also holds for other corners than $\mathbf{0}$.

As Owen [12] remarks, the QMC integration is superior to Monte Carlo integration, if $\max_j A_j < 1/(2r)$, provided $D_N^*(\mathbf{x}_1, \dots \mathbf{x}_N) = O(N^{-1+\epsilon})$. In view of applications of QMC integration of singular functions, we have to use LDS which avoid corners with small r.

In the case $\mathbf{h} = 0$, the lower bound $r \geq 1$ is obvious for (t,s)-sequences (for a definition, see, e.g., [9]) and for Halton sequences. Sobol' [16] shows that $r = 1$ holds for Sobol' sequences, and Owen [12] establishes $r = 1$ for Halton sequences. Hartinger, Kainhofer, and Ziegler [4] show $r = 1$ for generalized Niederreiter sequences. For all $\mathbf{h}$, i.e., $\min_{\mathbf{h} \in \{0,1\}^s} \|x_N\|_{\mathbf{h}}$, much less is known: Hartinger, Kainhofer, and Ziegler [4] establish that $r \leq 1 + \varepsilon$ for Halton sequences and $r \geq 3/2$ for the Faure sequence. In a randomized setting Owen [10] shows that the expected error of randomized QMC (under mild conditions on the moments of the integrand) is superior to Monte Carlo even if there exist point singularities with unknown locations. In this paper we show that $r \leq 1 + \epsilon$ holds also for almost all random-start Halton sequences.

Section 2 reviews the definition of random-start Halton sequences, that might be considered as Halton sequences started at an arbitrary point in the unit cube (the classical Halton sequence starts in $\mathbf{0}$). In view of an application of Schmidt's subspace theorem we formulate and discuss this theorem in section 3. The subspace theorem will be an essential part of the proof of the results in section 4, which establishes corner avoidance results for Halton sequences started in special points of the unit cube. In section 5 we show that the set of start points inducing Halton sequences that do not avoid all corners has Lebesgue measure zero.

2. Random-start Halton sequences. For a base $p \in \mathbb{Z}^+$ ($\mathbb{Z}^+$ denotes the set of positive integers) and an integer n with p-adic expansion $n = \sum_{r=1}^{l} a_r(n)p^{r-1}$, define the radical inverse function by $\Phi_p(n) = \sum_{r=1}^{l} a_r(n)p^{-r}$. The nth element of the s-dimensional Halton sequence [3] in relatively coprime bases $p_1, \ldots, p_s$ (typically the first s primes) is given by $\mathbf{x}_n = (\Phi_{p_1}(n), \ldots, \Phi_{p_s}(n))$.

Wang and Hickernell [18] propose to generalize the Halton sequence in the following way: Let $z \in [0, 1)$ be written as

$$z = \sum_{r=1}^{\infty} \frac{a_r(z)}{p^r}, \tag{3}$$

where for all $r_0 \in \mathbb{Z}^+$ there exists an $r \geq r_0$ such that $a_r(z) < p-1$ and $1 = \sum_{r=1}^{\infty} \frac{p-1}{p^r}$. The p-adic von Neumann–Kakutani transform is given by rightward-carry addition in base p, i.e., $T_p(z) = z \oplus \frac{1}{p} = \frac{1+a_m(z)}{p^m} + \sum_{r>m} \frac{a_r(z)}{p^r}$, where $m = \min\{r \in \mathbb{Z}^+ \mid u_r \neq p-1\}$. For relatively coprime bases $p_1, \ldots, p_s$, the s-dimensional $\mathbf{z}$-start Halton sequence is recursively defined by $\mathbf{z} = \mathbf{x}_0 = (x_0^{(1)}, \ldots, x_0^{(s)}) \in [0, 1]^s$ and $\mathbf{x}_{n+1} = (T_{p_1}(x_n^{(1)}), \ldots, T_{p_s}(x_n^{(s)}))$ for $n \in \mathbb{Z}_0^+$. Wang and Hickernell [18] show that for all $\mathbf{z} \in [0, 1]^s$ the induced Halton sequence is an LDS and propose a randomization of the Halton sequence by choosing $\mathbf{z}$ through a realization of a uniformly distributed random variable on $[0, 1]^s$ (therefore this construction is known as a random-start Halton sequence).

For later convenience, we give an equivalent p-adic definition of the $\mathbf{z}$-start Halton sequence: For a prime p, let $\mathbb{Z}_p$ denote the set of all p-adic integers. Supposing $z = \sum_{r=1}^{\infty} a_r(z)p^{r-1} \in \mathbb{Z}_p$, we define the extended radical inverse function $\Phi_p : \mathbb{Z}_p \to [0, 1]$, with $z \mapsto \sum_{r=1}^{\infty} a_r(z)p^{-r}$. Since $a_r(z) \in \{0, 1, \ldots, p-1\}$, the sum converges and Φ_p is a well-defined function. The z-start van der Corput sequence $(\phi_p(n; z))_{n\geq 1}$ for a starting point $z \in \mathbb{Z}_p$ is defined by $\phi_p(n; z) = \Phi_p(n + z)$. Finally, for distinct primes $p_1, \ldots, p_s$ and fixed p_i-adic integers z_i ($1 \leq i \leq s$), the $\mathbf{z}$-start Halton sequence $(\mathbf{x}_n)_{n\geq 1}$ is given by

$$\mathbf{x}_n = (\phi_{p_1}(n; z_1), \ldots, \phi_{p_s}(n; z_s)).$$

For $k \in \mathbb{Z}^+$, the sequence $(\phi_p(n; k))_{n\geq 1} = (\Phi_p(n+k))_{n\geq 1}$ is a shift of the van der Corput sequence by k elements. Therefore a $\mathbf{z}$-start Halton sequence with $\mathbf{z} \in \mathbb{Z}^s$ is called a finitely shifted Halton sequence. If there exists a coordinate in $\mathbf{z}$ such that $z_i \in \mathbb{Z}_p \setminus \mathbb{Z}$, the corresponding Halton sequence is said to be infinitely shifted.

Let us define the map $\tilde{\Phi}_p(z) : [0, 1] \to \mathbb{Z}_p$ with $z \mapsto \sum_{r=1}^{\infty} a_r(z)p^{r-1}$, where the coefficients are obtained by (3). Thus, we have $\Phi_p \circ \tilde{\Phi}_p = \mathrm{id}$. For $\mathbf{z} = (z_1, \ldots, z_s) \in [0, 1)^s$ and $n \in \mathbb{Z}^+$, the corresponding $\mathbf{z}$-start Halton sequence may now be written in the following form:

$$\mathbf{x}_n := \mathbf{x}_n^{(z_1,\ldots,z_s)} = \left(\phi_{p_1}(n; \tilde{\Phi}(z_1)), \ldots, \phi_{p_s}(n; \tilde{\Phi}(z_s))\right).$$

Remark. Note that we have defined the Halton sequence only for prime basis. A similar construction may be obtained for relative prime basis. Observe that for a prime p, $\mathbb{Z}_p$ may be defined by $\mathbb{Z}_p = \varprojlim \mathbb{Z}/p^k\mathbb{Z}$, the inverse limit of the rings $\mathbb{Z}/p^k\mathbb{Z}$. This inverse limit exists even if p is not a prime. Considered as a limit of rings, it does not give a suitable structure, i.e., in general $\mathbb{Z}_p$ is not a domain. Nevertheless, viewed as a limit of additive groups one gets a similar structure as in the prime case. Since in the definition above only the additive structure is needed, the definition is suitable to any relatively coprime basis.

3. The subspace theorem. Akin to the classical Halton sequence (cf. [4]) the backbone in the proof for finitely shifted Halton sequences (see section 4) is the p-adic subspace theorem due to Schlickewei [13] (see also [14, Chapter V, Theorem 1D]). In the literature there exist many versions of the subspace theorem. The most general form was established by Evertse and Schlickewei [1]. Because of technical reasons we only state a nonquantitative version of this theorem, which can be found in Schmidt's book [14, Chapter V]. Before we state the subspace theorem, we introduce some notations.

Let $\mathbb{Q}$ be the field of rationals; then there exist several absolute values on $\mathbb{Q}$. One of them is the so-called Archimedian absolute value $|\cdot|_\infty$, which is the usual absolute value. For any prime p we obtain an absolute value $|\cdot|_p$, with $|x|_p = p^{-\alpha}$, where $x = p^\alpha u/v$ with $p \nmid uv$; these absolute values are called non-Archimedian. The Archimedian and non-Archimedian absolute values form the set $M(\mathbb{Q})$, the canonical absolute value. Let K be some number field; then every absolute value can be extended to an absolute value on K, possibly in several ways. Absolute values obtained in this way are called Archimedian again, if they are extensions of the usual absolute value and are called non-Archimedian if they are induced by an absolute value of the form $|\cdot|_p$ with p a prime. The union of these absolute values is denoted by $M(K)$ and is called the set of canonical absolute values. Let $p \in \{\infty, 2, 3, 5, \dots\} = M(\mathbb{Q})$ (we identify p with $|\cdot|_p$) and let $\nu \in M(K)$; then we write $\nu|p$ if ν is induced by p. Again let $\nu \in M(K)$ with $\nu|p$. Then we denote by K_ν the (topological) closure of K with respect to ν. The index $n_\nu := [K_\nu : \mathbb{Q}_p]$ is called the local degree. Note that if $p = \infty$, then $n_\nu = 1, 2$ depending on whether the K_ν is real or complex. We also use the following notation. Let $\mathbf{x} \in K^s$; then we use the notation

$$\overline{|\mathbf{x}|} = \max_{\substack{1 \le i \le s \\ \nu \in M(K),\, \nu|\infty}} |x_i|_\nu.$$

Note that by linear forms we mean homogeneous polynomials of degree 1 over some number field K. We say that linear forms are linearly independent if they are linearly independent over their (fixed) field of coefficients. Now we can state a nonquantitative form of the subspace theorem (see [14, Chapter V, Theorem 1D]—note the misprint non-Archimedian instead of Archimedian).

THEOREM 2. *Let K be an algebraic number field, and let $S \subset M(K)$ be a finite set of absolute values containing all of the Archimedian ones. For each $\nu \in S$ let $L_{\nu,1}, \dots, L_{\nu,s}$ be s linearly independent (over K) linear forms in s variables with coefficients in K. Then for a given $\delta > 0$, the solutions of the inequality*

$$\prod_{\nu \in S} \prod_{i=1}^{s} |L_{\nu,i}(\mathbf{x})|_\nu^{n_\nu} < \overline{|\mathbf{x}|}^{-\delta}$$

with $\mathbf{x} \in \mathfrak{o}_K^s$ and $\mathbf{x} \neq \mathbf{0}$, where $\mathfrak{o}_K$ is the ring of integers of K, lie in finitely many proper subspaces of K^s.

In view of the next section we only need Theorem 2 in the case of $K = \mathbb{Q}$. Therefore we state also following version of the subspace theorem.

THEOREM 3. *Let $\{\infty\} \subset S \subset M(\mathbb{Q})$ be a finite set of absolute values. For each $\nu \in S$ let $L_{\nu,1}, \ldots, L_{\nu,s}$ be s rational, linearly independent linear forms in s variables. Then for a given $\delta > 0$, the solutions of the inequality*

$$\prod_{\nu \in S} \prod_{i=1}^{s} |L_{\nu,i}(\mathbf{x})|_\nu < \left(\max_{1 \le i \le s} |x_i| \right)^{-\delta}$$

with $\mathbf{x} \in \mathbb{Z}^s$ and $\mathbf{x} \neq \mathbf{0}$ lie in finitely many proper subspaces of $\mathbb{Q}^s$.

4. Finite shifts. In this section we show that almost all finite shifts preserve the corner avoidance properties of the Halton sequence in basis $p_1, \ldots, p_s$. The first step is to establish a relation between the hyperbolic distance and exact powers of basis elements dividing certain p-adic integers. For a p-adic integer z, we say α is the exact power of p dividing z (in symbols $p^\alpha \| z$), if $\alpha = \min\{r \ge 0 : a_{r+1}(z) \neq 0\}$.

LEMMA 1. *Let $\mathbf{h}$ be a corner of the unit cube and let $\mathbf{z} \in [0,1]^s$ such that $p_i^{\alpha_i} \| (\tilde{\Phi}_{p_i}(z_i) + h_i)$ for $i = 1, \ldots, s$. Then*

$$\prod_{i=1}^{s} p_i^{-\alpha_i - 1} \le \|\mathbf{z}\|_{\mathbf{h}} \le \prod_{i=1}^{s} p_i^{-\alpha_i}.$$

Proof. Obviously, it suffices to prove

$$p_i^{-\alpha_i - 1} \le |z_i - h_i| \le p_i^{-\alpha_i},$$

for $i = 1, \ldots, s$.

First, consider the case $h_i = 0$. For notational convenience omit the index i. Due to $p^\alpha \| \tilde{\Phi}(z)$, we have $\tilde{\Phi}(z) = \sum_{r=\alpha+1}^{\infty} a_r(z) p^{r-1}$ $(a_{\alpha+1}(z) \neq 0)$. Since Φ is the left inverse of $\tilde{\Phi}$, we obtain

$$z = \Phi \circ \tilde{\Phi}(z) = \sum_{r=\alpha+1}^{\infty} \frac{a_r(z)}{p^r}, \qquad a_{\alpha+1}(z) \neq 0.$$

Now, $p^{-\alpha-1} \le z \le p^{-\alpha}$ follows immediately by $0 \le a_r(z) \le p - 1$ for $r \ge 1$.

Consider the case $h_i = 1$. Omitting indices again, we have $p^\alpha \| (\tilde{\Phi}(z) + 1)$. Hence, $\tilde{\Phi}(z) + 1 = \sum_{r=\alpha+1}^{\infty} a_r(z) p^{r-1}$ $(a_{\alpha+1}(z) \neq 0)$, and

$$\tilde{\Phi}(z) = \sum_{r=1}^{\alpha} (p-1) p^{r-1} + (a_{\alpha+1}(z) - 1) p^\alpha + \sum_{r=\alpha+2}^{\infty} a_r(z) p^{r-1}.$$

Again, an application of Φ to $\tilde{\Phi}(z)$ yields

$$z = \Phi \circ \tilde{\Phi}(z) = \sum_{r=1}^{\alpha} \frac{p-1}{p^r} + \frac{a_{\alpha+1}(z) - 1}{p^{\alpha+1}} + \sum_{r=\alpha+2}^{\infty} \frac{a_r(z)}{p^r}$$

and furthermore

$$\begin{aligned} 1 - z &= \sum_{r=1}^{\infty} \frac{p-1}{p^r} - \sum_{r=1}^{\alpha} \frac{p-1}{p^r} - \frac{a_{\alpha+1}(z) - 1}{p^{\alpha+1}} - \sum_{r=\alpha+2}^{\infty} \frac{a_r(z)}{p^r} \\ &= \frac{p - a_{\alpha+1}(z)}{p^{\alpha+1}} + \sum_{r=\alpha+2}^{\infty} \frac{p - 1 - a_r(z)}{p^r}. \end{aligned}$$

Thus, $p^{-\alpha-1} \leq 1 - z \leq p^{-\alpha}$. □

THEOREM 4. *For* $\mathbf{k} = (k_1, \ldots, k_s) \in \mathbb{Z}^s$, *let*

$$\mathbf{x}_n = (x_n^{(1)}, \ldots, x_n^{(s)}) = (\phi_{p_1}(n; k_1), \ldots, \phi_{p_s}(n; k_s))$$

be the $\mathbf{k}$*-start Halton sequence in relatively coprime basis* $p_1, \ldots, p_s$. *Denote by* $K(L)$ *the set of all shift-vectors* $\mathbf{k}$ *with* $\|\mathbf{k}\|_\infty < L$ *such that the* $\mathbf{k}$*-start Halton sequence does not avoid all corners, i.e.,*

$$K(L) = \{\mathbf{k} \in ([0, L] \cap \mathbb{Z})^s \ : \ \exists \varepsilon > 0 \ \exists h \in \{0,1\}^s \ \forall n_0 \in \mathbb{Z}^+ \ \exists n \geq n_0 \\ : \ \|\mathbf{x}_n\|_{\mathbf{h}} < n^{-1-\epsilon}\}.$$

Then

$$\lim_{L \to \infty} \frac{\#K(L)}{L^s} = 0.$$

Proof. Let $1 \leq i \leq s$. Let us first assume that all quantities $k_i + h_i$ are the same, say equal to β. By the virtue of Lemma 1, it suffices to show that for every $\epsilon > 0$ there are only finitely many $n \in \mathbb{Z}^+$, such that

$$p_i^{\alpha_i} \| (n + k_i + h_i) = n + \beta \qquad \text{and} \qquad \prod_{i=1}^{s} p_i^{\alpha_i} \geq n^{1+\epsilon}. \tag{4}$$

Since our assumption we have $\prod_{i=1}^s p_i^{\alpha_i} | (n + \beta)$, and therefore (4) can hold only for finitely many n.

Now let us assume not all $k_i + h_i$ are the same. Let us write $x_i = n + k_i + h_i = C_i p_i^{\alpha_i}$. Note that C_i and p_i are relative prime. We may assume that the quantities $k_i + h_i$ are pairwise distinct. (If $k_i + h_i = k_j + h_j$ for two indices $i \neq j$, set $x_i = x_j = \tilde{C}_i p_i^{\alpha_i} p_j^{\alpha_j}$.) Fix some index j and assume that $\frac{C_1 \cdots C_s}{\prod_{i \neq j} x_i} \geq n^{-\epsilon/2}$. For n large, such that $n^{1+\epsilon/2} \geq n + k_j + h_j = x_j$, we obtain

$$n^{1+\epsilon} \geq x_j n^{\epsilon/2} \geq x_j \frac{\prod_{i \neq j} x_i}{\prod_{i=1}^s C_i} = \prod_{i=1}^{s} p_i^{\alpha_i}.$$

Now, assume that $\frac{C_1 \cdots C_s}{\prod_{i \neq j} x_i} \leq n^{-\epsilon/2}$. We now come to the key point of the proof, the application of the subspace theorem (Theorem 3). Let S be the set of absolute values corresponding to a prime that divides a basis together with the usual Archimedian absolute value. For every non-Archimedian absolute values ν we choose $L_{\nu,i}(x_1, \ldots, x_s) = x_i$. The linear forms corresponding to the Archimedian absolute value are $L_{\infty,j}(x_1, \ldots, x_s) = x_j$ and $L_{\infty,i}(x_1, \ldots, x_s) = x_j - x_i$ for $i \neq j$. For n large, such that $n^2 \geq \max\{n + k_i + h_i, (k_i + h_i)^{8s/\varepsilon}\}$ for all i, we obtain the following inequality:

$$\begin{aligned} \overline{|\mathbf{x}|}^{-\epsilon/8} &= \left(\max_{i=1,\ldots,s} |n + k_i + h_i| \right)^{-\epsilon/8} \\ &\geq n^{-\epsilon/4} \geq n^{-\epsilon/2} \max_{1 \leq i \leq s} (k_i + h_i)^s \\ &\geq \frac{C_1 \cdots C_s}{\prod_{i \neq j} x_i} \prod_{i \neq j} |k_j + h_j - k_i - h_i| \end{aligned}$$

$$= \frac{x_j}{p_j^{\alpha_j}} \prod_{i \neq j} \frac{|k_j + h_j - k_i - h_i|}{p_i^{\alpha_i}}$$

$$\geq \prod_{\nu \in S} \prod_{i=1}^{s} |L_{\nu,i}(\mathbf{x})|_\nu.$$

The last inequality holds since $|L_{\infty,i}(\mathbf{x})| = |x_j - x_i| = |k_j + h_j - k_i - h_i|$ in case of $j \neq i$ and $|L_{\infty,j}(\mathbf{x})| = |x_j|$. Furthermore, $|L_{p,i}(\mathbf{x})|_p = |C_i p_i^{\alpha_i}|_p = p^{-\alpha_i}$ if $p|p_i$, and $|L_{p,i}(\mathbf{x})|_p \leq 1$ otherwise.

By Theorem 2 the solutions of this inequality lie only in finitely many subspaces. Therefore the vector $\mathbf{x}$ satisfies the linear system (let T be an adequate subspace)

$$\begin{aligned} x_1 t_1 + \cdots + x_s t_s &= 0 \qquad && (\text{subspace } T), \\ x_j - x_i &= k_j + h_j - k_i - h_i \qquad && (i \neq j). \end{aligned} \tag{5}$$

System (5) may be written in the following form:

$$\begin{pmatrix} t_1 & t_2 & \cdots & t_{j-1} & t_j & t_{j+1} & \cdots & t_s \\ -1 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ 0 & -1 & \cdots & 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -1 & 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 & 1 & -1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_j \\ x_{j+1} \\ \vdots \\ x_s \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ k_j + h_j - k_1 - h_1 \\ k_j + h_j - k_2 - h_2 \\ \vdots \\ k_j + h_j - k_{j-1} - h_{j-1} \\ k_j + h_j - k_{j+1} - h_{j+1} \\ \vdots \\ k_j + h_j - k_s - h_s \end{pmatrix}.$$

Note that the $\mathbf{k}$-start Halton sequence avoids the corner $\mathbf{h}$ if and only if system (5) has infinitely many solutions. But the system has infinitely many solutions if and only if

$$\sum_{i=1}^{s} t_i = 0 \quad \text{and} \quad \sum_{i \neq j} t_i (k_i + h_i) = -t_j (k_j + h_j). \tag{6}$$

But (6) yields at most L^{s-1} possibilities for $(k_1 + h_1, \ldots, k_s + h_s)$ such that $\max_{1 \leq i \leq s}(k_i + h_i) \leq L$. Moreover, the subspace theorem yields only a finite number of t subspaces. Thus, we have at most tL^{s-1} different vectors $\mathbf{k}$ that do not admit corner avoidance. Hence, $K(L) \leq tL^{s-1}$ and $\lim_{L \to \infty} \frac{\#K(L)}{L^s} = 0$. □

5. Metric results. This section aims to provide metric results on corner properties of random-start Halton sequences. It is shown that (Lebesgue) almost all $\mathbf{z}$-start Halton sequences avoid corners. In the spirit of Lemma 1 we use the following notion to characterize points that are near the corners.

DEFINITION 1. *Let $f : \mathbb{Z}^+ \to \mathbb{R}^+$ be a fixed arithmetic function such that $f(n) \geq 1$ and*

$$\sum_{n=1}^{\infty} \frac{(\log(nf(n)))^{s-1}}{nf(n)} < \infty.$$

For $\mathbf{z} = (z_1, \ldots, z_s) \in \mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}$ and $n \geq 1$, let $\mathbf{x}_n = (\phi_{p_1}(n; z_1), \ldots, \phi_{p_s}(n; z_s))$ be the $\mathbf{z}$-start Halton sequence in relatively coprime basis $p_1, \ldots, p_s$. The point $\mathbf{x}_n$ is said to be close to the corner $\mathbf{h}$ if there exist $\alpha_1, \ldots, \alpha_s$ such that for $i = 1, \ldots, s$,

$$p_i^{\alpha_i} \| (z_i + h_i + n) \quad \text{and} \quad \prod_{i=1}^{s} p_i^{\alpha_i} > nf(n). \tag{7}$$

If (7) is fulfilled for infinitely many n, the $\mathbf{z}$-start Halton sequence $(\mathbf{x}_n)_{n\geq 1}$ is said to approach $\mathbf{h}$.

For the rest of this section we fix an arithmetic function f, which is assumed to fulfill the properties of Definition 1.

THEOREM 5. *For $\mathbf{z} = (z_1, \ldots, z_s) \in [0, 1)^s$ and $n \geq 1$, let*

$$(\mathbf{x}_n)_{n\geq 1} = (\mathbf{x}_n^{\mathbf{z}})_{n\geq 1} = \left(\phi_{p_1}(n; \tilde{\Phi}(z_1)), \ldots, \phi_{p_s}(n; \tilde{\Phi}(z_s))\right)_{n\geq 1}$$

be the $\mathbf{z}$-start Halton sequence in relatively coprime basis $p_1, \ldots, p_s$. Then the set

$$\mathcal{A} = \{\mathbf{z} \in [0, 1)^s \,:\, (\mathbf{x}_n^{\mathbf{z}})_{n\geq 1} \text{ approaches a corner } \mathbf{h}\}$$

has Lebesgue measure zero, i.e., $\lambda(\mathcal{A}) = 0$.

Proof. The proof will be done in several steps.

- *A tiling for $\mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}$.*
  For $l_1, \ldots, l_s \in \mathbb{Z}^+$, let us define the fundamental tile by

  $$X_{l_1,\ldots,l_s} := \{\mathbf{y} \in \mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s} \,:\, p^{\alpha_i} | y_i,\ i = 1, \ldots, s\}.$$

  A tiling of $\mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}$ is obtained by the tiles

  $$X_{l_1,\ldots,l_s} + (k_1, \ldots, k_s) := \{\mathbf{y} \in \mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s} \,:\, p^{\alpha_i} | (y_i - k_i),\ i = 1, \ldots, s\},$$

  for all integers k_i with at most α_i p_i-adic digits, i.e., $k_i < p_i^{l_i} \in \mathbb{Z}^+$ $(i = 1, \ldots, s)$.
  Since Φ_p is surjective we also have a tiling of $[0, 1]^s$. In most cases we will need a tiling induced by the fundamental tile $X_l := X_{l,\ldots,l}$, for a fixed $l \in \mathbb{Z}^+$.
  Remark. The tiles $X_{l_1,\ldots,l_s} + (k_1, \ldots, k_s)$ generate the σ-algebra of the Borel sets of the topological group $\mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}$. The Haar measure μ such that $\mu(\mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}) = 1$ is obtained if we allocate to each tile $X_{l_1,\ldots,l_s} + (k_1, \ldots, k_s)$ the measure $p_1^{-l_1} \cdots p_s^{-l_s}$ (see [5, pp. 202–203]). Moreover, the functions Φ and $\tilde{\Phi}$ are measurable and the Haar measure μ is induced by the Lebesgue measure λ and vice versa, i.e., $\mu(A) = \lambda(\tilde{\Phi}^{-1}(A))$ and $\lambda(B) = \mu(\Phi^{-1}(B))$ for all measurable (Borel) sets $A \subseteq \mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}$ and $B \subseteq [0, 1)^s$.

To prove that Φ is measurable we only need to show that every element of a generating system of the Borel sets of $[0,1)^s$ has as preimage a Borel set of $\mathbb{Z}_{p_1} \times \cdots \times \mathbb{Z}_{p_s}$. In fact, we have

$$\Phi^{-1}\left(\left[\frac{k_1}{p_1^{l_1}}, \frac{k_1+1}{p_1^{l_1}}\right) \times \cdots \times \left[\frac{k_s}{p_s^{l_s}}, \frac{k_s+1}{p_s^{l_s}}\right)\right) = X^-_{l_1,\ldots,l_s} + (k_1,\ldots,k_s),$$

where $X^-_{l_1,\ldots,l_s}$ is $X_{l_1,\ldots,l_s}$ without elements which have a component $y_i \in \mathbb{Z}_{p_i}$ of the form $y_i = \sum_{r=l_i+1}^{\infty}(p_i-1)p_i^r$. Loosely speaking, $X^-_{l_1,\ldots,l_s}$ is $X_{l_1,\ldots,l_s}$ without right upper border. It is easily seen that this right upper border is a Borel set, which shows that Φ is measurable, since $X_{l_1,\ldots,l_s}$ is also a Borel set.

Similarly we have

$$\tilde{\Phi}^{-1}\left(X_{l_1,\ldots,l_s} + (k_1,\ldots,k_s)\right) = \left[\frac{k_1}{p_1^{l_1}}, \frac{k_1+1}{p_1^{l_1}}\right] \times \cdots \times \left[\frac{k_s}{p_s^{l_s}}, \frac{k_s+1}{p_s^{l_s}}\right].$$

Since the right side is a Borel set, $\tilde{\Phi}$ is measurable.

- *Let $N \in \mathbb{Z}^+$ be fixed and $l \in \mathbb{Z}^+$ be large. Consider the tiling induced by X_l; if $\mathbf{x}_N^{\mathbf{z}}$ is close to the corner $\mathbf{0}$, then the whole tile corresponding to $\mathbf{z}$ is close to the corner $\mathbf{0}$.*

  Define $\Phi(\mathbf{z}) := (\Phi_{p_1}(z_1), \ldots \Phi_{p_s}(z_s))$. First, observe that for $i = 1,\ldots,s$ and $a_i \in \mathbb{Z}_{p_i}$, we have

$$\left(p_i^{\alpha_i} | (\tilde{\Phi}_{p_i}(z_i) + N)\right) \quad \Longrightarrow \quad \left(p_i^{\alpha_i} | (\tilde{\Phi}_{p_i}(z_i) + a_i p_i^{\alpha_i} + N)\right). \tag{8}$$

  For $i = 1,\ldots,s$, let $\bar{k}_i = \sum_{r=0}^{\alpha_i} a_{r+1}(z_i)p_i^r$. From (8), we conclude that if $\mathbf{x}_N^{\mathbf{z}}$ is close to $\mathbf{0}$, then $\mathbf{x}_N^{\Phi(\bar{\mathbf{k}})}$ is close to $\mathbf{0}$. Thus, all points of the form $\mathbf{x}_N^{\Phi(\bar{\mathbf{k}}+\mathbf{y})}$ with $y \in X_{\alpha_1,\ldots,\alpha_s}$ are close to $\mathbf{0}$.

  The same holds for all points of the form $\mathbf{x}_N^{\Phi(\bar{\mathbf{k}}+\mathbf{y})}$ with $y \in X_{\beta_1,\ldots,\beta_s}$, where $\beta_i \geq \alpha_i$ $(i = 1,\ldots,s)$, in particular for $y \in X_l$ with $l \geq \max\{\alpha_1,\ldots,\alpha_s\}$.

- *Reduction of the problem to a counting problem.*

  Let us define the set

$$\mathcal{A}_N := \{\mathbf{z} \in [0,1)^s \ : \ \mathbf{x}_N^{\mathbf{z}} \text{ is close to } \mathbf{0}\}.$$

  The idea is to find a Borel measureable set $\mathcal{A}'_N \supset \mathcal{A}_N$, such that we are able to estimate the measure of $\mathcal{A}'_N$ by solving a counting problem. We denote the set of s-tuples of integers that possess p_i-adic expansions with at most l digits by $\mathcal{F}_{(l)}$, i.e.,

$$\mathcal{F}_{(l)} = \mathbb{Z}^s \cap \left([0,p_1^l) \times \cdots \times [0,p_s^l)\right),$$

  and consider the set

$$\mathcal{C}_{N,l} = \{(k_1,\ldots,k_s) \in \mathcal{F}_{(l)} \ : \ \mathbf{x}_N^{\Phi(k_1,\ldots,k_s)} \text{ is close to } \mathbf{0}\}.$$

  Let $l > \frac{\log(Nf(N))}{\log 2}$; thus $l \geq \max\{\alpha_1,\ldots,\alpha_s\}$. We know that with $\mathbf{k} \in \mathcal{C}_{N,l}$ all $\mathbf{x}_N^{\Phi(\mathbf{k}+y)}$ are close to $\mathbf{0}$ provided $y \in X_l$. Therefore we define

$$\mathcal{A}'_N = \bigcap_{l=\lfloor \frac{\log(Nf(N))}{\log 2} \rfloor + 1}^{\infty} \mathcal{A}'_{N,l} \quad \text{and} \quad \mathcal{A}'_{N,l} = \Phi\left(\bigcup_{\mathbf{k} \in \mathcal{C}_{N,l}} \mathbf{k} + X_l\right).$$

Since the sets $\Phi(\mathbf{k}+X_l)$ are disjunct intervals with side lengths p_i^{-l}, we have

$$\lambda(\mathcal{A}'_{N,l}) = \frac{\sharp\mathcal{C}_{N,l}}{(p_1\cdots p_s)^l}.$$

- *Reduction to compute a volume.*
We define the set $\mathcal{H}_{N,l} = \{\mathbf{x}_N^{\Phi(\mathbf{k})} : \mathbf{k} \in \mathcal{F}_{(l)}\}$ and the following tiling of $[0,1)^s$:

$$T_{\mathbf{k},l} = \left[\frac{k_1}{p_1^l}, \frac{k_1+1}{p_1^l}\right) \times \cdots \times \left[\frac{k_s}{p_s^l}, \frac{k_s+1}{p_s^l}\right), \qquad \mathbf{k} \in \mathcal{F}_{(l)}.$$

We know that distinct elements of $\mathcal{H}_{N,l}$ lie in distinct sets $T_{\mathbf{k},l}$. Suppose $\mathbf{u}, \mathbf{v} \in \mathcal{F}_{(l)}$ and $\mathbf{x}_N^{\Phi(\mathbf{u})}, \mathbf{x}_N^{\Phi(\mathbf{v})} \in T_{\mathbf{k},l}$. Then for $i = 1, \ldots, s$, $p_i^l | ((u_i + N) - (v_i + N))$; thus $u_i - v_i \geq p_i^l$ or $u_i - v_i = 0$. By $0 \leq u_i, v_i < p_i^l$, we have $\mathbf{u} = \mathbf{v}$.
In order to compute $\sharp\mathcal{C}_{N,l}$ we have to compute the number of elements of $\mathcal{H}_{N,l}$, which have hyperbolic distance less than $\frac{1}{Nf(N)}$. As each element of $\mathcal{H}_{N,l}$ lies in one and only one tile $T_{\mathbf{k},l}$, we get an upper bound on $\sharp\mathcal{C}_{N,l}$ by computing the number of tiles $T_{\mathbf{k},l}$ possessing a left lower corner with hyperbolic distance less than $\frac{1}{Nf(N)}$, i.e.,

$$\sharp\mathcal{C}_{N,l} \leq \sharp\left\{\mathbf{k} \in \mathcal{F}_{(l)} : ||\Phi(\mathbf{k})||_{\mathbf{0}} \leq \frac{1}{Nf(N)}\right\}.$$

Thus, we have reduced the problem to count lattice points contained in some body. The number of lattice points contained in a body is about the volume of the body divided by the volume of the fundamental parallelotope plus the number of lattice points that lie near the border. Here, to each lattice point there is a parallelotope with side lengths p_i^{-l} $(i = 1, \ldots, s)$ attached. Let R denote the number of points lying on one of the hyperplanes characterized by $\{\mathbf{z} \in [0,1]^s : x_i = 0\}$ for some $i = 1, \ldots, s$. Furthermore, define the sets

$$\mathcal{C}_{Nf(N)} = \left\{\mathbf{z} \in [0,1)^s \ : \ z_1 \cdots z_s < \frac{1}{Nf(N)}\right\}$$

and

$$\delta\mathcal{C}_{Nf(N)} = \left\{\mathbf{z} \in [0,1)^s \ : \ z_1 \cdots z_s = \frac{1}{Nf(N)}\right\}.$$

Set $\varepsilon := \sqrt{p_1^{-2l} + \cdots + p_s^{-2l}}$. The area near the border is given by

$$\Delta\mathcal{C}_{Nf(N)} = \{\mathbf{z} \in (\mathbb{R}^+)^s \ : \ |\delta\mathcal{C}_{Nf(N)} - \mathbf{z}| < \varepsilon\} \setminus \mathcal{C}_{Nf(N)},$$

where $|\delta\mathcal{C}_{Nf(N)} - \mathbf{z}|$ denotes the minimal Euclidean distance from $\delta\mathcal{C}_{Nf(N)}$ to $\mathbf{z}$.
Thus, a bound for $\sharp\mathcal{C}_{N,l}$ is given by

$$\sharp\mathcal{C}_{N,l} \leq \left(\lambda(\mathcal{C}_{Nf(N)}) + \lambda(\Delta\mathcal{C}_{Nf(N)})\right)(p_1\cdots p_s)^l + R,$$

where R is the number of lattice points lying in $\mathcal{C}_{Nf(N)} \cap \{\mathbf{z} \in [0,1)^s : z_i = 1\}$ for some $i \in \{1, \ldots, s\}$, i.e., the number of lattice points which lie on the remaining borders. By the definition of R we find that

$$R \leq (p_1\cdots p_s)^l \sum_{j=1}^{s} p_j^{-l} \leq (p_1\cdots p_s)^l \sqrt{s}\varepsilon.$$

Since $\delta\mathcal{C}_{Nf(N)}$ is convex, we may project it on each hyperplane $x_i = 0$ to get a bound for the volume of $\Delta C_{Nf(N)}$ of the form $s\epsilon$. This yields

$$\sharp\mathcal{C}_{N,l} \leq \left(\lambda(\mathcal{C}_{Nf(N)}) + (s+\sqrt{s})\varepsilon\right)(p_1 \cdots p_s)^l.$$

- *Computation of the volume.*
We claim that

$$\lambda(\mathcal{C}_k) = \frac{1}{k}\sum_{j=0}^{s-1}\frac{(\log k)^j}{j!} < \frac{(\log k)^{s-1}}{k}e.$$

Let us denote by $C_k^{(j)}$ the cylinder $\{(z_1, \dots z_s) \in [0,1)^s : z_1 \cdots z_j \leq 1/k\}$. Once we have proved $\lambda\left(C_k^{(j+1)} \setminus C_k^{(j)}\right) = \frac{(\log k)^j}{k(j!)}$, the claim follows by induction,

$$\begin{aligned}
&\lambda\left(C_k^{(j+1)} \setminus C_k^{(j)}\right) \\
&\quad= \int_{z_1=\frac{1}{k}}^1 \int_{z_2=\frac{1}{kz_1}}^1 \cdots \int_{z_j=\frac{1}{kz_1\cdots z_{j-1}}}^1 \int_{z_{j+1}=0}^{\frac{1}{kz_1\cdots z_j}} \int_{z_{j+2}=0}^1 \cdots \int_{z_s=0}^1 1 \; dz_s \cdots dz_1 \\
&\quad= \int_{z_1=\frac{1}{k}}^1 \int_{z_2=\frac{1}{kz_1}}^1 \cdots \int_{z_j=\frac{1}{kz_1\cdots z_{j-1}}}^1 \frac{1}{kz_1\cdots z_j}\; dz_j \cdots z_1 \\
&\quad= \int_{z_1=\frac{1}{k}}^1 \int_{z_2=\frac{1}{kz_1}}^1 \cdots \int_{z_{j-1}=\frac{1}{kz_1\cdots z_{j-2}}}^1 \frac{\log(kz_1\cdots z_{j-1})}{kz_1\cdots z_{j-1}}\; dz_{j-1}\cdots dz_1 \\
&\quad= \int_{z_1=\frac{1}{k}}^1 \int_{z_2=\frac{1}{kz_1}}^1 \cdots \int_{z_{j-2}=\frac{1}{kz_1\cdots z_{j-3}}}^1 \frac{(\log(kz_1\cdots z_{j-2}))^2}{2kz_1\cdots z_{j-2}}\; dz_{j-2}\cdots dz_1 \\
&\quad= \cdots = \frac{(\log k)^j}{k(j!)}.
\end{aligned}$$

- *The probability that the Nth point of a Halton sequence is close to* $\mathbf{0}$.
Remember that $\mathcal{A}'_N := \bigcap_{l=l_0}^{\infty}\mathcal{A}'_{N,l}$, where $l_0 = \left\lfloor\frac{\log(Nf(N))}{\log 2}\right\rfloor + 1$. Since $\mathcal{A}'_{N,l+1} \subset \mathcal{A}'_{N,l}$, the set A'_N is measurable.
The volume of $A'_{N,l}$ may now be estimated by

$$\lambda(\mathcal{A}'_{N,l}) \leq e\frac{(\log(Nf(N)))^{s-1}}{Nf(N)} + (s+\sqrt{s})\varepsilon.$$

Since $\varepsilon \to 0$ as $l \to \infty$, we conclude that

$$\lambda(\mathcal{A}'_N) \leq e\frac{(\log(Nf(N)))^{s-1}}{Nf(N)}.$$

- *The probability that the random-start Halton sequence is approaching* $\mathbf{0}$.
Let us define the sets

$$\mathcal{B}_0 := \{\mathbf{z} : (\mathbf{x}_n^{\mathbf{z}})_{n\geq 1} \text{ is approaching } \mathbf{0}\} \quad \text{and}$$

$$\mathcal{B}_0^{(k)} := \{\mathbf{z} : (\mathbf{x}_n^{\mathbf{z}})_{n\geq 1} \text{ has at least } k \text{ points that are close to } \mathbf{0}\}.$$

Hence, $\mathcal{B}_0^{(k+1)} \subset \mathcal{B}_0^{(k)}$ and

$$\bigcap_{k=1}^{\infty} \mathcal{B}_0^{(k)} = \mathcal{B}_0.$$

If in a sequence $(\mathbf{x}_n)$ there are k points close to $\mathbf{0}$, then the index N_k of the kth point $\mathbf{x}_{N_k}$ that is close to $\mathbf{0}$ is in the set $\{N \in \mathbb{Z}^+ : N \geq k\}$. Therefore

$$\mathcal{B}_0^{(k)} \subset \bigcup_{N=k}^{\infty} \mathcal{A}'_N.$$

If we take into account the definition of $\mathcal{B}_0$, then

$$\mathcal{B}_0 \subset \mathcal{B}'_0 := \bigcap_{k=1}^{\infty} \bigcup_{N=k}^{\infty} \mathcal{A}'_N.$$

Since the Lebesgue measure is continuous, we find

$$\mathcal{B}'_0 = \lim_{k\to\infty} \sum_{N=k}^{\infty} \lambda(\mathcal{A}'_N) \leq \lim_{k\to\infty} \sum_{N=k}^{\infty} e\frac{(\log(Nf(N)))^{s-1}}{Nf(N)} = 0,$$

by the assumption $\sum_{N=1}^{\infty} \frac{(\log(Nf(N)))^{s-1}}{Nf(N)} < \infty$. Finally, $\lambda(\mathcal{B}'_0) = 0$ includes $\lambda(\mathcal{B}_0) = 0$ by the completeness of the Lebesgue measure.

- *Deduction from one corner to all corners.*
  If a sequence $(\mathbf{x}_n^{\mathbf{z}})_{n\geq 1}$ approaches a corner $\mathbf{h} \in \{0,1\}^s$, then the sequence $\left(\mathbf{x}_n^{\mathbf{z}'}\right)_{n\geq 1}$ with $z'_i = \Phi_{p_i}(\tilde{\Phi}_{p_i}(x_i) - h_i)$ approaches the corner $\mathbf{0}$. Thus, the probability for a random-start Halton sequence to approach any corner is 2^s times the probability for approaching $\mathbf{0}$. Hence, $\lambda(\mathcal{A}) = 0$. ∎

Finally, let us give the following corollary.

COROLLARY 1. *Almost all random-start Halton sequences avoid all corners, i.e., for any fixed $\epsilon > 0$ there exists a constant $c = c(\epsilon, p_1, \ldots, p_s) > 0$ such that for any corner $\mathbf{h} \in \{0,1\}^s$ and all $n \geq 1$ the condition*

$$\|\mathbf{x}_n\|_{\mathbf{h}} > cn^{-1-\epsilon}$$

is fulfilled.

Proof. The result is a direct consequence of Theorem 5 and Lemma 1 combined with the fact that

$$\sum_{n=1}^{\infty} \frac{\log(n^{1+\epsilon})^{s-1}}{n^{1+\epsilon}} = \sum_{n=1}^{\infty} \frac{(1+\epsilon)^{s-1}\log(n)^{s-1}}{n^{1+\epsilon}} < \infty. \qquad ∎$$

Remark. Theorem 4 could also be proved by similar methods (counting lattice points in hyperbolic areas) as used in the proof of Theorem 5. However, we gave a "Diophantine proof" of Theorem 4, because it is completely different and, moreover, it provides more information on the structure whose points avoid corners. For example, it is possible to show that the (nonrandomized) Halton sequence avoids all corners with the "Diophantine approach" (cf. [4]).

On the other hand we could also use Theorem 4 in the proof of Theorem 5. Indeed instead of counting lattice points one could apply Theorem 4 directly to estimate the

number of starting points that come close to a corner. But in this case we had to restrict the proof to the case $f(n) = n^{\epsilon}$.

Lastly, note that in Corollary 1 the factor $\frac{c}{n^{1+\epsilon}}$ could be replaced by $\frac{c}{n(\log n)^s \log_2 n \cdots \log_{l-1} n(\log_l n)^{1+\epsilon}}$, where $\log_k n$ is the k times iterated logarithm of n.

REFERENCES

[1] J.-H. Evertse and H. P. Schlickewei, *A quantitative version of the absolute subspace theorem*, J. Reine Angew. Math., 548 (2002), pp. 21–127.

[2] H. Faure, *Discrépances de suites associées à un système de numération (en dimension un)*, Bull. Soc. Math. France, 109 (1981), pp. 143–182.

[3] J. H. Halton, *On the efficiency of certain quasi-random sequences of points in evaluating multidimensional integrals*, Numer. Math., 2 (1960), pp. 84–90.

[4] J. Hartinger, R. Kainhofer, and V. Ziegler, *On the corner avoidance properties of various low-discrepancy sequences*, Integers, 5 (3) (2005), A10.

[5] E. Hewitt and K. A. Ross, *Abstract Harmonic Analysis,* I, *Structure of Topological Groups. Integration Theory, Group Representations*, Grundlehren Math. Wiss. 115, Springer-Verlag, Berlin, 1963.

[6] E. Hlawka, *Uniform distribution modulo* 1 *and numerical analysis*, Compositio Math., 16 (1964), pp. 92–105.

[7] H. Niederreiter, *Point sets and sequences with small discrepancy*, Monatsh. Math., 104 (1987), pp. 273–337.

[8] H. Niederreiter, *Low-discrepancy and low-dispersion sequences*, J. Number Theory, 30 (1988), pp. 51–70.

[9] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS-NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.

[10] A. B. Owen, *Randomized qmc and Point Singularities*, Technical report, Stanford University, Palo Alto, CA, 2004.

[11] A. B. Owen, *Multidimensional variation for quasi-Monte Carlo*, in International Conference on Statistics in Honour of Professor Kai-Tai Fang's 65th Birthday, J. Fan and G. Li, eds., Hong Kong, 2005, pp. 49–74.

[12] A. B. Owen, *Halton sequences avoid the origin*, SIAM Rev., 48 (2006), pp. 487–503.

[13] H. P. Schlickewei, *Die p-adische Verallgemeinerung des Satzes von Thue-Siegel-Roth-Schmidt*, J. Reine Angew. Math., 288 (1976), pp. 86–105.

[14] W. M. Schmidt, *Diophantine Approximations and Diophantine Equations*, Lecture Notes in Math. 1467, Springer-Verlag, Berlin, 1991.

[15] I. M. Sobol', *On the distribution of points in a cube and the approximate evaluation of integrals*, USSR Comput. Math. Math. Phys., 7 (1967), pp. 86–112.

[16] I. M. Sobol', *On the use of uniformly distributed sequences for approximate computations of improper integrals*, in Theory of Cubature Formulas and Applications to Certain Problems in Mathematical Physics, S. Sobolev, ed., Novosibirsk Nauka, 1973, pp. 62–66.

[17] S. Tezuka and T. Tokuyama, *A note on polynomial arithmetic analogue of Halton sequences*, ACM Trans. Model. Comput. Simul., 4 (1994), pp. 279–284.

[18] X. Wang and F. J. Hickernell, *Randomized Halton sequences*, Math. Comput. Modelling, 32 (2000), pp. 887–899.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1122–1141

ANALYSIS OF PROFILE FUNCTIONS FOR GENERAL LINEAR REGULARIZATION METHODS*

BERND HOFMANN† AND PETER MATHÉ‡

Abstract. The stable approximate solution of ill-posed linear operator equations in Hilbert spaces requires regularization. Tight bounds for the noise-free part of the regularization error are constitutive for bounding the overall error. Norm bounds of the noise-free part which decrease to zero along with the regularization parameter are called profile functions and are the subject of our analysis. The interplay between properties of the regularization and certain smoothness properties of solution sets, which we shall describe in terms of sourcewise representations, is crucial for the decay of associated profile functions. On the one hand, we show that a given decay rate is possible only if the underlying true solution has appropriate smoothness. On the other hand, if smoothness fits the regularization, then decay rates are easily obtained. If smoothness does not fit, then we will measure this in terms of some distance function. Tight bounds for these allow us to obtain profile functions. Finally we study the most realistic case when smoothness is measured with respect to some operator which is related to the one governing the original equation only through a link condition. In many parts the analysis is done on geometric basis, extending classical concepts of linear regularization theory in Hilbert spaces. We emphasize intrinsic features of linear ill-posed problems which are frequently hidden in the classical analysis of such problems.

Key words. linear ill-posed problems, regularization, distance function, convergence rates, index function, source condition, qualification, range inclusion

AMS subject classifications. Primary, 47A52; Secondary, 65J20, 65F22, 65R30

DOI. 10.1137/060654530

1. Introduction.

We study noisy linear operator equations

$$y^\delta = Ax^\dagger + \delta\xi \qquad (\|\xi\| \le 1)\,, \tag{1.1}$$

where $A : X \to Y$ is some bounded linear operator mapping between infinite-dimensional separable Hilbert spaces X and Y and $\delta > 0$ denotes the noise level. The spaces X and Y are equipped with norms $\|\cdot\|$. The same norm symbol is also used for associated operator norms.

We assume that A is *injective* and that the range $\mathcal{R}(A)$ is *not closed* in Y. Then the *linear operator equation* $Ax = y$ has a unique solution $x = x^\dagger \in X$, for every $y \in \mathcal{R}(A)$, but the equation is *ill-posed* since A^{-1} is an unbounded operator. Thus regularization is required in order to find stable approximate solutions of the operator equation based on noisy data $y^\delta \in Y$. We consider general linear regularization schemes based on a family of piecewise continuous functions $g_\alpha(t)$ $(0 < t \le a := \|A^*A\|)$ for regularization parameters $0 < \alpha \le \overline{\alpha}$. The family g_α determines the *regularization method*. Once a regularization g_α is chosen, the approximate solution to (1.1) is given by

$$x_\alpha^\delta = g_\alpha(A^*A)A^*y^\delta.$$

*Received by the editors March 17, 2006; accepted for publication (in revised form) November 17, 2006; published electronically May 10, 2007.
http://www.siam.org/journals/sinum/45-3/65453.html

†Department of Mathematics, Chemnitz University of Technology, 09107 Chemnitz, Germany (hofmannb@mathematik.tu-chemnitz.de). This author was supported by Deutsche Forschungsgemeinschaft (DFG) under grant HO1454/7-1.

‡Weierstraß Institute for Applied Analysis and Stochastics, Mohrenstraße 39, 10117 Berlin, Germany (mathe@wias-berlin.de).

For such approximate solution x_α^δ we obtain an obvious error bound, using the intermediate quantity $x_\alpha = g_\alpha(A^*A)A^*y = g_\alpha(A^*A)A^*Ax^\dagger$, as

$$(1.2)\quad e(x^\dagger,\alpha,\delta) := \|x_\alpha^\delta - x^\dagger\| \le \|x^\dagger - x_\alpha\| + \delta\|g_\alpha(A^*A)A^*\| \qquad \text{for all} \quad 0<\alpha\le\overline{\alpha}.$$

The second summand on the right is independent of the underlying true solution. Let us denote by $r_\alpha(t) := 1 - t\,g_\alpha(t)$ $(0 < t \le a)$ the residual or bias functions related to the regularization method g_α, thus $\|x^\dagger - x_\alpha\| = \|r_\alpha(A^*A)x^\dagger\|$. In these terms, tight bounds on the norm of the residual are constitutive for the accuracy of the regularized solution. Bounds which are increasing functions in $\alpha > 0$ will give rise to what we call profile functions.

The outline is as follows. In section 2 we recall the basic underlying quantities, namely general linear regularization methods for operator equations in Hilbert space and the concept of solution smoothness in terms of general source conditions. Then, in section 3 we associate profile functions to any given regularization and to any set of smooth solutions and discuss their existence. The rate at which profile functions decay to zero turns out to be crucial and is the objective of our analysis. It will become clear that this rate depends on the underlying regularization as well as on the solution smoothness. In section 4 we indicate situations when maximal rates of decay occur, regardless of the underlying solution smoothness, namely due to the limited qualification of the regularization method. We close this part by showing that decay rates imply solution smoothness.

The constructive part of obtaining explicit descriptions of profile functions, as dependent on the qualification of the regularization and smoothness properties of the solution with respect to the operator A, is carried out in sections 5 and 6 for several degrees of generality. We start in section 5 with the easiest case, when solution smoothness is measured in terms of general source conditions given through functions of A^*A. This is then extended to the situation where a source condition is satisfied only approximately, measured in terms of a specific concept of distance functions. Tight upper bounds for such distance functions imply profile functions.

We close the analysis with section 6 discussing the situation when solution smoothness is measured with respect to a self-adjoint operator $G : X \to X$ with nonclosed range which is different from A^*A. In this case an assumption, linking A^*A and G, will allow us to draw conclusions on the decay rate of the associated profile functions.

In many parts the analysis is done on a geometric basis, extending classical concepts as those used in the theory of linear ill-posed equations in Hilbert space. By doing so we not only extend previous results to a more general situation, but we aim at emphasizing intrinsic features of the problems under consideration. Such features are often hidden in the classical analysis of linear ill-posed problems.

2. General linear regularization methods and general smoothness. As mentioned in the introduction, profile functions will be assigned to regularization methods and solution sets of (1.1). We start with the notion of a general linear regularization scheme. Then we turn to the description of solution smoothness in terms of general source conditions.

The basic underlying objects are index functions, and we recall the following definition, as known in the literature (e.g., [7, 14, 3]).

DEFINITION 2.1. *A real function $\varphi(t)$ $(0 < t \le \bar{t})$ is called an* index function *if it is continuous, strictly increasing, and satisfies the limit condition* $\lim_{t\to 0+}\varphi(t) = 0$.

2.1. General regularization methods.

DEFINITION 2.2. *A family of functions $g_\alpha(t)$ $(0 < t \le a)$, defined for parameters $0 < \alpha \le \overline{\alpha}$, is called a* regularization *if they are piecewise continuous in α and the following three properties are satisfied:*

(i) *For each $0 < t \le a$ there is convergence $|r_\alpha(t)| \to 0$ as $\alpha \to 0$.*

(ii) *There is a constant γ_1 such that $|r_\alpha(t)| \le \gamma_1$ for all $0 < \alpha \le \overline{\alpha}$.*

(iii) *There is a constant γ_* such that $\sqrt{t}\,|g_\alpha(t)| \le \gamma_*/\sqrt{\alpha}$ for all $0 < \alpha \le \overline{\alpha}$.*

Example 2.3. The most famous method of regularization is the *Tikhonov method* with $g_\alpha(t) = 1/(t+\alpha)$, which satisfies the properties of Definition 2.2 for the constants $\gamma_1 = 1$ and $\gamma_* = 1/2$ and arbitrarily large $\overline{\alpha} > 0$.

Example 2.4. Another common regularization method is *spectral cut-off*, which is given as

$$g_\alpha(t) = \begin{cases} 0 & (0 < t < \alpha) \\ 1/t & (\alpha \le t \le a) \end{cases} \quad \text{with respective residual } r_\alpha(t) = \begin{cases} 1 & (0 < t < \alpha) \\ 0 & (\alpha \le t \le a). \end{cases}$$

Obviously this obeys the properties from Definition 2.2 with $\gamma_1 = \gamma_* = 1$. Also for that method, the upper bound $\overline{\alpha}$ for the regularization parameter can be selected arbitrarily.

Example 2.5. Iterative regularization methods, as, for instance, *Landweber iteration*, where for some $0 < \mu < 1/\|A^*A\|$ we let

$$x_n^\delta := \mu \sum_{j=0}^{n-1} (I - \mu A^*A)^j A^* y^\delta, \quad n = 1, 2, \dots,$$

are conform to this approach when assigning $n := \lfloor 1/\alpha \rfloor$ $(0 < \alpha < 1)$. Thus with this identification we obtain $g_\alpha(t) := 1/t\left(1 - (1-\mu t)^{\lfloor 1/\alpha \rfloor}\right)$ and the corresponding residual $r_\alpha(t) := (1-\mu t)^{\lfloor 1/\alpha \rfloor}$ $(0 < \alpha < 1)$, hence obviously $\gamma_1 = 1$. It remains to bound γ_*. Bernoulli's inequality yields $1 - n\mu t \le (1-\mu t)^n$, which can be used to bound

$$\sqrt{t} g_\alpha(t) = 1/\sqrt{t}\,(1 - (1-\mu t)^n) \le (1/t\,(1 - (1-\mu t)^n))^{1/2} \le \sqrt{\mu n}\,.$$

By the definition of n this yields $\gamma_* = \sqrt{\mu}$.

The above requirements (i)–(iii) are made to ensure convergence of regularization methods for any given element $x^\dagger \in X$. However, these are not enough to describe rates of convergence.

As introduced in the papers [13, 14, 15, 16], we measure the *qualification* of any regularization method in terms of index functions ψ.

DEFINITION 2.6. *Let $\psi(t)$ $(0 < t \le a)$ be an index function. A regularization g_α for the operator equation* (1.1) *is said to have qualification ψ with constant $\gamma \in (0, \infty)$ if*

$$\sup_{0<t\le a} |r_\alpha(t)|\,\psi(t) \le \gamma\,\psi(\alpha) \qquad \text{for all} \quad 0 < \alpha \le a\,. \tag{2.1}$$

This definition generalizes the concept of qualification of a regularization method as a finite number or infinity, as, for example, used in [5]. We remark that a first systematic discussion of the interrelations between solution smoothness and the traditional concept of qualification was given in [25, 26].

For Tikhonov regularization (see Example 2.3) we can give sufficient conditions for ψ being a qualification in different ways, as this is formulated in the following proposition. For more details and proofs we refer to [15, 16, 3].

PROPOSITION 2.7. *The index function $\psi(t)$ $(0 < t \le a)$ is a qualification of Tikhonov regularization with constant $\gamma = 1$ if either* (a) *$\psi(t)/t$ is nonincreasing on $(0, a]$ or* (b) *$\psi(t)$ is concave on $(0, a]$.*

If there exists an argument $\hat{t} \in (0, a)$ such that (c) *$\psi(t)/t$ is nonincreasing on $(0, \hat{t}\,]$ or* (d) *$\psi(t)$ is concave on $(0, \hat{t}\,]$, then ψ is a qualification with constant $\gamma = \psi(a)/\psi(\hat{t}\,)$.*

2.2. Measuring solution smoothness. In a wide sense the smoothness of expected solutions $x^\dagger$ to (1.1) can be written as a property of the form $x^\dagger \in M$ with $M \subseteq \mathcal{R}(G)$ for some "smoothing" linear operator $G : X \to X$, where G is assumed to be positive self-adjoint with nonclosed range $\mathcal{R}(G)$ (see also [3, 18]). Specifically, here we shall assume that the solution $x^\dagger$ belongs to a set

$$G_\tau(R) := \{x \in X : \ x = \tau(G)w, \quad \|w\| \le R\} \tag{2.2}$$

with some index function $\tau(t)$ $(0 < t \le \|G\|)$.

As the following lemma asserts, such a set is closed in X and even compact whenever G is compact.

LEMMA 2.8. *For a positive self-adjoint bounded linear operator $G : X \to X$ and an index function $\tau(t)$ $(0 < t \le \|G\|)$ the set $G_\tau(R)$ from* (2.2) *is closed in X. Moreover, $G_\tau(R)$ is a compact subset of X whenever G is a compact operator.*

Proof. First we show that $G_\tau(R)$ is a closed subset in X. We show that the image $\{x \in X : x = Gw, \ w \in X, \ \|w\| \le R\}$ of the centered ball with radius R in X with respect to any bounded positive self-adjoint linear operator $G : X \to X$ is a closed subset of X. Since $\tau(G)$ has the same properties as a consequence of the boundedness of any index function τ, this shows the closedness of $G_\tau(R)$. Consider a convergent sequence of images $Gx_n \to y_0 \in X$ with $\|x_n\| \le R$. Since any closed ball in X is weakly precompact and weakly closed, there is a weakly convergent subsequence $x_{n_k} \rightharpoonup x_0$ with $\|x_0\| \le R$. Since every continuous operator G is also weakly continuous and hence weakly closed, this implies the weak convergence $Gx_{n_k} \rightharpoonup Gx_0$, thus $y_0 = Gx_0$ which shows the required closedness. Moreover, for compact G it is evident that $\tau(G) : X \to X$ is a compact operator and then $G_\tau(R)$ is a precompact subset of X. Since $G_\tau(R)$ is closed in X, this implies the compactness and proves the lemma. ∎

In our analysis below for index functions τ we shall assign pairs (G, τ) Hilbert spaces X_τ^G having $G_\tau(1)$ as their unit balls. In particular, we use the shortcut $H := A^*A$ and consider Hilbert spaces X_φ^H for index functions φ with the set $H_\varphi(1)$ as unit ball, where we define

$$H_\varphi(R) := \{x \in X : \ x = \varphi(A^*A)w, \quad \|w\| \le R\}\,. \tag{2.3}$$

Corresponding norms will be denoted by $\|\cdot\|_{X_\tau^G}$ and $\|\cdot\|_{X_\varphi^H}$, respectively. This construction is basically due to [6].

3. Profile functions. In this section we shall introduce the notion of a profile function, discuss the problem of existence, and show that their decay is related to smoothness of the underlying solution $x^\dagger$ of (1.1).

3.1. Definition and existence. Having chosen a linear regularization g_α, and having fixed a set $M \subset X$ of possible solutions to (1.1) we assign profile functions as follows.

DEFINITION 3.1. *An index function* $f\colon (0,\overline{\alpha}] \to (0,\infty)$ *is called* profile function for (M, g_α) *whenever*

$$\sup_{x\in M} \|r_\alpha(A^*A)x\| \le f(\alpha) \qquad \textit{for all} \quad 0<\alpha\le\overline{\alpha}\,. \tag{3.1}$$

In the definition we suppress the dependence of profile functions f on the operator A, governing (1.1). If $M := \{x\} \in X$ is a singleton, then we shall write (x, g_α) instead of $(\{x\}, g_\alpha)$. Note that the bound (3.1) is required only for $\alpha \le \overline{\alpha}$, which is useful for asymptotic considerations as $\delta \to 0$ in (1.1).

The character of possible profile functions f for (M, g_α) is closely connected with three ingredients and their interplay. In this context, properties of the regularization g_α as first component and of the set $M \subset X$ expressing the solution smoothness as second components meet as third component the smoothing behavior of the operator A in (1.1), which leads to the nonclosedness of the range $\mathcal{R}(A)$.

Remark 3.2. Once a profile function $f(\alpha)$ as above is found, together with property (iii) of Definition 2.2, we may then continue the estimate (1.2) to derive

$$e(x^\dagger,\alpha,\delta) \le f(\alpha) + \frac{\gamma_*\,\delta}{\sqrt{\alpha}} \qquad \text{for all} \quad 0<\alpha\le\overline{\alpha}\,, \tag{3.2}$$

uniformly for $x^\dagger \in M$. The bound on the right in (3.2) can be balanced with respect to the choice of α depending on δ. To this end we consider the index function

$$\Theta(\alpha) := \sqrt{\alpha}\, f(\alpha) \qquad (0<\alpha\le\overline{\alpha}).$$

Let $\alpha_* = \alpha_*(\delta) = \Theta^{-1}(\delta)$ $(0 < \delta \le \Theta(\bar\alpha))$. Then we obtain uniformly for $x^\dagger \in M$ that

$$e(x^\dagger,\alpha_*,\delta) \le (1+\gamma_*) f(\alpha_*). \tag{3.3}$$

Thus the function $f(\Theta^{-1}(\delta))$ yields a *convergence rate* of the regularization g_α for $x^\dagger$ as $\delta \to 0$. This rate is achieved by an a priori parameter choice $\alpha_* = \alpha_*(\delta)$.

First we shall establish that profile functions exist for any regularization g_α and compact subsets $M \subset X$.

PROPOSITION 3.3. *Let* g_α *be any regularization and* $M \subset X$ *be compact. Then there is a profile function for* (M, g_α).

Proof. From the properties (i) and (ii) of Definition 2.2, we deduce for $\alpha \to 0$ pointwise convergence $r_\alpha(A^*A)x \to 0$ for all $x \in X$ (see, e.g., [5, Theorem 4.1]). This convergence is uniform on compact sets $M \subset X$. Hence we have

$$h(\alpha) := \sup_{x\in M} \|r_\alpha(A^*A)x\| \to 0 \quad \text{as } \alpha\to 0.$$

Its increasing majorant $\bar h(\alpha) := \sup_{0<s\le\alpha} h(s)$, which is well-defined for sufficiently small positive α, satisfies $\lim_{\alpha\to 0} \bar h(\alpha) = 0$. If $\bar h(\alpha)$ is continuous and nonvanishing, then it is a profile function. Otherwise, suppose $\bar h(s) = 0$ for some $s > 0$. We fix some $t > 0$ with $\bar h(t) > 0$ and let

$$\hat h(x) := \begin{cases} \bar h(x), & x > t,\\ \bar h(t), & s < x \le t,\\ x/s\ \bar h(t), & 0 < x \le s, \end{cases}$$

which, when continuous, defines an index function. ∎

Thus if G is compact and τ is an index function, then for any regularization g_α there are profile functions for $(G_\tau(R), g_\alpha)$, where the sets $G_\tau(R)$ were defined in (2.2).

On the other hand, there cannot exist profile functions for (M, g_α), where $M := \{x \in X : \|x\| \le 1\}$ is the unit ball in X. Their existence would imply that $\|r_\alpha(A^*A)\|$ tends to zero as $\alpha \to 0$ and hence that the range $\mathcal{R}(A)$ were closed, which would be contrary to the ill-posedness of the problem under consideration (see, e.g., [23] and [5, Chapter 3.1]). More generally, extending this argument, profile functions cannot exist for (M, g_α), whenever M possesses an interior point.

However, there are profile functions for noncompact sets. In Proposition 5.1, profile functions for $(H_\varphi(R), g_\alpha)$ will be obtained, where the operator A may be compact (ill-posedness of type II in the sense of Nashed [19]) or noncompact (ill-posedness of type I). In the latter case this yields noncompact sets $M = H_\varphi(R)$. Another specific example of profile functions for the noncompact set $M = \{x \in L^\infty(0,1) : \|x\|_{L^\infty(0,1)} \le R\} \subset X = L^2(0,1)$ for the Tikhonov regularization and multiplication operators A mapping in $L^2(0,1)$ can be taken from [10]. This is not by chance and some explanation will be given in Remark 5.2. Roughly speaking, if smoothness properties of M are appropriate for the underlying operator A from (1.1), then profile functions exist for (M, g_α), regardless of their compactness. In this respect, compactness of M may be viewed as universal (problem independent) smoothness.

3.2. Decay rates yield solution smoothness. To exhibit the fact that a decay rate of a profile function implies solution smoothness in the sense of section 2.2, we start with the following result, which extends analysis in [21], and we also refer to the recent monograph [1]. We recall that the operator $H = A^*A$ admits a spectral resolution with a family $(E_\lambda)_{0<\lambda\le a}$ of projections, which is assumed to be such that $\lambda \mapsto \|E_\lambda x^\dagger\|^2$ is left continuous, thus representing a (spectral) measure. We start with the following technical result from [21, Lemma 2.1]; see also [5, Proof of Proposition 4.13].

LEMMA 3.4. *Let $g_\alpha(t)$ $(0 < t \le a,\ 0 < \alpha \le \bar\alpha)$ be a regularization with constant γ_*. If $0 < t \le \min\{\alpha, a\}$, then $\left|r_{(4\gamma_*^2\alpha)}(t)\right| \ge 1/2$.*

The above lemma yields the following estimate.

LEMMA 3.5. *Let g_α be a regularization with constant γ_* as in property* (iii) *of Definition* 2.2. *The following estimate holds true:*

$$\|r_{(4\gamma_*^2\alpha)}x^\dagger\| \ge \frac{1}{2}\left(\int_0^\alpha d\|E_\lambda x^\dagger\|^2\right)^{1/2} \qquad \text{for all} \quad 0 < \alpha \le \min\left\{a, \bar\alpha/4\gamma_*^2\right\}. \tag{3.4}$$

Before turning to the main result of this section we state the following lemma.

LEMMA 3.6. *Suppose $\varphi(t)$ $(0 < t \le \bar t)$ is an index function. There is a sequence $f_n(t)$ $(0 < t \le \bar t)$ of step functions of the form $\sum_{j=1}^m c_j \chi_{(0,\alpha_j)}(t)$ converging to $1/\varphi(t)$ pointwise and $f_n(t) \le 1/\varphi(t)$.*

Proof. Given any such φ and $n \in \mathbb{N}$ large enough $n \ge n_0$, we let $f(t) = 1/\varphi(t)$ and truncate at $t_n = f^{-1}(n) < \bar t$ to obtain $g^n(t)$ $(0 \le t \le \bar t)$, which is a nonincreasing bounded continuous function on the closed interval $[0, \bar t]$. Thus there is a step function $f_n(t)$ of the required form, satisfying $|f_n(t) - g^n(t)| \le 1/n$. The sequence $f_n(t)$ $(0 < t \le \bar t)$, $n = n_0, n_0+1, \ldots$, converges pointwise to f. ∎

Given a regularization g_α with constant γ_* and any index function $h(t)$ $(0 < t \le$

a), we can assign a nonnegative measure Φ_h on $(0,a]$ by letting

$$\Phi_h[0,\alpha) := h(4\gamma_*^2\alpha) \quad (0 < 4\gamma_*^2\alpha \le a).$$

With this notation we can formulate the following result.

THEOREM 3.7. *Let $g_\alpha(t)$ $(0 < t \le a)$ for the parameters $0 < \alpha \le \bar\alpha$ be a regularization with constant γ_*. We assume that the index function $f(\alpha)$ $(0 < \alpha \le \bar\alpha)$ is a profile function for $(x^\dagger, g_\alpha)$ with associated measure $\Phi = \Phi_{f^2}$, restricted to the interval $J_* := (0, \min\{a, \bar\alpha/4\gamma_*^2\}]$. Then the following assertions are true:*

(a) *If ψ is any index function such that $1/\psi \in L^2(J_*, d\Phi)$, then necessarily $x^\dagger \in X_\psi^H$.*

(b) *Let ψ be an index function for which $t \mapsto 1/(\psi^2((f^2)^{-1}(t))) \in L^1_{\rm loc}(J_*, dt)$, i.e., it is locally integrable. Then $x^\dagger \in X_\psi^H$.*

Proof. Using Lemma 3.5 and the fact that $f(\alpha)$ $(0 < \alpha \le \bar\alpha)$ is assumed to be a profile function for $(x^\dagger, g_\alpha)$ we conclude that the estimate

$$\frac{1}{4}\int_0^\alpha d\|E_\lambda x^\dagger\|^2 \le \|r_{(4\gamma_*\alpha)}x^\dagger\|^2 \le f^2(4\gamma_*^2\alpha) = \int_0^\alpha d\Phi(\lambda) \quad (\alpha \in J_*) \tag{3.5}$$

is valid.

Now let ψ be any index function such that $1/\psi(t) \in L^2(J_*, d\Phi)$. By Lemma 3.6 we can find a sequence $f_n(t)$ of step functions on J_*, converging to $1/\psi^2(t)$ pointwise. Using (3.5) and the particular form of f_n we deduce that

$$\frac{1}{4}\int_{J_*} f_n(\lambda) d\|E_\lambda x^\dagger\|^2 \le \int_{J_*} f_n(\lambda) d\Phi(\lambda) \le \int_{J_*} \frac{1}{\psi^2(\lambda)} d\Phi(\lambda).$$

By Fatou's lemma we conclude that also $1/\psi(t) \in L^2(J_*, d\|E_\lambda x^\dagger\|^2)$ and

$$\|1/\psi\|_{L^2(J_*, d\|E_\lambda x^\dagger\|^2)} \le 2\|1/\psi\|_{L^2(J_*, d\Phi)}.$$

Consequently,

$$\begin{aligned}
\|x^\dagger\|_{X_\psi^H}^2 &= \int_0^a \frac{1}{\psi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 \\
&= \int_{J_*} \frac{1}{\psi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 + \int_{(0,a]\setminus J_*} \frac{1}{\psi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 \\
&\le 4\|1/\psi\|^2_{L^2(J_*, d\Phi)} + \frac{1}{\min_{\lambda\in(0,a]\setminus J_*}\psi^2(\lambda)}\|x^\dagger\|^2 < \infty,
\end{aligned} \tag{3.6}$$

because the second summand on the right is finite, which proves assertion (a).

We use a change of measure to establish assertion (b). The proof is complete. ∎

Remark 3.8. If the interval J_* coincides with $(0,a]$, then the second summand on the right in (3.6) does not appear and we get a bound $\|x^\dagger\|_{X_\psi^H} \le 2\|1/\psi\|_{L^2((0,a], d\Phi)}$. The following elementary observation is useful.

LEMMA 3.9. *Suppose ψ, ψ_1 and f, f_1 are pairs of index functions which are related by some common strictly increasing function g as $f(t) = f_1(g(t))$ and $\psi(t) = \psi_1(g(t))$ on the respective domains of the definition. Then it holds true that $f(\psi^{-1}(t)) = f_1(\psi_1^{-1}(t))$.*

Theorem 3.7 also covers cases which were known before, like the ones discussed in the following examples.

Example 3.10 (see [21]). If the profile function f for $(x^\dagger, g_\alpha)$ is a monomial $f(\alpha) = \alpha^\nu$ for some $\nu > 0$, then we can draw the following conclusion. For every monomial $\psi(t) = t^\mu$ we obtain $1/\psi^2((f^2)^{-1}(t)) = t^{-\mu/\nu}$, which is integrable on every finite interval for $\mu < \nu$. Hence we deduce that necessarily $x^\dagger \in X_\psi^H$ for all $0 < \mu < \nu$.

Example 3.11 (see [12, Theorem 8]). If the profile function f for $(x^\dagger, g_\alpha)$ is of logarithmic type, say $f(\alpha) = \log^{-\nu}(1/\alpha)$ $(0 < \alpha < 1)$ for some $\nu > 0$, then by using Lemma 3.9 we also deduce that necessarily $x^\dagger \in X_\psi^H$ for all functions $\psi(t) = \log^{-\mu}(1/t)$ $(0 < t < 1)$ with $\mu < \nu$, because both are related to the respective functions from Example 3.10 through $g(t) := \log^{-1}(1/t)$ $(0 < t < 1)$.

4. Lower bounds for profile functions. In general profile functions $f(\alpha)$ can decrease to zero arbitrarily fast as α tends to zero. This is, for instance, the case when g_α is chosen as spectral cut-off in Example 2.4 and $x^\dagger$ is an eigenelement of A^*A, in which case $\|r_\alpha(A^*A)x^\dagger\| \equiv 0$ for α small enough.

However, for many regularization methods there is a maximal speed of convergence $\|r_\alpha(A^*A)x^\dagger\| \to 0$ as $\alpha \to 0$, for any $x^\dagger \neq 0$, regardless of its smoothness. This phenomenon is related to *saturation*, as was studied, e.g., in [21, 22], and in more generality in [13], from which the present approach is taken. The impact of limited qualification on profile functions can be seen under an additional convexity assumption.

THEOREM 4.1. *Let g_α be any regularization with residual r_α. Suppose that for all $0 < t \le a$ the functions*

$$\alpha \mapsto |r_\alpha(t)| \qquad (0 < \alpha \le \overline{\alpha}) \tag{4.1}$$

are increasing, and for all $0 < \alpha \le \overline{\alpha}$ the functions

$$t \mapsto |r_\alpha(t)|^2 \quad (0 < t \le a) \tag{4.2}$$

are convex. Let $\bar\psi$ be given as

$$\bar\psi(\alpha) := \inf_{0<t\le a} |r_\alpha(t)| \quad (0 < \alpha \le \overline{\alpha}). \tag{4.3}$$

Then for each $0 \neq x \in X$ we have

$$\bar\psi(\alpha) \le \frac{1}{\|x\|}\|r_\alpha(A^*A)x\| \qquad \textit{for all} \quad 0 < \alpha \le \overline{\alpha}. \tag{4.4}$$

Hence $\bar\psi$ is a nondecreasing lower bound to any profile function for (x_0, g_α) uniformly for all elements $x_0 \in X$ of the unit sphere, i.e., with $\|x_0\| = 1$.

Sketch of a proof. To prove that $\bar\psi$ is a lower bound to any profile function for (x_0, g_α) we use a Jensen-type inequality (see, e.g., [13]), which yields that under (4.2) we have

$$\bar\psi(\alpha) \le \left|r_\alpha(\|Ax\|^2/\|x\|^2)\right| \le \frac{\|r_\alpha(A^*A)x\|}{\|x\|} \qquad \text{for all} \quad 0 < \alpha \le \overline{\alpha}.$$

Moreover, under (4.1) the function $\bar\psi$ is nondecreasing. This completes the proof. ∎

Remark 4.2. In many cases, the above function $\bar\psi(\alpha)$ turns out to be a qualification of the regularization g_α. In such a case it is maximal qualification.

We shall exhibit the above result in the following examples.

Example 4.3. For Tikhonov regularization as in Example 2.3 we easily verify that the assumptions are satisfied. We conclude that $\bar\psi(\alpha) = \alpha/(\alpha + a)$ with $\bar\psi(\alpha) \ge \alpha/(2a)$ $(0 < \alpha < a)$. In this case this corresponds to the maximal qualification which is $\psi(\alpha) = \alpha$.

Example 4.4. The n-fold iterated Tikhonov regularization, which has $r_\alpha(t) = (\alpha/(\alpha+t))^n$ as its residual function, also satisfies the assumptions from Theorem 4.1 and $\bar\psi(\alpha) = (\alpha/(\alpha + a))^n \ge (\alpha/(2a))^n$. This method corresponds to the maximal known qualification $\psi(\alpha) = \alpha^n$.

As in [13] we close with the following example, which is interesting as it shows that regularization, which has arbitrary classical qualification in the form $\psi(t) = t^q$ for any $0 < q < \infty$, still has a limited rate of decay for the profile functions, although these can decay exponentially fast.

Example 4.5. Landweber iteration from Example 2.5 also satisfies all the assumptions. The function $\bar\psi$, letting $0 < b := 1/(1 - \mu a) < \infty$, turns out to be $\bar\psi(\alpha) = (1 - \mu a)^{\lfloor 1/\alpha \rfloor} \ge \exp(-b/\alpha)$ $(0 < \alpha < 1)$.

Finally we stress that spectral cut-off, as in Example 2.4, does not fulfill the above assumptions. Moreover, formally we would obtain the lower bound $\bar\psi(\alpha) \equiv 0$, which is trivial.

Remark 4.6. Lower bounds for profile functions are related to the saturation phenomenon as we shall briefly sketch. The following estimate is shown in the cause of the proof of the theorem in [13]:

$$\sup_{\|\xi\| \le 1} e(x^\dagger, g_\alpha, \delta) \ge \max\left\{\|r_\alpha(A^*A)x^\dagger\|, \delta/\sqrt{\alpha}\right\} \quad (0 < \alpha \le \bar\alpha). \tag{4.5}$$

Thus, if $\bar\psi(\alpha)$ is a lower bound as in (4.4), then for any $x^\dagger$ with $\|x^\dagger\| = 1$ we derive that

$$\sup_{\|\xi\| \le 1} e(x^\dagger, g_\alpha, \delta) \ge \max\left\{\bar\psi(\alpha), \delta/\sqrt{\alpha}\right\} \ge \bar\psi(\Theta^{-1}(\delta)) \quad (0 < \alpha \le \bar\alpha)$$

with $\Theta(t) := \sqrt{t}\,\bar\psi(t)$ $(0 < t \le \overline{\alpha})$. Hence, the function $\bar\psi(\Theta^{-1}(\delta))$ is a lower bound for the error at $x^\dagger$, no matter how smooth the true solution $x^\dagger \in X$ was.

The functions $\bar\psi$ derived in Examples 4.3–4.5 can be seen to be the saturation rates caused by the limited qualifications of the underlying regularization methods.

5. Impact of solution smoothness.

As stressed earlier, the behavior of profile functions is determined by both the chosen regularization g_α and the underlying solution smoothness. As introduced in section 2.2, we measure this in terms of smoothness conditions of the form $x^\dagger \in G_\tau(R)$, see (2.2), determined by an operator G and an index function τ. The impact of such a smoothness assumption on the decay rate of profile functions is best seen if G is a function of A^*A.

5.1. G as a function of A^*A.

To obtain profile functions f for the regularization method g_α, the concept of *general source conditions*, as expressed in

$$x^\dagger = \psi(A^*A)\, w \qquad (w \in X,\ \|w\| \le R), \tag{5.1}$$

for some index functions $\psi(t)$ $(0 < t \le a)$ was used recently (see, e.g., [12, 14, 15, 24]). We note that (5.1) is a specific smoothing condition (2.2) with $\tau(G) = \psi(A^*A)$ (cf. [3] for further discussion of such conditions).

We are going to find profile functions f uniformly for sets $H_\psi(R)$, as defined by formula (2.3), provided the corresponding function ψ is a qualification of the chosen regularization g_α.

PROPOSITION 5.1. *Let the index function ψ be a qualification of the regularization method g_α with constant $0 < \gamma < \infty$. Then uniformly for each $x^\dagger \in H_\psi(R)$ the inequality*

$$\|x_\alpha - x^\dagger\| \le \gamma\, R\, \psi(\alpha) \qquad \text{for all} \quad 0 < \alpha \le a \tag{5.2}$$

is valid. Hence $f(\alpha) := \gamma\, R\, \psi(\alpha)$ is a profile function for $(H_\psi(R), g_\alpha)$.

Proof. From spectral theory (see, e.g., [5, formula (2.47)]) we have with (5.1) that

$$\|x_\alpha - x^\dagger\| = \|r_\alpha(A^*A)\, x^\dagger\| = \|r_\alpha(A^*A)\, \psi(A^*A)\, w\| \le R \sup_{0<t\le a} |r_\alpha(t)|\, \psi(t)\,.$$

Taking into account inequality (2.1), this yields (5.2) and proves the proposition. ∎

Remark 5.2. This proposition can be reformulated as follows. Suppose that we are given a pair (M, g_α) of a solution set M and a regularization g_α. If we can find an index function ψ on $(0, a]$ that is *both* a qualification for g_α and a smoothness for M, i.e., $M \subseteq H_\psi(R)$ for some R, then there is a profile function for (M, g_α). In addition the index function ψ provides a decay rate. Although this is a simple observation it explains the existence of profile functions for noncompact sets M, as discussed at the end of section 3.1.

5.2. Approximate source conditions. An important extension of the above concept is obtained by relaxing requirement (5.1). In this context, we restrict ourselves to a fixed index function $\varphi(t)$ $(0 < t \le a)$ as *benchmark function*. We suppose that the solution $x^\dagger \in X$ of (1.1) is not smooth enough to satisfy a condition (5.1) with φ instead of ψ even if $R \ge 0$ is arbitrary large. The injectivity of A implies the injectivity of $\varphi(A^*A)$ for any index function φ. Hence the range $\mathcal{R}(\varphi(A^*A))$ is dense in X. Consequently, for all $0 \le R < \infty$ the element $x^\dagger$ satisfies such a general source condition in an approximate manner as $x^\dagger = \varphi(A^*A)\, w + \xi \quad (\|w\| \le R,\ \xi \in X)$, where the norm of the perturbation $\|\xi\|$ tends to zero as R tends to infinity.

In the following we shall confine to this situation, when

$$x^\dagger \notin \mathcal{R}(\varphi(A^*A)). \tag{5.3}$$

The quality of the approximation of $x^\dagger$ by elements from $H_\varphi(1)$ can be expressed by favor of the *distance function*

(5.4)
$$\rho_{x^\dagger}(t) = \rho_{x^\dagger}^{(H,\varphi)} := \operatorname{dist}(tx^\dagger, H_\varphi(1)) = \inf\left\{\|tx^\dagger - \varphi(H)v\| : v \in X, \|v\| \le 1\right\} \quad (t > 0).$$

If the reference to the benchmark (H, φ) is clear, as in the following lemma, then we shall omit the superscript.

LEMMA 5.3. *Under the assumption* (5.3) *the functions $\rho_{x^\dagger}(t)$ and $\rho_{x^\dagger}(t)/t$ $(t > 0)$ are both index functions. Moreover, we have $\lim_{t\to\infty} \rho_{x^\dagger}(t) = \infty$.*

Proof. The idea of the proof is standard in regularization theory. For each $t > 0$ the value $\rho_{x^\dagger}(t)/t = \operatorname{dist}(x^\dagger, H_\varphi(1/t))$ is obtained from constrained minimization, and Lagrange multipliers can be used to determine this value. Hence, given $x^\dagger \in X$ let

$$F_{x^\dagger}(\lambda) := \|x^\dagger - \varphi(A^*A)v\|^2 + \lambda\|v\|^2.$$

At given λ its minimizer with respect to $v \in X$ is

$$v_\lambda := \left[\varphi^2(A^*A) + \lambda I\right]^{-1} \varphi(A^*A)x^\dagger,$$

which has to obey the side constraint $\chi(\lambda) = 1/t$, where setting

$$\chi(\lambda) := \| \left[\varphi^2(A^*A) + \lambda I\right]^{-1} \varphi(A^*A)x^\dagger \|. \tag{5.5}$$

Based on the injectivity of $\varphi(A^*A)$, spectral calculus yields that the function $\chi(\lambda)$ ($\lambda > 0$) is positive, continuous, and strictly decreasing to zero as $\lambda \to \infty$. Moreover, under (5.3) we have $\lim_{\lambda \to 0+} \chi(\lambda) = \infty$. Therefore for all $t > 0$ the function $\lambda(t) := \chi^{-1}(1/t)$ exists and is an index function. Hence we obtain

$$\rho_{x^\dagger}(t)/t = \|x^\dagger - \varphi(A^*A)v_{\lambda(t)}\| = \lambda(t) \, \| \left[\varphi^2(A^*A) + \lambda(t) I\right]^{-1} x^\dagger \| \quad (t > 0), \tag{5.6}$$

which is the composition of two index functions in t. As a consequence, both functions $\rho_{x^\dagger}(t)/t$ and $\rho_{x^\dagger}(t)$ have that property. On the other hand, we have

$$\lim_{t \to \infty} \rho_{x^\dagger}(t) = \lim_{t \to \infty} t \left(\frac{\rho_{x^\dagger}(t)}{t} \right) = \infty \,,$$

since $\rho_{x^\dagger}(t)/t$ as an index function cannot tend to zero as $t \to \infty$. This completes the proof. ∎

Remark 5.4. By using distance functions of the form

$$d(R) := \operatorname{dist}(x^\dagger, H_\varphi(R)) \, = \, R\rho_{x^\dagger}(1/R) \quad (0 < R < \infty), \tag{5.7}$$

error estimates for the Tikhonov regularization were obtained in [9] and [4]; see also [2, 8, 11] for variants thereof. The fundamental estimate for profile functions under approximate source conditions is as follows.

THEOREM 5.5. *Let g_α be a regularization method with qualification φ and constant γ. If the solution $x^\dagger$ to* (1.1) *obeys* (5.3)*, then*

$$\|x_\alpha - x^\dagger\| \le \max\{\gamma, \gamma_1\} \, \frac{1}{t} \, (\rho_{x^\dagger}(t) + \varphi(\alpha)) \quad \textit{for all} \quad t > 0 \quad \textit{and} \quad 0 < \alpha \le a. \tag{5.8}$$

Thus the function

$$f(\alpha) := 2 \max\{\gamma, \gamma_1\} \, \frac{\varphi(\alpha)}{\rho_{x^\dagger}^{-1}(\varphi(\alpha))} \quad (0 < \alpha \le a) \tag{5.9}$$

is a profile function for $(x^\dagger, g_\alpha)$.

Proof. First we establish (5.8). For any $v \in X$ with $\|v\| \le 1$ we can estimate

$$\begin{aligned}
\|x_\alpha - x^\dagger\| &= \frac{1}{t} \|r_\alpha(A^*A)tx^\dagger\| \\
&= \frac{1}{t} \|r_\alpha(A^*A)\, tx^\dagger - r_\alpha(A^*A)\, \varphi(A^*A)v + r_\alpha(A^*A)\, \varphi(A^*A)v\| \\
&\le \frac{1}{t} \left(\|r_\alpha(A^*A)(tx^\dagger - \varphi(A^*A)v)\| \, + \, \|r_\alpha(A^*A)\, \varphi(A^*A)v\| \right) \\
&\le \frac{1}{t} \left(\gamma_1 \|tx^\dagger - \varphi(A^*A)v\| \, + \, \|r_\alpha(A^*A)\, \varphi(A^*A)\| \right) \\
&\le \frac{1}{t} \left(\gamma_1 \, \|tx^\dagger - \varphi(A^*A)v\| + \gamma \, \varphi(\alpha) \right).
\end{aligned}$$

Since this estimate remains true if we substitute $\|tx^\dagger - \varphi(A^*A)v\|$ by its infimum over all v from the unit ball of X and since φ is a qualification of the used regularization method, we obtain

$$\|x^\dagger - x_\alpha\| \le \max\{\gamma, \gamma_1\} \, \frac{1}{t} \, (\rho_{x^\dagger}(t) + \varphi(\alpha)) \quad \text{for all} \quad t > 0 \quad \text{and} \quad 0 < \alpha \le a,$$

which proves estimate (5.8). Since this estimate is valid for all $t > 0$ and we have by Lemma 5.3 for the index function $\rho_{x^\dagger}$ the limit condition $\lim_{t\to\infty} \rho_{x^\dagger}(t) = \infty$, we can equate the two terms in brackets of the right-hand side of (5.8). Taking into account the strict monotonicity of function $\rho_{x^\dagger}(t)$ $(t > 0)$, (5.9) is yielded. ∎

Remark 5.6. We notice that the upper bound in (5.8) cannot be improved by other values of t, because it is the balance of a strictly increasing function $\rho_{x^\dagger}(t)/t$ and a decreasing, with respect to t, function $\varphi(\alpha)/t$.

We also mention that the same arguments yield a slightly different bound

$$\|x^\dagger - x_\alpha\| \le \frac{(\gamma + \gamma_1)}{t} \max\{\rho_{x^\dagger}(t), \varphi(\alpha)\} \quad \text{for all} \quad t > 0 \quad \text{and} \quad 0 < \alpha \le a,$$

which is better if the constants γ and γ_1 differ. This implies that in all estimates below the expression $2\max\{\gamma, \gamma_1\}$ can be replaced by $(\gamma + \gamma_1)$.

Remark 5.7. Since the denominator $\rho^{-1}_{x^\dagger}(\varphi(\alpha))$ in (5.9) expresses an index function tending to zero as α tends to zero, the decay rate of $f(\alpha) \to 0$ as $\alpha \to 0$ is always *lower* than the corresponding rate of the benchmark function φ, i.e., $\varphi(\alpha) = o(f(\alpha))$ as $\alpha \to 0$. In particular, one has to choose a sufficiently good benchmark function and a regularization with high enough qualification to achieve by that way the best possible rate for given $x^\dagger$.

5.3. Approximate source conditions for solutions in sourcewise representation. It is worthwhile to discuss the situation when $x^\dagger$ has a sourcewise representation (5.1), but the benchmark function φ is chosen in such a way that $x^\dagger \notin \mathcal{R}(\varphi(A^*A))$. This can happen in the following case only.

LEMMA 5.8. *Suppose $x^\dagger$ obeys* (5.1). *If $x^\dagger \notin \mathcal{R}(\varphi(A^*A))$, then necessarily $(\varphi/\psi)(t) \to 0$ as $t \to 0$.*

Proof. Suppose $\varphi(t) \ne o(\psi(t))$. Then there is $C < \infty$ such that $\psi(t) \le C\varphi(t)$ for small $0 < t \le \bar{t}$. Given $0 < \varepsilon \le \bar{t}$ we can bound

$$\begin{aligned}
\int_\varepsilon^a \frac{1}{\varphi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 &= \int_\varepsilon^{\bar{t}} \frac{1}{\varphi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 + \int_{\bar{t}}^a \frac{1}{\varphi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 \\
&\le C^2 \int_\varepsilon^{\bar{t}} \frac{1}{\psi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 + \sup_{\lambda \ge \bar{t}} \frac{\psi^2(\lambda)}{\varphi^2(\lambda)} \int_{\bar{t}}^a \frac{1}{\psi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 \\
&\le \max\left\{C^2, \sup_{\lambda \ge \bar{t}} \frac{\psi^2(\lambda)}{\varphi^2(\lambda)}\right\} \int_\varepsilon^a \frac{1}{\psi^2(\lambda)} d\|E_\lambda x^\dagger\|^2 \\
&\le \max\left\{C^2, \sup_{\lambda \ge \bar{t}} \frac{\psi^2(\lambda)}{\varphi^2(\lambda)}\right\} \|x^\dagger\|^2_{X^H_\psi}.
\end{aligned}$$

Letting $\varepsilon \to 0$ we obtain $\|x^\dagger\|_{X^H_\varphi} < \infty$, thus $x^\dagger \in \mathcal{R}(\varphi(A^*A))$, which completes the proof. ∎

If, slightly stronger but geometrically intuitive, we assume that the quotient $(\varphi/\psi)(t)$ is strictly increasing, then we can give a clear picture of the resulting function $\rho_{x^\dagger}(t)$ for $t > 0$ sufficiently small.

THEOREM 5.9. *We suppose that $x^\dagger$ obeys* (5.1) *and that the quotient $(\varphi/\psi)(t)$ is an index function for $0 < t \le a$. Then we can estimate the distance function as*

$$\rho_{x^\dagger}(t) \le \varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1}(Rt)\right) \qquad \text{for all} \quad 0 < t \le \frac{1}{R}\frac{\varphi(a)}{\psi(a)}. \tag{5.10}$$

Proof. The proof is carried out using the analysis from the proof of Lemma 5.3, and we shall make use of the notation introduced there. According to the proof of Lemma 5.3, let $\lambda(t) := \chi^{-1}(1/t)$ $(t > 0)$ with function χ from (5.5). Then for $x^\dagger = \psi(A^*A)v$ with $\|v\| \le R$, representation (5.6) allows for the following bound:

$$\begin{aligned}\rho_{x^\dagger}(t) &= t\lambda(t)\| \left[\varphi^2(A^*A) + \lambda(t)I\right]^{-1} x^\dagger\| \le (Rt)\lambda(t)\| \left[\varphi^2(A^*A) + \lambda(t)I\right]^{-1} \psi(A^*A)\| \\ &= (Rt) \sup_{0<s\le a} \frac{\lambda(t)\psi(s)}{\varphi^2(s) + \lambda(t)} = (Rt) \sup_{0<u\le\varphi^2(a)} \frac{\lambda(t)}{u + \lambda(t)} \psi((\varphi^2)^{-1}(u))\,,\end{aligned}$$

where we make the crucial observation that $u \mapsto \lambda(t)/(u + \lambda(t))$ is the residual of Tikhonov regularization. To continue we introduce the auxiliary function

$$\kappa(s) := \frac{\psi((\varphi^2)^{-1}(s))}{\sqrt{s}} = \left(\frac{\psi}{\varphi}\right)\left((\varphi^2)^{-1}(s)\right) \qquad (0 < s \le \varphi^2(a)). \tag{5.11}$$

It is clear that $1/\kappa(s)$ is an index function, hence $\lim_{s\to 0+} \kappa(s) = \infty$. Also, the function $\kappa(u)/\sqrt{u}$ is decreasing whenever κ is. Hence Proposition 2.7(a) applies and allows us to conclude the estimate

$$\rho_{x^\dagger}(t) \le (Rt)\psi((\varphi^2)^{-1}(\lambda(t))) \qquad (t > 0)\,, \tag{5.12}$$

noting that $\psi((\varphi^2)^{-1}(s))$ for sufficiently small $s > 0$ is an index function.

Next we shall establish for sufficiently small $t > 0$ an upper bound $\tilde\lambda(t)$ for $\lambda(t)$ which then will yield estimate (5.10). Indeed, let $\tilde\lambda(t)$ be obtained as inverse

$$\tilde\lambda(t) = \kappa^{-1}(1/(Rt)). \tag{5.13}$$

It is enough to show that $\lambda(t) \le \tilde\lambda(t)$. To this end notice that κ was decreasing, hence $u \mapsto (\psi((\varphi^2)^{-1}(u))\sqrt{u})/u$ is so, and we derive, again using arguments as above, that for $0 < t \le \frac{1}{R}\frac{\varphi(a)}{\psi(a)}$ the estimate

$$\begin{aligned}\kappa(\tilde\lambda(t)) \le \frac{1}{Rt} = \frac{\chi(\lambda(t))}{R} &\le \| \left[\varphi^2(A^*A) + \lambda(t)I\right]^{-1} \varphi(A^*A)\psi(A^*A)\| \\ &\le \frac{1}{\lambda(t)} \sup_{0<u\le\varphi^2(a)} \frac{\lambda(t)}{u + \lambda(t)} \psi((\varphi^2)^{-1}(u))\sqrt{u} \le \frac{1}{\lambda(t)} \psi((\varphi^2)^{-1}(\lambda(t)))\sqrt{\lambda(t)} \\ &= \kappa(\lambda(t))\,.\end{aligned}$$

Consequently, $\lambda(t) \le \tilde\lambda(t)$ and we arrive at

$$\rho_{x^\dagger}(t) \le (Rt)\psi((\varphi^2)^{-1}(\lambda(t))) \le (Rt)\psi((\varphi^2)^{-1}(\tilde\lambda(t))) = \sqrt{\kappa^{-1}\left(\frac{1}{Rt}\right)}. \tag{5.14}$$

It is a routine matter to check that both versions in the right-hand side of (5.14) are equal. Indeed, starting from the identity $\psi(u)/\varphi(u) = \psi(u)/\varphi(u)$, a variable substitution $u := (\varphi/\psi)^{-1}(Rt)$ yields

$$\frac{1}{Rt} = \frac{\psi\left((\varphi/\psi)^{-1}(Rt)\right)}{\varphi\left((\varphi/\psi)^{-1}(Rt)\right)} = \kappa\left(\varphi^2\left(\left(\frac{\varphi}{\psi}\right)^{-1}(Rt)\right)\right),$$

completing the proof. ∎

COROLLARY 5.10. *Suppose that φ is a qualification for g_α with constant γ. Under the assumptions of Theorem 5.9 there is some $\overline{\alpha} > 0$ such that*

$$f(\alpha) := 2\max\{\gamma, \gamma_1\}\, R\,\psi(\alpha) \qquad (0 < \alpha \le \overline{\alpha}) \tag{5.15}$$

is a profile function for $(H_\psi(R), g_\alpha)$.

Proof. For proving that (5.15) is a profile function for $(H_\psi(R), g_\alpha)$ we use the estimate (5.8) of Theorem 5.5 and the bound (5.10) which together yield for some sufficiently small $\bar{t} > 0$ the error bound

$$\|x_\alpha - x^\dagger\| \le \max\{\gamma, \gamma_1\}\, \frac{1}{t}\left(\varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1}(Rt)\right) + \varphi(\alpha)\right) \quad (0 < t \le \bar{t},\ 0 < \alpha \le a).$$

Then for sufficiently small $\alpha > 0$ there is some $t_* = t_*(\alpha) \in (0, \bar{t}\,]$ satisfying the equation

$\varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1}(Rt_*)\right) = \varphi(\alpha)$, namely $t_* = \varphi(\alpha)/(R\,\psi(\alpha))$ implying

$$\|x_\alpha - x^\dagger\| \le 2\,\max\{\gamma, \gamma_1\}\,\frac{\varphi(\alpha)}{t_*} = 2\,\max\{\gamma, \gamma_1\}\; R\,\psi(\alpha).$$

This, however, completes the proof. ∎

Example 5.11. For monomials $\varphi(t) = t^\nu$ and $\psi(t) = t^\eta$ with $\nu, \eta > 0$, everything can be made explicit. Lemma 5.8 states that (5.3) is valid if and only if $0 < \eta < \nu$, which in the case of monomials is equivalent to saying that $(\varphi/\psi)(t)$ is an index function. We obtain the bound $\rho_{x^\dagger}(t) \le (Rt)^{\nu/(\nu-\eta)}$.

The global properties required for the quotient function φ/ψ on $(0, a]$ are rather strong assumptions in Theorem 5.9 used for obtaining the estimate (5.15) in Corollary 5.10. On the other hand, in [11] and [4] by a completely different technique there have been developed error estimates of type (5.15) with some other constant which only need local properties of φ/ψ on an arbitrarily small interval $(0, \varepsilon]$. In order to show that our approach is powerful enough to work with such weaker assumptions, we conclude this section with a local variant of Theorem 5.9 yielding the results of Corollary 5.10 with different constant under the local assumption on the quotient function.

THEOREM 5.12. *We suppose that $x^\dagger$ obeys (5.1) and that φ, ψ are index functions on $(0, a]$. Moreover, it is assumed that there exists some $0 < \varepsilon \le a$ such that the quotient function φ/ψ is an index function on the interval $(0, \varepsilon]$. Then with the constants $C_\varepsilon = \frac{\psi(a)}{\psi(\varepsilon)} \ge 1$ and $K_\varepsilon = \frac{\psi(a)}{\psi(\varepsilon)}\,\frac{\varphi(a)}{\varphi(\varepsilon)} \ge 1$ we can estimate the distance function as*

$$\rho_{x^\dagger}(t) \le \frac{C_\varepsilon}{K_\varepsilon}\,\varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1}(R\,K_\varepsilon\,t)\right) \qquad \text{for all} \quad 0 < t \le \bar{t} \tag{5.16}$$

and sufficiently small $\bar{t} > 0$. If, moreover, φ is a qualification for g_α with constant γ, then there is $\bar{\alpha} > 0$ such that the function

$$f(\alpha) := 2\max\{\gamma, \gamma_1\}\, K_\varepsilon\, R\,\psi(\alpha) \quad (0 < \alpha \le \bar{\alpha}) \tag{5.17}$$

is a profile function for $(H_\psi(R), g_\alpha)$.

Sketch of a proof. We follow the proof of Theorem 5.9, but the local version of the estimate (5.12) is obtained using Proposition 2.7(c) with $\hat{t} = \varphi^2(\varepsilon)$ as

$$\rho_{x^\dagger}(t) \le (Rt)\, C_\varepsilon\, \psi((\varphi^2)^{-1}(\lambda(t)))$$

for sufficiently small $t > 0$. Moreover, instead of (5.13) in the local variant we have to set

$$\tilde{\lambda}(t) \colon = \kappa^{-1}(1/(R\,K_\varepsilon\, t))\,,$$

which is well-defined for $t \in (0, \bar{t}\,]$ with $\bar{t} > 0$ sufficiently small. Then as in the original proof it can be shown that $\lambda(t) \le \tilde{\lambda}(t)$ for $0 < t \le \bar{t}$ again based on Proposition 2.7(c). Precisely, we have the estimate

$$\begin{aligned}
\kappa(\tilde{\lambda}(t)) &= \frac{1}{RK_\varepsilon t} = \frac{\chi(\lambda(t))}{RK_\varepsilon} \le \frac{1}{K_\varepsilon} \left\| \left[\varphi^2(A^*A) + \lambda(t) I\right]^{-1} \varphi(A^*A)\psi(A^*A) \right\| \\
&= \frac{1}{K_\varepsilon \lambda(t)} \sup_{0<u\le\varphi^2(a)} \frac{\lambda(t)}{u+\lambda(t)} \psi((\varphi^2)^{-1}(u))\sqrt{u} \le \frac{1}{\lambda(t)} \psi((\varphi^2)^{-1}(\lambda(t)))\sqrt{\lambda(t)} \\
&= \kappa(\lambda(t))\,.
\end{aligned}$$

Finally we arrive at

$$\rho_{x^\dagger}(t) \le (Rt)\, C_\varepsilon\, \psi((\varphi^2)^{-1}(\lambda(t))) \le (Rt)\, C_\varepsilon\, \psi((\varphi^2)^{-1}(\tilde{\lambda}(t))) = \frac{C_\varepsilon}{K_\epsilon} \sqrt{\kappa^{-1}\left(\frac{1}{RK_\varepsilon t}\right)}$$

which proves (5.16). For proving (5.17) we use the estimate (5.8) of Theorem 5.5 yielding here for sufficiently small $t > 0$ and $\alpha > 0$, and since $\frac{C_\varepsilon}{K_\varepsilon} \le 1$,

$$\|x_\alpha - x^\dagger\| \le \max\{\gamma, \gamma_1\}\, \frac{1}{t} \left(\varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1} (R\,K_\varepsilon\, t) \right) + \varphi(\alpha) \right).$$

Now we choose $t_* = t_*(\alpha)$ such that the equation

$$\varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1} (R\,K_\varepsilon\, t_*) \right) = \varphi(\alpha)$$

holds. This is possible for sufficiently small $\alpha > 0$ and yields $t_* = \frac{\varphi(\alpha)}{\psi(\alpha)}\, \frac{1}{R\,K_\varepsilon}$. Hence we obtain the profile function (5.17) as required. ∎

6. Linking scales by range inclusions. Since the initial study of linear inverse problems in Hilbert scales (see [20]), it is well known that the operator G measuring smoothness of the solution $x^\dagger$ must be linked to the operator A governing (1.1) in order to obtain error bounds. There are various ways to establish such a link and we will investigate its impact on profile functions next.

Again we start with the benchmark function φ and assume in addition that

$$x^\dagger \in G_\tau(R) \tag{6.1}$$

with $G_\tau(R)$ defined by (2.2). Moreover, we impose the following *link condition*; precisely, that there is an index function $\sigma(t)$ $(0 < t \le \|G\|)$ and a constant $C < \infty$ such that

$$\|\sigma(G)v\| \le C\|\varphi(A^*A)v\| \qquad \text{for all} \quad v \in X. \tag{6.2}$$

Remark 6.1. There is an extensive analysis in [3] of linking conditions in various ways. In particular, it is shown as Proposition 2.1 in [3] that the validity of condition (6.2) with some positive C is equivalent to the *range inclusion*

$$\mathcal{R}(\sigma(G)) \subseteq \mathcal{R}(\varphi(A^*A))\,. \tag{6.3}$$

We mention the following consequence of (6.2) (see, e.g., [17]). Given Hilbert spaces X and Z with $Z \subset X$ let $J\colon Z \to X$ be the canonical embedding, leaving elements from Z invariant.

LEMMA 6.2. *Under* (6.2) *the canonical embedding* $J_{G,\sigma}^{H,\varphi}\colon X_\sigma^G \to X_\varphi^H$ *is norm bounded by* C *and we have*

$$G_\sigma(R) \subseteq H_\varphi(C\,R). \tag{6.4}$$

Proof. It is well known that for any pair S, T of operators a relation $\|Sv\| \le \|Tv\|$ implies $\|T^{-1}v\| \le \|S^{-1}v\|$, whenever the right-hand sides are finite. Thus (6.2) implies for any $x \in X_\sigma^G$ with $\|x\|_{X_\sigma^G} \le 1$ that $\|x\|_{X_\varphi^H} \le C$ and hence (6.4). ∎

We will distinguish two scenarios and start with the easier one.

PROPOSITION 6.3. *Let* g_α *be a regularization which has qualification* φ *with constant* γ *and assume that* $x^\dagger$ *obeys* (6.1). *If condition* (6.2) *is valid for an index function* σ, *and if there is* $K < \infty$ *such that* $\tau(t)/\sigma(t) \le K$ $(0 < t \le \|G\|)$, *then the function*

$$f(\alpha) := \gamma\,C\,K\,R\,\varphi(\alpha) \qquad (0 < \alpha \le a) \tag{6.5}$$

is a profile function for $(G_\tau(R), g_\alpha)$.

Proof. From $\tau(t)/\sigma(t) \le K$ $(0 < t \le a)$ we deduce from Lemma 6.2 that $G_\tau(R) \subseteq G_\sigma(KR)$, which is equivalent to $\|\tau(G)x\| \le K\,\|\sigma(G)x\|$ for all $x \in X$. Furthermore, in the light of Lemma 6.2, the link condition (6.2) implies $G_\sigma(KR) \subseteq H_\varphi(CKR)$, and any profile function for $H_\varphi(CKR)$ provides us with a profile function for $G_\tau(R)$, such that the proof can be completed using Proposition 5.1. ∎

Thus we are left with the case when

$$(\sigma/\tau)(t) \to 0 \quad \text{as} \quad t \to 0. \tag{6.6}$$

Then we have $X_\sigma^G \subset X_\tau^G$ and the canonical embedding $J_{G,\sigma}^{G,\tau}\colon X_\sigma^G \to X_\tau^G$ is norm bounded. The question is whether one can use condition (6.2) to draw conclusions for the behavior of profile functions in this case.

Suppose we assume a linking condition (6.3), but smoothness is measured as $x^\dagger \in G_\tau(R)$ with respect to a different index function τ. Can we establish an index function ψ, assigned to a triplet (σ, τ, φ), such that the following range implication holds true:

$$\mathcal{R}(\sigma(G)) \subseteq \mathcal{R}(\varphi(H)) \quad \Longrightarrow \quad \mathcal{R}(\tau(G)) \subseteq \mathcal{R}(\psi(H))? \tag{6.7}$$

In specific situations this problem was already posed (cf. [11, formula (5.10), p. 815]) and partially answered previously (cf. [3, Corollary 2.3]). Most prominently, the Heinz–Kato inequality (see [5, Proposition 8.21]) yields

$$\mathcal{R}(G) \subseteq \mathcal{R}(H) \quad \Longrightarrow \quad \mathcal{R}(G^\mu) \subseteq \mathcal{R}(H^\mu)$$

for $0 < \mu \le 1$ as a consequence of operator monotonicity. In fact this can be extended to more general situations in which operator monotone functions occur. It is convenient to draw the following diagram.

$$\begin{array}{ccccc} G{:}X_\sigma^G & \xrightarrow{J_{G,\sigma}^{G,\tau}} & X_\tau^G & \xrightarrow{J_{G,\tau}^{I}} & X \\ \Big\downarrow J_{G,\sigma}^{H,\varphi} & & \Big\downarrow J_{G,\tau}^{H,\psi} & & \Big\downarrow I \\ H{:}X_\varphi^H & \xrightarrow{J_{H,\varphi}^{H,\psi}} & X_{\psi?}^H & \xrightarrow{J_{H,\psi}^{I}} & X \end{array} \tag{6.8}$$

Under (6.6) the upper row shows embeddings which are bounded. Using Lemma 6.2 the embedding $J_{G,\sigma}^{H,\varphi}$ is norm bounded by C, provided the link condition (6.2) holds true. Plainly the identity $I\colon X \to X$ has norm equal to one. The question addressed in this diagram is whether we can describe an index function ψ such that the corresponding embedding $J_{G,\tau}^{H,\psi}$ is norm bounded, say by some constant $L < \infty$. Diagram (6.8) also suggests that the resulting function ψ will describe smoothness not covered by φ, and approximate source conditions must be used to obtain results.

Remark 6.4. If the embedding $J_{G,\tau}^{H,\psi}$ were norm bounded, say by some constant $L < \infty$, then $G_\tau(R) \subseteq H_\psi(LR)$, and any profile function for $(H_\psi(LR), g_\alpha)$ would also be a profile function for $(G_\tau(R), g_\alpha)$.

As the diagram (6.8) clearly indicates, interpolation properties may help to find suitable index function ψ. The implication (6.7) of range inclusions is indeed true if operator monotonicity occurs and we shall mention the following result from [17].

THEOREM 6.5. *Let $x^\dagger$ obey* (6.1). *We assume that G and A^*A are linked by* (6.2), *where we suppose that σ is such that there is an extension $\sigma(t)$ $(0 < t \le b)$ with $\sigma(b) \ge \varphi(a)$ and this extension is an index function. Moreover, given an index function $\tau(t)$ $(0 < t \le \|G\|)$ we assign the index function*

$$\psi(t) := \tau(\sigma^{-1}(\varphi(t))) \quad (0 < t \le a). \tag{6.9}$$

Then the implication (6.7) *is satisfied whenever the function $\tau^2((\sigma^2)^{-1}(t))$ $(0 < t \le \varphi^2(a))$ is operator monotone and $(\sigma/\tau)(t)$ $(0 < t \le \|G\|)$ is an index function.*

Precisely, the norm bound

$$\|J_{G,\tau}^{H,\psi}\colon X_\tau^G \to X_\psi^H\| \le \max\{1, C\} \tag{6.10}$$

with C from (6.2) *is valid.*

Now we return to the analysis of profile functions. To establish these functions the full strength of the implication (6.7) is not necessary. But we shall also indicate its strength in Theorem 6.11. However, the function ψ from (6.9) will occur, nonetheless.

There are in principle two ways to use the link conditions (6.2) or (6.3), respectively, to obtain profile functions. One can either transfer all information to the scale generated by the operator G or to the scale generated by $H := A^*A$. Both ways finally provide the same asymptotic results but under assumptions of different strength. We start with the first approach which requires weaker assumptions.

LEMMA 6.6. *The link condition* (6.2) *implies*

$$\rho_{x^\dagger}^{(H,\varphi)}(t) \le \frac{1}{C}\rho_{x^\dagger}^{(G,\sigma)}(Ct) \qquad \textit{for all} \quad 0 < t < \infty. \tag{6.11}$$

Proof. Plainly, condition (6.2) yields $G_\sigma(1/C) \subseteq H_\varphi(1)$ and we obtain

$$\begin{aligned}\rho_{x^\dagger}^{(H,\varphi)}(t) &= \operatorname{dist}(tx^\dagger, H_\varphi(1)) \le \operatorname{dist}(tx^\dagger, G_\sigma(1/C)) \\ &= \frac{1}{C}\operatorname{dist}(Ctx^\dagger, G_\sigma(1)) = \frac{1}{C}\rho_{x^\dagger}^{(G,\sigma)}(Ct). \qquad \square\end{aligned}$$

With this preparation we can state the main result of this section.

THEOREM 6.7. *Let g_α be a regularization method with qualification φ and constant γ for the operator equation* (1.1) *with solution $x^\dagger$ the smoothness of which is characterized by the conditions* (5.3) *and* (6.1) *with some index functions φ and τ. We suppose the link condition* (6.2) *with some index function σ for connecting A^*A and G. If the function*

$$(\sigma/\tau)(t) \ \ (0 < t \le \|G\|) \quad \text{is an index function}, \tag{6.12}$$

then there is some $\bar\alpha > 0$ for which the function $\psi(t)$ $(0 < t \le \bar\alpha)$ from (6.9) *is an index function and*

$$f(\alpha) := 2\,\max\{\gamma, \gamma_1\}\,\max\{1, C\}R\,\psi(\alpha) \quad (0 < \alpha \le \bar\alpha) \tag{6.13}$$

is a profile function for $(G_\tau(R), g_\alpha)$.

Remark 6.8. Assume (6.3) instead of (6.2). Let $C := \|(\varphi(A^*A))^{-1}\tau(G)\| < \infty$. Then the function f from (6.13) is a profile function with the constant C.

Proof of Theorem 6.7. For an arbitrary $x^\dagger \in G_\tau(R)$ using the bound (5.8) and Lemma 6.6 we obtain for all $0 < \alpha \le a$ that

$$\|x_\alpha - x^\dagger\| \le \frac{\max\{\gamma, \gamma_1\}}{t}\left(\rho_{x^\dagger}^{(H,\varphi)}(t) + \varphi(\alpha)\right) \le \max\{\gamma, \gamma_1\}\frac{1}{t}\left(\frac{1}{C}\rho_{x^\dagger}^{(G,\sigma)}(Ct) + \varphi(\alpha)\right).$$

By exploiting Theorem 5.9 in the scale generated by G we can continue and bound

$$\|x_\alpha - x^\dagger\| \le \frac{\max\{\gamma, \gamma_1\}}{t}\left(\frac{1}{C}\sigma\left(\left(\frac{\sigma}{\tau}\right)^{-1}(RCt)\right) + \varphi(\alpha)\right) \tag{6.14}$$

for $0 < \alpha \le \frac{1}{RC}\left(\frac{\sigma}{\tau}\right)(\|G\|)$. There is some $0 < \bar\alpha \le \|G\|/C$ for which we can equate both summands on the right of formula (6.14) whenever $0 < \alpha \le \bar\alpha$. This leads to

$$t_* = t_*(\alpha) := \frac{1}{R}\frac{\varphi(\alpha)}{\tau(\sigma^{-1}(C\varphi(\alpha)))} \quad (0 < \alpha \le \bar\alpha).$$

Moreover, by (6.12) we have that $\tau(\sigma^{-1}(Ct)) \le \max\{1, C\}\,\tau(\sigma^{-1}(t))$ for $0 < t \le \overline{\alpha}$. Thus we can estimate for $0 < \alpha \le \bar\alpha$

$$\|x_\alpha - x^\dagger\| \le 2\,\max\{\gamma, \gamma_1\}\,\frac{\varphi(\alpha)}{t_*} \le 2\,\max\{\gamma, \gamma_1\}\,R\,\tau(\sigma^{-1}(C\varphi(\alpha)))\,.$$

Consequently, we obtain

$$\|x_\alpha - x^\dagger\| \le 2\,\max\{\gamma, \gamma_1\}\,\max\{1, C\}\;R\,\psi(\alpha) \qquad (0 < \alpha \le \bar\alpha),$$

completing the proof. $\square$

Remark 6.9. The results of Theorem 6.7 with an appropriately modified constant in (6.13) can also be obtained under the weaker assumption that

$$(\sigma/\tau)(t) \ \ (0 < t \le \varepsilon) \quad \text{is an index function}$$

for arbitrarily small $\varepsilon > 0$ instead of the global assumption (6.12). This is an immediate result of the opportunity of localization as outlined in Theorem 5.12 and its proof.

As mentioned above, we can also try to transfer the information from the link conditions (6.2) and (6.3) to the scale generated by $H = A^*A$.

We recall the definition of the function ψ in formula (6.9) in the context of Theorem 6.5. The following observation is useful.

LEMMA 6.10. *Let the functions τ, σ and φ, ψ be as in Theorem* 6.5. *If the quotient σ/τ is an index function on $(0, \|G\|]$, then φ/ψ is an index function on $(0, a]$.*

Proof. We assign $s = s(t) := \sigma^{-1}(\varphi(t))$ $(0 < t \le a)$, thus $s \in (0, b]$. With this identification we obtain

$$\frac{\varphi(t)}{\psi(t)} = \frac{\sigma(s)}{\psi(\varphi^{-1}(\sigma(s)))} = \frac{\sigma(s)}{\tau(\sigma^{-1}(\sigma(s)))} = \frac{\sigma(s)}{\tau(s)}. \qquad \square$$

Keeping this lemma in mind we can prove the following counterpart of Theorem 6.7.

THEOREM 6.11. *Assume that the regularization g_α has qualification φ with constant γ and that σ/τ is an index function on $(0, \|G\|]$. Under the assumptions of Theorem* 6.5*, in particular the operator monotonicity of the function $\tau^2((\sigma^2)^{-1}(t))$ $(0 < t \le \varphi^2(a))$, the function*

$$f(\alpha) = 2 \max\{\gamma, \gamma_1\} \max\{1, C\}\ R\psi(\alpha) \qquad (0 < \alpha \le a) \tag{6.15}$$

is a profile function for $(G_\tau(R), g_\alpha)$.

Proof. Let $L := \max\{1, C\}$. The estimate (6.10) of Theorem 6.5 yields the inclusion $G_\tau(R) \subset H_\psi(LR)$. Thus profile functions for $(H_\psi(LR), g_\alpha)$ are also profile functions for $(G_\tau(R), g_\alpha)$. By Lemma 6.10 the function $\varphi(t)/\psi(t)$ $(0 < t \le a)$ is an index function and we can apply Theorem 5.9 to bound the distance function $\rho_{x^\dagger}^{(H,\psi)}$ as

$$\rho_{x^\dagger}^{(H,\psi)}(t) \le \varphi\left(\left(\frac{\varphi}{\psi}\right)^{-1}(LRt)\right) \qquad \left(0 < t \le \frac{\varphi(a)}{LR\psi(a)}\right).$$

Corollary 5.10 provides us with the profile function as given in (6.15). $\square$

Example 6.12. Again, let us discuss the situation when the index functions are in the form of monomials; more precisely, we assume that $\sigma(t) = t^\mu,\ \tau(t) = t$. Then the operator monotonicity as required in Theorem 6.11 is fulfilled whenever $\mu \ge 1$, which can be deduced from the Heinz–Kato inequality. If the link condition (6.2) is assumed to hold for $\varphi(t) = t^\nu$, and if the regularization has qualification φ, then we arrive at a profile function $f(\alpha) = C\alpha^{\nu/\mu}$, uniformly for $x^\dagger$ satisfying (5.3) and (6.1).

REFERENCES

[1] A. B. BAKUSHINSKY AND M. YU. KOKURIN, *Iterative Methods for Approximate Solutions of Inverse Problems*, Springer, Dordrecht, The Netherlands, 2004.

[2] J. BAUMEISTER, *Stable Solution of Inverse Problems*, Vieweg, Braunschweig, Germany, 1987.

[3] A. BÖTTCHER, B. HOFMANN, U. TAUTENHAHN, AND M. YAMAMOTO, *Convergence rates for Tikhonov regularization from different kinds of smoothness conditions*, Appl. Anal., 85 (2006), pp. 555–578.

[4] D. DÜVELMEYER, B. HOFMANN, AND M. YAMAMOTO, *Range Inclusions and Approximate Source Conditions with General Benchmark Functions*, submitted, Preprint 2007-02, Technische Universität Chemnitz, Chemnitz, Germany, 2007.

[5] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*. Kluwer, Dordrecht, The Netherlands, 1996.
[6] M. Hegland, *An optimal order regularization method which does not use additional smoothness assumptions*, SIAM J. Numer. Anal., 29 (1992), pp. 1446–1461.
[7] M. Hegland, *Variable Hilbert scales and their interpolation inequalities with applications to Tikhonov regularization*, Appl. Anal., 59 (1995), pp. 207–223.
[8] B. Hofmann, *Approximate source conditions in Tikhonov–Phillips regularization and consequences for inverse problems with multiplication operators*, Math. Methods Appl. Sci., 29 (2006), pp. 351–371.
[9] B. Hofmann, D. Düvelmeyer, and K. Krumbiegel, *Approximate source conditions in Tikhonov regularization—New analytical results and numerical studies*, Math. Model. Anal., 11 (2006), pp. 41–56.
[10] B. Hofmann and G. Fleischer, *Stability rates for linear ill-posed problems with compact and noncompact operators*, Z. Anal. Anwendungen, 18 (1999), pp. 267–286.
[11] B. Hofmann and M. Yamamoto, *Convergence rates for Tikhonov regularization based on range inclusions*, Inverse Problems, 21 (2005), pp. 805–820.
[12] T. Hohage, *Regularization of exponentially ill-posed problems*, Numer. Funct. Anal. Optim., 21 (2000), pp. 439–464.
[13] P. Mathé, *Saturation of regularization methods for linear ill-posed problems in Hilbert spaces*, SIAM J. Numer. Anal., 42 (2004), pp. 968–973.
[14] P. Mathé and S. V. Pereverzev, *Geometry of linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.
[15] P. Mathé and S. V. Pereverzev, *Discretization strategy for linear ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 1263–1277.
[16] P. Mathé and S. V. Pereverzev, *Regularization of some linear ill-posed problems with discretized random noisy data*, Math. Comp., 75 (2006), pp. 1913–1929.
[17] P. Mathé and U. Tautenhahn, *Interpolation in variable Hilbert scales with application to inverse problems*, Inverse Problems, 22 (2006), pp. 2271–2297.
[18] M. T. Nair, S. V. Pereverzev, and U. Tautenhahn, *Regularization in Hilbert scales under general smoothing conditions*, Inverse Problems, 21 (2005), pp. 1851–1869.
[19] M. Z. Nashed, *A new approach to classification and regularization of ill-posed operator equations*, in Inverse and Ill-posed Problems, H. W. Engl and C. W. Groetsch, eds., Academic Press, Boston, 1987, pp. 53–75.
[20] F. Natterer, *Error bounds for Tikhonov regularization in Hilbert scales*, Applicable Anal., 18 (1984), pp. 29–37.
[21] A. Neubauer, *On converse and saturation results for regularization methods*, in Beiträge zur Angewandten Analysis und Informatik, E. Schock, ed., Shaker, Aachen, Germany, 1994, pp. 262–270.
[22] A. Neubauer, *On converse and saturation results for Tikhonov regularization of linear ill-posed problems*, SIAM J. Numer. Anal., 34 (1997), pp. 517–527.
[23] E. Schock, *Approximate solution of ill-posed equations: Arbitrarily slow convergence vs. superconvergence*, in Constructive Methods for the Practical Treatment of Integral Equations, G. Hämmerlin and K. H. Hoffmann, eds., Birkhäuser, Basel, 1985, pp. 234–243.
[24] U. Tautenhahn, *Optimality for ill-posed problems under general source conditions*, Numer. Funct. Anal. Optim., 19 (1998), pp. 377–398.
[25] G. Vainikko, *Solution Methods for Linear Ill-Posed Problems in Hilbert Spaces*, Tartu State University, Tartu, Estonia, 1982 (in Russian).
[26] G. M. Vainikko and A. Y. Veretennikov, *Iteration Procedures in Ill-Posed Problems*, Nauka, Moscow, 1986 (in Russian).

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1142–1174

A DISCRETE DUALITY FINITE VOLUME APPROACH TO HODGE DECOMPOSITION AND DIV-CURL PROBLEMS ON ALMOST ARBITRARY TWO-DIMENSIONAL MESHES*

SARAH DELCOURTE†, KOMLA DOMELEVO‡, AND PASCAL OMNES†

Abstract. We define discrete differential operators such as gradient, divergence, and curl, on general two-dimensional nonorthogonal meshes. These discrete operators verify discrete analogues of usual continuous theorems: discrete Green formulae, discrete Hodge decomposition of vector fields, and vector curls have a vanishing divergence and gradients have a vanishing curl. We apply these ideas to discretize div-curl systems. We give error estimates based on the reformulation of these systems into equivalent equations for the potentials. Numerical results illustrate the use of the method on several types of meshes, some of which are degenerating triangular meshes and nonconforming locally refined meshes.

Key words. discrete duality finite volume method, discrete Green formula, discrete Hodge decomposition, discrete differential operators, div-curl equations, arbitrary meshes, nonconforming meshes, degenerating meshes, convergence, error estimates

AMS subject classifications. 35Q60, 65N12, 65N15, 65N30, 78A30

DOI. 10.1137/060655031

1. Introduction.

Discretization schemes which are based on a discrete vector analysis satisfying discrete analogues of the usual continuous theorems lead to robust and efficient approximations of various physical models. Based on finite volume-like formulations, they provide discrete approximations of differential operators such as gradient, divergence, and curl.

Such schemes were, for example, constructed by Hyman, Shashkov, and co-workers, initially on logically rectangular grids. We refer to [13, 14] for the construction of the discrete operators and to [15] for the proof of a discrete Hodge decomposition. These schemes were then applied in several different circumstances (see, e.g., [16, 17]) and extended to unstructured [5] or even nonconforming grids [19], although on those types of meshes, to our knowledge, no discrete Hodge decomposition has been proved.

Our interests in this paper are related to other schemes based on a discrete vector analysis. These schemes were proposed by Nicolaides and co-workers to solve fluid mechanics problems [7], div-curl problems [20, 12], or Maxwell equations [21]. In these works, these so-called covolume schemes are restricted to locally equiangular triangular meshes in the two-dimensional case. Given such a primal triangular mesh, a dual mesh is constructed by joining the circumcenters of adjacent triangles. Thus the edges of the primal and dual meshes are orthogonal. In what follows, this property will be called "the orthogonality property." The necessity for the mesh to verify this property might in certain cases be a severe restriction, particularly with respect to mesh adaptivity.

*Received by the editors March 24, 2006; accepted for publication (in revised form) November 21, 2006; published electronically May 10, 2007.
http://www.siam.org/journals/sinum/45-3/65503.html

†Commissariat à l'Énergie Atomique de Saclay, DEN-DM2S-SFME, 91191 Gif-sur-Yvette Cedex, France (sarah.delcourte@cea.fr, pascal.omnes@cea.fr).

‡Mathématiques pour l'Industrie et la Physique, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex 4, France (komla@mip.ups-tlse.fr).

In [20], discrete field components are defined as normal to the edges of the primal mesh and, therefore, thanks to the orthogonality property, along the edges of the dual mesh. This single component is enough to permit the definition of a discrete divergence operator on the primal mesh and of a discrete curl operator on the dual mesh. Reciprocally, discrete analogues of the normal (with respect to the edges of the primal mesh) components of the gradients (resp., vector curls) are obtained over the edges with the help of scalar quantities defined at the circumcenters (resp., at the vertices) of the primal cells.

Due to the anisotropy of the media considered in [12], the authors are led to introduce both components of vector fields on the edges of the mesh, which allows them to define discrete divergence and curl operators on both the primal and dual meshes. Nevertheless, they keep on considering only the normal components of the discrete gradient and curl vectors, thus leaving the generalization of [20] incomplete.

In the present work, we extend the covolume ideas of Nicolaides to almost arbitrary two-dimensional meshes, including, in particular, nonconforming meshes. The only requirement on the mesh is that the dual cells (which are obtained in a different way) form a partition of the domain of computation. These meshes do not necessarily verify the orthogonality property, and we therefore discretize vector fields by their two components over so-called diamond-cells which are quadrilaterals whose vertices are the extremities of primal and associated dual edges. Like in [12], these two field components enable us to define discrete divergence and curl operators both on the primal and dual meshes. Reciprocally, and in contrast to [12], both components of discrete gradient and vector curl operators are defined over the diamond-cells with the help of scalar quantities given on both the primal and dual cells. Together with the definition of appropriate discrete scalar products, we establish that these discrete operators verify discrete properties which are analogous to those verified by their continuous counterparts: discrete Green formulae, discrete Hodge decomposition of vector fields, the divergence of vector curls, and the curl of gradients vanish. These results thus generalize those obtained in [12, 20], with the major novelty that they hold on a much wider class of meshes. Because of the discrete Green formulae, finite volume schemes based on these ideas have been named discrete duality finite volume (DDFV) schemes in [9] and their use has started with the construction and analysis of a finite volume method for the Laplace equation on almost arbitrary two-dimensional meshes [10]. Then, these ideas have been applied to the discretization of nonlinear elliptic equations [2], drift-diffusion and energy-transport models [6], and electro-cardiology problems [22].

In this article, we apply these ideas to the numerical solution of div-curl problems which occur, for example, in fluid dynamics and electro- and magnetostatics. Using the discrete Hodge decomposition of the discrete unknown vector field, this problem is recast into two discrete Laplace equations for the discrete potentials, just like in the continuous problem. Using results obtained in [10], we prove the convergence of the scheme provided the continuous potentials are smooth enough and under geometrical hypotheses related to the nondegeneracy of the diamond-cells.

This paper is organized as follows: in section 2, we explain the construction of the primal, dual, and diamond meshes and define our notations. In section 3 we construct the discrete differential operators, while section 4 is devoted to the proof of the properties of the discrete operators. Then, we apply these ideas in section 5 to discretize the div-curl problem and obtain error estimates. Several numerical experiments are reported in section 6 and conclusions are drawn in section 7.

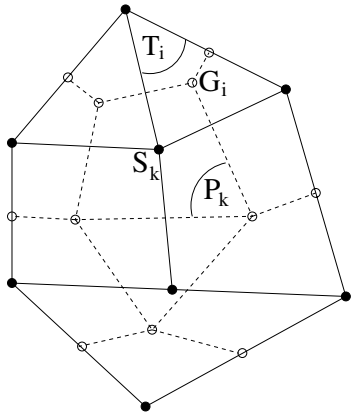


FIG. 2.1. *An example of a primal mesh and its associated dual mesh.*

2. Definitions and notations. Let Ω be a bounded polygon of $\mathbb{R}^2$, not necessarily simply connected, whose boundary is denoted by Γ. We suppose, in addition, that the domain has Q holes. Throughout the paper, we shall assume that $Q > 0$, but the results also hold for the case $Q = 0$.

Let Γ_0 denote the exterior boundary of Ω and let Γ_q, with $q \in [1, Q]$, be the interior polygonal boundaries of Ω, so that $\Gamma = \Gamma_0 \bigcup_{q\in[1,Q]} \Gamma_q$.

The domain Ω will be covered by three different meshes whose constructions are similar to those given in [10].

2.1. Construction of the primal mesh. We consider a first partition of Ω (named primal mesh) composed of elements T_i, with $i \in [1, I]$, which are supposed to be convex polygons. For each element T_i of the mesh, we associate a node G_i located inside T_i. This point may be the barycenter of T_i, but is not necessarily. The area of T_i is denoted by $|T_i|$. We shall denote by J the total number of edges of this mesh. Note that in the case of a nonconforming mesh, an edge is any segment whose extremities are nodes of the mesh. We also denote by J^Γ the number of edges which are located on the boundary Γ and we associate with each of these boundary edges its midpoint, also denoted by G_i with $i \in [I + 1, I + J^\Gamma]$. By a slight abuse of notations, we shall write $i \in \Gamma_q$ iff $G_i \in \Gamma_q$.

2.2. Construction of the dual mesh. We denote by S_k, with $k \in [1, K]$, the nodes of the polygons of the primal mesh. For each of these points, we associate a polygon denoted by P_k, obtained by joining the points G_i associated with the elements of the primal mesh (and possibly to the boundary edges) of which S_k is a node. The area of P_k is denoted by $|P_k|$. In what follows, we shall only consider the cases where the P_ks constitute a second partition of Ω, which we name dual mesh.[1] Figure 2.1 displays an example of a nonconforming primal mesh and its associated dual mesh.

Moreover, we suppose that the set $[1, K]$ is ordered so that when S_k is not on Γ, then $k \in [1, K - J^\Gamma]$, and when S_k is on Γ, then $k \in [K - J^\Gamma + 1, K]$. We shall also write $k \in \Gamma_q$ iff $S_k \in \Gamma_q$.

2.3. Construction of the diamond mesh. With each edge of the primal mesh, denoted by A_j (whose length is $|A_j|$), with $j \in [1, J]$, we associate a quadrilateral

[1] It may happen that the P_ks overlap, as seen on Figure 2 of [10].

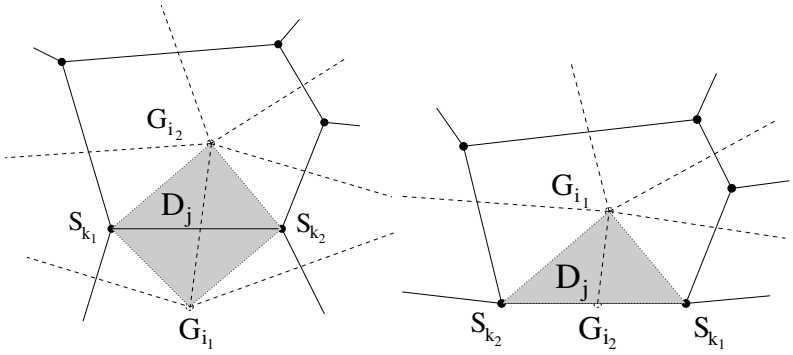


FIG. 2.2. *Examples of diamond-cells.*

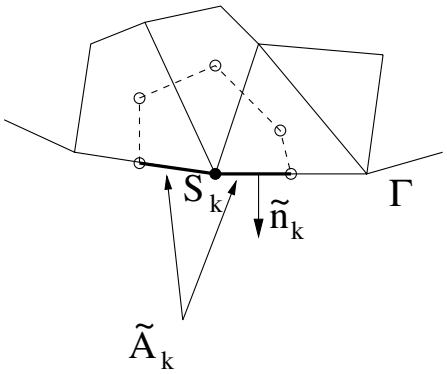


FIG. 2.3. *Definition of $\tilde{A}_k$ and $\tilde{\mathbf{n}}_k$ for the boundary nodes.*

named "diamond-cell" denoted by D_j. When A_j is not on the boundary, this cell is obtained by joining the points $S_{k_1(j)}$ and $S_{k_2(j)}$, which are the two nodes of A_j, with the points $G_{i_1(j)}$ and $G_{i_2(j)}$ associated with the elements of the primal mesh which share this edge. When A_j is on the boundary Γ, the cell D_j is obtained by joining the two nodes of A_j with the point $G_{i_1(j)}$ associated with the only element of the primal mesh of which A_j is an edge and to the point $G_{i_2(j)}$ associated with A_j (i.e., by convention, $i_2(j)$ is an element of $[I+1, I+J^\Gamma]$ when A_j is located on Γ). The cells D_j constitute a third partition of Ω, which we name "diamond mesh." The area of the cell D_j is denoted by $|D_j|$. Such cells are displayed on Figure 2.2.

Moreover, we suppose that the set $[1, J]$ is ordered so that when A_j is not on Γ, then $j \in [1, J-J^\Gamma]$, and when A_j is on Γ, then $j \in [J-J^\Gamma+1, J]$. We shall also write $j \in \Gamma_q$ iff $A_j \subset \Gamma_q$.

2.4. Definitions of geometrical elements. The unit vector normal to A_j is denoted by $\mathbf{n}_j$ and is oriented so that $\mathbf{G_{i_1(j)}G_{i_2(j)}} \cdot \mathbf{n}_j \geq 0$. We further denote by A'_j the segment $[G_{i_1(j)}G_{i_2(j)}]$ (whose length is $|A'_j|$) and by $\mathbf{n}'_j$ the unit vector normal to A'_j oriented so that $\mathbf{S_{k_1(j)}S_{k_2(j)}} \cdot \mathbf{n}'_j \geq 0$.

When $S_k \in \Gamma$ ($k \in [K-J^\Gamma+1, K]$), we define $\tilde{A}_k$ as the part of the boundary Γ which consists of the union of the halves of the two segments A_j located on Γ, of which S_k is a node and by $\tilde{\mathbf{n}}_k$ the exterior unit normal vector to $\tilde{A}_k$ (see Figure 2.3). We denote by $M_{i_\alpha(j)\,k_\beta(j)}$ the midpoint of the segment $[G_{i_\alpha(j)}S_{k_\beta(j)}]$, for each pair of

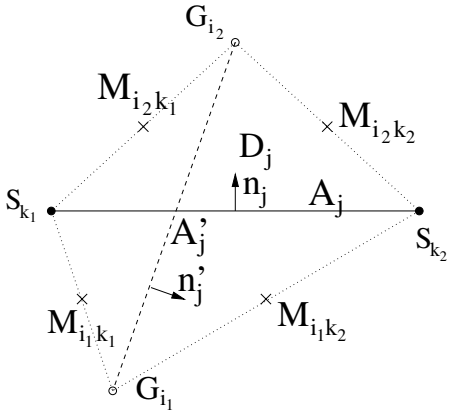


FIG. 2.4. *Notations for the diamond-cell.*

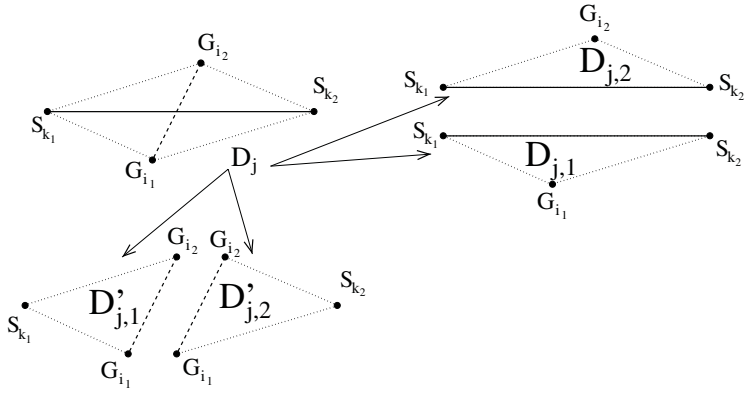


FIG. 2.5. *A diamond-cell may be split into two triangles in two distinct ways.*

integers (α, β) in $\{1;2\}^2$ (see Figure 2.4). We define for each $i \in [1, I]$ the set $\mathcal{V}(i)$ of integers $j \in [1, J]$ such that A_j is an edge of T_i and for each $k \in [1, K]$ the set $\mathcal{E}(k)$ of integers $j \in [1, J]$ such that S_k is a node of A_j.

We define for each $j \in [1, J]$ and each k such that $j \in \mathcal{E}(k)$ (resp., each i such that $j \in \mathcal{V}(i)$) the real-valued number s'_{jk} (resp., s_{ji}) whose value is $+1$ or -1 whether $\mathbf{n}'_j$ (resp., $\mathbf{n}_j$) points outwards or inwards P_k (resp., T_i). We define $\mathbf{n}'_{jk} := s'_{jk}\mathbf{n}'_j$ (resp., $\mathbf{n}_{ji} := s_{ji}\mathbf{n}_j$) and remark that $\mathbf{n}'_{jk}$ (resp., $\mathbf{n}_{ji}$) always points outwards P_k (resp., T_i).

For $j \in [1, J - J^\Gamma]$, as indicated on Figure 2.5, we also denote by $D_{j,1}$ and $D_{j,2}$, the triangles $S_{k_1(j)}G_{i_1(j)}S_{k_2(j)}$ and $S_{k_2(j)}G_{i_2(j)}S_{k_1(j)}$). In the same way, we denote by $D'_{j,1}$ and $D'_{j,2}$, the triangles $G_{i_2(j)}S_{k_1(j)}G_{i_1(j)}$ and $G_{i_1(j)}S_{k_2(j)}G_{i_2(j)}$.

The characteristic functions of the cells T_i and P_k will be denoted by θ_i^T and θ_k^P.

2.5. Definitions of discrete and continuous scalar products and norms. As will be seen in what follows, we shall associate with each point G_i ($i \in [1, I + J^\Gamma]$) and each vertex S_k ($k \in [1, K]$) discrete values. This leads us to the definition of the

following discrete scalar product for all $(\phi,\psi)=\big((\phi_i^T,\phi_k^P),(\psi_i^T,\psi_k^P)\big)\in\big(\mathbb{R}^I\times\mathbb{R}^K\big)^2$:

$$(\phi,\psi)_{T,P}:=\frac{1}{2}\left(\sum_{i\in[1,I]}|T_i|\,\phi_i^T\,\psi_i^T+\sum_{k\in[1,K]}|P_k|\,\phi_k^P\,\psi_k^P\right). \tag{2.1}$$

In the same way, we define a discrete scalar product on the diamond mesh for all $(\mathbf{u},\mathbf{v})=((\mathbf{u}_j),(\mathbf{v}_j))\in\big(\mathbb{R}^2\big)^J\times\big(\mathbb{R}^2\big)^J$

$$(\mathbf{u},\mathbf{v})_D:=\sum_{j\in[1,J]}|D_j|\,\mathbf{u}_j\cdot\mathbf{v}_j \tag{2.2}$$

and a discrete scalar product for the traces of $u\in\mathbb{R}^J$ and $\phi\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K$ on the boundaries Γ_q

$$(u,\phi)_{\Gamma_q,h}:=\sum_{j\in\Gamma_q}|A_j|\,u_j\times\frac{1}{4}\left(\phi^P_{k_1(j)}+2\phi^T_{i_2(j)}+\phi^P_{k_2(j)}\right)$$

and on Γ

$$(u,\phi)_{\Gamma,h}:=\sum_{q\in[0,Q]}(u,\phi)_{\Gamma_q,h}. \tag{2.3}$$

Further, for any $\phi\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K$, we define a discrete H^1 seminorm on the diamond mesh with the help of the discrete gradient operator (see (3.2)):

$$|\phi|_{1,D}:=\left(\boldsymbol{\nabla}_h^D\phi,\boldsymbol{\nabla}_h^D\phi\right)_D^{1/2}.$$

Finally, H^m is the space of functions v of $\mathrm{L}^2(\Omega)$ whose partial derivatives (in the distributional sense) $\partial^\alpha v$, with $|\alpha|\le m$, all belong to $\mathrm{L}^2(\Omega)$, while $||\cdot||_{m,\Omega}$ is the associated norm. The standard $\mathrm{L}^2(\Omega)$ inner product will be denoted by $(\cdot,\cdot)_\Omega$.

3. Construction of the discrete operators. In this section, we approach the gradient, divergence, and curl operators by discrete counterparts. We would like to stress that in two dimensions a distinction is usually made between the vector curl operator from $\mathbb{R}$ to $\mathbb{R}^2$, defined by $\boldsymbol{\nabla}\times\phi=(\frac{\partial\phi}{\partial y},-\frac{\partial\phi}{\partial x})^T$ and the scalar curl operator from $\mathbb{R}^2$ to $\mathbb{R}$, defined by $\nabla\times\mathbf{u}=\frac{\partial\mathbf{u}_y}{\partial x}-\frac{\partial\mathbf{u}_x}{\partial y}$.

Figure 3.1 shows the stencils of the different operators and their combinations: The stencil for the discrete gradient and vector curl operators simply consists of the four corners of the diamond-cell D_j. The stencil for the discrete divergence and scalar curl operators consists of the diamonds associated with the edges of the primal and dual cells. Arrows are displayed on Figure 3.1 to represent the normal and tangential components of the vector fields associated with the diamonds. The stencils for the discrete Laplacian on the primal and dual cells, respectively, consist of the black and white circles on the left and right part of the figure.

3.1. Construction of the discrete gradient and vector curl operators on the diamond-cells. We define the discrete gradient of a function ϕ by its values on the diamond-cells of the mesh. We follow [8, 10] and compute the mean-value of the

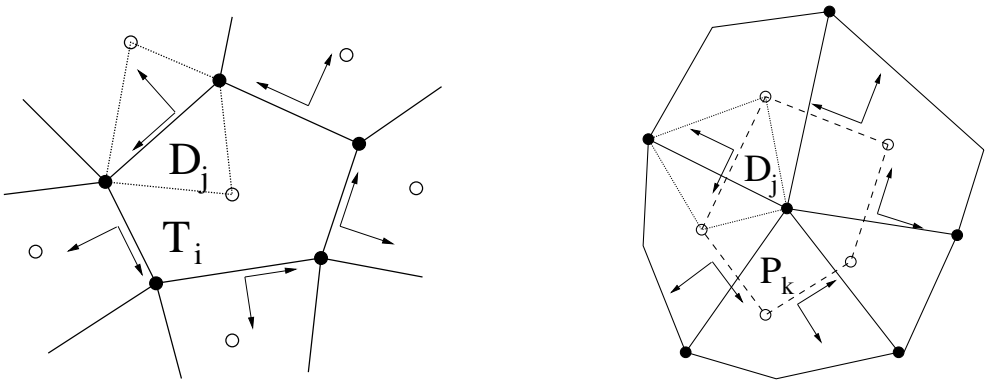


FIG. 3.1. *Stencils for the discrete operators. Left part: primal cell. Right part: dual cell.*

gradient of any function ϕ on such a cell D_j by the following formula:

$$
\begin{aligned}
|D_j| \left\langle \boldsymbol{\nabla}\phi_{|D_j} \right\rangle &= \int_{D_j} \boldsymbol{\nabla}\phi(\mathbf{x})\, d\mathbf{x} = \int_{\partial D_j} \phi(\xi)\mathbf{n}(\xi)\, d\xi \\
&= \sum_{(\alpha,\beta)} \int_{[G_{i_\alpha} S_{k_\beta}]} \phi(\xi)\, \mathbf{n}\, d\xi,
\end{aligned}
\tag{3.1}
$$

where $\mathbf{n}(\xi)$ stands for the outward unit normal vector to D_j at point ξ. The integrals in (3.1) can be approximated by the following formula:

$$
\int_{[GS]} \phi(\xi)\, d\xi \approx \ell_{GS}\, \frac{[\phi(G) + \phi(S)]}{2},
$$

where ℓ_{GS} denotes the length of the segment $[GS]$. Summing the contributions of the different vertices of D_j and using elementary geometrical equalities allows us to give *the definition* of the discrete gradient $\boldsymbol{\nabla}_h^D$ on D_j.

DEFINITION 3.1. *The discrete gradient $\boldsymbol{\nabla}_h^D$ is defined by its values over the diamond-cells D_j:*

$$
(\boldsymbol{\nabla}_h^D \phi)_j := \frac{1}{2\,|D_j|} \left\{ \left[\phi_{k_2}^P - \phi_{k_1}^P\right] |A_j'| \mathbf{n}_j' + \left[\phi_{i_2}^T - \phi_{i_1}^T\right] |A_j| \mathbf{n}_j \right\},
\tag{3.2}
$$

where we set $\phi_{k_\alpha}^P := \phi(S_{k_\alpha})$ and $\phi_{i_\alpha}^T := \phi(G_{i_\alpha})$, for $\alpha \in \{1;2\}$.

Note that formula (3.2) is exact for polynomials of degree one. Computing the discrete gradient only requires the values of ϕ at the nodes of the primal and dual meshes. The operator $\boldsymbol{\nabla}_h^D$ thus acts from $\mathbb{R}^{I+J^\Gamma} \times \mathbb{R}^K$ into $\left(\mathbb{R}^2\right)^J$.

In the same way, we may approach the vector curl operator $\boldsymbol{\nabla} \times \bullet = (\frac{\partial \bullet}{\partial y}, -\frac{\partial \bullet}{\partial x})^T$ by a discrete vector curl operator.

DEFINITION 3.2. *The discrete vector curl operator $\boldsymbol{\nabla}_h^D \times$ is defined by its values over the diamond-cells D_j:*

$$
(\boldsymbol{\nabla}_h^D \times \phi)_j := -\frac{1}{2\,|D_j|} \left\{ \left[\phi_{k_2}^P - \phi_{k_1}^P\right] |A_j'| \boldsymbol{\tau}_j' + \left[\phi_{i_2}^T - \phi_{i_1}^T\right] |A_j| \boldsymbol{\tau}_j \right\},
\tag{3.3}
$$

where the unit vectors $\boldsymbol{\tau}_j$ and $\boldsymbol{\tau}_j'$ are such that $(\mathbf{n}_j, \boldsymbol{\tau}_j)$ and $(\mathbf{n}_j', \boldsymbol{\tau}_j')$ are orthonormal positively oriented bases of $\mathbb{R}^2$.

Remark 3.3. In a connected domain, the discrete gradient and vector curl of a given $\phi = ((\phi_i^T), (\phi_k^P))$ vanish iff there exist two constants c^T and c^P, such that $\phi_i^T = c^T \ \forall\ i$ and $\phi_k^P = c^P \ \forall\ k$. The fact that c^T and c^P may differ from each other means that such a ϕ may, in general, present oscillations. However, in the applications studied in the present work, such oscillations never appear due to information on the mean-value of ϕ (see (4.16) and (5.7d)), or due to boundary conditions (4.17) and (5.8e).

3.2. Construction of the discrete divergence and scalar curl operators on the primal and dual meshes. Next, we choose to define the discrete divergence of a vector field $\mathbf{u}$ by its values both on the primal and dual cells of the mesh. A very natural way to do so on the primal cell T_i is to write

$$|T_i| \left\langle \nabla \cdot \mathbf{u}_{|T_i} \right\rangle = \int_{T_i} \nabla \cdot \mathbf{u}(\mathbf{x})\, d\mathbf{x} = \int_{\partial T_i} \mathbf{u}(\xi) \cdot \mathbf{n}(\xi) = \sum_{j \in \mathcal{V}(i)} \int_{A_j} \mathbf{u}(\xi) \cdot \mathbf{n}_{ji}\,,$$

where we recall that $\mathcal{V}(i)$ is the set of integers $j \in [1, J]$ such that A_j is an edge of T_i and that $\mathbf{n}_{ji}$ is the unit vector orthogonal to A_j pointing outward T_i. Supposing that the vector field $\mathbf{u}$ is given by both of the Cartesian components of its discrete values $\mathbf{u}_j$ on the diamond-cells D_j, and performing a similar computation over the cells P_k, we obtain *the definition* of the discrete divergence $\nabla_h^T \cdot$ on each T_i and the discrete divergence $\nabla_h^P \cdot$ on each P_k.

DEFINITION 3.4. *The discrete divergence* $\nabla_h^{T,P} \cdot := (\nabla_h^T \cdot, \nabla_h^P \cdot)$ *is defined by its values over the primal cells* T_i *and the dual cells* P_k*:*

$$\begin{aligned} (\nabla_h^T \cdot \mathbf{u})_i &:= \frac{1}{|T_i|} \sum_{j \in \mathcal{V}(i)} |A_j| \mathbf{u}_j \cdot \mathbf{n}_{ji}, \\ (\nabla_h^P \cdot \mathbf{u})_k &:= \frac{1}{|P_k|} \left(\sum_{j \in \mathcal{E}(k)} |A'_j| \mathbf{u}_j \cdot \mathbf{n}'_{jk} + \sum_{j \in \mathcal{E}(k) \cap [J - J^\Gamma + 1, J]} \frac{1}{2}\, |A_j| \mathbf{u}_j \cdot \mathbf{n}_j \right). \end{aligned} \tag{3.4}$$

Remark that if the node S_k is not on the boundary Γ (i.e., if $k \in [1, K - J^\Gamma]$), then the set $\mathcal{E}(k) \cap [J - J^\Gamma + 1, J]$ is empty. On the contrary, if P_k is a boundary dual cell, then the set $\mathcal{E}(k) \cap [J - J^\Gamma + 1, J]$ is composed of the two boundary edges which have S_k as a vertex. In this case, the quantity $\sum_{j \in \mathcal{E}(k) \cap [J - J^\Gamma + 1, J]} \frac{1}{2}\, |A_j| \mathbf{u}_j \cdot \mathbf{n}_j$ is an approximation of $\int_{\tilde{A}_k} \mathbf{u} \cdot \tilde{\mathbf{n}}_k(\xi)\, d\xi$ (see Figure 2.3).

For a given vector field $\mathbf{u}$, it is easily checked that these formulae are the exact mean-values of $\nabla \cdot \mathbf{u}$ over the primal and inner dual cells if $\mathbf{u}_j \cdot \mathbf{n}_{ji}$ and $\mathbf{u}_j \cdot \mathbf{n}'_{jk}$ represent the mean-values of $\mathbf{u} \cdot \mathbf{n}_{ji}$ over A_j and of $\mathbf{u} \cdot \mathbf{n}'_{jk}$ over A'_j. The operator $\nabla_h \cdot$ acts from $\left(\mathbb{R}^2\right)^J$ into $\mathbb{R}^I \times \mathbb{R}^K$.

In the same way, we may approach the scalar curl operator $\nabla \times \bullet = (\frac{\partial \bullet_y}{\partial x} - \frac{\partial \bullet_x}{\partial y})$ by a discrete scalar curl operator in the following definition.

DEFINITION 3.5. *The discrete scalar curl operator* $\nabla_h^{T,P} \times := (\nabla_h^T \times, \nabla_h^P \times)$ *is*

defined by its values over the primal cells T_i and the dual cells P_k:

$$(\nabla_h^T \times \mathbf{u})_i := \frac{1}{|T_i|} \sum_{j \in \mathcal{V}(i)} |A_j| \mathbf{u}_j \cdot \boldsymbol{\tau}_{ji},$$

$$(\nabla_h^P \times \mathbf{u})_k := \frac{1}{|P_k|} \left(\sum_{j \in \mathcal{E}(k)} |A_j^{'}| \mathbf{u}_j \cdot \boldsymbol{\tau}_{jk}^{'} + \sum_{j \in \mathcal{E}(k) \cap [J - J^\Gamma + 1, J]} \frac{1}{2} |A_j| \mathbf{u}_j \cdot \boldsymbol{\tau}_j \right). \tag{3.5}$$

4. Properties of the operators.

4.1. Discrete Green formulae. Here, we check that the discrete operators verify some discrete duality principles.

PROPOSITION 4.1. *The following discrete analogues of the Green formulae hold:*

$$(\nabla_h^{T,P} \cdot \mathbf{u}, \phi)_{T,P} = -(\mathbf{u}, \boldsymbol{\nabla}_h^D \phi)_D + (\mathbf{u} \cdot \mathbf{n}, \phi)_{\Gamma,h}, \tag{4.1}$$

$$(\nabla_h^{T,P} \times \mathbf{u}, \phi)_{T,P} = (\mathbf{u}, \boldsymbol{\nabla}_h^D \times \phi)_D + (\mathbf{u} \cdot \boldsymbol{\tau}, \phi)_{\Gamma,h} \tag{4.2}$$

for all $\mathbf{u} \in \left(\mathbb{R}^2\right)^J$ and all $\phi = (\phi^T, \phi^P) \in \mathbb{R}^{I+J^\Gamma} \times \mathbb{R}^K$, where the definitions (2.1), (2.2), *and* (2.3) *have been used.*

Proof. The proof of (4.1) may be found in [10] and is based on a discrete summation by parts. The proof of (4.2) follows exactly the same lines. ∎

4.2. Compositions of the discrete operators. The aim of this section is to verify a discrete analogue of the following continuous identities: $\nabla \cdot (\boldsymbol{\nabla} \times) = 0$, $\nabla \times \boldsymbol{\nabla} = 0$, and $\nabla \times \boldsymbol{\nabla} \times = -\nabla \cdot \boldsymbol{\nabla}$. For this, we start with a useful lemma.

LEMMA 4.2. *Recall that s_{ji} and s'_{jk} are defined in section* 2.4. *Then,*

$$\sum_{j \in \mathcal{V}(i)} s_{ji} \left(\phi^P_{k_2(j)} - \phi^P_{k_1(j)} \right) = 0 \quad \forall i \in [1, I], \tag{4.3}$$

$$\sum_{j \in \mathcal{E}(k)} s'_{jk} \left(\phi^T_{i_2(j)} - \phi^T_{i_1(j)} \right) = 0 \quad \forall\, k \in [1, K - J^\Gamma]. \tag{4.4}$$

Proof. Let us consider a given primal cell T_i. For each edge A_j of T_i, with $j \in \mathcal{V}(i)$, there are two possibilities for the orientation of $\mathbf{n}_j$ (see Figure 4.1): If $\mathbf{n}_j$ is the inward unit normal vector to T_i (case 1), then $s_{ji} = -1$ and $s_{ji}\,(\phi^P_{k_2(j)} - \phi^P_{k_1(j)}) = \phi^P_{k_1(j)} - \phi^P_{k_2(j)}$. If $\mathbf{n}_j$ is the outward unit normal vector to T_i (case 2), then $s_{ji} = 1$ and $s_{ji}\,(\phi^P_{k_2(j)} - \phi^P_{k_1(j)}) = \phi^P_{k_2(j)} - \phi^P_{k_1(j)}$; moreover, $S_{k_1}(j)$ and $S_{k_2}(j)$ are swapped. What appears finally is that, whatever the case, the value ϕ^P_k associated with the "left" vertex of the considered edge A_j appears in the sum (4.3) with a positive sign and the value ϕ^P_k associated with the "right" vertex of the considered edge A_j appears in the sum (4.3) with a negative sign. But each ϕ^P_k appears twice in that sum, once as the value associated with the "right" vertex of a given edge, and once as the value associated with the "left" vertex of the following edge, so that these two contributions cancel. This ends the proof of (4.3). The proof of (4.4) follows the same lines. ∎

Next, the following properties are direct consequences of the computation of the area $|D_j|$.

LEMMA 4.3.

$$\frac{|A_j|\,|A'_j|}{2\,|D_j|}\ \mathbf{n}_j\cdot\boldsymbol{\tau}'_j=1\quad\forall j\in[1,J], \tag{4.5}$$

$$\frac{|A_j|\,|A'_j|}{2\,|D_j|}\ \mathbf{n}'_j\cdot\boldsymbol{\tau}_j=-1\quad\forall j\in[1,J]. \tag{4.6}$$

We may now state the following results.

PROPOSITION 4.4. *Given any* $\phi=(\phi_i^T,\phi_k^P)\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K$, *there holds*

$$\left(\nabla_h^T\cdot(\boldsymbol{\nabla}_h^D\times\phi)\right)_i=0\quad\forall i\in[1,I], \tag{4.7}$$

$$\left(\nabla_h^P\cdot(\boldsymbol{\nabla}_h^D\times\phi)\right)_k=0\quad\forall k\in[1,K-J^\Gamma], \tag{4.8}$$

$$\left(\nabla_h^T\times(\boldsymbol{\nabla}_h^D\phi)\right)_i=0\quad\forall i\in[1,I], \tag{4.9}$$

$$\left(\nabla_h^P\times(\boldsymbol{\nabla}_h^D\phi)\right)_k=0\quad\forall k\in[1,K-J^\Gamma]. \tag{4.10}$$

Moreover, on each boundary dual cell P_k *(*$k\in[K-J^\Gamma+1,K]$*),* (4.8) *and* (4.10) *still hold if there exist for each boundary* Γ_q, *with* $q\in[0,Q]$, *two real numbers* (c_q^T,c_q^P) *such that* $\phi_i^T=c_q^T$ *and* $\phi_k^P=c_q^P$ *uniformly over* Γ_q.

Proof. Let us first prove (4.7); combining (3.4), (3.3), and the fact that $\mathbf{n}_{ji}\cdot\boldsymbol{\tau}_j=0$, we get

$$\begin{aligned}(\nabla_h^T\cdot(\boldsymbol{\nabla}_h^D\times\phi))_i&=\frac{1}{|T_i|}\sum_{j\in\mathcal{V}(i)}|A_j|(\boldsymbol{\nabla}_h^D\times\phi)_j\cdot\mathbf{n}_{ji}\\&=-\frac{1}{|T_i|}\sum_{j\in\mathcal{V}(i)}\frac{|A_j|\,|A'_j|}{2\,|D_j|}\,\mathbf{n}_j\cdot\boldsymbol{\tau}'_j\,s_{ji}\left(\phi^P_{k_2(j)}-\phi^P_{k_1(j)}\right)\quad\forall i\in[1,I].\end{aligned}$$

Applying (4.5) and (4.3) successively, we obtain

$$(\nabla_h^T\cdot(\boldsymbol{\nabla}_h^D\times\phi))_i=0\quad\forall i\in[1,I].$$

Equation (4.9) can be proved in a similar way.

Next, for each interior dual cell P_k, with $k\in[1,K-J^\Gamma]$, the set $\mathcal{E}(k)\cap[J-J^\Gamma+1,J]$ is empty, so that (4.8) and (4.10) can be proved like (4.7) and (4.9), using (4.6), (4.4) and the fact that $\mathbf{n}'_{jk}\cdot\boldsymbol{\tau}'_j=0$.

As far as the boundary dual cells P_k are concerned ($k\in[K-J^\Gamma+1,K]$), similar computations show that (see Figure 4.2 for the notations)

$$(\nabla_h^P\cdot(\boldsymbol{\nabla}_h^D\times\phi))_k=\frac{1}{|P_k|}\left(\phi^T_{I_2}-\phi^T_{I_1}\right)+\frac{1}{2\,|P_k|}\left(\phi^P_{K_1}-\phi^P_{K_2}\right). \tag{4.11}$$

If all ϕ_i^T are equal to the same constant c_q^T over Γ_q and if all ϕ_k^P are equal to the same constant c_q^P over Γ_q, then $\phi^T_{I_2}=\phi^T_{I_1}$ and $\phi^P_{K_1}=\phi^P_{K_2}$ so that

$$(\nabla_h^P\cdot(\boldsymbol{\nabla}_h^D\times\phi))_k=0,$$

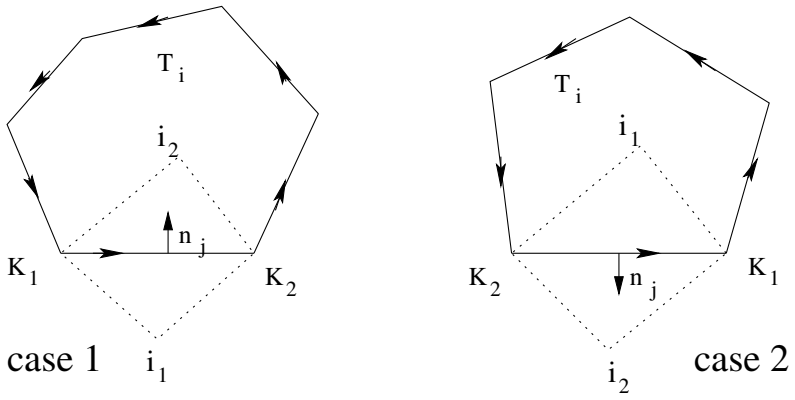


FIG. 4.1. *Two possibilities of orientation for each edge.*

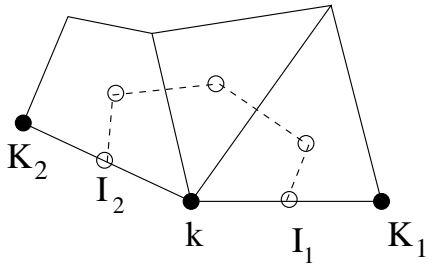


FIG. 4.2. *Notations for the boundary dual cells.*

for the boundary dual cells, and (4.10) for the boundary dual cells is proved in a similar way. ∎

PROPOSITION 4.5. *The following equalities hold:*

$$
\begin{aligned}
(\nabla_h^T \times \boldsymbol{\nabla}_h^D \times \phi)_i &= -(\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \phi)_i \quad \forall i \in [1, I], \\
(\nabla_h^P \times \boldsymbol{\nabla}_h^D \times \phi)_k &= -(\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \phi)_k \quad \forall k \in [1, K].
\end{aligned} \tag{4.12}
$$

Proof. These formulae follow immediately from the definitions (3.2), (3.3), (3.4), and (3.5) and from the equality $\boldsymbol{\tau}_j \cdot \boldsymbol{\tau}'_j = \mathbf{n}_j \cdot \mathbf{n}'_j \quad \forall j \in [1, J]$. ∎

4.3. Hodge's decomposition. In the continuous case, the Hodge decomposition for nonsimply connected domains reads

$$
(L^2)^2 = \boldsymbol{\nabla} V \overset{\perp}{\oplus} \boldsymbol{\nabla} \times W, \tag{4.13}
$$

with $V = \{\phi \in H^1 : \int_\Omega \phi = 0\}$ and $W = \{\psi \in H^1 : \psi_{|\Gamma_0} = 0, \quad \psi_{|\Gamma_q} = c_q \ \forall q \in [1, Q]\}$. To prove an analogous property in the discrete case, we rely on the following result.

LEMMA 4.6 (Euler's formula). *For a nonsimply connected bidimensional domain covered by a mesh with I elements, K vertices, J edges, and Q holes, there holds*

$$
I + K = J + 1 - Q. \tag{4.14}
$$

We may now state the following discrete Hodge decomposition.

THEOREM 4.7. *Let $(\mathbf{u}_j)_{j\in[1,J]}$ be a discrete vector field defined by its values on the diamond-cells D_j. There exist unique $\phi = (\phi_i^T, \phi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$, $\psi = (\psi_i^T, \psi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$, and $(c_q^T, c_q^P)_{q\in[1,Q]}$ such that*

$$\mathbf{u}_j = (\boldsymbol{\nabla}_h^D \phi)_j + (\boldsymbol{\nabla}_h^D \times \psi)_j \quad \forall j \in [1, J], \tag{4.15}$$

$$\sum_{i\in[1,I]} |T_i|\phi_i^T = \sum_{k\in[1,K]} |P_k|\phi_k^P = 0, \tag{4.16}$$

$$\psi_i^T = 0 \ \ \forall i \in \Gamma_0, \quad \psi_k^P = 0 \ \ \forall k \in \Gamma_0, \tag{4.17}$$

and

$$\forall q \in [1, Q], \qquad \psi_i^T = c_q^T \ \ \forall i \in \Gamma_q, \quad \psi_k^P = c_q^P \ \ \forall k \in \Gamma_q. \tag{4.18}$$

Moreover, the decomposition (4.15) *is orthogonal.*

Proof. There are $2(I + K + J^\Gamma) + 2Q$ unknowns corresponding to (ϕ_i^T, ϕ_k^P) and (ψ_i^T, ψ_k^P) and to the constants (c_q^T, c_q^P). On the other hand, $2J$ equations are given by (4.15), while (4.17) and (4.18) provide $2J^\Gamma$ constraints. Finally, (4.16) gives two supplementary equalities, so that the total number of equations is $2J + 2 + 2J^\Gamma$. Consequently, according to (4.14), there are as many equations as unknowns. Therefore, existence and uniqueness of the decomposition are equivalent, and we shall prove uniqueness through injectivity.

Proving the orthogonality of $(\boldsymbol{\nabla}_h^D \phi)$ and $(\boldsymbol{\nabla}_h^D \times \psi)$ for any (ϕ, ψ) verifying (4.17) and (4.18) amounts to showing $(\boldsymbol{\nabla}_h^D \times \psi, \boldsymbol{\nabla}_h^D \phi)_D = 0$. Thanks to (4.1), there holds

$$(\boldsymbol{\nabla}_h^D \times \psi, \boldsymbol{\nabla}_h^D \phi)_D = -(\nabla_h^{T,P} \cdot \boldsymbol{\nabla}_h^D \times \psi, \phi)_{T,P} + (\boldsymbol{\nabla}_h^D \times \psi \cdot \mathbf{n}, \phi)_{\Gamma,h}.$$

Next, thanks to Proposition 4.4, $\nabla_h^{T,P} \cdot \boldsymbol{\nabla}_h^D \times \psi$ vanishes on all primal and inner dual cells. Because ψ verifies (4.17) and (4.18), we infer from Proposition 4.4 that $\nabla_h^{T,P} \cdot \boldsymbol{\nabla}_h^D \times \psi$ also vanishes on the boundary dual cells. Finally, according to (3.3), we have

$$(\boldsymbol{\nabla}_h^D \times \psi)_j \cdot \mathbf{n}_j = -\frac{1}{2\,|D_j|} \left(\psi_{k_2}^P - \psi_{k_1}^P\right) |A_j'| \boldsymbol{\tau}_j' \cdot \mathbf{n}_j,$$

which also vanishes on the boundary because of (4.17) and (4.18). Thus, orthogonality is proved. In order to prove injectivity, we suppose $\mathbf{u}_j = 0 \ \forall j \in [1, J]$:

$$0 = (\boldsymbol{\nabla}_h^D \phi)_j + (\boldsymbol{\nabla}_h^D \times \psi)_j \quad \forall j \in [1, J]. \tag{4.19}$$

We carry out the scalar product of (4.19) with $|D_j|\ (\boldsymbol{\nabla}_h^D \phi)_j$ and sum over $j \in [1, J]$:

$$0 = (\boldsymbol{\nabla}_h^D \phi, \boldsymbol{\nabla}_h^D \phi)_D + (\boldsymbol{\nabla}_h^D \times \psi, \boldsymbol{\nabla}_h^D \phi)_D. \tag{4.20}$$

Thanks to the orthogonality proved above, (4.20) implies that $(\boldsymbol{\nabla}_h^D \phi, \boldsymbol{\nabla}_h^D \phi)_D = \sum_{j\in[1,J]} |D_j| |(\boldsymbol{\nabla}_h^D \phi)_j|^2 = 0$, so that $(\boldsymbol{\nabla}_h^D \phi)_j = 0 \ \forall$ j. Since the domain is connected, there exist two real constants α and β such that $\phi_k^P = \alpha \ \forall k \in [1, K]$ and $\phi_i^T = \beta \ \forall i \in [1, I + J^\Gamma]$. Equation (4.16) implies that these two constants vanish, so that

$$\phi_i^T = 0 \quad \forall i \in [1, I + J^\Gamma] \ \text{ and } \ \phi_k^P = 0 \quad \forall k \in [1, K].$$

Consequently, (4.19) is equivalent to $(\boldsymbol{\nabla}_h^D \times \psi)_j = 0 \; \forall j \in [1, J]$. Since the domain is connected, there exist two real constants α and β such as $\psi_k^P = \alpha \; \forall k \in [1, K]$ and $\psi_i^T = \beta \; \forall i \in [1, I + J^\Gamma]$. As $\psi = 0$ over Γ_0 these two constants vanish and

$$\psi_i^T = 0 \quad \forall i \in [1, I + J^\Gamma] \;\text{ and }\; \psi_k^P = 0 \quad \forall k \in [1, K] \,. \qquad \square$$

Remark 4.8. The two equalities in (4.16) are discrete analogues, respectively stated on the primal mesh and dual mesh, of the condition $\int_\Omega \phi = 0$ that appears in the definition of the space V in (4.13), while formulae (4.17) and (4.18) are discrete analogues of the boundary conditions that appear in the definition of W.

5. Numerical solution of the div-curl problem for nonsimply connected domains.

5.1. Discretization of the div-curl problem with normal boundary conditions. We are interested in the approximation of the following continuous problem: given f, g, σ, $(k_q)_{q\in[1,Q]}$, find $\mathbf{u}$ such that

$$\begin{cases} \nabla \cdot \mathbf{u} = f \;\text{ in }\; \Omega, \\ \nabla \times \mathbf{u} = g \;\text{ in }\; \Omega, \\ \mathbf{u} \cdot \mathbf{n} = \sigma \;\text{ on }\; \Gamma, \\ \int_{\Gamma_q} \mathbf{u} \cdot \boldsymbol{\tau} = k_q \quad \forall q \in [1, Q]. \end{cases} \tag{5.1}$$

A necessary condition for the existence of a solution to (5.1) is given by the formula

$$\int_\Omega f(\mathbf{x}) d\mathbf{x} = \int_\Gamma \sigma(\xi) \, d\xi. \tag{5.2}$$

We discretize the solution of this problem by a vector field $(\mathbf{u}_j)_{j\in[1,J]}$ defined by its values over the diamond-cells of the mesh. Using the discrete differential operators defined in section 3, and following [12], we write the following discrete equations:

$$\left(\nabla_h^T \cdot \mathbf{u}\right)_i = f_i^T \quad \forall i \in [1, I], \tag{5.3a}$$

$$\left(\nabla_h^P \cdot \mathbf{u}\right)_k = f_k^P \quad \forall k \in [1, K], \tag{5.3b}$$

$$\left(\nabla_h^T \times \mathbf{u}\right)_i = g_i^T \quad \forall i \in [1, I], \tag{5.3c}$$

$$\left(\nabla_h^P \times \mathbf{u}\right)_k = g_k^P \quad \forall k \in [1, K - J^\Gamma], \tag{5.3d}$$

$$\mathbf{u}_j \cdot \mathbf{n}_j = \sigma_j \quad \forall j \in [J - J^\Gamma + 1, J], \tag{5.3e}$$

$$(\mathbf{u} \cdot \boldsymbol{\tau}, 1)_{\Gamma_q, h} = k_q \quad \forall q \in [1, Q], \tag{5.3f}$$

$$\sum_{k\in\Gamma_q} |P_k| \;\; (\nabla_h^P \times \mathbf{u})_k = \sum_{k\in\Gamma_q} |P_k| \;\; g_k^P \quad \forall q \in [1, Q], \tag{5.3g}$$

where the following definitions have been used:

$$f_i^T = \frac{1}{|T_i|} \int_{T_i} f(\mathbf{x}) \, d\mathbf{x} \quad \forall i \in [1, I], \;\; f_k^P = \frac{1}{|P_k|} \int_{P_k} f(\mathbf{x}) \, d\mathbf{x} \quad \forall k \in [1, K], \tag{5.4}$$

$$g_i^T = \frac{1}{|T_i|} \int_{T_i} g(\mathbf{x}) \, d\mathbf{x} \quad \forall i \in [1, I], \;\; g_k^P = \frac{1}{|P_k|} \int_{P_k} g(\mathbf{x}) \, d\mathbf{x} \quad \forall k \in [1, K], \tag{5.5}$$

$$\sigma_j = \frac{1}{|A_j|} \int_{A_j} \sigma(\xi) \, d\xi \quad \forall j \in [J - J^\Gamma + 1, J]. \tag{5.6}$$

Using the discrete Hodge decomposition of $(\mathbf{u}_j)_{j\in[1,J]}$, problem (5.3) may be split into two independent problems involving the following potentials.

PROPOSITION 5.1. *Problem* (5.3) *can be split into two independent problems: Find* $(\phi_i^T, \phi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$ *such that*

$$(\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \phi)_i = f_i^T \quad \forall i \in [1, I], \tag{5.7a}$$

$$(\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \phi)_k = f_k^P \quad \forall k \in [1, K], \tag{5.7b}$$

$$(\boldsymbol{\nabla}_h^D \phi)_j \cdot \mathbf{n}_j = \sigma_j \quad \forall j \in [J - J^\Gamma + 1, J], \tag{5.7c}$$

$$\sum_{i\in[1,I]} |T_i| \phi_i^T = \sum_{k\in[1,K]} |P_k| \phi_k^P = 0. \tag{5.7d}$$

Find $(\psi_i^T, \psi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$ *and* $(c_q^T, c_q^P)_{q\in[1,Q]}$ *such that*

$$-(\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \psi)_i = g_i^T \quad \forall i \in [1, I], \tag{5.8a}$$

$$-(\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k = g_k^P \quad \forall k \in [1, K - J^\Gamma], \tag{5.8b}$$

$$(\boldsymbol{\nabla}_h^D \psi \cdot \mathbf{n}, 1)_{\Gamma_q,h} = -k_q \quad \forall q \in [1, Q], \tag{5.8c}$$

$$-\sum_{k\in\Gamma_q} |P_k| \ (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k = \sum_{k\in\Gamma_q} |P_k| \ g_k^P \quad \forall q \in [1, Q], \tag{5.8d}$$

$$\psi_i^T = \psi_k^P = 0 \quad \forall i \in \Gamma_0 \quad \forall k \in \Gamma_0, \tag{5.8e}$$

$$\forall q \in [1, Q] \ \ \psi_i^T = c_q^T \ \ \forall i \in \Gamma_q, \tag{5.8f}$$

$$\forall q \in [1, Q] \ \ \psi_k^P = c_q^P \ \ \forall k \in \Gamma_q. \tag{5.8g}$$

The vector $\mathbf{u}$ *is then reconstructed by*

$$\mathbf{u}_j = (\boldsymbol{\nabla}_h^D \phi)_j + (\boldsymbol{\nabla}_h^D \times \psi)_j \quad \forall j \in [1, J]. \tag{5.9}$$

Proof. First, the discrete Hodge decomposition of $(\mathbf{u}_j)_{j\in[1,J]}$ shows the existence of $(\phi_i^T, \phi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$, $(\psi_i^T, \psi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$, and $(c_q^T, c_q^P)_{q\in[1,Q]}$ such that (5.9), (5.7d), and (5.8e)–(5.8g) are verified. Next, (5.7a) is proved using (4.7):

$$f_i^T = (\nabla_h^T \cdot \mathbf{u})_i = (\nabla_h^T \cdot (\boldsymbol{\nabla}_h^D \phi + \boldsymbol{\nabla}_h^D \times \psi))_i = (\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \phi)_i \quad \forall i \in [1, I].$$

Similarly, using (4.8) and $\psi_i^T = c_q^T$ and $\psi_k^P = c_q^P \ \ \forall q \in [0, Q]$, we obtain (5.7b). As far as the boundary conditions are concerned, using (3.3) shows that

$$(\boldsymbol{\nabla}_h^D \times \psi)_j \cdot \mathbf{n}_j = -\frac{1}{2\,|D_j|} \left(\psi_{k_2} - \psi_{k_1}\right) |A_j'| \boldsymbol{\tau}_j' \cdot \mathbf{n}_j \quad \forall j \in [J - J^\Gamma + 1, J]. \tag{5.10}$$

Since $\psi_k^P = c_q^P \ \ \forall q \in [0, Q]$, we infer from (5.10)

$$(\boldsymbol{\nabla}_h^D \times \psi)_j \cdot \mathbf{n}_j = 0 \quad \forall j \in [J - J^\Gamma + 1, J],$$

so that (5.3e) and (5.9) imply (5.7c). Further, using (5.9), (5.3c)–(5.3d), (4.9), (4.10), and (4.12), we may prove (5.8a)–(5.8b). Moreover, there holds

$$(\boldsymbol{\nabla}_h^D \phi)_j \cdot \boldsymbol{\tau}_j = \frac{1}{2|D_j|} \left(\phi_{k_2(j)}^T - \phi_{k_1(j)}^T\right) |A_j'| \mathbf{n}_j' \cdot \boldsymbol{\tau}_j,$$

so that, using (4.6),

$$(\boldsymbol{\nabla}_h^D \phi \cdot \boldsymbol{\tau}, 1)_{\Gamma_q,h} = \sum_{j\in\Gamma_q} \frac{|A_j|\,|A_j'|}{2\,|D_j|}\mathbf{n}_j' \cdot \boldsymbol{\tau}_j \left(\phi_{k_2(j)}^T - \phi_{k_1(j)}^T\right) = -\sum_{j\in\Gamma_q} \left(\phi_{k_2(j)}^T - \phi_{k_1(j)}^T\right),$$

which vanishes because Γ_q is a closed contour. Thus, (5.3f) implies (5.8c) because $(\boldsymbol{\nabla}_h^D \times \psi)\cdot\boldsymbol{\tau}_j = -\boldsymbol{\nabla}_h^D\psi\cdot\mathbf{n}_j$. Finally, a computation similar to that which led to (4.11) shows that

$$(\nabla_h^P \times (\boldsymbol{\nabla}_h^D\phi))_k = \frac{1}{|P_k|}\left(\phi_{I_2}^T - \phi_{I_1}^T\right) + \frac{1}{2\,|P_k|}\left(\phi_{K_1}^P - \phi_{K_2}^P\right)$$

for boundary cells $k \in [K - J^\Gamma + 1, K]$ (see Figure 4.2 for the notations). Thus, when summing these contributions over a closed contour Γ_q, we obtain

$$\sum_{k\in\Gamma_q} |P_k|(\nabla_h^P \times (\boldsymbol{\nabla}_h^D\phi))_k = 0,$$

so that (5.3g) implies (5.8d). ∎

PROPOSITION 5.2. *Problems* (5.7) *and* (5.8) *both have a unique solution.*

Proof. As far as problem (5.7) is concerned, the existence and uniqueness of its solution have been proved in [10] if the following discrete equivalent of (5.2) is verified:

$$\sum_{i\in[1,I]} |T_i|\, f_i^T = \sum_{k\in[1,K]} |P_k|\, f_k^P = \sum_{j\in[J-J^\Gamma+1,J]} |A_j|\,\sigma_j,$$

which is the case here because, thanks to the definitions (5.4) and (5.6), we have

$$\sum_{i\in[1,I]} |T_i|\, f_i^T = \sum_{k\in[1,K]} |P_k|\, f_k^P = \int_\Omega f(\mathbf{x})\,d\mathbf{x} \text{ and } \sum_{j\in[J-J^\Gamma+1,J]} |A_j|\,\sigma_j = \int_\Gamma \sigma(\xi)\,d\xi.$$

As far as problem (5.8) is concerned, there are $I+K+J^\Gamma+2Q$ unknowns, while (5.8a) and (5.8b), respectively, provide I and $K-J^\Gamma$ equations. Equations (5.8c) and (5.8d) provide $2Q$ additional relations. Finally, boundary conditions (5.8e)–(5.8g) provide the last $2J^\Gamma$ equations. Since there are as many equations as unknowns, it suffices to check the injectivity of the system. Let us set $g_i^T = g_k^P = k_q = 0$ in system (5.8) and compute the following discrete scalar product $(\nabla_h^{T,P}\cdot\boldsymbol{\nabla}_h^D\psi,\psi)_{T,P}$ (see (2.1) for the definition). In this scalar product, the sum over the indices $i\in[1,I]$ and the sum over the indices $k\in[1,K-J^\Gamma]$ vanish, respectively, because of (5.8a) and (5.8b). Further, due to (5.8e), the contributions of the indices $k\in\Gamma_0$ also vanish, so that

$$(\nabla_h^{T,P}\cdot\boldsymbol{\nabla}_h^D\psi,\psi)_{T,P} = \frac{1}{2}\sum_{q\in[1,Q]}\sum_{k\in\Gamma_q} |P_k|\,(\nabla_h^P\cdot\boldsymbol{\nabla}_h^D\psi)_k\,\psi_k^P.$$

Further, (5.8g) implies that

$$(\nabla_h^{T,P}\cdot\boldsymbol{\nabla}_h^D\psi,\psi)_{T,P} = \frac{1}{2}\sum_{q\in[1,Q]} c_q^P \sum_{k\in\Gamma_q} |P_k|\,(\nabla_h^P\cdot\boldsymbol{\nabla}_h^D\psi)_k,$$

which vanishes due to (5.8d). Thanks to the discrete Green formula (4.2), there holds

$$(\nabla_h^{T,P}\cdot\boldsymbol{\nabla}_h^D\psi,\psi)_{T,P} = -(\boldsymbol{\nabla}_h^D\psi,\boldsymbol{\nabla}_h^D\psi)_D + (\boldsymbol{\nabla}_h^D\psi\cdot\mathbf{n},\psi)_{\Gamma,h} = 0. \tag{5.11}$$

Now, due to boundary conditions (5.8e)–(5.8g), we may write

$$(\boldsymbol{\nabla}_h^D\psi\cdot\mathbf{n},\psi)_{\Gamma,h}=\sum_{q\in[1,Q]}\frac{c_q^T+c_q^P}{2}\ (\boldsymbol{\nabla}_h^D\psi\cdot\mathbf{n},1)_{\Gamma_q,h}, \tag{5.12}$$

which vanishes thanks to (5.8c). Thus, (5.11), (5.12), and definition (2.2) imply that

$$(\boldsymbol{\nabla}_h^D\psi,\boldsymbol{\nabla}_h^D\psi)_D=\sum_{j\in[1,J]}|D_j||\boldsymbol{\nabla}_h^D\psi|^2=0.$$

Consequently, just like at the end of the proof of Theorem 4.7, we infer that

$$\psi_i^T=0\quad\forall i\in[1,I+J^\Gamma]\ \text{ and }\ \psi_k^P=0\quad\forall k\in[1,K],$$

which proves uniqueness and thus existence. ∎

5.2. The div-curl problem with tangential boundary conditions. We consider the following continuous problem: given f, g, σ, $(k_q)_{q\in[1,Q]}$, find $\mathbf{u}$ such that

$$\begin{cases}\nabla\cdot\mathbf{u}=f\ \text{ in }\ \Omega,\\ \nabla\times\mathbf{u}=g\ \text{ in }\ \Omega,\\ \mathbf{u}\cdot\boldsymbol{\tau}=\sigma\ \text{ on }\ \Gamma,\\ \int_{\Gamma_q}\mathbf{u}\cdot\mathbf{n}=k_q\ \ \forall q\in[1,Q].\end{cases}$$

A necessary condition for the existence of a solution to this system is given by Green's formula: $\int_\Omega g(\mathbf{x})d\mathbf{x}=\int_\Gamma\sigma(\xi)\,d\xi$. This problem is discretized like in section 5.1 by a vector field $(\mathbf{u}_j)_{j\in[1,J]}$ defined by its values over the diamond-cells. Using the discrete differential operators defined in section 3, we write the following discrete equations:

$$\begin{cases}\big(\nabla_h^T\cdot\mathbf{u}\big)_i=f_i^T\quad\forall i\in[1,I],\\ \big(\nabla_h^P\cdot\mathbf{u}\big)_k=f_k^P\quad\forall k\in[1,K-J^\Gamma],\\ \big(\nabla_h^T\times\mathbf{u}\big)_i=g_i^T\quad\forall i\in[1,I],\\ \big(\nabla_h^P\times\mathbf{u}\big)_k=g_k^P\quad\forall k\in[1,K],\\ \mathbf{u}_j\cdot\boldsymbol{\tau}_j=\sigma_j\quad\forall j\in[J-J^\Gamma+1,J],\\ (\mathbf{u}\cdot\mathbf{n},1)_{\Gamma_q,h}=k_q\quad\forall q\in[1,Q],\\ \sum_{k\in\Gamma_q}|P_k|\ (\nabla_h^P\cdot\mathbf{u})_k=\sum_{k\in\Gamma_q}|P_k|\ f_k^P\quad\forall q\in[1,Q].\end{cases} \tag{5.13}$$

Existence and uniqueness of the solution of (5.13) are proved similarly to section 5.1; the main difference is that the Hodge decomposition is modified in the following way.

THEOREM 5.3. *Let $(\mathbf{u}_j)_{j\in[1,J]}$ be a discrete vector field defined by its values on the diamond-cells D_j. There exist unique $\phi=(\phi_i^T,\phi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$, $\psi=(\psi_i^T,\psi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$, and $(c_q^T,c_q^P)_{q\in[1,Q]}$ such that*

$$\begin{gathered}\mathbf{u}_j=(\boldsymbol{\nabla}_h^D\psi)_j+(\boldsymbol{\nabla}_h^D\times\phi)_j\quad\forall j\in[1,J],\\ \sum_{i\in[1,I]}|T_i|\phi_i^T=\sum_{k\in[1,K]}|P_k|\phi_k^P=0,\\ \psi_i^T=0\quad\forall i\in\Gamma_0,\quad\psi_k^P=0\quad\forall k\in\Gamma_0,\end{gathered} \tag{5.14}$$

and

$$\forall q \in [1,Q], \quad \psi_i^T = c_q^T \quad \forall i \in \Gamma_q, \quad \psi_k^P = c_q^P \quad \forall k \in \Gamma_q.$$

Moreover, the decomposition (5.14) *is orthogonal.*

Further, problem (5.13) decouples into two independent subproblems involving the following potentials.

PROPOSITION 5.4. *Problem* (5.13) *can be split into two independent problems. Find* $(\phi_i^T, \phi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$ *such that*

$$\begin{cases} -(\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \phi)_i = g_i^T \quad \forall i \in [1,I], \\ -(\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \phi)_k = g_k^P \quad \forall k \in [1,K], \\ -(\boldsymbol{\nabla}_h^D \phi)_j \cdot \mathbf{n}_j = \sigma_j \quad \forall j \in [J - J^\Gamma + 1, J], \\ \sum_{i\in[1,I]} |T_i| \phi_i^T = \sum_{k\in[1,K]} |P_k| \phi_k^P = 0. \end{cases}$$

Find $(\psi_i^T, \psi_k^P)_{i\in[1,I+J^\Gamma],k\in[1,K]}$ *and* $(c_q^T, c_q^P)_{q\in[1,Q]}$

$$\begin{cases} (\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \psi)_i = f_i^T \quad \forall i \in [1,I], \\ (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k = f_k^P \quad \forall k \in [1, K - J^\Gamma], \\ (\boldsymbol{\nabla}_h^D \psi \cdot \mathbf{n}, 1)_{\Gamma_q} = k_q \quad \forall q \in [1,Q], \\ \sum_{k\in\Gamma_q} |P_k| \ (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k = \sum_{k\in\Gamma_q} |P_k| \ f_k^P \quad \forall q \in [1,Q], \\ \psi_i^T = \psi_k^P = 0 \quad \forall i \in \Gamma_0, \quad \forall k \in \Gamma_0, \\ \forall q \in [1,Q], \ \ \psi_i^T = c_q^T \quad \forall i \in \Gamma_q, \\ \forall q \in [1,Q], \ \ \psi_k^P = c_q^P \quad \forall k \in \Gamma_q. \end{cases}$$

The vector $\mathbf{u}$ *is then reconstructed by*

$$\mathbf{u}_j = (\boldsymbol{\nabla}_h^D \psi)_j + (\boldsymbol{\nabla}_h^D \times \phi)_j \quad \forall j \in [1,J].$$

5.3. Error estimate for the div-curl problem. Unlike in [20], we shall derive estimates for the potentials involved in the Hodge decomposition of $\mathbf{u}$; indeed we shall rely on similar estimates which have been obtained in [10]. For the sake of simplicity, we shall restrict ourselves to the case where all diamond-cells are convex; the case of nonconvex diamond-cells requires additional hypotheses similar to those given in [10]. We shall obtain error estimates under the following hypothesis (see Figures 2.5 and 5.1 for the notations).

Hypothesis 5.5. There exists an angle τ^*, strictly lower than π and independent of the mesh, such that the following hold:

1. For any *interior* diamond-cell D_j, the smallest in the maximum angle of the couple of triangles $(D_{j,1}, D_{j,2})$ *or* in the maximum angle of the couple of triangles $(D'_{j,1}, D'_{j,2})$ is bounded by τ^*:

$$\min\left(\max(\alpha_1, \beta_1, \mu_1 + \mu_2, \alpha_2, \beta_2, \nu_1 + \nu_2), \max(\mu_1, \nu_1, \alpha_1 + \alpha_2, \mu_2, \nu_2, \beta_1 + \beta_2)\right) \leq \tau^*.$$

2. The greatest angle of any boundary cell D_j is bounded by the angle τ^*.

Obtaining error estimates usually relies on regularity assumptions on the solution of the problem. In order to apply results given in [10], we shall assume regularity of

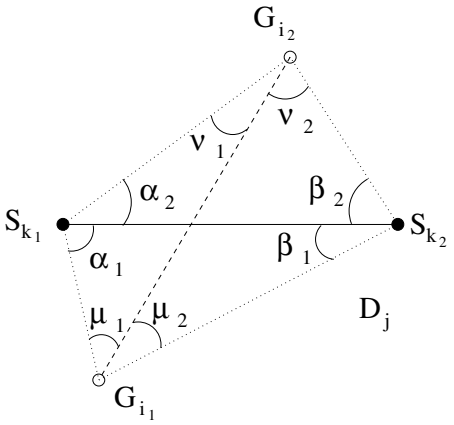


FIG. 5.1. *Notations for section* 5.3.

the potentials given by the following proposition.

PROPOSITION 5.6. *Let* (f, g, σ) *belong to* $\mathrm{L}^2(\Omega)^2 \times \mathrm{H}^{1/2}(\Gamma)$*, and let* $(k_q)_{q\in[1,Q]}$ *be a set of given real numbers; let* $\hat{\mathbf{u}}$ *be the exact solution of problem* (5.1)*. Then, there exist* $\hat{\phi}$ *and* $\hat{\psi}$ *both in* $\mathrm{H}^1(\Omega)$ *and a set of real numbers* $(C_q)_{q\in[1,Q]}$ *such that*

$$\hat{\mathbf{u}} = \boldsymbol{\nabla}\hat{\phi} + \boldsymbol{\nabla} \times \hat{\psi},$$

where $\hat{\phi}$ *is the solution of*

$$\begin{cases} \Delta\hat{\phi} = \nabla \cdot \hat{\mathbf{u}} = f \ \ in \ \ \Omega, \\ \boldsymbol{\nabla}\hat{\phi} \cdot \mathbf{n} = \hat{\mathbf{u}} \cdot \mathbf{n} = \sigma \ \ on \ \ \Gamma, \\ \int_\Omega \hat{\phi} = 0, \end{cases} \tag{5.15}$$

and $\hat{\psi}$ *is the solution of*

$$\begin{cases} -\Delta\hat{\psi} = \nabla \times \hat{\mathbf{u}} = g \ \ in \ \ \Omega, \\ \hat{\psi}_{|\Gamma_0} = 0; \ \ \hat{\psi}_{|\Gamma_q} = C_q \ \ \forall\, q \in [1, Q], \\ \int_{\Gamma_q} \boldsymbol{\nabla}\hat{\psi} \cdot n = -k_q. \end{cases} \tag{5.16}$$

Proof. The Hodge decomposition of $\hat{\mathbf{u}}$ and the determination of $\hat{\phi}$ and $\hat{\psi}$ through (5.15) and (5.16) are direct consequences of [11, Theorem 3.2 and Corollary 3.1]. ∎

Hypothesis 5.7. We suppose that the potentials $\hat{\phi}$ and $\hat{\psi}$ given by Proposition 5.6 belong to $\mathrm{H}^2(\Omega)$.

We remark that due to reentrant corners related to the internal polygonal boundaries Γ_q, the H^2 regularity of the potentials is not a consequence of the regularity of the data (f, g, σ).

Obviously, we may relate the L^2 error between the solution $\hat{\mathbf{u}}$ of (5.1) and the discrete solution $(\mathbf{u}_j)_{j\in[1,J]}$ of (5.3) to the errors between the solutions $\hat{\phi}$ and $\hat{\psi}$ of (5.15) and (5.16) and the discrete solutions (ϕ_i^T, ϕ_k^P) and (ψ_i^T, ψ_k^P) defined in Proposition 5.1,

respectively, by (5.7) and (5.8). Indeed we see that

$$\sum_{j\in[1,J]}\int_{D_j}|\mathbf{u}_j-\hat{\mathbf{u}}(\mathbf{x})|^2\,d\mathbf{x}\le 2\Bigg(\sum_{j\in[1,J]}\int_{D_j}\left|(\boldsymbol{\nabla}_h^D\phi)_j-\boldsymbol{\nabla}\hat{\phi}(\mathbf{x})\right|^2 d\mathbf{x} + \sum_{j\in[1,J]}\int_{D_j}\left|(\boldsymbol{\nabla}_h^D\psi)_j-\boldsymbol{\nabla}\hat{\psi}(\mathbf{x})\right|^2 d\mathbf{x}\Bigg). \tag{5.17}$$

5.3.1. Equivalent finite element formulations for the potentials. In order to evaluate the errors on the potentials, we follow [10] and rewrite (5.7) and (5.8) in terms of equivalent (nonconforming) finite element formulations. Recalling that the points $M_{i_\alpha(j)\,k_\beta(j)}$ are illustrated on Figure 2.4, we construct the following functions.

PROPOSITION 5.8. *Let $(\phi_i^T,\phi_k^P)\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K$ be given; there exists a function ϕ_h defined by*

$$(\phi_h)_{|D_j}\in P^1(D_j)\quad\forall j\in[1,J],$$

$$\phi_h(M_{i_\alpha(j)\,k_\beta(j)})=\frac{1}{2}(\phi^T_{i_\alpha(j)}+\phi^P_{k_\beta(j)})\quad\forall j\in[1,J],\quad\forall(\alpha\,,\,\beta)\in\{1;2\}^2\,. \tag{5.18}$$

Moreover, we have the following essential property:

$$(\boldsymbol{\nabla}\phi_h)_{|D_j}=(\boldsymbol{\nabla}_h^D\phi)_j\quad\forall j\in[1,J]. \tag{5.19}$$

Proof. The proof is given in [10]. We recall that the definition of ϕ_h through the four equalities contained in (5.18) is possible because $(M_{i_1k_1}M_{i_1k_2}M_{i_2k_2}M_{i_2k_1})$ is a parallelogram and $\phi_h(M_{i_1k_1})+\phi_h(M_{i_2k_2})=\phi_h(M_{i_1k_2})+\phi_h(M_{i_2k_1})$. ∎

DEFINITION 5.9. *We shall denote by L the linear operator which associates ϕ_h, defined by Proposition* 5.8, *with a given $(\phi_i^T,\phi_k^P)\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K$. Further, the solution of* (5.7) *is in the following space:*

$$V_N:=\left\{(\phi_i^T,\phi_k^P)\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K\ /\ \sum_{i\in[1,I]}|T_i|\phi_i^T=\sum_{k\in[1,K]}|P_k|\phi_k^P=0\right\}.$$

The solution of (5.8) *is in the following space:*

$$V_D:=\Big\{(\phi_i^T,\phi_k^P)\in\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K\ /\ \phi_i^T=\phi_k^P=0\ \forall\, i\in\Gamma_0\ \ \forall\, k\in\Gamma_0\ \ \textit{and}$$

$$\exists\,(c^T_{q,\phi},c^P_{q,\phi})\in(\mathbb{R}^2)^Q\ \ \textit{s.t.}\ \ \phi_i^T=c^T_{q,\phi}\ \forall\, i\in\Gamma_q,\ \textit{and}\ \phi_k^P=c^P_{q,\phi}\ \forall\, k\in\Gamma_q\ \ \forall\, q\in[1,Q]\Big\}.$$

Remark 5.10. It is easily proved that the linear operator L introduced in Definition 5.9 is injective over V_N and over V_D. Thus, for any Φ_h in $L(V_N)$ or in $L(V_D)$, there exists a unique $\Phi=(\Phi_i^T,\Phi_k^P)$ in $\mathbb{R}^{I+J^\Gamma}\times\mathbb{R}^K$, either in V_N or in V_D such that $\Phi_h=L(\Phi)$. The values (Φ_i^T,Φ_k^P) are used in the definitions of Φ_h^* and $\tilde{\Phi}_h$ associated with Φ_h, respectively, by (5.22) and (5.23).

With these definitions, we may state the following result.

PROPOSITION 5.11. *Problem* (5.7) *amounts to finding $\phi_h\in L(V_N)$, such that*

$$a_h(\phi_h,\Phi_h)=\ell_N(\Phi_h)\quad\forall\ \Phi_h\in L(V_N) \tag{5.20}$$

with

$$
\begin{aligned}
a_h(\phi_h, \Phi_h) &:= \sum_{j\in[1,J]} \int_{D_j} \boldsymbol{\nabla}\phi_h \cdot \boldsymbol{\nabla}\Phi_h\,(\mathbf{x})d\mathbf{x},\\
\ell_N(\Phi_h) &:= -\int_\Omega f\Phi_h^*(\mathbf{x})d\mathbf{x} \;+ \int_\Gamma \sigma\,\tilde{\Phi}_h(\xi)\,d\xi,
\end{aligned}
\tag{5.21}
$$

where Φ_h^* *is defined over* Ω *by*

$$
\Phi_h^*(\mathbf{x}) := \frac{1}{2}\left(\sum_{i\in[1,I]} \Phi_i^T\theta_i^T(\mathbf{x}) + \sum_{k\in[1,K]} \Phi_k^P\theta_k^P(\mathbf{x})\right)
\tag{5.22}
$$

and $\tilde{\Phi}_h$ *is defined over* Γ *by*

$$
\tilde{\Phi}_h(\xi) = \sum_{j\in[1,J]} \frac{1}{4}\left(\Phi^P_{k_1(j)} + 2\Phi^T_{i_2(j)} + \Phi^P_{k_2(j)}\right)\theta_j^\Gamma(\xi),
\tag{5.23}
$$

where we recall that θ_i^T, θ_k^P, *and* θ_j^Γ *are, respectively, the characteristic functions of the cells* T_i, P_k *and of the boundary edge* A_j.

Proof. Let us suppose that $\phi \in V_N$ is the solution of (5.7); then multiplying the first equation by $\frac{1}{2}|T_i|\Phi_i^T$, the second equation by $\frac{1}{2}|P_k|\Phi_k^P$, and summing over all $i \in [1, I]$ and all $k \in [1, K]$ yields

$$
(\nabla_h^{T,P}\cdot\boldsymbol{\nabla}_h^D\phi, \Phi)_{T,P} = (f, \Phi)_{T,P}.
\tag{5.24}
$$

Thanks to the discrete Green formula (4.1), we may write the left-hand side of (5.24) in the following way:

$$
\begin{aligned}
-(\boldsymbol{\nabla}_h^D\phi, \boldsymbol{\nabla}_h^D\Phi)_D + (\boldsymbol{\nabla}_h^D\phi\cdot\mathbf{n}, \Phi)_{\Gamma,h} \;&= -\sum_{j\in[1,J]} |D_j|\,(\boldsymbol{\nabla}_h^D\phi)_j\cdot(\boldsymbol{\nabla}_h^D\Phi)_j\\
&\quad + \sum_{j\in[J-J^\Gamma+1,J]} |A_j|\,(\boldsymbol{\nabla}_h^D\phi)_j\cdot\mathbf{n}_j \times \frac{1}{4}\left(\Phi^P_{k_1(j)} + 2\Phi^T_{i_2(j)} + \Phi^P_{k_2(j)}\right).
\end{aligned}
$$

Next, thanks to (5.19), and because $(\boldsymbol{\nabla}_h^D\phi)_j\cdot(\boldsymbol{\nabla}_h^D\Phi)_j$ is a constant over D_j, we may write

$$
-\sum_{j\in[1,J]} |D_j|\,(\boldsymbol{\nabla}_h^D\phi)_j\cdot(\boldsymbol{\nabla}_h^D\Phi)_j = -\sum_{j\in[1,J]}\int_{D_j}\boldsymbol{\nabla}\phi_h\cdot\boldsymbol{\nabla}\Phi_h\,(\mathbf{x})d\mathbf{x}.
$$

Moreover, according to the boundary conditions given by (5.7c),

$$
|A_j|\,(\boldsymbol{\nabla}_h^D\phi)_j\cdot\mathbf{n}_j = |A_j|\,\sigma_j = \int_{A_j}\sigma(\xi)d\xi,
$$

so that

$$
|A_j|\,(\boldsymbol{\nabla}_h^D\phi)_j\cdot\mathbf{n}_j \times \frac{1}{4}\left(\Phi^P_{k_1} + 2\Phi^T_{i_2} + \Phi^P_{k_2}\right) = \int_{A_j}\sigma\,\left(\tilde{\Phi}_h\right)_{|A_j}(\xi)\,d\xi.
$$

Finally, the left-hand side of (5.24) is equal to

$$-a_h(\phi_h, \Phi_h) + \int_\Gamma \sigma\, \tilde{\Phi}_h(\xi)\, d\xi.$$

By (5.4), and because $\Phi_i^T \theta_i^T(\mathbf{x})_{|T_i} = \Phi_i^T$ and $\Phi_k^P \theta_k^P(\mathbf{x})_{|P_k} = \Phi_k^P$, the right-hand side of (5.24) is equal to

$$\int_\Omega f(\mathbf{x})\, \frac{1}{2} \left(\sum_{i\in[1,I]} \Phi_i^T \theta_i^T(\mathbf{x}) + \sum_{k\in[1,K]} \Phi_k^P \theta_k^P(\mathbf{x}) \right) d\mathbf{x},$$

which ends this part of the proof.

Conversely, let $\phi_h \in L(V_N)$ satisfy (5.20) for all $\Phi_h \in L(V_N)$; then $\phi = L^{-1}(\phi_h)$ satisfies (5.7d) by definition of V_N. Further, we prove that the boundary condition (5.7c) is verified along each boundary edge $j_0 \in [J - J^\Gamma + 1, J]$ by considering its corresponding basis element $\Phi_0 \in V_N$ defined by (recall that the index $i_2(j_0)$ is associated with the unknown located at the center of the segment A_{j_0})

$$\forall i \in [1, I + J^\Gamma],\ (\Phi_0)_i^T = \delta_i^{i_2(j_0)} \ \text{ and } \ \forall k \in [1, K],\ (\Phi_0)_k^P = 0.$$

Then, defining $(\Phi_0)_h = L(\Phi_0)$, we obviously have the following properties:

$$(\boldsymbol{\nabla}(\Phi_0)_h)_{|D_j} = 0 \ \text{ if } \ j \neq j_0 \quad \text{and} \quad (\boldsymbol{\nabla}(\Phi_0)_h)_{|D_{j_0}} = \frac{1}{2\,|D_{j_0}|}\, |A_{j_0}|\, \mathbf{n}_{j_0}$$

and

$$(\Phi_0)_h^*(\mathbf{x}) = 0 \ \ \forall \mathbf{x} \in \Omega \quad \text{and} \quad (\tilde{\Phi}_0)_h(\xi) = \frac{1}{2} \theta_{j_0}^\Gamma(\xi) \ \ \forall \xi \in \Gamma.$$

We thus have

$$\sum_{j\in[1,J]} \int_{D_j} \boldsymbol{\nabla}\phi_h \cdot \boldsymbol{\nabla}(\Phi_0)_h\, (\mathbf{x}) d\mathbf{x} = \frac{1}{2}\, |A_{j_0}|\, (\boldsymbol{\nabla}\phi_h)_{|D_{j_0}} \cdot \mathbf{n}_{j_0} = \frac{1}{2}\, |A_{j_0}|\, (\boldsymbol{\nabla}_h^D \phi)_{j_0} \cdot \mathbf{n}_{j_0}$$

and

$$-\int_\Omega f(\Phi_0)_h^*(\mathbf{x}) d\mathbf{x} \ + \int_\Gamma \sigma\, (\tilde{\Phi}_0)_h(\xi)\, d\xi \ = \int_{A_{j_0}} \frac{1}{2} \sigma(\xi)\, d\xi \ = \frac{1}{2}\, |A_{j_0}|\, \sigma_{j_0}.$$

Finally, writing (5.20) for $(\Phi_0)_h$ proves that ϕ satisfies the boundary condition

$$(\boldsymbol{\nabla}_h^D \phi)_{j_0} \cdot \mathbf{n}_{j_0} = \sigma_{j_0} \quad \forall\, j_0 \in [J - J^\Gamma + 1, J].$$

Next, in order to prove (5.7a) for any primal cell $i_0 \in [1, I]$, we consider its corresponding basis element $\Phi_1 \in V_N$ defined by

$$\forall i \in [1, I + J^\Gamma],\ (\Phi_1)_i^T = \delta_i^{i_0} - \frac{|T_{i_0}|}{|\Omega|}, \ \text{ and } \ \forall k \in [1, K],\ (\Phi_1)_k^P = 0.$$

Then, defining $(\Phi_1)_h = L(\Phi_1)$ and according to (5.20), we may write

$$\sum_{j\in[1,J]} \int_{D_j} \boldsymbol{\nabla}\phi_h \cdot \boldsymbol{\nabla}(\Phi_1)_h\, (\mathbf{x}) d\mathbf{x} = -\int_\Omega f(\Phi_1)_h^*(\mathbf{x}) d\mathbf{x} \ + \int_\Gamma \sigma\, (\tilde{\Phi}_1)_h(\xi)\, d\xi. \tag{5.25}$$

To evaluate the left-hand side of (5.25), we consider $\Phi \in \mathbb{R}^{I+J^\Gamma} \times \mathbb{R}^K$ such that

$$\forall i \in [1, I + J^\Gamma],\ \Phi_i^T = \delta_i^{i_0} \ \text{ and } \ \forall k \in [1, K],\ \Phi_k^P = 0.$$

Note that $\Phi \notin V_N$ but that its discrete gradient (see (3.2)) obviously equals that of Φ_1. Thanks to this equality and to (5.19), we have

$$\sum_{j\in[1,J]} \int_{D_j} \boldsymbol{\nabla}\phi_h \cdot \boldsymbol{\nabla}(\Phi_1)_h\,(\mathbf{x})d\mathbf{x} = (\boldsymbol{\nabla}_h^D\phi, \boldsymbol{\nabla}_h^D\Phi_1)_D = (\boldsymbol{\nabla}_h^D\phi, \boldsymbol{\nabla}_h^D\Phi)_D,$$

which, in turn, can be transformed, thanks to (4.1), into

$$-(\nabla_h^{T,P} \cdot \boldsymbol{\nabla}_h^D\phi, \Phi)_{T,P} + (\boldsymbol{\nabla}_h^D\phi \cdot \mathbf{n}, \Phi)_{\Gamma,h}.$$

Thanks to the definition of Φ, this quantity reduces to the contribution of i_0, which proves that the left-hand side of (5.25) may be written

$$\sum_{j\in[1,J]} \int_{D_j} \boldsymbol{\nabla}\phi_h \cdot \boldsymbol{\nabla}(\Phi_1)_h\,(\mathbf{x})d\mathbf{x} = -\frac{1}{2}\,|T_{i_0}|\,(\nabla_h^T \cdot \boldsymbol{\nabla}_h^D\phi)_{i_0}. \tag{5.26}$$

Next, we compute the right-hand side of (5.25)

$$\begin{aligned}
-\int_\Omega f(\Phi_1)_h^*(\mathbf{x})d\mathbf{x} &= -\frac{1}{2}\sum_{i\in[1,I]} \int_{T_i} \left(\delta_i^{i_0} - \frac{|T_{i_0}|}{|\Omega|}\right) f(\mathbf{x})\,d\mathbf{x} \\
&= -\frac{1}{2}\int_{T_{i_0}} f(\mathbf{x})\,d\mathbf{x} + \frac{1}{2}\frac{|T_{i_0}|}{|\Omega|}\int_\Omega f(\mathbf{x})\,d\mathbf{x};
\end{aligned}$$

$$\int_\Gamma \sigma\,(\tilde{\Phi}_1)_h(\xi)\,d\xi = \sum_{j\in[J-J^\Gamma+1,J]} \int_{A_j} \sigma(\xi)\,\frac{1}{4}\left(-2\frac{|T_{i_0}|}{|\Omega|}\right)\,d\xi = -\frac{1}{2}\frac{|T_{i_0}|}{|\Omega|}\int_\Gamma \sigma(\xi)d\xi,$$

so that the right-hand side of (5.25) equals

$$-\frac{1}{2}\int_{T_{i_0}} f(\mathbf{x})\,d\mathbf{x} + \frac{1}{2}\frac{|T_{i_0}|}{|\Omega|}\int_\Omega f(\mathbf{x})\,d\mathbf{x} - \frac{1}{2}\frac{|T_{i_0}|}{|\Omega|}\int_\Gamma \sigma(\xi)d\xi.$$

Because of (5.2), the last two terms in the previous sum cancel and we get

$$-\int_\Omega f(\Phi_1)_h^*(\mathbf{x})d\mathbf{x} + \int_\Gamma \sigma\,(\tilde{\Phi}_1)_h(\xi)\,d\xi = -\frac{1}{2}\int_{T_{i_0}} f(\mathbf{x})\,d\mathbf{x} = -\frac{1}{2}\,|T_{i_0}|\,f_{i_0}^T. \tag{5.27}$$

Comparing (5.25), (5.26), and (5.27), we infer that

$$(\nabla_h^T \cdot \boldsymbol{\nabla}_h^D\phi)_{i_0} = f_{i_0}^T.$$

In a similar way, we can prove (5.7b) for any dual cell $k_0 \in [1, K]$ by considering its corresponding basis element $\Phi_2 \in V_N$, defined by

$$\forall i \in [1, I + J^\Gamma],\ (\Phi_2)_i^T = 0 \ \text{ and } \ \forall k \in [1, K],\ (\Phi_2)_k^P = \delta_k^{k_0} - \frac{|P_{k_0}|}{|\Omega|},$$

which ends the proof of the equivalence. ∎

PROPOSITION 5.12. *Problem* (5.8) *is equivalent to finding* $\psi_h \in L(V_D)$*, such that* $\forall \Psi_h \in L(V_D)$*,*

$$a_h(\psi_h, \Psi_h) = \ell_D(\Psi_h) \tag{5.28}$$

with

$$\ell_D(\Psi_h) := \int_\Omega g \Psi_h^*(\mathbf{x}) d\mathbf{x} \; - \sum_{q \in [1,Q]} k_q \left(\frac{c_{q,\Psi}^T + c_{q,\Psi}^P}{2} \right).$$

Proof. Let us suppose that $\psi \in V_D$ is the solution of (5.8); then we may compute the following discrete scalar product:

$$\begin{aligned} -(\nabla_h^{T,P} \cdot \boldsymbol{\nabla}_h^D \psi, \Psi)_{T,P} = & -\frac{1}{2} \sum_{i \in [1,I]} |T_i| (\nabla_h^T \cdot \boldsymbol{\nabla}_h^D \psi)_i \Psi_i^T \\ & -\frac{1}{2} \sum_{k \in [1, K - J^\Gamma]} |P_k| (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k \Psi_k^P \\ & -\frac{1}{2} \sum_{k \in [K - J^\Gamma + 1, K]} |P_k| (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k \Psi_k^P. \end{aligned} \tag{5.29}$$

Due to (5.8a)–(5.8b), the sum of the first two terms on the right-hand side of (5.29) equals

$$\frac{1}{2} \sum_{i \in [1,I]} |T_i| g_i^T \Psi_i^T + \frac{1}{2} \sum_{k \in [1, K - J^\Gamma]} |P_k| g_k^P \Psi_k^P.$$

Next, using the fact that Ψ^P is equal to a constant $c_{q,\Psi}^P$ over each Γ_q and vanishes over Γ_0, we may write, according to (5.8d),

$$\begin{aligned} -\sum_{k \in [K - J^\Gamma + 1, K]} |P_k| (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k \Psi_k^P &= -\sum_{q \in [1,Q]} c_{q,\Psi}^P \sum_{k \in \Gamma_q} |P_k| (\nabla_h^P \cdot \boldsymbol{\nabla}_h^D \psi)_k \\ &= \sum_{q \in [1,Q]} c_{q,\Psi}^P \sum_{k \in \Gamma_q} |P_k| g_k^P = \sum_{k \in [K - J^\Gamma + 1, K]} |P_k| g_k^P \Psi_k^P. \end{aligned}$$

Finally, (5.29) may be rewritten in the following way:

$$-(\nabla_h^{T,P} \cdot \boldsymbol{\nabla}_h^D \psi, \Psi)_{T,P} = (g, \Psi)_{T,P}. \tag{5.30}$$

Using the discrete Green formula (4.1), the left-hand side of (5.30) is equal to

$$(\boldsymbol{\nabla}_h^D \psi, \boldsymbol{\nabla}_h^D \Psi)_D - (\boldsymbol{\nabla}_h^D \psi \cdot \mathbf{n}, \Psi)_{\Gamma,h}.$$

Like previously, the first of these terms equals $a_h(\psi_h, \Psi_h)$. Next, using the fact that Ψ^P (resp., Ψ^T) is equal to a constant $c_{q,\Psi}^P$ (resp., $c_{q,\Psi}^T$) over each Γ_q and vanishes over Γ_0, and using (5.8c), there holds

$$(\boldsymbol{\nabla}_h^D \psi \cdot \mathbf{n}, \Psi)_{\Gamma,h} = \sum_{q \in [1,Q]} \left(\frac{c_{q,\Psi}^T + c_{q,\Psi}^P}{2} \right) \sum_{\Gamma_q} (\boldsymbol{\nabla}_h^D \psi)_j \cdot \mathbf{n}_j = -\sum_{q \in [1,Q]} k_q \left(\frac{c_{q,\Psi}^T + c_{q,\Psi}^P}{2} \right),$$

which shows that the left-hand side of (5.30) is equal to

$$a_h(\psi_h, \Psi_h) + \sum_{q\in[1,Q]} k_q \left(\frac{c^T_{q,\Psi} + c^P_{q,\Psi}}{2} \right).$$

This ends the first part of the proof.

Conversely, if $\psi_h \in L(V_D)$ satisfies (5.28) for all $\Psi_h \in L(V_D)$, then $\psi = L^{-1}(\psi_h)$ verifies (5.8e), (5.8f), and (5.8g) by definition of V_D. Next, in order to prove (5.8a) for any primal cell $i_0 \in [1, I]$, let us consider its associated basis element $\Psi_1 \in V_D$ defined through

$$(\Psi_1)^T_i = \delta^{i_0}_i \quad \forall\, i \in [1, I + J^\Gamma] \text{ and } (\Psi_1)^P_k = 0 \quad \forall\, k \in [1, K].$$

Applying (5.28) for $\Psi_h = L(\Psi_1)$ and using (5.19), (4.1), and (5.5) shows that (5.8a) is verified for the considered $i_0 \in [1, I]$. Equality (5.8b) can be proved in the same way for any dual cell $k_0 \in [1, K - J^\Gamma]$ by considering its associated basis element $\Psi_2 \in V_D$ defined through

$$(\Psi_2)^T_i = 0 \quad \forall\, i \in [1, I + J^\Gamma] \text{ and } (\Psi_2)^P_k = \delta^{k_0}_k \quad \forall\, k \in [1, K].$$

Next, let us consider an internal boundary Γ_{q_0} with $q_0 \in [1, Q]$ and let us consider $\Psi_3 \in V_D$ which vanishes everywhere but on Γ_{q_0}, where it has a constant value:

$$(\Psi_3)^T_i = (\Psi_3)^P_k = 0\, \forall\, i \in [1, I],\ \forall\, k \in [1, K] \text{ and } (\Psi_3)^T_i = \delta^{q_0}_q\ \forall\, i \in \Gamma_q,\ \forall\, q \in [0, Q].$$

Applying (5.28) for $\Psi_h = L(\Psi_3)$ and using (5.19) and (4.1) shows that (5.8c) is verified for the considered $q_0 \in [1, Q]$. In the same way, we prove (5.8d) for a given $q_0 \in [1, Q]$ by choosing $\Psi_4 \in V_D$ defined through

$$(\Psi_4)^T_i = 0 \quad \forall\, i \in [1, I],\ (\Psi_4)^P_k = 0 \quad \forall\, k \in [1, K - J^\Gamma],$$

$$(\Psi_4)^T_i = \delta^{q_0}_q \quad \forall\, i \in \Gamma_q \text{ and } (\Psi_4)^P_k = -\delta^{q_0}_q \quad \forall\, k \in \Gamma_q,\ \forall\, q \in [0, Q].$$

This ends the proof of Proposition 5.12. ∎

5.3.2. Error estimates for the potentials. We may now turn to error estimates for the potentials $\hat{\phi}$ and $\hat{\psi}$. First, given the equivalent finite element formulation stated by Proposition 5.11 (resp., Proposition 5.12), we may study the numerical error concerning $\hat{\phi}$ (resp., $\hat{\psi}$) in a traditional way by noting that a_h acts on $\mathrm{H}^1 + L(V_N)$ (resp., $\mathrm{H}^1 + L(V_D)$), on which we define $|x|_{1,h} := \sqrt{a_h(x,x)}$, and by using the so-called "Strang second lemma" [24]:

$$|\hat{\phi} - \phi_h|_{1,h} \le 2 \inf_{\omega_h \in L(V_N)} |\hat{\phi} - \omega_h|_{1,h} + \sup_{\omega_h \in L(V_N)} \frac{|a_h(\hat{\phi}, \omega_h) - \ell_N(\omega_h)|}{|\omega_h|_{1,h}} \tag{5.31}$$

and

$$|\hat{\psi} - \psi_h|_{1,h} \le 2 \inf_{\omega_h \in L(V_D)} |\hat{\psi} - \omega_h|_{1,h} + \sup_{\omega_h \in L(V_D)} \frac{|a_h(\hat{\psi}, \omega_h) - \ell_D(\omega_h)|}{|\omega_h|_{1,h}}\,. \tag{5.32}$$

The first term in (5.31) and (5.32) is named the "interpolation error," while the second is called the "consistency error."

Interpolation error for $\hat{\phi}$. We start with the following proposition.

PROPOSITION 5.13. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *there exists a constant $C(\tau^*)$ depending only on τ^* such that*

$$\inf_{\omega_h \in L(V_N)} |\hat{\phi} - \omega_h|_{1,h} \leq C(\tau^*)\, h\, ||\hat{\phi}||_{2,\Omega}\,. \tag{5.33}$$

Proof. Consider the pointwise projection of the exact solution onto $\mathbb{R}^{I+J^\Gamma} \times \mathbb{R}^K$:

$$\forall i \in [1, I + J^\Gamma]\,,\ \ (\Pi\hat{\phi})_i^T = \hat{\phi}(G_i) \quad \text{and} \quad \forall k \in [1, K]\,,\ \ (\Pi\hat{\phi})_k^P = \hat{\phi}(S_k)\,.$$

Then, this element is itself projected onto V_N in the following way:

$$\forall i \in [1, I + J^\Gamma],\ \ (\tilde{\Pi}\hat{\phi})_i^T = (\Pi\hat{\phi})_i^T - \frac{\displaystyle\sum_{i\in[1,I]} |T_i|(\Pi\hat{\phi})_i^T}{|\Omega|}$$

$$\forall k \in [1, K],\ \ (\tilde{\Pi}\hat{\phi})_k^P = (\Pi\hat{\phi})_k^P - \frac{\displaystyle\sum_{k\in[1,K]} |P_k|(\Pi\hat{\phi})_k^P}{|\Omega|}.$$

Obviously, $\tilde{\Pi}\hat{\phi}$ and $\Pi\hat{\phi}$ have the same discrete gradient so that the interpolation error in (5.33) is bounded in the following way:

$$\inf_{\omega_h \in L(V_N)} |\hat{\phi} - \omega_h|_{1,h} \leq |\hat{\phi} - L(\tilde{\Pi}\hat{\phi})|_{1,h} = |\hat{\phi} - L(\Pi\hat{\phi})|_{1,h}.$$

Finally, an upper bound for $|\hat{\phi} - L(\Pi\hat{\phi})|_{1,h}$ has been given in [10] and is based on the relation between $L(\Pi\hat{\phi})$ and the standard Lagrange P^1 interpolants on the pairs $(D_{j,1}, D_{j,2})$ and $(D'_{j,1}, D'_{j,2})$. It leads to the estimation (5.33). Hypothesis 5.5 is here to ensure that the so-called maximum angle condition [3, 18] is verified for at least one of the pairs of triangles $(D_{j,1}, D_{j,2})$ or $(D'_{j,1}, D'_{j,2})$. ∎

Consistency error for $\hat{\phi}$. Let $\omega_h = L(\omega)$. Thanks to (5.21), we start by writing

$$a_h(\hat{\phi}, \omega_h) - \ell_N(\omega_h) = \left[a_h(\hat{\phi}, \omega_h) + (f, \omega_h)_\Omega - \int_\Gamma \sigma\, \tilde{\omega}_h(\xi)\, d\xi\right] - (f, \omega_h - \omega_h^*)_\Omega. \tag{5.34}$$

The last term in (5.34) can be bounded by the following lemma.

LEMMA 5.14. *If all diamond-cells are convex, there exists a constant C independent of the grid such that*

$$|(f, \omega_h - \omega_h^*)_\Omega| \leq Ch||f||_{0,\Omega}|\omega_h|_{1,h}. \tag{5.35}$$

Proof. The proof is identical to that given in [10] for homogeneous Dirichlet conditions. ∎

Then, we follow [10] with a slight modification due to nonhomogeneous Neumann boundary conditions. We divide each *interior* diamond-cell D_j (with $j \in [1, J - J^\Gamma]$) either into $D_{j,1} \cup D_{j,2}$, or into $D'_{j,1} \cup D'_{j,2}$ (see Figure 2.5). Note that this choice is local to D_j and does not influence the choice which can be made for the division of $D_{j'}$, for $j' \neq j$. Boundary diamond-cells are such that $D_{j,1} = D_j$ and $D_{j,2} = \emptyset$ and will never be split into $D'_{j,1} \cup D'_{j,2}$. To simplify notations, we shall write $\mathcal{T}_{j,\alpha}$

to represent either $D_{j,\alpha}$ or $D'_{j,\alpha}$. Further, we define $RT(\boldsymbol{\nabla}\hat{\phi})$, the Raviart–Thomas interpolation of $\boldsymbol{\nabla}\hat{\phi}$ on each $\mathcal{T}_{j,\alpha}$ (see [23]), by

$$RT(\boldsymbol{\nabla}\hat{\phi})_{|\mathcal{T}_{j,\alpha}} \in (P_0(\mathcal{T}_{j,\alpha}))^2 \oplus \begin{pmatrix} x \\ y \end{pmatrix} P_0(\mathcal{T}_{j,\alpha}) \quad \text{and} \quad \int_s RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n}\, d\xi = \int_s \boldsymbol{\nabla}\hat{\phi}.\mathbf{n}\, d\xi$$

for any edge s of $\mathcal{T}_{j,\alpha}$ whose normal exterior unit vector is denoted by $\mathbf{n}$. We can state the following lemma.

LEMMA 5.15. *Let $\hat{\phi}$ be the solution of* (5.15) *and let $\omega_h \in L(V_N)$. Denote by $\langle \omega_h \rangle_{j,\alpha}$ the average value of ω_h over $\mathcal{T}_{j,\alpha}$. Then, if all diamond-cells are convex, we have*

$$a_h(\hat{\phi}, \omega_h) + (f, \omega_h)_\Omega - \int_\Gamma \sigma\, \tilde{\omega}_h(\xi)\, d\xi = \sum_{j\in[1,J]} \sum_{\alpha=1}^{2} \int_{\mathcal{T}_{j,\alpha}} \left[(\boldsymbol{\nabla}\hat{\phi} - RT(\boldsymbol{\nabla}\hat{\phi})) \cdot \boldsymbol{\nabla}\omega_h - f \left(\langle \omega_h \rangle_{j,\alpha} - \omega_h \right) \right] d\mathbf{x}. \tag{5.36}$$

Proof. By definition, $RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n}$ is a constant on each edge of $\mathcal{T}_{j,\alpha}$. In addition, on two neighboring triangles $\mathcal{T}_{j,\alpha}$, the values of $RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n}$ on both sides of their common edge are opposite of each other, because of the orientation of the normal vector $\mathbf{n}$. By noting $\mathcal{S}$ the set of all the edges of all the $\mathcal{T}_{j,\alpha}$, $\mathbf{n}$ the normal unit vector to an edge s in $\mathcal{S}$, and $[\omega_h]_s$ the jump of ω_h through s, then

$$\begin{aligned} \sum_{j\in[1,J]} \sum_{\alpha=1}^{2} \int_{\partial\mathcal{T}_{j,\alpha}} RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n}\, \omega_h\, d\xi &= \sum_{s\in\mathcal{S},\ s\not\subset\Gamma} RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n} \int_s [\omega_h]_s\, d\xi \\ &\quad + \sum_{s\in\mathcal{S},\ s\subset\Gamma} RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n} \int_s \omega_h\, d\xi. \end{aligned} \tag{5.37}$$

Since ω_h is in $L(V_N)$, then $[\omega_h]_s$ is a polynomial of degree one, which vanishes at the midpoint of s (by construction of the functions of $L(V_N)$). Its integral on s is thus null. Further, there is an obvious one-to-one correspondence between a given $s \in \mathcal{S},\ s \subset \Gamma$, and some boundary edge A_j, with $j \in [J - J^\Gamma + 1, J]$ because boundary diamond-cells are such that $D_j = D_{j,1} = \mathcal{T}_{j,\alpha}$, with $\alpha = 1$. Therefore, for such $s \in \mathcal{S},\ s \subset \Gamma$, there exists a unique $j \in [J - J^\Gamma + 1, J]$ such that

$$RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n} = \frac{1}{|A_j|} \int_{A_j} RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n}_j = \frac{1}{|A_j|} \int_{A_j} \boldsymbol{\nabla}\hat{\phi} \cdot \mathbf{n}_j = \frac{1}{|A_j|} \int_{A_j} \sigma(\xi)\, d\xi.$$

Further, on this A_j, the function ω_h is a polynomial of degree one, whose integral is easy to compute:

$$\int_s \omega_h\, d\xi = \frac{|A_j|}{4} \left(\omega_{k_1}^P + 2\omega_{i_2}^T + \omega_{k_2}^P \right).$$

Recalling the definition (5.23) of the piecewise constant function $\tilde{\omega}_h$, we may write

$$\sum_{j\in[1,J]} \sum_{\alpha=1}^{2} \int_{\partial\mathcal{T}_{j,\alpha}} RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n}\, \omega_h\, d\xi = \sum_{s\in\mathcal{S},\ s\subset\Gamma} RT(\boldsymbol{\nabla}\hat{\phi}) \cdot \mathbf{n} \int_s \omega_h\, d\xi = \int_\Gamma \sigma\tilde{\omega}_h(\xi) d\xi.$$

But we may also write the previous equality in the following way:

$$\sum_{j\in[1,J]}\sum_{\alpha=1}^{2}\left(\int_{\mathcal{T}_{j,\alpha}}\nabla\cdot(RT(\boldsymbol{\nabla}\hat{\phi}))\,\omega_h\,d\mathbf{x}+\int_{\mathcal{T}_{j,\alpha}}RT(\boldsymbol{\nabla}\hat{\phi})\cdot\boldsymbol{\nabla}\omega_h\,d\mathbf{x}\right)=\int_\Gamma\sigma\tilde{\omega}_h(\xi)d\xi.$$

By subtracting this equality from $a_h(\hat{\phi},\omega_h)$, we obtain

$$\begin{aligned}a_h(\hat{\phi},\omega_h)-\int_\Gamma\sigma\tilde{\omega}_h(\xi)d\xi=&\sum_{j\in[1,J]}\sum_{\alpha=1}^{2}\int_{\mathcal{T}_{j,\alpha}}(\boldsymbol{\nabla}\hat{\phi}-RT(\boldsymbol{\nabla}\hat{\phi}))\cdot\boldsymbol{\nabla}\omega_h\,d\mathbf{x}\\&-\sum_{j\in[1,J]}\sum_{\alpha=1}^{2}\int_{\mathcal{T}_{j,\alpha}}\nabla\cdot(RT(\boldsymbol{\nabla}\hat{\phi}))\,\omega_h\,d\mathbf{x}.\end{aligned}\tag{5.38}$$

Let us note $\langle\omega_h\rangle_{j,\alpha}$ the mean value of ω_h on $\mathcal{T}_{j,\alpha}$. Since $\nabla\cdot(RT(\boldsymbol{\nabla}\hat{\phi}))$ is by construction a constant on $\mathcal{T}_{j,\alpha}$, we may write the following series of equalities:

$$\begin{aligned}\int_{\mathcal{T}_{j,\alpha}}\nabla\cdot(RT(\boldsymbol{\nabla}\hat{\phi}))\,\omega_h\,d\mathbf{x}&=\langle\omega_h\rangle_{j,\alpha}\int_{\mathcal{T}_{j,\alpha}}\nabla\cdot(RT(\boldsymbol{\nabla}\hat{\phi}))\,d\mathbf{x}\\&=\langle\omega_h\rangle_{j,\alpha}\int_{\partial\mathcal{T}_{j,\alpha}}RT(\boldsymbol{\nabla}\hat{\phi})\cdot\mathbf{n}\,d\xi=\langle\omega_h\rangle_{j,\alpha}\int_{\partial\mathcal{T}_{j,\alpha}}\boldsymbol{\nabla}\hat{\phi}\cdot\mathbf{n}\,d\xi\\&=\langle\omega_h\rangle_{j,\alpha}\int_{\mathcal{T}_{j,\alpha}}\Delta\hat{\phi}\,d\mathbf{x}=\langle\omega_h\rangle_{j,\alpha}\int_{\mathcal{T}_{j,\alpha}}f\,d\mathbf{x}.\end{aligned}\tag{5.39}$$

Equality (5.36) follows from (5.38) and (5.39). ∎

The first term in the right-hand side of (5.34) can be bounded by the following lemma.

LEMMA 5.16. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *there exists a constant C independent of the grid such that*

$$\left|a_h(\hat{\phi},\omega_h)+(f,\omega_h)_\Omega-\int_\Gamma\sigma\tilde{\omega}_h(\xi)\,d\xi\right|\leq C\frac{h}{\sin\tau^*}|\omega_h|_{1,h}\left(||f||_{0,\Omega}+||\hat{\phi}||_{2,\Omega}\right).\tag{5.40}$$

Proof. By virtue of Lemma 5.15, bounding the left-hand side of (5.40) amounts to bounding the right-hand side of (5.36). This was performed in [10]. Again, Hypothesis 5.5 is here to ensure the maximum angle condition needed by the Raviart–Thomas interpolation of $\boldsymbol{\nabla}\hat{\phi}$; see [1]. ∎

We end the consistency error estimation with the following proposition.

PROPOSITION 5.17. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *then there exists a constant C independent of the grid such that*

$$\sup_{\omega_h\in L(V_N)}\frac{|a_h(\hat{\phi},\omega_h)-\ell_N(\omega_h)|}{|\omega_h|_{1,h}}\leq C\frac{h}{\sin\tau^*}\left(||f||_{0,\Omega}+||\hat{\phi}||_{2,\Omega}\right).\tag{5.41}$$

Proof. The result follows from (5.34), (5.35), and (5.40). ∎

Interpolation error for $\hat{\psi}$. Next, given the equivalent finite element formulation stated by Proposition 5.12, we may study the numerical error concerning ψ in a very analogous way: The interpolation error is bounded by choosing $\omega_h = L(\Pi\hat{\psi})$ with $\Pi\hat{\psi} \in V_D$ defined by

$$\forall i \in [1, I + J^\Gamma]\,, \;\; (\Pi\hat{\psi})_i^T = \hat{\psi}(G_i) \quad \text{and} \quad \forall k \in [1, K]\,, \;\; (\Pi\hat{\psi})_k^P = \hat{\psi}(S_k)$$

and we obtain a result analogous to (5.33).

PROPOSITION 5.18. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *then there exists a constant $C(\tau^*)$ depending only on τ^* such that*

$$\inf_{\omega_h \in L(V_D)} |\hat{\psi} - \omega_h|_{1,h} \leq C(\tau^*)\, h\, ||\hat{\psi}||_{2,\Omega}. \tag{5.42}$$

Consistency error for $\hat{\psi}$. Concerning the consistency error, we may prove a result analogous to (5.36).

LEMMA 5.19. *Let $\hat{\psi}$ be the solution of* (5.16) *and let $\omega_h \in L(V_D)$. Then, if all diamond-cells are convex, we have*

$$\begin{aligned} &a_h(\hat{\psi}, \omega_h) - (g, \omega_h)_\Omega + \sum_{q \in [1,Q]} k_q \left(\frac{c_{q,\omega}^T + c_{q,\omega}^P}{2} \right) \\ &\qquad = \sum_{j \in [1,J]} \sum_{\alpha=1}^{2} \int_{\mathcal{T}_{j,\alpha}} \left[(\boldsymbol{\nabla}\hat{\psi} - RT(\boldsymbol{\nabla}\hat{\psi})) \cdot \boldsymbol{\nabla}\omega_h + g \left(\langle \omega_h \rangle_{j,\alpha} - \omega_h \right) \right] d\mathbf{x}. \end{aligned} \tag{5.43}$$

Proof. We first write for $\hat{\psi}$ an equality analogous to (5.37). For the same reasons as in the proof of Lemma 5.15, this amounts to evaluating the boundary part:

$$\sum_{j \in [1,J]} \sum_{\alpha=1}^{2} \int_{\partial \mathcal{T}_{j,\alpha}} RT(\boldsymbol{\nabla}\hat{\psi}) \cdot \mathbf{n}\, \omega_h \, d\xi = \sum_{q \in [1,Q]} \sum_{j \in \Gamma_q} RT(\boldsymbol{\nabla}\hat{\psi}) \cdot \mathbf{n}_j \int_{A_j} \omega_h \, d\xi$$

$$= \sum_{q \in [1,Q]} \left(\frac{c_{q,\omega}^T + c_{q,\omega}^P}{2} \right) \sum_{j \in \Gamma_q} |A_j| RT(\boldsymbol{\nabla}\hat{\psi}) \cdot \mathbf{n}_j = \sum_{q \in [1,Q]} \left(\frac{c_{q,\omega}^T + c_{q,\omega}^P}{2} \right) \sum_{j \in \Gamma_q} \int_{A_j} \boldsymbol{\nabla}\hat{\psi} \cdot \mathbf{n}_j$$

$$= \sum_{q \in [1,Q]} \left(\frac{c_{q,\omega}^T + c_{q,\omega}^P}{2} \right) \int_{\Gamma_q} \boldsymbol{\nabla}\hat{\psi} \cdot \mathbf{n}_j = - \sum_{q \in [1,Q]} k_q \left(\frac{c_{q,\omega}^T + c_{q,\omega}^P}{2} \right).$$

The end of the proof of (5.43) follows exactly the same lines as that of (5.36) and is thus skipped. ∎

Next, bounding the right-hand side of (5.43) is performed like in [10], and we obtain a result analogous to (5.41)

PROPOSITION 5.20. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *then there exists a constant C independent of the grid such that*

$$\sup_{\omega_h \in L(V_D)} \frac{|a_h(\hat{\psi}, \omega_h) - \ell_D(\omega_h)|}{|\omega_h|_{1,h}} \leq C \frac{h}{\sin \tau^*} \left(||g||_{0,\Omega} + ||\hat{\psi}||_{2,\Omega} \right). \tag{5.44}$$

To conclude subsection 5.3.2, estimates (5.31), (5.33), and (5.41) on the one hand and (5.32), (5.42), and (5.44) on the other hand allow us to state the following theorem.

THEOREM 5.21. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *then there exists a constant* $C(\tau^*)$ *independent of the grid such that*

$$|\hat{\phi} - \phi_h|_{1,h} \leq C(\tau^*)h \left(||f||_{0,\Omega} + ||\hat{\phi}||_{2,\Omega} \right) \tag{5.45}$$

and

$$|\hat{\psi} - \psi_h|_{1,h} \leq C(\tau^*)h \left(||g||_{0,\Omega} + ||\hat{\psi}||_{2,\Omega} \right). \tag{5.46}$$

To conclude section 5.3, Theorem 5.21, along with (5.17) and (5.19), leads to the following theorem.

THEOREM 5.22. *If all diamond-cells are convex and under Hypotheses* 5.5 *and* 5.7, *then there exists a constant* $C(\tau^*)$ *independent of the grid such that*

$$\left(\sum_{j \in [1,J]} \int_{D_j} |\mathbf{u}_j - \hat{\mathbf{u}}(\mathbf{x})|^2 \, d\mathbf{x} \right)^{1/2} \leq C(\tau^*)h \left(||f||_{0,\Omega} + ||g||_{0,\Omega} + ||\hat{\phi}||_{2,\Omega} + ||\hat{\psi}||_{2,\Omega} \right).$$

6. Numerical results. We test the finite volume method over different types of meshes and define the discrete relative L^2 error by

$$e^2(h) := \frac{\sum_j |D_j| \, |\mathbf{u} - \Pi\hat{\mathbf{u}}|_j^2}{\sum_j |D_j| \, |\Pi\hat{\mathbf{u}}|_j^2},$$

where $(\Pi\hat{\mathbf{u}})_j$ is the value of the exact solution at the barycenter of D_j (noted B_j):

$$\forall \, j \in [1, J], (\Pi\hat{\mathbf{u}})_j = \hat{\mathbf{u}}(B_j).$$

For the first three families of meshes (triangular unstructured, nonconforming, and degenerating triangular), the domain of computation is the unit square $\Omega = [0;1] \times [0;1]$. We choose the data f, g and the boundary conditions so that the analytical solution is given by

$$\hat{\mathbf{u}}(x, y) = \begin{pmatrix} \exp(x) \, \cos(\pi y) + \pi \, \sin(\pi x) \, \cos(\pi y) \\ -\pi \, \exp(x) \, \sin(\pi y) - \pi \, \cos(\pi x) \, \sin(\pi y) \end{pmatrix}.$$

This means that the exact potentials are given by

$$\hat{\phi}(x, y) = \exp(x) \, \cos(\pi y) \quad \text{and} \quad \hat{\psi}(x, y) = \sin(\pi x) \, \sin(\pi y).$$

In addition, we always choose the points G_i associated with the control volumes of the primal mesh to be the barycenters of the cell T_i.

6.1. Unstructured meshes. First of all, we consider a family of six unstructured grids made up of increasingly small triangles. The first two of these grids are represented on the left and central parts of Figure 6.1. The numerical errors in the discrete L^2 norm are presented in logarithmic scale on the right part of Figure 6.1, on which we also plotted a straight line of slope 1. We remark, as proved previously, a first order convergence of the presented scheme.

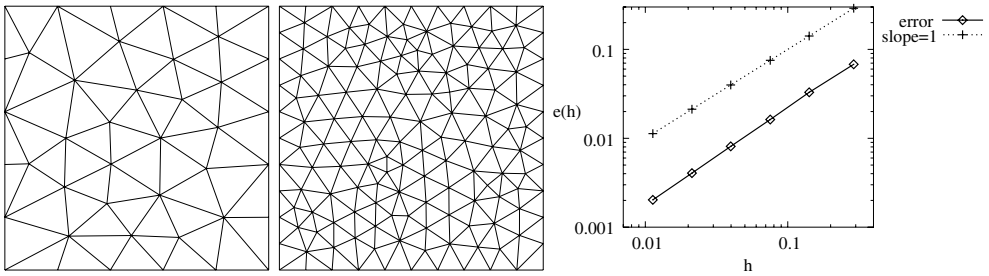


FIG. 6.1. *Unstructured triangular meshes.*

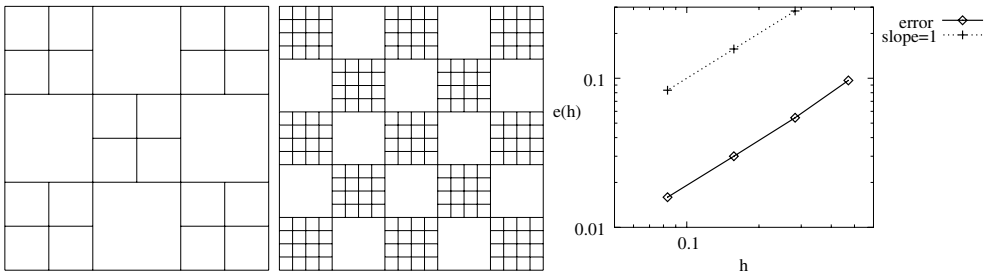


FIG. 6.2. *Nonconforming square meshes.*

6.2. Nonconforming meshes. Next, we consider the nonconforming family of meshes constructed in the following way. Let n be a nonzero integer. We split Ω into $(2^n+1)\times(2^n+1)$ identical squares. Then, every other square is itself divided into $2^n\times 2^n$ identical subsquares. We choose $n\in[1;4]_N$. The left and central parts of Figure 6.2 display the first two of these meshes. Of course, this family of meshes is not of practical use, but constitutes, in our opinion, a good choice in order to test the applicability of the presented method on arbitrarily locally refined nonconforming meshes. A zoom on the most distorted diamond-cell for this type of mesh (with $n=2$) is displayed on Figure 6.3. Comparing this figure with Figure 5.1, we infer that

$$\max(\alpha_1,\beta_1,\mu_1+\mu_2,\alpha_2,\beta_2,\nu_1+\nu_2)=\beta_2,$$

which is always lower than $\frac{3\pi}{4}$ for all values of n. Moreover, it is easily checked that the maximum angle of every boundary diamond-cell equals $\frac{\pi}{2}$, so that this family of meshes satisfies Hypothesis 5.5 with an angle $\tau^*=\frac{3\pi}{4}$. The discrete L^2 error is displayed in logarithmic scale on the right part of Figure 6.2, together with a reference straight line with a slope equal to one. We observe, on this family of nonconforming, locally refined meshes, a first order convergence in the discrete L^2 norm.

6.3. Degenerating meshes. The third family is made up of grids of increasingly flat triangles built in the following way. Let n be a nonzero integer. We divide Ω into 4^n horizontal stripes of the same height and divide each of these stripes into similar triangles (except those at both ends) so that there are 2^n bases of triangles in the width of a stripe and choose $n\in[1;6]_N$. The left and central parts of Figure 6.4 represent the first two of these grids. The numerical errors in the L^2 norm are presented in logarithmic scale on the right part of Figure 6.4, as well as a straight line of slope 1.5. Although such a family of meshes does not verify Hypothesis 5.5 (due to boundary diamond-cells), we observe a superconvergence of the method in this case,

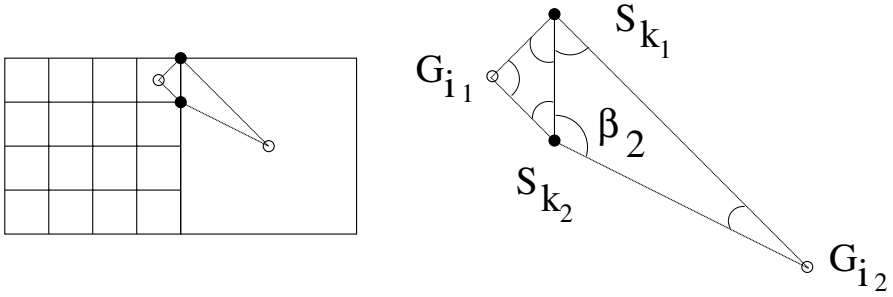


FIG. 6.3. *Zoom on a diamond-cell for the locally refined meshes with* $n = 2$.

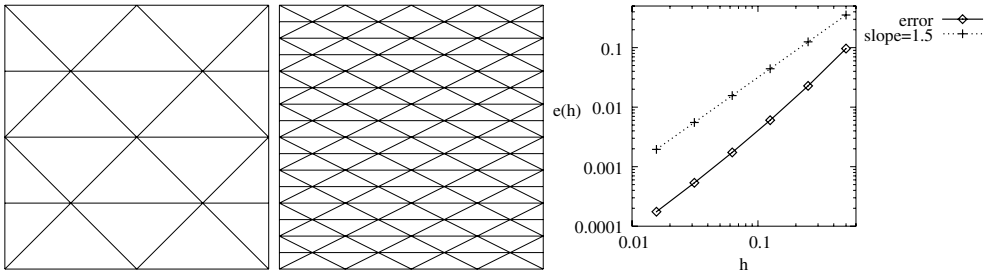


FIG. 6.4. *Degenerating triangular meshes.*

which is due to the fact, as shown in [10], that almost all diamond-cells (except those at the boundary) are parallelograms.

6.4. Nonsimply connected domains. Here, the domain of computation is $\Omega = [0,1]^2 \setminus [1/3, 2/3]^2$ and the data and boundary conditions are chosen so that the analytic solution is given by

$$\hat{\mathbf{u}}(x,y) = \begin{pmatrix} \exp(x)\,\cos(\pi y) + 3\pi\,\sin(3\pi x)\,\cos(3\pi y) \\ -\pi\,\exp(x)\,\sin(\pi y) - 3\pi\,\cos(3\pi x)\,\sin(3\pi y) \end{pmatrix}.$$

This means that the exact potentials are given by

$$\hat{\phi}(x,y) = \exp(x)\,\cos(\pi y) \quad \text{and} \quad \hat{\psi}(x,y) = \sin(3\pi x)\,\sin(3\pi y).$$

We compute the numerical solution on a family of five increasingly fine triangular meshes. The first two of the meshes are displayed on the left and central parts of Figure 6.5. The numerical errors in the L^2 norm are presented in logarithmic scale on the right part of Figure 6.5, as well as a straight line of slope 1. We observe the first order convergence of the scheme on this type of nonconvex meshes when the solution is regular enough, which is not the case of the last example.

6.5. Nonconvex domains and less regular solutions. Here, the domain of computation is $\Omega =]-1/2; 1/2[^2 \backslash]0; 1/2[^2$ and the data and boundary conditions are chosen so that the analytic solution, expressed in polar coordinates centered on $(0,0)$, is given by

$$\hat{\mathbf{u}}(r,\theta) = \boldsymbol{\nabla}\left(r^{2/3}\cos\left(\frac{2}{3}\theta\right)\right),$$

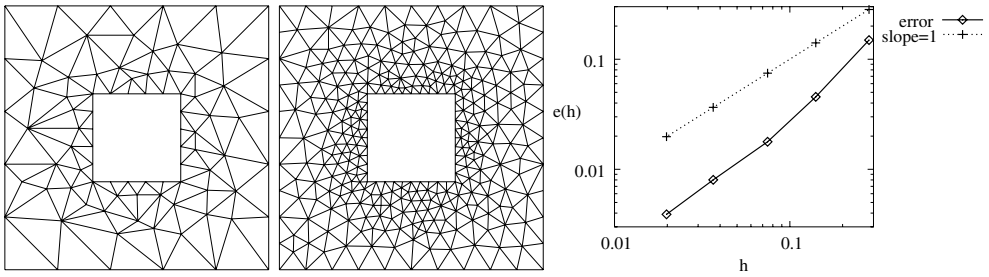


FIG. 6.5. *Nonsimply connected meshes.*

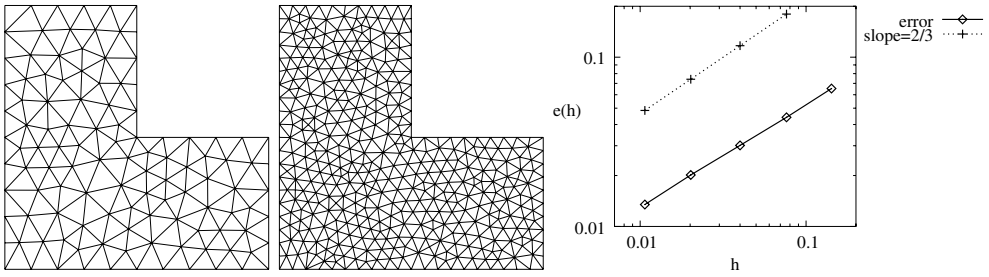


FIG. 6.6. *Nonconvex meshes.*

that is to say $\hat{\phi}(r,\theta) = r^{2/3}\cos(\frac{2}{3}\theta)$ and $\hat{\psi} = 0$. Note that $\hat{\phi}$ is still in H^1 but not in H^2, so that the error estimate derived in section 5.3 is not valid. More precisely, $\hat{\phi} \in (H^{1+s}(\Omega))^2$ with $s < 2/3$. We use a family of five unstructured triangular grids. The first two meshes of this family are displayed on the left and central parts of Figure 6.6, while the error curve in the discrete L^2 norm is shown on the right part of Figure 6.6, together with a reference line of slope 2/3. The order of convergence of the scheme seems to be 2/3 in this case, like that obtained in [4].

7. Conclusion. We have proposed new discretizations of differential operators such as divergence, gradient, and curl on almost arbitrary two-dimensional meshes. These discrete operators verify discrete properties analogous to their continuous counterparts. We have applied these ideas to approximate the solution of two-dimensional div-curl problems and have given error estimations for the resulting scheme. Finally, we have demonstrated the possibilities of the method by providing a series of numerical tests. Extensions of these ideas to problems with inhomogeneous and/or anisotropic and/or discontinuous coefficients and to the discretization of Stokes-like problems are currently being investigated.

REFERENCES

[1] G. ACOSTA AND R. G. DURÁN, *The maximum angle condition for mixed and nonconforming elements: Application to the Stokes equations*, SIAM J. Numer. Anal., 37 (1999), pp. 18–36.

[2] B. ANDREIANOV, F. BOYER, AND F. HUBERT, *Discrete duality finite volume schemes for Leray-Lions-type elliptic problems on general 2D meshes*, Numer. Methods Partial Differential Equations, 23 (2007), pp. 145–195.

[3] I. BABUSKA AND A. K. AZIZ, *On the angle condition in the finite element method*, SIAM J. Numer. Anal., 13 (1976), pp. 214–226.

[4] J. H. Bramble and J. E. Pasciak, *A new approximation technique for div-curl systems*, Math. Comp., 73 (2004), pp. 1739–1762.
[5] J. C. Campbell, J. M. Hyman, and M. J. Shashkov, *Mimetic finite difference operators for second-order tensors on unstructured grids*, Comput. Math. Appl., 44 (2002), pp. 157–173.
[6] C. Chainais-Hillairet, *Finite volume schemes for two dimensional drift-diffusion and energy-transport models*, in Finite Volumes for Complex Applications, IV, F. Benkhaldoun, D. Ouazar, and S. Raghay, eds., Hermes Science Publishing, London, UK, 2005, pp. 13–22.
[7] S. Choudhury and R. A. Nicolaides, *Discretization of incompressible vorticity-velocity equations on triangular meshes*, Internat. J. Numer. Methods Fluids, 11 (1990), pp. 823–833.
[8] Y. Coudière, J.-P. Vila, and P. Villedieu, *Convergence rate of a finite volume scheme for a two-dimensional convection-diffusion problem*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 493–516.
[9] S. Delcourte, K. Domelevo, and P. Omnes, *Discrete duality finite volume method for second order elliptic problems*, in Finite Volumes for Complex Applications, IV, F. Benkhaldoun, D. Ouazar, and S. Raghay, eds., Hermes Science Publishing, London, UK, 2005, pp. 447–458.
[10] K. Domelevo and P. Omnes, *A finite volume method for the Laplace equation on almost arbitrary two-dimensional grids*, M2AN Math. Model. Numer. Anal., 39 (2005), pp. 1203–1249.
[11] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.
[12] X. H. Hu and R. A. Nicolaides, *Covolume techniques for anisotropic media*, Numer. Math., 61 (1992), pp. 215–234.
[13] J. M. Hyman and M. Shashkov, *Natural discretizations for the divergence, gradient, and curl on logically rectangular grids*, Comput. Math. Appl., 33 (1997), pp. 81–104.
[14] J. M. Hyman and M. Shashkov, *Adjoint operators for the natural discretizations for the divergence, gradient and curl on logically rectangular grids*, Appl. Numer. Math., 25 (1997), pp. 413–442.
[15] J. M. Hyman and M. Shashkov, *The orthogonal decomposition theorems for mimetic finite difference methods*, SIAM J. Numer. Anal., 36 (1999), pp. 788–818.
[16] J. M. Hyman and M. Shashkov, *Mimetic discretizations for Maxwell's equations*, J. Comput. Phys., 151 (1999), pp. 881–909.
[17] J. Hyman, J. Morel, M. Shashkov, and S. Steinberg, *Mimetic finite difference methods for diffusion equations*, Comput. Geosci., 6 (2002), pp. 333–352.
[18] P. Jamet, *Estimations d'erreur pour des éléments finis droits presque dégénérés*, RAIRO Analyse Numérique, 10 (1976), pp. 43–60.
[19] K. Lipnikov, J. Morel, and M. Shashkov, *Mimetic finite difference methods for diffusion equations on non-orthogonal non-conformal meshes*, J. Comput. Phys., 199 (2004), pp. 589–597.
[20] R. A. Nicolaides, *Direct discretization of planar div-curl problems*, SIAM J. Numer. Anal., 29 (1992), pp. 32–56.
[21] R. A. Nicolaides and D.-Q. Wang, *Convergence analysis of a covolume scheme for Maxwell's equations in three dimensions*, Math. Comp., 67 (1998), pp. 947–963.
[22] C. Pierre, *Modélisation et Simulation de L'activité Électrique du Cœur dans le Thorax, Analyse Numérique et Méthodes de Volumes Finis*, Ph.D. thesis, University of Nantes, Nantes, France, 2005.
[23] P.-A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of the Finite Element Methods, Lecture Notes in Math. 606, I. Galligani and E. Magenes, eds., Springer-Verlag, Berlin, 1977, pp. 292–315.
[24] G. Strang, *Variational crimes in the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 689–710.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1175–1199

B-SPLINE-BASED MONOTONE MULTIGRID METHODS*

MARKUS HOLTZ† AND ANGELA KUNOTH†

Abstract. For the efficient numerical solution of elliptic variational inequalities on closed convex sets, multigrid methods based on piecewise linear finite elements have been investigated over the past decades. Essential to their success is the appropriate approximation of the constraint set on coarser grids which is based on function values for piecewise linear finite elements. On the other hand, there are a number of problems which profit from higher order approximations. Among these are the problem of pricing American options, formulated as a parabolic boundary value problem involving Black–Scholes' equation with a free boundary. In addition to computing the free boundary (the optimal exercise price of the option) of particular importance are accurate pointwise derivatives of the value of the stock option up to order two, the so-called Greek letters. In this paper, we propose a monotone multigrid method for discretizations in terms of B-splines of arbitrary order to solve elliptic variational inequalities on a closed convex set. In order to maintain monotonicity (upper bound) and quasi optimality (lower bound) of the coarse grid corrections, we propose an optimized coarse grid correction (OCGC) algorithm which is based on B-spline expansion coefficients. We prove that the OCGC algorithm is of optimal complexity of the degrees of freedom of the coarse grid and, therefore, the resulting monotone multigrid method is of asymptotically optimal multigrid complexity. Finally, the method is applied to a standard model for the valuation of American options. In particular, it is shown that a discretization based on B-splines of order four enables us to compute the second derivative of the value of the stock option to high precision.

Key words. variational inequality, linear complementary problem, monotone multigrid method, cardinal higher order B-spline, system of linear inequalities, optimized coarse grid correction algorithm, optimal complexity, convergence rates, American option, Greek letters, high precision

AMS subject classifications. 65M55, 35J85, 65N30, 65D07

DOI. 10.1137/050642575

1. Introduction.

The motivation for this paper stems from an application in Mathematical Finance, the fair pricing of American options. In a standard model, this problem can be formulated as a parabolic boundary value problem involving Black–Scholes' equation [BS] with a free boundary. In addition to computing the free boundary (the optimal exercise price of the option), pointwise higher order derivatives of the solution (the value of the stock option) are particularly important. These so-called Greek letters are needed with high precision as they play a crucial role as hedge parameters in the analysis of market risks. Thus, a discretization in terms of higher order basis functions is preferable.

On the other hand, for the fast numerical solution of the resulting (semidiscrete) elliptic variational inequality, the method of choice is the monotone multigrid method developed in [Ko1, Ko2]. Multigrid methods have been proposed previously for such problems using second order discretizations (i.e., standard finite difference stencils or piecewise linear finite elements) in different variants [BC, HM, Ho, Ma] where, however, not all of them have assured, consequently, that the obstacle criterion is met. Using piecewise linear finite element ansatz functions, geometric considerations

*Received by the editors October 13, 2005; accepted for publication (in revised form) November 13, 2006; published electronically May 14, 2007. This work has been supported in part by the Deutsche Forschungsgemeinschaft SFB 611, Universität Bonn.

http://www.siam.org/journals/sinum/45-3/64257.html

†Institut für Numerische Simulation, Universität Bonn, Wegelerstr. 6, 53115 Bonn, Germany (holtz@ins.uni-bonn.de, www.ins.uni-bonn.de/∼holtz, kunoth@ins.uni-bonn.de, www.ins.uni-bonn.de/∼kunoth).

based on point values are used in [Ko1] to represent the problem-inherent obstacles on coarser grids in such a way that a violation of the obstacle is excluded. The difficulty to correctly identifying coarse grid approximations has also been the motivation for a cascadic multigrid algorithm for variational inequalities in [BBS] for which, however, no convergence theory is yet available.

In this paper, we generalize the monotone multigrid (MMG) method from [Ko1, Ko2] to discretizations involving higher order B-splines. One of the key ingredients of an MMG method are restrictions of the obstacle to coarser grids which satisfy the (upper) bound imposed by the obstacle (monotonicity) as well as a lower one which corresponds to the condition of quasi optimality in [Ko1]. We formulate the construction of coarse grid approximations as a linear constrained optimization problem with respect to the B-spline expansion coefficients. Our construction heavily profits from properties of B-splines [Bo, Sb]. In particular, we present with our optimized coarse grid correction (OCGC) algorithm a method to construct monotone and quasi-optimal coarse grid approximations to the obstacle function in optimal complexity of the coarse grid for B-spline basis functions of any degree.

Building the OCGC scheme into the MMG method, our higher order MMG method is shown to be of optimal multigrid complexity. Moreover, following the arguments in [Ko1], we can prove that our method is globally convergent and reduces asymptotically to a linear subspace correction method once the contact set has been identified [HzK]. Hence, we can expect particular robustness of the scheme and full multigrid efficiency in the asymptotic range in the numerical experiments. This is confirmed by computations for an American option pricing problem in terms of cubic B-splines. Details about the derivation of the problem of fair pricing American options and its formulation as a free boundary value problem and corresponding results can be found in [WHD, Hz]. Of course, once higher order MMG methods are available, they may be applied to other obstacle problems like Signorini's problem which has been solved using piecewise linear hat functions in [Kr].

This paper is structured as follows. In section 2 we introduce monotone multigrid methods, recollect the main features of B-splines, and specify a B-spline-based projected Gauss–Seidel relaxation as a smoothing component of the scheme. In section 3 the crucial ingredients of the higher order MMG schemes, suitable restriction operators for the obstacle function, are presented for B-spline functions of arbitrary degree in the univariate case. Their construction for higher spatial dimensions is presented in section 4 using tensor products. In section 5 some short remarks concerning the convergence theory for B-spline-based MMG schemes are made. Finally, in section 6 we present a numerical example of pricing American options. The convergence behavior of the projected Gauss–Seidel and the multigrid schemes is compared for basis functions of different orders. We conclude with an estimation of asymptotic multigrid convergence rates which exhibit full multigrid efficiency for the truncated version.

2. MMG methods.

2.1. Elliptic variational inequalities and linear complementary problems. Let Ω be a domain in $\mathbb{R}^d$ and $\mathcal{J}(v) := \frac{1}{2}a(v,v) - f(v)$ a quadratic functional induced by a continuous, symmetric, and H_0^1– elliptic bilinear form $a(\cdot,\cdot) : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ and a linear functional $f : H_0^1(\Omega) \to \mathbb{R}$. As usual, $H_0^1(\Omega)$ is the subspace of functions belonging to the Sobolev space $H^1(\Omega)$ with zero trace on the boundary. We consider the constrained minimization problem

$$\text{find } u \in \mathcal{K} : \quad \mathcal{J}(u) \le \mathcal{J}(v) \quad \text{for all } v \in \mathcal{K} \tag{2.1}$$

on the closed and convex set

$$\mathcal{K} := \{v \in H_0^1(\Omega) : \ v(x) \le g(x) \ \text{ for all } x \in \Omega\} \subset H_0^1(\Omega).$$

The function $g \in H_0^1(\Omega)$ represents an upper obstacle for the solution $u \in H_0^1(\Omega)$. Lower obstacles can be treated in the obvious analogous way. If g satisfies $g(x) \ge 0$ for all $x \in \partial\Omega$, then problem (2.1) admits a unique solution $u \in \mathcal{K}$ by the Lax–Milgram theorem. It is well-known that (2.1) can be rewritten as a variational inequality; see, e.g., [EO, KS]: find $u \in \mathcal{K}$: $a(u, v-u) \ge f(v-u)$ for all $v \in \mathcal{K}$ or, equivalently, as a *linear complementary problem*

$$\begin{aligned} \mathcal{L}u &\ge f, \\ u &\le g, \\ (u-g)(\mathcal{L}u - f) &= 0 \end{aligned} \tag{2.2}$$

almost everywhere in Ω. Here $\mathcal{L} : H_0^1(\Omega) \to H^{-1}(= (H_0^1(\Omega))')$ is the Riesz operator defined by $\langle \mathcal{L}u, v\rangle := a(u, v)$ for all $v \in H_0^1(\Omega)$.

Discretizing in a finite dimensional spline space S_L of piecewise polynomials on a grid Δ_L with uniform grid spacing h_L leads to the discrete formulation of (2.1),

$$\text{find } u_L \in \mathcal{K}_L : \quad \mathcal{J}(u_L) \le \mathcal{J}(v_L) \quad \text{ for all } v_L \in \mathcal{K}_L \tag{2.3}$$

on the closed and convex set $\mathcal{K}_L := \{v_L \in S_L : \ v_L(x) \le g_L(x) \ \text{ for all } x \in \Omega\} \subset S_L$, or, equivalently,

$$\begin{aligned} \mathcal{L}_L u_L &\ge f_L, \\ u_L &\le g_L, \\ (u_L - g_L)(\mathcal{L}_L u_L - f_L) &= 0. \end{aligned} \tag{2.4}$$

In [BHR] regularity $u \in H^{5/2-\epsilon}(\Omega)$ of the solution u to (2.2) is shown for arbitrary $\epsilon > 0$. Moreover, error estimates $\|u - u_L\|_{H^1(\Omega)} = O(h_L)$ and $\|u - u_L\|_{H^1(\Omega)} = O(h_L^{3/2-\epsilon})$ are proved in the case of piecewise linear (respectively, piecewise quadratic) functions, provided the functions f, g are sufficiently regular.

2.2. The MMG-algorithm. For solving (2.3) numerically, a now-popular method is the MMG method [Ko1]. By adding a projection step and employing specific restriction operators, it can be implemented as a variant of a standard multigrid scheme. Let $S_1 \subset S_2 \subset \cdots \subset S_L \subset H_0^1(\Omega)$ be a nested sequence of finite dimensional spaces, and let $u_L^\nu \in S_L$ be the approximation in the νth iteration of the MMG method. The basic multigrid idea is that the error $v_L := u_L - u_L^{\nu,1}$ between the smoothed iterate $u_L^{\nu,1} := \mathcal{S}(u_L^\nu)$ ($\mathcal{S}$ always being the standard Gauss–Seidel iteration) and the exact solution u_L can be approximated without essential loss of information on a coarser grid Δ_{L-1}. We explain how this is realized in the case of a linear complementary problem for two grids Δ_L and Δ_{L-1}. Introducing the defect $d_L := f_L - \mathcal{L}_L u_L^{\nu,1}$, (2.4) can be written as

$$\begin{aligned} \mathcal{L}_L v_L &\ge d_L, \\ v_L &\le g_L - u_L^{\nu,1}, \\ (v_L - g_L + u_L^{\nu,1})(\mathcal{L}_L v_L - d_L) &= 0. \end{aligned} \tag{2.5}$$

On a coarser grid Δ_{L-1}, the defect problem can now be approximated by

$$\begin{aligned} \mathcal{L}_{L-1} v_{L-1} &\geq d_{L-1}, \\ v_{L-1} &\leq g_{L-1}, \\ (v_{L-1} - g_{L-1})(\mathcal{L}_{L-1} v_{L-1} - d_{L-1}) &= 0, \end{aligned}$$

where $d_{L-1} := r\, d_L$ and $g_{L-1} := \tilde{r}(g_L - u_L^{\nu,1})$ with (different) restriction operators $r, \tilde{r} : S_L \to S_{L-1}$. The solution v_{L-1} of the coarse grid problem is then used as an approximation to the error v_L. It is first transported back to the fine grid by a prolongation operator p and is then added to the approximation $u_L^{\nu,1}$. It is important that the restriction $\tilde{r}$ is chosen such that the new iterate satisfies the constraint

$$u_L^{\nu,2} := u_L^{\nu,1} + p v_{L-1} \leq g_L \tag{2.6}$$

on the fine grid. Applying this idea recursively on several different grids, one obtains the MMG method for linear complementary problems.

ALGORITHM 2.1. **MMG**$_\ell$ *(νth cycle on level $\ell \geq 1$).*
Let $u_\ell^\nu \in S_\ell$ be a given approximation.

1. *A priori smoothing and projection:* $u_\ell^{\nu,1} := (\mathcal{P} \circ \mathcal{S}(u_\ell^\nu))^{\eta_1}$.
2. *Coarse grid correction:*
$$\begin{aligned} d_{\ell-1} &:= r(f_\ell - \mathcal{L}_\ell u_\ell^{\nu,1}), \\ g_{\ell-1} &:= \tilde{r}(g_\ell - u_\ell^{\nu,1}), \\ \mathcal{L}_{\ell-1} &:= r \mathcal{L}_\ell p. \end{aligned}$$
If $\ell = 1$, solve exactly the linear complementary problem
$$\begin{aligned} \mathcal{L}_{\ell-1} v &\geq d_{\ell-1}, \\ v &\leq g_{\ell-1}, \\ (v - g_{\ell-1})(\mathcal{L}_{\ell-1} v - d_{\ell-1}) &= 0, \end{aligned}$$
and set $v_{\ell-1} := v$.
If $\ell > 1$, do γ steps of **MMG**$_{\ell-1}$ *with initial value $u_{\ell-1}^0 := 0$ and solution $v_{\ell-1}$.*
Set $u_\ell^{\nu,2} := u_\ell^{\nu,1} + p v_{\ell-1}$.
3. *A posteriori smoothing and projection:* $u_\ell^{\nu,3} := (\mathcal{P} \circ \mathcal{S}(u_\ell^{\nu,2}))^{\eta_2}$.
Set $u_\ell^{\nu+1} := u_\ell^{\nu,3}$.

The number of a priori and a posteriori smoothing steps is denoted by η_1 and η_2, respectively. For $\gamma = 1$ one obtains a V-cycle, for $\gamma = 2$ a W-cycle. $\mathcal{P}$ denotes a projection operator defined in (2.7) and (2.11).

Condition (2.6) leads to an inner approximation of the solution set $\mathcal{K}_L$ and ensures that the multigrid scheme is robust [Ko1]. Striving for optimal multigrid efficiency, satisfaction of the constraint should *not* be checked by interpolating v_ℓ back to the finest grid. Instead, special restriction operators $\tilde{r}$ are needed for the obstacle function. A corresponding construction for B-splines of general order k will be introduced in sections 3 and 4. Next we discuss the projection step for general order B-splines.

2.3. A B-spline-based projected Gauss–Seidel scheme. Since the operator $\mathcal{L}$ is symmetric positive definite and continuous piecewise linear functions are used for discretization, the discrete form (2.4) can be solved by the projected Gauss–Seidel scheme; see, e.g., [Cr]. Given an iterate u_L^ν, a standard Gauss–Seidel sweep $\bar{u}_L^\nu := \mathcal{S}(u_L^\nu)$ is supplemented by a projection $u_L^{\nu+1} = \mathcal{P}\, \bar{u}_L^\nu$ into the convex set $\mathcal{K}_L$. If

S_L consists of hat functions, the projection can be defined for given grid points $\{\theta_i\}_i$ by

$$\mathcal{P}\, v_L(\theta_i) := \min\{v_L(\theta_i),\, g_L(\theta_i)\}. \tag{2.7}$$

For higher order functions v_L, the difficulty arises already in the univariate case that for given $x \in [\theta_i, \theta_{i+1}]$ the estimate

$$\min\{v_L(\theta_i), v_L(\theta_{i+1})\} \le v_L(x) \le \max\{v_L(\theta_i), v_L(\theta_{i+1})\} \tag{2.8}$$

is no longer valid. Thus, controlling function values on grid points is not a sufficient criterion in this case. Instead, we propose here a construction using higher order B-splines, which compares B-spline expansion coefficients instead of function values and heavily profits from the fact that B-splines are nonnegative. We begin with the univariate case. For readers' convenience, we recall the relevant facts about B-spline bases from [Bo].

DEFINITION 2.2 (B-spline basis functions). *For $k \in \mathbb{N}$ and $n \in \mathbb{N}$ let $T := \{\theta_i\}_{i=1,\dots,n+k}$ be an* expanded knot sequence *with uniform grid spacing h_L in the interior of the interval $I := [a, b]$ of the form*

$$\theta_1 = \cdots = \theta_k = a < \theta_{k+1} < \cdots < \theta_n < b = \theta_{n+1} = \cdots = \theta_{n+k}. \tag{2.9}$$

Then the B-spline basis functions $N_{i,k}$ of order k are recursively defined for $i = 1, \dots, n$ by

$$\begin{aligned} N_{i,1}(x) &= \begin{cases} 1 & \text{if } x \in [\theta_i, \theta_{i+1}) \\ 0 & \text{else,} \end{cases} \\ N_{i,k}(x) &= \frac{x - \theta_i}{\theta_{i+k-1} - \theta_i}\, N_{i,k-1}(x) + \frac{\theta_{i+k} - x}{\theta_{i+k} - \theta_{i+1}}\, N_{i+1,k-1}(x) \end{aligned} \tag{2.10}$$

for $x \in I$.

It is known that $\operatorname{supp} N_{i,k} \subseteq [\theta_i,\, \theta_{i+k}]$ (local support), $N_{i,k}(x) \ge 0$ for all $x \in I$ (nonnegativity), and $N_{i,k} \in C^{k-2}(I)$ (differentiability) holds. Moreover, the set $\Sigma_L := \{N_{1,k}, \dots, N_{n,k}\}$ constitutes a locally independent and unconditionally stable basis with respect to $\|\cdot\|_{L_p}$, $1 \le p \le \infty$ for the finite dimensional space $S_L = \mathcal{N}_{k,T} := \operatorname{span} \Sigma_L$ of the splines of order k.

LEMMA 2.3. *If the B-spline coefficients of $v_L, g_L \in \mathcal{N}_{k,T} = S_L$ satisfy $v_i \le g_i$ for all $i = 1, \dots, n$, then $v_L(x) \le g_L(x)$ holds for all $x \in I$.*

Proof. Using the representation $v_L = \sum_{i=1}^n v_i\, N_{i,k}$ and $g_L = \sum_{i=1}^n g_i\, N_{i,k}$ and the nonnegativity $N_{i,k}(x) \ge 0$ for all $x \in I$, we deduce that $g_L(x) - v_L(x) = \sum_{i=1}^n (g_i - v_i)\, N_{i,k}(x) \ge 0$ for all $x \in I$. ∎

Here and in section 5, we use the subscript i in $v_i = (v_L)_i$ to denote B-spline expansion coefficients.

The projection can now be defined for B-spline functions of general order k similar to (2.7), but now involving expansion coefficients by setting

$$\mathcal{P}\, v_i := \min\{v_i,\, g_i\}. \tag{2.11}$$

Using the same arguments as in [Cr], the resulting projected Gauss–Seidel scheme still converges since the discrete solution set $\{\mathbf{v} \in \mathbb{R}^n : v_i \le g_i \text{ for } i = 1, \dots, n\}$ describes a cuboid in $\mathbb{R}^n$. Moreover, if the problem is nondegenerate, the contact set,

defined by all coefficients for which equality holds, is identified after a finite number of iterations [Cr, EO].

We treat the multivariate case by taking tensor products. Specifying the domain Ω as $\Omega := \prod_{\ell=1}^{d}[a_\ell, b_\ell] \subset \mathbb{R}^d$, the ith d-dimensional tensor product B-spline of order k on a tensorized extended knot sequence $T^{(d)}$ is defined by

$$N_{i,k}^{(d)}(x) := \prod_{\ell=1}^{d} N_{i_\ell,k}(x_\ell), \qquad x \in \Omega, \tag{2.12}$$

where $i := (i_1, \ldots, i_d)$ denotes a multi-index. Defining S_L in analogy to the univariate case, the result of Lemma 2.3 immediately carries over to the d-dimensional setting.

3. Construction of monotone and quasi-optimal obstacle approximations. In this section, the second essential ingredient for our B-spline-based MMG methods is provided, the construction of so-called *monotone* and *quasi-optimal coarse grid approximations* of the obstacle function, which lead to suitable restriction operators $\tilde{r}$. We begin with the univariate case; the extension to d dimensions follows in section 4. We consider in what follows only two grids, as the generalization to several grids is obvious. Given an obstacle function $\tilde{S}$ which is defined on a fine grid $\Delta \subset I$, we provide an approximation S with respect to a coarser grid T which satisfies

1. $S(x) \le \tilde{S}(x)$ for all $x \in I$;
2. $S(x) \ge L_k(x)$ for all $x \in I$ and a still-to-be-specified lower barrier $L_k(x)$ provided in section 3.2;
3. $S \approx \tilde{S}$ with respect to a target functional F_k defined below in (3.10).

The first condition ensures the monotonicity and robustness of the multigrid scheme, the second an asymptotical reduction of the method to a linear relaxation, and the third an efficient coarse grid correction. As the construction is used as a component of the monotone multigrid scheme striving for optimal computational multigrid complexity, it also has to satisfy

4. the number of arithmetic operations must be of order $O(n)$, where n denotes the number of degrees of freedom on the coarse grid.

Specifically, let T be an extended knot sequence with grid spacing H as in (2.9) and let $\Delta := \{\tilde{\theta}_i\}_{i=1,\ldots,\tilde{n}+k}$ be a *finer* knot sequence

$$\tilde{\theta}_1 = \cdots = \tilde{\theta}_k = a < \tilde{\theta}_{k+1} < \cdots < \tilde{\theta}_{\tilde{n}} < b = \tilde{\theta}_{\tilde{n}+1} = \cdots = \tilde{\theta}_{\tilde{n}+k} \tag{3.1}$$

with grid spacing $h = \frac{1}{2}H$. It is defined such that $\theta_i = \tilde{\theta}_{2i-k}$ for $i = k, \ldots, n+1$ and $\frac{1}{2}(\theta_{i-1} + \theta_i) = \tilde{\theta}_{2i-k-1}$ for $i = k+1, \ldots, n+1$. Then it holds that

$$\tilde{n} = 2n + 1 - k. \tag{3.2}$$

The corresponding spline spaces are $\mathcal{N}_{k,\Delta}$ and $\mathcal{N}_{k,T}$ with member functions $N_{i,k,\Delta}$ and $N_{i,k,T}$, respectively. Now let the obstacle function on the fine grid $\tilde{S} \in \mathcal{N}_{k,\Delta}$ and its approximation $S \in \mathcal{N}_{k,T}$ be expanded as

$$\tilde{S} = \sum_{i=1}^{\tilde{n}} \tilde{c}_i\, N_{i,k,\Delta} =: \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}, \qquad S = \sum_{i=1}^{n} c_i\, N_{i,k,T} =: \mathbf{c}^T \mathbf{N}_{k,T}. \tag{3.3}$$

There is a natural *prolongation operator* p from $\mathcal{N}_{k,T}$ to $\mathcal{N}_{k,\Delta}$ for B-splines $N_{i,k,T}$ in terms of their refinement or mask coefficients [Bo, Sb]. In the special case $H = 2\,h$

considered here the refinement relation is given by

$$N_{i,k,T} = \sum_{j=0}^{k} a_j \, N_{2i-k+j,k,\Delta} \tag{3.4}$$

with the *subdivision* or *mask* coefficients

$$a_j := 2^{1-k} \binom{k}{j} \quad \text{for } j = 0, \ldots, k. \tag{3.5}$$

In step 2 of Algorithm 2.1, we choose the restriction r as the adjoint of p, following [Ha]. However, for the obstacle function the restriction operator r cannot be used since it does not satisfy condition (2.6).

3.1. Monotone coarse grid approximations. There is a vast amount of literature, see, e.g., [DV, Mv, Pi] especially from approximation theory, dealing with monotone approximations to a given function g. The function $\hat{g}$ is a *monotone* (or *one-sided*) *lower approximation* to g if $\hat{g}(x) \le g(x)$ for all $x \in I$. There the number n of degrees of freedom of the function $\hat{g}$ is chosen such that a given approximation accuracy can be reached. In contrast to these studies, the question here is different, since the number n of degrees of freedom is given by the mesh size H.

DEFINITION 3.1 (monotone coarse grid approximation). *For knot sequences T and Δ from* (2.9) *and* (3.1), *respectively, we call $S \in \mathcal{N}_{k,T}$ a* monotone lower coarse grid approximation *to $\tilde{S} \in \mathcal{N}_{k,\Delta}$ if $S(x) \le \tilde{S}(x)$ holds for all $x \in I$.*

For hat functions such approximations are constructed in [Ma, Ko1]. A corresponding construction for higher order functions has, to our knowledge, not been provided so far. In view of Lemma 2.3 we propose here to control B-spline expansion coefficients.

THEOREM 3.2 (monotone coarse grid approximation). *Let $\tilde{S} \in \mathcal{N}_{k,\Delta}$ be an upper obstacle with $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}$ for a given order k and the knot sequence Δ from* (3.1). *Then $S \in \mathcal{N}_{k,T}$ with $S = \mathbf{c}^T \mathbf{N}_{k,T}$ defined on the knot sequence T from* (2.9) *is a monotone lower coarse grid approximation to $\tilde{S}$ if the inequality system*

$$A_k \, \mathbf{c} \le \tilde{\mathbf{c}} \tag{3.6}$$

is satisfied. The two-slanted *matrix A_k is defined by*

$$A_k := \begin{pmatrix}
a_{k-1} & a_{k-3} & \ddots & & & & \\
a_k & a_{k-2} & & & & & \\
 & a_{k-1} & & a_0 & & & \\
 & a_k & & a_1 & & & \\
 & & \ddots & a_2 & & & \\
 & & & \vdots & & & \\
 & & & a_{k-1} & \ddots & & \\
 & & & a_k & & & \\
 & & & & \ddots & a_0 & \\
 & & & & & a_1 &
\end{pmatrix} \in \mathbb{R}^{\tilde{n} \times n}$$

with the subdivision coefficients a_j from (3.5) *and has maximal rank.*

Proof. The proof relies on the subdivision property (3.4) and on the nonnegativity of B-splines. We only consider the case k even as the other case is analogous. Substituting (3.4) into (3.3) and sorting according to the basis functions $N_{i,k,\Delta}$ leads to

$$S(x) = \sum_{\substack{i=1\\ i \text{ odd}}}^{\tilde{n}} \left(a_{k-1}\, c_{(i+1)/2} + a_{k-3}\, c_{(i+3)/2} + \cdots + a_1\, c_{(i+k-1)/2}\right) N_{i,k,\Delta}(x) + \sum_{\substack{i=2\\ i \text{ even}}}^{\tilde{n}-1} \left(a_k\, c_{i/2} + a_{k-2}\, c_{(i+2)/2} + \cdots + a_0\, c_{(i+k)/2}\right) N_{i,k,\Delta}(x),$$

where all c_j with $j < 1$ or $j > n$ are treated as zero. Defining the coefficients

$$d_i := \begin{cases} \tilde{c}_i - \left(a_{k-1}\, c_{(i+1)/2} + a_{k-3}\, c_{(i+3)/2} + \cdots + a_1\, c_{(i+k-1)/2}\right) & \text{if } i \text{ is odd,} \\ \tilde{c}_i - \left(a_k\, c_{i/2} + a_{k-2}\, c_{(i+2)/2} + \cdots + a_0\, c_{(i+k)/2}\right) & \text{if } i \text{ is even,} \end{cases}$$

which can be written in compact matrix/vector form as

$$d_i = \tilde{c}_i - (A_k\, \mathbf{c})_i \tag{3.7}$$

(involving the ith component of the vector $A_k\, \mathbf{c}$), we obtain

$$\tilde{S}(x) - S(x) = \sum_{i=1}^{\tilde{n}} d_i\, N_{i,k,\Delta}(x). \tag{3.8}$$

By Lemma 2.3 we have $\tilde{S}(x) - S(x) \geq 0$ for all $x \in I$, provided $d_i \geq 0$ holds for all $i = 1, \ldots, \tilde{n}$. By (3.7), we obtain the inequality system (3.6). Since the B-splines form bases for $\mathcal{N}_{k,T}$ and $\mathcal{N}_{k,\Delta}$, the matrix A_k has full rank for each k. ∎

Example 3.3. In the special case of continuous, piecewise linear functions ($k = 2$), C^1-smooth; piecewise quadratic ($k = 3$); and C^2-smooth, piecewise cubic ($k = 4$) splines, one has

$$A_2 = \begin{pmatrix} 1 & & & & & \\ \frac{1}{2} & \frac{1}{2} & & & & \\ & 1 & & & & \\ & \frac{1}{2} & \frac{1}{2} & & & \\ & \ddots & & & & \\ & & & \frac{1}{2} & & \\ & & & 1 & & \\ & & & \frac{1}{2} & \frac{1}{2} & \\ & & & & 1 \end{pmatrix} \in \mathbb{R}^{(2n-1)\times n}, \quad A_3 = \frac{1}{4} \begin{pmatrix} 3 & 1 & & & & \\ 1 & 3 & & & & \\ & 3 & 1 & & & \\ & 1 & 3 & & & \\ & \ddots & & \ddots & & \\ & & & 3 & 1 & \\ & & & 1 & 3 & \\ & & & & 3 & 1 \\ & & & & 1 & 3 \end{pmatrix} \in \mathbb{R}^{(2n-2)\times n},$$

$$A_4 = \frac{1}{8} \begin{pmatrix} 4 & 4 & & & & \\ 1 & 6 & 1 & & & \\ & 4 & 4 & & & \\ & 1 & 6 & 1 & & \\ & \ddots & & \ddots & & \\ & & 1 & 6 & 1 & \\ & & & 4 & 4 & \\ & & & 1 & 6 & 1 \\ & & & & 4 & 4 \end{pmatrix} \in \mathbb{R}^{(2n-3)\times n}.$$

TABLE 3.1
The values β_k and γ_k for orders $k = 2, 4, 6, 8$.

k	2	4	6	8
β_k	1	$\frac{2}{3}$	$\frac{17}{30}$	$\frac{166}{315}$
γ_k	0	$\frac{1}{3}$	$\frac{13}{30}$	$\frac{149}{315}$

3.2. Quasi-optimal coarse grid approximations. Now we can immediately derive a monotone lower coarse approximation.

PROPOSITION 3.4. *The spline $L_k := \mathbf{q}^T \mathbf{N}_{k,T} \in \mathcal{N}_{k,T}$ with coefficients*

$$q_i := \min\{\tilde{c}_{2i-k}, \ldots, \tilde{c}_{2i}\} \quad \text{for } i = 1, \ldots, n \tag{3.9}$$

(leaving out $\tilde{c}_j$ in the right-hand side if $j < 1$ or $j > \tilde{n}$) is a monotone lower coarse grid approximation to $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta} \in \mathcal{N}_{k,\Delta}$.

Proof. As all row sums of A_k are equal to one, the vector $\mathbf{q} := (q_1, \ldots, q_n)^T$ defined in (3.9) obviously satisfies the inequality system $A_k\,\mathbf{q} \le \tilde{\mathbf{c}}$ so that the assertion directly follows from Theorem 3.2. ∎

Remark 3.5. In the special case $k = 2$, the restriction operator $\hat{r} : \mathcal{N}_{2,\Delta} \to \mathcal{N}_{2,T}$, $\tilde{S} \mapsto L_2$ induced by Proposition 3.4 coincides with the restriction operator from [Ma].

As is illustrated in Figures 3.1 and 3.2 for the cases $k = 2$ and $k = 3$, the approximation L_k can be further improved in many cases. This will be the subject of the next subsections: there $\mathbf{q}$ is interpreted as a componentwise lower barrier for the B-spline coefficients $\mathbf{c}$ of the desired coarse grid approximation.

DEFINITION 3.6 (quasi-optimal coarse grid approximation). *We call a monotone lower coarse grid approximation $S = \mathbf{c}^T \mathbf{N}_{k,T}$ to the spline $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}$* quasi-optimal *if it is an improvement over L_k in the sense that $\mathbf{c} \ge \mathbf{q}$ holds with $\mathbf{q}$ defined in* (3.9).

3.3. A linear optimization problem. Aiming at improving the coarse grid approximation L_k from Proposition 3.4, we define an *optimal* monotone and quasi-optimal coarse grid approximation $S = \mathbf{c}^T \mathbf{N}_{k,T}$ to a given $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}$ by formulating a linear optimization problem. We choose a target functional F_k which estimates the sum of the distances from approximation to obstacle on all coarse grid points, i.e.,

$$F_k(\mathbf{c}) := \sum_{\theta \in T} |\tilde{S}(\theta) - S(\theta)|. \tag{3.10}$$

LEMMA 3.7. *The function F_k defined in* (3.10) *is a linear function $\mathbb{R}^n \to \mathbb{R}$ of the form*

$$F_k(\mathbf{c}) = \boldsymbol{\xi}^T \mathbf{c} + \eta, \tag{3.11}$$

where

$$\boldsymbol{\xi} := -A_k^T \mathbf{s}_k \in \mathbb{R}^n, \quad \mathbf{s}_k := (\beta_k, \gamma_k, \beta_k, \ldots)^T \in \mathbb{R}^{\tilde{n}}, \text{ and } \eta := \mathbf{s}_k^T \tilde{\mathbf{c}} \in \mathbb{R}. \tag{3.12}$$

The values β_k and γ_k can be computed explicitly: for odd k we have $\beta_k = \gamma_k = \frac{1}{2}$, and for even $k = 2, 4, 6, 8$ the values are displayed in Table 3.1.

Proof. By Theorem 3.2 we have $|\tilde{S}(x) - S(x)| = \tilde{S}(x) - S(x)$ for all $x \in I$. Using (3.8) we obtain

$$F_k(\mathbf{c}) = \sum_{\theta \in T} \Big(\tilde{S}(\theta) - S(\theta)\Big) = \sum_{\theta \in T} \sum_{i=1}^{\tilde{n}} d_i\, N_{i,k,\Delta}(\theta) = \sum_{i=1}^{\tilde{n}} d_i \sum_{\theta \in T} N_{i,k,\Delta}(\theta). \tag{3.13}$$

Abbreviating $(\tilde{\mathbf{s}}_k)_i := \sum_{\theta\in T} N_{i,k,\Delta}(\theta)$, we next show that $\tilde{\mathbf{s}}_k$ coincides with $\mathbf{s}_k$ defined in (3.12). In fact, $\sum_{\theta\in\Delta} N_{i,k,\Delta}(\theta) = 1$ is easily shown by induction for $k \in \mathbb{N}$. For odd k we can use a simple symmetry argument to conclude $(\tilde{\mathbf{s}}_k)_i = \frac{1}{2}$. For even k two cases must be distinguished according to the position of $N_{i,k,\Delta}$. Evaluating the B-spline on coarse grid points leads to $(\tilde{\mathbf{s}}_k)_i = \beta_k$ if $\theta_{i+k/2} \in T$ and $(\tilde{\mathbf{s}}_k)_i = \gamma_k$ in the other case. For orders $k = 2, 4, 6, 8$, the concrete values β_k and γ_k are displayed in Table 3.1. Thus, we have $(\mathbf{s}_k)_i = (\tilde{\mathbf{s}}_k)_i$ and employing (3.7) in (3.13) leads to (3.11), i.e., $F_k(\mathbf{c}) = \sum_{i=1}^{\tilde{n}} (\mathbf{s}_k)_i\,(\tilde{c}_i - (A_k\,\mathbf{c})_i) = \mathbf{s}_k^T\tilde{\mathbf{c}} - \mathbf{s}_k^T\,A_k\,\mathbf{c} = \boldsymbol{\xi}^T\mathbf{c} + \eta$. ∎

We can now define an optimal monotone and quasi-optimal coarse grid approximation as the solution of the linear optimization problem

$$\begin{aligned} &\text{Minimize the target functional} && F_k(\mathbf{c}) = \boldsymbol{\xi}^T\mathbf{c} + \eta \\ &\text{with respect to the constraints} && A_k\,\mathbf{c} \le \tilde{\mathbf{c}} \quad\text{and}\quad \mathbf{c} \ge \mathbf{q}. \end{aligned} \tag{3.14}$$

Here $A_k \in \mathbb{R}^{\tilde{n}\times n}$, $\tilde{\mathbf{c}} \in \mathbb{R}^{\tilde{n}}$, and $\mathbf{q} \in \mathbb{R}^n$ are defined as before with $\tilde{n} = 2n - k + 1$ and $\boldsymbol{\xi} \in \mathbb{R}^n$ and $\eta \in \mathbb{R}$ are given as in (3.12). The upper inequality guarantees the monotonicity of the approximation by Theorem 3.2, while the second one ensures quasi optimality by Proposition 3.4.

3.4. Solution of the linear optimization problem. Via the linear optimization formulation (3.14) a (with respect to the target functional F_k) optimal monotone and quasi-optimal coarse grid approximation may now be obtained, in principle, by the *simplex algorithm*; see, e.g., [Sj]. Here the point $\mathbf{q} \in \mathbb{R}^n$ could be used as a starting corner by Proposition 3.4. In a multigrid scheme, however, the simplex algorithm should not be used because the optimal complexity $O(n)$ would be destroyed. As shown next, a direct solution for $k = 2$ can be obtained by the *Fourier–Motzkin elimination*; see, e.g., [Sj]. For the general case $k > 2$ we present afterwards an approximate solution algorithm which can be applied in optimal complexity.

LEMMA 3.8 (direct solution for hat functions). *For $k = 2$ and given $\tilde{\mathbf{c}} \in \mathbb{R}^{\tilde{n}}$, the solution of the linear optimization problem* (3.14) *is recursively given by*

$$\begin{aligned} c_1 &:= \min\{\tilde{c}_1,\, 2\tilde{c}_2 - q_2\} \\ c_i &:= \min\{2\,\tilde{c}_{2i-2} - c_{i-1},\, \tilde{c}_{2i-1},\, 2\,\tilde{c}_{2i} - q_{i+1}\} \quad \textit{for } i = 2, \dots, n-1, \\ c_n &:= \min\{2\,\tilde{c}_{2n-2} - c_{n-1},\, \tilde{c}_{2n-1}\} \end{aligned} \tag{3.15}$$

with $q_i = \min\{\tilde{c}_{2i-2},\, \tilde{c}_{2i-1},\, \tilde{c}_{2i}\}$ for $i = 1, \dots, n$ defined in (3.9). *In particular, $S = \mathbf{c}^T\,\mathbf{N}_{k,T}$ is a monotone and quasi-optimal coarse grid approximation to the obstacle $\tilde{S} = \tilde{\mathbf{c}}^T\,\mathbf{N}_{2,\Delta}$.*

Proof. First, the n conditions $-\mathbf{c} \le -\mathbf{q}$ are integrated into the inequality system $A_2\,\mathbf{c} \le \tilde{\mathbf{c}}$ from Theorem 3.2. Then, Fourier–Motzkin elimination is applied to the resulting $(3n-1) \times n$ inequality system so that we obtain the solution range

$$\begin{aligned} q_1 &\le c_1 \le \min\{\tilde{c}_1,\, 2\tilde{c}_2 - q_2\}, \\ q_i &\le c_i \le \min\{2\tilde{c}_{2i-2} - c_{i-1},\, \tilde{c}_{2i-1},\, 2\tilde{c}_{2i} - q_{i+1}\} \quad \text{for } i = 2, \dots, n-1, \\ q_n &\le c_n \le \min\{2\tilde{c}_{2n-2} - c_{n-1},\, \tilde{c}_{2n-1}\}. \end{aligned}$$

Because of (3.9), $q_1 \le \min\{\tilde{c}_1, 2\tilde{c}_2 - q_2\}$ holds. To minimize the target function F_2 given by Lemma 3.7, all coefficients c_i must be chosen as large as possible which leads to (3.15). ∎

Remark 3.9. The restriction operator $\tilde{r} : \mathcal{N}_{2,\Delta} \to \mathcal{N}_{2,T}$, $\tilde{S} \mapsto S$, implied by Lemma 3.8, corresponds to the restriction operator from [Ko1] which is derived by

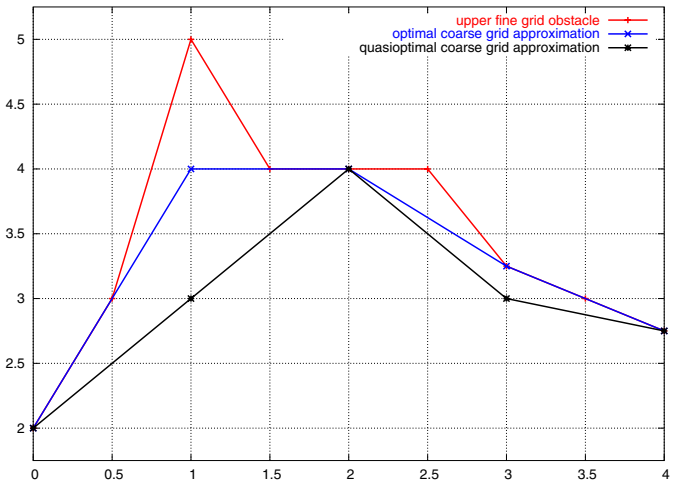


FIG. 3.1. *Continuous piecewise linear upper obstacle function on the fine grid* $[0,4] \cap \mathbb{Z}/2$ *and coarse grid approximations according to Lemma* 3.8 *and Proposition* 3.4, *respectively, on the coarse grid* $[0,4] \cap \mathbb{Z}$.

geometric considerations. It is an improvement of the restriction operator $\hat{r}$ from Remark 3.5 or [Ma] since $\tilde{r}(\tilde{S}) \geq \hat{r}(\tilde{S})$ holds for all $\tilde{S} \in \mathcal{N}_{2,\Delta}$.

In Figure 3.1 a continuous, piecewise linear, upper obstacle function, the optimal coarse grid approximation according to Lemma 3.8 and the coarse grid approximation according to Proposition 3.4 are displayed. The improvement of the simple approximation L_2 is clearly visible. Since the band width of A_k increases with increasing order k, and since the Fourier–Motzkin elimination is only suited for small matrices or for matrices with mainly zero entries [Sj], a different approach must be found to solve the linear optimization problem in the higher order case $k > 2$.

To simplify the notation we define in addition to (3.5) that $a_j := 0$ for $j > k$ and $j < 0$.

THEOREM 3.10 (optimized coarse grid correction (OCGC) scheme). *Let* $\tilde{S} \in \mathcal{N}_{k,\Delta}$ *be given with* $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}$. *Let* $L_k \in \mathcal{N}_{k,T}$ *with* $L_k = \mathbf{q}^T \mathbf{N}_{k,T}$ *be as in* (3.9) *and define*

$$\tilde{b}_j = \tilde{b}_j(c_1, \ldots, c_{\lfloor (j+k)/2 \rfloor - 1}) := \tilde{c}_j - \sum_{\nu=1}^{\lfloor (j+k)/2 \rfloor - 1} a_{j+k-2\nu}\, c_\nu \tag{3.16}$$

for $j = 1, \ldots, \tilde{n}$ *and* $\tilde{b}_j := \infty$ *for* $j < 1$. *Let* $\hat{b}_{m,i} := \infty$ *for* $m > \tilde{n}$ *or* $m < 1$ *and*

$$\hat{b}_{m,i} = \hat{b}_{m,i}(c_1, \ldots, c_{i-1}) := \tilde{c}_m - \sum_{\nu=1}^{i-1} a_{m+k-2\nu}\, c_\nu - \sum_{\nu=i+1}^{\lfloor (m+k)/2 \rfloor} a_{m+k-2\nu}\, q_\nu \tag{3.17}$$

for $i = 1, \ldots, n$ *and* $m = 2i - k + 2, \ldots, 2i$, *where* $q_j := 0$ *for* $j > n$. *Further, let the*

vector $\mathbf{c}$ *be recursively defined by*

$$c_i := \min\left\{ \frac{\tilde{b}_{2i-k}}{a_0}, \frac{\tilde{b}_{2i-k+1}}{a_1}, \frac{\hat{b}_{2i-k+2,i}}{a_2}, \ldots, \frac{\hat{b}_{2i,i}}{a_k} \right\} \quad \textit{for } i = 1, \ldots, n. \tag{3.18}$$

Then $S = \mathbf{c}^T \mathbf{N}_{k,T} \in \mathcal{N}_{k,T}$ *is a monotone and quasi-optimal coarse grid approximation to* $\tilde{S}$*, i.e.,*

$$L_k(x) \le S(x) \le \tilde{S}(x) \quad \textit{for all } x \in I.$$

Proof. We only consider the case k odd as the other case is analogous.

We first derive conditions which guarantee monotonicity (3.6) of the approximation. Moving all entries $a_{i+k-2j}c_j$ of the inequality system (3.6) except for the rightmost nonzero ones in each row to the right-hand side leads to

$$\begin{pmatrix} a_0 & 0 & & \\ a_1 & 0 & & \\ 0 & a_0 & 0 & \\ 0 & a_1 & 0 & \\ & & \ddots & \\ & & 0 & a_0 \\ 0 & \cdots & 0 & a_1 \end{pmatrix} \begin{pmatrix} c_{\ell+1} \\ c_{\ell+2} \\ \vdots \\ c_n \end{pmatrix} \le \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \\ \tilde{b}_3 \\ \tilde{b}_4 \\ \vdots \\ \tilde{b}_{\tilde{n}-1} \\ \tilde{b}_{\tilde{n}} \end{pmatrix} \tag{3.19}$$

with $\ell := \lfloor (k-1)/2 \rfloor$ and the new right-hand side coefficients $\tilde{b}_i$ defined in (3.16). From (3.19) we immediately obtain that the inequality system $A_k\mathbf{c} \le \tilde{\mathbf{c}}$ is satisfied for arbitrary $c_1, \ldots, c_\ell$ if

$$c_i \le \min\left\{ \frac{\tilde{b}_{2i-k}}{a_0}, \frac{\tilde{b}_{2i-k+1}}{a_1} \right\} \quad \text{for } i = \ell+1, \ldots, n \tag{3.20}$$

holds.

Second, we derive conditions which ensure quasi-optimality $\mathbf{c} \ge \mathbf{q}$ of the approximation. For an arbitrary $j \in \{\ell+1, \ldots, n\}$ the first inequality of (3.20) and definition (3.16) imply

$$a_0 c_j \le \tilde{b}_{2j-k} = \tilde{c}_{2j-k} - \sum_{\nu=1}^{j-1} a_{2j-2\nu}\, c_\nu.$$

For every $i \in \{1, \ldots, j-1\}$, we therefore obtain the condition

$$a_{2j-2i} c_i \le \tilde{c}_{2j-k} - \sum_{\substack{\nu=1 \\ \nu \ne i}}^{j} a_{2j-2\nu}\, c_\nu.$$

When we determine c_i, we can assume that the c_ν's for $\nu = 1, \ldots, i-1$ are already computed. For the c_ν, $\nu = i+1, \ldots, j$, which are yet to be determined, demanding quasi-optimality $c_\nu \ge q_\nu$ leads to

$$a_{2j-2i} c_i \le \tilde{c}_{2j-k} - \sum_{\nu=1}^{i-1} a_{2j-2\nu}\, c_\nu - \sum_{\nu=i+1}^{j} a_{2j-2\nu}\, q_\nu = \hat{b}_{2j-k,i} \tag{3.21}$$

with $\hat{b}_{j,i}$ defined in (3.17). Analogously we get

$$a_{2j-2i+1}c_i \le \hat{b}_{2j-k+1,i} \tag{3.22}$$

for $i < j$ using the second inequality of (3.20). Because of $a_m = 0$ for $m > k$, the inequalities (3.21) and (3.22) only apply for $i+1 \le j \le i+\ell$ so that we obtain the conditions

$$c_i \le \min\left\{ \frac{\hat{b}_{2i-k+2,i}}{a_2}, \ldots, \frac{\hat{b}_{2i,i}}{a_k} \right\} \tag{3.23}$$

for $i = 1, \ldots, n$. Then both (3.20) and (3.23) are satisfied by defining c_i, $i = 1, \ldots, n$, as in (3.18) which completes the proof. ∎

Remark 3.11. If one only aims at a coarse grid approximation S which is monotone by construction, one could use the relation (3.20) and replace the inequality by an equality sign. However, in many cases the as-large-as-possible choice of the components c_i according to (3.20) then has to be balanced to preserve monotonicity by very small, maybe even negative components c_j, $j > i$, which leads to undesirable oscillations in the solution. This is avoided by taking in addition the lower bounds into consideration.

Example 3.12. In the case $k = 2$ the recursion (3.18) recovers the direct solution

$$c_i = \min\{2\,\tilde{c}_{2i-2} - c_{i-1},\ \tilde{c}_{2i-1},\ 2\,\tilde{c}_{2i} - q_{i+1}\} \tag{3.24}$$

from Lemma 3.8. For $k = 3$ the recursion (3.18) simplifies to

$$c_i = \min\left\{4\,\tilde{c}_{2i-3} - 3\,c_{i-1},\ \tfrac{4}{3}\tilde{c}_{2i-2} - \tfrac{1}{3}c_{i-1},\ \tfrac{4}{3}\tilde{c}_{2i-1} - \tfrac{1}{3}q_{i+1},\ 4\,\tilde{c}_{2i} - 3\,q_{i+1}\right\}. \tag{3.25}$$

In the case $k = 4$, one obtains

$$\begin{aligned} c_i := \min\big\{&8\tilde{c}_{2i-4} - c_{i-2} - 6c_{i-1},\ 2\tilde{c}_{2i-3} - c_{i-1},\ \tfrac{4}{3}\tilde{c}_{2i-2} - \tfrac{1}{6}(c_{i-1} + q_{i+1}),\\ &2\tilde{c}_{2i-1} - q_{i+1},\ 8\tilde{c}_{2i} - 6q_{i+1} - q_{i+2}\big\}, \end{aligned} \tag{3.26}$$

where we use the notation that all terms in (3.24)–(3.26) which involve c_j with $j < 1$ or q_j with $j > n$ have to be omitted.

Using (3.2) and exploiting the fact that the number of nonzero terms in each of the sums in the definitions (3.16) and (3.17) is bounded by k, the above algorithm works in optimal complexity.

THEOREM 3.13. *For fixed $k \in \mathbb{N}$, the costs of the OCGC algorithm is restricted by $O(n)$ operations.*

Next, we visualize the effect of our algorithm. In Figure 3.2, one can see a C^1-smooth, piecewise quadratic upper obstacle, the coarse grid approximation obtained by the OCGC algorithm, the coarse grid approximation $L_3 \in \mathcal{N}_{3,T}$ according to Proposition 3.4, and the optimal coarse grid approximation obtained by the simplex algorithm. (Recall, however, that the simplex algorithm does not yield the solution in optimal complexity.) The improvement of the OCGC approximation over the spline L_3 is clearly visible. There is no difference of our OCGC approximation to the optimal coarse grid approximation obtained by the simplex method, except for a slight variation in the interval [0,2]. This difference seems to be caused by boundary effects which has been confirmed in further numerical experiments. As expected, smooth parts of the obstacle are very well approximated, while variations of the obstacle

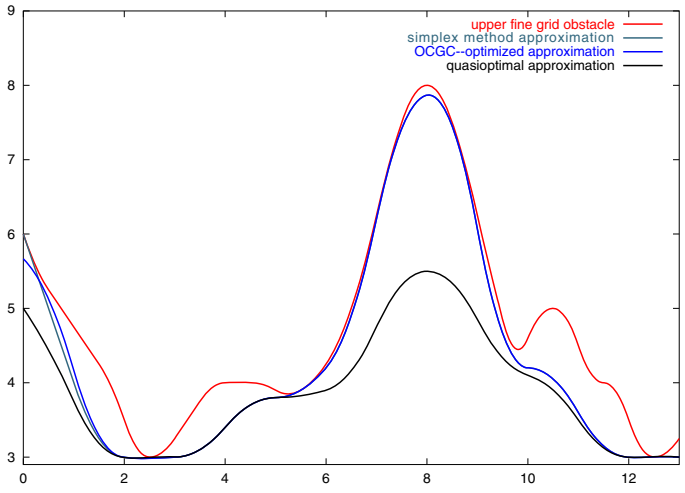


FIG. 3.2. *Right: C^1-smooth, quadratic upper obstacle function on the fine grid $\Delta := [0,13] \cap \mathbb{Z}/2$ with OCGC-optimized quadratic restriction, the optimal coarse grid approximation obtained by the simplex method and lower quasi–optimal barrier L_3, all three of which are defined on the coarse grid $T := [0,13] \cap \mathbb{Z}$.*

of higher frequency can only be partly approximated as it is visible in the interval $[10,12]$. In this example, the control polygon of the B-spline coefficients of the OCGC approximation (which is not displayed here) is partly above the control polygon of the obstacle function, although by construction the OCGC approximation always lies below the obstacle. This indicates that the result of our OCGC algorithm is superior to alternative methods in which monotone approximations are obtained via monotone restrictions of control polygons.

4. Higher spatial dimensions. In the multivariate case $\Omega \subset \mathbb{R}^d$, using (2.12), a d-dimensional spline $S : \Omega \to \mathbb{R}$ of order k can be represented by

$$S(x) = \sum_{i \in \mathbb{I}_c} c_i \, N^{(d)}_{i,k,T}(x) =: \mathbf{c}^T \, \mathbf{N}^{(d)}_{k,T}(x), \quad x \in \Omega, \tag{4.1}$$

with coefficients $\mathbf{c} \in \mathbb{R}^{n^d}$ and indices from $\mathbb{I}_c := \{i \in \mathbb{N}^d : 1 \le i_m \le n,\ m = 1, \dots, d\}$. The two-scale relation (3.4) attains the multivariate refinement relation

$$N^{(d)}_{i,k,T} = \sum_{j \in J} a^{(d)}_j \, N^{(d)}_{2i-k+j,k,\Delta} \tag{4.2}$$

with the index set $J := \{j \in \mathbb{N}^d : 0 \le j_m \le k \text{ for } m = 1, \dots, d\}$ and the subdivision coefficients

$$a^{(d)}_j := 2^{(1-k)d} \prod_{\nu=1}^{d} \binom{k}{j_\nu} \quad \text{for } j \in J. \tag{4.3}$$

The extension of Theorem 3.2 then reads as follows.

THEOREM 4.1 (monotone coarse grid approximation). *The spline* $S = \mathbf{c}^T \mathbf{N}_{k,T}^{(d)}$ *is a monotone coarse grid approximation to the upper obstacle* $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}^{(d)}$ *if their B-spline expansion coefficients satisfy the linear inequality system*

$$A_k^{(d)}\, \mathbf{c} \;\le\; \tilde{\mathbf{c}} \tag{4.4}$$

with the tensor product matrix $A_k^{(d)} := A_k \otimes \ldots \otimes A_k \in \mathbb{R}^{(\tilde{n}\times n)^d}$ *and* A_k *as in* (3.6).

Proof. The proof follows by the same arguments as in the univariate case, by using the refinement relation (4.2) and applying the multivariate version of Lemma 2.3 to

$$\tilde{S}(x) - S(x) = \sum_{i\in\mathbb{I}_f} (A_k^{(d)}\, \mathbf{c} \,-\, \tilde{\mathbf{c}})_i \, N_{i,k,\Delta}^{(d)}(x),$$

where $\mathbb{I}_f := \{i \in \mathbb{N}^d : \, 1 \le i_m \le \tilde{n},\ m = 1, \ldots, d\}$ using the nonnegativity of (tensor product) B-splines. ∎

Example 4.2. In the special case of C^2-smooth, piecewise cubic ($k = 4$) splines on a two-dimensional domain, the system (4.4) reads

$$\begin{pmatrix}
\frac{4}{8}A_4 & \frac{4}{8}A_4 & & & & & & \\
\frac{1}{8}A_4 & \frac{6}{8}A_4 & \frac{1}{8}A_4 & & & & & \\
 & \frac{4}{8}A_4 & \frac{4}{8}A_4 & & & & & \\
 & \frac{1}{8}A_4 & \frac{6}{8}A_4 & & & & & \\
 & & & \ddots & \ddots & & & \\
 & & & & & \frac{6}{8}A_4 & \frac{1}{8}A_4 & \\
 & & & & & \frac{4}{8}A_4 & \frac{4}{8}A_4 &
\end{pmatrix}
\begin{pmatrix} c_{1,1} \\ \vdots \\ c_{1,n} \\ \vdots \\ \vdots \\ c_{n,1} \\ \vdots \\ c_{n,n} \end{pmatrix}
\le
\begin{pmatrix} \tilde{c}_{1,1} \\ \vdots \\ \tilde{c}_{1,\tilde{n}} \\ \tilde{c}_{2,1} \\ \vdots \\ \tilde{c}_{2,\tilde{n}} \\ \vdots \\ \tilde{c}_{\tilde{n},1} \\ \vdots \\ \tilde{c}_{\tilde{n},\tilde{n}} \end{pmatrix}.$$

As all rows in the system (4.4) sum to one we immediately obtain from Theorem 4.1 the following generalization of Proposition 3.4.

PROPOSITION 4.3. *The spline* $L_k := \mathbf{q}^T\, \mathbf{N}_{k,T}^{(d)}$ *with expansion coefficients*

$$q_i := \min\{\tilde{c}_j : \; 2i_m - k \le j_m \le 2i_m,\ m = 1, \ldots, d\} \quad \text{for } i \in \mathbb{I}_c \tag{4.5}$$

(leaving out $\tilde{c}_j$ *in the right-hand side if* $j_m < 1$ *or* $j_m > \tilde{n}$*) is a monotone coarse grid approximation to the obstacle function* $\tilde{S} = \tilde{\mathbf{c}}^T \mathbf{N}_{k,\Delta}^{(d)}$.

In the special case $k = 2$ Fourier–Motzkin elimination can be applied to the inequality system (4.4) with the constraint $\mathbf{c} \ge \mathbf{q}$ as in the univariate case to obtain Lemma 3.8 for arbitrary d.

LEMMA 4.4 (direct solution for hat functions). *Define the sum* $s_{\ell,i}$ *of all neighboring coarse grid coefficients* c_j *to a given fine grid coefficient* $\tilde{c}_\ell$ *except* c_i *by*

$$s_{\ell,i} := \sum_{\{j\in\mathbb{I}_c:\, j\ne i,\, |\theta_{j+1} - \tilde{\theta}_{\ell+1}| < H\}} c_j \quad \text{for } \ell \in \mathbb{I}_f,\ i \in \mathbb{I}_c,$$

with the Euclidean distance $|\cdot|$ *and the mesh size* H *of the coarse grid. Define* $\bar{s}_{\ell,i}$ *as* $s_{\ell,i}$ *with the modification that all coefficients* c_j *in the sum which are not yet known*

are replaced by q_j given from (4.5). *Then, for $k = 2$, a monotone and quasi-optimal coarse grid approximation is recursively given by*

$$c_i := \min\{2^{\sum_{m=1}^d |\ell_m-(2i_m-1)|}\,\tilde{c}_\ell - \bar{s}_{\ell,i} :\ 2i_m - 2 \le \ell_m \le 2i_m \ \ \textit{for}\ \ m = 1,\dots,d\}$$

for $i \in \mathbb{I}_c$, leaving out $\tilde{c}_j$ in the right-hand side if $j_m < 1$ or $j_m > \tilde{n}$.

To improve the approximation from Proposition 4.3 in the case $k > 2$, the OCGC algorithm can be applied recursively with respect to the dimension d as follows.

THEOREM 4.5 (optimized coarse grid correction (OCGC) scheme for $d > 1$). *The OCGC algorithm applied dimension-recursively to the multivariate inequality system* (4.4) *provides in optimal complexity of $O(n^d)$ arithmetic operations a coarse grid approximation S which satisfies the monotonicity and quasi-optimality condition*

$$L_k(x) \le S(x) \le \tilde{S}(x) \quad \textit{for all } x \in \Omega.$$

Proof. We provide the proof for the case $k = 3$; the other cases follow immediately by exchanging the system matrix $A_3^{(d)}$ by $A_k^{(d)}$. The $(n \times \tilde{n})^d$ tensor product matrix $A_3^{(d)}$ can be written as a $n \times \tilde{n}$ matrix of $(n \times \tilde{n})^{(d-1)}$ tensor product matrices

$$A_3^{(d)} := \begin{pmatrix}
\frac{3}{4}A_3^{(d-1)} & \frac{1}{4}A_3^{(d-1)} & & & & & \\
\frac{1}{4}A_3^{(d-1)} & \frac{3}{4}A_3^{(d-1)} & & & & & \\
 & \frac{3}{4}A_3^{(d-1)} & \frac{1}{4}A_3^{(d-1)} & & & & \\
 & \frac{1}{4}A_3^{(d-1)} & \frac{3}{4}A_3^{(d-1)} & & & & \\
 & & & \ddots & \ddots & & \\
 & & & & \frac{3}{4}A_3^{(d-1)} & \frac{1}{4}A_3^{(d-1)} & \\
 & & & & \frac{1}{4}A_3^{(d-1)} & \frac{3}{4}A_3^{(d-1)} &
\end{pmatrix}.$$

Accordingly, we define $\mathbf{c}_{i\bullet} := (c_{ij})_{i=1,\dots,n,j\in\mathbb{N}^{(d-1)}:\,1\le j_l\le n} \in \mathbb{R}^{n^{(d-1)}}$. Applying the OCGC algorithm for $d = 1$ (3.25) to this formulation, we obtain for $i = 1,\dots n$ the conditions

$$\begin{aligned} A_3^{(d-1)}\mathbf{c}_{i\bullet} \le \min\{&4\,\tilde{c}_{2i-3\bullet} - 3A_3^{(d-1)}\,c_{i-1\bullet}, \tfrac{4}{3}\tilde{c}_{2i-2\bullet} - \tfrac{1}{3}A_3^{(d-1)}c_{i-1\bullet}, \\ &\tfrac{4}{3}\tilde{c}_{2i-1\bullet} - \tfrac{1}{3}A_3^{(d-1)}q_{i+1\bullet}, 4\,\tilde{c}_{2i\bullet} - 3A_3^{(d-1)}\,q_{i+1\bullet}\} \end{aligned}$$

for $\mathbf{c}_{i\bullet} \in \mathbb{R}^{n^{(d-1)}}$. Each of these conditions can again be reduced by one dimension by a further application of the OCGC algorithm until the whole system is solved. As the complexity of the OCGC algorithm is $O(n)$, the overall complexity is given by $O(n^d)$ arithmetic operations. ∎

Example 4.6. In the case $k = 3$ and $d = 2$ the multivariate OCGC algorithm is defined as follows:

(1) For given $\tilde{\mathbf{c}} \in \mathbb{R}^{\tilde{n}^2}$ determine $\mathbf{q} \in \mathbb{R}^{n^2}$ by (4.5).

(2) For $i = 1,\dots,n$

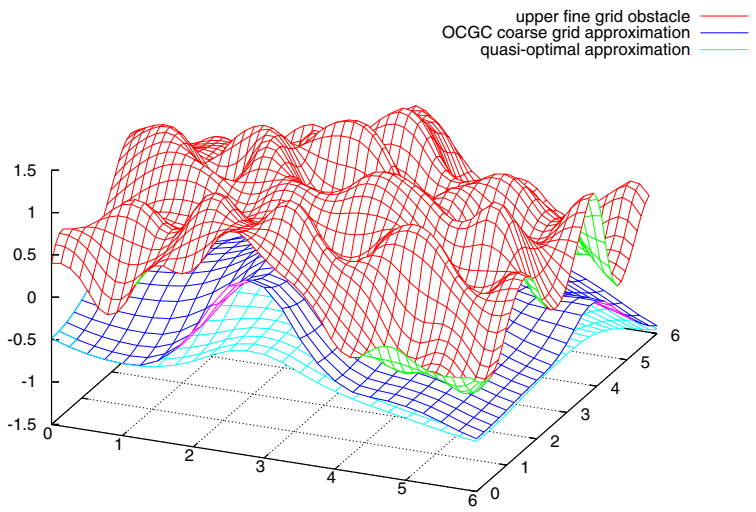


FIG. 4.1. *Two-dimensional C^1-smooth, quadratic, upper obstacle function defined on fine grid $[0,6]^2 \cap (\mathbb{Z}/2)^2$ and coarse grid approximations from Proposition* 4.3 *(quasi-optimal) and Theorem* 4.5 *(OCGC) on the coarse grid $[0,6]^2 \cap \mathbb{Z}^2$.*

define $\mathbf{g} \in \mathbb{R}^n$ by $g_j := q_{i,j}$ and $\mathbf{f} \in \mathbb{R}^{\tilde{n}}$ by $f_j := \min\{4\,\tilde{c}_{2i-3,j} - 3(A_3\, c_{i-1}\bullet)_j,$
$\frac{4}{3}\tilde{c}_{2i-2,j} - \frac{1}{3}(A_3 c_{i-1}\bullet)_j, \frac{4}{3}\tilde{c}_{2i-1,j} - \frac{1}{3}(A_3 q_{i+1}\bullet)_j, 4\,\tilde{c}_{2i,j} - 3(A_3\, q_{i+1}\bullet)_j\}$,
solve the univariate problem $A_3\mathbf{e} \le \mathbf{f}, \quad \mathbf{e} \ge \mathbf{g}$ by the 1d-OCGC Algorithm,
set $c_{i,j} := e_j$.

The splines which correspond to the coefficient vector $\mathbf{q}$ and $\mathbf{c}$ from Proposition 4.3 and Theorem 4.5, respectively, are displayed in Figure 4.1 for a given upper obstacle function defined on the fine grid $[0,6]^2 \cap (\mathbb{Z}/2)^2$.

The resulting MMG method in the multivariate case can now be implemented by adding the projection operator (2.11) and the obstacle approximation from Theorem 4.5 to a standard multigrid method. The standard multigrid method for tensor products of higher order B-splines is described, e.g., in [Hö, HRW] for the case $d > 1$.

5. Convergence theory for B-spline-based MMG methods. It is shown in [Ko1] that MMG methods are globally convergent and asymptotically reducing to a linear subspace correction method, provided nodal basis functions and monotone and quasi-optimal restriction operators $\tilde{r}$ are used. Because of the lack of such restriction operators for smooth functions, the MMG method has so far been restricted to hat functions. Using B-splines as basis functions, we have already transferred the scheme to functions of general smoothness in section 2. Suitable restriction operators have been constructed in sections 3 and 4. We have established in the extended version of this paper [HzK] that all convergence results from [Ko1] can be transferred to B-spline basis functions, using their expansion coefficients instead of function values.

6. Numerical example. To present a numerical example from the area of Mathematical Finance, we choose the domain $\Omega_{\mathcal{L}} := \mathbb{R}^+ \times [0,T)$, the differential

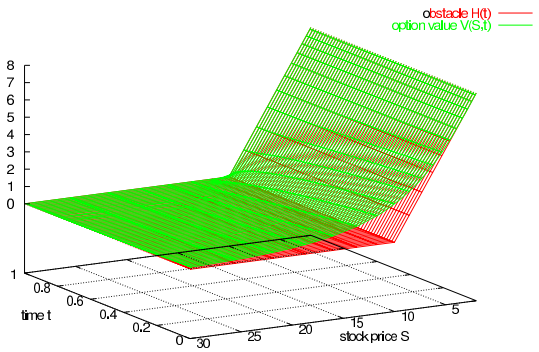


FIG. 6.1. *Solution $V(S,t)$ of the linear complementary problem* (6.2).

operator

$$\mathcal{L} := \frac{\partial}{\partial t} + \frac{1}{2}\sigma^2 S^2 \frac{\partial^2}{\partial S^2} + rS\frac{\partial}{\partial S} - r, \tag{6.1}$$

and the function $\mathcal{H}(S) := (K-S)^+$. We consider the linear complementary problem to find $V = V(S,t) \in H^1(\Omega_{\mathcal{L}})$, such that

$$\begin{aligned} [(\mathcal{L}\,V)(S,t)]\,(V(S,t) - \mathcal{H}(S)) &= 0, \\ (\mathcal{L}\,V)(S,t) &\le 0, \\ V(S,t) &\ge \mathcal{H}(S) \end{aligned} \tag{6.2}$$

holds for all $(S,t) \in \Omega_{\mathcal{L}}$, with boundary data $V(S,t) = 0$ for $S \to \infty$, $V(S,t) = \mathcal{H}(S)$ for $S \to 0$ and final data $V(S,T) = \mathcal{H}(S)$ for $S \in \mathbb{R}^+$.

As it is shown in [WHD], the solution V describes the fair value of an *American put option* with strike price K and maturity T which depends on an underlying stock with value S and volatility σ. No analytical solution is known for the problem (6.2) so that one has to resort to numerical solution schemes. In the numerical experiments we used for the linear complementary problem (6.2), the parameters $K = 10$ for the strike price, $T = 1$ for maturity, $\sigma = 0.6$ for volatility, and $r = 2.5\%$ for the interest rate. The numerical solution V and the obstacle function $\mathcal{H}$ are displayed in Figure 6.1 in the case of $M = N = 64$ grid points in space and time.

If the obstacle function is set to minus infinity, the solution V describes the fair value of a *European put option* (see [WHD]). In that case an analytical solution is known and given by the famous Black–Scholes formula; see [BS].

Using a Crank–Nicolson finite difference scheme for the time discretization and at least continuous piecewise finite elements for the space discretization, the method converges quadratically. Employing higher order finite element functions, the derivatives of the solution V which provide important hedge parameters in the option pricing context can be determined by direct differentiation of the basis functions. Using B-spline

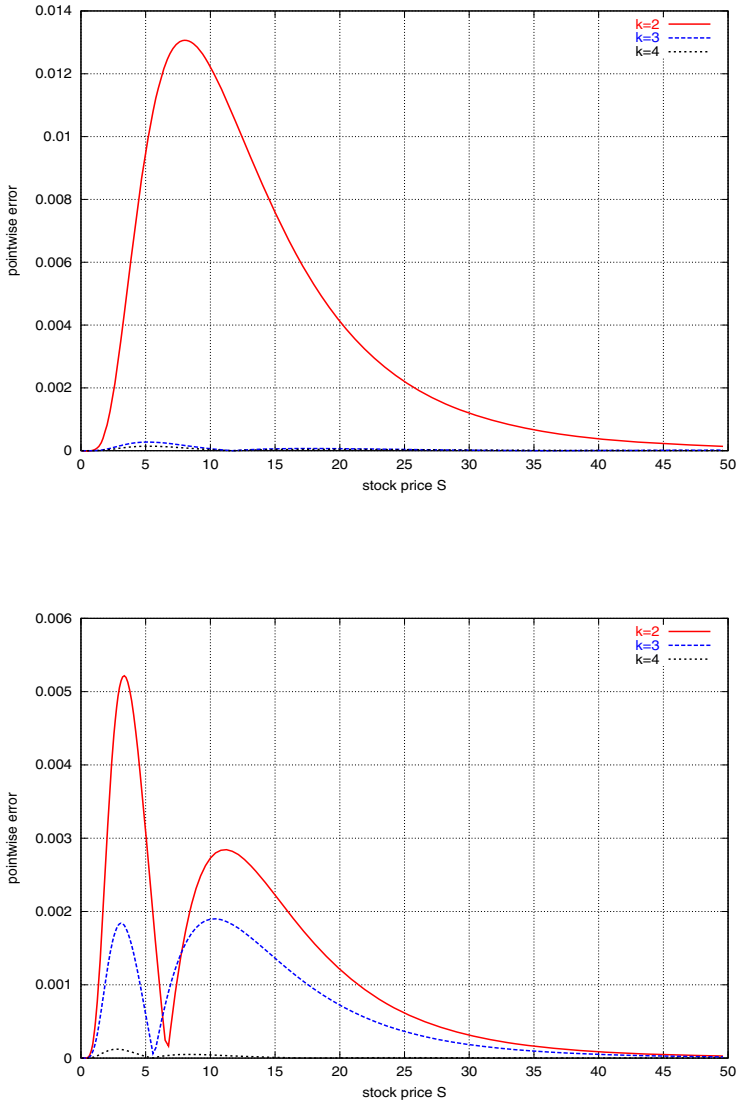


FIG. 6.2. *Comparison of pointwise error for* Delta *and* Gamma *at time* $t = 0$ *for orders* $k = 2, 3, 4$ *and* $N = M = 275$.

bases of order k we obtain all derivatives up to the $(k-2)$th derivative in quadratic convergence. In particular, pointwise derivatives, the so-called Greek letters, can be computed up to high accuracy. These results, as well as extensive discussion, can be found in [Hz]. As an illustration of the impressible difference a variable order k may offer, we display in Figure 6.2 only the pointwise errors of $Delta := \frac{\partial V}{\partial S}$ and $Gamma := \frac{\partial^2 V}{\partial S^2}$.

In view of this application, we would like to point out that our higher order MMG method could also be applied to the valuation of basket options, at least for small

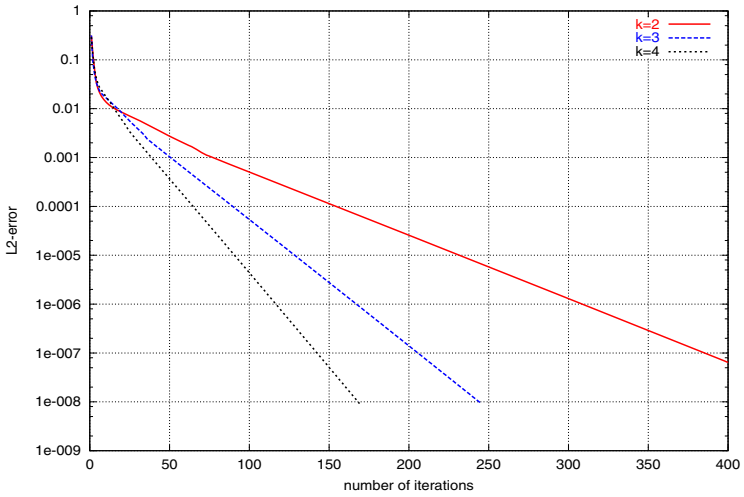


FIG. 6.3. *PSOR–iteration history of one time step with* $M = 256$.

baskets with $d = 2$ or $d = 3$. Similar to the univariate case, the multivariate Black–Scholes equation can be transformed into a multivariate heat diffusion problem, as shown in [RW].

6.1. Convergence behavior of Gauss–Seidel and MMG schemes. In the following, only one time step of problem (6.2) is considered to analyze the performance of the multigrid scheme. In Figure 6.3, the iteration errors of the projected Gauss–Seidel scheme are displayed for different orders k. The impact of the order k is clearly visible. Next we compare the convergence behavior of the following methods:

PSOR	Projected Gauss–Seidel scheme,
MMG	Monotone multigrid method with optimized approximation of the obstacle according to Lemma 3.8 and Theorem 3.10,
TrMMG	Truncated version of the monotone multigrid method [Ko1] with optimized approximation of the obstacle according to Lemma 3.8 and Theorem 3.10,
MMG (q-opt)	Monotone multigrid method with simple approximation of the obstacle according to Proposition 3.4,
TrMMG (q-opt)	Truncated version of the monotone multigrid method with simple approximation of the obstacle according to Proposition 3.4,
MG	Linear multigrid method applied to the unrestricted problem.

To analyze the influence of the order k on the convergence behavior, the case $k = 2$ is systematically compared to the case $k = 3$. For $k > 3$ similar results are expected. In the experiments of the finance parameters used in the previous section, the finest level $L = 7$ and a random initial guess have been chosen. To make sure that the iteration does not terminate too early, we have selected independently of the discretization error the stopping criterion

$$\|u_L^{\nu+1} - u_L^{\nu}\|_\infty \leq 10^{-12},$$

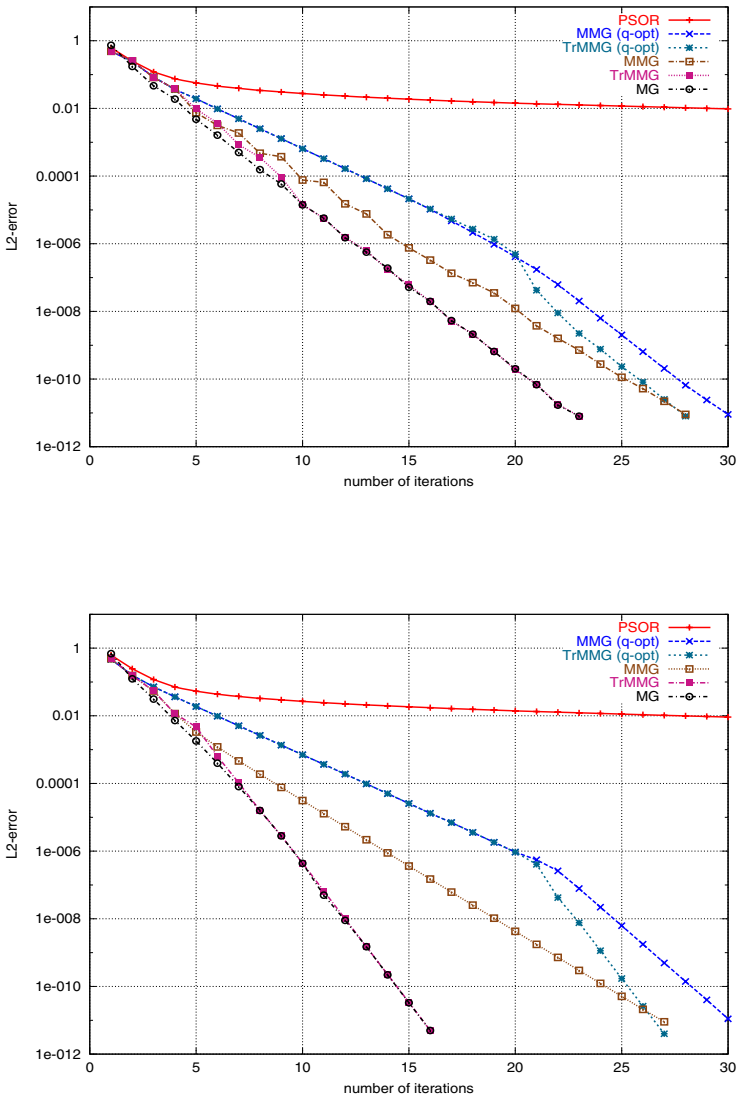


FIG. 6.4. *Iteration history for hat functions ($k = 2$) (top) and for C^1-smooth basis functions ($k = 3$) (bottom).*

where u_L^ν denotes the νth iterate on the finest grid L.

The numerical results are summarized in Figure 6.4 and Table 6.1. In the third column in Table 6.1, the number ν_0 of iterations needed to identify the contact set $K^\bullet(u_L)$ is displayed. In the next column $\sharp$`It.`, we list the number of iterations which are needed to solve the problem up to machine accuracy. To compare the costs of the schemes, we employ the definition of a work unit (`WU`) from [BC]. A work unit $\texttt{WU} = \texttt{WU}_L$ denotes the costs of one iteration step of the projected Gauss–Seidel scheme on the finest grid L. The costs $\texttt{WU}_\ell$ of one iteration step on level $\ell \leq L$ is then given

	Scheme	1 smoothing step			2 smoothing steps		
		ν_0	♯ It.	♯WU	ν_0	♯ It.	♯ WU
$k=2$	PSOR	134	403	403	—	—	—
	MMG (q-opt)	6	30	59.06	5	21	82.69
	TrMMG (q-opt)	7	28	55.13	5	17	66.94
	MMG	7	28	55.13	5	14	55.13
	TrMMG	7	23	45.28	5	13	51.19
$k=3$	PSOR	103	447	447	—	—	—
	MMG (q-opt)	5	31	61.03	4	20	78.75
	TrMMG (q-opt)	6	27	53.16	4	17	66.94
	MMG	5	27	53.16	4	14	55.13
	TrMMG	5	16	31.5	4	11	43.31

TABLE 6.1

Number of iterations needed to identify the contact set and to compute the solution up to machine accuracy and the cost in work units.

by

$$\mathtt{WU}_\ell = 2^{L-\ell}\,\mathtt{WU}_L.$$

The number of work units which is needed to reach the stop criteria is displayed in the last column ♯WU in Table 6.1.

The numerical results show that already one or two smoothing steps are sufficient with regard to cost and accuracy. In comparison to the Gauss–Seidel relaxation, the cost is substantially reduced in the multigrid schemes. The truncated versions TrMMG and TrMMG (q-opt) converge in all cases faster than the standard versions MMG or MMG (q-opt). Moreover, multigrid methods with an optimized approximation of the obstacle according to Lemma 3.8 or Theorem 3.10 converge faster than the simple approximations according to Proposition 3.4. For hat functions, this corresponds to the results in [Ko1]. For the higher order case, this indicates the quality of the OCGC approximations from section 3.4. The contact set is identified correctly by all methods within only a few iterations.

Considering the above results within the time discretization when solving the instationary problem, we wish to point out that the average number of iterations per time step is much smaller. This is due to the fact that the solution of the previous time step serves as a good initial guess. Therefore, we can expect that the asymptotic phase dominates the convergence behavior of the multigrid scheme. The asymptotic multigrid rates are discussed in the following section.

6.2. Multigrid convergence rates. The convergence rate ρ_ℓ of a multigrid scheme with $\ell+1$ levels is given by

$$\|u_\ell^{\nu+1} - u_\ell\|_{\ell_2} \le \rho_\ell \|u_\ell^{\nu} - u_\ell\|_{\ell_2}.$$

Here $u_\ell \in S_\ell$ denotes the exact solution and $u_\ell^\nu \in S_\ell$ the approximate solution in the νth iteration step. A scheme is said to have multigrid convergence if ρ_ℓ is bounded independently of the grid size by a constant $\rho_\infty < 1$.

The asymptotic convergence rates are estimated for the V-cycle of the truncated version TrMMG with $\ell+1$ levels according to

$$\rho_\ell \approx \frac{\|u_\ell^{\nu^*+1} - u_\ell^{\nu^*}\|_{\ell_2}}{\|u_\ell^{\nu^*} - u_\ell^{\nu^*-1}\|_{\ell_2}}.$$

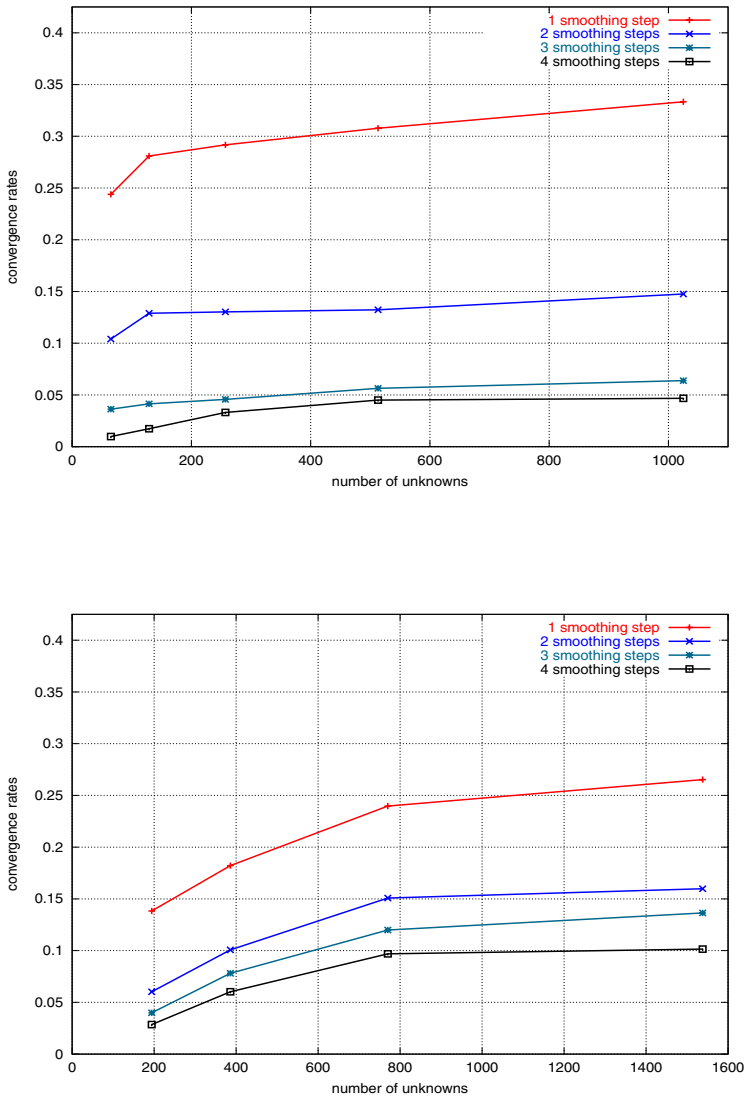


FIG. 6.5. *Asymptotic convergence rates for the case $k = 2$ (left) and $k = 3$ (right) depending on the number M of unknowns.*

Here ν^* is chosen such that $\|u_\ell^{\nu^*+1} - u_\ell^{\nu^*}\|_{\ell_2} \leq 10^{-12}$. In Figure 6.5 the results are displayed on the left-hand side for continuous, piecewise linear basis functions and on the right-hand side for C^1-smooth, piecewise quadratic basis functions. We recover the favorable convergence rates of standard multigrid schemes which are bounded in our case by $\rho_\infty \approx 0.31$ $(k = 2)$ and $\rho_\infty \approx 0.27$ $(k = 3)$ in the case of only one smoothing step on each refinement level.

Acknowledgments. We would like to thank Michael Griebel for pointing out the problem of deriving MMG methods based on higher order basis functions and their possible application to the computation of American options. We also want to thank Rolf Krause for helpful discussions on MMG methods.

REFERENCES

[BC] A. Brandt and C. W. Cryer, *Multigrid algorithms for the solution of linear complementarity problems arising from free boundary problems*, SIAM J. Sci. Statist. Comput., 4 (1983), pp. 655–684.

[BHR] F. Brezzi, W. W. Hager, and P. A. Raviart, *Error estimates for the finite element solution of variational inequalities*, Numer. Math., 28 (1977), pp. 431–443.

[BBS] H. Blum, D. Braess, and F. T. Suttmeier, *A cascadic multigrid algorithm for variational inequalities*, Comput. Vis. Sci., 7 (2004), pp. 153–157.

[Bo] C. de Boor, *A Practical Guide to Splines*, revised edition, Appl. Math. Sci. 27, Springer-Verlag, New York, 2001.

[BS] F. Black and M. Scholes, *The pricing of options and corporate liabilities*, J. Polit. Econ., 81 (1973), pp. 637–654.

[Cr] C. W. Cryer, *The solution of a quadratic programming problem using systematic overrelaxation*, SIAM J. Control, 9 (1971), pp. 385–392.

[DV] R. DeVore, *One-sided approximation of functions*, J. Approximation Theory, 1 (1968), pp. 11–25.

[EO] C. M. Elliott and J. K. Ockendon, *Weak and Variational Methods for Moving Boundary Problems*, Res. Notes in Math. 59, Pitman, Boston, 1982.

[Ha] W. Hackbusch, *Multigrid Methods and Applications*, Springer Ser. Comput. Math. 4, Springer-Verlag, Berlin, 1985.

[Hö] K. Höllig, *Finite Element Methods with B-Splines*, Frontiers Appl. Math. 26, SIAM, Philadelphia, 2003.

[HRW] K. Höllig, U. Reif, and J. Wipper, *Multigrid methods with web-splines*, Numer. Math., 91 (2002), pp. 237–256.

[HM] W. Hackbusch and H.-D. Mittelmann, *On multigrid methods for variational inequalities*, Numer. Math., 42 (1983), pp. 65–76.

[Hz] M. Holtz, *The Computation of American Option Price Sensitivities using a Monotone Multigrid Method for Higher Order B-spline Discretizations*, manuscript, 2004, submitted, in revision.

[HzK] M. Holtz and A. Kunoth, *B-Spline-based Monotone Multigrid Methods—Extended Version*, manuscript, 2005, available online at http://www.iam.uni-bonn.de/~kunoth/papers/papers.html.

[Ho] R. H. W. Hoppe, *Multigrid algorithms for variational inequalities*, SIAM J. Numer. Anal., 24 (1987), pp. 1046–1065.

[Ko1] R. Kornhuber, *Monotone multigrid methods for elliptic variational inequalities,* I, Numer. Math., 69 (1994), pp. 167–184.

[Ko2] R. Kornhuber, *Adaptive Monotone Multigrid Methods for Nonlinear Variational Problems*, B. G. Teubner, Stuttgart, Germany, 1997.

[Kr] R. Krause, *Monotone Multigrid Methods for Signorini's Problem with Friction*, Dissertation, FU Berlin, Berlin, Germany, 2001.

[KS] D. Kinderlehrer and G. Stampacchia, *An Introduction to Variational Inequalities and their Applications*, Academic Press, New York, 1980.

[Ma] J. Mandel, *A multilevel iterative method for symmetric, positive definite linear complementarity problems*, Appl. Math. Optim., 11 (1984), pp. 77–95.

[Mv] E. R. Matveev, *On a super one-sided spline approximation of functions of several variables*, Izvestiya VUZ. Matematika, 32 (1988), pp. 49–54.

[Pi] A. M. Pinkus, *On L^1-Approximation*, Cambridge Tracts in Math. 93, Cambridge University Press, Cambridge, UK, 1989.

[RW] C. Reisinger and G. Wittum, *On multigrid for anisotropic equations and variational inequalities: Pricing multi-dimensional European and American options*, Comput. Vis. Sci., 7 (2004), pp. 189–197.

[Sb] I. J. Schoenberg, *Contributions to the problem of approximation of equidistant data by analytic functions*, Quart. Appl. Math., 4 (1946), pp. 45–99.

[Sj] A. Schrijver, *Theory of Linear and Integer Programming*, John Wiley and Sons, Chichester, UK, 1986.

[WHD] P. Wilmott, S. Howison, and J. Dewynne, *The Mathematics of Financial Derivatives*, Cambridge University Press, Cambridge, UK, 1995.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1200–1225

SMOOTHNESS EQUIVALENCE PROPERTIES OF MANIFOLD-VALUED DATA SUBDIVISION SCHEMES BASED ON THE PROJECTION APPROACH*

GANG XIE† AND THOMAS P.-Y. YU†

Abstract. Interpolation of manifold-valued data is a fundamental problem which has applications in many fields. The linear subdivision method is an efficient and well-studied method for interpolating or approximating real-valued data in a multiresolution fashion. A natural way to apply a linear subdivision scheme $\overline{S}$ to interpolate manifold-valued data is to first embed the manifold at hand into an Euclidean space and construct a projection operator P that maps points from the ambient space to a closest point on the embedded surface, and then consider the *nonlinear* subdivision operator $S := P \circ \overline{S}$. When applied to symmetric spaces such as S^{n-1}, $SO(n)$, $SL(n)$, $SE(n)$, $G(n,k)$ the projection method can also be carried out in such a way that the resulting schemes enjoy natural coordinate invariance properties and robust numerical implementations. Despite such nice features, the mathematical analysis of such nonlinear subdivision schemes is in its infancy. In this article, we attack the so-called smoothness equivalence conjecture, which asserts that the smoothness property of S is exactly the same as that of $\overline{S}$. We show that in the cases of S^{n-1}, $SO(n)$, and related manifolds, we have a proximity condition of the form $|(S-\overline{S})y|_\infty \cdot \sum_{i=1}^{p-1} |\Delta^i y|_\infty |\Delta^{p-i} y|_\infty$, where p is the accuracy order of $\overline{S}$. Armed with this proximity condition and other known approximation theoretic results, we can establish the result that the Hölder smoothness exponent of S is always as high as that of $\overline{S}$—no matter how high the latter is.

Key words. linear subdivision schemes, nonlinear subdivision schemes, Riemannian manifold, Lie groups, sphere, $SO(n)$, smoothness, approximation order, closest point projection, interpolation, singular value decomposition

AMS subject classifications. 26A15, 26A16, 26A18, 41A05, 42C40

DOI. 10.1137/060652944

1. Introduction.

Given a smooth manifold $\mathcal{M}$ and a sequence of points $p_i \in \mathcal{M}$ on the manifold (called a control polygon), a fundamental problem is to find a smooth curve that either interpolates or approximates the control polygon. A typical application in robotics and computer vision is rigid body motion interpolation, in which the manifold is the Lie group $\mathcal{M} = SE(3)$ [2]. In the curve design problem on the sphere described in [27], we have $\mathcal{M} = \mathbb{S}^2$. In the "grand tour" method for visualizing p-dimensional data points based on ortho-projection of the points to two-dimensional subspaces [1], we have the Grassmannian manifold $\mathcal{M} = G(p,2)$. Time series that take values on a Grassmannian manifold $G(n,k)$ also arise in array signal processing. There is also a lot of interest in numerical analysis for interpolation in Lie groups, especially in the geometric methods for numerical solution of ODEs; see, e.g., [20, 18, 17].

Some, if not many, of the existing methods for the manifold interpolation problem appear to be quite specific to the manifold at hand. On the other hand, in certain applications it is quite desirable that the curve be given in various levels of detail; in this case subdivision-based methods would be very attractive.

*Received by the editors February 25, 2006; accepted for publication (in revised form) October 31, 2006; published electronically May 16, 2007.
http://www.siam.org/journals/sinum/45-3/65294.html

†Department of Mathematics, Drexel University, 3141 Chestnut Street, 206 Korman Center, Philadelphia, PA 19104 (gangxie2006@gmail.com, yut@drexel.edu; http://www.math.drexel.edu/∼tyu). This research was partially supported by the National Science Foundation grants CCF 9984501 (CAREER Award) and DMS 0512673.

In recent papers [30, 24, 31], various general subdivision methods have been proposed which can be used to subdivide data that takes values on a manifold (or data that obeys some other forms of nonlinear constraints). A common feature of these methods is that they are all based on adapting a linear subdivision scheme to a manifold, resulting in nonlinear algorithms that are easy to implement but difficult to analyze. Since in each case the method is based on taking a *linear* subdivision scheme and applying it to subdivision of data obeying *nonlinear* constraints, we refer to such a method as a *linearization* method. Here, we discuss the following three classes of linearization methods:

- *Tangent plane approach*: In this approach, the basic operation is that for any given $p \in \mathcal{M}$ there is a map f_p with inverse f_p^{-1} such that the map points in a neighborhood of p back and forth between the manifold and the tangent plane $T_p\mathcal{M}$. Then a subdivided point is obtained from a group of points in the coarser scale by first mapping these points to the tangent plane based at one of these points using the corresponding map f_p, then applying the linear subdivision rule in $T_p\mathcal{M}$ (a linear space) and mapping the subdivided point in the tangent plane back to the manifold using f_p^{-1}. The specific f_p and f_p^{-1} proposed in [24] are the logarithmic and exponential maps (in either the setting of a Riemannian manifold or Lie group; see, e.g., [3, 11]); we refer to this linearization method as the *Log-Exp scheme*.
- *Factorization-geodesic scheme*: The mask of any linear subdivision scheme can be factorized (in a nonunique way) into a number of two point weighted averages [30, Theorem 1]. On any Riemannian manifold, the notion of geodesic is well defined, and so one can perform subdivision based on replacing the weighted averages by "weighted geodesic averages." See [30] for details.
- *Projection approach*: In this approach $\mathcal{M}$ is supposed to be immersed or embedded in an Euclidean space $\mathbb{R}^n$; as such, we view $\mathcal{M}$ as a subset of $\mathbb{R}^n$. Also a projection operator that maps points in a neighborhood of $\mathcal{M}$ onto $\mathcal{M}$ is chosen. A natural example of such an operator would be the one that maps a $p \in \mathbb{R}^n$ to a point in $\mathcal{M}$ *closest* to p. We refer to this as the *closest point projection scheme*. A subdivision step is based on first applying the linear subdivision rule in $\mathbb{R}^n$, resulting in points that are typically not in $\mathcal{M}$, then followed by applying the projection operator to force these subdivided points back to $\mathcal{M}$.

Any of the above linearization methods can be used in conjunction with an arbitrary linear subdivision scheme.

The Log-Exp scheme and the factorization-geodesic schemes are both based on geodesics and are *intrinsic* in nature, whereas the projection approach is *extrinsic* in nature: in the projection approach, one may have two isometric embeddings of a given Riemannian manifold $(\mathcal{M}, g)$ into two Euclidean spaces, and the (closest point, say) projection scheme may result in two different curves on $\mathcal{M}$ for the same set of initial points on $\mathcal{M}$. From a practical point of view, the projection method is probably the most natural to use for manifolds with a "natural" embedding[1] into an Euclidean space, e.g., $\mathbb{S}^n \hookrightarrow \mathbb{R}^{n+1}$, $SO(n) \hookrightarrow SL(n) \hookrightarrow GL(n) \hookrightarrow \mathbb{R}^{n\times n}$, or $SE(n) \hookrightarrow GL(n+1) \hookrightarrow \mathbb{R}^{(n+1)\times(n+1)}$.

[1] If G is a group that acts transitively on $\mathcal{M}$, some authors call an embedding $\Phi : \mathcal{M} \to \mathbb{R}^n$ *natural* if there is a smooth group homomorphism $\rho : G \to SE(n)$ such that $\phi(g \cdot x) = \rho(g) \cdot \Phi(x)$ for all $x \in \mathcal{M}$ and $g \in G$.

1.1. Linear subdivision schemes. We recall in this section some of the basic definitions, notions, and notation related to linear subdivision schemes. We keep the exposition to the minimum, as there are plenty of references on this topic; see, for example, [4, 14, 25, 6, 7] and the references therein.

In the simplest setting, a linear stationary subdivision scheme is defined by a linear operator S on $\ell(\mathbb{Z}) := \{x \mid x : \mathbb{Z} \to \mathbb{R}\}$ of the form

$$(Sx)(2k) = \sum_{i\in\mathbb{Z}} x(i)\, a_e(k-i), \quad (Sx)(2k+1) = \sum_{i\in\mathbb{Z}} x(i)\, a_o(k-i), \tag{1.1}$$

where a_e and a_o are two finitely supported real-valued sequences such that $\sum_i a_e(i) = \sum_i a_o(i) = 1$. This operator is usually written by analysts in the following more compact form: $(Sx)(k) = \sum_{i\in\mathbb{Z}} x(i)\, \mathbf{a}(k-2i)$. Here, $\mathbf{a}$ is called the *mask* of the subdivision scheme and can be easily assembled from the a_o and a_e above.

If $a_e = \delta_0$ (the Kronecker sequence), then we say S is *interpolatory*. Interpolatory subdivision schemes were first studied in [12, 9].

A subdivision operator S is meant to be *iterated*. Moreover, for any initial sequence $v : \mathbb{Z} \to \mathbb{R}$, one is supposed to visualize $S^j v$ as a function on the grid $2^{-j}\mathbb{Z}$, as opposed to a function on $\mathbb{Z}$ as our mathematical notation may unduly suggest. We say S is *convergent* if for any $v \in \ell(\mathbb{Z})$ the sequence $f_j := \sum_{k\in\mathbb{Z}} v_{j,k} 1_{[2^{-j}k,\, 2^{-j}(k+1))}$, $j = 0, 1, 2, \ldots$, $v_j := S^j v$, converges uniformly on compact sets to a limit function; we denote this (necessarily unique) limit function by $S^\infty v$. For a convergent subdivision operator S, we define its critical Hölder smoothness (a.k.a. L^∞-Lipschitz smoothness) by

$$s_\infty(S) := \inf_{v\in\ell^\infty} \sup\{\alpha : S^\infty v \in \operatorname{Lip}\alpha\}. \tag{1.2}$$

While a subdivision scheme S as defined above operates on scalars, one can apply it in a componentwise fashion to m-vectors; this, in particular, gives a practical curve drawing algorithm with input being a coarse control polygon in $\mathbb{R}^m$ ($m = 2$ or 3). When we later write Sy, where y is a sequence of m-vectors, this is to be interpreted as the componentwise application of S to y.

For any (finite or infinite) sequence of m-vectors $y = (y_i)_{i\in\mathcal{I}}$, $y_i \in \mathbb{R}^m$, $\mathcal{I} = \mathbb{Z}$ or $\{1, \ldots, n\}$, we write

$$|y|_\infty := \sup_{i\in\mathcal{I}} \|y_i\|_2$$

and define its difference sequences by $(\Delta y)_i := y_i - y_{i-1}$, $(\Delta^k y)_i := (\Delta^{k-1}y)_i - (\Delta^{k-1}y)_{i-1}$, $k > 1$, where i ranges through the appropriate set of indices when $\mathcal{I}$ is finite. We denote by $\ell^\infty(\mathbb{Z} \to \mathbb{R}^m)$, or simply ℓ^∞ when there is no source of confusion, the space of all sequences $y : \mathbb{Z} \to \mathbb{R}^m$ such that $|y|_\infty$ is finite.

An interpolatory subdivision scheme S is said to have approximation order p ($\in \mathbb{Z}_+$) if for any bounded C^p function $f : \mathbb{R} \to \mathbb{R}$ the interpolant of f on the grid $h\mathbb{Z}$ defined by

$$f_h := (S^\infty v)(\cdot/h), \quad v_k = f(kh), \tag{1.3}$$

satisfies

$$\|f - f_h\|_\infty = O(h^p), \quad h \to 0.$$

For an interpolatory scheme S, it is not hard to show that, based on simple twists of the arguments already presented by Dubuc [12], S has approximation order p if and only if S reproduces the polynomial space Π_{p-1}, i.e., $S(p\,|_{\mathbb{Z}}) = p\,|_{\frac{1}{2}\mathbb{Z}}$ for all $p \in \Pi_{p-1}$. This condition imposes a set of linear conditions on the mask of S.

If $s_\infty(S) > p$, then S must have approximation order $p+1$; the converse is far from the truth: an interpolatory scheme can have an arbitrarily high approximation order but an arbitrarily low smoothness.

1.2. Smoothness equivalence. Intuitively, both the Log-Exp scheme and the factorization-geodesic scheme try to use geodesics to mimic weighted averages (1.1) on a manifold locally. Hence, one may expect that the nonlinearities presented in the resulting schemes are rather weak *at fine subdivision levels*. This seems to be true in the sense that the nonlinearity at fine scales is too weak to affect either the smoothness or the approximation order. We discuss smoothness in this paper and approximation order in a companion paper [35].

From many computational experiments both the Log-Exp scheme and the factorization-geodesic scheme are observed empirically to produce limit paths on the manifold with Hölder regularity exactly the same as that produced by the underlying linear scheme [24, 36]. This is the so-called *smoothness equivalence* property and is conjectured to hold for both the Log-Exp scheme and the factorization-geodesic scheme when applied in conjunction with any linear subdivision scheme on any C^∞ Riemannian manifold.

The smoothness equivalence conjecture is true for both Log-Exp and factorization-geodesic schemes if $\mathcal{M}$ is an Abelian Lie group (e.g., the n-torus $\mathbb{T}^n$) viewed as a Riemannian manifold with a bi-invariant metric. (Caution: For $\mathbb{T}^n$, we are referring to the so-called flat torus; when $n = 2$, it is diffeomorphic, but not isometric, to the bagel-like torus as drawn in $\mathbb{R}^3$ with the induced Riemannian metric from $\mathbb{R}^3$. We still believe that the conjecture is true for the "bagel-like torus," just that the proof is not going to be as trivial.)

The goal of this paper is to study the smoothness equivalence properties of the closest point projection scheme, which is *not* based on geodesics. Our empirical experiments suggest that the corresponding smoothness equivalence conjecture for closest point projection schemes also holds true. We note that, even in the case of the circle $\mathbb{S}^1$ ($\cong \mathbb{T}^1$), the proof is, unlike that for Log-Exp or factorization-geodesic, already not trivial.

1.3. Warming up on the circle. As a warmup, we consider here the case when $\mathcal{M} = \mathbb{S}^1 = \{x \in \mathbb{R}^2 : \|x\|_2 = 1\}$ and the linear subdivision schemes are Deslauriers–Dubuc $2L$-point interpolatory subdivision schemes. Here the metric on $\mathbb{S}^1$ is induced by the standard Euclidean metric of $\mathbb{R}^2$. In this case the geodesic distance is just the arc length on the circle, and the Log-Exp and factorization-geodesic schemes are essentially the same. If $p_{0,k} = [\cos(\theta_{0,k}), \sin(\theta_{0,k})]^T \in \mathbb{S}^1$, and we assume for simplicity that all $\theta_{0,k} \in (-\delta, \delta)$ for a not so big δ, then either scheme generates the subdivided points $p_{j,k} = [\cos(\theta_{j,k}), \sin(\theta_{j,k})]^T$ based on *linearly* subdividing the *angles* $\theta_{j,k}$ using the linear rule:

$$\theta_{j+1,2k} = \theta_{j,k}, \quad \theta_{j+1,2k+1} = \sum_{i=-L+1}^{L} w_i \theta_{j,k+i}. \tag{1.4}$$

Here the w_i's are the mask entries of the $2L$-point Deslauriers–Dubuc scheme; e.g., $w_0 = w_1 = 9/16$, $w_{-1} = w_2 = -1/16$ if $L = 2$. In this case, the Log-Exp scheme

produces a limit function $f : \mathbb{R} \to \mathbb{S}^1$ of the form $f(t) = [\cos(\theta(t)), \sin(\theta(t))]^T$, where $\theta(t)$ is the scalar-valued limit function obtained by applying the Deslauriers–Dubuc scheme to the initial *scalar* data $(\theta_{0,k})_{k\in\mathbb{Z}}$. Thus smoothness equivalence between the linear subdivision scheme and the associated "nonlinear" scheme is readily clear.

On the other hand, the nonlinear subdivision scheme based on closest point projection gives: $p_{j+1,2k} = p_{j,k}$, $p_{j+1,2k+1} = \sum_{i=-L+1}^{L} w_i p_{j,k+i} / \| \sum_{i=-L+1}^{L} w_i p_{j,k+i} \|_2$, or

$$\theta_{j+1,2k} = \theta_{j,k}, \quad \theta_{j+1,2k+1} = \arctan \frac{\sum_{i=-L+1}^{L} w_i \sin(\theta_{j,k+i})}{\sum_{i=-L+1}^{L} w_i \cos(\theta_{j,k+i})}.$$

The smoothness equivalence property is far less obvious in this case. Using Taylor's theorem, one can show that

$$\arctan \frac{\sum_{i=-L+1}^{L} w_i \sin(\theta_{j,k+i})}{\sum_{i=-L+1}^{L} w_i \cos(\theta_{j,k+i})} = \sum_{i=-L+1}^{L} w_i \theta_{j,k+i} + O\left(|\Delta\theta_j|_\infty^3\right). \tag{1.5}$$

Using the proximity and perturbation arguments detailed in this article, the above estimate allows us to conclude that the closest point projection scheme on $\mathbb{S}^1$ shares the same Hölder regularity of the underlying linear subdivision scheme *if the linear scheme has a critical Hölder regularity* $\leqslant 3$. However, when $L \geqslant 4$, the Deslauriers–Dubuc schemes have critical Hölder regularity > 3, and the above estimate is insufficient for proving smoothness equivalence. We can still use (1.5) to conclude that the nonlinear scheme has critical Hölder smoothness no less than 3, but we can no longer conclude that the nonlinear scheme is as smooth as the linear scheme.

We note that (1.5) is a kind of proximity condition between a linear and a nonlinear scheme, similar to, but not the same as, the proximity results in [30]. Using either a general result in [30] or the first part of Theorem 3.7, we have

$$\frac{\sum_{i=-L+1}^{L} w_i p_{j,k+i}}{\| \sum_{i=-L+1}^{L} w_i p_{j,k+i} \|} = \sum_{i=-L+1}^{L} w_i p_{j,k+i} + O\left(|\Delta p_j|_\infty^2\right). \tag{1.6}$$

This is yet another proximity condition. While (1.6) may seem more natural than (1.5), the former is not as powerful as the latter: (1.6) allows us to conclude smoothness equivalence only when the linear scheme has smoothness $\leqslant 2$.

Regardless, neither (1.6) nor (1.5) is nearly good enough to prove the smoothness equivalence conjecture of the $\mathbb{S}^1$-closest point projection scheme based on Deslauriers–Dubuc schemes for an *arbitrary* L, as it is well known that the smoothness of Deslauriers-Dubuc schemes grows unboundedly as $L \to \infty$.

1.4. Contributions and organization of this paper. This paper aims to attack the *arbitrary degree* smoothness equivalence conjectures as described in previous sections. We prove that smoothness equivalence holds when the closest point projection scheme is applied in conjunction with an *arbitrary* linear interpolatory subdivision scheme to data that takes values on $\mathbb{S}^n$, $SO(n)$, and their direct products.

These results stand in contrast to the recent low degree smoothness equivalence results in [30, 29, 36]. Related low degree smoothness equivalence results of nonlinear subdivision schemes arising from other applications can be found in [33, 32, 8].

The idea—and limitation—of [30, 36] are that by bounding the nonlinearity using a quadratic term, one can show smoothness equivalence when—and only when—the

linear scheme has smoothness $\leqslant 2$. Wallner's preprint [29] shows how one may go from "$\leqslant 2$" to "$\leqslant 3$" for certain linear schemes used in conjunction with the factorization-geodesic and the projection schemes.

The key discovery reported in this paper is that the polynomial reproducibility property of a linear interpolatory subdivision scheme $\overline{S}$ can lend itself to a useful "factorization" of the nonlinearity (section 3.1): by bounding the nonlinearity of the $\mathbb{S}^n$-closest point projection scheme expressed in *specific* quadratic terms that involve the interpolatory subdivision mask (see (3.7)-(3.9)), we can somehow lift the quadratic proximity bound $|(S-\overline{S})y|_\infty = O(|\Delta y|_\infty^2)$ into a higher order proximity bound

$$(*) \qquad |(S-\overline{S})y|_\infty \leqslant B_p \sum_{i=1}^{p-1} |\Delta^i y|_\infty |\Delta^{p-i} y|_\infty$$

when $\overline{S}$ *reproduces* Π_{p-1}.

Once $(*)$ is proved, we have a strong enough proximity condition that allows us, when it is combined with the perturbation theorem [8, Theorem 3.3] and a "bootstrapping" argument that relies on the interpolatory property of the subdivision scheme, to prove the desired arbitrary degree smoothness equivalence result. See section 3.2.

In section 3.3, we show how to relax the closest point projection operator to a near-closest projection operation without jeopardizing the arbitrary degree smoothness equivalence property.

In section 4, we show how to extend the main smoothness equivalence result to $SO(n)$, $SE(n)$, and related manifolds.

We discuss in section 5 possible extensions of our results.

In section 2, we streamline some of the general proximity results in [30, 29]. The proofs of a number of lemmas and theorems are recorded in the appendix.

2. General proximity results. The following theorem is a restatement of [29, Theorem 2]. It is an extension of [30, Theorem 2].

THEOREM 2.1. *Let* $T_1, T_2 : \ell^\infty \to \ell^\infty$. *Suppose that there exist* $\delta, A > 0, \mu \in (0,1)$, *and* $\alpha > 1$ *such that for any* $p \in \ell^\infty$ *with* $|\Delta p|_\infty < \delta$ *we have*

$$|\Delta T_1^j p|_\infty \leqslant \mu^j |\Delta p|_\infty \quad \forall j \in \mathbb{N}, \tag{2.1}$$

$$|T_1 p - T_2 p|_\infty \leqslant A |\Delta p|_\infty^\alpha.$$

Then for any $\epsilon > 0$ *there exists* $\delta' > 0$ *such that when* $|\Delta p|_\infty < \delta'$,

$$|\Delta T_2^j p|_\infty \leqslant (\mu + \epsilon)^j |\Delta p|_\infty \quad \forall j \in \mathbb{N}.$$

Our goal is to extend the above theorem to Theorem 2.4. For this purpose, we need the next two lemmas.

LEMMA 2.2. *Let* $T_1, T_2 : \ell^\infty \to \ell^\infty$. *Suppose that there exist* $C, \delta, A > 0$ *and* $\alpha > 1$ *such that for all* $p \in \ell^\infty$ *with* $|\Delta p|_\infty < \delta$ *we have*

$$|\Delta T_1 p|_\infty \leqslant C |\Delta p|_\infty, \tag{2.2}$$

$$|T_1 p - T_2 p|_\infty \leqslant A |\Delta p|_\infty^\alpha. \tag{2.3}$$

Then there exist $C' > 0$ *and* $\delta' > 0$ *such that when* $|\Delta p|_\infty < \delta'$,

$$|\Delta T_2 p|_\infty \leqslant C' |\Delta p|_\infty. \tag{2.4}$$

Proof. See Appendix A.1 for the proof.

The following lemma is an extension of [29, Lemma 2].

LEMMA 2.3. *Let $T_1, T_2 : \ell^\infty \to \ell^\infty$. Suppose that there exist $C, A, \delta > 0$ and $\alpha > 1$ such that for all $p \in \ell^\infty$ with $|\Delta p|_\infty < \delta$ we have*

$$|\Delta T_1 p|_\infty \leqslant C|\Delta p|_\infty, \tag{2.5}$$

$$|T_1 p - T_2 p|_\infty \leqslant A|\Delta p|_\infty^\alpha. \tag{2.6}$$

If T_1 is bounded and linear, then for any $j \in \mathbb{N}$ there exist $\delta_j, C_j > 0$ such that when $|\Delta p|_\infty < \delta_j$,

$$|T_1^j p - T_2^j p|_\infty \leqslant C_j|\Delta p|_\infty^\alpha. \tag{2.7}$$

Proof. See Appendix A.2 for the proof.

THEOREM 2.4. *Let $T_1, T_2 : \ell^\infty \to \ell^\infty$. Suppose that there exist $C, A, \delta > 0$, $\mu \in (0, 1)$, and $\alpha > 1$ such that for all $p \in \ell^\infty$ with $|\Delta p|_\infty < \delta$ we have*

$$|\Delta T_1^j p|_\infty \leqslant C\mu^j|\Delta p|_\infty \quad \forall j \in \mathbb{N}, \tag{2.8}$$

$$|T_1 p - T_2 p|_\infty \leqslant A|\Delta p|_\infty^\alpha. \tag{2.9}$$

If at least one of T_1 and T_2 is bounded and linear, then for any $\epsilon > 0$ there exist $\delta', C' > 0$ such that when $|\Delta p|_\infty < \delta'$,

$$|\Delta T_2^j p|_\infty \leqslant C'(\mu + \epsilon)^j|\Delta p|_\infty \quad \forall j \in \mathbb{N}.$$

Proof. See Appendix A.3 for the proof.

Remark 2.5. In Theorem 2.4, we make the assumption that one of T_1 and T_2 is bounded and linear to accommodate the assumption (2.8) on T_1, which is weaker than (2.1) in Theorem 2.1.

3. Closest point projection scheme for $\mathbb{S}^{m-1}$-valued data. Let $\overline{S}$ be a linear interpolatory subdivision operator defined by

$$(\overline{S}y)_{2k} = y_k, \quad (\overline{S}y)_{2k+1} = \sum_{i=1}^{n} w_i\, y_{k+i+\ell}, \tag{3.1}$$

where $\ell \in \mathbb{Z}$ is fixed. We also assume that $\overline{S}$ reproduces Π_{p-1}. By linearity and the well-posedness of Lagrange interpolation, it is necessary that $p \leqslant n$, and there is a unique mask with $p = n$. Since we are in the business of proving smoothness equivalence, we assume that $\overline{S}$ has at least some smoothness: $s_\infty(\overline{S}) > 0$. This, in turn, implies that (i) $p \geqslant 1$, (ii)

$$\sum_{i=1}^{n} w_i = 1, \tag{3.2}$$

and (iii) there exists a subdivision operator $\overline{S}^{[1]}$ such that $\overline{S}^{[1]} \circ \Delta = \Delta \circ \overline{S}$. The special cases of Deslauriers–Dubuc schemes [12, 9] correspond to $n = 2L$ (i.e., n is even), $\ell = -L$, and $p = n$.

In the z-transform domain, the polynomial reproduction property is equivalent to the existence of a polynomial $b(z)$ such that

$$\sum_{i=1}^{n} w_i' z^{2i} + z^{2n'+1} = (1+z)^p\, b(z), \tag{3.3}$$

where $w_i' = w_{n+1-i}$ and $n' = n + \ell$. By taking derivatives of both sides of (3.3) and evaluating at $z = -1$, we get the sum rules

$$\sum_{i=1}^{n} \binom{2i}{\gamma} w_i' = \binom{2n'+1}{\gamma}, \quad \gamma = 0, \ldots, p-1, \tag{3.4}$$

where $\binom{u}{v}$ is the standard binomial coefficient, i.e.,

$$\binom{u}{v} = \begin{cases} \frac{u!}{v!(u-v)!} & \text{if } u \geqslant v \geqslant 0, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\mathbb{S}^{m-1}$ be the unit sphere in $\mathbb{R}^m$, i.e., $\mathbb{S}^{m-1} = \{x \in \mathbb{R}^m : \|x\| = 1\}$. Let $P : \mathbb{R}^m \backslash \{0\} \to \mathbb{S}^{m-1}$ be the closest point projection operator onto the sphere, i.e., $P(x) = x/\|x\|$.

Consider the nonlinear subdivision operator S for $\mathbb{S}^{m-1}$-valued data defined by

$$(Sy)_{2k} = y_k, \quad (Sy)_{2k+1} = P\left(\sum_{i=1}^{n} w_i y_{k+i+\ell}\right). \tag{3.5}$$

Remark 3.1. The subdivision operation (3.5) is well defined as long as $y : \mathbb{Z} \to \mathbb{S}^{m-1}$ is such that $\sum_{i=1}^{n} w_i\, y_{k+i+\ell} \neq 0$ for all $k \in \mathbb{Z}$. We show in Lemma 3.3 that as long as $|\Delta y|_\infty$ is small enough, $S^j y$ is well defined for all j.

Let $x_1, \ldots, x_n \in \mathbb{S}^{m-1}$ with $\sum_{i=1}^{n} w_i\, x_i \neq 0$. For any $y \in \mathbb{R}^m \backslash \{0\}$,

$$\left\| y - P(y) \right\| = \left\| y - \frac{y}{\|y\|} \right\| = \left| 1 - \|y\| \right| \leqslant \left| 1 - \|y\|^2 \right|. \tag{3.6}$$

Combined with the facts that $x_i \in \mathbb{S}^{m-1}$ and $\sum_i w_i = 1$, we have

$$\left\| \sum_{i=1}^{n} w_i x_i - P\left(\sum_{i=1}^{n} w_i x_i\right) \right\| = \left| 1 - \left\| \sum_{i=1}^{n} w_i x_i \right\| \right| \leqslant \left| 1 - \left\| \sum_{i=1}^{n} w_i x_i \right\|^2 \right| \tag{3.7}$$

$$= \left| \sum_{i=1}^{n} w_i \langle x_i, x_i \rangle - \sum_{i=1}^{n} \sum_{j=1}^{n} w_i w_j \langle x_i, x_j \rangle \right| \tag{3.8}$$

$$= \left| \sum_{i=1}^{n} \sum_{j=1}^{n} c_{i,j}^0 \langle x_i, x_j \rangle \right|,$$

where the $n \times n$ matrix $C_0 = (c_{i,j}^0)$ is given by

$$c_{i,j}^0 := \begin{cases} w_i - w_i^2, & i = j, \\ -w_i w_j, & i \neq j. \end{cases} \tag{3.9}$$

We shall come back to this matrix after a few remarks.

Remark 3.2. Using the identity $\langle x_i, x_j\rangle = 1 - \|x_i - x_j\|^2/2$, we can also rewrite the upper bound (3.7) as

(3.10)
$$\left\|\sum_{i=1}^n w_i x_i - P\left(\sum_{i=1}^n w_i x_i\right)\right\| \leqslant \left|1 - \left\|\sum_{i=1}^n w_i x_i\right\|^2\right| = \left|\frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n w_i w_j \|x_i - x_j\|^2\right|.$$

LEMMA 3.3 (well-definedness of $S^j y$). *There exist $\delta^* > 0$ and $K > 0$ such that for any $y : \mathbb{Z} \to \mathbb{S}^{m-1}$ with $\|\Delta y\|_\infty \leqslant \delta^*$, $|\Delta S^j y|_\infty \leqslant K\delta^*$ and $S^j y$ is well defined for all $j = 1, 2, \ldots$.*

Proof. The proof is easy, and we just give the main idea: (3.10) says that as long as $|\Delta y|_\infty$ is small enough, $(Sy)_k$ stays away from the origin for every k. Since we assume $s_\infty(\overline{S}) > 0$, $|\Delta \overline{S}^j y|_\infty = O(2^{-j\nu})$ for any $0 < \nu < \min(1, s_\infty(\overline{S}))$, with (hidden constant) $\propto |\Delta y|_\infty$. Again by (3.10), S and $\overline{S}$ satisfy the proximity condition (2.9) in Theorem 2.4 (with $\alpha = 2$); hence by the theorem, $|\Delta S^j y|_\infty = O(2^{-j\nu^-})$ for any $\nu^- < \nu$, again with (hidden constant) $\propto |\Delta y|_\infty$. This means that if $|\Delta y|_\infty$ is small enough to begin with, then all $|\Delta S^j y|_\infty$ are small (in fact, decay that exponentially fast as j increases) and all $S^j y$ are well defined. ∎

Remark 3.4. We define convergence of $S^j y$ and the limit function $S^\infty y$ analogous to the linear case. Similar to (1.2), we define

$$s_\infty(S) := \inf_y \sup\{\alpha : S^\infty y \in \operatorname{Lip}\alpha\},$$

where the infimum is taken over all sequences y for which $S^\infty y$ is well defined. Since, for any such y, $|\Delta S^j y|_\infty$ decays and the smoothness of $S^\infty y$ does not depend on the behavior of $S^j y$ at coarse scales, there is no difference if we take the infimum over all y such that $|\Delta y|_\infty \leqslant \delta$, for any $0 < \delta \leqslant \delta^*$, where δ^* is given by Lemma 3.3.

3.1. Main proximity theorem. In this section, we show that the difference between Sy and $\overline{S}y$ can be bounded by the magnitudes of high order differences of y.

LEMMA 3.5. *Let $y_1, \ldots, y_n, z_1, \ldots, z_n \in \mathbb{R}^m$. Suppose*

$$y_i = \sum_{j=1}^n f_{i,j} z_j, \quad i = 1, \ldots, n.$$

Let $F = (f_{i,j})$ and $A_0 = (a^0_{i,j})$ be $n \times n$ real matrices. Then

$$\sum_{i=1}^n\sum_{k=1}^n a^0_{i,k}\langle y_i, y_k\rangle = \sum_{i=1}^n\sum_{k=1}^n a^1_{i,k}\langle z_i, z_k\rangle,$$

where matrix $A_1 = (a^1_{i,k})$ is given by

$$A_1 = F^T A_0 F.$$

The proof is straightforward and we omit it.

Define $d^0_i = x_i$, $i = 1, \ldots, n$, and for $k = 1, \ldots, n$,

$$d^k_i = \begin{cases} d^{k-1}_i, & i = 1, \ldots, k \\ d^{k-1}_i - d^{k-1}_{i-1}, & i = k+1, \ldots, n. \end{cases} \tag{3.11}$$

It follows from Lemma 3.5 and (3.11) that we can further rewrite (3.8) as follows:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}^{0}\langle x_i, x_j\rangle = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}^{k}\langle d_i^k, d_j^k\rangle, \quad k = 1,\dots,n-1, \tag{3.12}$$

where $C_k = (c_{i,j}^k)$ is given by

$$C_k = F_{k-1}^T C_{k-1} F_{k-1}, \quad k = 1,\dots,n-1,$$

and

$$F_{k-1} = \begin{pmatrix} I_{k-1} & & & \\ & 1 & & \\ & \vdots & \ddots & \\ & 1 & \cdots & 1 \end{pmatrix}. \quad (I_{k-1} \text{ is the } (k-1)\times(k-1) \text{ identity matrix.}) \tag{3.13}$$

The next result says that the matrix C_k has many vanishing entries when the interpolatory subdivision mask reproduces polynomials; in particular, if $n = p$, i.e., the interpolatory mask has the highest possible order of polynomial reproducibility, then C_{n-1} has the following form:

$$C_{n-1} = \begin{bmatrix} 0 & 0 & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & 0 & * \\ \vdots & \vdots & \ddots & * & \vdots \\ \vdots & 0 & * & \cdots & \vdots \\ 0 & * & \cdots & \cdots & * \end{bmatrix}.$$

LEMMA 3.6 (vanishing property of C_k). *If the interpolatory subdivision mask that defines the matrices C_k reproduces Π_{p-1}, then $c_{i,j}^k = 0$ if $i = 1$ or $j = 1$ or $i, j \leqslant k$, $i + j \leqslant p + 1$.*

Proof. See Appendix A.4 for the proof.

Lemma 3.6 leads to our main proximity result, as follows.

THEOREM 3.7. *Let $\overline{S}$ be a linear interpolatory subdivision scheme which reproduces Π_{p-1}. There are constants $B_2,\dots,B_p$ such that for any $y : \mathbb{Z} \to \mathbb{S}^{m-1}$ such that* (3.5) *is well defined, we have*

$$\text{if } p \geqslant 1, \quad |\overline{S}y - Sy|_\infty \leqslant B_2 |\Delta y|_\infty^2; \tag{3.14}$$

$$\text{if } p \geqslant 3, \quad |\overline{S}y - Sy|_\infty \leqslant B_k \sum_{i=1}^{k-1} |\Delta^i y|_\infty |\Delta^{k-i} y|_\infty, \quad k = 3,\dots,p. \tag{3.15}$$

Proof. Let $1 \leqslant k \leqslant p$. By Lemma 3.6, $c_{i,j}^k = 0$ when $i = 1$ or $j = 1$ or $i+j \leqslant k+1$. When $k = 1$, we have

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}^1\langle d_i^1, d_j^1\rangle\right| = \left|\sum_{i=2}^{n}\sum_{j=2}^{n} c_{i,j}^1\langle d_i^1, d_j^1\rangle\right| \leqslant \sum_{i=2}^{n}\sum_{j=2}^{n} |c_{i,j}^1| \left\|d_i^1\right\| \left\|d_j^1\right\| \leqslant B_2 |\Delta x|_\infty^2,$$

where B_2 is a constant that depends on $w_1,\dots,w_n$.

If $p \geqslant 3$, then for $3 \leqslant k \leqslant p$,

$$\left|\sum_{i=1}^{n}\sum_{j=1}^{n} c_{i,j}^{k}\,\langle d_i^k, d_j^k\rangle\right| = \left|\sum_{\substack{i+j>k+1\\ i>1,j>1}} c_{i,j}^{k}\,\langle d_i^k, d_j^k\rangle\right| \leqslant \sum_{\substack{i+j>k+1\\ i>1,j>1}} |c_{i,j}^{k}|\,\|d_i^k\|\,\|d_j^k\|$$

$$\leqslant B_k \sum_{i=1}^{k-1} |\Delta^i x|_\infty |\Delta^{k-i} x|_\infty,$$

where B_k is a constant that depends on $w_1, \ldots, w_n$.

Combined with (3.8) and (3.12), the above two estimates yield

$$\text{if } p \geqslant 1, \quad \left\|\sum_{i=1}^{n} w_i x_i - P\left(\sum_{i=1}^{n} w_i x_i\right)\right\|_\infty \leqslant B_2 |\Delta x|_\infty^2;$$

$$\text{if } p \geqslant 3, \quad \left\|\sum_{i=1}^{n} w_i x_i - P\left(\sum_{i=1}^{n} w_i x_i\right)\right\|_\infty \leqslant B_k \sum_{i=1}^{k-1} |\Delta^i x|_\infty |\Delta^{k-i} x|_\infty, \quad k = 3, \ldots, p\,.$$

The above can be applied to any n consecutive entries of an infinite sequence y, so we have proved the theorem. □

3.2. Smoothness equivalence. We now prove the main smoothness equivalence result, as follows.

THEOREM 3.8. *If $\overline{S}$ is a linear interpolatory subdivision scheme with $s_\infty(\overline{S}) > 0$, then*

$$s_\infty(S) \geqslant s_\infty(\overline{S}).$$

To prove this theorem, we need to first recall two results.

The first one is well known; see, e.g., [5, 10], which basically says that one can characterize the Hölder smoothness of a continuous function based on its samples at dyadic points. If $f : \mathbb{R} \to \mathbb{R}$ is bounded and continuous, then for any $\alpha > 0$,

$$\begin{aligned} f \in \text{Lip}\,\alpha \quad &\Longleftrightarrow \quad \exists\, r \in \mathbb{Z}_+,\, r > \alpha, \text{ s.t. } |(\Delta^r f_j)_k|_\infty = O(2^{-j\alpha}) \\ &\Longleftrightarrow \quad \forall\, r \in \mathbb{Z}_+,\, r > \alpha,\, |(\Delta^r f_j)_k|_\infty = O(2^{-j\alpha}), \end{aligned} \tag{3.16}$$

where $f_j := f|_{2^{-j}\mathbb{Z}}$, i.e., $(f_j)_k := f(2^{-j}k)$. These equivalences also imply that the critical Hölder regularity exponent of f can be determined from the exact asymptotic decay rate of $|(\Delta^r f_j)_k|_\infty$ for a large enough differencing order r, i.e.,

$$\sup\{\alpha : f \in \text{Lip}\,\alpha\} = \sup\left\{\alpha : |(\Delta^r f_j)_k|_\infty = O(2^{-j\alpha})\right\}. \tag{3.17}$$

The second result is the perturbation theorem [8, Theorem 3.3]. Originally derived to meet the needs of the analysis of a specific nonlinear subdivision algorithm, the theorem has been proved to be useful in the analysis of other nonlinear subdivision algorithms as well; see [33, 32, 23, 22, 36]. We restate this result in a form convenient to us, as follows.

THEOREM 3.9 (see [8, Theorem 3.3]). *Let $\overline{S}$ be a linear subdivision operator with $s_\infty(\overline{S}) > 0$. Let S be a (linear or nonlinear) subdivision operator that maps $\mathcal{D}(S) \subseteq \ell^\infty$ into itself. Let $y \in \mathcal{D}(S)$. If there exists $\nu > 0$ such that*

$$\left|(S - \overline{S}\,)S^j y\right|_\infty = O(2^{-j\nu}),$$

then S is convergent and $s_\infty(S, y) := \sup\{\alpha : S^\infty y \in \operatorname{Lip}\alpha\} \geqslant \min(\nu, s_\infty(\overline{S}))$.

Proof of Theorem 3.8. By Remark 3.4, it suffices to prove $s_\infty(S, y) := \sup\{\alpha : S^\infty y \in \operatorname{Lip}\alpha\} \geqslant s_\infty(\overline{S})$ for those y such that $|\Delta y|_\infty$ is small.

1° For any $0 < \gamma < \min(1, s_\infty(\overline{S}))$, there exists a constant $C > 0$ such that

$$|\Delta \overline{S}^j y|_\infty \leqslant C\, 2^{-\gamma j} |\Delta y|_\infty \quad \forall\, y \in \ell^\infty.$$

For any $\gamma' \in (0, \gamma)$, let $\epsilon := 2^{-\gamma'} - 2^{-\gamma} > 0$. Then it follows from (3.14) and Theorem 2.4 that there exist $\delta_{\gamma'} > 0$ and $C' > 0$ such that

(3.18)
$$|\Delta S^j y|_\infty \leqslant C'\,(2^{-\gamma} + \epsilon)^j |\Delta y|_\infty = C'\, 2^{-\gamma' j} |\Delta y|_\infty \quad \forall\, y : \mathbb{Z} \to \mathbb{S}^{m-1} \text{ s.t. } |\Delta y|_\infty < \delta_{\gamma'}.$$

2° It suffices to consider a fixed y with $|\Delta y|_\infty$ small enough so that (3.18) can be applied and all $S^j y$ are well defined. Recall Lemma 3.3 and Remark 3.4.

By (3.14) in Theorem 3.7, we have

$$|S(S^j y) - \overline{S}(S^j y)|_\infty \leqslant B_2 |\Delta S^j y|_\infty^2 \overset{(3.18)}{=} O(2^{-2\gamma' j}).$$

Then, by Theorem 3.9, we have $s_\infty(S, y) \geqslant \min(2\gamma', s_\infty(\overline{S}))$. Since γ' can be arbitrarily close to $\min(1, s_\infty(\overline{S}))$, we get

$$s_\infty(S, y) \geqslant \min(2, s_\infty(\overline{S})). \tag{3.19}$$

Thus the theorem is proved if $s_\infty(\overline{S}) \leqslant 2$. From now on we assume $s_\infty(\overline{S}) > 2$.

3° Let q be the unique integer such that

$$p \geqslant q + 1 \geqslant s_\infty(\overline{S}) > q \geqslant 2. \tag{3.20}$$

(Recall that $\overline{S}$ reproduces Π_{p-1} with $p \geqslant s_\infty(\overline{S})$ according to the theory of linear subdivision.)

We use induction to prove

$$s_\infty(S, y) \geqslant q. \tag{3.21}$$

Let $2 \leqslant k < q$. Assume that we have proved $s_\infty(S, y) \geqslant k$. By (3.16),

$$\begin{aligned} &|\Delta^\ell S^j y|_\infty = O(2^{-\ell j}), \quad \ell = 1, \dots, k-1, \\ &|\Delta^k S^j y|_\infty = O(2^{-(k-\epsilon)j}). \end{aligned} \tag{3.22}$$

Since $p > k + 1$, Theorem 3.7 applies, and we have

$$|\overline{S} S^j y - S^{j+1} y|_\infty \leqslant B_{k+1} \sum_{\ell=1}^{k} |\Delta^\ell S^j y|_\infty |\Delta^{k+1-\ell} S^j y|_\infty \overset{(3.22)}{=} O(2^{-(k+1-\epsilon)j}). \tag{3.23}$$

So by Theorem 3.9,

$$s_\infty(S, y) \geqslant \min(k + 1 - \epsilon, s_\infty(\overline{S})) \overset{(3.20)}{=} k + 1 - \epsilon$$

for any $\epsilon > 0$, which implies $s_\infty(S, y) \geqslant k + 1$. This completes the induction.

Since we have now proved (3.21), both (3.22) and (3.23) hold for $k = q$. Using Theorem 3.9 one more time gives

$$s_\infty(S, y) \geqslant \min(q + 1, s_\infty(\overline{S})) \overset{(3.20)}{=} s_\infty(\overline{S}).$$

This completes the proof. ∎

3.3. Near-closest projections onto the sphere. In this section, we show that for projections that are close to the closest point projection onto the sphere, the associated nonlinear schemes are also at least as smooth as the underlying interpolatory linear scheme. Such a result is to be expected: recall that the starting point of our proximity result is the simple inequality in (3.6); it is clear that we can relax P a little to achieve essentially the same upper bound on the right-hand side of (3.6). It is also clear that for any such projection operator, all the arguments for our smoothness equivalence result pertaining to the nonlinear subdivision operator $S = P \circ \overline{S}$ go through verbatim.[2]

In order for $S = P \circ \overline{S}$ to have a chance of being convergent, for any sequence y such that all $S^j y$ are well defined, $|\Delta S^j y|_\infty$ must converge to zero as $j \to \infty$. Also, by (3.10), as long as consecutive points in a sequence $z : \mathbb{Z} \to \mathbb{S}^{m-1}$ are sufficiently close to each other, the points in $\overline{S}z$ can be made as close to the sphere as we want. Therefore, we need to study only the property of P in a neighborhood of the sphere as far as the smoothness analysis of S is concerned.

LEMMA 3.10. *Let $P : \mathbb{R}^m \to \mathbb{S}^{m-1}$. If there exist $\delta, C > 0$ such that when $\big|\|x\| - 1\big| < \delta$,*

$$\cos(\angle(x, P(x))) \geqslant \frac{1 + \|x\|^2 - C(1 - \|x\|^2)^2}{2\|x\|}, \tag{3.24}$$

then $\|P(x) - x\| \leqslant \sqrt{C}\,\big|1 - \|x\|^2\big|$ when $\big|\|x\| - 1\big| < \delta$.

Proof. Since $\langle x, P(x)\rangle = \|x\|\,\|P(x)\|\cos(\angle(x, P(x))) = \|x\|\cos(\angle(x, P(x)))$, it follows from (3.24) that

$$\langle x, P(x)\rangle \geqslant \frac{1 + \|x\|^2 - C(1 - \|x\|^2)^2}{2}.$$

Hence

$$\begin{aligned}
\|P(x) - x\|^2 &= \langle P(x) - x, P(x) - x\rangle \\
&= 1 - 2\langle x, P(x)\rangle + \|x\|^2 \\
&\leqslant 1 - (1 + \|x\|^2 - C(1 - \|x\|^2)^2) + \|x\|^2 \\
&= C(1 - \|x\|^2)^2.
\end{aligned}$$

Thus $\|P(x) - x\| \leqslant \sqrt{C}\,\big|1 - \|x\|^2\big|$. ∎

When P is the closest point projection, we always have $\angle(x, P(x)) = 0$. The above lemma shows that as long as P satisfies (3.24), the same bound (3.8) applies

[2] We abuse notation and extend the definition of P to a map that maps sequences of m-vectors to sequences of points on the sphere; i.e. if y is a sequence of non-zero m-vectors, $P(y)$ is the sequence $P(y)_i = y_i/\|y_i\|$. We will abuse notation in a similar manner later without mention.

with an adjustment of the hidden constant. Consequently, we have the following result.

THEOREM 3.11. *For any projection operator P satisfying* (3.24) *and any interpolatory linear subdivision scheme $\overline{S}$, the corresponding nonlinear subdivision S_P satisfies $s_\infty(S_P) \geqslant s_\infty(\overline{S})$.*

4. Projection scheme for $SO(m)$-valued data and extensions. In this section, we first extend our smoothness equivalence result to the Lie group of special orthogonal matrices:

$$SO(m) = \{Y \in \mathbb{R}^{m\times m} : YY^T = I,\ \det(Y) = 1\}.$$

In order to use the projection approach for data taking values in $SO(m)$, we need to (i) embed $SO(m)$ into an Euclidean space and (ii) define a projection operator P from the Euclidean space to the embedded surface. From a practical point of view, we would also like such a P to be efficiently computable.

There is a natural way to embed $SO(m)$ in $\mathbb{R}^{m^2}$: simply treat every matrix in $SO(m)$ as a point in $\mathbb{R}^{m^2}$. It is not hard to prove that such a procedure indeed defines a smooth embedding; so $SO(m)$ now "looks like" a $m(m-1)/2$-dimensional curved surface in $\mathbb{R}^{m^2}$, and we shall decide how to project a given point outside of this surface onto the surface.

For $X_1, X_2 \in \mathbb{R}^{m\times m}$, define $\langle , \rangle$ by

$$\langle X_1, X_2\rangle := \operatorname{trace}(X_1 X_2^T), \tag{4.1}$$

where $\operatorname{trace}(X) = \sum_{i=1}^m x_{i,i}$ is the trace of a matrix $X = (x_{i,j})$. This inner product induces the so-called Frobenius norm:

$$\|X\|_F := \left(\sum_{i=1}^m \sum_{j=1}^m x_{i,j}^2\right)^{1/2}. \tag{4.2}$$

Recall that for any orthogonal matrices U, V,

$$\|UXV\|_F^2 = \|X\|_F^2. \tag{4.3}$$

If we identify $\mathbb{R}^{m\times m}$ with $\mathbb{R}^{m^2}$, then (4.1) and (4.2) are just the most standard inner product and Euclidean norm (respectively) in $\mathbb{R}^{m^2}$.

We can also extend the definitions of Δ and $|\cdot|_\infty$ to sequences with entries in $\mathbb{R}^{m\times m}$. For example, $|Y|_\infty := \sup_i \|Y_i\|_F$.

The closest point projection onto $SO(m)$ of a matrix with positive determinant can be computed efficiently using its singular value decomposition (SVD); a now-classical reference for (especially the computational aspect of) SVD is [15].

PROPOSITION 4.1. *Let $A \in \mathbb{R}^{m\times m}$ with $\det(A) > 0$. Then*

$$P(A) := \operatorname*{argmin}_{X\in SO(m)} \|A - X\|_F = UV^T = (AA^T)^{-1/2}A, \tag{4.4}$$

where $A = U\Sigma V^T$ is an SVD of A.

Proof. Let $A = U\Sigma V^T$ be an SVD of A, where U, V are orthogonal matrices and D is a diagonal matrix with positive diagonal entries $\sigma_1 \geqslant \cdots \geqslant \sigma_m > 0$. Then

$$\begin{aligned}\operatorname*{argmin}_{X\in SO(m)} \|A - X\|_F &= \operatorname*{argmin}_{X\in SO(m)} \|U\Sigma V^T - X\|_F \\ &= U\left(\operatorname*{argmin}_{X\in SO(m)} \|\Sigma - U^T X V\|_F\right) V^T \\ &= U\left(\operatorname*{argmin}_{R\in SO(m)} \|\Sigma - R\|_F\right) V^T.\end{aligned}$$

It is easy to show that $\operatorname{argmin}_{R\in SO(m)} \|\Sigma - R\|_F = I$:

$$\begin{aligned}\|\Sigma - R\|_F^2 &= \sum_{i=1}^m \left[(\sigma_i - R_{ii})^2 + \sum_{j\neq i} R_{ij}^2\right] = \sum_{i=1}^m \left[(\sigma_i - R_{ii})^2 + (1 - R_{ii}^2)\right] \\ &= \sum_{i=1}^m (\sigma_i^2 + 1 - 2\sigma_i R_{ii}).\end{aligned}$$

Since $R_{ii} \leqslant 1$, the right-hand side is minimized when $R_{ii} = 1$ which, since $R \in SO(m)$, also implies $R = I$. Notice also that $(AA^T)^{-1/2}A = (U\Sigma^2 U^T)^{-1/2}U\Sigma V^T = (U\Sigma^{-1}U^T)U\Sigma V^T = UV^T$. ∎

Remark 4.2. Proposition 4.1 is essentially a result published in [2, 28] and seems to be known to others as well. We present our proof anyway not only because it is short and elementary but also because we want to address a subtle point. First of all, we note that the projection operator defined by (4.4) has the invariance property: $P(R_1 A R_2) = R_1 P(A) R_2$ for any $R_1, R_2 \in SO(m)$. It implies that our resulting subdivision algorithm has the desirable property that it does not depend on the artificial choice of orthogonal coordinate system for representing m-dimensional rotations. Our presentation above is more elementary than that in [2] because we are unconcerned with invariance at the beginning, simply think of $SO(m)$ as a regular surface in $\mathbb{R}^{m^2}$, and use the plain Euclidean metric in $\mathbb{R}^{m^2}$. The approach in [2], instead, considers $SO(m)$ as a subgroup $GL(m)$ and uses a (left-)invariant Riemannian metric of $GL(m)$ in defining "closest." While the two different points of view yield the same projector $P(A) = (AA^T)^{-1/2}A$, the coincidence is due to (4.3) and is specific to $SO(m)$. We will revisit this issue in section 4.2.

Let $X = (X_1, \ldots, X_n)$ with $X_i \in SO(m)$ and $w_1, \ldots, w_n$ be as in (3.1). Then it follows from $\sum_{i=1}^n w_i = 1$ that

$$\left\|\sum_{i=1}^n w_i X_i - X_1\right\|_F = \left\|\sum_{i=1}^n w_i (X_i - X_1)\right\|_F \leqslant \left(\sum_{j=1}^n |w_j|\right) \max_i \|X_i - X_1\|_F. \tag{4.5}$$

Since $\det(X)$ is a continuous function of $X \in \mathbb{R}^{m\times m}$ and $\det(X_1) = 1 > 0$, it follows from (4.5) that there exists $\delta > 0$ such that when $\max_i \|X_i - X_1\|_F < \delta$,

$$\det\left(\sum_{i=1}^n w_i X_i\right) > 0,$$

and consequently Proposition 4.1 can be applied to define $P(\sum_{i=1}^n w_i X_i)$.

Let $\sum_{i=1}^n w_i X_i = U\Sigma V^T$ be the SVD of $\sum_{i=1}^n w_i X_i$, where U, V are orthogonal matrices and Σ is a diagonal matrix with positive diagonal entries $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_m > 0$. Then it follows from Proposition 4.1 that

(4.6)
$$\left\|\sum_{i=1}^n w_i X_i - P\left(\sum_{i=1}^n w_i X_i\right)\right\|_F = \|U\Sigma V^T - UV^T\|_F = \|\Sigma - I\|_F = \left(\sum_{\ell=1}^m (\sigma_\ell - 1)^2\right)^{1/2}$$
$$\leqslant \sqrt{m}\max_\ell |\sigma_\ell - 1| \leqslant \sqrt{m}\max_\ell |\sigma_\ell^2 - 1|.$$

Since $(\sum_{i=1}^n w_i X_i)V = U\Sigma$, it follows that $(\sum_{i=1}^n w_i X_i)v_\ell = \sigma_\ell u_\ell$, where $u_\ell, v_\ell \in \mathbb{S}^{m-1}$ are the columns of U, V respectively. So

$$\sigma_\ell^2 = \|\sigma_\ell u_\ell\|^2 = \left\|\left(\sum_{i=1}^n w_i X_i\right) v_\ell\right\|^2 = \left\langle \left(\sum_{i=1}^n w_i X_i\right) v_\ell, \left(\sum_{i=1}^n w_i X_i\right) v_\ell \right\rangle$$
$$= \sum_{i=1}^n \sum_{j=1}^n w_i w_j v_\ell^T X_j^T X_i v_\ell.$$

Note that $\|(X_i - X_j)v_\ell\|^2 = \langle X_i v_\ell - X_j v_\ell, X_i v_\ell - X_j v_\ell\rangle = 2 - 2v_\ell^T X_j^T X_i v_\ell$; hence

$$\sigma_\ell^2 = \sum_{i=1}^n \sum_{j=1}^n w_i w_j \left(1 - \frac{1}{2}\|(X_i - X_j)v_\ell\|^2\right)$$
$$= 1 - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n w_i w_j \|(X_i - X_j)v_\ell\|^2.$$

Thus

$$|\sigma_\ell^2 - 1| = \left|\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n w_i w_j \|(X_i - X_j)v_\ell\|^2\right|.$$

Now fix ℓ and let $x_i := X_i v_\ell$, $i = 1, \ldots, n$. Then

$$|\sigma_\ell^2 - 1| = \left|\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n w_i w_j \|x_i - x_j\|^2\right|. \tag{4.7}$$

Note that the right-hand side of (4.7) looks exactly the same as the right-hand side of (3.10). Following exactly the same arguments there, we can show the following: If $p \geqslant 1$,

$$\left|\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n w_i w_j \|x_i - x_j\|^2\right| \leqslant \sum_{i=2}^n \sum_{j=2}^n |c_{i,j}^1| \left\|d_i^1\right\| \left\|d_j^1\right\|. \tag{4.8}$$

If $p \geqslant 3$, then for $3 \leqslant k \leqslant p$,

$$\left|\frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n w_i w_j \|x_i - x_j\|^2\right| \leqslant \sum_{\substack{i+j>k+1 \\ i>1, j>1}} |c_{i,j}^k| \left\|d_i^k\right\| \left\|d_j^k\right\|, \tag{4.9}$$

where $c_{i,j}^k$ and d_i^k are as defined in section 3.

Parallel to the definition of d_i^k, we define $D_i^0 = D_i$, $i = 1, \ldots, n$, and for $k = 1, \ldots, n$,

$$D_i^k = \begin{cases} D_i^{k-1}, & i = 1, \ldots, k, \\ D_i^{k-1} - D_{i-1}^{k-1}, & i = k+1, \ldots, n. \end{cases}$$

Then for any k and i,

$$d_i^k = D_i^k v_\ell.$$

So

$$\|d_i^k\| = \|D_i^k v_\ell\| \leqslant \|D_i^k\|_2 \|v_\ell\| = \|D_i^k\|_2 \leqslant \|D_i^k\|_F, \tag{4.10}$$

where $\|.\|_2$ denotes the 2-norm of a matrix and we used the fact that $\|Y\|_2 \leqslant \|Y\|_F$ for any matrix Y.

Combining (4.10) with (4.7), (4.8), and (4.9), we get

$$(p \geqslant 1) \qquad |\sigma_\ell^2 - 1| \leqslant \sum_{i=2}^{n} \sum_{j=2}^{n} |c_{i,j}^1| \left\|D_i^1\right\|_F \left\|D_j^1\right\|_F \leqslant B_2 |\Delta X|_\infty^2$$

and for $3 \leqslant k \leqslant p$,

$$(p \geqslant 3) \qquad |\sigma_\ell^2 - 1| \leqslant \sum_{\substack{i+j>k+1 \\ i>1, j>1}} |c_{i,j}^k| \|D_i^k\|_F \|D_j^k\|_F \leqslant B_k \sum_{i=1}^{k-1} |\Delta^i X|_\infty |\Delta^{k-i} X|_\infty,$$

where $B_2, B_3, \ldots, B_p$ are constants that depend only on $w_1, \ldots, w_n$. Combining this with (4.6), we get

$$\left\| \sum_{i=1}^{n} w_i X_i - P\left(\sum_{i=1}^{n} w_i X_i \right) \right\|_F \tag{4.11}$$

$$\leqslant \begin{cases} \sqrt{m}\, B_2 |\Delta X|_\infty^2 & \text{if } p \geqslant 1 \\ \sqrt{m}\, B_k \sum_{i=1}^{k-1} |\Delta^i X|_\infty |\Delta^{k-i} X|_\infty, & k = 3, \ldots, p, \quad \text{if } p \geqslant 3. \end{cases} \tag{4.12}$$

This is essentially the same as (3.14) and (3.15) in Theorem 3.7, on which the proof of Theorem 3.8 is based; this also means that we have proved the following claim.

THEOREM 4.3. *For any interpolating linear subdivision $\overline{S}$, the corresponding closest point projection scheme S for $SO(m)$-valued data satisfies $s_\infty(S) \geqslant s_\infty(\overline{S})$.*

4.1. Extensions to related Lie groups. We consider rigid body displacements:

$$SE(m) = \{T_{A,b} : \mathbb{R}^m \to \mathbb{R}^m \mid T_{A,b}(x) = Ax + b,\ A \in SO(m),\ b \in \mathbb{R}^m\}.$$

There is a standard way to smoothly embed this matrix Lie group into $\mathbb{R}^{(m+1)\times(m+1)}$:

$$T_{A,b} \mapsto \begin{bmatrix} A & b \\ 0 & 1 \end{bmatrix}.$$

This embedding is also a group homomorphism from $SE(m)$ to the general linear group $GL(m+1)$, as $T_{A',b'} \circ T_{A,b} = T_{A'A,A'b+b}$ and

$$\begin{bmatrix} A' & b' \\ 0 & 1 \end{bmatrix} \begin{bmatrix} A & b \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} A'A & A'b+b' \\ 0 & 1 \end{bmatrix}.$$

So, once again, we are in the situation as discussed in Remark 4.2: for the purpose of constructing a subdivision scheme for $SE(m)$-valued data based on a *linear* subdivision scheme $\overline{S}$, we take the point of view that $SE(m)$ is a regular surface in the *linear space* $\mathbb{R}^{(m+1)\times(m+1)}$; however, for the purpose of constructing a projection operator P with a sensible invariance property, we should take the point of view that $SE(m)$ is embedded in the Lie group $GL(m)$ and define a projection operator based on a (left-) invariant metric with respect to the group operation. In this case, the projection operator is given by

$$P\left(\begin{bmatrix} A & b \\ 0 & 1 \end{bmatrix}\right) = \begin{bmatrix} UV^T & b \\ 0 & 1 \end{bmatrix}, \tag{4.13}$$

where $A = U\Sigma V^T$ is an SVD of A. However, again as in the case of $SO(m)$, it does not quite matter whether we think of "closest point projection" in terms of the standard Euclidean metric in $\mathbb{R}^{m^2}$ or an invariant metric in $GL(m+1)$.

After all, the most important fact is that the nonlinear subdivision operator $S = P \circ \overline{S}$ with P given by (4.13) and $\overline{S}$ a linear interpolatory subdivision scheme acting componentwise is well defined when applied to any sequence $Y : \mathbb{Z} \to SE(m)$ with a small enough $|\Delta Y|_\infty$; moreover, S has the desirable property that it is invariant under any change of orthogonal reference frame for representing rigid motions in an m-dimensional space.

Motivated by motion design, we are also interested in direct products of $SO(m)$; e.g., one can model the combined motion of 17 major human joints as an element in

$$\underbrace{SO(3)}_{\text{neck}} \times \underbrace{SO(3)\times SO(3)}_{\text{shoulders}} \times \underbrace{SO(2)\times SO(2)}_{\text{elbows}} \times \underbrace{SO(3)\times SO(3)}_{\text{wrists}}$$
$$\times \underbrace{SO(3)\times SO(3)}_{\text{hips}} \times \underbrace{SO(2)\times SO(2)}_{\text{knees}} \times \underbrace{SO(3)\times SO(3)}_{\text{ankles}}$$
$$\times \underbrace{SO(3)\times SO(3)\times SO(3)\times SO(3)}_{\text{spine}}.$$

See [19, Figure 3.2] for a graphical illustration.

It is obvious that we can extend the projection approach to subdivide data taking values on such a direct product. It is also obvious that we can extend Theorem 4.3 to $SE(m)$ and such direct products. For the record, let us state it formally, as follows.

THEOREM 4.4. *Let $\mathcal{M} = SE(m)$ or $\prod_{i=1}^{k} SO(m_i) \times \prod_{j=1}^{k'} SE(n_j)$. For any interpolating linear subdivision $\overline{S}$, the corresponding closest point projection scheme $S = P \circ \overline{S}$ for $\mathcal{M}$-valued data satisfies $s_\infty(S) \geqslant s_\infty(\overline{S})$.*

4.2. $SL(m)$. We recall once again the dual view first discussed in Remark 4.2. This time we consider the matrix Lie group of all measure- and orientation-preserving linear transformations:

$$SL(m) := \{Y \in \mathbb{R}^{m\times m} : \det(Y) = 1\}.$$

$SL(m)$ has a natural embedding as a regular hypersurface in $\mathbb{R}^{m^2}$, but is also a subgroup of $GL(m)$. Unlike the cases of $SO(m)$ and $SE(m)$, the two points of view give different projectors. Using the latter setup which offers invariance, the resulting projector P is given by (see [13])

$$A \mapsto A/\det A^{1/m}. \tag{4.14}$$

Note that $P(UAV) = UP(A)V$ for all $U, V \in SL(m)$. On the other hand, solving $\min_{X \in SL(m)} \|A - X\|_F$ is in principle a straightforward application of the method of Lagrange multipliers but gives a very complicated projector—which also lacks invariance—even in dimension $m = 2$.

Our computational experiment (akin to those "smoothness equivalence experiments" found in [33, 23, 24]) clearly indicates that the nonlinear subdivision scheme $S = P \circ \overline{S}$ enjoys the same smoothness equivalence property as in the case of Theorems 3.8, 4.3, 4.4. A proof is yet to be found.

5. Conclusions and discussions. Interpolation of manifold-valued data is a fundamental problem that has applications in many fields. The linear subdivision method is an efficient and very well-studied method for interpolating or approximating real-valued data in a multiresolution fashion. We described in section 1 three sets of approaches for adapting a linear subdivision scheme to subdivide manifold-valued data. The mathematical analysis of such nonlinear subdivision schemes is at its infancy. We mentioned a number of articles which offer some low degree smoothness equivalence results for certain nonlinear subdivision schemes. To the best of our knowledge, this is the first article that attacks the arbitrary degree smoothness equivalence conjectures.

We suspect that Theorems 3.8, 4.3, and 4.4 can be extended to any C^∞ k-dimensional regular surface in $\mathbb{R}^n$ with any near-closest projection operator.

We mention a recent smoothness *non*-equivalence result in the nonlinear subdivision literature: in [37], a seemingly nonadaptive nonlinear subdivision scheme is shown to have a fairly strong data-dependent property, *unlike* any linear subdivision scheme or weakly nonlinear subdivision schemes such as the ones studied in this article. Specifically, it is shown in [37] that a nonlinear convexity-preserving subdivision scheme produces limit curves with critical Hölder regularity depending on the initial data, and the regularity can be anywhere between 1 and 2.

We discuss potential applications of our results in the following two seemingly unrelated problems:

- *Conics-reproducing subdivision scheme.* A standard complaint of standard B-splines and standard linear subdivision schemes is that they can reproduce only polynomials and not conic sections. The industrial standard NURBS uses rational B-splines because rational polynomials can reproduce conics while polynomials cannot. But it is widely argued that NURBS methods lack some of the key advantages of subdivision methods.[3] A linear but *non-stationary* 4-point interpolatory scheme is derived in [14, 26], which reproduces $\mathrm{span}(1, x, \cos(x), \sin(x))$ instead of the usual $\Pi_3 = \mathrm{span}(1, x, x^2, x^3)$, and hence can reproduce circles when the initial control polygon is sampled uniformly from a circle. Given the result in this paper, it seems as though a better way to solve this problem is to use the projection approach. When one demands that a subdivision scheme exactly reproduce a circle, or any

[3]The debate, however, is mostly on surface modeling, not curve modeling.

conic section, or any other prespecified shape $\mathcal{C}$, our proposed method is to use a nonlinear but stationary scheme of the form $S = P_{\mathcal{C}} \circ \overline{S}$, as opposed to the linear but nonstationary scheme proposed in [14, 26]. The projection approach seems more general and flexible: it does not require uniform sampling, it can be used in conjunction with any interpolatory subdivision scheme (not just 4-point), and, for the circle at least, Theorem 3.8 says that the nonlinear scheme is as smooth as the underlying linear scheme. (The exact Hölder regularity of the specific non-stationary 4-point scheme in [14, 26] is not known, but the scheme is shown to be at least C^1.)

- *Normal multiresolution of curves.* Underlying the method of normal multiresolution of curves in [8] is a nonlinear subdivision scheme of almost exactly the same form as those studied in this paper, i.e., $S = P \circ \overline{S}$. The key difference is that the P's in this article are such that $P(y)_i$ is dependent only on y_i, whereas the P in [8] is more data-adaptive: $P(y)_{2i} = y_{2i}$ and $P(y)_{2i+1} =$ an intersection point of $\mathcal{C}$ with the line passing through y_{2i+1} and normal to the line $\overline{y_{2i}\, y_{2i+2}}$. See [8, Figure 2] for a graphical illustration. (Here $\mathcal{C}$ is a planar curve subject to a normal multiresolution analysis.) It is conjectured that the parametrization induced by a normal multisolution has exactly the same smoothness as that of the underlying interpolatory subdivision scheme. Similar to the other low degree smoothness equivalence results in [29, 30, 36], Daubechies et al. prove the smoothness equivalence only when (in the notation of this paper) $s_\infty(\overline{S}) \leqslant 2$. It seems possible to adapt the ideas and results in this paper to solve the full smoothness equivalence conjecture pertaining to normal multiresolution.

Yet another extension is to consider Hermite subdivision schemes on manifolds. See, e.g., [20] for the interests of Hermite interpolation in Lie groups arising from geometric integration of ODEs. Hermite subdivision schemes in the linear setting are quite well studied; see [21, 34, 16] and the references therein. It is not hard to construct Hermite subdivision schemes on Lie groups, but the analysis of such schemes is likely to be difficult.

Finally, it is needless to say that similar smoothness equivalence results for the more intrinsic linearization methods described in section 1 are waiting to be developed.

Appendix.

A.1. Proof of Lemma 2.2. Since $|\Delta x - \Delta y|_\infty \leqslant 2|x-y|_\infty$ for any $x, y \in \ell^\infty$, it follows that

$$|\Delta T_2 p - \Delta T_1 p|_\infty \leqslant 2|T_2 p - T_1 p|_\infty.$$

Combining this with (2.3), we get

$$|\Delta T_2 p - \Delta T_1 p|_\infty \leqslant 2A|\Delta p|_\infty^\alpha \quad \forall\, p \in \ell^\infty,\ |\Delta p|_\infty < \delta.$$

Thus

$$|\Delta T_2 p|_\infty \leqslant |\Delta T_1 p|_\infty + 2A|\Delta p|_\infty^\alpha \quad \forall\, p \in \ell^\infty,\ |\Delta p|_\infty < \delta.$$

Since we have (2.2), it follows that for all $p \in \ell^\infty$ with $|\Delta p|_\infty < \min(\delta, 1)$, we have

$$|\Delta T_2 p|_\infty \leqslant C|\Delta p|_\infty + 2A|\Delta p|_\infty^\alpha = (C + 2A|\Delta p|_\infty^{\alpha-1})|\Delta p|_\infty < (C+2A)|\Delta p|_\infty. \tag{A.1}$$

Therefore (2.4) holds for $C' = C + 2A$ and $\delta' = \min(\delta, 1)$. ∎

A.2. Proof of Lemma 2.3. We use induction. For $j = 1$, (2.7) follows immediately from (2.6) by choosing $\delta_1 = \delta$ and $C_1 = A$. Now suppose that (2.7) holds for some $j \geqslant 1$.

It follows from Lemma 2.2 that there exist $C' > 1$ and $\delta' > 0$ such that when $|\Delta p|_\infty < \delta'$,

$$|\Delta T_2 p|_\infty \leqslant C' |\Delta p|_\infty.$$

Thus for any $j \in \mathbb{N}$,

$$|\Delta T_2^j p|_\infty \leqslant C'^j |\Delta p|_\infty < \delta \quad \text{if } |\Delta p|_\infty < \min(\delta, \delta') C'^{-j}.$$

Since T_1 is bounded and linear, it follows from (2.6) that for $q \in \ell^\infty$ satisfying $|\Delta q|_\infty < \delta$ we have

$$|T_1 p - T_2 q|_\infty \leqslant |T_1 p - T_1 q|_\infty + |T_1 q - T_2 q|_\infty \leqslant |T_1|_\infty |p - q|_\infty + A|\Delta q|^\alpha.$$

Hence if $|\Delta p|_\infty < \min(\min(\delta, \delta') C'^{-j}, \delta_j)$, then

$$\begin{aligned} |T_1^{j+1} p - T_2^{j+1} p|_\infty &\leqslant |T_1|_\infty |T_1^j p - T_2^j p|_\infty + A|\Delta T_2^j p|_\infty^\alpha \\ &\leqslant |T_1|_\infty |T_1^j p - T_2^j p|_\infty + AC'^{j\alpha} |\Delta p|_\infty^\alpha \\ &\leqslant (|T_1|_\infty C_j + AC'^{j\alpha}) |\Delta p|_\infty^\alpha. \end{aligned}$$

This means that (2.7) holds for $j + 1$. By induction, (2.7) holds for any $j \in \mathbb{N}$. ∎

A.3. Proof of Theorem 2.4. Theorem 2.1 already covers the case of $C \leqslant 1$, so we can assume $C > 1$.

For any $\epsilon \in (0, 1 - \mu)$, we can find $N \in \mathbb{N}$ such that $C^{1/N} < 1 + \epsilon/\mu$. So $C\mu^N < (\mu + \epsilon)^N < 1$. Hence when $|\Delta p|_\infty < \delta$,

$$|\Delta T_1^{jN} p|_\infty \leqslant C\mu^{jN} |\Delta p|_\infty \leqslant (C\mu^N)^j |\Delta p|_\infty \quad \forall j \in \mathbb{N}.$$

It follows from (2.8), (2.9), and Lemma 2.2 that there exist $\tilde{C}, \tilde{\delta} > 0$ such that when $|\Delta p|_\infty < \tilde{\delta}$,

$$|\Delta T_2 p|_\infty \leqslant \tilde{C} |\Delta p|_\infty. \tag{A.2}$$

Since one of T_1 and T_2 is bounded and linear, it follows from (2.8), (A.2), (2.9), and Lemma 2.3 that there exist $C_N > 0$ and $\delta_N > 0$ such that when $|\Delta p|_\infty < \delta_N$,

$$|T_1^N p - T_2^N p|_\infty \leqslant C_N |\Delta p|_\infty^\alpha.$$

Now we can apply Theorem 2.1 to operators T_1^N and T_2^N. We have that for any $0 < \epsilon_0 < (\mu + \epsilon)^N - C\mu^N$ there exists $0 < \delta_0 < \delta$ such that

$$|\Delta T_2^N p|_\infty \leqslant (C\mu^N + \epsilon_0) |\Delta p|_\infty \leqslant (\mu + \epsilon)^N |\Delta p|_\infty \quad \text{if } |\Delta p|_\infty < \delta_0.$$

Therefore for $k = 0, 1, \ldots,$ and $r = 0, 1, \ldots, N - 1$

$$|\Delta T_2^{kN+r} p|_\infty \leqslant (\mu + \epsilon)^{kN} |\Delta T_2^r p|_\infty \quad \text{if } |\Delta T_2^r p|_\infty < \delta_0. \tag{A.3}$$

It follows from (A.2) that

$$|\Delta T_2^r p|_\infty \leqslant \tilde{C}^r |\Delta p|_\infty < \delta_0 \quad \text{if } |\Delta p|_\infty < \min(\delta_0, \tilde{\delta}) \tilde{C}^{-r}. \tag{A.4}$$

Combining (A.3) and (A.4), we have

$$|\Delta T_2^{kN+r}p|_\infty \leqslant (\mu+\epsilon)^{kN}\tilde{C}^r|\Delta p|_\infty \leqslant (\mu+\epsilon)^{kN+r}(\mu+\epsilon)^{-N}\tilde{C}^N|\Delta p|_\infty$$

if $|\Delta p| < \delta' := \min(\delta_0, \tilde{\delta})\tilde{C}^{-r}$. Let $C' = (\mu+\epsilon)^{-N}\tilde{C}^N$. Then when $|\Delta p|_\infty < \delta'$,

$$|\Delta T_2^j p|_\infty \leqslant C'(\mu+\epsilon)^j|\Delta p|_\infty$$

for any $j \in \mathbb{N}$ and $|\Delta p|_\infty < \delta'$. ∎

A.4. Proof of Lemma 3.6. By (3.9), $c^0_{i,j} = c^0_{j,i}$ and for $j = 1, \dots, n$,

$$\sum_{i=1}^n c^0_{i,j} = c^0_{j,j} + \sum_{\substack{i=1\\ i\neq j}}^n c^0_{i,j} = w_j - w_j^2 - \sum_{\substack{i=1\\ i\neq j}}^n w_i w_j = w_j - w_j^2 - w_j(1-w_j) = 0. \tag{A.5}$$

Thus the first row and first column of C_0 both sum to zero.

LEMMA A.1. *Let $A_0 = (a^0_{i,j})$ be an $n \times n$ real matrix. For $k = 0, \dots, n-2$, define*

$$A_{k+1} = F_k^T A_k F_k,$$

where F_k is defined by (3.13). *Then for $k = 0, \dots, n-2$, $A_{k+1} = (a^{k+1}_{\alpha,\beta})$ is given by*

$$a^{k+1}_{\alpha,\beta} = \begin{cases} \sum_{i=1}^n \sum_{j=1}^n \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} a^0_{i,j}, & \alpha,\beta \leqslant k+1, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha+k}{k}\binom{j-1}{\beta-1} a^0_{i,j}, & \alpha > k+1, \beta \leqslant k+1, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-1}{\alpha-1}\binom{j-\beta+k}{k} a^0_{i,j}, & \alpha \leqslant k+1, \beta > k+1, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha+k}{k}\binom{j-\beta+k}{k} a^0_{i,j}, & \alpha,\beta > k+1. \end{cases} \tag{A.6}$$

Proof. We prove (A.6) by induction. It follows from $A_1 = F_0^T A_0 F_0$ that

$$a^1_{\alpha,\beta} = \sum_{i=\alpha}^n \sum_{j=\beta}^n a^0_{i,j} = \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha}{0}\binom{j-\beta}{0} a^0_{i,j}.$$

This shows that (A.6) is true for $k = 0$.

Suppose that (A.6) is true for $k = q-1$; i.e.,

$$a^q_{\alpha,\beta} = \begin{cases} \sum_{i=1}^n \sum_{j=1}^n \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} a^0_{i,j}, & \alpha,\beta \leqslant q, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha+q-1}{q-1}\binom{j-1}{\beta-1} a^0_{i,j}, & \alpha > q, \beta \leqslant q, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-1}{\alpha-1}\binom{j-\beta+q-1}{q-1} a^0_{i,j}, & \alpha \leqslant q, \beta > q, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha+q-1}{q-1}\binom{j-\beta+q-1}{q-1} a^0_{i,j}, & \alpha,\beta > q. \end{cases} \tag{A.7}$$

It follows from $A_{q+1} = F_q^T A_q F_q$ that

$$a_{\alpha,\beta}^{q+1} = \begin{cases} a_{\alpha,\beta}^q, & \alpha,\beta \leqslant q, \\ \sum_{s=\alpha}^{n} a_{s,\beta}^q, & \alpha > q, \beta \leqslant q, \\ \sum_{t=\beta}^{n} a_{\alpha,t}^q, & \alpha \leqslant q, \beta > q, \\ \sum_{s=\alpha}^{n}\sum_{t=\beta}^{n} a_{s,t}^q, & \alpha,\beta > q. \end{cases} \tag{A.8}$$

Substituting (A.7) into (A.8), we get

$$a_{\alpha,\beta}^{q+1} = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} a_{i,j}^0, & \alpha,\beta \leqslant q, \\ \sum_{s=\alpha}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-s+q-1}{q-1}\binom{j-1}{\beta-1} a_{i,j}^0, & \alpha > q, \beta \leqslant q, \\ \sum_{t=\beta}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-1}{\alpha-1}\binom{j-t+q-1}{q-1} a_{i,j}^0, & \alpha \leqslant q, \beta > q, \\ \sum_{s=\alpha}^{n}\sum_{t=\beta}^{n}\sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-s+q-1}{q-1}\binom{j-t+q-1}{q-1} a_{i,j}^0, & \alpha,\beta > q. \end{cases}$$

Using the following identities on combinatorial numbers,

$$\sum_{s=\alpha}^{n} \binom{i-s+q-1}{q-1} = \binom{i-\alpha+q}{q}, \quad \sum_{t=\beta}^{n} \binom{j-t+q-1}{q-1} = \binom{j-\beta+q}{q},$$

we get

$$a_{\alpha,\beta}^{q+1} = \begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} a_{i,j}^0, & \alpha,\beta \leqslant q, \\ \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-\alpha+q}{q}\binom{j-1}{\beta-1} a_{i,j}^0, & \alpha > q, \beta \leqslant q, \\ \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-1}{\alpha-1}\binom{j-\beta+q}{q} a_{i,j}^0, & \alpha \leqslant q, \beta > q, \\ \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-\alpha+q}{q}\binom{j-\beta+q}{q} a_{i,j}^0, & \alpha,\beta > q. \end{cases} \tag{A.9}$$

It can be easily verified that (A.9) agrees with (A.6) when $k = q$. This concludes the proof. □

It follows from Lemma A.1 that for $k = 1, \dots, n-1$,

$$c^k_{\alpha,\beta} = \begin{cases} \sum_{i=1}^n \sum_{j=1}^n \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} c^0_{i,j}, & \alpha, \beta \leqslant k, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha+k-1}{k-1}\binom{j-1}{\beta-1} c^0_{i,j}, & \alpha > k, \beta \leqslant k, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-1}{\alpha-1}\binom{j-\beta+k-1}{k-1} c^0_{i,j}, & \alpha \leqslant k, \beta > k, \\ \sum_{i=1}^n \sum_{j=1}^n \binom{i-\alpha+k-1}{k-1}\binom{j-\beta+k-1}{k-1} c^0_{i,j}, & \alpha, \beta > k. \end{cases} \tag{A.10}$$

Combining this with (A.5), we get for $k = 1, \dots, n-1$ and $\ell = 1, \dots, n$ that

$$c^k_{\ell,1} = c^k_{1,\ell} = \left\{ \begin{array}{ll} \sum_{i=1}^n \sum_{j=1}^n \binom{j-1}{\ell-1} c^0_{i,j}, & \ell \leqslant k \\ \sum_{i=1}^n \sum_{j=1}^n \binom{j-\ell+k-1}{k-1} c^0_{i,j}, & \ell > k \end{array} \right\} = 0. \tag{A.11}$$

For any $z \in \mathbb{R}$ and $\ell \in \mathbb{N} \cup \{0\}$, the generalized binomial coefficient $\binom{z}{\ell}$ is defined as

$$\binom{z}{\ell} = \frac{z(z-1)\cdots(z-\ell+1)}{\ell!}.$$

For each fixed ℓ, let $q_\ell(z) = \binom{z}{\ell}$. Then $q_\ell(z)$ is the unique polynomial in z of degree ℓ satisfying

$$q_\ell(0) = q_\ell(1) = \cdots = q_\ell(\ell-1) = 0, \quad q_\ell(\ell) = 1.$$

Furthermore, $q_0(z), q_1(z), \dots, q_\ell(z)$ form a basis of the polynomial space Π_ℓ. So for each $\gamma \in \mathbb{N} \cup \{0\}$, there exist constants $\tau^\gamma_0, \dots, \tau^\gamma_\gamma$ satisfying

$$\binom{n - \frac{z}{2}}{\gamma} = \sum_{j=0}^{\gamma} \tau^\gamma_j \binom{z}{j}.$$

Combining this with (3.4), we get for $\gamma = 0, \dots, p-1$

$$\sum_{i=1}^n \binom{n-i}{\gamma} w'_i = \sum_{i=1}^n \sum_{j=0}^{\gamma} \tau^\gamma_j \binom{2i}{j} w'_i = \sum_{j=0}^{\gamma} \tau^\gamma_j \binom{2n'+1}{j} = \binom{n-n'-\frac{1}{2}}{\gamma}. \tag{A.12}$$

More generally, for each $\gamma_1, \gamma_2 \in \mathbb{N} \cup \{0\}$, there exist constants $\tau^{\gamma_1,\gamma_2}_0, \dots, \tau^{\gamma_1,\gamma_2}_{\gamma_1+\gamma_2}$ satisfying

$$\binom{n - \frac{z}{2}}{\gamma_1}\binom{n - \frac{z}{2}}{\gamma_2} = \sum_{j=0}^{\gamma_1+\gamma_2} \tau^{\gamma_1,\gamma_2}_j \binom{z}{j} \qquad \forall z \in \mathbb{R}.$$

Together with (3.4), we get for $\gamma_1 + \gamma_2 \leqslant p-1$

$$\begin{aligned} \sum_{i=1}^n \binom{n-i}{\gamma_1}\binom{n-i}{\gamma_2} w'_i &= \sum_{i=1}^n \sum_{j=0}^{\gamma_1+\gamma_2} \tau^{\gamma_1,\gamma_2}_j \binom{2i}{j} w'_i = \sum_{j=0}^{\gamma_1+\gamma_2} \tau^{\gamma_1,\gamma_2}_j \binom{2n'+1}{j} \\ &= \binom{n-n'-\frac{1}{2}}{\gamma_1}\binom{n-n'-\frac{1}{2}}{\gamma_2}. \end{aligned} \tag{A.13}$$

Therefore it follows from (A.10), (3.9), (A.12), and (A.13) that for $\alpha, \beta \leqslant k$ and $\alpha + \beta \leqslant p + 1$ we have

$$\begin{aligned}
c^k_{\alpha,\beta} &= \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} c^0_{i,j} \\
&= \sum_{i=1}^{n} \binom{i-1}{\alpha-1}\binom{i-1}{\beta-1} w_i - \sum_{i=1}^{n}\sum_{j=1}^{n} \binom{i-1}{\alpha-1}\binom{j-1}{\beta-1} w_i w_j \\
&= \sum_{i=1}^{n} \binom{n-i}{\alpha-1}\binom{n-i}{\beta-1} w'_i - \left(\sum_{i=1}^{n} \binom{n-i}{\alpha-1} w'_i\right)\left(\sum_{j=1}^{n} \binom{n-j}{\beta-1} w'_j\right) \\
&= \binom{n-n'-\frac{1}{2}}{\alpha-1}\binom{n-n'-\frac{1}{2}}{\beta-1} - \binom{n-n'-\frac{1}{2}}{\alpha-1}\binom{n-n'-\frac{1}{2}}{\beta-1} \\
&= 0.
\end{aligned}$$

We have completed the proof of Lemma 3.6. □

Acknowledgment. We thank Tom Duchamp for many stimulating discussions.

REFERENCES

[1] D. Asimov, *The grand tour: A tool for viewing multidimensional data*, SIAM J. Sci. Stat. Comp., 6 (1985), pp. 128–143.
[2] C. Belta and V. Kumar, *An SVD-based projection method for interpolation on SE(3)*, IEEE Trans. Robotics Automat., 18 (2002), pp. 334–345.
[3] W. M. Boothby, *An Introduction to Differentiable Manifolds and Riemannian Geometry*, 2nd ed., Academic Press, New York, 2002.
[4] A. S. Cavaretta, W. Dahmen, and C. A. Micchelli, *Stationary Subdivision*, Mem. Amer. Math. Soc. 453, AMS., Providence, RI, 1991.
[5] Z. Ciesielski, *Approximation of splines and its application to Lipschitz classes and to stochastic processes*, in Approximation Theory of Functions, Proceedings of the Conference in Kaluga, 1975, Nauka, Moscow, 1977, pp. 397–400.
[6] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* I. *Existence and global regularity of solutions*, SIAM J. Math. Anal., 22 (1991), pp. 1388–1410.
[7] I. Daubechies and J. C. Lagarias, *Two-scale difference equations* II. *Local regularity, infinite products of matrices and fractals*, SIAM J. Math. Anal., 23 (1992), pp. 1031–1079.
[8] I. Daubechies, O. Runborg, and W. Sweldens, *Normal multiresolution approximation of curves*, Constructive Approximation, 3 (2004), pp. 399–463.
[9] G. Deslauriers and S. Dubuc, *Symmetric iterative interpolation processes*, Constr. Approx., 5 (1989), pp. 49–68.
[10] Z. Ditzian, *Moduli of smoothness using discrete data*, J. Approx. Theory, 49 (1987), pp. 115–129.
[11] M. P. do Carmo, *Riemannian Geometry* (translated by F. Flaherty), Birkhäuser Boston, Cambridge, MA, 1992.
[12] S. Dubuc, *Interpolation through an iterative scheme*, J. Math. Anal. Appl., 114 (1986), pp. 185–204.
[13] T. Duchamp, *personal communication*, 2005.
[14] N. Dyn and D. Levin, *Subdivision schemes in geometric modelling*, Acta Numer., 11 (2002), pp. 73–144.
[15] G. Golub and C. van Loan, *Matrix Computations*, 3rd ed., The Johns Hopkins University Press, London, 1996.
[16] B. Han, T. P.-Y. Yu, and Y. Xue, *Non-interpolatory Hermite subdivision schemes*, Math. Comput., 74 (2005), pp. 1345–1367.
[17] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numer., 9 (2000), pp. 215–365.

[18] A. Iserles and S. P. Nørsett, *On the solution of linear differential equations on Lie groups*, Phil. Trans. Royal Soc. A, 357 (1999), pp. 983–1019.
[19] V. Ivancevic, *Symplectic rotational geometry in human biomechanics*, SIAM Rev., 46 (2004), pp. 455–474.
[20] A. Marthinsen, *Interpolation in Lie groups*, SIAM J. Numer. Anal., 37 (1999), pp. 269–285.
[21] J. L. Merrien, *A family of Hermite interpolants by bisection algorithms*, Numer. Algorithms, 2 (1992), pp. 187–200.
[22] P. Oswald, *Smoothness of a nonlinear subdivision scheme*, in Curves and Surface Fitting: Saint-Malo 2002, A. Cohen, J.-L. Merrien, and L. L. Schumaker, eds., Nashboro Press, Brentwood, TN, 2003, pp. 323–332.
[23] P. Oswald, *Smoothness of nonlinear median-interpolation subdivision*, Adv. Comput. Math., 20 (2004), pp. 401–423.
[24] I. Ur Rahman, I. Drori, V. C. Stodden, D. L. Donoho, and P. Schröder, *Multiscale representations for manifold-valued data*, Multiscale Model. Simul., 4 (2005), pp. 1201–1232.
[25] O. Rioul, *Simple regularity criteria for subdivision schemes*, SIAM J. Math. Anal., 23 (1992), pp. 1544–1576.
[26] S. Schaefer, *A Factored, Interpolatory Subdivision Scheme for Surfaces of Revolution*, Master's thesis, Department of Computer Science, Rice University, Houston, TX, 2003.
[27] C. H. Sequin and J. A. Yen, *Fair and robust curve interpolation on the sphere*, in Sketches and Application (SIGGRAPH'01), 2001, p. 182.
[28] J. Wallner, *Gliding spline motions and applications*, Comput. Aided Geom. Design, 21 (2004), pp. 3–21.
[29] J. Wallner, *Smoothness analysis of subdivision schemes by proximity*, Constr. Approx., 24 (2006), pp. 289–318.
[30] J. Wallner and N. Dyn, *Convergence and C^1 analysis of subdivision schemes on manifolds by proximity*, Comput. Aided Geom. Design, 22 (2005), pp. 593–622.
[31] J. Wallner and H. Pottmann, *Intrinsic subdivision with smooth limits for graphics and animation*, ACM Trans. Graphics, 25 (2005), pp. 356–374.
[32] G. Xie and T. P.-Y. Yu, *On a linearization principle for nonlinear p-mean subdivision schemes*, in Advances in Constructive Approximation, M. Neamtu and E. B. Saff, eds., Nashboro Press, Brentwood, TN, 2004, pp. 519–533.
[33] G. Xie and T. P.-Y. Yu, *Smoothness analysis of nonlinear subdivision schemes of homogeneous and affine invariant type*, Construct. Approx., 22 (2005), pp. 219–254.
[34] T. P.-Y. Yu, *On the regularity analysis of interpolatory Hermite subdivision schemes*, J. Math. Anal. Appl., 302 (2005), pp. 201–216.
[35] T. P.-Y. Yu, *Approximation order equivalence properties of manifold-valued data subdivision schemes*, 2006, in preparation.
[36] T. P.-Y. Yu, *Cutting corners on the sphere*, in Wavelets and Splines: Athens 2005, G. Chen and M.-J. Lai, eds., Nashboro Press, Brentwood, TN, 2006, pp. 496–506.
[37] T. P.-Y. Yu, *How data dependent is a nonlinear subdivision scheme? A case study based on convexity preserving subdivision*, SIAM J. Numer. Anal., 44 (2006), pp. 936–948.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1226–1245

STEPSIZE CONDITIONS FOR GENERAL MONOTONICITY IN NUMERICAL INITIAL VALUE PROBLEMS*

M. N. SPIJKER†

Abstract. For Runge–Kutta methods and linear multistep methods, much attention has been paid, in the literature, to special nonlinear stability properties indicated by the terms total-variation-diminishing (TVD), strong-stability-preserving (SSP), and monotonicity. Stepsize conditions, guaranteeing these properties, were studied, e.g., by Shu and Osher [*J. Comput. Phys.*, 77 (1988), pp. 439–471], Gottlieb, Shu, and Tadmor [*SIAM Rev.*, 43 (2001), pp. 89–112], Hundsdorfer and Ruuth [*Monotonicity for Time Discretizations*, Dundee Conference Report NA/217 2003, University of Dundee, Dundee, UK, 2003, pp. 85–94], Higueras [*J. Sci. Comput.*, 21 (2004), pp. 193–223] and [*SIAM J. Numer. Anal.*, 43 (2005), pp. 924–948], Spiteri and Ruuth [*SIAM J. Numer. Anal.*, 40 (2002), pp. 469–491], Gottlieb [*J. Sci. Comput.*, 25 (2005), pp. 105–128], and Ferracina and Spijker [*SIAM J. Numer. Anal.*, 42 (2004), pp. 1073–1093] and [*Math. Comp.*, 74 (2005), pp. 201–219]. In the present paper, we obtain a special stepsize condition guaranteeing the above properties, for a generic numerical process. This condition is best possible in a well defined and natural sense. It is applicable to the important class of general linear methods, and it can also be used to answer some open questions, for methods of which the above stability properties were studied earlier.

Key words. initial value problem, method of lines, ordinary differential equation, general linear method, total-variation-diminishing, strong-stability-preserving, monotonicity

AMS subject classifications. 65L05, 65L06, 65L20, 65M20

DOI. 10.1137/060661739

1. Introduction.

1.1. Maximal stepsize-coefficients for monotonicity.

Consider an initial value problem for a system of ordinary differential equations of type

$$\frac{d}{dt}U(t) = f(t, U(t)) \quad (t \geq 0), \quad U(0) = u_0. \tag{1.1}$$

We shall deal with step-by-step-methods for finding numerical approximations u_n to the true solution values $U(n\Delta t)$, where Δt denotes a positive stepsize and $n = 1, 2, 3, \ldots$.

The general Runge–Kutta method (RKM) for computing u_n can be written in the form

$$y_i = u_{n-1} \; + \; \Delta t \cdot \sum_{j=1}^{s} a_{ij}\, f((n-1+c_j)\Delta t, y_j) \quad (1 \leq i \leq s+1), \tag{1.2.a}$$

$$u_n = y_{s+1}. \tag{1.2.b}$$

Here $a_{i,j}$ and c_j are parameters defining the method, whereas y_i $(1 \leq i \leq s)$ are intermediate approximations used for computing $u_n = y_{s+1}$ from u_{n-1}. If $a_{ij} = 0$ (for $j \geq i$), the method is called *explicit*.

In the following, $\mathbb{V}$ stands for the vector space on which the differential equation is defined, and $\|\cdot\|$ denotes a convex function on $\mathbb{V}$ (i.e., $\|\lambda v + (1-\lambda)w\| \leq \lambda\|v\| + (1-$

*Received by the editors June 1, 2006; accepted for publication (in revised form) November 17, 2006; published electronically May 22, 2007.
http://www.siam.org/journals/sinum/45-3/66173.html
†Mathematical Institute, Leiden University, P. O. Box 9512, NL-2300-RA Leiden, The Netherlands (spijker@math.leidenuniv.nl).

$\lambda)\|w\|$ for $0 \le \lambda \le 1$ and $v, w \in \mathbb{V}$). Much attention has been paid in the literature to the property

$$\|y_i\| \le \|u_{n-1}\| \quad (\text{for } 1 \le i \le s+1). \tag{1.3}$$

Clearly, (1.3) implies $\|u_n\| \le \|u_{n-1}\|$. The latter property, as well as property (1.3), is often referred to by the term *monotonicity* or *strong stability*; it is of particular importance in situations where (1.1) results from (method of lines) semidiscretizations of time-dependent partial differential equations. Choices for $\|\cdot\|$, which occur in that context, include, e.g., the *supremum norm* $\|x\| = \|x\|_\infty = \sup_i |\xi_i|$ and the *total variation seminorm* $\|x\| = \|x\|_{TV} = \sum_i |\xi_{i+1} - \xi_i|$ (for vectors x with components ξ_i). Numerical processes satisfying $\|u_n\|_{TV} \le \|u_{n-1}\|_{TV}$ play a special role in the solution of hyperbolic conservation laws and are called *total variation diminishing*; cf., e.g., Harten [13], Shu [28], Shu and Osher [30], LeVeque [26], and Hundsdorfer and Verwer [21]. We note that, for practical calculations, special importance has been attached to the inequality $\|y_i\| \le \|u_{n-1}\|$ being fulfilled for *all* i with $1 \le i \le s+1$ (rather than just for $i = s+1$); see, e.g., Shu [29] and Gottlieb [8].

Conditions on Δt which guarantee (1.3) were given in the literature, mainly for autonomous differential equations (i.e., f is independent of t). These conditions apply, however, equally well to general f and we discuss them below for that case. In many papers one starts from an assumption about f which, for given $\tau_0 > 0$, essentially amounts to

$$\|v + \tau_0 f(t, v)\| \le \|v\| \quad (\text{for } t \in \mathbb{R},\, v \in \mathbb{V}). \tag{1.4}$$

Assumption (1.4) means that the forward Euler method is monotonic with stepsize τ_0. It can be interpreted as a condition on the manner in which the semidiscretization is performed, in case $\frac{d}{dt}U(t) = f(t, U(t))$ stands for a semidiscrete version of a partial differential equation.

In the literature, *stepsize-coefficients* c were determined such that monotonicity, in the sense of (1.3), is present for all Δt with

$$0 < \Delta t \le c \cdot \tau_0. \tag{1.5}$$

For explicit RKMs, this was done by rewriting the right-hand members of (1.2.a) as convex combinations of forward Euler steps; see, e.g., Shu and Osher [30], Spiteri and Ruuth [31], and Ruuth [27]. For more general RKMs, stepsize-coefficients were obtained, e.g., in Gottlieb, Shu, and Tadmor [10], Higueras [14, 16], and Ferracina and Spijker [6, 7]. We note that, in the context of discretizations for hyperbolic problems, the above coefficients c are sometimes called *CFL coefficients*; see, e.g., Gottlieb and Shu [9] and Shu [29].

The linear multistep method (LMM) for computing u_n can be written in the form

$$u_n = \sum_{j=1}^{k} \alpha_j\, u_{n-j} \;+\; \Delta t \cdot \sum_{j=0}^{k} \beta_j\, f((n-j)\Delta t, u_{n-j}). \tag{1.6}$$

Here α_j, β_j are parameters defining the method and u_n is computed from $u_{n-k}, \ldots, u_{n-1}$. If $\beta_0 = 0$, the method is called *explicit*.

Monotonicity has been studied for (1.6) in the sense of the inequality

$$\|u_n\| \le \max_{1 \le j \le k} \|u_{n-j}\|. \tag{1.7}$$

For explicit LMMs, stepsize-coefficients c, with the property that (1.4), (1.5) guarantee (1.7), were determined by rewriting the right-hand member of (1.6) as a convex combination of forward Euler steps; see, e.g., Shu [28]. Stepsize-coefficients, relevant to more general LMMs, were given, e.g., in Gottlieb, Shu, and Tadmor [10], Hundsdorfer and Ruuth [19], and Hundsdorfer, Ruuth, and Spiteri [20].

Clearly, the larger c is, the less restrictive is condition (1.5). For any given method, the *maximal stepsize-coefficient* c, with the property that (1.4), (1.5) imply monotonicity, is thus an important and characteristic quantity. When comparing the computational efficiency of different methods, it is natural to take these characteristic quantities into account.

Special attention was paid to the problem of determining, for any given RKM, the corresponding maximal stepsize-coefficient; in Higueras [14] and Ferracina and Spijker [6, 7] conditions were given under which this coefficient equals the famous coefficient $R(A, b)$, which was introduced by Kraaijevanger [24]. For completeness, we note also that much attention was paid to the related, but different, problem of optimizing, over given *classes* of RKMs or LMMs, the *special* stepsize-coefficients obtainable via convex combinations of Euler steps; see, e.g., Shu [28], Shu and Osher [30], Gottlieb and Shu [9], Gottlieb [8], Spiteri and Ruuth [31], and Ruuth [27].

Both RKMs and LMMs are examples of methods belonging to the important and very large class of *general linear methods* (GLMs), introduced by Butcher [3], and studied extensively in the literature; see, e.g., Butcher [4, 5], Hairer, Nørsett, and Wanner [12], Hairer and Wanner [11], and the references therein. No theory seems to be available in the literature for determining maximal stepsize-coefficients for arbitrary GLMs.

In this paper, we determine the maximal stepsize-coefficient for a generic numerical process. This result enables us to obtain maximal stepsize-coefficients for arbitrary GLMs and to gain new insights for numerical methods of which the monotonicity properties were studied earlier.

For completeness we note that, already in Burrage and Butcher [2], monotonicity of GLMs was studied, but, for seminorms $\|\cdot\|$ generated by (pseudo) inner products, excluding, e.g., the seminorm $\|\cdot\|_{TV}$. This paper deals with arbitrary convex functions $\|\cdot\|$; as a result, our analysis is largely different from the one in the paper just mentioned.

1.2. Scope of the paper. Section 2 contains our theory for the generic numerical process mentioned above. In section 2.1, we specify GLMs as well as the generic numerical process and characterize them by a pair of matrices S, T. We also give a formal definition of monotonicity.

In section 2.2, we introduce in an algebraic way a coefficient $c(S, T)$, which can be viewed as a generalization of Kraaijevanger's coefficient $R(A, b)$. We state typical properties of $c(S, T)$ in Theorem 2.2. This theorem extends earlier results about $R(A, b)$.

In section 2.3 we state, without proof, the basic results of the paper, Theorems 2.4 and 2.7. These theorems specify situations in which the maximal stepsize-coefficient, for monotonicity of the generic numerical process, is equal to $c(S, T)$. Theorem 2.7 has a wider scope than Theorem 2.4, but the latter theorem has a more simple structure and is of independent interest.

Section 3 contains examples and applications of the theory given in section 2. In section 3.1, we focus on arbitrary GLMs. Theorem 3.1 tells us that the maximal stepsize-coefficient for these methods equals $c(S, T)$. Corollaries 3.3 and 3.4, respec-

tively, show that $c(S,T)$ is not only relevant to monotonicity, but also to a *discrete maximum principle* and *numerical contractivity* of GLMs.

In section 3.2, we apply the preceding theory to RKMs, LMMs, and a class of multistep-multistage methods (MMMs). We arrive at conclusions supplementing earlier results about these methods. In particular, we find (optimal) second order and third order MMMs which we have not seen elsewhere.

In section 3.3, we apply material from section 2 to the interesting class of *additive Runge–Kutta methods*. In this way we obtain Theorem 3.6, which answers an open and fundamental question about these methods.

Section 4 contains the proof of Theorems 2.4 and 2.7. In section 4.1, we prove $c(S,T)$ to be a stepsize-coefficient for the generic numerical process, and in section 4.2 we prove it to be maximal.

2. A theory for monotonicity.

2.1. Monotonicity in a general setting.

2.1.1. General linear methods. The GLM for solving (1.1) depends on parameters c_j $(1 \le j \le m)$ and parameter matrices $S = (s_{i,j}) \in \mathbb{R}^{m\times l}$, $T = (t_{i,j}) \in \mathbb{R}^{m\times m}$, where $1 \le l \le m$. The method can be written in the following form:

$$\text{(2.1.a)} \quad y_i = \sum_{j=1}^{l} s_{ij}\, u_j^{(n-1)} \;+\; \Delta t \cdot \sum_{j=1}^{m} t_{ij}\, f((n-1+c_j)\Delta t,\, y_j) \quad (1 \le i \le m),$$

$$\text{(2.1.b)} \quad u_i^{(n)} = y_{m-l+i} \quad (1 \le i \le l).$$

Here $u_i^{(n-1)}$ are input vectors available at the nth step of the method, whereas y_i are (intermediate) approximations used for computing the input vectors $u_i^{(n)}$ for the next step; cf., e.g., Butcher [4, pp. 336–338] and [5, p. 358] and Hairer, Nørsett, and Wanner [12, p. 390] for related representations of GLMs.

Obviously, the RKM (1.2) is an example of (2.1), with $l = 1,\ m = s+1$, $u_i^{(n)} = u_n \simeq U(n\cdot\Delta t)$, and $s_{i1} = 1,\ t_{ij} = a_{ij}$ (for $1 \le j \le s$), $t_{ij} = 0$ (for $j = s+1$).

The LMM (1.6) is another example of (2.1), with $l = k,\ m = k+1$, and $u_i^{(n)} = u_{n-l+i}$ $(1 \le i \le l)$, $y_i = u_{n-m+i}$ $(1 \le i \le m)$. Method (1.6) can be written in the form (2.1) with $c_j = j-k$, $S = \binom{I}{A}$, $T = \binom{O}{B}$, where I denotes the $k\times k$ identity matrix, O the $k\times(k+1)$ zero matrix and $A = (\alpha_k, \dots, \alpha_1)$, $B = (\beta_k, \dots, \beta_0)$.

We denote the vector space on which the differential equation is defined again by $\mathbb{V}$, and assume $\|\cdot\|$ to be a convex function on $\mathbb{V}$. We will say that method (2.1) is *monotonic* (for the stepsize Δt, function f, and convex function $\|.\|$) if

$$\|y_i\| \;\le\; \max_{1\le j\le l} \|u_j^{(n-1)}\| \quad (\text{for } 1 \le i \le m), \tag{2.2}$$

whenever $u_i^{(n-1)}$ and y_i satisfy (2.1.a). Note that the inequalities (2.2) reduce to (1.3) or (1.7), respectively, if method (2.1) stands for (1.2) or (1.6) in the way just indicated.

In the following, we shall assume that the parameters $s_{i,j}$ satisfy

$$s_{i\,1} + s_{i\,2} + \cdots + s_{i\,l} = 1 \quad (1 \le i \le m). \tag{2.3}$$

This condition is fulfilled if (2.1) stands, in the above way, for (1.2) or (1.6) (provided

$\sum_j \alpha_j = 1$). Moreover, the condition can be seen to be no essential restriction for the general process (2.1): any (preconsistent) GLM can be transformed into an equivalent method satisfying (2.3); see Butcher [5, pp. 358–360] for transformations of GLMs.

2.1.2. A generic numerical process with a simple form. The relations (2.1.a) can be rewritten a bit more compactly. Defining

(2.4)
$$x_i = u_i^{(n-1)} \text{ (for } 1 \le i \le l), \quad f_i(v) = f((n-1+c_i)\Delta t, v) \text{ (for } 1 \le i \le m,\ v \in \mathbb{V}),$$

the relations (2.1.a) reduce to

$$y_i = \sum_{j=1}^{l} s_{ij}\, x_j \ + \ \Delta t \cdot \sum_{j=1}^{m} t_{ij}\, f_j(y_j) \quad (1 \le i \le m). \tag{2.5}$$

Furthermore, when $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ satisfies (1.4), then definition (2.4) implies

$$\|v + \tau_0\, f_i(v)\| \ \le \ \|v\| \quad (\text{for } 1 \le \ i \le \ m \text{ and } v \in \mathbb{V}). \tag{2.6}$$

In the rest of section 2 we shall deal with (2.5) rather than (2.1.a), not only because (2.5) has a more simple form, but also because this widens, in a natural way, the range of applications: in section 3.3 we shall apply our results, to be obtained for the generic process (2.5), to numerical methods which, strictly speaking, are *not* of the form (2.1).

We shall interpret $x_i \in V$ and $y_i \in V$ as *input* and *output vectors*, respectively, of the process (2.5). In the general situation (2.3), (2.5), (2.6), we shall focus on the bound

$$\|y_i\| \ \le \ \max_{1 \le j \le l} \|x_j\| \quad (\text{for } 1 \le i \le m), \tag{2.7}$$

and we will say that process (2.5) is *monotonic* (for the stepsize Δt, functions f_i, and convex function $\|.\|$) if (2.7) holds whenever x_i and y_i satisfy (2.5). Clearly, when (2.5) stands for (2.1.a) via the relations (2.4), then monotonicity of (2.5) corresponds to monotonicity as defined above for the GLM.

In section 2.3 we shall present the basic results of the paper, the best possible stepsize conditions which guarantee monotonicity of process (2.5). In formulating these results we need a coefficient which we first introduce in section 2.2.

2.2. The coefficient $c(S,T)$. Throughout this section we denote by $S \in \mathbb{R}^{m \times l}$ and $T \in \mathbb{R}^{m \times m}$ arbitrary matrices, with property (2.3). In section 2.3 we shall use a coefficient $c(S,T)$ which can be adjoined to S and T. The definition of this coefficient involves the following condition, in which γ denotes a real variable:

$$I + \gamma\, T \text{ is invertible and } (I + \gamma\, T)^{-1} \big[\ S \ \ \gamma\, T\ \big] \ge 0. \tag{2.8}$$

Here I denotes the $m \times m$ identity matrix, and $\big[\ S \ \ \gamma\, T\ \big]$ stands for the $m \times (l+m)$ matrix whose first l columns equal to those of S and whose last m columns equal those of $\gamma\, T$. The inequality in (2.8) should be interpreted entrywise; all inequalities for matrices occurring below are to be interpreted in the same way.

Definition 2.1 (the coefficient $c(S,T)$). *We define $c(S,T) = 0$ if there is no $\gamma > 0$ satisfying* (2.8)*; otherwise*

$$c(S,T) = \sup\{\gamma \ : \ \gamma \text{ satisfies (2.8)}\}.$$

The previous definition may seem to appear out of the blue. The author was led to it, however, by important earlier work of Kraaijevanger [24] and Higueras [16]. In case (2.1) stands for the RKM (1.2) in the way indicated in section 2.1.1, then $c(S,T)$ can be seen to reduce to the coefficient introduced and denoted by $R(A,b)$ in Kraaijevanger [24]; see also section 3.2.1 of this paper. In Higueras [16] the original conditions used by Kraaijevanger for defining his coefficient were simplified to an elegant form which has a resemblance to condition (2.8).

By Definition 2.1 we have $c(S,T) \geq 0$. Part (i) of Theorem 2.2 makes it relatively easy to see whether $c(S,T)$ is zero or not. If $c(S,T) > 0$, part (ii) of Theorem 2.2 is useful for simplifying the (numerical) computation of $c(S,T)$; e.g., by using a bisection-type algorithm as in Ferracina and Spijker [6, section 4.3] and Kraaijevanger [24, p. 498].

In part (i) of Theorem 2.2 we use, for any given matrix $M = (m_{ij})$, the notation $\mathrm{Inc}(M)$ to denote the *incidence matrix* of M (which has the same dimensions as M), given by

$$\mathrm{Inc}(M) = (\widetilde{m}_{ij}), \quad \text{with } \widetilde{m}_{ij} = 1 \text{ (if } m_{ij} \neq 0) \text{ and } \widetilde{m}_{ij} = 0 \text{ (if } m_{ij} = 0).$$

THEOREM 2.2 (properties of $c(S,T)$).

(i) $c(S,T) > 0$ *if and only if* $S \geq 0,\ T \geq 0,\ Inc(T\,S) \leq Inc(S)$, *and* $Inc(T^2) \leq Inc(T)$.

(ii) *Suppose* $0 < \gamma < \infty$ *with* $\gamma \leq c(S,T)$. *Let* $D = \mathrm{diag}(\delta_1, \ldots, \delta_m)$, *where* $0 \leq \delta_i \leq 1$. *Then* (2.8) *holds, with* T *replaced by* TD.

We note that part (ii) of the theorem is already nontrivial and useful in the simple case where D equals the identity matrix I. The theorem can be viewed as an extension (and improvement) of earlier results in the literature; for related results concerning $R(A,b)$, see Kraaijevanger [24, Theorem 4.2, Lemma 4.4], Higueras [15, Proposition 2.11], and Horváth [18, Theorem 4].

Below we shall prove Theorem 2.2 using Lemma 2.3. We think the lemma is of independent interest: it gives an interesting interpretation of (2.8). We shall use for $x \in \mathbb{R}^n$ (with components ξ_i) and arbitrary $A = (a_{i,j}) \in \mathbb{R}^{m\times n}$ the notations $\|x\|_\infty = \max_i |\xi_i|$, $\|A\|_\infty = \max_{x\neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty}$, and the well known formula $\|A\|_\infty = \max_i \sum_j |a_{i,j}|$.

LEMMA 2.3 (interpretation of (2.8)). *Let* $0 < \gamma < \infty$. *Then* (2.8) *holds if and only if* $(I + \gamma\,T)$ *is invertible and* $\|(I + \gamma\,T)^{-1}\big[\ S\ \ \gamma\,T\ \big]\|_\infty \leq 1$.

Proof of Lemma 2.3. For any integer $q \geq 1$ we denote by e_q the vector in $\mathbb{R}^q$ with all components equal to 1. We assume that $0 < \gamma < \infty$ and $I + \gamma\,T$ is invertible.

In view of (2.3) we have $Se_l = e_m$. Introducing the matrices

$$P = (p_{i,j}) = (I + \gamma\,T)^{-1}(\gamma T),\ \ Q = (q_{i,j}) = (I + \gamma\,T)^{-1},\ \ R = (r_{i,j}) = Q\,S, \tag{2.9}$$

we thus have $R = (I - P)S$ and $\big[\ R\ \ P\ \big]e_{l+m} = R\,e_l + P\,e_m = e_m$. Consequently,

$$\sum_{j=1}^{l} r_{i,j} + \sum_{j=1}^{m} p_{i,j} = 1 \quad (\text{for } 1 \leq i \leq m). \tag{2.10}$$

If (2.8) holds, then all $r_{i,j},\ p_{i,j}$ are nonnegative, so that (2.10) implies $\|\big[\ R\ \ P\ \big]\|_\infty \leq 1$. Conversely, if $\|\big[\ R\ \ P\ \big]\|_\infty \leq 1$, then $\sum_{j=1}^{l} |r_{i,j}| + \sum_{j=1}^{m} |p_{i,j}| \leq \sum_{j=1}^{l} r_{i,j} + \sum_{j=1}^{m} p_{i,j}$. The last inequality proves that all $r_{i,j},\ p_{i,j}$ are nonnegative, so that (2.8) holds. □

We now turn to the proof of Theorem 2.2.

Proof of Theorem 2.2(i). In view of part (ii) of Theorem 2.2 (to be proved below), we have $c(S,T) > 0$ if and only if there is a $\gamma_0 > 0$ such that the matrix $M(\gamma) = (I+\gamma T)^{-1}\left[\begin{array}{cc} S & \gamma T \end{array}\right]$ is nonnegative for all $\gamma \in [0,\gamma_0]$. Therefore, we can assume with no loss of generality that $S \geq 0,\ T \geq 0$.

We have, for $\gamma > 0$ sufficiently small, $M(\gamma) = \{\sum_{k=0}^{\infty}(\gamma T)^{2k}\}\left[\begin{array}{cc} (I-\gamma T)S & (I-\gamma T)\gamma T \end{array}\right]$. It follows that $M(\gamma) \geq 0$ for $\gamma \downarrow 0$ if and only if $\mathrm{Inc}(TS) \leq \mathrm{Inc}(S)$ and $\mathrm{Inc}(TT) \leq \mathrm{Inc}(T)$, which proves (i).

Proof of Theorem 2.2(ii). Let γ_i be any finite values with $0 \leq \gamma_i \leq c(S,T)$ and put $\Gamma = \mathrm{diag}(\gamma_1 \ldots \gamma_m)$. In order to prove statement (ii), it is enough to assume $c(S,T) > 0$ and to show that (2.8) holds with matrix γT replaced throughout by the product $T\Gamma$.

Choose any finite γ satisfying (2.8) with $0 < \gamma \leq c(S,T)$ and put $E = \mathrm{diag}(\varepsilon_1 \ldots \varepsilon_m)$, where $\varepsilon_i = (\gamma - \gamma_i)\gamma^{-1}$. In order to prove the invertibility of $I + T\Gamma$, we write

$$I + T\Gamma = (I+\gamma T)(I-X), \text{ with } X = PE \text{ and } P \text{ as in (2.9)}.$$

In view of Lemma 2.3, we have $\|P\|_\infty \leq 1$, so that $\|X\|_\infty \leq \|E\|_\infty$.

First, consider the special case where $c(S,T) < \infty$ and $\gamma_i = c(S,T)$ (for $1 \leq i \leq m$). Choosing the above γ sufficiently close to $c(S,T)$, we can arrange that $\|E\|_\infty < 1$, which implies that in this special case $I + T\Gamma = I + c(S,T)\,T$ is invertible. Using a continuity argument it follows that (2.8) holds with $\gamma = c(S,T)$.

Next, consider again the general case of arbitrary finite γ_i with $0 \leq \gamma_i \leq c(S,T) \leq \infty$. In view of the above, we can choose a positive γ satisfying (2.8), with $\gamma \geq \max_i \gamma_i$. With this γ we have $0 \leq X = PE \leq P$, which implies that the spectral radii of X and P satisfy $\mathrm{spr}(X) \leq \mathrm{spr}(P) \leq \|P\|_\infty \leq 1$. In view of (2.9), the matrix $Q = I - P$ is invertible, so that P has no eigenvalue equal to 1. Applying the Perron–Frobenius theory (see, e.g., Horn and Johnson [17, p. 503]), it follows that $\mathrm{spr}(P) < 1$. Hence $\mathrm{spr}(X) < 1$, so that $I + T\Gamma = (I+\gamma T)(I-X)$ has an inverse equal to $(I-X)^{-1}(I+\gamma T)^{-1}$. Using $(I+T\Gamma)^{-1} = \left(\sum_0^\infty X^k\right)(I+\gamma T)^{-1}$, it follows that (2.8) is valid with γT replaced by $T\Gamma$. □

2.3. Stepsize-coefficients for monotonicity. In this section we give, without proof, the basic results of the paper, Theorems 2.4 and 2.7. Throughout the section, $S \in \mathbb{R}^{m\times l}$ and $T \in \mathbb{R}^{m\times m}$ are again arbitrary matrices satisfying (2.3). We study stepsize conditions $0 < \Delta t \leq c \cdot \tau_0$, guaranteeing monotonicity of process (2.5) when f_i satisfies (2.6). The following inequality will be of crucial importance:

$$c \ \leq \ c(S,T). \tag{2.11}$$

Our first result is as follows.

THEOREM 2.4 (monotonicity for arbitrary f_i satisfying (2.6)). *Consider numerical process* (2.5). *Let* τ_0, c *be given with* $0 < \tau_0 < \infty$, $0 < c \leq \infty$. *Then each of the following statements* (2.12), (2.13) *is equivalent to* (2.11)*:*

(2.12) *Condition* $0 < \Delta t \leq c \cdot \tau_0$ *implies monotonicity whenever* $\mathbb{V}$ *is a vector space,* $\|\cdot\|$ *a convex function on* $\mathbb{V}$, *and arbitrary* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfy* (2.6)*;*

(2.13) *Condition* $0 < \Delta t \leq c \cdot \tau_0$ *implies monotonicity when* $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$, *and arbitrary* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfy* (2.6).

Clearly, (2.12) is a priori a stronger statement than (2.13). Accordingly, the essence of Theorem 2.4 is that the (algebraic) property (2.11) implies the (strong) statement (2.12), whereas already the (weaker) statement (2.13) implies (2.11).

The theorem highlights the importance of the quantity $c(S,T)$: Theorem 2.4 shows that, with respect to the situations specified in (2.12), (2.13), *the maximal stepsize-coefficient* c, with the property that condition $0 < \Delta t \leq c \cdot \tau_0$ guarantees monotonicity, is equal to $c(S,T)$.

Our second result, Theorem 2.7, deals with important situations not adequately covered by Theorem 2.4: it is often *not* natural to allow, as in Theorem 2.4, that all functions f_i are different from each other. For instance, if in method (2.1) we have $c_i = c_j$ for some $i \neq j$, or if the differential equation is autonomous, then (2.1) is represented by a process (2.5) with $f_i = f_j$ for some, or all, indices $i \neq j$. In section 3.3 we will come across another situation where the functions f_i in (2.5) are not independent of each other. In order to cover all of such cases, we consider index sets I_q with $I_q \subset \{1, \ldots, m\}$ (for $1 \leq q \leq r$) and functions $f_i : \mathbb{V} \to \mathbb{V}$ (for $1 \leq i \leq m$), such that

(2.14)
$I_1, \ldots, I_r$ are nonempty and mutually disjoint, with $I_1 \cup \cdots \cup I_r = \{1, \ldots, m\}$,

(2.15) $f_i = f_j$ whenever i and j belong to the same index set I_q.

According to Theorem 2.4, also when (2.14), (2.15) hold, inequality (2.11) is *sufficient* in order that condition $0 < \Delta t \leq c \cdot \tau_0$ implies monotonicity of numerical process (2.5); but the following counterexample shows that, under the assumptions (2.14), (2.15), the *maximal stepsize-coefficient* $c = c_{max}$ can be larger than $c(S,T)$.

EXAMPLE 2.5. *Consider process* (2.5) *with* $l = 1$, $m = 2$, *and* $s_{i,1} = 1$, $t_{i,1} = 2$, $t_{i,2} = -1$ *(for* $i = 1, 2$*). Suppose* (2.14), (2.15) *with* $r = 1$, $I_1 = \{1, 2\}$. *Since condition* $T \geq 0$ *in Theorem* 2.2(i) *is violated, we have* $c(S,T) = 0$. *But, with* $f_1 = f_2 = f$, *the process reduces to the (backward Euler) method* $y_1 = x_1 + \Delta t\, f(y_1)$, *which is again of the form* (2.5), *with* $\tilde{l} = \tilde{m} = 1$ *and* $c(\tilde{S}, \tilde{T}) = \infty$. *In line with Theorem* 2.4, *the maximal stepsize-coefficient* $c = c_{max}$ *such that condition* $0 < \Delta t \leq c \cdot \tau_0$ *implies monotonicity (for the original process with* $m = 2$ *and* (2.6), (2.14), (2.15) *in force), is equal to* $c_{max} = \infty > c(S,T) = 0$.

Theorem 2.7 will make clear that the inequality $c_{max} > c(S,T)$, in Example 2.5, is an anomaly related to reducibility of the method. We shall use the following formal definition of reducibility and irreducibility, with regard to index sets $I_1, \ldots, I_r$ satisfying (2.14).

DEFINITION 2.6 (reducibility and irreducibility). *Process* (2.5) *is called reducible with respect to* $I_1, \ldots, I_r$, *if indices* i, j, q *exist with the following properties:* $i \in I_q$, $j \in I_q$, *and* $i \neq j$, *whereas the* i*th and the* j*th row of the matrix* $\begin{bmatrix} S & T \end{bmatrix}$ *are equal to each other. Process* (2.5) *is called irreducible with respect to* $I_1, \ldots, I_r$, *if such indices* i, j, q *do not exist.*

Clearly, if $r < m$ and there is reducibility with respect to $I_1, \ldots, I_r$, then process (2.5), with f_i satisfying (2.15), is equivalent to a process (2.5) with a smaller value of m.

THEOREM 2.7 (monotonicity when f_i satisfy (2.6), (2.15)). *Assume* (2.14) *and irreducibility of process* (2.5) *with respect to* $I_1, \ldots, I_r$. *Let* τ_0, c *be given with* $0 < \tau_0 < \infty$, $0 < c \leq \infty$. *Then each of the following statements* (2.16), (2.17) *is equivalent*

to (2.11)*:*

(2.16) *Condition* $0 < \Delta t \le c \cdot \tau_0$ *implies monotonicity whenever* $\mathbb{V}$ *is a vector space,* $\|\cdot\|$ *a convex function on* $\mathbb{V}$*, and functions* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfy* (2.6), (2.15);

(2.17) *Condition* $0 < \Delta t \le c \cdot \tau_0$ *implies monotonicity whenever* $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$*, and functions* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfy* (2.6), (2.15).

Theorem 2.7 implies that, in the situations specified by (2.16) and (2.17), the *maximal stepsize-coefficient* c, such that condition $0 < \Delta t \le c \cdot \tau_0$ implies monotonicity, is still equal to $c(S,T)$, provided there is irreducibility with respect to the relevant index sets.

The following counterexample shows that the dimension of the space $\mathbb{V}$ in statements (2.13) and (2.17) cannot be replaced, in general, by an integer smaller than m.

EXAMPLE 2.8. *Consider numerical process* (2.5) *with* $l = 1$, $m = 2$*, and* $s_{1,1} = s_{2,1} = 1$, $t_{1,1} = 1/3$, $t_{1,2} = 8/3$, $t_{2,1} = 0$, $t_{2,2} = 1$*. A straightforward calculation yields* $c(S,T) = 3/5$*. On the other hand, it can be proved that propositions* (2.13) *and* (2.17) *would be valid with* $c = 3/2 > c(S,T)$*, if the space* $\mathbb{V} = \mathbb{R}^m = \mathbb{R}^2$ *would be replaced by* $\mathbb{V} = \mathbb{R}^1$*.*

Theorem 2.4 can formally be viewed as a special case of Theorem 2.7; the latter theorem with $r = m$ reduces to the former. We have formulated Theorem 2.4 separately in view of its importance and simplicity: it does not need (2.14), (2.15) or Definition 2.6. Furthermore, in section 4, where the theorems are proved, we will see that it is convenient to focus first on Theorem 2.4 and to use (arguments used in the proof of) that theorem for proving Theorem 2.7.

3. Examples and applications.

3.1. Applications to arbitrary GLMs. In this section we consider method (2.1). We assume (2.3) and give some results which follow readily from the above theory. We focus on stepsize-coefficients c such that

(3.1) Condition $0 < \Delta t \le c \cdot \tau_0$ implies monotonicity whenever $\mathbb{V}$ is a vector space, $\|\cdot\|$ a convex function on $\mathbb{V}$, and functions $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ satisfy (1.4).

In the following theorem, the columns of the matrix $T = (t_{ij})$ are denoted by T_i $(1 \le i \le m)$ and the rows of the $m \times (l+m)$ matrix $[\,S \;\; T\,]$ by R_i $(1 \le i \le m)$.

THEOREM 3.1 (monotonicity of GLMs). *Consider method* (2.1)*, and given* $\tau_o > 0$.

(i) *Let* $c \le c(S,T)$*. Then statement* (3.1) *is valid.*

(ii) *Assume the method is irreducible in the sense that* $R_i \ne R_j$ *for all* i, j *with* $i \ne j$, $T_i \ne 0$, $T_j \ne 0$, $c_i = c_j$*. Then, conversely, statement* (3.1) *implies that* $c \le c(S,T)$.

Note that the irreducibility assumption in (ii) is trivially fulfilled if the method is *nonconfluent*, i.e., if $c_i \ne c_j$ (for all $i \ne j$). Moreover, in case $c_i = c_j$ (for some $i \ne j$), the assumption of irreducibility is *no* strong restriction, because any given method, violating the assumption, is equivalent to a method (with a smaller number of stages) which is irreducible. The theorem highlights the importance of $c(S,T)$ for (irreducible) GLMs: it implies that the *maximal stepsize-coefficient* c, with property (3.1), is equal to $c(S,T)$.

Proof of Theorem 3.1. In order to prove (i), it is enough to apply Theorem 2.4 to process (2.5), where x_i and f_i are defined via (2.4).

For proving (ii), note that, when $T_k = 0$, the value of the parameter c_k is *irrelevant* to monotonicity of method (2.1). We can thus arrange, without loss of generality, that $c_k \neq c_i$ (for $T_k = 0$ and $i \neq k$). Under the irreducibility assumption in (ii), we thus have

$$R_i \neq R_j \quad \text{whenever} \quad i \neq j,\ c_i = c_j.$$

In order to apply Theorem 2.7, we specify index sets I_q by the following requirement: indices i, j belong to the same index set, if and only if $c_i = c_j$. Since process (2.5) is now irreducible in the sense of Definition 2.6, we can apply Theorem 2.7. For proving (ii) it is thus enough to show that (3.1) (for the GLM) implies (2.16) (for process (2.5)).

In order to prove the last implication, we assume (3.1) and suppose x_i, y_i satisfy (2.5) with $0 < \Delta t \leq c \cdot \tau_0$ and functions $f_i : \mathbb{V} \to \mathbb{V}$ satisfying (2.6), (2.15).

We shall show that (2.1) holds with $u_i^{(n-1)} = x_i$ and some function f satisfying (1.4). In defining f we use the notations $t_i = (n-1+c_i)\Delta t$, $\alpha = \min t_i$, $\beta = \max t_i$. We put $f(t_i, v) = f_i(v)$, and extend this function to a function $f : \mathbb{R} \times \mathbb{V} \to \mathbb{V}$ by linear interpolation for $\alpha \leq t \leq \beta$, and by setting $f(t, v) = f(\alpha, v)$ (for $t < \alpha$) and $f(t, v) = f(\beta, v)$ (for $t > \beta$). This function f satisfies (1.4); and (2.1) holds with $u_i^{(n-1)} = x_i$.

Applying (3.1), it follows that (2.2)—and therefore also (2.7)—is fulfilled. This implies (2.16) and concludes the proof. □

Remark 3.2. Theorems 2.4 and 2.7, used in the above proof, can also be applied to prove a variant of Theorem 3.1 tuned to *autonomous* differential equations. In such a variant, property (3.1) is modified by including that f is independent of t, and the irreducibility condition in (ii) becomes: $R_i \neq R_j$ whenever $i \neq j$, $T_i \neq 0$, $T_j \neq 0$.

The subsequent corollaries to Theorem 3.1 involve two properties different in appearance from (2.2).

Property 1 (discrete maximum principle). Let $\mathbb{V} = \mathbb{R}^N$, $N \geq 1$. Suppose vectors y_i, $u_i^{(n-1)} \in \mathbb{R}^N$ satisfy (2.1.a). We denote the components of these vectors by y_{pi} and $u_{pi}^{(n-1)}$, respectively ($1 \leq p \leq N$). The property

(3.2)
$$\min_{1\leq j\leq l} \min_{1\leq q\leq N} u_{qj}^{(n-1)} \ \leq\ y_{pi} \ \leq\ \max_{1\leq j\leq l} \max_{1\leq q\leq N} u_{qj}^{(n-1)} \quad (\text{for } 1 \leq i \leq m,\ 1 \leq p \leq N)$$

can be interpreted as a *discrete maximum principle*. It is of importance in the solution of partial differential equations (via the method of lines) and can be associated with the absence of undesirable overshoots and undershoots; see, e.g., Hundsdorfer and Verwer [21, pp. 9 and 118]. Below we denote the components of $f(t, x) \in \mathbb{R}^N$ by $f_p(t, x)$ ($1 \leq p \leq N$).

Corollary 3.3 (discrete maximum principle for GLMs). *Let $f : \mathbb{R} \times \mathbb{R}^N \to \mathbb{R}^N$ and $\tau_0 > 0$ be such that, for $x \in \mathbb{R}^N$ with components ξ_p,*

$$\min_{1\leq q\leq N} \xi_q \ \leq\ \xi_p + \tau_0 \cdot f_p(t, x) \ \leq\ \max_{1\leq q\leq N} \xi_q \quad (\textit{for } 1 \leq p \leq N).$$

Then (3.2) *holds, whenever $u_i^{(n-1)}$ and y_i satisfy* (2.1.a) *with $0 < \Delta t \leq c(S, T) \cdot \tau_0$.*

Proof of Corollary 3.3. Define the convex functions $\|x\|_+ = \max_p \xi_p$ and $\|x\|_- = -\min_p \xi_p$ for $x \in \mathbb{V}$. The assumption in the corollary, about f and τ_0, can be

rewritten as

$$\|x+\tau_0 f(t,x)\|_- \le \|x\|_- , \quad \|x+\tau_0 f(t,x)\|_+ \le \|x\|_+,$$

so that (1.4) holds with $\|\cdot\|$ equal to $\|\cdot\|_-$ and $\|\cdot\|_+$, respectively.

Assume (2.1.a) with $0 < \Delta t \le c(S,T)\cdot\tau_0$. Choosing $c = c(S,T)$ and applying Theorem 3.1(i), we get from (3.1) the inequalities $\|y_i\|_+ \le \max_{1\le j\le l}\|u_i^{(n-1)}\|_+$ and $\|y_i\|_- \le \max_{1\le j\le l}\|u_i^{(n-1)}\|_-$, which imply (3.2). □

Property 2 (contractivity). Let $\|.\|$ be a convex function on the vector space $\mathbb{V}$. We consider the *contractivity* property

$$\|\tilde y_i - y_i\| \le \max_{1\le j\le l}\|\tilde u_j^{(n-1)} - u_j^{(n-1)}\| \quad (\text{for } 1\le i\le m), \tag{3.3}$$

where $u_i^{(n-1)}$, y_i and $\tilde u_i^{(n-1)}$, $\tilde y_i$ satisfy (2.1.a) with the same stepsize $\Delta t > 0$. Contractivity of numerical processes were studied earlier in various frameworks; cf., e.g., Kraaijevanger [24] Hairer and Wanner [11].

COROLLARY 3.4 (contractivity for GLMs). *Let $f : \mathbb{R}\times\mathbb{V}\to\mathbb{V}$ and $\tau_0 > 0$ be such that $\|\tilde v - v + \tau_0\cdot(f(t,\tilde v) - f(t,v))\| \le \|\tilde v - v\|$ (for $t\in\mathbb{R}$ and v, $\tilde v\in\mathbb{V}$). Then* (3.3) *holds, whenever $u_i^{(n-1)}$, y_i and $\tilde u_i^{(n-1)}$, $\tilde y_i$ satisfy* (2.1.a) *with $0 < \Delta t \le c(S,T)\cdot\tau_0$.*

Proof of Corollary 3.4. The corollary follows from Theorem 3.1, using arguments similar to those in Burrage and Butcher [2, p. 190]: We introduce the auxiliary space $\mathbb{W} = \mathbb{V}\times\mathbb{V}$ and put $\|w\| = \|\tilde v - v\|$, $g(t,w) = (f(t,\tilde v), f(t,v))$ (for $w = (\tilde v, v)$ with $\tilde v, v\in\mathbb{V}$). The above assumption, about f and τ_0, implies that $\|w+\tau_0\cdot g(t,w)\| \le \|w\|$ (for $w\in\mathbb{W}$).

Let $u_i^{(n-1)}$, y_i and $\tilde u_i^{(n-1)}$, $\tilde y_i$ satisfy (2.1.a). Defining $U_i = (\tilde u_i^{(n-1)}, u_i^{(n-1)})$, $Y_i = (\tilde y_i, y_i)$, we have $Y_i = \sum_j s_{ij} U_j + \Delta t\cdot\sum_j t_{ij} g((n-1+c_j)\Delta t,\ Y_j)$ and $\|Y_i\| = \|\tilde y_i - y_i\|$, $\|U_i\| = \|\tilde u_i^{(n-1)} - u_i^{(n-1)}\|$. An application of Theorem 3.1(i) (to the space $\mathbb{W}$ and the function g) proves the proposition. □

3.2. Applications to RKMs, LMMs, and a MMM. We illustrate the preceding theory by applying it to some concrete numerical methods.

3.2.1. Runge–Kutta methods. Consider method (1.2). We denote by A_{s+1} the $(s+1)\times s$ matrix with entries a_{ij} and by A_s the matrix of order s obtained from A_{s+1} by omitting its last row. By E_{s+1} and E_s, respectively, we denote the $(s+1)\times 1$ and the $s\times 1$ matrix with all entries equal to 1. In section 2.1.1, method (1.2) was already represented as a GLM of form (2.1), with $l = 1$, $m = s+1$ and

$$S = E_{s+1},\ T = [\,A_{s+1}\ O\,].$$

Monotonicity of this GLM amounts to (1.3). Hence, according to Theorem 3.1, the largest stepsize-coefficient c, such that (3.1) holds for the RKM, is essentially equal to $c(S,T)$. Below we reformulate this result in a more explicit form.

For S, T just defined, it follows easily, similarly as in Higueras [16], that (2.8) is equivalent to the following condition:

(3.4)
$I+\gamma A_s$ is invertible and $A_{s+1}(I+\gamma A_s)^{-1} \ge O$, $E_{s+1} \ge \gamma A_{s+1}(I+\gamma A_s)^{-1}E_s$.

In view of Definition 2.1, it thus follows, after a simple application of Theorem 2.2(i), that $c(S,T) = \Gamma$, where

$$\Gamma = \sup\{\gamma :\ \gamma \text{ satisfies (3.4)}\}\ (\text{if } A_{s+1}\ge O) \quad \text{and}\ \ \Gamma = 0\ \ (\text{otherwise}). \tag{3.5}$$

We denote the rows of the $s \times s$ matrix A_s by $r_1, \ldots, r_s$. Applying Theorem 3.1, we immediately arrive at the following two conclusions:

(i) *For method* (1.2)*, statement* (3.1) *is valid with* $c = \Gamma$*, where* Γ *is given by* (3.5).

(ii) *Assume the RKM is irreducible in the sense that* $r_i \neq r_j$ *for all* i, j *with* $i \neq j$, $c_i = c_j$. *Then the value* $c = \Gamma$ *in conclusion* (i) *is optimal, in that* (3.1) *is not valid with* $c > \Gamma$.

These results imply that for (irreducible) RKMs the *maximal stepsize-coefficient* c with property (3.1) equals Γ. Statements (i) and (ii) supplement related material in Higueras [14, 16] and Ferracina and Spijker [6, 7]. The irreducibility condition in (ii) is essentially weaker than in these papers, whereas the monotonicity property (3.1) in (i) and (ii) is stronger than in (most of) the papers.

Definition (3.5) can be viewed as a smooth variant of similar definitions in the papers just mentioned. For many RKMs, the corresponding value of Γ is explicitly known, because it equals the coefficient introduced, and denoted by $R(A, b)$, in Kraaijevanger [24]; the equality $\Gamma = R(A, b)$ is an easy consequence of Theorem 2.2(ii). For various interesting RKMs, the actual value of $R(A, b)$ was studied and computed in the last mentioned paper; see also Higueras [14] and Ferracina and Spijker [6].

Versions of (i) and (ii) tuned to *autonomous* differential equations can easily be obtained by applying the variant of Theorem 3.1 mentioned in Remark 3.2. In these versions, the assumption on f in (3.1) includes that f is independent of t, and the irreducibility condition on the RKM becomes: $r_i \neq r_j$ (whenever $i \neq j$).

3.2.2. Linear multistep methods. Consider method (1.6), with $\sum_1^k \alpha_i = 1$. In section 2.1.1, the method was represented as a GLM of form (2.1), with $l = k,\ m = k + 1,\ c_j = j - k,\ S = \binom{I}{A},\ T = \binom{O}{B}$, where $A = (\alpha_k, \ldots, \alpha_1),\ B = (\beta_k, \ldots, \beta_0)$. This GLM is irreducible in the sense of Theorem 3.1, because $c_i \neq c_j$ (for $i \neq j$). Its monotonicity amounts to (1.7). Theorem 3.1 thus implies that the largest c, for which the LMM has property (3.1), is equal to $c = c(S, T)$.

In order to find a convenient expression for $c(S, T)$, we consider, for $\gamma > 0$, condition (2.8) with S, T as defined above. One easily sees (using $\sum_1^k \alpha_i = 1$) that (2.8) is equivalent to the requirement that $\beta_0 \geq 0$ and $\alpha_i \geq 0,\ \beta_i \geq 0,\ \alpha_i - \gamma\,\beta_i \geq 0\ \ (1 \leq i \leq k)$. By Definition 2.1, we obtain $c(S, T) = \Gamma$, where

$$\Gamma = \min_{1 \leq i \leq k} \alpha_i/\beta_i \ \text{ (if all } \alpha_i, \beta_i \text{ are nonnegative)} \quad \text{and} \quad \Gamma = 0 \ \text{ (otherwise).} \tag{3.6}$$

Here we use the convention that $a/0 = \infty$ for all $a \geq 0$. In view of the above, we have the following conclusions:

(i) *For method* (1.6)*, statement* (3.1) *is valid with* $c = \Gamma$*, where* Γ *is given by* (3.6).

(ii) *The value* $c = \Gamma$ *in conclusion* (i) *is optimal, in that* (3.1) *is not valid with* $c > \Gamma$.

Statements (i) and (ii) imply that the *maximal stepsize-coefficient* c with property (3.1) equals Γ. Results similar to (i) were given earlier; see, e.g., Shu [28], Gottlieb, Shu, and Tadmor [10], Hundsdorfer and Ruuth [19], and Hundsdorfer, Ruuth, and Spiteri [20].

As an illustration we consider the following LMM, taken from Shu [28]:

$$u_n = \tfrac{3}{4}\, u_{n-1} + \tfrac{1}{4}\, u_{n-3} + \tfrac{3}{2}\, \Delta t\, f((n-1)\Delta t, u_{n-1}). \tag{3.7}$$

For this second order method, we have $\Gamma = 1/2$, so that (3.1) holds with $c = 1/2$. In Gottlieb, Shu, and Tadmor [10], the method was proved to be optimal, in that there exists no explicit second order method (1.6), with $k = 3$ and $\Gamma > 1/2$. In view of statement (ii), it follows that (3.7) is even *optimal in a wider and more fundamental sense* than stated in the last paper: there exists no explicit second order method (1.6), with $k = 3$, satisfying (3.1) with $c > 1/2$.

Versions of (i) and (ii) for *autonomous* differential equations follow again from the variant of Theorem 3.1 mentioned in Remark 3.2. No explicit irreducibility assumption, about the LMM, is needed in these versions.

3.2.3. A multistep-multistage method. We illustrate Theorems 2.2 and 3.1 with the method

$$v_n = \gamma_1\, u_{n-1} + \gamma_2\, u_{n-2} \;+\; \Delta t \cdot [\delta_0\, f_n + \delta_1\, f_{n-1} + \delta_2\, f_{n-2} + \delta_3\, g_n], \tag{3.8.a}$$

$$u_n = \alpha_1\, u_{n-1} + \alpha_2\, u_{n-2} \;+\; \Delta t \cdot [\beta_0\, f_n + \beta_1\, f_{n-1} + \beta_2\, f_{n-2} + \beta_3\, g_n]. \tag{3.8.b}$$

Here v_n is an intermediate approximation used for computing u_n from u_{n-1}, u_{n-2}, and $f_n = f(n\Delta t,\, u_n)$, $g_n = f((n-\sigma)\Delta t,\, v_n)$. We assume $\alpha_1 + \alpha_2 = \gamma_1 + \gamma_2 = 1$ and, in order to prevent reducibility, that the coefficient vectors $(\alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2, \beta_3)$ and $(\gamma_1, \gamma_2, \delta_0, \delta_1, \delta_2, \delta_3)$ are different. Method (3.8) can be viewed as a modified LMM or a two-step RKM; cf., e.g., Butcher [5] and Jackiewicz and Tracogna [22]. In Gottlieb, Shu, and Tadmor [10, pp. 102 and 103], methods of type (3.8) were explored which are *explicit*, i.e., $\beta_0 = \delta_0 = \delta_3 = 0$.

The general method (3.8) can be written as a GLM (2.1), with $l = 2$, $m = 4$ and with matrices S, T, determined by α_i, β_i, γ_i, δ_i, such that $y_1 = v_n$, $y_2 = u_{n-2}$, $y_3 = u_{n-1}$, $y_4 = u_n$. The monotonicity relation (2.2) reduces to $\max\{\|v_n\|, \|u_n\|\} \le \max\{\|u_{n-1}\|, \|u_{n-2}\|\}$. Applying Theorem 3.1, we conclude that the largest c, for which method (3.8) satisfies (3.1), is equal to $c(S,T)$. By combining this conclusion with Theorem 2.2(i), we arrive after a short calculation at the following proposition.

PROPOSITION 3.1. *For method* (3.8)*, a positive c exists with property* (3.1)*, if and only if all coefficients α_i, β_i, γ_i, δ_i are nonnegative, with $\frac{\alpha_i}{\beta_i + \beta_3\,\gamma_i} > 0$, $\frac{\gamma_i}{\delta_i + \delta_0\,\alpha_i} > 0$ $(i = 1, 2)$ and $\frac{\beta_{i-1}}{\beta_3\,\delta_{i-1}} > 0$, $\frac{\delta_i}{\delta_0\,\beta_i} > 0$ $(i = 1, 2, 3)$.*

Here we use again the convention that $a/0 = \infty > 0$ for $a \ge 0$.

With E_2 we denote the class of all *explicit* methods (3.8) with *order of accuracy* 2. We consider the problem of determining a method in the class which is *optimal*, in that it has property (3.1) with a value c which is maximal in E_2. In view of Theorem 3.1, this problem amounts to finding the maximum of $c(S,T)$ over the class E_2. According to Definition 2.1, this maximum can be computed by performing an optimization, with objective function γ and search variables α_i, β_i, γ_i, δ_i, σ, γ, under the constraints (2.8), supplemented by the order conditions. We performed a numerical search along these lines (using MATLAB) and obtained an (optimal) method of form (3.8), for which we found that the nonzero parameters can be represented (up to 13 decimal digits) as follows:

$$\alpha_1 = 2(\sqrt{2}-1),\;\; \alpha_2 = 3-2\sqrt{2},\;\; \beta_1 = \beta_3 = 2-\sqrt{2},\;\; \gamma_1 = 1,\;\; \delta_1 = \sqrt{2}/2,\;\; \sigma = 1-\sqrt{2}/2.$$

This method is of order 2 and it satisfies (3.1) with stepsize-coefficient $c = \sqrt{2}$.

Second order methods with a larger stepsize-coefficient can be found in the class G_2 of *general second order* methods (3.8). By a numerical search in G_2, similar to the above, we arrived at an (optimal) method with the following nonzero parameters:

$$\alpha_1 = \gamma_1 = 1,\;\; \beta_0 = \beta_1 = \delta_1 = \delta_3 = 1/4,\;\; \beta_3 = 1/2,\;\; \sigma = 1/2.$$

This method is of order 2 and it satisfies (3.1) with stepsize-coefficient $c = 4$. Note that the method is equivalent to two applications of the trapezoidal rule (TR) starting from u_{n-1} and using stepsize $\Delta t/2$; we refer to Lenferink [25, p. 180] for a related interesting optimality property of the TR.

We also performed a similar numerical search in the class E_3 of all *explicit third order* methods (3.8). Our search resulted in an (optimal) method for which the nonzero parameters can be represented (up to 13 decimal digits) as follows:

$$\alpha_1 = 6\sqrt{3} - 10, \;\; \alpha_2 = 11 - 6\sqrt{3}, \;\; \beta_1 = 4 - 2\sqrt{3}, \;\; \beta_2 = 2 - \sqrt{3}, \;\; \beta_3 = 6 - 3\sqrt{3},$$

$$\gamma_1 = 2/3, \;\; \gamma_2 = 1/3, \;\; \delta_1 = (1 + \sqrt{3})/3, \;\; \sigma = 1 - \sqrt{3}/3.$$

The method is of order 3 and satisfies (3.1) with stepsize-coefficient $c = \sqrt{3} - 1 \approx 0.732$. This result extends an earlier numerical search in Gottlieb, Shu, and Tadmor [10, pp. 102 and 103], where a special method of class E_3 was found which can be implemented, at the cost of an additional function evaluation $\tilde{f}_{n-1}$, such that (3.1) holds with $c \approx 0.473$.

We also did a numerical search in the class G_3 of *general third order* methods (3.8). The best method we could find has a coefficient $\delta_0 = 0$ and it satisfies (3.1) with stepsize-coefficient $c \approx 3.233$, but we did not succeed in finding simple closed-form expressions, similarly as above, for the parameters specifying the method.

We do not go into details here of higher order methods. We just refer to Gottlieb, Shu, and Tadmor [10, p. 103] for an interesting proposition about explicit fourth order methods, and note that fifth order methods with $\delta_0 = 0$ exist satisfying (3.1) with positive c.

3.3. Applications to additive RKMs. Numerical methods of the form

$$\text{(3.9.a)} \;\; y_i = u_{n-1} \;+\; \Delta t \cdot \sum_{j=1}^{s} a_{ij}\, f(y_j) \;+\; \Delta t \cdot \sum_{j=1}^{s} \widehat{a}_{ij}\, \widehat{f}(y_j) \quad (1 \le i \le s+1),$$

$$\text{(3.9.b)} \; u_n = y_{s+1}$$

have been considered for the efficient solution of equations $\frac{d}{dt}U(t) = f(U(t)) + \widehat{f}(U(t))$, where f and $\widehat{f}$ have different stiffness properties; cf., e.g., Ascher, Ruuth, and Spiteri [1] and Kennedy and Carpenter [23]. The methods are known as *additive Runge–Kutta methods*; and also as *implicit-explicit (IMEX) methods* in case the RKM with coefficients a_{ij} is implicit and the one with $\widehat{a}_{ij}$ explicit. Furthermore, methods of the form (3.9) have been studied under the name of *perturbed Runge–Kutta methods*, in the context of solving semidiscrete versions of hyperbolic problems. In that situation, (3.9) is equivalent to a *Shu–Osher implementation* of a standard RKM where some a_{ij} are negative; cf. Higueras [15, 16].

In the last mentioned papers, monotonicity, in the sense of (1.3), was studied for (3.9), under the assumption

$$\text{(3.10)} \qquad \|v + \tau_0 f(v)\| \le \|v\|, \quad \|v + \widehat{\tau}_0 \widehat{f}(v)\| \le \|v\| \qquad (\text{for all } v \in \mathbb{V});$$

a stepsize $(\Delta t)^*$ was presented with the following crucial property:

(3.11) Condition $0 < \Delta t \le (\Delta t)^*$ implies monotonicity of (3.9) whenever $\mathbb{V}$ is a vector space, $\|.\|$ a convex function on $\mathbb{V}$, and functions f, $\widehat{f} : \mathbb{V} \to \mathbb{V}$ satisfy (3.10).

In order to specify $(\Delta t)^*$, we introduce the $(s+1) \times s$ matrices $A = (a_{ij})$, $\widehat{A} = (\widehat{a}_{ij})$, and the $(s+1) \times (s+1)$ matrices $K = [A\ O]$, $\widehat{K} = [\widehat{A}\ O]$. We define $\mathcal{R}$ to be the set of all pairs $(\gamma, \widehat{\gamma}) \in \mathbb{R}^2$ such that

$$I + \gamma K + \widehat{\gamma}\,\widehat{K} \text{ is invertible and } (I + \gamma K + \widehat{\gamma}\,\widehat{K})^{-1}\big[\, E \ \ \gamma K \ \ \widehat{\gamma}\,\widehat{K} \,\big] \geq 0. \tag{3.12}$$

Here E stands for the $(s+1) \times 1$ matrix with all entries equal to 1 and the inequality in (3.12) is to be interpreted entrywise. For a given $\tau_0 > 0$, $\widehat{\tau}_0 > 0$, we put

$$\begin{aligned} &(\Delta t)^* = 0 \quad \text{if there is no pair } (\gamma, \widehat{\gamma}) \text{ in } \mathcal{R} \text{ with } \gamma\,\tau_0 = \widehat{\gamma}\,\widehat{\tau}_0 > 0; \text{ otherwise} \\ &(\Delta t)^* = \sup\{\tau :\ \tau = \gamma\,\tau_0 = \widehat{\gamma}\,\widehat{\tau}_0 > 0 \text{ with } (\gamma, \widehat{\gamma}) \text{ in } \mathcal{R}\}. \end{aligned} \tag{3.13}$$

The following theorem follows immediately from the material in Higueras [15].

THEOREM 3.5 (Higueras, 2006). *Consider method* (3.9) *and let* $\tau_0 > 0$, $\widehat{\tau}_0 > 0$ *be given. Then statement* (3.11) *is valid, with* $(\Delta t)^*$ *defined by* (3.13).

In Higueras [15], sets $\mathcal{R}$ were computed for a series of important additive RKMs. For any given τ_0, $\widehat{\tau}_0$, these sets allow the immediate calculation of $(\Delta t)^*$ defined by (3.13). One may be tempted to view these sets as important characteristics of the underlying methods, and to compare the efficiency of different methods by taking (the magnitude of) the corresponding sets $\mathcal{R}$ into account. However, if (3.11) would also be valid for some $(\Delta t)^*$ which is *greater* than the one given by (3.13), such a use of these sets might be misleading. The natural question arises of whether the value $(\Delta t)^*$, given in the above theorem, is best possible. We think this fundamental question has not yet been answered in the literature.

By applying the theorems of section 2, one can recover the above theorem and essentially answer the question just raised (in the positive); we have the following theorem.

THEOREM 3.6 (upper bound for $(\Delta t)^*$ in (3.11)). *Let* (3.9) *be irreducible, in the sense that the first* s *rows of the* $(s+1) \times 2\,s$ *matrix* $\big[\, A \ \ \widehat{A} \,\big]$ *are different from each other. Let* $\tau_0 > 0$, $\widehat{\tau}_0 > 0$ *be given. If* $(\Delta t)^*$ *is such that statement* (3.11) *holds, then* $(\Delta t)^*$ *cannot exceed the value given in* (3.13).

Proof of Theorems 3.5 *and* 3.6 *using the theory of section* 2. (i) Let $\tau_0 > 0$, $\widehat{\tau}_0 > 0$ be given. We shall relate (3.9) to a numerical process of the form (2.5): we put $l = 1$, $m = 2(s+1)$, and $S = (s_{ij}) = \binom{E}{E}$, $T = (t_{ij}) = \binom{K\ \delta\widehat{K}}{K\ \delta\widehat{K}}$, where $\delta = \tau_0/\widehat{\tau}_0$. We define index sets $I_1 = \{1, \ldots, s\}$, $I_2 = \{s+1\}$, $I_3 = \{s+2, \ldots, 2s+1\}$, and $I_4 = \{2(s+1)\}$.

Let y_i, u_{n-1} satisfy (3.9.a) with f, $\widehat{f}$ as in (3.10). Then $x_1 = u_{n-1}$ and y_i, with

$$y_{s+1+i} = y_i \ \ (\text{for } 1 \leq i \leq s+1), \tag{3.14}$$

can be seen to fulfill (2.5), with some functions f_i satisfying (2.6).

Conversely, let x_i, y_i fulfill (2.5) with f_i satisfying (2.6), (2.15). Then (3.14) holds, so that y_i and $u_{n-1} = x_1$ satisfy (3.9.a) with some f, $\widehat{f}$ as in (3.10).

(ii) In view of the above, it follows that (3.11) holds, with $(\Delta t)^* = c \cdot \tau_0$, as soon as (2.12) is in force for process (2.5). According to Theorem 2.4, we can choose $c = c(S,T)$. Hence (3.11) is valid with $(\Delta t)^* = c(S,T) \cdot \tau_0$. A straightforward calculation shows that $c(S,T) \cdot \tau_0$ is equal to the value $(\Delta t)^*$ defined by (3.13). This proves Theorem 3.5.

(iii) Assume (3.11) holds for some $(\Delta t)^*$. We see now that property (2.16) must be valid for process (2.5), with $c = (\Delta t)^*/\tau_0$. The irreducibility assumption in Theorem

3.6 implies that process (2.5) is irreducible with respect to $I_1, \dots, I_4$, so that Theorem 2.7 can be applied. It follows that the largest value c in (2.16) equals $c(S,T)$, which implies $(\Delta t)^*/\tau_0 \le c(S,T)$. Using again that $c(S,T)\cdot\tau_0$ equals the value defined by (3.13), we arrive at Theorem 3.6. □

Remark 3.7. The set $\mathcal{R}$ has the following interesting property:

$$\text{If } (\gamma, \widehat{\gamma}) \in \mathcal{R}, \text{ then } (\beta, \widehat{\beta}) \in \mathcal{R} \text{ whenever } 0 \le \beta \le \gamma,\ 0 \le \widehat{\beta} \le \widehat{\gamma}. \tag{3.15}$$

This can be proved by defining S, T similarly as in the above proof and applying Theorem 2.2(ii).

For related material, see Higueras [15].

4. Proof of Theorems 2.4 and 2.7.

4.1. Sufficiency of the inequality (2.11). In order to write (2.5) and similar relations more concisely, we introduce some notations relevant to the vector space $\mathbb{V}$. For any integer $n \ge 1$ and vectors $x_1, \dots, x_n \in \mathbb{V}$, we denote the vector in $\mathbb{V}^n$ with components x_i by

$$x = [x_i] = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{V}^n.$$

Furthermore, we denote with a boldface letter the linear operators from $\mathbb{V}^n$ to $\mathbb{V}^m$ determined in a natural way by $m \times n$ matrices: for any matrix $A = (a_{i,j}) \in \mathbb{R}^{m\times n}$ and $x = [x_i] \in \mathbb{V}^n$ we define $\boldsymbol{A}(x) = y$, where $y = [y_i] \in \mathbb{V}^m$ is given by $y_i = \sum_{j=1}^n a_{ij}\, x_j$ $(1 \le i \le m)$.

We combine the vectors x_i and y_i, occurring in (2.5), into the vectors $x = [x_i] \in \mathbb{V}^l$ and $y = [y_i] \in \mathbb{V}^m$, respectively. Furthermore, for given functions $f_i : \mathbb{V} \to \mathbb{V}$ $(1 \le i \le m)$, we define a function F, from $\mathbb{V}^m$ to $\mathbb{V}^m$, by $F(y) = [f_i(y_i)] \in \mathbb{V}^m$ for $y = [y_i] \in \mathbb{V}^m$. With these notations, the relations (2.5) can be written as an equality in $\mathbb{V}^m$:

$$y \;=\; \boldsymbol{S}x \;+\; \Delta t\cdot\; \boldsymbol{T}\, F(y). \tag{4.1}$$

The simple Lemma 4.1 will be quite useful, in the present section for proving that (2.11) implies (2.12) and (2.16), and in the next section for proving that (2.13) and (2.17) imply (2.11). In the lemma we use the notations (2.9) and we relate (4.1), with f_i satisfying (2.6), (2.15), to the conditions

$$y \;=\; \boldsymbol{R}x \;+\; \boldsymbol{P}\, z, \quad \text{with } \|z_i\| \le \|y_i\| \ (1 \le i \le m), \tag{4.2.a}$$

$$z_i = z_j, \quad \text{whenever } y_i = y_j \text{ and } i, j \text{ belong to the same index set } I_q. \tag{4.2.b}$$

LEMMA 4.1 (reformulation of (4.1) with f_i satisfying (2.6), (2.15)). *Let* $\tau_0 > 0$, $\Delta t > 0$, $\gamma = \Delta t/\tau_0$, *and* $I + \gamma T$ *be invertible. Assume* (2.14), *and let* $x = [x_i] \in \mathbb{V}^l$ *and* $y = [y_i] \in \mathbb{V}^m$ *be given. Then* (4.1) *holds for some* $f_i : \mathbb{V} \to \mathbb{V}$ *satisfying* (2.6), (2.15), *if and only if there exists a vector* $z = [z_i] \in \mathbb{V}^m$ *such that* (4.2) *holds.*

Proof of Lemma 4.1. Assume (4.1), (2.6), (2.15), and define $z_i = y_i + \tau_0\, f_i(y_i)$, $z = [z_i] \in \mathbb{V}^m$. Applying (2.6), (2.9), and the equality $y = \mathbf{S}\, x + \gamma\, \mathbf{T}\, (-y + z)$, we arrive at (4.2.a); and by applying (2.15) we obtain (4.2.b).

Conversely, suppose (4.2.a) and (4.2.b) hold. For $i \in I_q$ we define $f_i : \mathbb{V} \to \mathbb{V}$ by

$$f_i(v) = (1/\tau_0)(-y_j + z_j) \ \text{ (if } v = y_j,\ j \in I_q), \ \text{ and } \ f_i(v) = 0 \ \text{ (otherwise) .}$$

Using again (2.9), it follows easily that (4.1) holds with f_i satisfying (2.6), (2.15). ∎

Proof that inequality (2.11) *implies* (2.12) *and* (2.16). Assume $0 < \tau_0 < \infty$, $0 < c \le c(S,T)$. We shall prove (2.12), which is enough because (2.16) follows from (2.12).

Let $\mathbb{V}$, $\|.\|$, f_i be as assumed in (2.12), and suppose x_i, y_i satisfy (2.5) with $0 < \Delta t \le c \cdot \tau_0$. We put $\gamma = \Delta t/\tau_0$ so that $0 < \gamma \le c(S,T)$. Applying Theorem 2.2(ii), it thus follows that γ satisfies (2.8), so that $I + \gamma T$ is invertible.

Since (4.1) holds with f_i satisfying (2.6), we can apply Lemma 4.1, with the trivial index sets $I_q = \{q\}$ for $1 \le q \le m$. It follows that (4.2.a) holds, so that, with the notations (2.9),

$$y_i \;=\; \sum_{j=1}^{l} r_{ij}\, x_j \;+\; \sum_{j=1}^{m} p_{ij}\, z_j, \quad \|z_i\| \le \|y_i\| \quad (1 \le i \le m).$$

In view of (2.8), (2.9), we have $r_{ij} \ge 0$, $p_{ij} \ge 0$; similarly, as in the proof of Lemma 2.3 we have (2.10), i.e., $\sum_j r_{ij} + \sum_j p_{ij} = 1$.

We denote the column vector in $\mathbb{R}^l$ with components $\|x_i\|$ by $[\|x_i\|]$, and we use a similar notation with regard to y_i and z_i. Using the convexity of the function $\|\cdot\|$, it thus follows that $[\|y_i\|] \le R\,[\|x_i\|] + P\,[\|z_i\|] \le Q\,S\,[\|x_i\|] + (I - Q)\,[\|y_i\|]$, i.e., $Q\,[\|y_i\|] \le Q\,S\,[\|x_i\|]$. Multiplying the last inequality by the matrix $Q^{-1} = I + \gamma T$ (which is nonnegative, in view of Theorem 2.2(i)), we get

$$\|y_i\| \;\le\; \sum_{j=1}^{l} s_{ij}\, \|x_j\| \quad (1 \le i \le m). \tag{4.3}$$

Using (2.3) and the nonnegativity of s_{ij} (cf. Theorem 2.2(i)), we obtain (2.7). ∎

4.2. Necessity of the inequality (2.11). In proving that (2.13) and (2.17) imply (2.11), we shall use the following lemma.

LEMMA 4.2 (invertibility of $I + \gamma T$). *Let $\tau_0 > 0$, $\Delta t > 0$ be given and $\gamma = \Delta t/\tau_0$. Assume $\mathbb{V} = \mathbb{R}^m$, $\|\cdot\| = \|\cdot\|_\infty$, and let $I_1, \ldots, I_r$ be index sets as in* (2.14). *Suppose process* (2.5) *is monotonic for all functions f_i satisfying* (2.6), (2.15). *Then $I + \gamma T$ is invertible.*

Proof of Lemma 4.2. Suppose $(I + \gamma T)\,\eta = 0$ for some vector $\eta = [\eta_i] \in \mathbb{R}^m$. Define $f_i(v) = -(1/\tau_0)\,v$ (for all $v \in \mathbb{V}$). Then (2.5) is satisfied by the vectors $x_i = 0$ $(1 \le i \le l)$ and $y_i = \eta_i\, e_m$ $(1 \le i \le m)$, where e_m is the vector in $\mathbb{R}^m$ with all components equal to 1. Since the functions f_i satisfy (2.6), (2.15), it follows that $|\eta_i| = \|y_i\| \le \max_j \|x_j\| = 0$, so that $\eta = 0$. ∎

Proof that (2.13) *implies* (2.11). Let τ_0, c be given with $0 < \tau_0 < \infty$, $0 < c \le \infty$, and assume (2.3), (2.13). We choose $\Delta t = \gamma\,\tau_0$, where γ is an arbitrary finite value with $0 < \gamma \le c$; and we define $\mathbb{V} = \mathbb{R}^m$.

An application of Lemma 4.2, with the trivial index sets $I_q = \{q\}$ (for $1 \le q \le m$), shows that the matrix $I + \gamma T$ is invertible. We thus can use the notations (2.9) and apply Lemma 4.1 (again with the trivial index sets), so as to conclude that, for any $x \in \mathbb{V}^l$ and $y,\, z \in \mathbb{V}^m$, the relations

$$y \;=\; \boldsymbol{R}\,x \;+\; \boldsymbol{P}\,z, \;\; \text{with} \;\; \|z_j\|_\infty \le \|y_j\|_\infty \;(1 \le j \le m), \tag{4.4}$$

imply that

$$\|y_j\|_\infty \;\le\; \max_{1 \le k \le l} \|x_k\|_\infty \quad (\text{for } 1 \le j \le m). \tag{4.5}$$

Below we shall use this implication for proving that

$$\| [\, R \;\; P \,] \|_\infty \le 1. \tag{4.6}$$

By Lemma 2.3, inequality (4.6) implies that $\gamma \le c(S,T)$. Since γ was chosen arbitrarily in $(0,c]$, the last inequality implies (2.11).

In proving (4.6), we shall use the notation $\mathrm{sgn}(\alpha) = 1$ (for $\alpha \ge 0$), $\mathrm{sgn}(\alpha) = -1$ (for $\alpha < 0$). We put $x_{ij} = \mathrm{sgn}(r_{ij})$, $z_{ij} = \mathrm{sgn}(p_{ij})$, where r_{ij}, p_{ij} are the entries of R and P, and we consider the special vectors x_j, $z_j \in \mathbb{V} = \mathbb{R}^m$ with components x_{ij} and z_{ij}, respectively $(1 \le i \le m)$. We define $x \in \mathbb{V}^l$ and y, $z \in \mathbb{V}^m$ by $x = [x_j]$, $z = [z_j]$, $y = [y_j] = \boldsymbol{R}x + \boldsymbol{P}z$, and denote the components of the vectors y_j by y_{ij} $(1 \le i \le m)$.

The relations (4.4) hold, with these special vectors, because

$$\|y_j\|_\infty \ge y_{jj} = \sum_k r_{jk}\, x_{jk} \;+\; \sum_k p_{jk}\, z_{jk} \;=\; \sum_k |r_{jk}| \;+\; \sum_k |p_{jk}|,$$

and, in view of (2.10),

$$\|z_j\|_\infty = 1 \;=\; \sum_k r_{jk} \;+\; \sum_k p_{jk} \;\le\; \sum_k |r_{jk}| \;+\; \sum_k |p_{jk}|.$$

Since (4.4) implies (4.5), we obtain $\sum_k |r_{jk}| \;+\; \sum_k |p_{jk}| \;\le\; 1$, i.e., (4.6). □

Proof that (2.17) *implies* (2.11). (i) Assume (2.3), (2.14) and irreducibility with respect to the index sets under consideration. Let τ_0, c be given with $0 < \tau_0 < \infty$, $0 < c \le \infty$, and assume (2.17). We choose $\Delta t = \gamma\, \tau_0$, where γ is an arbitrary finite value with $0 < \gamma \le c$ and we define $\mathbb{V} = \mathbb{R}^m$.

Similar to the proof above, $I + \gamma\, T$ is invertible, and for any $x \in \mathbb{V}^l$ and $y,\ z \in \mathbb{V}^m$, the implication

$$(4.4) \;\text{ and }\; (4.2.\mathrm{b}) \;\;\Rightarrow\;\; (4.5)$$

is valid. For completing the present proof, it is again enough to deduce (from the last implication) that (4.6) holds.

Below we shall denote by x_j, y_j, z_j the special vectors in $\mathbb{V} = \mathbb{R}^m$, with components x_{ij}, y_{ij}, z_{ij}, used in the previous proof that (2.13) implies (2.11).

(ii) First, assume that $y_i \ne y_j$ whenever indices $i \ne j$ belong to the same index set. Clearly, under this assumption (4.2.b) holds. Furthermore, just as in the previous proof, we have (4.4) so that (4.5) is valid. This again implies (4.6).

(iii) Next, assume the last assumption is violated, i.e., there are indices s, q belonging to the same index set, with $s \ne q$ and $y_s = y_q$. In this situation, we modify (only) the qth component of our special vectors x_j, y_j, z_j into $\tilde{x}_{qj} = \xi_j$, $\tilde{y}_{qj} = \eta_j$, and $\tilde{z}_{qj} = \zeta_j$, respectively. Here $\xi = [\xi_j] \in \mathbb{R}^l$, $\eta = [\eta_j] \in \mathbb{R}^m$, and $\zeta = [\zeta_j] \in \mathbb{R}^m$ are vectors such that

$$\eta \;=\; R\,\xi \;+\; P\,\zeta, \;\; \text{with } \|\xi\|_\infty \le 1,\; \|\zeta\|_\infty \le 1, \tag{4.7.a}$$

$$\eta_i \ne \eta_j \text{ whenever } i \ne j \text{ belong to the same index set.} \tag{4.7.b}$$

We will show that such vectors exist in part (iv) of the proof. In order to distinguish the original vectors x_j, y_j, z_j from the modified ones, we denote the latter by $\tilde{x}_j$, $\tilde{y}_j$, $\tilde{z}_j$, respectively.

Clearly, for $\tilde{x} = [\tilde{x}_j]$, $\tilde{y} = [\tilde{y}_j]$, $\tilde{z} = [\tilde{z}_j]$, the equality $\tilde{y} = \boldsymbol{R}\tilde{x} + \boldsymbol{P}\tilde{z}$ holds. Furthermore, $\tilde{y}_i \neq \tilde{y}_j$, whenever $i \neq j$ belong to the same index set. Finally (using $|y_{ss}| = |y_{sq}| \leq \|\tilde{y}_q\|_\infty$), we see that $\|\tilde{z}_j\|_\infty \leq 1 \leq \|\tilde{y}_j\|_\infty$ $(1 \leq j \leq m)$. The modified vectors thus satisfy (4.4), (4.2.b). Consequently, they satisfy (4.5), which implies $\sum_k |r_{jk}| + \sum_k |p_{jk}| \leq 1$ (for all $j \neq q$). By interchanging the role of s and q, we see that the last inequality is also valid for all $j \neq s$. Hence, (4.6) holds.

(iv) In view of the irreducibility assumption, the polynomials $f_i(\lambda) = \sum_{k=1}^{l} s_{ik}\,\lambda^k + \gamma \cdot \sum_{k=1}^{m} t_{ik}\,\lambda^{l+k}$ satisfy $f_i \neq f_j$, if $i \neq j$ belong to the same index set. It follows that, for sufficiently small $\lambda > 0$, the vectors $\xi = [\xi_j]$, $\eta = [\eta_j]$, $\zeta = [\zeta_j]$, with $\xi_k = \lambda^k$ $(1 \leq k \leq l)$, $\eta_k = f_k(\lambda)$ $(1 \leq k \leq m)$, $\zeta_k = \lambda^{l+k} + f_k(\lambda)$ $(1 \leq k \leq m)$, satisfy (4.7). □

Acknowledgments. The author thanks Dr. Karel in 't Hout and Dr. Jaap van de Griend for helpful discussions related to the MATLAB calculations in section 3.2.3.

REFERENCES

[1] U. M. Ascher, S. J. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[2] K. Burrage and J. C. Butcher, *Nonlinear stability of a general class of differential equation methods*, BIT, 20 (1980), pp. 185–203.

[3] J. C. Butcher, *On the convergence of numerical solutions to ordinary differential equations*, Math. Comp., 20 (1966), pp. 1–10.

[4] J. C. Butcher, *The Numerical Analysis of Ordinary Differential Equations*, John Wiley, Chichester, UK, 1987.

[5] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, John Wiley, Chichester, UK, 2003.

[6] L. Ferracina and M. N. Spijker, *Stepsize restrictions for the total-variation-diminishing property in general Runge–Kutta methods*, SIAM J. Numer. Anal., 42 (2004), pp. 1073–1093.

[7] L. Ferracina and M. N. Spijker, *An extension and analysis of the Shu-Osher representation of Runge-Kutta methods*, Math. Comp., 74 (2005), pp. 201–219.

[8] S. Gottlieb, *On high order strong stability preserving Runge-Kutta and multi step time discretizations*, J. Sci. Comput., 25 (2005), pp. 105–128.

[9] S. Gottlieb and C.-W. Shu, *Total variation diminishing Runge-Kutta schemes*, Math. Comp., 67 (1998), pp. 73–85.

[10] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[11] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations.* II. *Stiff and Differential-Algebraic Problems*, 2nd ed., Springer-Verlag, Berlin, 1996.

[12] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations.* I. *Nonstiff Problems*, Springer-Verlag, Berlin, 1987.

[13] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[14] I. Higueras, *On strong stability preserving time discretization methods*, J. Sci. Comput., 21 (2004), pp. 193–223.

[15] I. Higueras, *Strong stability for additive Runge–Kutta methods*, SIAM J. Numer. Anal., 44 (2006), pp. 1735–1758.

[16] I. Higueras, *Representations of Runge–Kutta methods and strong stability preserving methods*, SIAM J. Numer. Anal., 43 (2005), pp. 924–948.

[17] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, UK, 1988.

[18] Z. Horváth, *Positivity of Runge-Kutta and diagonally split Runge-Kutta methods*, Appl. Numer. Math., 28 (1998), pp. 309–326.

[19] W. H. Hundsdorfer and S. J. Ruuth, *Monotonicity for Time Discretizations*, Dundee Conference Report NA/217 2003, D. F. Griffiths and G. A. Watson, eds., University of Dundee, Dundee, UK, 2003, pp. 85–94.

[20] W. H. Hundsdorfer, S. J. Ruuth, and R. J. Spiteri, *Monotonicity-preserving linear multistep methods*, SIAM J. Numer. Anal., 41 (2003), pp. 605–623.
[21] W. H. Hundsdorfer and J. G. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin, 2003.
[22] Z. Jackiewicz and S. Tracogna, *A general class of two-step Runge–Kutta methods for ordinary differential equations*, SIAM J. Numer. Anal., 32 (1995), pp. 1390–1427.
[23] C. A. Kennedy and M. H. Carpenter, *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math., 44 (2003), pp. 139–181.
[24] J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT, 31 (1991), pp. 482–528.
[25] H. W. J. Lenferink, *Contractivity-preserving implicit linear multistep methods*, Math. Comp., 56 (1991), pp. 177–199.
[26] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.
[27] S. J. Ruuth, *Global optimization of explicit strong-stability-preserving Runge-Kutta methods*, Math. Comp., 75 (2006), pp. 183–207.
[28] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Statist. Comput., 9 (1988), pp. 1073–1084.
[29] C.-W. Shu, *A survey of strong stability preserving high order time discretizations*, in Collected Lectures on the Preservation of Stability under Discretization, D. Estep and S. Tavener, eds., SIAM, Philadelphia, 2002, pp. 51–65.
[30] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[31] R. J. Spiteri and S. J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1246–1268

ROBIN–ROBIN DOMAIN DECOMPOSITION METHODS FOR THE STOKES–DARCY COUPLING*

MARCO DISCACCIATI†, ALFIO QUARTERONI‡, AND ALBERTO VALLI§

Abstract. In this paper we consider a coupled system made of the Stokes and Darcy equations, and we propose some iteration-by-subdomain methods based on Robin conditions on the interface. We prove the convergence of these algorithms, and for suitable finite element approximations we show that the rate of convergence is independent of the mesh size h. Special attention is paid to the optimization of the performance of the methods when both the kinematic viscosity ν of the fluid and the hydraulic conductivity tensor K of the porous medium are very small.

Key words. Stokes and Darcy equations, domain decomposition, Robin interface condition, finite element approximation

AMS subject classifications. 65N55, 65N30, 35M20, 35Q35

DOI. 10.1137/06065091X

1. Introduction and problem setting.

Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain, decomposed in two nonintersecting subdomains Ω_f and Ω_p separated by an interface Γ, i.e., $\overline{\Omega} = \overline{\Omega_f} \cup \overline{\Omega_p}$, $\Omega_f \cap \Omega_p = \emptyset$, and $\overline{\Omega_f} \cap \overline{\Omega_p} = \Gamma$.

We are interested in the case in which Γ is a surface separating an upper domain Ω_f filled by a fluid, from a lower domain Ω_p formed by a porous medium. We assume that the fluid contained in Ω_f has an upper fixed surface (i.e., we do not consider the free surface fluid case) and can filtrate through the porous medium beneath.

The motion of the fluid in Ω_f is modeled by the Stokes equations:

$$-\boldsymbol{\nabla} \cdot \mathsf{T}(\mathbf{u}_f, p_f) = \mathbf{f}, \quad \nabla \cdot \mathbf{u}_f = 0 \qquad \text{in } \Omega_f, \tag{1}$$

where $\mathsf{T}(\mathbf{u}_f, p_f) = 2\,\nu\,\mathsf{D}(\mathbf{u}_f) - p_f \mathsf{I}$ is the stress tensor, and $\mathsf{D}(\mathbf{u}_f) = \frac{1}{2}(\nabla \mathbf{u}_f + \nabla^T \mathbf{u}_f)$ is the deformation tensor; as usual, ∇ and $\nabla\cdot$ denote the gradient operator and the divergence operator, respectively, with respect to the space coordinates. The parameter $\nu > 0$ is the kinematic viscosity of the fluid, while $\mathbf{u}_f$ and p_f denote the fluid velocity and pressure, respectively. We suppose ν to be constant in the whole domain Ω_f.

In the lower domain Ω_p we define the piezometric head $\varphi = z + p_p/(\rho g)$, where z is the elevation from a reference level, p_p the pressure of the fluid in Ω_p, $\rho > 0$ the density of the fluid (assumed to be constant in the whole domain Ω), and $g > 0$ the gravity acceleration.

The flow in Ω_p is modeled by the equations:

$$\mathbf{u}_p = -\frac{K}{n} \nabla \varphi, \quad \nabla \cdot \mathbf{u}_p = 0 \qquad \text{in } \Omega_p, \tag{2}$$

*Received by the editors January 26, 2006; accepted for publication January 17, 2007; published electronically May 22, 2007.
http://www.siam.org/journals/sinum/45-3/65091.html
†J. Radon Institute for Computational and Applied Mathematics (RICAM), Altengerberstraße 69, A-4040 Linz, Austria (marco.discacciati@oeaw.ac.at).
‡CMCS-IACS, EPFL, CH-1015 Lausanne, Switzerland, and MOX, Politecnico di Milano, P.zza Leonardo da Vinci 32, I-20133 Milano, Italy (alfio.quarteroni@epfl.ch).
§Dipartimento di Matematica, Università di Trento, Via Sommarive 14, I-38050 Povo (Trento), Italy (valli@science.unitn.it).

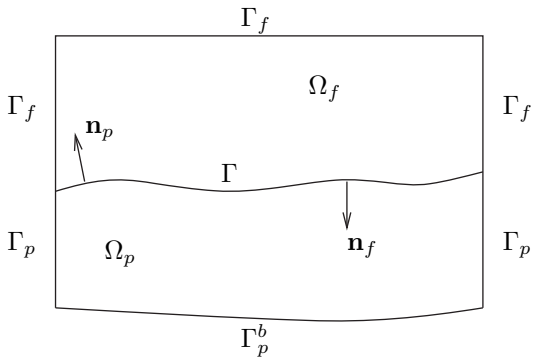


FIG. 1. *Schematic representation of a* 2D *vertical section of the computational domain.*

where $\mathbf{u}_p$ is the fluid velocity, and $n > 0$ is the volumetric porosity. The tensor K is the hydraulic conductivity $K = \mathrm{diag}\,(K_1, \ldots, K_d)$, and we suppose that $K_i \in L^\infty(\Omega_p)$ and $\inf_{\Omega_p} K_i > 0$, $i = 1, \ldots, d$. In the following we shall denote $\mathsf{K} = K/n = \mathrm{diag}\,(K_i/n)$ $(i = 1, \ldots, d)$. The first equation in (2) is Darcy's law.

For the sake of simplicity, we adopt homogenous boundary conditions. We impose the no-slip condition $\mathbf{u}_f = \mathbf{0}$ on $\Gamma_f = \partial\Omega_f \setminus \Gamma$ for the Stokes problem (1), while, for the Darcy problem (2), we set the piezometric head $\varphi = 0$ on the lateral surface Γ_p, and we require a slip condition on Γ_p^b: $\mathbf{u}_p \cdot \mathbf{n}_p = 0$ on Γ_p, where $\partial\Omega_p = \Gamma \cup \Gamma_p^b \cup \Gamma_p$ (see Figure 1). The vectors $\mathbf{n}_p$ and $\mathbf{n}_f$ denote the unit outward normal vectors to the surfaces $\partial\Omega_p$ and $\partial\Omega_f$, respectively; in particular, we have $\mathbf{n}_f = -\mathbf{n}_p$ on Γ. In the following we shall indicate $\mathbf{n} = \mathbf{n}_f$ for simplicity of notation. We also assume that the boundary $\partial\Omega$ and the interface Γ are piecewise smooth manifolds.

Other boundary conditions (see, e.g., [6, 7, 13, 10, 11]) could also be considered, and all of the results in this paper would remain true without essential changes in the proofs.

We supplement the Stokes and Darcy problems with the following matching conditions on Γ (see [12]):

$$\mathbf{u}_p \cdot \mathbf{n} = \mathbf{u}_f \cdot \mathbf{n}, \tag{3}$$

$$-\varepsilon \boldsymbol{\tau}_j \cdot (\mathsf{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}) = \nu \mathbf{u}_f \cdot \boldsymbol{\tau}_j, \qquad j = 1, \ldots, d-1, \tag{4}$$

$$-\mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}) = g\varphi_{|\Gamma}, \tag{5}$$

where $\boldsymbol{\tau}_j$ $(j = 1, \ldots, d-1)$ are linear independent unit tangential vectors to the interface Γ, and ε represents the characteristic length of the pores of the porous medium.

Conditions (3)–(5) impose the continuity of the normal velocity on Γ, as well as that of the normal component of the normal stress, but they allow the pressure to be discontinuous across the interface.

This problem has been studied in several works. In [8, 6, 7] the mathematical and numerical analysis of the coupled problem was carried out, in the case in which the Darcy equation is replaced by a scalar elliptic problem for the sole piezometric head φ. The analysis of the coupled problem in its original form (1)–(2) has been considered in [13, 10], and the recent works [18, 11] address the analysis and preconditioning of mortar discretizations of the Stokes–Darcy problem.

A domain decomposition method of the Dirichlet–Neumann type based on the choice of the fluid normal velocity across Γ as an interface variable was proposed and analyzed in [6, 7]. A similar approach, using the trace of φ on Γ as an interface

variable, has been studied in [8]. After proving that this method is equivalent to a preconditioned Richardson algorithm for the Steklov–Poincaré interface equation associated to the Stokes–Darcy problem, it was proved that the convergence rate of the algorithm is independent of the mesh parameter h, for suitable conforming finite element approximations of the coupled problem. An extension to the time-dependent case has been presented in [9].

The previous results indicate that, in the steady case, preconditioners of the Dirichlet–Neumann type may be sensitive to the variation of the viscosity ν and of the entries of the hydraulic conductivity K, downgrading the convergence rate of the algorithm.

In this work we extend some preliminary results contained in [8], by presenting improved domain decomposition methods based on Robin interface conditions. The aim is twofold: first, to propose an algorithm whose rate of convergence does not deteriorate as ν and the entries of K become smaller and smaller, and second, to devise an algorithm that is more "symmetric" with respect to the treatment of either Ω_f and Ω_p, namely, being based on solvers that treat simultaneously (i.e., in parallel) the two subdomains.

After having presented in section 2 the weak formulation of the coupled problem, in section 3 we introduce two methods, based on a multiplicative and on an additive paradigm, respectively. Then, in section 4 the convergence analysis of the algorithms is developed. Finally, some numerical results are presented in section 5.

The first algorithm has optimal convergence properties with respect to ν and K. On the other hand, the second algorithm, which indeed for small values of ν and K does not outperform the Dirichlet–Neumann scheme, is interesting for its parallel nature. Moreover, its convergence analysis is rather simple and is based on the fact that the so-called Robin-to-Dirichlet and Robin-to-Neumann maps are symmetric and positive, uniformly with respect to the mesh size h. These important properties seem to be yet overlooked in the literature and could be revealed to be very useful also in different contexts.

2. Weak form of the coupled problem. From now on, instead of (2), we will take the following scalar formulation of the Darcy problem:

$$-\nabla \cdot (\mathsf{K}\nabla\varphi) = 0 \quad \text{in } \Omega_p. \tag{6}$$

Accordingly, (3) becomes

$$-\mathsf{K}\nabla\varphi \cdot \mathbf{n} = \mathbf{u}_f \cdot \mathbf{n} \quad \text{on } \Gamma. \tag{7}$$

We define the following functional spaces:

$$H_f = \{\mathbf{v} \in (H^1(\Omega_f))^d |\ \mathbf{v} = \mathbf{0} \text{ on } \Gamma_f\}, \quad Q = L^2(\Omega_f), \tag{8}$$

$$H_p = \{\psi \in H^1(\Omega_p) |\ \psi = 0 \text{ on } \Gamma_p^b\} \tag{9}$$

and the bilinear forms

$$a_f(\mathbf{v}, \mathbf{w}) = 2\nu \int_{\Omega_f} \mathsf{D}(\mathbf{v}) : \mathsf{D}(\mathbf{w}) \qquad \forall \mathbf{v}, \mathbf{w} \in (H^1(\Omega_f))^d, \tag{10}$$

$$b_f(\mathbf{v}, q) = -\int_{\Omega_f} q\, \nabla \cdot \mathbf{v} \qquad \forall \mathbf{v} \in (H^1(\Omega_f))^d, \quad \forall q \in Q, \tag{11}$$

$$a_p(\varphi, \psi) = \int_{\Omega_p} \nabla\psi \cdot \mathsf{K}\nabla\varphi \qquad \forall \varphi, \psi \in H^1(\Omega_p). \tag{12}$$

The coupling conditions (4), (5), and (7) can be incorporated in the weak formulation of the global problem as natural conditions on Γ. In particular, we can write the following weak saddle-point formulation of the coupled Stokes–Darcy problem:

Find $(\mathbf{u}_f, p_f) \in H_f \times Q$, $\varphi \in H_p$ such that

$$a_f(\mathbf{u}_f, \mathbf{v}) + b_f(\mathbf{v}, p_f) + g\, a_p(\varphi, \psi) + \int_\Gamma g\, \varphi(\mathbf{v}\cdot\mathbf{n}) - \int_\Gamma g\, \psi(\mathbf{u}_f\cdot\mathbf{n}) + \int_\Gamma \sum_{j=1}^{d-1} \frac{\nu}{\varepsilon}(\mathbf{u}_f\cdot\boldsymbol{\tau}_j)(\mathbf{v}\cdot\boldsymbol{\tau}_j) = \int_{\Omega_f} \mathbf{f}\cdot\mathbf{v} \qquad \forall \mathbf{v} \in H_f,\ \psi \in H_p, \tag{13}$$

$$b_f(\mathbf{u}_f, q) = 0 \qquad \forall q \in Q. \tag{14}$$

Using Brezzi's theory of saddle-point problems [2], we can guarantee that the coupled problem (13)–(14) has a unique solution (see [8, 6, 13]).

In the rest of the paper, instead of (4) we shall adopt the following simplified condition on the interface:

$$\mathbf{u}_f\cdot\boldsymbol{\tau}_j = 0 \quad \text{on } \Gamma \qquad (j = 1, \ldots, d-1), \tag{15}$$

and, consequently, we will use the functional space:

$$H_f^\tau = \{\mathbf{v} \in H_f |\ \mathbf{v}\cdot\boldsymbol{\tau}_j = 0 \text{ on } \Gamma,\ j = 1, \ldots, d-1\}. \tag{16}$$

This simplification is acceptable from the physical viewpoint, since the term in (4) involving the normal derivative of $\mathbf{u}_f$ is multiplied by ε and the velocity itself can be supposed at least of order $O(\varepsilon)$ in the neighborhood of Γ, so that the left-hand side can be approximated by zero. We point out that this simplification does not dramatically influence the coupling of the two subproblems, since (4) is not strictly speaking a coupling condition but only a boundary condition for the fluid problem in Ω_f. In any case, all of the results in the paper are still true for the more general interface condition (4), provided H_f^τ is replaced by H_f and the bilinear form $a_f(\mathbf{w}, \mathbf{v})$ by $a_f(\mathbf{w}, \mathbf{v}) + \int_\Gamma \sum_{j=1}^{d-1} \frac{\nu}{\varepsilon}(\mathbf{w}\cdot\boldsymbol{\tau}_j)(\mathbf{v}\cdot\boldsymbol{\tau}_j)$.

Remark 2.1. In [6, 7] we considered another simplified form of (4), i.e., $\boldsymbol{\tau}_j \cdot (\mathsf{T}(\mathbf{u}_f, p_f)\cdot\mathbf{n}) = 0$ on Γ. Although not completely precise from the physical point of view, this simplified condition is perfectly acceptable from the mathematical viewpoint for the setup and analysis of solution methods for the coupled problem.

3. Iterative domain decomposition methods for solving the coupled problem. In this section we propose new iterative methods to compute the solution of the coupled problem which exploit the decoupled structure of the problem, thus requiring one at each step to solve independently the fluid and the groundwater subproblems, i.e., using as building blocks a Stokes solver and an elliptic solver.

As we have already remarked, the numerical performances of the domain decomposition methods of the Dirichlet–Neumann type presented in [6, 7] strongly depend on the fluid viscosity ν and on the entries of the hydraulic conductivity K. More precisely, the convergence rate of the algorithm deteriorates as ν and the entries of K decrease. The following numerical example illustrates the situation.

Example 3.1. We consider the computational domain $\Omega \subset \mathbb{R}^2$, with $\Omega_f = (0,1)\times(1,2)$, $\Omega_p = (0,1)\times(0,1)$, and $\Gamma = (0,1)\times\{1\}$, and choose the parameter $g = 1$; moreover, we assume that the hydraulic conductivity tensor K is a multiple of the identity tensor, namely, a scalar function. Boundary conditions and the right-hand

TABLE 1
Iterations using PCG with the Dirichlet–Neumann preconditioner with respect to several values of ν and K and of the grid parameter h ($h_1 \approx 0.14$ and $h_i = h_1/2^{i-1}$, $i = 2, 3, 4$).

ν	K	h_1	h_2	h_3	h_4
1	1	5	5	5	5
10^{-1}	10^{-1}	10	10	8	8
10^{-2}	10^{-1}	13	15	14	14
10^{-3}	10^{-2}	19	49	60	55
10^{-4}	10^{-3}	20	58	143	167
10^{-6}	10^{-4}	20	56	138	202

side $\mathbf{f}$ are chosen in such a way that the exact solution of the coupled Stokes–Darcy problem is $\mathbf{u}_f = (y^2 - 2y + 1, x^2 - x)^T$, $p_f = 2\nu(x + y - 1) + 1/(3\mathsf{K})$, $\varphi = (x(1 - x)(y-1) + y^3/3 - y^2 + y)/\mathsf{K} + 2x\nu$, with ν and K constant in Ω_f and Ω_p, respectively. Table 1 reports the number of iterations obtained for several choices of ν and K and four different grid sizes, using the preconditioned conjugate gradient (PCG) method on the interface equation, with the preconditioner characterized by the Dirichlet–Neumann method. A tolerance of 10^{-9} has been imposed on the relative increment. Taylor–Hood finite elements have been used to approximate the Stokes problem and quadratic Lagrangian elements for the Darcy equation (6).

Such small values of ν and K are quite realistic for real-life physical flows. This fact motivates our interest to set up new algorithms that are more robust to parameter variations.

3.1. Iterative methods based on Robin interface conditions. We present two possible domain decomposition methods based on the adoption of Robin interface conditions, i.e., proper linear combinations of the coupling conditions (5) and (7).

3.1.1. A sequential Robin–Robin method. We consider a sequential Robin–Robin (sRR) method, which at each iteration requires one to solve a Darcy problem in Ω_p followed by a Stokes problem in Ω_f, both with Robin conditions on Γ. Precisely, the algorithm reads as follows.

Having assigned a trace function $\eta^0 \in L^2(\Gamma)$ and two acceleration parameters $\gamma_f \geq 0$ and $\gamma_p > 0$, for each $k \geq 0$:

(i) find $\varphi^{k+1} \in H_p$ such that

$$\gamma_p a_p(\varphi^{k+1}, \psi) + \int_\Gamma g\varphi^{k+1}_{|\Gamma}\psi_{|\Gamma} = \int_\Gamma \eta^k \psi_{|\Gamma} \qquad \forall\ \psi \in H_p. \tag{17}$$

This corresponds to imposing the following interface condition (in weak, or natural, form) for the Darcy problem:

$$-\gamma_p \mathsf{K}\nabla\varphi^{k+1} \cdot \mathbf{n} + g\varphi^{k+1}_{|\Gamma} = \eta^k \qquad \text{on } \Gamma. \tag{18}$$

(ii) Then find $(\mathbf{u}_f^{k+1}, p_f^{k+1}) \in H_f^\tau \times Q$ such that

$$\begin{aligned} &a_f(\mathbf{u}_f^{k+1}, \mathbf{v}) + b_f(\mathbf{v}, p_f^{k+1}) + \gamma_f \int_\Gamma (\mathbf{u}_f^{k+1} \cdot \mathbf{n})(\mathbf{v} \cdot \mathbf{n}) \\ &\quad = \int_\Gamma \Big(\frac{\gamma_f}{\gamma_p}\eta^k - \frac{\gamma_f + \gamma_p}{\gamma_p} g\varphi^{k+1}_{|\Gamma}\Big)(\mathbf{v} \cdot \mathbf{n}) + \int_{\Omega_f} \mathbf{f} \cdot \mathbf{v} \quad \forall\ \mathbf{v} \in H_f^\tau, \\ &b_f(\mathbf{u}_f^{k+1}, q) = 0 \qquad \forall\ q \in Q. \end{aligned} \tag{19}$$

This corresponds to imposing on the Stokes problem the following matching conditions on Γ (still in natural form):

$$\begin{aligned}
&\mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_f^{k+1},p_f^{k+1})\cdot\mathbf{n})+\gamma_f\mathbf{u}_f^{k+1}\cdot\mathbf{n}=\frac{\gamma_f}{\gamma_p}\eta^k-\frac{\gamma_f+\gamma_p}{\gamma_p}g\varphi_{|\Gamma}^{k+1}\\
&\qquad=-g\varphi_{|\Gamma}^{k+1}-\gamma_f\mathsf{K}\nabla\varphi^{k+1}\cdot\mathbf{n},\\
&\mathbf{u}_f^{k+1}\cdot\boldsymbol{\tau}_j=0,\qquad j=1,\dots,d-1.
\end{aligned}\tag{20}$$

(iii) Finally, set

$$\begin{aligned}
\eta^{k+1}&=-\mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_f^{k+1},p_f^{k+1})\cdot\mathbf{n})+\gamma_p\mathbf{u}_f^{k+1}\cdot\mathbf{n}\\
&=(\gamma_f+\gamma_p)(\mathbf{u}_f^{k+1}\cdot\mathbf{n})+\frac{\gamma_f+\gamma_p}{\gamma_p}g\varphi_{|\Gamma}^{k+1}-\frac{\gamma_f}{\gamma_p}\eta^k\in L^2(\Gamma).
\end{aligned}\tag{21}$$

Concerning the solvability of problem (19), we note first that using the trace theorem and the Korn inequality (see, e.g., [3, p. 416]), there exist two constants $\kappa_1,\kappa_2>0$ such that

$$\int_\Gamma|\mathbf{u}_f\cdot\mathbf{n}|^2\le\kappa_1\left(\int_{\Omega_f}(|\mathbf{u}_f|^2+|\nabla\mathbf{u}_f|^2)\right)\le\kappa_2\int_{\Omega_f}|\mathsf{D}(\mathbf{u}_f)|^2.\tag{22}$$

Therefore, the bilinear form

$$a_f(\mathbf{u}_f,\mathbf{v})+\gamma_f\int_\Gamma(\mathbf{u}_f\cdot\mathbf{n})(\mathbf{v}\cdot\mathbf{n})$$

is continuous and coercive in $H_f^\tau\times H_f^\tau$. Moreover, the bilinear form $b_f(\mathbf{v},p)$ satisfies an inf–sup condition on the space $H_f^\tau\times Q$ (see, e.g., [17, pp. 157–158]). Then, for every $\mathbf{f}\in(L^2(\Omega_f))^d$, $\eta^k\in L^2(\Gamma)$, and $\varphi_{|\Gamma}^{k+1}\in L^2(\Gamma)$, there exists a unique solution of problem (19).

If the sRR method converges, in the limit we recover the solution $(\mathbf{u}_f,p_f)\in H_f^\tau\times Q$ and $\varphi\in H_p$ of the coupled Stokes–Darcy problem. Indeed, denoting by φ^* the limit of the sequence φ^k in $H^1(\Omega_p)$ and by $(\mathbf{u}_f^*,p_f^*)$ that of $(\mathbf{u}_f^k,p_f^k)$ in $(H^1(\Omega_f))^d\times Q$, we obtain

$$-\gamma_p\mathsf{K}\nabla\varphi^*\cdot\mathbf{n}+g\varphi_{|\Gamma}^*=-\mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_f^*,p_f^*)\cdot\mathbf{n})+\gamma_p\mathbf{u}_f^*\cdot\mathbf{n}\quad\text{on }\Gamma,\tag{23}$$

so that, as a consequence of (20), we have

$$(\gamma_f+\gamma_p)\mathbf{u}_f^*\cdot\mathbf{n}=-(\gamma_f+\gamma_p)\mathsf{K}\nabla\varphi^*\cdot\mathbf{n}\quad\text{on }\Gamma,$$

yielding, since $\gamma_f+\gamma_p\neq0$, $\mathbf{u}_f^*\cdot\mathbf{n}=-\mathsf{K}\nabla\varphi^*\cdot\mathbf{n}$ on Γ and also, from (23), that $\mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_f^*,p_f^*)\cdot\mathbf{n})=-g\varphi_{|\Gamma}^*$ on Γ. Thus, the two interface conditions (5) and (7) are satisfied, and we can conclude that the limit functions $\varphi^*\in H_p$ and $(\mathbf{u}_f^*,p_f^*)\in H_f^\tau\times Q$ are the solutions of the coupled Stokes–Darcy problem.

The proof of convergence will be given in section 4.1.

3.1.2. A parallel Robin–Robin method. We consider now a parallel Robin–Robin (pRR) algorithm. The idea behind this new method resembles that for a Neumann–Neumann scheme. However, the latter cannot be considered straightforwardly in our case, since we would not be able to guarantee the correct regularity of the data for each subproblem, as we shall point out more precisely in Remark 3.1.

The pRR algorithm that we propose reads as follows: Let $\mu^k \in L^2(\Gamma)$ be an assigned trace function on Γ, and let γ_1, γ_2 be two positive parameters; then, for $k \geq 0$,

(i) find $(\mathbf{u}_f^{k+1}, p_f^{k+1}) \in H_f^\tau \times Q$ such that

$$\begin{aligned} &a_f(\mathbf{u}_f^{k+1}, \mathbf{v}) + b_f(\mathbf{v}, p_f^{k+1}) - \gamma_1 \int_\Gamma (\mathbf{u}_f^{k+1} \cdot \mathbf{n})(\mathbf{v} \cdot \mathbf{n}) \\ &\qquad = \int_\Gamma \mu^k (\mathbf{v} \cdot \mathbf{n}) + \int_{\Omega_f} \mathbf{f} \cdot \mathbf{v} \qquad \forall\, \mathbf{v} \in H_f^\tau, \\ &b_f(\mathbf{u}_f^{k+1}, q) = 0 \qquad \forall\, q \in Q, \end{aligned} \tag{24}$$

and, at the same time, find $\varphi^{k+1} \in H_p$ such that

$$a_p(\varphi^{k+1}, \psi) + \frac{1}{\gamma_1} \int_\Gamma g \varphi_{|\Gamma}^{k+1} \psi_{|\Gamma} = -\frac{1}{\gamma_1} \int_\Gamma \mu^k \psi_{|\Gamma} \qquad \forall\, \psi \in H_p. \tag{25}$$

Remark that on the interface Γ we are imposing the matching conditions

$$\begin{aligned} &\mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f^{k+1}, p^{k+1}) \cdot \mathbf{n}) - \gamma_1 \mathbf{u}_f^{k+1} \cdot \mathbf{n} = \mu^k \\ &\qquad = -g \varphi_{|\Gamma}^{k+1} + \gamma_1 \mathsf{K} \nabla \varphi^{k+1} \cdot \mathbf{n}, \\ &\mathbf{u}_f^{k+1} \cdot \boldsymbol{\tau}_j = 0, \quad j = 1, \ldots, d-1. \end{aligned} \tag{26}$$

(ii) As a second step, find $(\widehat{\boldsymbol{\omega}}^{k+1}, \widehat{\pi}^{k+1}) \in H_f^\tau \times Q$ such that

$$\begin{aligned} &a_f(\widehat{\boldsymbol{\omega}}^{k+1}, \mathbf{v}) + b_f(\mathbf{v}, \widehat{\pi}^{k+1}) + \gamma_2 \int_\Gamma (\widehat{\boldsymbol{\omega}}^{k+1} \cdot \mathbf{n})(\mathbf{v} \cdot \mathbf{n}) \\ &\qquad = \gamma_2 \int_\Gamma \widehat{\sigma}^{k+1} (\mathbf{v} \cdot \mathbf{n}) \qquad \forall\, \mathbf{v} \in H_f^\tau, \\ &b_f(\widehat{\boldsymbol{\omega}}^{k+1}, q) = 0 \qquad \forall\, q \in Q, \end{aligned} \tag{27}$$

and find $\widehat{\chi}^{k+1} \in H_p$ such that

$$a_p(\widehat{\chi}^{k+1}, \psi) + \frac{1}{\gamma_2} \int_\Gamma g \widehat{\chi}_{|\Gamma}^{k+1} \psi_{|\Gamma} = \int_\Gamma \widehat{\sigma}^{k+1} \psi_{|\Gamma} \qquad \forall\, \psi \in H_p, \tag{28}$$

where

$$\widehat{\sigma}^{k+1} = \mathbf{u}_f^{k+1} \cdot \mathbf{n} + \mathsf{K} \nabla \varphi^{k+1} \cdot \mathbf{n} = \mathbf{u}_f^{k+1} \cdot \mathbf{n} + \frac{1}{\gamma_1} (g \varphi_{|\Gamma}^{k+1} + \mu^k) \in L^2(\Gamma). \tag{29}$$

Note that on the interface Γ we are now imposing the matching conditions

$$\begin{aligned} &\mathbf{n} \cdot (\mathsf{T}(\widehat{\boldsymbol{\omega}}^{k+1}, \widehat{\pi}^{k+1}) \cdot \mathbf{n}) + \gamma_2 \widehat{\boldsymbol{\omega}}^{k+1} \cdot \mathbf{n} = \gamma_2 \widehat{\sigma}^{k+1} \\ &\qquad = g \widehat{\chi}_{|\Gamma}^{k+1} - \gamma_2 \mathsf{K} \nabla \widehat{\chi}^{k+1} \cdot \mathbf{n}, \\ &\widehat{\boldsymbol{\omega}}^{k+1} \cdot \boldsymbol{\tau}_j = 0, \quad j = 1, \ldots, d-1. \end{aligned} \tag{30}$$

(iii) Finally, set

$$\begin{aligned} \mu^{k+1} &= \mu^k - \theta [\mathbf{n} \cdot (\mathsf{T}(\widehat{\boldsymbol{\omega}}^{k+1}, \widehat{\pi}^{k+1}) \cdot \mathbf{n}) + g \widehat{\chi}_{|\Gamma}^{k+1}] \\ &= \mu^k - \theta [\gamma_2 (\widehat{\sigma}^{k+1} - \widehat{\boldsymbol{\omega}}^{k+1} \cdot \mathbf{n}) + g \widehat{\chi}_{|\Gamma}^{k+1}] \in L^2(\Gamma), \end{aligned} \tag{31}$$

where $\theta > 0$ is a further acceleration parameter.

Before moving to the convergence analysis of the pRR method (24)–(31), a few remarks are in order.

Concerning the well-posedness of problem (24), since the inf–sup condition is satisfied (see [17, pp. 157–158]), and thanks to (22), the bilinear form

$$a_f(\mathbf{u}_f, \mathbf{v}) - \gamma_1 \int_\Gamma (\mathbf{u}_f \cdot \mathbf{n})(\mathbf{v} \cdot \mathbf{n})$$

is coercive in $H_f^\tau \times H_f^\tau$ provided

$$\gamma_1 < \frac{2\nu}{\kappa_2}. \tag{32}$$

As regards the consistency of the algorithm, note that if we find a fixed point μ^*, from (31) we have (again denoting the limit functions by an upper $*$)

$$\gamma_2(\widehat{\boldsymbol{\omega}}^* \cdot \mathbf{n} - \widehat{\sigma}^*) = g\widehat{\chi}^*_{|\Gamma} \qquad \text{on } \Gamma \tag{33}$$

and also, equivalently,

$$\frac{1}{\gamma_2} g\widehat{\chi}^*_{|\Gamma} - \widehat{\sigma}^* = \frac{2}{\gamma_2} g\widehat{\chi}^*_{|\Gamma} - \widehat{\boldsymbol{\omega}}^* \cdot \mathbf{n} \qquad \text{on } \Gamma. \tag{34}$$

Therefore, if we multiply (28) by g, sum the resulting equation to (27), and use relations (33) and (34), we obtain

$$\begin{aligned} a_f(\widehat{\boldsymbol{\omega}}^*, \mathbf{v}) + b_f(\mathbf{v}, \widehat{\pi}^*) + \int_\Gamma g\widehat{\chi}^*_{|\Gamma}(\mathbf{v} \cdot \mathbf{n}) + g a_p(\widehat{\chi}^*, \psi) \\ - \int_\Gamma g(\widehat{\boldsymbol{\omega}}^* \cdot \mathbf{n})\psi_{|\Gamma} + \int_\Gamma \frac{2g^2}{\gamma_2} \widehat{\chi}^*_{|\Gamma} \psi_{|\Gamma} = 0 \qquad \forall (\mathbf{v}, \psi) \in H_f^\tau \times H_p. \end{aligned}$$

Taking $\mathbf{v} = \widehat{\boldsymbol{\omega}}^*$ and $\psi = \widehat{\chi}^*$, we find

$$a_f(\widehat{\boldsymbol{\omega}}^*, \widehat{\boldsymbol{\omega}}^*) + g a_p(\widehat{\chi}^*, \widehat{\chi}^*) + \int_\Gamma \frac{2g^2}{\gamma_2} (\widehat{\chi}^*_{|\Gamma})^2 = 0\,;$$

hence, $\widehat{\chi}^* = 0$ in Ω_p, and $\widehat{\boldsymbol{\omega}}^* = \mathbf{0}$ in Ω_f thanks to the Korn inequality.

The interface equation (30) gives $\widehat{\sigma}^* = 0$ on Γ; hence, $\mathbf{u}_f^* \cdot \mathbf{n} = -\mathsf{K}\nabla\varphi^* \cdot \mathbf{n}$ on Γ. Moreover, using (26), we obtain $\mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f^*, p_f^*) \cdot \mathbf{n}) = -g\varphi^*_{|\Gamma}$ on Γ. Thus, the two interface conditions (5) and (7) are fulfilled, so that the solutions $(\mathbf{u}_f^*, p_f^*) \in H_f^\tau \times Q$ and $\varphi^* \in H_p$ (corresponding to the fixed point μ^*) satisfy the coupled Stokes–Darcy problem.

Our aim is now to prove that the map generating the sequence μ^k is a contraction in $L^2(\Gamma)$. We shall address this point in section 4.2.

Remark 3.1. A Neumann–Neumann method corresponding to the choice of the normal velocity $\mathbf{u}_f \cdot \mathbf{n}$ as an interface variable would involve the following steps. For an assigned function $\lambda^k \in H_{00}^{1/2}(\Gamma)$, with $\int_\Gamma \lambda^k = 0$ (we refer to [14] for a definition of the trace space $H_{00}^{1/2}(\Gamma)$), first solve a Stokes problem in Ω_f with boundary conditions $\mathbf{u}_f^{k+1} \cdot \mathbf{n} = \lambda^k$, $\mathbf{u}_f^{k+1} \cdot \boldsymbol{\tau}_j = 0$ on Γ, and a Darcy problem in Ω_p imposing $-\mathsf{K}\nabla\varphi^{k+1} \cdot \mathbf{n} = \lambda^k$ on Γ. Then, similarly to (29), we have to compute $\widehat{\sigma}^{k+1} = -\mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f^{k+1}, p_f^{k+1}) \cdot \mathbf{n}) - g\varphi^{k+1}_{|\Gamma}$ on Γ. Here we would have $\widehat{\sigma}^{k+1} \in H^{-1/2}(\Gamma)$. Therefore, this regularity of $\widehat{\sigma}^{k+1}$ would not be enough to guarantee the solvability of the subsequent Darcy problem, which would demand one to impose $g\widehat{\chi}^{k+1}_{|\Gamma} = \widehat{\sigma}^{k+1}$ as a boundary condition

on Γ. Thus, a Neumann–Neumann method does not guarantee that the regularity of the interface data is preserved at each iteration and that the sequence λ^k generated by the algorithm is in $H_{00}^{1/2}(\Gamma)$.

Of course one may speculate that this issue of lack of regularity is not relevant at the finite dimensional level, for instance, for finite element approximation. However, the difficulty is only hidden, and we should expect that it will show up as the mesh parameter h goes to 0.

4. Convergence analysis. In what follows, for either an open set or a manifold D, we denote the norm in the Sobolev space $H^s(D)$, $s \geq -1$, by $\|\cdot\|_{s,D}$.

4.1. Convergence of the sRR method. We prove that the sequences φ^k and $(\mathbf{u}_f^k, p_f^k)$ generated by the sRR method (17)–(21) converge in $H^1(\Omega_p)$ and $(H^1(\Omega_f))^d \times Q$, respectively. As a consequence, the sequence η^k is convergent in the dual space $H^{-1/2}(\Gamma)$ and weakly convergent in $L^2(\Gamma)$.

The proof of convergence that we are presenting follows the guidelines of the theory by Lions [15] for the Robin–Robin method (see also [17, section 4.5]).

We denote by $\mathbf{e}_u^k = \mathbf{u}_f^k - \mathbf{u}_f$, $e_p^k = p_f^k - p_f$, and $e_\varphi^k = \varphi^k - \varphi$ the errors at the kth step. Remark that, thanks to the linearity, the functions $(\mathbf{e}_u^k, e_p^k)$ satisfy problem (19) with $\mathbf{f} = \mathbf{0}$, while e_φ^k is a solution to (17). Moreover, we assume that $\gamma_p = \gamma_f$, and we denote by γ their common value.

Finally, let us point out that the solutions $(\mathbf{u}_f, p_f) \in H_f^\tau \times Q$ and $\varphi \in H_p$ of the coupled Stokes–Darcy problem satisfy $\mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}) \in H^{1/2}(\Gamma)$ (as it is equal to $-g\varphi_{|\Gamma}$ on Γ), and $\nabla\varphi \cdot \mathbf{n} \in L^2(\Gamma)$ (as it is equal to $-\mathsf{K}^{-1}\mathbf{u}_f \cdot \mathbf{n}$ on Γ); i.e., these functions enjoy a better regularity than one might usually expect. Therefore, the interface conditions (18) and (20) for the error functions hold in $L^2(\Gamma)$.

Let us come to the proof of convergence. Choosing $\psi = e_\varphi^{k+1}$ in (17), and using the identity

$$AB = \frac{1}{4}[(A+B)^2 - (A-B)^2],$$

we have

$$\begin{aligned} g\, a_p(e_\varphi^{k+1}, e_\varphi^{k+1}) &= \frac{1}{\gamma}\int_\Gamma (\eta^k - g e_{\varphi|\Gamma}^{k+1}) g e_{\varphi|\Gamma}^{k+1} \\ &= \frac{1}{4\gamma}\int_\Gamma (\eta^k)^2 - \frac{1}{4\gamma}\int_\Gamma (\eta^k - 2g e_{\varphi|\Gamma}^{k+1})^2. \end{aligned} \tag{35}$$

Similarly, taking $\mathbf{v} = \mathbf{e}_u^{k+1}$ in (19) and using (21), we have

$$\begin{aligned} a_f(\mathbf{e}_u^{k+1}, \mathbf{e}_u^{k+1}) &= \frac{1}{\gamma}\int_\Gamma (\eta^k - 2g e_{\varphi|\Gamma}^{k+1} - \gamma \mathbf{e}_u^{k+1}\cdot\mathbf{n})(\gamma \mathbf{e}_u^{k+1}\cdot\mathbf{n}) \\ &= \frac{1}{4\gamma}\int_\Gamma (\eta^k - 2g e_{\varphi|\Gamma}^{k+1})^2 - \frac{1}{4\gamma}\int_\Gamma (\eta^k - 2g e_{\varphi|\Gamma}^{k+1} - 2\gamma \mathbf{e}_u^{k+1}\cdot\mathbf{n})^2 \\ &= \frac{1}{4\gamma}\int_\Gamma (\eta^k - 2g e_{\varphi|\Gamma}^{k+1})^2 - \frac{1}{4\gamma}\int_\Gamma (\eta^{k+1})^2. \end{aligned} \tag{36}$$

Adding (35) and (36), we find

$$g\, a_p(e_\varphi^{k+1}, e_\varphi^{k+1}) + a_f(\mathbf{e}_u^{k+1}, \mathbf{e}_u^{k+1}) + \frac{1}{4\gamma}\int_\Gamma (\eta^{k+1})^2 = \frac{1}{4\gamma}\int_\Gamma (\eta^k)^2.$$

Summing over k from $k=0$ to $k=N$, with $N \geq 1$, we finally obtain

$$\sum_{k=0}^{N} \left(g\, a_p(e_\varphi^{k+1}, e_\varphi^{k+1}) + a_f(\mathbf{e}_u^{k+1}, \mathbf{e}_u^{k+1}) \right) + \frac{1}{4\gamma} \int_\Gamma (\eta^{N+1})^2 = \frac{1}{4\gamma} \int_\Gamma (\eta^0)^2.$$

Thus, the series

$$\sum_{k=0}^{\infty} \left(g\, a_p(e_\varphi^{k+1}, e_\varphi^{k+1}) + a_f(\mathbf{e}_u^{k+1}, \mathbf{e}_u^{k+1}) \right)$$

is convergent, and the errors e_φ^k and $\mathbf{e}_u^k$ tend to zero in $H^1(\Omega_p)$ and $(H^1(\Omega_f))^d$, respectively. The convergence of the pressure error e_p^k to 0 in Q is then a well-known consequence of the convergence of the velocity.

4.1.1. Interpretation of the sRR method as an alternating direction scheme. The sRR method can be interpreted as an alternating direction scheme (see [1]; see also [8]). For technical reasons, to make precise this statement let us assume that a flux boundary condition $\mathsf{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n} = \mathbf{g}$ is imposed on the top of the fluid domain Ω_f, $\mathbf{g}$ being a given vector function. Moreover, we assume that the interface Γ is smooth, say, a $\mathcal{C}^2$-manifold with a boundary.

Then introduce the spaces

$$\begin{aligned}
&\widehat{H}_f = \{\mathbf{v} \in (H^1(\Omega_f))^d |\ \mathbf{v} = \mathbf{0} \text{ on the lateral boundary of } \Omega_f\},\\
&\widehat{H}_f^\tau = \{\mathbf{v} \in \widehat{H}_f |\ \mathbf{v} \cdot \boldsymbol{\tau}_j = 0 \text{ on } \Gamma,\ j = 1, \ldots, d-1\},\\
&\widehat{H}_f^{\tau,n} = \{\mathbf{v} \in \widehat{H}_f^\tau |\ \mathbf{v} \cdot \mathbf{n} = 0 \ \text{ on } \Gamma\}, \quad H_p^0 = \{\psi \in H_p |\ \psi = 0 \ \text{ on } \Gamma_p\},
\end{aligned}$$

and define the operator S_f as

$$S_f : H_{00}^{1/2}(\Gamma) \to (H_{00}^{1/2}(\Gamma))', \quad \chi \to S_f\chi = \mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_\chi, p_\chi) \cdot \mathbf{n}),$$

where $(\mathbf{u}_\chi, p_\chi) \in \widehat{H}_f^\tau \times Q$ satisfies

$$\begin{aligned}
&a_f(\mathbf{u}_\chi, \mathbf{v}) + b_f(\mathbf{v}, p_\chi) = 0 \qquad \forall\ \mathbf{v} \in \widehat{H}_f^{\tau,n}(\Omega_f),\\
&b_f(\mathbf{u}_\chi, q) = 0 \qquad \forall\ q \in Q,
\end{aligned}$$

with $\mathbf{u}_\chi \cdot \mathbf{n} = \chi$ on Γ.

In a similar way, for each $\eta \in (H_{00}^{1/2}(\Gamma))'$ define the operator S_p as

$$S_p : (H_{00}^{1/2}(\Gamma))' \to H_{00}^{1/2}(\Gamma), \quad \eta \to S_p\eta = g\varphi_{\eta|\Gamma},$$

where $\varphi_\eta \in H_p^0$ is the solution to

$$a_p(\varphi_\eta, \psi) = \langle \eta, \psi_{|\Gamma} \rangle_\Gamma \qquad \forall\ \psi \in H_p^0,$$

where $\langle \cdot, \cdot \rangle_\Gamma$ denotes the duality pairing between $(H_{00}^{1/2}(\Gamma))'$ and $H_{00}^{1/2}(\Gamma)$. As a consequence, we have $-\mathsf{K}\nabla\varphi_\eta \cdot \mathbf{n} = \eta$ on Γ.

Since for each $\varphi \in H_p^0$ we have $S_p(-\mathsf{K}\nabla\varphi \cdot \mathbf{n}) = g\varphi_{|\Gamma}$, the first step (19) of our procedure corresponds to imposing on Γ

$$\begin{aligned}
-\gamma_p \mathsf{K}\nabla\varphi^{k+1} \cdot \mathbf{n} + g\varphi_{|\Gamma}^{k+1} &= -\gamma_p \mathsf{K}\nabla\varphi^{k+1} \cdot \mathbf{n} + S_p(-\mathsf{K}\nabla\varphi^{k+1} \cdot \mathbf{n})\\
&= (\gamma_p I + S_p)(-\mathsf{K}\nabla\varphi^{k+1} \cdot \mathbf{n}) = \eta^k\ ;
\end{aligned}$$

hence

$$-\mathsf{K}\nabla\varphi^{k+1}\cdot\mathbf{n} = (\gamma_p I + S_p)^{-1}\eta^k. \tag{37}$$

On the other hand, the right-hand side in (20) can be written as

$$\begin{aligned} -g\varphi_{|\Gamma}^{k+1} - \gamma_f\mathsf{K}\nabla\varphi^{k+1}\cdot\mathbf{n} &= S_p(\mathsf{K}\nabla\varphi^{k+1}\cdot\mathbf{n}) - \gamma_f\mathsf{K}\nabla\varphi^{k+1}\cdot\mathbf{n} \\ &= -(\gamma_f I - S_p)\mathsf{K}\nabla\varphi^{k+1}\cdot\mathbf{n} \\ &= (\gamma_f I - S_p)(\gamma_p I + S_p)^{-1}\eta^k. \end{aligned} \tag{38}$$

In an analogous way, still denoting by $(\mathbf{u}_f^{k+1}, p_f^{k+1})$ the solution to (19) with $\mathbf{f} = \mathbf{0}$ and H_f^τ replaced by $\widehat{H}_f^\tau$, one has $S_f(\mathbf{u}_f^{k+1}\cdot\mathbf{n}) = \mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_f^{k+1}, p_f^{k+1})\cdot\mathbf{n})$. Then, the left-hand side in (20) can be written as

$$\begin{aligned} \mathbf{n}\cdot(\mathsf{T}(\mathbf{u}^{k+1}, p^{k+1})\cdot\mathbf{n}) + \gamma_f\mathbf{u}^{k+1}\cdot\mathbf{n} &= S_f(\mathbf{u}^{k+1}\cdot\mathbf{n}) + \gamma_f\mathbf{u}^{k+1}\cdot\mathbf{n} \\ &= (\gamma_f I + S_f)(\mathbf{u}^{k+1}\cdot\mathbf{n}). \end{aligned} \tag{39}$$

Using (38) and (39), the interface condition (20) becomes

$$\mathbf{u}^{k+1}\cdot\mathbf{n} = (\gamma_f I + S_f)^{-1}(\gamma_f I - S_p)(\gamma_p I + S_p)^{-1}\eta^k. \tag{40}$$

In conclusion, our iterative procedure (with homogeneous data $\mathbf{f}$ and $\mathbf{g}$) can be written as

$$\begin{aligned} \eta^{k+1} &= -\mathbf{n}\cdot(\mathsf{T}(\mathbf{u}^{k+1}, p^{k+1})\cdot\mathbf{n}) + \gamma_p\mathbf{u}^{k+1}\cdot\mathbf{n} \\ &= -S_f(\mathbf{u}^{k+1}\cdot\mathbf{n}) + \gamma_p\mathbf{u}^{k+1}\cdot\mathbf{n} \\ &= (\gamma_p I - S_f)\mathbf{u}^{k+1}\cdot\mathbf{n} \\ &= (\gamma_p I - S_f)(\gamma_f I + S_f)^{-1}(\gamma_f I - S_p)(\gamma_p I + S_p)^{-1}\eta^k. \end{aligned} \tag{41}$$

This is an alternating direction scheme, à la Peaceman and Rachford (see [16]), that has been deeply analyzed. Sufficient conditions for convergence are that $\gamma_f = \gamma_p$ and that the operators S_f and S_p are bounded and strictly positive in a given Hilbert space. These do not apply in the present situation, as the operators S_f and S_p act from a space into its dual. In fact, we can prove only that the iteration operator is nonexpansive but not a contraction in $(H_{00}^{1/2}(\Gamma))'$.

On the other hand, it is worthy to note that the convergence of this alternating direction scheme can be easily proved in the discrete case, as the matrices that correspond to the finite dimensional Steklov–Poincaré operators S_f and S_p are in fact symmetric and positive definite.

To illustrate how the proof of convergence works, we consider a suitable modification of the iteration scheme. Let us introduce the operators $J_- : H_{00}^{1/2}(\Gamma) \to (H_{00}^{1/2}(\Gamma))'$ and $J_+ : (H_{00}^{1/2}(\Gamma))' \to H_{00}^{1/2}(\Gamma)$ defined as follows:

$$\begin{aligned} (J_-\chi, \mu)_{-1/2,00,\Gamma} &= \langle\mu, \chi\rangle_\Gamma \qquad \forall\ \chi \in H_{00}^{1/2}(\Gamma),\ \mu \in (H_{00}^{1/2}(\Gamma))', \\ (J_+\eta, \xi)_{1/2,00,\Gamma} &= \langle\eta, \xi\rangle_\Gamma \qquad \forall\ \eta \in (H_{00}^{1/2}(\Gamma))',\ \xi \in H_{00}^{1/2}(\Gamma). \end{aligned}$$

(Here and in what follows we are denoting by $(\cdot,\cdot)_{1/2,00,\Gamma}$ and $(\cdot,\cdot)_{-1/2,00,\Gamma}$ the scalar products in $H^{1/2}(\Gamma)$ and $(H_{00}^{1/2}(\Gamma))'$, respectively, and by $\|\cdot\|_{1/2,00,\Gamma}$ and $\|\cdot\|_{-1/2,00,\Gamma}$ the associated norms.)

The existence of these operators is guaranteed by the Riesz representation theorem. Moreover, it is easily verified that $\|J_-\chi\|_{-1/2,00,\Gamma} = \|\chi\|_{1/2,00,\Gamma}$, $\|J_+\eta\|_{1/2,00,\Gamma} = \|\eta\|_{-1/2,00,\Gamma}$ (and consequently the operator norms are $\|J_-\| = \|J_+\| = 1$), and $(J_-\chi,\eta)_{-1/2,00,\Gamma} = (\chi, J_+\eta)_{1/2,00,\Gamma}$.

We consider the following iterative scheme:

$$\eta^{k+1} = (\gamma J_- - S_f)(\gamma J_- + S_f)^{-1} J_-(\gamma J_+ - S_p)(\gamma J_+ + S_p)^{-1} J_+ \eta^k. \tag{42}$$

This represents a slight modification of (41), in which we have inserted the operators J_- and J_+ instead of the identity I, and we have taken $\gamma_p = \gamma_f = \gamma$. The convergence of (42) is a consequence of the contraction mapping theorem (see the appendix).

Remark 4.1. One could argue that the iterative scheme (42) is not relevant with the problem at hand, since it is not equivalent to (41). Indeed, (42) converges to our original problem with slightly modified interface conditions, which read

$$\begin{aligned} &\gamma\, J_-(\mathbf{u}_f \cdot \mathbf{n}) + \mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n}) = -\gamma\, J_- J_+(\mathsf{K}\nabla\varphi \cdot \mathbf{n}) - J_-(g\varphi_{|\Gamma}) \quad \text{on } \Gamma, \\ &\gamma\, J_+ J_-(\mathbf{u}_f \cdot \mathbf{n}) - J_+(\mathbf{n} \cdot (\mathsf{T}(\mathbf{u}_f, p_f) \cdot \mathbf{n})) = -\gamma\, J_+(\mathsf{K}\nabla\varphi \cdot \mathbf{n}) + g\varphi_{|\Gamma} \quad \text{on } \Gamma. \end{aligned}$$

The operators J_- and J_+ have the role of assuring that the functions on either side are in the same trace space.

The problem of equalization of trace spaces can be encountered in other domain decompositions of heterogeneous problems as well. For these cases, the procedure that we have advocated here (and the associated convergence proof) might be useful.

4.2. Convergence of the pRR method. We turn now to the proof of convergence of the parallel method (24)–(31). Our aim is to prove that the map $\mu^k \to \mu^{k+1}$ defined through (24)–(31) is a contraction in $L^2(\Gamma)$. As a consequence of linearity, in the whole section we can assume without restriction that $\mathbf{f} = \mathbf{0}$. In order to introduce a suitable representation of this map, we define several interface operators.

Let $\mathcal{H}_S$ be the *Robin-to-Dirichlet* map for the *Stokes problem*,

$$\mathcal{H}_S : L^2(\Gamma) \to L^2(\Gamma), \qquad \mu \to \mathcal{H}_S\mu = \mathbf{u}_\mu \cdot \mathbf{n}, \tag{43}$$

where $(\mathbf{u}_\mu, p_\mu) \in H_f^\tau \times Q$ is the solution to (24) with $\mathbf{f} = \mathbf{0}$ and the Robin boundary datum μ.

Define $\mathcal{H}_D$ as the *Robin-to-Neumann* operator for the *Darcy scalar problem*,

$$\mathcal{H}_D : L^2(\Gamma) \to L^2(\Gamma), \qquad \mu \to \mathcal{H}_D\mu = \frac{1}{\gamma_1}(g\varphi_{\mu|\Gamma} + \mu), \tag{44}$$

where $\varphi_\mu \in H_p$ is the solution to (25) corresponding to the Robin boundary datum μ.

Moreover, let $\mathcal{K}_S$ be the *Robin-to-Neumann* operator for the *Stokes problem*,

$$\mathcal{K}_S : L^2(\Gamma) \to L^2(\Gamma), \qquad \sigma \to \mathcal{K}_S\sigma = \gamma_2(\sigma - \boldsymbol{\omega}_\sigma \cdot \mathbf{n}), \tag{45}$$

where $(\boldsymbol{\omega}_\sigma, \pi_\sigma) \in H_f^\tau \times Q$ is the solution to (27) with the Robin boundary datum σ.

Finally, $\mathcal{K}_D$ denotes the *Robin-to-Dirichlet* operator for the *Darcy scalar problem*,

$$\mathcal{K}_D : L^2(\Gamma) \to L^2(\Gamma), \qquad \sigma \to \mathcal{K}_D\sigma = g\chi_{\sigma|\Gamma}, \tag{46}$$

$\chi_\sigma \in H_p$ being the solution to (28) with the Robin boundary datum σ.

By means of these operators, we reformulate (29) as

$$\widehat{\sigma}^{k+1} = \mathcal{H}_S\mu^k + \mathcal{H}_D\mu^k = (\mathcal{H}_S + \mathcal{H}_D)\mu^k$$

and the relaxation step (31) as

$$\begin{aligned}\mu^{k+1} &= \mu^k - \theta(\mathcal{K}_S\widehat{\sigma}^{k+1} + \mathcal{K}_D\widehat{\sigma}^{k+1}) = \mu^k - \theta(\mathcal{K}_S + \mathcal{K}_D)(\mathcal{H}_S + \mathcal{H}_D)\mu^k \\ &= [I - \theta(\mathcal{K}_S + \mathcal{K}_D)(\mathcal{H}_S + \mathcal{H}_D)]\mu^k.\end{aligned}$$

PROPOSITION 4.1. *The operators defined in* (43)–(46) *enjoy the following properties:*

1. *$\mathcal{H}_S$ and $\mathcal{K}_D$ are symmetric, continuous, and nonnegative in $L^2(\Gamma)$;*
2. *$\mathcal{H}_D$ and $\mathcal{K}_S$ are symmetric, continuous, and coercive in $L^2(\Gamma)$.*

Proof. 1. We consider first the operator $\mathcal{H}_S$. For every η and μ, letting $\mathbf{u}_\eta \cdot \mathbf{n} = \mathcal{H}_S\eta$ and $\mathbf{u}_\mu \cdot \mathbf{n} = \mathcal{H}_S\mu$, we have

$$\begin{aligned}\int_\Gamma (\mathcal{H}_S\mu)\eta &= \int_\Gamma \mathbf{u}_\mu \cdot \mathbf{n}\,\eta = a_f(\mathbf{u}_\eta, \mathbf{u}_\mu) - \gamma_1 \int_\Gamma (\mathbf{u}_\eta \cdot \mathbf{n})(\mathbf{u}_\mu \cdot \mathbf{n}) \\ &= \int_\Gamma \mu\, \mathbf{u}_\eta \cdot \mathbf{n} = \int_\Gamma \mu\, (\mathcal{H}_S\eta);\end{aligned}$$

therefore, $\mathcal{H}_S$ is symmetric.

Now, taking $\mathbf{v} = \mathbf{u}_\mu$ in (24) (with $\mathbf{f} = \mathbf{0}$), thanks to (22) we have

$$\begin{aligned}2\nu \int_{\Omega_f} |\mathsf{D}(\mathbf{u}_\mu)|^2 = a_f(\mathbf{u}_\mu, \mathbf{u}_\mu) &= \gamma_1 \int_\Gamma |\mathbf{u}_\mu \cdot \mathbf{n}|^2 + \int_\Gamma \mu\, \mathbf{u}_\mu \cdot \mathbf{n} \\ &\le \gamma_1\kappa_2 \int_{\Omega_f} |\mathsf{D}(\mathbf{u}_\mu)|^2 + \kappa_2^{1/2}\|\mu\|_{0,\Gamma}\|\mathsf{D}(\mathbf{u}_\mu)\|_{0,\Omega_f}.\end{aligned}$$

Therefore, for $\gamma_1 < (2\nu)/\kappa_2$, one has $\|\mathsf{D}(\mathbf{u}_\mu)\|_{0,\Omega_f} \le \kappa_3\|\mu\|_{0,\Gamma}$, with $\kappa_3 = \kappa_2^{1/2}/(2\nu - \gamma_1\kappa_2)$. Hence, from (22), $\mathcal{H}_S$ is a continuous operator.

Finally, for $\gamma_1 < (2\nu)/\kappa_2$ we have

$$\int_\Gamma (\mathcal{H}_S\mu)\,\mu = 2\nu \int_{\Omega_f} |\mathsf{D}(\mathbf{u}_\mu)|^2 - \gamma_1 \int_\Gamma |\mathbf{u}_\mu \cdot \mathbf{n}|^2 \ge (2\nu - \gamma_1\kappa_2) \int_{\Omega_f} |\mathsf{D}(\mathbf{u}_\mu)|^2 \ge 0;$$

hence, $\mathcal{H}_S$ is a nonnegative operator.

We consider now the operator $\mathcal{K}_D$. We denote by χ_σ and χ_ξ the solutions to (28) with data σ and ξ, respectively. Thus, $\mathcal{K}_D\sigma = g\chi_{\sigma|\Gamma}$ and $\mathcal{K}_D\xi = g\chi_{\xi|\Gamma}$. Then using (28) we have

$$\begin{aligned}\int_\Gamma (\mathcal{K}_D\sigma)\xi = \int_\Gamma g\chi_{\sigma|\Gamma}\xi &= g\, a_p(\chi_\xi, \chi_\sigma) + \frac{g^2}{\gamma_2}\int_\Gamma \chi_{\xi|\Gamma}\,\chi_{\sigma|\Gamma} \\ &= \int_\Gamma g\sigma\chi_{\xi|\Gamma} = \int_\Gamma \sigma(\mathcal{K}_D\xi),\end{aligned}$$

which proves the symmetry of $\mathcal{K}_D$.

Now if we take in (28) the test function $\psi = \chi_\sigma$, we find

$$a_p(\chi_\sigma, \chi_\sigma) + \frac{1}{\gamma_2}\int_\Gamma g\chi_{\sigma|\Gamma}^2 = \int_\Gamma \sigma\chi_{\sigma|\Gamma} \le \left(\int_\Gamma \sigma^2\right)^{1/2}\left(\int_\Gamma \chi_{\sigma|\Gamma}^2\right)^{1/2};$$

consequently, since $a_p(\chi_\sigma, \chi_\sigma) \ge 0$, we have $g\|\chi_{\sigma|\Gamma}\|_{0,\Gamma} \le \gamma_2\|\sigma\|_{0,\Gamma}$; i.e., $\mathcal{K}_D$ is a continuous operator.

Finally, $\mathcal{K}_D$ is nonnegative, since

$$\int_\Gamma (\mathcal{K}_D\sigma)\,\sigma = \int_\Gamma g\chi_{\sigma|\Gamma}\,\sigma = g\,a_p(\chi_\sigma,\chi_\sigma) + \frac{g^2}{\gamma_2}\int_\Gamma \chi^2_{\sigma|\Gamma} \geq 0 \quad \forall\sigma\in L^2(\Gamma).$$

2. Consider now the operator $\mathcal{H}_D$. For all μ and η we denote by φ_μ and φ_η the solutions of (25) corresponding to the data μ and η, respectively, so that $\mathcal{H}_D\mu = (g\varphi_{\mu|\Gamma}+\mu)/\gamma_1$ and $\mathcal{H}_D\eta = (g\varphi_{\eta|\Gamma}+\eta)/\gamma_1$. Then, proceeding as we did for the operator $\mathcal{K}_D$, we have

$$\begin{aligned}\int_\Gamma (\mathcal{H}_D\mu)\,\eta &= \frac{1}{\gamma_1}\int_\Gamma (\mu\,\eta + g\varphi_{\mu|\Gamma}\,\eta)\\ &= \frac{1}{\gamma_1}\int_\Gamma \mu\,\eta - \frac{g^2}{\gamma_1}\int_\Gamma \varphi_{\eta|\Gamma}\varphi_{\mu|\Gamma} - g\,a_p(\varphi_\eta,\varphi_\mu)\\ &= \frac{1}{\gamma_1}\int_\Gamma \mu\,\eta + \frac{g}{\gamma_1}\int_\Gamma \mu\,\varphi_{\eta|\Gamma} = \int_\Gamma \mu\,(\mathcal{H}_D\eta);\end{aligned}$$

thus, $\mathcal{H}_D$ is symmetric.

Moreover, taking $\psi = \varphi_\mu$ in (25), the continuity of $\mathcal{H}_D$ easily follows from the estimate:

$$a_p(\varphi_\mu,\varphi_\mu) + \frac{g}{\gamma_1}\int_\Gamma \varphi^2_{\mu|\Gamma} = -\frac{1}{\gamma_1}\int_\Gamma \mu\varphi_{\mu|\Gamma} \leq \frac{1}{\gamma_1}\left(\int_\Gamma \mu^2\right)^{1/2}\left(\int_\Gamma \varphi^2_{\mu|\Gamma}\right)^{1/2},$$

which yields $\|\varphi_{\mu|\Gamma}\|_{0,\Gamma} \leq g^{-1}\|\mu\|_{0,\Gamma}$, as $a_p(\varphi_\mu,\varphi_\mu)\geq 0$.

Finally, let us show that $\mathcal{H}_D$ is a coercive operator. Recalling its definition, we have

$$\begin{aligned}a_p(\varphi_\mu,\varphi_\mu) &= -\frac{1}{\gamma_1}\int_\Gamma g\varphi^2_{\mu|\Gamma} - \frac{1}{\gamma_1}\int_\Gamma \mu\varphi_{\sigma|\Gamma} = -\int_\Gamma (\mathcal{H}_D\mu)\,\varphi_{\mu|\Gamma}\\ &= -\frac{1}{g}\int_\Gamma (\mathcal{H}_D\mu)(\gamma_1\mathcal{H}_D\mu - \mu) = \frac{1}{g}\int_\Gamma (\mathcal{H}_D\mu)\,\mu - \frac{\gamma_1}{g}\int_\Gamma (\mathcal{H}_D\mu)^2.\end{aligned}$$

Consequently, since $a_p(\varphi_\mu,\varphi_\mu)\geq \kappa_3\int_{\Omega_p}|\nabla\varphi_\mu|^2$ for a suitable constant $\kappa_3>0$, there exists a constant $q_1>0$ such that

$$\int_\Gamma (\mathcal{H}_D\mu)\,\mu \geq q_1\left(\int_\Gamma (\mathcal{H}_D\mu)^2 + \int_{\Omega_p}|\nabla\varphi_\mu|^2\right).$$

On the other hand, using the trace inequality and the Poincaré inequality,

$$\begin{aligned}\int_\Gamma \mu^2 &= \int_\Gamma (\gamma_1\mathcal{H}_D\mu - g\varphi_{\mu|\Gamma})^2 \leq 2\gamma_1^2\int_\Gamma (\mathcal{H}_D\mu)^2 + 2g^2\int_\Gamma \varphi^2_{\mu|\Gamma}\\ &\leq Q_1\left(\int_\Gamma (\mathcal{H}_D\mu)^2 + \int_{\Omega_p}|\nabla\varphi_\mu|^2\right)\end{aligned}$$

where $Q_1>0$ is a suitable constant. The coerciveness of $\mathcal{H}_D$ now follows.

Turning now to the operator $\mathcal{K}_S$, its symmetry can be proved as we did for $\mathcal{H}_S$. Moreover, taking $\mathbf{v} = \boldsymbol{\omega}_\sigma$ in (27) (where $\boldsymbol{\omega}_\sigma$ is the solution with datum σ), one has

$$a_f(\boldsymbol{\omega}_\sigma,\boldsymbol{\omega}_\sigma) + \gamma_2\int_\Gamma (\boldsymbol{\omega}_\sigma\cdot\mathbf{n})^2 = \gamma_2\int_\Gamma \sigma\,\boldsymbol{\omega}_\sigma\cdot\mathbf{n}.$$

Since $a_f(\boldsymbol{\omega}_\sigma, \boldsymbol{\omega}_\sigma) \geq 0$, this yields

$$\int_\Gamma (\boldsymbol{\omega}_\sigma \cdot \mathbf{n})^2 \leq \int_\Gamma \sigma\, \boldsymbol{\omega}_\sigma \cdot \mathbf{n} \leq \left(\int_\Gamma \sigma^2 \right)^{1/2} \left(\int_\Gamma (\boldsymbol{\omega}_\sigma \cdot \mathbf{n})^2 \right)^{1/2},$$

and this proves that the operator $\mathcal{K}_S$ is continuous.

Finally, using the definition (45) of $\mathcal{K}_S$, we have

$$\begin{aligned} a_f(\boldsymbol{\omega}_\sigma, \boldsymbol{\omega}_\sigma) &= -\gamma_2 \int_\Gamma (\boldsymbol{\omega}_\sigma \cdot \mathbf{n})^2 + \gamma_2 \int_\Gamma \sigma\, \boldsymbol{\omega}_\sigma \cdot \mathbf{n} = \int_\Gamma (\mathcal{K}_S \sigma)\, \boldsymbol{\omega}_\sigma \cdot \mathbf{n} \\ &= \int_\Gamma (\mathcal{K}_S \sigma)\, (\sigma - \gamma_2^{-1} \mathcal{K}_S \sigma) = \int_\Gamma (\mathcal{K}_S \sigma)\, \sigma - \gamma_2^{-1} \int_\Gamma (\mathcal{K}_S \sigma)^2. \end{aligned}$$

Therefore, since $a_f(\boldsymbol{\omega}_\sigma, \boldsymbol{\omega}_\sigma) = 2\nu \int_{\Omega_f} |\mathsf{D}(\boldsymbol{\omega}_\sigma)|^2$, there exists a constant $q_2 > 0$ such that

$$\int_\Gamma (\mathcal{K}_S \sigma)\, \sigma \geq q_2 \left(\int_{\Omega_f} |\mathsf{D}(\boldsymbol{\omega}_\sigma)|^2 + \int_\Gamma (\mathcal{K}_S \sigma)^2 \right).$$

On the other hand, by the trace and the Korn inequalities, we have

$$\begin{aligned} \int_\Gamma \sigma^2 &= \int_\Gamma (\boldsymbol{\omega}_\sigma \cdot \mathbf{n} + \gamma_2^{-1} \mathcal{K}_S \sigma)^2 \leq 2 \int_\Gamma (\boldsymbol{\omega}_\sigma \cdot \mathbf{n})^2 + 2\gamma_2^{-2} \int_\Gamma (\mathcal{K}_S \sigma)^2 \\ &\leq Q_2 \left(\int_{\Omega_f} |\mathsf{D}(\boldsymbol{\omega}_\sigma)|^2 + \int_\Gamma (\mathcal{K}_S \sigma)^2 \right) \end{aligned}$$

for a suitable constant $Q_2 > 0$. Thus, the operator $\mathcal{K}_S$ is coercive. ∎

It follows from Proposition 4.1 that the operators $\mathcal{H} = \mathcal{H}_S + \mathcal{H}_D$ and $\mathcal{K} = \mathcal{K}_S + \mathcal{K}_D$ are both *symmetric, continuous, and coercive* on $L^2(\Gamma)$.

To prove the convergence of the pRR iterative scheme, we shall apply the following abstract result whose proof is similar to that of Theorem 4.2.5 in [17].

THEOREM 4.1. *Let X be a (real) Hilbert space and X' its dual. We consider a linear invertible continuous operator $\mathcal{Q}\colon X \to X'$, which can be split as $\mathcal{Q} = \mathcal{Q}_1 + \mathcal{Q}_2$, where both $\mathcal{Q}_1$ and $\mathcal{Q}_2$ are linear operators. Take $\mathcal{Z} \in X'$, let $x \in X$ be the unknown solution to the equation $\mathcal{Q}x = \mathcal{Z}$, and consider for its solution the preconditioned Richardson method*

$$x^{k+1} = x^k + \theta \mathcal{N}(\mathcal{Z} - \mathcal{Q}x^k), \quad k \geq 0, \tag{47}$$

θ being a positive relaxation parameter and $\mathcal{N} : X' \to X$ a suitable scaling operator. Suppose that the following conditions are satisfied:

1. *$\mathcal{Q}_i$ $(i = 1, 2)$ are continuous and coercive;*
2. *$\mathcal{N}$ is symmetric, continuous, and coercive.*

Then there exists $\theta_{max} > 0$ such that for each $\theta \in (0, \theta_{max})$ and for any given $x^0 \in X$ the sequence (47) converges in X to the solution of problem $\mathcal{Q}x = \mathcal{Z}$.

We can now prove the main result of this section.

COROLLARY 4.1. *Under the constraint (32), the pRR iterative method (24), (25), (27), (28), (31) converges to the solution $(\mathbf{u}_f, p_f) \in H_f^\tau \times Q$, $\varphi \in H_p$ of the coupled Stokes–Darcy problem for any choice of the initial guess $\mu^0 \in L^2(\Gamma)$ and for suitable values of the relaxation parameter θ.*

Proof. It follows from Theorem 4.1, whose hypotheses are satisfied thanks to Proposition 4.1. ∎

5. Finite element approximation and numerical results. We consider a regular family of triangulations $\mathcal{T}_h$ of the domain $\overline{\Omega_f} \cup \overline{\Omega_p}$ depending on a positive parameter $h > 0$, made up of triangles if $d = 2$ or of tetrahedra in the 3-dimensional case. We assume that the triangulations $\mathcal{T}_{fh}$ and $\mathcal{T}_{ph}$ induced on the subdomains Ω_f and Ω_p are compatible on Γ; i.e., they share the same edges (if $d = 2$) or faces (if $d = 3$) therein. The family of triangulations induced on Γ will be denoted by $\mathcal{B}_h$.

Several choices of finite element spaces can be made to approximate the coupled problem (13)–(14). For the sake of exposition, we will consider the following conforming spaces ($d = 2, 3$):

$$H_{fh} = \{\mathbf{v}_h \in (X_{fh})^d |\, \mathbf{v}_h = \mathbf{0} \text{ on } \Gamma_f\},$$

with

$$X_{fh} = \{v_h \in \mathcal{C}^0(\overline{\Omega_f}) |\, v_{h|T} \in \mathbb{P}_2(T)\ \forall T \in \mathcal{T}_{fh}\},$$

and

$$Q_h = \{q_h \in \mathcal{C}^0(\overline{\Omega_f}) |\, q_{h|T} \in \mathbb{P}_1(T),\ \forall T \in \mathcal{T}_{fh}\};$$

moreover, H^τ_{fh} will be an internal approximation of H^τ_f.

On the other hand, we set

$$H_{ph} = \{\psi_h \in X_{ph} |\, \psi_h = 0 \text{ on } \Gamma^b_p\},$$

with

$$X_{ph} = \{\psi_h \in \mathcal{C}^0(\overline{\Omega_p}) |\, \psi_{h|T} \in \mathbb{P}_2(T)\ \forall T \in \mathcal{T}_{ph}\}.$$

Finally, we define

$$\Lambda_h = \{\eta_h \in L^2(\Gamma) \,|\, \eta_{h|\tau} \in \mathbb{P}_2(\tau)\ \forall \tau \in \mathcal{B}_h\};$$

in particular, we have that $\mathbf{v}_h \cdot \mathbf{n} \in \Lambda_h$ for each $\mathbf{v}_h \in H_{fh}$ and $\psi_{h|\Gamma} \in \Lambda_h$ for each $\psi_h \in H_{ph}$.

We will now present the discrete counterpart of the sRR and pRR algorithms.

5.1. The discrete sRR method. The finite element discretization of the coupled Stokes–Darcy problem (13)–(16) reads as follows:

Find $(\mathbf{u}_{fh}, p_{fh}) \in H^\tau_{fh} \times Q_h$, $\varphi_h \in H_{ph}$ such that

$$\begin{aligned} &a_f(\mathbf{u}_{fh}, \mathbf{v}_h) + b_f(\mathbf{v}_h, p_{fh}) + g\, a_p(\varphi_h, \psi_h) + \int_\Gamma g\, \varphi_h(\mathbf{v}_h \cdot \mathbf{n}) \\ &\qquad - \int_\Gamma g\, \psi_h(\mathbf{u}_{fh} \cdot \mathbf{n}) = \int_{\Omega_f} \mathbf{f} \cdot \mathbf{v}_h \qquad \forall \mathbf{v}_h \in H^\tau_{fh},\ \psi_h \in H_{ph}, \end{aligned} \tag{48}$$

$$b_f(\mathbf{u}_{fh}, q_h) = 0 \qquad \forall q_h \in Q_h. \tag{49}$$

The sRR algorithm on the discrete problem (48)–(49) becomes, taking a trace function $\eta^0_h \in \Lambda_h$ and considering two acceleration parameters $\gamma_f \geq 0$ and $\gamma_p > 0$, for each $k \geq 0$,

(i) find $\varphi^{k+1}_h \in H_{ph}$ such that

$$\gamma_p a_p(\varphi^{k+1}_h, \psi_h) + \int_\Gamma g \varphi^{k+1}_{h|\Gamma} \psi_{h|\Gamma} = \int_\Gamma \eta^k_h \psi_{h|\Gamma} \qquad \forall\ \psi_h \in H_{ph}. \tag{50}$$

(ii) Then find $(\mathbf{u}_{fh}^{k+1}, p_{fh}^{k+1}) \in H_{fh}^{\tau} \times Q_h$ such that

$$
\begin{aligned}
&a_f(\mathbf{u}_{fh}^{k+1}, \mathbf{v}_h) + b_f(\mathbf{v}_h, p_{fh}^{k+1}) + \gamma_f \int_\Gamma (\mathbf{u}_{fh}^{k+1} \cdot \mathbf{n})(\mathbf{v}_h \cdot \mathbf{n}) \\
&\quad = \int_\Gamma \Big(\frac{\gamma_f}{\gamma_p} \eta_h^k - \frac{\gamma_f + \gamma_p}{\gamma_p} g \varphi_{h|\Gamma}^{k+1} \Big) (\mathbf{v}_h \cdot \mathbf{n}) + \int_{\Omega_f} \mathbf{f} \cdot \mathbf{v}_h \quad \forall\, \mathbf{v}_h \in H_{fh}^{\tau}, \\
&b_f(\mathbf{u}_{fh}^{k+1}, q_h) = 0 \qquad \forall\, q_h \in Q_h.
\end{aligned} \tag{51}
$$

(iii) Finally, set

$$
\eta_h^{k+1} = (\gamma_f + \gamma_p)(\mathbf{u}_{fh}^{k+1} \cdot \mathbf{n}) + \frac{\gamma_f + \gamma_p}{\gamma_p} g \varphi_{h|\Gamma}^{k+1} - \frac{\gamma_f}{\gamma_p} \eta_h^k \in \Lambda_h. \tag{52}
$$

For $\gamma_p = \gamma_f$, the convergence of this algorithm to the solution of (48)–(49) can be proved as we did in section 4.1 to show the convergence of (17)–(20) to the solution of problems (13)–(16). Moreover, it is also possible to prove the convergence of the alternating direction scheme (see section 4.1.1), as the discrete Steklov–Poincaré operators are positive definite (however, in principle the proof of convergence cannot assure that the rate of convergence is independent of the mesh size h).

For the numerical tests we have exploited the interpretation of the method in terms of ADI iterations (section 4.1.1) in order to obtain some guidelines for the choice of the relaxation parameters, at least for the case of our interest, that is, when ν and the entries of K are very small (we recall that in this case the convergence rate of the Dirichlet–Neumann method deteriorates).

In particular, considering (41), we are led to investigate the behavior of the eigenvalues, say δ_f^j and δ_p^j, of the operators S_f and S_p, respectively; in fact, if we can estimate

$$
\max_j \left| \frac{\gamma_p - \delta_f^j}{\gamma_f + \delta_f^j} \right| \cdot \max_j \left| \frac{\gamma_f - \delta_p^j}{\gamma_p + \delta_p^j} \right|, \tag{53}
$$

this could be taken as a rough estimate of the convergence rate of the algorithm.

Assuming that K is a constant multiple of the identity, we proved that in the limit $\nu \to 0$ and $\mathsf{K} \to 0$ (for a fixed mesh size h) $\delta_f^j \to 0$ while $\delta_p^j \to \infty$ [8]. Thus, for small values of ν and K the ratio (53) behaves like γ_p/γ_f. This provides a first indication for the choice of the relaxation parameters; i.e., one should take $\gamma_f > \gamma_p > 0$. Moreover, γ_f and γ_p should not be taken too large to avoid possible increases of the condition numbers of the Stokes and Darcy stiffness matrices in (50) and (51), respectively. A reasonable trade-off is to choose both parameters approximately equal to 10^{-1}.

For the numerical tests, we take the same setting as in Example 3.1. In Table 2 we report the number of iterations obtained using the sRR method for some small values of ν and K and for four different computational grids. A convergence test based on the relative increment of the trace of the discrete normal velocity on the interface

TABLE 2

Number of iterations using the sRR method with respect to ν, K, and four different grid sizes h ($h_1 \approx 0.14$ and $h_i = h_1/2^{i-1}$, $i = 2, 3, 4$); the acceleration parameters are $\gamma_f = 0.3$ and $\gamma_p = 0.1$.

ν	K	h_1	h_2	h_3	h_4
10^{-4}	10^{-3}	19	19	19	19
10^{-6}	10^{-4}	20	20	20	20
10^{-6}	10^{-7}	20	20	20	20

$\mathbf{u}_{fh}^k \cdot \mathbf{n}_{|\Gamma}$ has been considered with tolerance 10^{-9}. In all computations we have taken $\gamma_f = 0.3$ and $\gamma_p = 0.1$.

5.2. The discrete pRR method. The pRR algorithm designed on (48)–(49) reads as follows: Let $\mu_h^0 \in \Lambda_h$ be a discrete trace function on Γ, and let $\gamma_1, \gamma_2 > 0$ be two positive relaxation parameters; then for $k \geq 0$

(i) find $(\mathbf{u}_{fh}^{k+1}, p_{fh}^{k+1}) \in H_{fh}^\tau \times Q_h$ such that

$$\begin{aligned} &a_f(\mathbf{u}_{fh}^{k+1}, \mathbf{v}_h) + b_f(\mathbf{v}_h, p_{fh}^{k+1}) - \gamma_1 \int_\Gamma (\mathbf{u}_{fh}^{k+1} \cdot \mathbf{n})(\mathbf{v}_h \cdot \mathbf{n}) \\ &\qquad = \int_\Gamma \mu_h^k (\mathbf{v}_h \cdot \mathbf{n}) + \int_{\Omega_f} \mathbf{f} \cdot \mathbf{v}_h \qquad \forall\, \mathbf{v}_h \in H_{fh}^\tau, \\ &b_f(\mathbf{u}_{fh}^{k+1}, q_h) = 0 \qquad \forall\, q_h \in Q_h, \end{aligned} \tag{54}$$

and find $\varphi_h^{k+1} \in H_{ph}$ such that

$$a_p(\varphi_h^{k+1}, \psi_h) + \frac{1}{\gamma_1} \int_\Gamma g \varphi_{h|\Gamma}^{k+1} \psi_{h|\Gamma} = -\frac{1}{\gamma_1} \int_\Gamma \mu_h^k \psi_{h|\Gamma} \qquad \forall\, \psi_h \in H_{ph}. \tag{55}$$

(ii) Then find $(\widehat{\boldsymbol{\omega}}_h^{k+1}, \widehat{\pi}_h^{k+1}) \in H_{fh}^\tau \times Q_h$ such that

$$\begin{aligned} &a_f(\widehat{\boldsymbol{\omega}}_h^{k+1}, \mathbf{v}_h) + b_f(\mathbf{v}_h, \widehat{\pi}_h^{k+1}) + \gamma_2 \int_\Gamma (\widehat{\boldsymbol{\omega}}_h^{k+1} \cdot \mathbf{n})(\mathbf{v}_h \cdot \mathbf{n}) \\ &\qquad = \gamma_2 \int_\Gamma \widehat{\sigma}_h^{k+1} (\mathbf{v}_h \cdot \mathbf{n}) \qquad \forall\, \mathbf{v}_h \in H_{fh}^\tau, \\ &b_f(\widehat{\boldsymbol{\omega}}_h^{k+1} q_h) = 0 \qquad \forall\, q_h \in Q_h, \end{aligned} \tag{56}$$

and find $\widehat{\chi}_h^{k+1} \in H_{ph}$ such that

$$a_p(\widehat{\chi}_h^{k+1}, \psi_h) + \frac{1}{\gamma_2} \int_\Gamma g \widehat{\chi}_{h|\Gamma}^{k+1} \psi_{h|\Gamma} = \int_\Gamma \widehat{\sigma}_h^{k+1} \psi_{h|\Gamma} \qquad \forall\, \psi_h \in H_{ph}, \tag{57}$$

where

$$\widehat{\sigma}_h^{k+1} = \mathbf{u}_{fh}^{k+1} \cdot \mathbf{n} + \frac{1}{\gamma_1}(g \varphi_{h|\Gamma}^{k+1} + \mu_h^k) \in \Lambda_h. \tag{58}$$

(iii) Finally, update μ_h^k as follows:

$$\mu_h^{k+1} = \mu_h^k - \theta[\gamma_2(\widehat{\sigma}_h^{k+1} - \widehat{\boldsymbol{\omega}}_h^{k+1} \cdot \mathbf{n}) + g \widehat{\chi}_{h|\Gamma}^{k+1}] \in \Lambda_h, \tag{59}$$

where $\theta > 0$ is an acceleration parameter.

As for the continuous case, this iterative scheme can be reformulated in terms of suitable interface operators on Λ_h. Precisely, let $\mathcal{H}_{Sh}$ and $\mathcal{K}_{Dh}$ be the discrete Robin-to-Dirichlet maps:

$$\begin{aligned} &\mathcal{H}_{Sh} : \Lambda_h \to \Lambda_h, \qquad \mu_h \to \mathcal{H}_{Sh}\mu_h = \mathbf{u}_{\mu_h} \cdot \mathbf{n}, \\ &\mathcal{K}_{Dh} : \Lambda_h \to \Lambda_h, \qquad \sigma_h \to \mathcal{K}_{Dh}\sigma_h = g\chi_{\sigma_h|\Gamma}, \end{aligned}$$

where $(\mathbf{u}_{\mu_h}, p_{\mu_h}) \in H_{fh}^\tau \times Q_h$ is the solution to (54) with $\mathbf{f} = \mathbf{0}$ and Robin boundary datum μ_h, while $\chi_{\sigma_h} \in H_{ph}$ is the solution of (57) with boundary datum σ_h on Γ.

Then consider the discrete Robin-to-Neumann operators

$$\mathcal{H}_{Dh} : \Lambda_h \to \Lambda_h, \qquad \mu_h \to \mathcal{H}_{Dh}\mu_h = \frac{1}{\gamma_1}(g\varphi_{\mu_h|\Gamma} + \mu_h),$$
$$\mathcal{K}_{Sh} : \Lambda_h \to \Lambda_h, \qquad \sigma_h \to \mathcal{K}_{Sh}\sigma_h = \gamma_2(\sigma_h - \boldsymbol{\omega}_{\sigma_h} \cdot \mathbf{n}),$$

where $\varphi_{\mu_h} \in H_{ph}$ is the solution of (55) with boundary datum μ_h, and $(\boldsymbol{\omega}_{\sigma_h}, \pi_{\sigma_h})$ is the solution of (56) with boundary datum σ_h.

Finally, we denote by $(\widetilde{\mathbf{u}}_h, \widetilde{p}_h) \in H^{\tau}_{fh} \times Q_h$ the solution of (54) with null boundary conditions, so that $\mathbf{u}^{k+1}_{fh} \cdot \mathbf{n} = \mathcal{H}_{Sh}\mu^k_h + \widetilde{\mathbf{u}}_h \cdot \mathbf{n}$ for all $k \geq 0$.

Then (58) becomes

$$\widehat{\sigma}^{k+1}_h = \mathcal{H}_{Sh}\mu^k_h + \mathcal{H}_{Dh}\mu^k_h + \widetilde{\mathbf{u}}_h \cdot \mathbf{n}.$$

Problem (48)–(49) can be associated with the discrete interface problem

$$\text{Find } \mu_h \in \Lambda_h : \quad (\mathcal{H}_{Sh} + \mathcal{H}_{Dh})\mu_h = -\widetilde{\mathbf{u}}_h \cdot \mathbf{n} \qquad \text{on } \Gamma. \tag{60}$$

Thus the discrete pRR method can be interpreted as the following preconditioned Richardson scheme to solve (60):

$$\mu^{k+1}_h = \mu^k_h - \theta(\mathcal{K}_{Sh} + \mathcal{K}_{Dh})[\widetilde{\mathbf{u}}_h \cdot \mathbf{n} + (\mathcal{H}_{Sh} + \mathcal{H}_{Dh})\mu^k_h], \qquad k \geq 0, \tag{61}$$

the preconditioner being

$$P = (\mathcal{K}_{Sh} + \mathcal{K}_{Dh})^{-1}. \tag{62}$$

The convergence of (61) is proved as done in section 4.2 for the infinite dimensional case; besides, its rate of convergence is independent of the mesh size h, as it depends only on the continuity and coerciveness constants of the operators $\mathcal{H}_{Sh}$, $\mathcal{H}_{Dh}$, $\mathcal{K}_{Sh}$, and $\mathcal{K}_{Dh}$, which are all independent of h.

Moreover, since the operators $\mathcal{H}_{Sh}$ and $\mathcal{H}_{Dh}$ are symmetric, we can use the PCG method to compute the solution of (60) using the same preconditioner (62).

More generally, we consider the following (variable) preconditioner:

$$P_k = (\sigma^k_1\mathcal{K}_{Sh} + \sigma^k_2\mathcal{K}_{Dh})^{-1}, \tag{63}$$

where σ^k_1 and σ^k_2 are two suitable acceleration coefficients (possibly depending on the iteration k).

The choice of the coefficients γ_1, γ_2, σ^k_1, and σ^k_2 to accelerate convergence is not straightforward. In our numerical experiments we have adopted two different strategies. First, we have used the PCG method with $P^{-1} = \sigma_1\mathcal{K}_{Sh} + \sigma_2\mathcal{K}_{Dh}$ with a suitable choice of the acceleration coefficients. Second, we have considered the preconditioner P^{-1}_k as in (63) in the framework of a Richardson method, and we have computed σ^k_1 and σ^k_2 according to an Aitken acceleration procedure (see, e.g., [5, 4]).

More precisely, the algorithm reads: Let r^0_h be the residual of (60) computed with respect to an initial datum $\mu^0_h \in \Lambda_h$, and let $z^0_h = P^{-1}_0 r^0_h$. Then for $k \geq 0$

1. compute the local preconditioned residuals $z^k_{Dh} = \mathcal{K}_{Dh}r^k_h$, $z^k_{Sh} = \mathcal{K}_{Sh}r^k_h$;
2. solve the linear system

$$A^T_k A_k \begin{pmatrix} \sigma^k_1 \\ \sigma^k_2 \end{pmatrix} = -A^T_k(\mu^k_h - \mu^{k-1}_h),$$

where A_k is the two column matrix $A_k = (z^k_{Sh} - z^{k-1}_{Sh};\ z^k_{Dh} - z^{k-1}_{Dh})$.

TABLE 3

Number of iterations using the PCG method with the pRR preconditioner P as in (62)*, with respect to ν and the grid size h ($h_1 \approx 0.14$ and $h_i = h_1/2^{i-1}$, $i = 2, 3, 4$).*

ν	K	γ_1	γ_2	σ_1	σ_2	h_1	h_2	h_3	h_4
1	1	0.5	0.5	1	1	11	12	11	12
10^{-1}	1	10^{-1}	1	1	1	27	28	29	28
10^{-2}	1	10^{-2}	1	1	1	68	76	72	64

TABLE 4

Number of iterations using the Aitken-accelerated Richardson method with the pRR preconditioner P_k as in (63)*; in the last two columns we indicate the mean value of the absolute values of the parameters σ_1^k and σ_2^k generated by the method. The h_i are as in Table* 3*.*

ν	K	γ_1	γ_2	Grid size	Iter.	$\lvert\bar{\sigma}_1\rvert$	$\lvert\bar{\sigma}_2\rvert$
1	1	0.5	0.5	h_1	10	2.68	0.64
				h_2	10	2.67	0.66
				h_3	10	2.66	0.67
				h_4	10	2.66	0.68
10^{-1}	1	10^{-1}	1	h_1	12	1.53	0.13
				h_2	11	1.50	0.13
				h_3	11	1.54	0.13
				h_4	12	1.50	0.12
10^{-2}	1	10^{-2}	1	h_1	23	0.90	0.06
				h_2	23	0.95	0.04
				h_3	23	0.96	0.06
				h_4	23	0.94	0.06
10^{-3}	1	10^{-3}	1	h_1	47	0.33	0.07
				h_2	47	0.38	0.04
				h_3	50	0.37	0.03
				h_4	52	0.38	0.03
10^{-1}	10^{-1}	10^{-1}	10	h_1	23	0.90	0.06
				h_2	23	0.95	0.04
				h_3	23	0.96	0.06
				h_4	23	0.94	0.06
10^{-2}	10^{-1}	10^{-2}	10^2	h_1	40	0.25	0.02
				h_2	39	0.26	0.01
				h_3	40	0.30	0.01
				h_4	44	0.27	0.01

This corresponds to minimizing

$$\|(\mu_h^k - \mu_h^{k-1}) + \sigma_1(z_{Sh}^k - z_{Sh}^{k-1}) + \sigma_2(z_{Dh}^k - z_{Dh}^{k-1})\|$$

over all possible values of σ_1 and σ_2.

3. Finally, update $z_h^{k+1} = \sigma_1^k z_{Sh}^k + \sigma_2^k z_{Dh}^k$, $r_h^{k+1} = r_h^k - (\mathcal{H}_{Sh} + \mathcal{H}_{Dh}) z_h^{k+1}$, and $\mu_h^{k+1} = \mu_h^k + z_h^{k+1}$.

For the numerical tests, we have considered the same settings as in Example 3.1. A tolerance of 10^{-9} has been imposed on the relative increment, and a maximal number of iterations `maxit = 300` has been required.

Table 3 reports the number of iterations obtained using the PCG method for three values of ν and four different grids. It is apparent that the rate of convergence deteriorates as ν goes to 0. We have noticed a similar behavior for small values of K as well.

The Richardson–Aitken strategy gives better results, as shown in Table 4. However, the Dirichlet–Neumann algorithm still turns out to be more efficient in this respect (see Table 1).

6. Appendix. We present here the proof of the convergence of the (modified) sRR scheme (42).

THEOREM 6.1. *Let us assume that the interface* Γ *is smooth, say, a* $\mathcal{C}^2$*-manifold with a boundary. Then for each* $\gamma > 0$ *the operator* $(\gamma J_- - S_f)(\gamma J_- + S_f)^{-1}$ *is a contraction in* $(H_{00}^{1/2}(\Gamma))'$*, and the operator* $(\gamma J_+ - S_p)(\gamma J_+ + S_p)^{-1}$ *is a contraction in* $H_{00}^{1/2}(\Gamma)$*.*

Proof. We have

$$\begin{aligned}
\|(\gamma J_- - S_f)(\gamma J_- + S_f)^{-1}\|^2 &= \sup_{\mu\neq 0} \frac{\|(\gamma J_- - S_f)(\gamma J_- + S_f)^{-1}\mu\|^2_{-1/2,00,\Gamma}}{\|\mu\|^2_{-1/2,00,\Gamma}} \\
&= \sup_{\chi\neq 0} \frac{\|(\gamma J_- - S_f)\chi\|^2_{-1/2,00,\Gamma}}{\|(\gamma J_- + S_f)\chi\|^2_{-1/2,00,\Gamma}} \\
&= \sup_{\chi\neq 0} \frac{\gamma^2\|J_-\chi\|^2_{-1/2,00,\Gamma} - 2\gamma(S_f\chi, J_-\chi)_{-1/2,00,\Gamma} + \|S_f\chi\|^2_{-1/2,00,\Gamma}}{\gamma^2\|J_-\chi\|^2_{-1/2,00,\Gamma} + 2\gamma(S_f\chi, J_-\chi)_{-1/2,00,\Gamma} + \|S_f\chi\|^2_{-1/2,00,\Gamma}} \\
&= \sup_{\chi\neq 0} \frac{\gamma^2\|\chi\|^2_{1/2,00,\Gamma} - 2\gamma\langle S_f\chi,\chi\rangle_\Gamma + \|S_f\chi\|^2_{-1/2,00,\Gamma}}{\gamma^2\|\chi\|^2_{1/2,00,\Gamma} + 2\gamma\langle S_f\chi,\chi\rangle_\Gamma + \|S_f\chi\|^2_{-1/2,00,\Gamma}}.
\end{aligned}$$

We prove now that S_f is positive and bounded; that is, there exist two positive constants C_1 and C_2 such that

$$\langle S_f\chi,\chi\rangle_\Gamma \geq C_1\|\chi\|^2_{1/2,00,\Gamma}, \qquad \|S_f\chi\|^2_{-1/2,00,\Gamma} \leq C_2\|\chi\|^2_{1/2,00,\Gamma}. \tag{64}$$

In fact, using the Korn and the trace inequality in $H_{00}^{1/2}(\Gamma)$ we have

$$\begin{aligned}
\langle S_f\chi,\chi\rangle_\Gamma &= \langle \mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_\chi,p_\chi)\cdot\mathbf{n}), \mathbf{u}_\chi\cdot\mathbf{n}\rangle_\Gamma \\
&= \left\langle \mathsf{T}(\mathbf{u}_\chi,p_\chi)\cdot\mathbf{n}, \mathbf{n}(\mathbf{u}_\chi\cdot\mathbf{n}) + \sum_{j=1}^{d-1}\boldsymbol{\tau}_j(\mathbf{u}_\chi\cdot\boldsymbol{\tau}_j)\right\rangle_\Gamma \\
&\quad (\text{as } \mathbf{u}_\chi\cdot\boldsymbol{\tau}_j = 0 \text{ on } \Gamma) \\
&= \int_{\Omega_f} \boldsymbol{\nabla}\cdot[\mathsf{T}(\mathbf{u}_\chi,p_\chi)\cdot\mathbf{u}_\chi] = 2\nu\int_{\Omega_f}|\mathsf{D}(\mathbf{u}_\chi)|^2 \\
&\geq c_1\|\mathbf{u}_\chi\|^2_{1,\Omega_f} \geq c_2\|\mathbf{u}_{\chi|\Gamma}\|^2_{1/2,00,\Gamma}.
\end{aligned}$$

The regularity assumption on Γ yields $\|\mathbf{u}_\chi\cdot\mathbf{n}\|_{1/2,00,\Gamma} \leq c_3\|\mathbf{u}_{\chi|\Gamma}\|_{1/2,00,\Gamma}$; hence,

$$\langle S_f\chi,\chi\rangle_\Gamma \geq C_1\|\mathbf{u}_\chi\cdot\mathbf{n}\|^2_{1/2,00,\Gamma} = C_1\|\chi\|^2_{1/2,00,\Gamma}.$$

Moreover, the regularity assumption on Γ also yields

$$\|\mathbf{n}\cdot(\mathsf{T}(\mathbf{u}_\chi,p_\chi)\cdot\mathbf{n})\|_{-1/2,00,\Gamma} \leq c_4\|\mathsf{T}(\mathbf{u}_\chi,p_\chi)\cdot\mathbf{n}\|_{-1/2,00,\Gamma};$$

therefore, the trace inequality in $(H_{00}^{1/2}(\Gamma))'$ and the a priori estimate for the solution of the Stokes problem give

$$\begin{aligned}
\|S_f\chi\|^2_{-1/2,00,\Gamma} &\leq c_4^2\|\mathsf{T}(\mathbf{u}_\chi,p_\chi)\cdot\mathbf{n}\|^2_{-1/2,00,\Gamma} \\
&\leq c_5\|\mathsf{T}(\mathbf{u}_\chi,p_\chi)\|^2_{0,\Omega_f} \leq C_2\|\chi\|^2_{1/2,00,\Gamma},
\end{aligned}$$

so that both inequalities in (64) are proved.

Consequently, setting $q_0 = (C_2 - 2\gamma C_1 + \gamma^2)/(C_2 + 2\gamma C_1 + \gamma^2)$, we can easily prove that

$$\sup_{\chi \neq 0} \frac{\gamma^2 \|\chi\|^2_{1/2,00,\Gamma} - 2\gamma \langle S_f \chi, \chi \rangle_\Gamma + \|S_f \chi\|^2_{-1/2,00,\Gamma}}{\gamma^2 \|\chi\|^2_{1/2,00,\Gamma} + 2\gamma \langle S_f \chi, \chi \rangle_\Gamma + \|S_f \chi\|^2_{-1/2,00,\Gamma}} \leq q_0 < 1.$$

The proof that $\|(\gamma J_+ - S_p)(\gamma J_+ + S_p)^{-1}\| < 1$ can be done in a similar way. In fact, using the trace inequality in $(H^{1/2}_{00}(\Gamma))'$ we have

$$\begin{aligned}
\langle \eta, S_p \eta \rangle_\Gamma &= -g \langle \mathsf{K} \nabla \varphi_\eta \cdot \mathbf{n}, \varphi_{\eta|\Gamma} \rangle_\Gamma \\
&= g \int_{\Omega_p} \nabla \cdot [\varphi_\eta \mathsf{K} \nabla \varphi_\eta] = g \int_{\Omega_p} \nabla \varphi_\eta \cdot \mathsf{K} \nabla \varphi_\eta \\
&= g \int_{\Omega_p} \mathsf{K}^{-1} \mathsf{K} \nabla \varphi_\eta \cdot \mathsf{K} \nabla \varphi_\eta \geq c_6 \int_{\Omega_p} |\mathsf{K} \nabla \varphi_\eta|^2 \\
&\geq C_3 \|\mathsf{K} \nabla \varphi_\eta \cdot \mathbf{n}\|^2_{-1/2,00,\Gamma} = C_3 \|\eta\|^2_{-1/2,00,\Gamma}.
\end{aligned}$$

Moreover, by the trace inequality in $H^{1/2}_{00}(\Gamma)$ and the a priori estimate for the solution of the Laplace equation, we obtain

$$\|S_p \eta\|^2_{1/2,00,\Gamma} = \|g \varphi_{\eta|\Gamma}\|^2_{1/2,00,\Gamma} \leq c_7 \|\varphi_\eta\|^2_{1,\Omega_p} \leq C_4 \|\eta\|^2_{-1/2,00,\Gamma}.$$

These two inequalities permit one to repeat for the operator S_p the same procedure used for the operator S_f. ∎

REFERENCES

[1] V. I. Agoshkov and V. I. Lebedev, *Variational algorithms of the domain decomposition method*, Sov. J. Numer. Anal. Math. Modelling, 5 (1990), pp. 27–46. Originally published in Russian as preprint no. 54, Dept. Numer. Math. URSS Acad. Sci. Moscow, 1983.

[2] F. Brezzi, *On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers*, Rev. Française Automat. Informat. Recherche Opérationnelle, Sér. Rouge, 8 (1974), pp. 129–151.

[3] R. Dautray and J.-L. Lions, *Mathematical Analysis and Numerical Methods for Science and Technology*, Vol. 2., Springer, Berlin, 1988.

[4] S. Deparis, M. Discacciati, and A. Quarteroni, *A domain decomposition framework for fluid-structure interaction problems*, in Proceedings of ICCFD3, Computational Fluid Dynamics 2004, C. Groth and D. W. Zingg, eds., Springer, New York, 2006, pp. 41–58.

[5] S. Deparis, *Numerical Analysis of Axisymmetric Flows and Methods for Fluid-Structure Interaction Arising in Blood Flow Simulation*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland, 2004.

[6] M. Discacciati and A. Quarteroni, *Analysis of a domain decomposition method for the coupling of Stokes and Darcy equations*, in Proceedings of ENUMATH 2001, F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, eds., Numer. Math. Adv. Appl., Springer, Milan, 2003, pp. 3–20.

[7] M. Discacciati and A. Quarteroni, *Convergence analysis of a subdomain iterative method for the finite element approximation of the coupling of Stokes and Darcy equations*, Comput. Vis. Sci., 6 (2004), pp. 93–103.

[8] M. Discacciati, *Domain Decomposition Methods for the Coupling of Surface and Groundwater Flows*, Ph.D. thesis, Ecole Polytechnique Fédérale de Lausanne, Switzerland, 2004.

[9] M. Discacciati, *Iterative methods for Stokes/Darcy coupling*, in Domain Decomposition Methods in Science and Engineering, R. Kornhuber, R. Hoppe, J. Périaux, O. Pironneau, O. Widlund, and J. Xu, eds., Lect. Notes Comput. Sci. Eng. 40, Springer, Berlin, 2004, pp. 563–570.

[10] J. C. Galvis and M. Sarkis, *Inf-sup for coupling Stokes-Darcy*, in Proceedings of the XXV Iberian and Latin American Congress on Computational Methods in Engineering, CILAMCE XXV, Universidade Federal de Pernambuco, Recife, 2004.

[11] J. C. Galvis and M. Sarkis, *Balancing domain decomposition methods for mortar coupling Stokes-Darcy systems*, in Domain Decomposition Methods in Science and Engineering XVI, O. B. Widlund and D. E. Keyes, eds., Lect. Notes Comput. Sci. Eng. 55, Springer, Berlin, 2007, pp. 373–380.

[12] W. Jäger and A. Mikelić, *On the interface boundary condition of Beavers, Joseph, and Saffman*, SIAM J. Appl. Math., 60 (2000), pp. 1111–1127.

[13] W. J. Layton, F. Schieweck, and I. Yotov, *Coupling fluid flow with porous media flow*, SIAM J. Numer. Anal., 40 (2003), pp. 2195–2218.

[14] J.-L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications*, Vol. 1, Dunod, Paris, 1968.

[15] P.-L. Lions, *On the Schwarz alternating method* III: *A variant for non-overlapping subdomains*, in Proceedings of the Third International Symposium on Domain Decomposition Methods for Partial Differential Equations, T. F. Chan, R. Glowinski, J. Périaux, and O. B. Widlund, eds., SIAM, Philadelphia, 1990, pp. 202–231.

[16] D. W. Peaceman and H. H. Rachford, Jr., *The numerical solution of parabolic and elliptic differential equations*, J. Soc. Indust. Appl. Math., 3 (1955), pp. 28–41.

[17] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford Science, Oxford, 1999.

[18] B. Rivière and I. Yotov, *Locally conservative coupling of Stokes and Darcy flows*, SIAM J. Numer. Anal., 42 (2005), pp. 1959–1977.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1269–1286

NEW FINITE ELEMENT METHODS IN COMPUTATIONAL FLUID DYNAMICS BY $H(\mathrm{DIV})$ ELEMENTS*

JUNPING WANG† AND XIU YE‡

Abstract. In this paper, the authors present two formulations for the Stokes problem which make use of the existing $H(\mathrm{div})$ elements of the Raviart–Thomas type originally developed for the second-order elliptic problems. In addition, two new $H(\mathrm{div})$ elements are constructed and analyzed particularly for the new formulations. Optimal-order error estimates are established for the corresponding finite element solutions in various Sobolev norms. The finite element solutions feature a full satisfaction of the continuity equation when existing Raviart–Thomas-type elements are employed in the numerical scheme.

Key words. finite element methods, Stokes problem

AMS subject classifications. Primary, 65N15, 65N30, 76D07; Secondary, 35B45, 35J50

DOI. 10.1137/060649227

1. Introduction.

This paper is concerned with numerical solutions of incompressible fluid flow problems by finite element methods. Our objective is to introduce a finite element scheme with attention paid to the discretization of the mass continuity equation. For illustrative purposes, we show how the method works for the Stokes problem, which seeks a pair of unknown functions $(\mathbf{u}; p)$ satisfying

$$-\nu\Delta\mathbf{u} + \nabla p = \mathbf{f} \quad \text{in } \Omega, \tag{1.1}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{in } \Omega, \tag{1.2}$$

$$\mathbf{u} = 0 \quad \text{on } \partial\Omega, \tag{1.3}$$

where ν denotes the fluid viscosity; Δ, ∇, and $\nabla\cdot$ denote the Laplacian, gradient, and divergence operators, respectively; $\Omega \subset \mathbb{R}^d$ is the region occupied by the fluid; $\mathbf{f} = \mathbf{f}(\mathbf{x}) \in (L^2(\Omega))^d$ is the unit external volumetric force acting on the fluid at $\mathbf{x} \in \Omega$.

The commonly used finite element methods for the Stokes problem (1.1)–(1.3) are based on a variational equation which is obtained by testing the momentum equation (1.1) by functions in $(H_0^1(\Omega))^d$ and the continuity equation (1.2) by functions in $L^2(\Omega)$ (see section 2 for their definition). The corresponding finite element method requires a pair of finite element spaces which are conforming in $(H_0^1(\Omega))^d \times L^2(\Omega)$ and satisfy the inf-sup condition of Babuška [2] and Brezzi [3]. These constraints result in finite element approximations, denoted by $(\mathbf{u}_h; p_h)$, which hardly satisfy the continuity equation

$$\nabla \cdot \mathbf{u}_h(\mathbf{x}) = 0 \quad \forall\, \mathbf{x} \in \Omega. \tag{1.4}$$

Readers are referred to [8, 19, 21] for more details regarding the approximation methods and their properties.

*Received by the editors January 5, 2006; accepted for publication (in revised form) February 1, 2007; published electronically May 22, 2007.
http://www.siam.org/journals/sinum/45-3/64922.html

†Division of Mathematical Sciences, National Science Foundation, Arlington, VA 22230 (jwang@nsf.gov). This author's research was supported in part by the NSF IR/D program. This research was initiated while the author was visiting NASA Center for Computational Sciences (NCCS) at GSFC through the NASA/ASEE Faculty Fellowship Program.

‡Department of Mathematics, University of Arkansas at Little Rock, Little Rock, AR 72204 (xxye@ualr.edu).

The recent development in discontinuous Galerkin methods [1, 4, 5, 6, 10, 11, 13] provides new means of solving the Stokes equations numerically. However, the corresponding finite element solutions are usually totally discontinuous and fail to satisfy the continuity equation (1.4) in the classical sense [12, 22, 24, 27].

The continuity equation (1.4) requires the numerical solution $\mathbf{u}_h$ to be a member of the Sobolev space $H(\mathrm{div};\Omega)$. Therefore, the discontinuous Galerkin methods [12, 22, 24, 27] appear to be noncompetitive when (1.4) needs to be satisfied. On the other hand, the $(H_0^1)^d \times L^2$ conforming finite element methods require the total continuity of $\mathbf{u}_h$, which is too much to satisfy for (1.4). Therefore, it seems that the $H(\mathrm{div})$ elements of Raviart–Thomas type [25, 7, 8, 17] might be good candidates for producing new numerical schemes that satisfy (1.4).

The goal of this paper is to present a method that demonstrates the use of $H(\mathrm{div})$ elements in solving the Stokes problem. Our main contribution is on the development of a new formulation for the Stokes problem which makes use of the existing $H(\mathrm{div})$ elements in numerical schemes. Optimal-order error estimates are derived for the resulting $H(\mathrm{div})$ finite element approximations. In addition, two new families of $H(\mathrm{div})$ elements are proposed and analyzed in this article.

This paper is organized as follows. In section 2, we introduce some preliminaries and notations for Sobolev spaces. A new variational formula is presented in section 3 for the Stokes problem. In section 4, we present a $H(\mathrm{div})$ finite element method by using the variational formula developed in section 3. In section 5, we establish some optimal-order error estimates for the new finite element approximations in H^1 and L^2 norms. Finally, in section 6, we review some representatives of $H(\mathrm{div})$ elements, followed with a detailed description of two new $H(\mathrm{div})$ elements.

2. Preliminaries and notations. Let D be any domain in $\mathbb{R}^d, d = 2, 3$. For simplicity, the method will be presented for two-dimensional problems only. An extension to higher-dimensional problems can be made formally for general polyhedral domains.

We use standard definitions for the Sobolev spaces $H^s(D)$ and their associated inner products $(\cdot,\cdot)_{s,D}$, norms $\|\cdot\|_{s,D}$, and seminorms $|\cdot|_{s,D}$ for $s \geq 0$. For example, for any integer $s \geq 0$, the seminorm $|\cdot|_{s,D}$ is given by

$$|v|_{s,D} = \left(\sum_{|\alpha|=s} \int_D |\partial^\alpha v|^2 dD \right)^{\frac{1}{2}},$$

with the usual notation

$$\alpha = (\alpha_1, \alpha_2), \quad |\alpha| = \alpha_1 + \alpha_2, \quad \partial^\alpha = \partial_{x_1}^{\alpha_1} \partial_{x_2}^{\alpha_2}.$$

The Sobolev norm $\|\cdot\|_{m,D}$ is given by

$$\|v\|_{m,D} = \left(\sum_{j=0}^{m} |v|_{j,D}^2 \right)^{\frac{1}{2}}.$$

The space $H^0(D)$ coincides with $L^2(D)$, for which the norm and the inner product are denoted by $\|\cdot\|_D$ and $(\cdot,\cdot)_D$, respectively. When $D = \Omega$, we shall drop the subscript D in the norm and inner product notation. We also use $L_0^2(\Omega)$ to denote the subspace of $L^2(\Omega)$ consisting of functions with mean value zero.

The space $H(\mathrm{div};\Omega)$ is defined as the set of vector-valued functions on Ω which, together with their divergence, are square integrable; i.e.,

$$H(\mathrm{div};\Omega) = \left\{\mathbf{v} :\ \mathbf{v} \in (L^2(\Omega))^2, \nabla\cdot\mathbf{v} \in L^2(\Omega)\right\}.$$

The norm in $H(\mathrm{div};\Omega)$ is defined by

$$\|\mathbf{v}\|_{H(\mathrm{div};\Omega)} = \left(\|\mathbf{v}\|^2 + \|\nabla\cdot\mathbf{v}\|^2\right)^{\frac{1}{2}}.$$

Let $K \subset \Omega$ be a triangle or quadrilateral. For any smooth vector-valued functions $\mathbf{w}$ and $\mathbf{v}$, it follows from the divergence theorem that

$$\int_K (-\Delta\mathbf{w})\cdot\mathbf{v}dK = (\nabla\mathbf{w},\nabla\mathbf{v})_K - \int_{\partial K}\frac{\partial\mathbf{w}}{\partial\mathbf{n}_K}\cdot\mathbf{v}\ ds, \tag{2.1}$$

where ds represents the boundary element, $\mathbf{n}_K$ is the outward normal direction on ∂K, and

$$(\nabla\mathbf{w},\nabla\mathbf{v})_K = \sum_{i,j=1}^{2}\int_K \frac{\partial w_i}{\partial x_j}\frac{\partial v_i}{\partial x_j}dK.$$

Let τ_K be the tangential direction to ∂K so that $\mathbf{n}_K$ and τ_K form a right-hand coordinate system. It follows from the representation

$$\mathbf{v} = (\mathbf{v}\cdot\mathbf{n}_K)\mathbf{n}_K + (\mathbf{v}\cdot\tau_K)\tau_K$$

that

$$\frac{\partial\mathbf{w}}{\partial\mathbf{n}_K}\cdot\mathbf{v} = \frac{\partial(\mathbf{w}\cdot\mathbf{n}_K)}{\partial\mathbf{n}_K}(\mathbf{v}\cdot\mathbf{n}_K) + \frac{\partial(\mathbf{w}\cdot\tau_K)}{\partial\mathbf{n}_K}(\mathbf{v}\cdot\tau_K). \tag{2.2}$$

3. A variational formula. For simplicity, we let $\nu = 1$ for the fluid viscosity in the Stokes equation (1.1). Furthermore, we assume that Ω is a plane polygonal domain without cracks.

Let $\mathcal{T}_h$ be a finite element partition of the domain Ω with mesh size h. Assume that the partition $\mathcal{T}_h$ is shape regular so that the routine inverse inequality in finite elements holds true (see [9]). Define the finite element spaces V_h and W_h for the velocity and pressure variables, respectively, by

$$V_h = \{\mathbf{v} \in H(\mathrm{div};\Omega) :\ \mathbf{v}|_K \in V_r(K) \quad \forall K \in \mathcal{T}_h;\ \mathbf{v}\cdot\mathbf{n}|_{\partial\Omega} = 0\}$$

$$W_h = \{q \in L_0^2(\Omega) :\ q|_K \in W_m(K) \quad \forall K \in \mathcal{T}_h\},$$

where $\mathbf{n}$ is the outward normal direction on the boundary of Ω, $V_r(K)$ is a space of vector-valued polynomials on the element K with index $r \geq 1$, and $W_m(K)$ is a set of polynomials on the element K with index $m \geq 0$. Examples of $V_r(K)$ and $W_m(K)$ will be given in section 6.

To derive a weak formulation, we multiply the equation (1.1) by any $\mathbf{v} \in V_h$ and use (2.1) to obtain

$$\sum_{K\in\mathcal{T}_h}\left((\nabla\mathbf{u},\nabla\mathbf{v})_K - \int_{\partial K}\frac{\partial\mathbf{u}}{\partial\mathbf{n}_K}\cdot\mathbf{v}\ ds\right) - (p,\nabla\cdot\mathbf{v}) = (\mathbf{f},\mathbf{v}), \tag{3.1}$$

where we have also used the integration by parts to deduce

$$\int_\Omega \nabla p \cdot \mathbf{v} d\Omega = -(p, \nabla \cdot \mathbf{v}).$$

The fact that $\mathbf{v} \in V_h$ implies that $\mathbf{v} \cdot \mathbf{n}_K$ is continuous across each interior boundary. Thus, it follows from (2.2) that

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \frac{\partial \mathbf{u}}{\partial \mathbf{n}_K} \cdot \mathbf{v}\ ds = \sum_{K\in\mathcal{T}_h} \int_{\partial K} \frac{\partial (\mathbf{u}\cdot\tau_K)}{\partial \mathbf{n}_K}\ \mathbf{v}\cdot\tau_K\ ds. \tag{3.2}$$

Introduce the following notation:

$$(\nabla_h \mathbf{u}, \nabla_h \mathbf{v}) = \sum_{K\in\mathcal{T}_h} (\nabla \mathbf{u}, \nabla \mathbf{v})_K.$$

By substituting (3.2) into (3.1) we obtain

$$(\nabla_h \mathbf{u}, \nabla_h \mathbf{v}) - (p, \nabla\cdot\mathbf{v}) - \sum_{K\in\mathcal{T}_h} \int_{\partial K} \frac{\partial (\mathbf{u}\cdot\tau_K)}{\partial \mathbf{n}_K}\ \mathbf{v}\cdot\tau_K ds = (\mathbf{f}, \mathbf{v}), \tag{3.3}$$

which is the basis of our first equation in the new variational form. Our second equation can be derived from testing (1.2) against any $q \in W_h$, yielding

$$(\nabla\cdot\mathbf{u}, q) = 0. \tag{3.4}$$

We now reformulate the boundary integrals in (3.3). Let e be an interior edge shared by two elements K_1 and K_2, and let $\mathbf{n}_1$ and $\mathbf{n}_2$ be unit normal vectors on e pointing exterior to K_1 and K_2, respectively. Denote by τ_1 and τ_2 the two tangential directions which make the right-hand coordinate systems with $\mathbf{n}_1$ and $\mathbf{n}_2$, respectively. We define the average $\{\!\{\cdot\}\!\}$ and jump $[\![\cdot]\!]$ on e for vector-valued functions $\mathbf{w}$ as follows:

$$\begin{aligned} \{\!\{\varepsilon(\mathbf{w})\}\!\} &= \frac{1}{2}\left(\mathbf{n}_1\cdot\nabla(\mathbf{w}\cdot\tau_1)|_{\partial K_1} + \mathbf{n}_2\cdot\nabla(\mathbf{w}\cdot\tau_2)|_{\partial K_2}\right), \\ [\![\mathbf{w}]\!] &= \mathbf{w}|_{\partial K_1}\cdot\tau_1 + \mathbf{w}|_{\partial K_2}\cdot\tau_2. \end{aligned}$$

For boundary edge $e = \partial K_1 \cap \partial\Omega$, the above two operations must be modified by

$$\{\!\{\varepsilon(\mathbf{w})\}\!\} = \mathbf{n}_1\cdot\nabla(\mathbf{w}\cdot\tau_1)|_{\partial K_1}, \quad [\![\mathbf{w}]\!] = \mathbf{w}|_{\partial K_1}\cdot\tau_1.$$

Let $\mathcal{E}_h$ denote the union of the boundaries of all elements K in $\mathcal{T}_h$. For sufficiently smooth $\mathbf{u}$ (e.g., $\mathbf{u} \in H^{\frac{3}{2}+\epsilon}(\Omega)$ for some $\epsilon > 0$), it is not hard to see that

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \frac{\partial (\mathbf{u}\cdot\tau_K)}{\partial \mathbf{n}_K}\ \mathbf{v}\cdot\tau_K ds = \sum_{e\in\mathcal{E}_h} \int_e \{\!\{\varepsilon(\mathbf{u})\}\!\}[\![\mathbf{v}]\!] ds.$$

Substituting the above into (3.3) we obtain

$$(\nabla_h \mathbf{u}, \nabla_h \mathbf{v}) - (\nabla\cdot\mathbf{v}, p) - \sum_{e\in\mathcal{E}_h} \int_e \{\!\{\varepsilon(\mathbf{u})\}\!\}[\![\mathbf{v}]\!] ds = (\mathbf{f}, \mathbf{v}). \tag{3.5}$$

Let $V(h) = V_h + (H^s(\Omega)\cap H_0^1(\Omega))^2$, with $s > \frac{3}{2}$. Denote by

$$a_o(\mathbf{u}, \mathbf{v}) = (\nabla_h \mathbf{u}, \nabla_h \mathbf{v}) - \sum_{e\in\mathcal{E}_h} \int_e \{\!\{\varepsilon(\mathbf{u})\}\!\}[\![\mathbf{v}]\!] ds$$

and

$$b(\mathbf{v}, q) = (\nabla \cdot \mathbf{v}, q)$$

two bilinear forms on $V(h) \times V(h)$ and $V(h) \times L_0^2(\Omega)$. With the conditions specified in this paper, it can be proved that the exact solution $(\mathbf{u}; p)$ of the Stokes problem in 2D belongs to $V(h)$ for some $s > \frac{3}{2}$. Readers are referred to [20, 15, 14, 23] for details. As a result, it follows from (3.5) and (3.4) that the exact solution of the 2D Stokes problem satisfies the following variational equations:

$$a_o(\mathbf{u}, \mathbf{v}) - b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in V_h, \tag{3.6}$$

$$b(\mathbf{u}, q) = 0 \qquad \forall q \in W_h. \tag{3.7}$$

However, it is not clear if the same statement can be made for the Stokes problem in three-dimensional space without assuming a smooth boundary $\partial\Omega$ or a convex polyhedral domain Ω [16, 18].

4. Finite element schemes. Our goal of this section is to propose two finite element schemes based on two modifications of the weak formulation (3.6)–(3.7) for the Stokes problem (1.1)–(1.3). To this end, let us introduce a symmetric bilinear form on $V(h) \times V(h)$ as follows:

$$a_s(\mathbf{w}, \mathbf{v}) = a_o(\mathbf{w}, \mathbf{v}) + \sum_{e \in \mathcal{E}_h} \int_e \left(\alpha h_e^{-1} [\![\mathbf{w}]\!][\![\mathbf{v}]\!] - \{\!\{\varepsilon(\mathbf{v})\}\!\}[\![\mathbf{w}]\!] \right) ds,$$

where $\alpha > 0$ is a parameter to be determined later, and h_e is the length of the edge e. For the exact solution $(\mathbf{u}; p)$ of the Stokes problem, we clearly have

$$a_s(\mathbf{u}, \mathbf{v}) = a_o(\mathbf{u}, \mathbf{v}) \qquad \forall \mathbf{v} \in V_h.$$

Therefore, it follows from (3.6) and (3.7) that

$$a_s(\mathbf{u}, \mathbf{v}) - b(\mathbf{v}, p) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in V_h, \tag{4.1}$$

$$b(\mathbf{u}, q) = 0 \qquad \forall q \in W_h. \tag{4.2}$$

The corresponding finite element scheme for (1.1)–(1.3) seeks $(\mathbf{u}_h; p_h) \in V_h \times W_h$ such that

$$a_s(\mathbf{u}_h, \mathbf{v}) - b(\mathbf{v}, p_h) = (\mathbf{f}, \mathbf{v}) \quad \forall \mathbf{v} \in V_h, \tag{4.3}$$

$$b(\mathbf{u}_h, q) = 0 \qquad \forall q \in W_h. \tag{4.4}$$

To investigate the properties of the above numerical scheme, we introduce two norms $|\!|\!| \cdot |\!|\!|_1$ and $|\!|\!| \cdot |\!|\!|$ for the set $V(h)$ as follows:

$$|\!|\!|\mathbf{v}|\!|\!|_1^2 = |\mathbf{v}|_{1,h}^2 + \sum_{e \in \mathcal{E}_h} h_e^{-1} \|[\![\mathbf{v}]\!]\|_e^2, \tag{4.5}$$

$$|\!|\!|\mathbf{v}|\!|\!|^2 = |\!|\!|\mathbf{v}|\!|\!|_1^2 + \sum_{e \in \mathcal{E}_h} h_e \|\{\!\{\varepsilon(\mathbf{v})\}\!\}\|_e^2, \tag{4.6}$$

where $|\mathbf{v}|_{1,h}^2 = \sum_{K \in \mathcal{T}_h} |\mathbf{v}|_{1,K}^2$ and $\|\mathbf{v}\|_e^2 = \displaystyle\int_e \mathbf{v} \cdot \mathbf{v} ds$.

Let K be an element with e as an edge. It is well known that there exists a constant C such that for any function $g \in H^1(K)$

$$\|g\|_e^2 \leq C\left(h_K^{-1}\|g\|_K^2 + h_K\|\nabla g\|_K^2\right). \tag{4.7}$$

In particular, for any $\mathbf{v} \in V_h$, we have

$$h_e\|\{\!\{\varepsilon(\mathbf{v})\}\!\}\|_e^2 \leq C\left(\|\nabla \mathbf{v}\|_K^2 + h_K^2\|\nabla^2\mathbf{v}\|_K^2\right). \tag{4.8}$$

The standard inverse inequality can be employed to the last term of the above inequality, yielding

$$h_e\|\{\!\{\varepsilon(\mathbf{v})\}\!\}\|_e^2 \leq C\|\nabla \mathbf{v}\|_K^2 \tag{4.9}$$

for some constant C independent of the mesh size h. Consequently, there is a constant C such that

$$|\!|\!|\mathbf{v}|\!|\!| \leq C_0|\!|\!|\mathbf{v}|\!|\!|_1 \qquad \forall \mathbf{v} \in V_h. \tag{4.10}$$

The following result is concerned with the ellipticity of the bilinear form $a_s(\cdot,\cdot)$ in $V_h \times V_h$.

LEMMA 4.1. *There exists a constant α_0 independent of h such that for any $\mathbf{v} \in V_h$ we have*

$$a_s(\mathbf{v},\mathbf{v}) \geq \alpha_0|\!|\!|\mathbf{v}|\!|\!|^2, \tag{4.11}$$

provided that α is sufficiently large.

Proof. It follows from the Cauchy–Schwarz inequality that there is a constant C such that

$$\begin{aligned}
\left|\sum_{e\in\mathcal{E}_h}\int_e \{\!\{\varepsilon(\mathbf{w})\}\!\}[\![\mathbf{v}]\!]ds\right| &\leq C\left(\sum_{e\in\mathcal{E}_h} h_e\|\{\!\{\varepsilon(\mathbf{w})\}\!\}\|_e^2\right)^{\frac{1}{2}}\left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2\right)^{\frac{1}{2}}\\
&\leq C|\mathbf{w}|_{1,h}\left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2\right)^{\frac{1}{2}}\\
&\leq \frac{1}{2}|\mathbf{w}|_{1,h}^2 + C\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2,
\end{aligned}$$

where we have used the estimate (4.9) in the second line. Using the above inequality and (4.10), we obtain

$$\begin{aligned}
a_s(\mathbf{v},\mathbf{v}) &= (\nabla_h\mathbf{v},\nabla_h\mathbf{v}) + \alpha\sum_{e\in\mathcal{E}_h} h_e^{-1}\int_e [\![\mathbf{v}]\!]^2 ds - 2\sum_{e\in\mathcal{E}_h}\int_e \{\!\{\varepsilon(\mathbf{v})\}\!\}[\![\mathbf{v}]\!]ds\\
&\geq |\mathbf{v}|_{1,h}^2 + \alpha\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2 - \frac{1}{2}|\mathbf{v}|_{1,h}^2 - C\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2\\
&= \frac{1}{2}|\mathbf{v}|_{1,h}^2 + (\alpha - C)\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2 \geq \alpha_1|\!|\!|\mathbf{v}|\!|\!|_1^2 \geq \alpha_0|\!|\!|\mathbf{v}|\!|\!|^2,
\end{aligned}$$

with $\alpha_1 = \min(\frac{1}{2}, \alpha - C)$ and $\alpha_0 = \alpha_1/C_0$. For example, one may have $\alpha_0 = 1/(2C_0)$ if the parameter α is chosen so that $\alpha \geq C + \frac{1}{2}$. ∎

In the rest of the paper, we assume that the parameter α is chosen so that (4.11) holds true for the symmetric bilinear from $a_s(\cdot,\cdot)$. The proof of Lemma 4.1 indicates that the value of α depends upon the constant in the inverse inequality for finite element functions. Therefore, the value of α for which $a_s(\cdot,\cdot)$ is coercive is mesh-dependent. Existing results for saddle-point problems indicate that it is theoretically and computationally important to have the coercivity (4.11). Therefore, the mesh dependence of the parameter α makes the finite element scheme (4.3)–(4.4) conditionally interesting in practical computation.

To overcome the difficulty on the parameter selection, we introduce a second finite element scheme which is parameter-insensitive. To this end, we define a nonsymmetric bilinear form on $V(h)\times V(h)$ as follows:

$$a_{ns}(\mathbf{w},\mathbf{v}) = a_o(\mathbf{w},\mathbf{v}) + \sum_{e\in\mathcal{E}_h}\int_e \left(\alpha h_e^{-1}[\![\mathbf{w}]\!][\![\mathbf{v}]\!] + \{\!\{\varepsilon(\mathbf{v})\}\!\}[\![\mathbf{w}]\!]\right) ds.$$

Similar to the bilinear form $a_s(\cdot,\cdot)$, for the exact solution $(\mathbf{u};p)$ of the Stokes problem we have

$$a_{ns}(\mathbf{u},\mathbf{v}) = a_o(\mathbf{u},\mathbf{v}) \qquad \forall \mathbf{v}\in V_h.$$

Consequently, the solution of the Stokes problem satisfies the following variational equations:

$$a_{ns}(\mathbf{u},\mathbf{v}) - b(\mathbf{v},p) = (\mathbf{f},\mathbf{v}) \quad \forall \mathbf{v}\in V_h, \tag{4.12}$$

$$b(\mathbf{u},q) = 0 \qquad \forall q\in W_h. \tag{4.13}$$

Our second finite element scheme for (1.1)–(1.3) seeks $(\mathbf{u}_h;p_h)\in V_h\times W_h$ such that

$$a_{ns}(\mathbf{u}_h,\mathbf{v}) - b(\mathbf{v},p_h) = (\mathbf{f},\mathbf{v}) \quad \forall \mathbf{v}\in V_h, \tag{4.14}$$

$$b(\mathbf{u}_h,q) = 0 \qquad \forall q\in W_h. \tag{4.15}$$

To see the coercivity of the bilinear form $a_{ns}(\cdot,\cdot)$, we use its definition and (4.10) to obtain

$$\begin{aligned} a_{ns}(\mathbf{v},\mathbf{v}) &= (\nabla_h\mathbf{v},\nabla_h\mathbf{v}) + \alpha\sum_{e\in\mathcal{E}_h} h_e^{-1}\int_e [\![\mathbf{v}]\!]^2 ds \\ &\geq \min(1,\alpha)|\!|\!|\mathbf{v}|\!|\!|_1^2 \geq \min(1,\alpha)C_0^{-1}|\!|\!|\mathbf{v}|\!|\!|^2, \end{aligned}$$

where $\mathbf{v}\in V_h$. Thus, the coercivity (4.11) holds true for the bilinear form $a_{ns}(\cdot,\cdot)$ with any value of $\alpha>0$.

The following is a result on the boundedness of the bilinear forms $a_s(\cdot,\cdot)$ and $a_{ns}(\cdot,\cdot)$.

LEMMA 4.2. *There exists a constant C independent of h such that*

$$|a_i(\mathbf{w},\mathbf{v})| \leq C|\!|\!|\mathbf{w}|\!|\!||\!|\!|\mathbf{v}|\!|\!| \quad \forall \mathbf{w},\mathbf{v}\in V(h), \tag{4.16}$$

where $i=s,ns$.

Proof. For simplicity, we shall present the analysis for $a_s(\cdot,\cdot)$ only. By the definition of $a_s(\mathbf{w},\mathbf{v})$ and the Schwarz inequality, there exists a constant C such that

$$\begin{aligned}|a_s(\mathbf{w},\mathbf{v})| \le C \Bigg\{ & |\mathbf{w}|_{1,h}|\mathbf{v}|_{1,h} + \left(\sum_{e\in\mathcal{E}_h} h_e \|\{\!\{\varepsilon(\mathbf{w})\}\!\}\|_e^2\right)^{\frac12} \left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2\right)^{\frac12} \\ & + \left(\sum_{e\in\mathcal{E}_h} h_e \|\{\!\{\varepsilon(\mathbf{v})\}\!\}\|_e^2\right)^{\frac12} \left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{w}]\!]\|_e^2\right)^{\frac12} \\ & + \alpha \left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{w}]\!]\|_e^2\right)^{\frac12} \left(\sum_{e\in\mathcal{E}_h} h_e^{-1}\|[\![\mathbf{v}]\!]\|_e^2\right)^{\frac12} \Bigg\} \\ & \le C|||\mathbf{w}|||\,|||\mathbf{v}|||,\end{aligned}$$

which proves the desired boundedness. ∎

5. Error estimates. The first goal of this section is to derive an optimal-order error estimate for the pressure in $L^2(\Omega)$ and the velocity in the norm given by (4.6). The second goal is to derive an optimal-order error estimate for the velocity approximation in the L^2-norm for the symmetric scheme (4.3)–(4.4).

Assumption 1. *There exists an operator* $\Pi_h : (H_0^1(\Omega))^2 \to V_h$ *such that*

$$b(\mathbf{v}-\Pi_h\mathbf{v}, q) = 0 \qquad \forall q \in W_h. \tag{5.1}$$

In addition, the operator Π_h *is assumed to satisfy the following:*

$$|\mathbf{v}-\Pi_h\mathbf{v}|_{s,K} \le Ch^{t-s}|\mathbf{v}|_{t,K} \quad \forall K\in\mathcal{T}_h,\ \ s=0,1, \tag{5.2}$$

where the constant C *depends only on the shape of* K *and* $1\le t\le r+1$.

From (5.2) and the inequality (4.7) it is not hard to see that

$$|||\mathbf{v}-\Pi_h\mathbf{v}|||_1 \le C\|\mathbf{v}\|_1 \quad \forall \mathbf{v}\in(H_0^1(\Omega))^2.$$

Thus, it follows from $|||\mathbf{v}|||_1 = |\mathbf{v}|_1 \le \|\mathbf{v}\|_1$ and the triangle inequality that

$$|||\Pi_h\mathbf{v}|||_1 \le C\|\mathbf{v}\|_1. \tag{5.3}$$

For our finite element formulations, the inf-sup condition given in Brezzi's framework would read as follows: There exists a positive constant β, independent of h, such that

$$\sup_{\mathbf{v}\in V_h} \frac{b(\mathbf{v},q)}{|||\mathbf{v}|||} \ge \beta\|q\| \quad \forall q\in W_h. \tag{5.4}$$

To verify (5.4), we first use the operator Π_h to obtain

$$\sup_{\mathbf{v}\in V_h} \frac{b(\mathbf{v},q)}{|||\mathbf{v}|||} \ge \sup_{\mathbf{v}\in(H_0^1(\Omega))^2} \frac{b(\Pi_h\mathbf{v},q)}{|||\Pi_h\mathbf{v}|||} = \sup_{\mathbf{v}\in(H_0^1(\Omega))^2} \frac{b(\mathbf{v},q)}{|||\Pi_h\mathbf{v}|||}. \tag{5.5}$$

Observe that by using (5.3), and (4.10), we have for all $\mathbf{v}\in(H_0^1(\Omega))^2$

$$|||\Pi_h\mathbf{v}||| \le C|||\Pi_h\mathbf{v}|||_1 \le C\|\mathbf{v}\|_1. \tag{5.6}$$

Thus, substituting (5.6) into the inequality (5.5) gives

$$\sup_{\mathbf{v}\in V_h} \frac{b(\mathbf{v},q)}{|||\mathbf{v}|||} \ge C^{-1} \sup_{\mathbf{v}\in(H_0^1(\Omega))^2} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_1} \ge \beta\|q\|,$$

where we have used the inf-sup condition for the continuous case [19, 8].

5.1. Error estimates in $H^1 \times L^2$. The error analysis requires the L^2 projection from $L_0^2(\Omega)$ to the finite element space W_h, which is denoted by Q_h. In addition, the following error equations turn out to be useful:

$$a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{v}) - b(\mathbf{v}, p - p_h) = 0 \qquad \forall \mathbf{v} \in V_h, \tag{5.7}$$

$$b(\mathbf{u} - \mathbf{u}_h, q) = 0 \qquad \forall q \in W_h. \tag{5.8}$$

These error equations can be obtained by subtracting (4.3)–(4.4) from (4.1)–(4.2). Similar error equations hold true for the nonsymmetric scheme (4.14)–(4.15) with $a_s(\cdot,\cdot)$ being replaced by $a_{ns}(\cdot,\cdot)$. Now we are in a position to present an error estimate for the new finite element approximations.

THEOREM 5.1. *Let* $(\mathbf{u}; p)$ *be the solution of* (1.1)–(1.3) *and* $(\mathbf{u}_h; p_h) \in V_h \times W_h$ *be obtained from either* (4.3)–(4.4) *or* (4.14)–(4.15). *Assume that Assumption* 1 *holds true. Then, there exists a constant* C *independent of* h *such that*

$$|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\| \le C \left(|||\mathbf{u} - \Pi_h \mathbf{u}||| + \|p - Q_h p\| \right). \tag{5.9}$$

Proof. Let

$$\xi_h = \mathbf{u}_h - \Pi_h \mathbf{u}, \qquad \eta_h = p_h - Q_h p$$

be the error between the finite element solution $(\mathbf{u}_h; p_h)$ and the projection $(\Pi_h \mathbf{u}; Q_h p)$ of the exact solution. Denote by

$$\xi = \mathbf{u} - \Pi_h \mathbf{u}, \qquad \eta = p - Q_h p$$

the error between the exact solution $(\mathbf{u}; p)$ and its projection. It follows from the error equations (5.7) and (5.8) that

$$a(\xi_h, \mathbf{v}) - b(\mathbf{v}, \eta_h) = a(\xi, \mathbf{v}) - b(\mathbf{v}, \eta), \tag{5.10}$$

$$b(\xi_h, q) = b(\xi, q) = 0 \tag{5.11}$$

for any $\mathbf{v} \in V_h$ and $q \in W_h$. Here and in what follows of this section, $a(\cdot,\cdot)$ denotes either $a_s(\cdot,\cdot)$ or $a_{ns}(\cdot,\cdot)$.

By letting $\mathbf{v} = \xi_h$ in (5.10) and $q = \eta_h$ in (5.11), the sum of (5.10) and (5.11) gives

$$a(\xi_h, \xi_h) = a(\xi, \xi_h) - b(\xi_h, \eta).$$

Thus, it follows from the coercivity (4.11) and the boundedness (4.16) that

$$\alpha_0 |||\xi_h|||^2 \le C(|||\xi||| |||\xi_h||| + \|\eta\| |||\xi_h|||),$$

which implies the following:

$$|||\xi_h||| \le C \left(|||\xi||| + \|\eta\| \right).$$

The above estimate can be rewritten as

$$|||\mathbf{u}_h - \Pi_h \mathbf{u}||| \le C \left(|||\mathbf{u} - \Pi_h \mathbf{u}||| + \|p - Q_h p\| \right). \tag{5.12}$$

Now using the triangle inequality and the error estimate (5.12) we get

$$|||\mathbf{u} - \mathbf{u}_h||| \le C \left(|||\mathbf{u} - \Pi_h \mathbf{u}||| + \|p - Q_h p\| \right), \tag{5.13}$$

which completes the estimate for the velocity approximation.

It remains to estimate the pressure approximation p_h. To this end, we use the discrete inf-sup condition (5.4) to obtain

$$\begin{aligned}
\|p_h - Q_h p\| &\le \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{b(\mathbf{v}, p_h - Q_h p)}{|||\mathbf{v}|||} \\
&= \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{b(\mathbf{v}, p_h - p) + b(\mathbf{v}, p - Q_h p)}{|||\mathbf{v}|||} \\
&= \frac{1}{\beta} \sup_{\mathbf{v}\in V_h} \frac{a(\mathbf{u} - \mathbf{u}_h, \mathbf{v}) + b(\mathbf{v}, p - Q_h p)}{|||\mathbf{v}|||} \\
&\le C \sup_{\mathbf{v}\in V_h} \frac{1}{|||\mathbf{v}|||} |||\mathbf{v}|||(|||\mathbf{u} - \mathbf{u}_h||| + \|p - Q_h p\|) \\
&\le C(|||\mathbf{u} - \mathbf{u}_h||| + \|p - Q_h p\|),
\end{aligned}$$

which, together with (5.13), implies that

$$\|p_h - Q_h p\| \le C\left(|||\mathbf{u} - \Pi_h \mathbf{u}||| + \|p - Q_h p\|\right).$$

The error estimate for the pressure approximation is then completed by combining the above inequality with the standard triangle inequality. ∎

5.2. An L^2-error estimate for the velocity approximation. Consider only the finite element approximate solutions arising from the symmetric finite element scheme. To derive an L^2-error estimate for the velocity approximation, we seek $(\mathbf{w}; \lambda) \in (H_0^1(\Omega))^2 \times L_0^2(\Omega)$ satisfying

$$\begin{aligned}
-\Delta \mathbf{w} + \nabla \lambda &= \mathbf{u} - \mathbf{u}_h \qquad \text{in } \Omega, \\
\nabla \cdot \mathbf{w} &= 0 \qquad \text{in } \Omega, \\
\mathbf{w} &= 0 \qquad \text{on } \partial\Omega.
\end{aligned}$$

Note that for any $(\mathbf{v}; q) \in V(h) \times L_0^2(\Omega)$ the solution $(\mathbf{w}; \lambda)$ satisfies

$$a_s(\mathbf{w}, \mathbf{v}) - b(\mathbf{v}, \lambda) = (\mathbf{u} - \mathbf{u}_h, \mathbf{v}), \tag{5.14}$$

$$b(\mathbf{w}, q) = 0. \tag{5.15}$$

Assume that the Stokes problem has the $H^2(\Omega) \times H^1(\Omega)$-regularity property in the sense that the solution $(\mathbf{w}; \lambda) \in (H^2(\Omega))^2 \times H^1(\Omega)$ and the following a priori estimate holds true:

$$\|\mathbf{w}\|_2 + \|\lambda\|_1 \le C\|\mathbf{u} - \mathbf{u}_h\|. \tag{5.16}$$

In addition, we assume that the finite element space V_h and the projection operator Π_h have the following property:

$$|||\mathbf{w} - \Pi_h \mathbf{w}||| \le Ch\|\mathbf{w}\|_2. \tag{5.17}$$

With these assumptions, it is not hard to see that there exists a constant C independent of h such that

$$|||\mathbf{w} - \Pi_h \mathbf{w}||| + \|\lambda - Q_h \lambda\| \le Ch\|\mathbf{u} - \mathbf{u}_h\|. \tag{5.18}$$

It must be pointed out that the $H^2 \times H^1$-regularity property assumption stated as above requires that the polygonal domain Ω be convex. For nonconvex but smooth

domains, the regularity (5.16) can be proved to be valid. However, isoparametric elements would need to be employed in the finite element scheme in order to maintain optimal-order error estimates in either H^1- or L^2-norms.

THEOREM 5.2. *Let* $(\mathbf{u}_h; p_h) \in V_h \times W_h$ *and* $(\mathbf{u}; p)$ *be the solutions of* (4.3)–(4.4) *and* (1.1)–(1.3)*, respectively. Assume that Assumption* 1 *and the estimate* (5.17) *hold true and that the Stokes problem* (1.1)–(1.3) *has the* $H^2(\Omega) \times H^1(\Omega)$*-regularity property. Then there exists a constant* C *independent of* h *such that*

$$\|\mathbf{u} - \mathbf{u}_h\| \le Ch(|||\mathbf{u} - \Pi_h\mathbf{u}||| + \|p - Q_h p\|). \tag{5.19}$$

Proof. By letting $\mathbf{v} = \mathbf{u} - \mathbf{u}_h$ in (5.14) we arrive at

$$a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{w}) - b(\mathbf{u} - \mathbf{u}_h, \lambda) = \|\mathbf{u} - \mathbf{u}_h\|^2. \tag{5.20}$$

Notice that

$$b(\mathbf{u} - \mathbf{u}_h, \lambda) = b(\mathbf{u} - \mathbf{u}_h, \lambda - Q_h\lambda) \tag{5.21}$$

and

$$\begin{aligned} a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{w}) &= a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{w} - \Pi_h\mathbf{w}) + a_s(\mathbf{u} - \mathbf{u}_h, \Pi_h\mathbf{w}) \\ &= a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{w} - \Pi_h\mathbf{w}) + b(\Pi_h\mathbf{w}, p - p_h) \\ &= a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{w} - \Pi_h\mathbf{w}) + b(\Pi_h\mathbf{w} - \mathbf{w}, p - p_h). \end{aligned} \tag{5.22}$$

Substituting (5.21) and (5.22) into (5.20) we obtain

$$\|\mathbf{u} - \mathbf{u}_h\|^2 = a_s(\mathbf{u} - \mathbf{u}_h, \mathbf{w} - \Pi_h\mathbf{w}) + b(\Pi_h\mathbf{w} - \mathbf{w}, p - p_h) - b(\mathbf{u} - \mathbf{u}_h, \lambda - Q_h\lambda).$$

Thus,

$$\|\mathbf{u} - \mathbf{u}_h\|^2 \le C\left(|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\|\right)\ \left(|||\mathbf{w} - \Pi_h\mathbf{w}||| + \|\lambda - Q_h\lambda\|\right).$$

Substituting (5.18) into the above estimate we obtain

$$\|\mathbf{u} - \mathbf{u}_h\|^2 \le Ch\left(|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\|\right)\|\mathbf{u} - \mathbf{u}_h\|,$$

which implies that

$$\|\mathbf{u} - \mathbf{u}_h\| \le Ch\left(|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\|\right).$$

The above inequality and the error estimate (5.9) imply

$$\|\mathbf{u} - \mathbf{u}_h\| \le Ch(|||\mathbf{u} - \Pi_h\mathbf{u}||| + \|p - Q_h p\|).$$

This completes the proof of the theorem. ∎

6. Examples of $H(\mathrm{div})$ elements. Let us recall that the error estimates established in section 5 are based on the following three properties:

B1. $V_h \subset H(\mathrm{div}; \Omega)$,

B2. Assumption 1 and the estimate (5.17) as described in section 5, and

B3. the $H^2 \times H^1$-regularity property assumption for the Stokes problem.

The last property (B3) is required only for the L^2-error estimate for the velocity approximation. This means that any finite element pair $V_h \times W_h$ satisfying properties B1–B2 is applicable for the formulations presented earlier in this manuscript.

Denote by $P_k(K)$ the space of polynomials of degree $\leq k$ and

$$P_{k_1,k_2}(K) = \left\{ p(x_1,x_2) : \quad p(x_1,x_2) = \sum_{0\leq i\leq k_1, 0\leq j\leq k_2} a_{ij} x_1^i x_2^j \right\}.$$

$P_{k_1,k_2,k_3}(K)$ is defined similarly in three-dimensional spaces. Define $Q_k(K)$ as follows:

$$Q_k(K) = \begin{cases} P_{k,k}(K) & \text{for } d = 2, \\ P_{k,k,k}(K) & \text{for } d = 3. \end{cases}$$

Observe that the finite element pair $V_h \times W_h$ is constructed from local elements $V_r(K)$ and $W_m(K)$ as described in section 3. Therefore, it suffices to specify the local pair $V_r(K) \times W_m(K)$ for each example to be presented.

6.1. Existing elements. All of the existing $H(\mathrm{div})$ elements designed for the second-order elliptic problems (e.g., see [25, 8, 7, 17, 19]) satisfy properties B1–B2, except the estimate (5.17) for the lowest-order Raviart–Thomas element on triangles and quadrilaterals. Therefore, there are plenty of finite element spaces applicable to the new formulation of the Stokes problem. For illustrative purposes, we mention three examples. Readers are referred to the book by Brezzi and Fortin [8] for more examples of the $H(\mathrm{div})$ element.

6.1.1. Raviart–Thomas elements on triangles or tetrahedra: $RT_k(K)$. Let $k \geq 1$ be any integer. For any triangular or tetrahedral element K, the local element $V_r(K) \times W_m(K)$ is defined by

$$V_k(K) = (P_k(K))^d \oplus \mathbf{x}P_k(K), \qquad W_k(K) = P_k(K),$$

where $d = 2$ if K is a triangle and $d = 3$ if K is a tetrahedron. The projection operator Π_h satisfying all of the required properties is given locally on each element K. For example, the restriction of Π_h on the element K, denoted by Π_K, is defined as follows:

$$\int_{\partial K} (\mathbf{v} - \Pi_K \mathbf{v}) \cdot \mathbf{n} q ds = 0 \quad \forall q \in P_k(\partial K),$$
$$\int_K (\mathbf{v} - \Pi_K \mathbf{v}) \cdot q dK = 0 \quad \forall q \in (P_{k-1}(K))^d, k \geq 1.$$

6.1.2. BDM elements on triangles or tetrahedra: $BDM_k(K)$ [8]. Let $k \geq 1$ be any integer. For any triangular or tetrahedral element K, the local element $V_r(K) \times W_m(K)$ is defined by setting $r = m + 1 = k$ and

$$V_k(K) = (P_k(K))^d, \quad W_{k-1}(K) = P_{k-1}(K).$$

On a triangular element K, the local projection operator $\Pi_K : (H^1(K))^2 \to V_k(K)$ is defined by

$$\int_{\partial K} (\mathbf{v} - \Pi_K \mathbf{v}) \cdot \mathbf{n} q ds = 0 \quad \forall q \in P_k(\partial K),$$
$$\int_K (\mathbf{v} - \Pi_K \mathbf{v}) \cdot \nabla q dK = 0 \quad \forall q \in P_{k-1}(K),$$
$$\int_K (\mathbf{v} - \Pi_K \mathbf{v}) \cdot \mathrm{curl}(b_K q) dK = 0 \quad \forall q \in P_{k-2}(K), k \geq 2,$$

where b_K is the bubble function defined on K. On a tetrahedral element K, the corresponding local projection Π_K is given by

$$\begin{aligned}\int_{\partial K}(\mathbf{v}-\Pi_K\mathbf{v})\cdot\mathbf{n}qds=0 \quad \forall q\in P_k(\partial K),\\ \int_K(\mathbf{v}-\Pi_K\mathbf{v})\cdot\nabla qdK=0 \quad \forall q\in P_{k-1}(K),\\ \int_K(\mathbf{v}-\Pi_K\mathbf{v})\cdot qdK=0 \quad \forall q\in \Phi_k(K),\end{aligned}$$

where

$$\Phi(K)=\{\phi\in(P_k(K))^3:\quad \nabla\cdot\phi=0,\ \phi\cdot\mathbf{n}=0 \text{ on } \partial K\}.$$

6.1.3. BDM elements on quadrilaterals: $BDM_{[k]}(K)$. It is sufficient to describe the element on the unit square. Let $k\geq 1$ be any integer. The local element $V_r(K)\times W_m(K)$ is defined by

$$\begin{aligned}V_k(K)&=(P_k(K))^2\oplus \operatorname{curl}(x_1^{k+1}x_2)\oplus\operatorname{curl}(x_1x_2^{k+1}),\\ W_{k-1}(K)&=P_{k-1}(K).\end{aligned}$$

On the unit square element K, the local projection operator $\Pi_K:(H^1(K))^2\to V_k(K)$ is defined by

$$\begin{aligned}\int_{\partial K}(\mathbf{v}-\Pi_K\mathbf{v})\cdot\mathbf{n}qds=0 \quad \forall q\in P_k(\partial K),\\ \int_K(\mathbf{v}-\Pi_K\mathbf{v})\cdot\mathbf{w}dK=0 \quad \forall\mathbf{w}\in(P_{k-2}(K))^2, k\geq 2.\end{aligned}$$

6.1.4. Error estimates for the existing elements. Recall that the velocity V_h and the pressure space W_h are defined, respectively, by

$$V_h=\{\mathbf{v}\in H(\operatorname{div};\Omega):\quad \mathbf{v}|_K\in V_r(K)\quad \forall K\in\mathcal{T}_h;\ \mathbf{v}\cdot\mathbf{n}|_{\partial\Omega}=0\} \tag{6.1}$$

and

$$W_h=\{q\in L_0^2(\Omega):\ q|_K\in W_m(K)\quad \forall K\in\mathcal{T}_h\}. \tag{6.2}$$

For the existing $H(\operatorname{div})$ elements listed above, we have $V_r(K)=RT_k(K)$, $BDM_k(K)$, or $BDM_{[k]}(K)$ and $W_m(K)=P_k(K)$, $P_{k-1}(K)$, or $P_{k-1}(K)$, respectively. The projection operator Π_h is given by

$$(\Pi_h\mathbf{v})\big|_K=\Pi_K(\mathbf{v}\big|_K). \tag{6.3}$$

The definition of Π_h implies that

$$b(\mathbf{v}-\Pi_h\mathbf{v},q)=0\quad \forall q\in W_h. \tag{6.4}$$

Furthermore, it has been proved in [8] that (5.2) and (5.17) hold true for Π_h defined in (6.3). Therefore, properties B1–B2 are well justified.

Let Q_h be the L^2 projection from $L_0^2(\Omega)$ to W_h. It is not hard to see that W_h has the following local approximation properties: For $BDM_k(K)$ and $BDM_{[k]}(K)$

$$|p-Q_hp|_{s,K}\leq Ch^{k-s}|p|_{k,K}\quad \forall K\in\mathcal{T}_h,\ \ s=0,1, \tag{6.5}$$

and for $RT_k(K)$

$$|p - Q_h p|_{s,K} \le Ch^{k+1-s}|p|_{k+1,K} \quad \forall K \in \mathcal{T}_h,\ s = 0, 1. \tag{6.6}$$

The constant C in (6.5)–(6.6) depends only on k and the shape of K.

The following result follows from (5.17), (6.5)–(6.6), and Theorems 5.1 and 5.2.

PROPOSITION 6.1. *Let* $(\mathbf{u}; p)$ *be the solution of* (1.1)–(1.3) *and* $(\mathbf{u}_h; p_h) \in V_h \times W_h$ *be obtained from either* (4.3)–(4.4) *or* (4.14)–(4.15). *Assume that* $(\mathbf{u}; p) \in (H^{t+1}(\Omega))^2 \times H^t(\Omega)$ *for some* $1 \le t \le k$. *Then there exists a constant* C *independent of* h *such that for* $BDM_k(K)$, $BDM_{[k]}(K)$, *and* $RT_k(K)$

$$|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\| \le Ch^t(\|\mathbf{u}\|_{t+1} + \|p\|_t). \tag{6.7}$$

Furthermore, if the $H^2 \times H^1$*-regularity property holds true for the Stokes problem, then there is a constant* C *such that the finite element approximation* $(\mathbf{u}_h; p_h)$ *from the symmetric formulation has the following error estimate:*

$$\|\mathbf{u} - \mathbf{u}_h\| \le Ch^{t+1}(\|\mathbf{u}\|_{t+1} + \|p\|_t).$$

We comment that the above error estimates hold true for all of the $H(\mathrm{div})$ elements listed in [8].

6.2. New elements. Stability and accuracy are two main factors in the construction of new finite elements. For the variational schemes presented in this paper, the stability part is realized by a combination of the inf-sup condition and the coercivity for the corresponding bilinear forms. The accuracy part is characterized by Assumption 1 and a balanced pressure space W_h. For example, Proposition 6.1 indicates that the Raviart–Thomas element can be used to approximate the solution of the Stokes equations, which is convergent as the mesh size decreases (note that we do not know any convergence when $RT_0(K)$ is used). However, the Raviart–Thomas elements do not appear to be well balanced, because both the velocity and the pressure unknowns are approximated by polynomials of order k. In contrast, the BDM elements offer a better/optimal combination for the solution of the Stokes problem. But the $BDM_{[k]}(K)$ element on rectangles is constructed in an awkward way by involving the curl of some polynomials. We feel that better constructed elements should be explored on rectangles and cubes for solving the Stokes problem. For this purpose, we would like to propose some alternatives on rectangles and cubes which are suitable for approximating the solution of the Stokes problem. These elements can be used on quadrilaterals through local transformations as described in [26].

6.2.1. A new element on rectangles: $NE1_k(K)$. We illustrate the construction of the new $NE1_k(K)$ element on the unit square $K = [0, 1] \times [0, 1]$. Let $k \ge 1$ be any integer. We define local elements $V_r(K) \times W_m(K)$ by

$$V_k(K) = (Q_k(K))^2, \quad W_{k-1}(K) = Q_{k-1}(K).$$

For the first component of $\mathbf{v} = (v_1, v_2)$, we define an operator $\Pi_{K,1} : H^1(K) \to Q_k(K)$ as follows:

$$\int_e (v_1 - \Pi_{K,1} v_1)\phi ds = 0 \quad \forall \phi \in P_k(e),\ e = \text{west}, \text{east}, \tag{6.8}$$

$$\int_K (v_1 - \Pi_{K,1} v_1)\psi dK = 0 \quad \forall \psi \in P_{k-2,k}(K), \tag{6.9}$$

where $e =$ west means that e is the west edge (i.e., $e = \{(0, x_2) : \ x_2 \in [0, 1]\}$) of the unit square; the east edge is defined accordingly.

The system (6.8) involves exactly $2(k + 1)$ linear equations and (6.9) involves $(k - 1)(k + 1)$ linear equations. The total number of equations is given by

$$2(k + 1) + (k - 1)(k + 1) = (k + 1)^2,$$

which is the same as the total number of degrees of freedom for a polynomial in $Q_k(K)$. The following proposition shows that the linear systems (6.8) and (6.9) uniquely determine the projection $\Pi_{K,1} v_1$.

PROPOSITION 6.2. *Let $v \in Q_k(K)$ be such that*

$$\int_e v\phi ds = 0 \quad \forall \phi \in P_k(e), e = \text{west}, \text{east}, \tag{6.10}$$

$$\int_K v\psi dK = 0 \quad \forall \psi \in P_{k-2,k}(K). \tag{6.11}$$

Then we must have $v \equiv 0$.

Proof. The condition (6.10) implies that $v = 0$ at the east and west edges of the unit square K. Thus, there is a polynomial $g = g(x_1, x_2) \in P_{k-2,k}(K)$ such that $v = x_1(1 - x_1)g$. Substitute $v = x_1(1 - x_1)$ into (6.11), and then let $\psi = g$. It follows that $g \equiv 0$. This shows that $v \equiv 0$. □

The projection of the second component of $\mathbf{v}$, denoted by $\Pi_{K,2} v_2$, can be defined in a similar fashion. The local projection operator is then given by

$$\Pi_K v = (\Pi_{K,1} v_1, \Pi_{K,2} v_2).$$

It is not hard to show that such a defined projection satisfies all of the conditions required in the previous sections. As a result, the element $NE1_k(K)$ can be used to approximate the Stokes problem.

6.2.2. A new element on cubes: $NE2_k(K)$. Again, we shall describe details only on the unit cube $K = [0, 1]^3$. Let $k \geq 1$ be an integer. A straightforward extension of the $NE1_k$ element to three-dimensional space is given by

$$V_k(K) = (Q_k(K))^3, \quad W_{k-1}(K) = Q_{k-1}(K).$$

Our goal here is to show that the above extension actually works. To this end, it suffices to construct a projection operator Π_K which satisfies the required properties.

Let $\mathbf{v} = (v_1, v_2, v_3) \in \left(H^1(\Omega)\right)^3$ be a vector-valued function. For each component v_i, we define its projection to $Q_k(K)$ as follows:

$$\int_{e_i} (v_i - \Pi_{K,i} v_i)\phi ds = 0 \quad \forall \phi \in Q_k(e_i), \tag{6.12}$$

$$\int_K (v_i - \Pi_{K,i} v_i)\psi dK = 0 \quad \forall \psi \in P_{k_1,k_2,k_3}(K), \tag{6.13}$$

where $e_i = \{(x_1, x_2, x_3) : \ x_j \in [0, 1], j \neq i; \ x_i = 0 \text{ or } 1\}$ are the two faces of the cube K which are orthogonal to the x_i-axis, and $k_i = k - 2, k_j = k$ for $j \neq i$.

There are $2(k+1)^2$ linear equations from the condition (6.12) and $(k-1)(k+1)^2$ linear equations from the condition (6.13). The total number of linear equations is then given by

$$2(k + 1)^2 + (k - 1)(k + 1)^2 = (k + 1)^3,$$

which is the same as the total number of degrees of freedom for a polynomial in the space $V_k(K)$.

A similar argument as in the previous subsection for $NE1_k(K)$ can be applied to show that the projection $\Pi_{K,i}v_i$ is uniquely determined by (6.12) and (6.13). Furthermore, the local projection given by

$$\Pi_K \mathbf{v} = (\Pi_{K,1}v_1, \Pi_{K,2}v_2, \Pi_{K,3}v_3)$$

can be verified to satisfy all of the properties required by the convergence theory developed in previous sections for the new finite element methods.

6.2.3. Another new element on rectangles: $NE3_k(K)$. Again for simplicity, we shall describe the new element on the unit square $K = [0,1]^2$. This element will be a simplified version of $NE1_k(K)$ but with the same order of accuracy.

Let $k \geq 1$ be any integer, and define

$$\begin{aligned} V_k(K) &= (P_k(K) \oplus \{x_1 x_2^k\}) \times (P_k(K) \oplus \{x_2 x_1^k\}), \\ W_{k-1}(K) &= P_{k-1}(K). \end{aligned}$$

For the first component of $\mathbf{v} = (v_1, v_2)$, we define its projection $\Pi_{K,1}v_1 \in P_k(K) \oplus \{x_1 x_2^k\}$ by using the following equations:

$$\int_e (v_1 - \Pi_{K,1}v_1)\phi ds = 0 \quad \forall \phi \in P_k(e), e = \text{west}, \text{east}, \tag{6.14}$$

$$\int_K (v_1 - \Pi_{K,1}v_1)\psi dK = 0 \quad \forall \psi \in P_{k-2}(K). \tag{6.15}$$

There are $2(k+1)$ equations from the condition (6.14) and $\frac{1}{2}(k-1)k$ equations from the condition (6.15). The total number of linear equations is given by

$$2(k+1) + \frac{1}{2}(k-1)k = \frac{1}{2}(k+1)(k+2) + 1,$$

which is the same as the total number of degrees of freedom for functions in the space $P_k(K) \oplus \{x_1 x_2^k\}$. Using the same technique as in the analysis for $NE1_k(K)$, it can be proved that $\Pi_{K,1}v_1$ is uniquely determined by (6.14) and (6.15). The projection of the second component of $\mathbf{v}$ can be determined in a similar way. The resulting local projection $\Pi_K \mathbf{v} = (\Pi_{K,1}v_1, \Pi_{K,2}v_2)$ satisfies all of the properties required in the convergence theory.

6.2.4. Error estimates for the new elements. First, we define the velocity space V_h by

$$V_h = \{\mathbf{v} \in H(\text{div}; \Omega) : \ \mathbf{v}|_K \in V_r(K) \quad \forall K \in \mathcal{T}_h; \ \mathbf{v} \cdot \mathbf{n}|_{\partial\Omega} = 0\} \tag{6.16}$$

and the pressure space W_h by

$$W_h = \{q \in L_0^2(\Omega) : \ q|_K \in W_m(K) \quad \forall K \in \mathcal{T}_h\}, \tag{6.17}$$

where $V_r(K) = NE1_k(K), NE2_k(K)$, or $NE3_k(K)$ and $W_m(K) = Q_{k-1}(K), Q_{k-1}(K)$, or $P_{k-1}(K)$, respectively. For any $\mathbf{v} \in (H_0^1(\Omega))^d$, with $d = 2, 3$, define $\Pi_h \mathbf{v} \in V_h$ by

$$(\Pi_h \mathbf{v})|_K = \Pi_K \mathbf{v} \quad \forall\, K \in \mathcal{T}_h, \tag{6.18}$$

where Π_K is the corresponding local projection operator on each element. From the construction of Π_K, it is easy to see that it is indeed true that $\Pi_h \mathbf{v} \in V_h$, and, moreover, one has

$$b(\mathbf{v} - \Pi_h \mathbf{v}, q) = 0 \quad \forall q \in W_h \tag{6.19}$$

and that properties B1–B2 are satisfied for the three new elements $NE1_k(K)$, $NE2_k(K)$, and $NE3_k(K)$. Similar to Proposition 6.1, we have the following convergence estimates.

PROPOSITION 6.3. *Let* $(\mathbf{u}; p)$ *be the solution of* (1.1)–(1.3) *and* $(\mathbf{u}_h; p_h) \in V_h \times W_h$ *be obtained from either* (4.3)–(4.4) *or* (4.14)–(4.15) *by using the new elements described in this subsection. Assume that* $(\mathbf{u}; p) \in (H^{t+1}(\Omega))^d \times H^t(\Omega)$ *for some* $\frac{1}{2} < t \leq k$. *Then there exists a constant* C *independent of* h *such that*

$$|||\mathbf{u} - \mathbf{u}_h||| + \|p - p_h\| \leq Ch^t(\|\mathbf{u}\|_{t+1} + \|p\|_t),$$

and for the symmetric formulation we also have

$$\|\mathbf{u} - \mathbf{u}_h\| \leq Ch^{t+1}(\|\mathbf{u}\|_{t+1} + \|p\|_t),$$

provided that the $H^2 \times H^1$*-regularity property holds true for the Stokes problem.*

We point out that, unlike the existing $H(\mathrm{div})$ elements, the new $H(\mathrm{div})$ elements described in this section do not yield numerical velocities that satisfy the continuity equation (1.4) in the classical sense. However, the numerical approximations arising from the new elements indeed conserve mass locally on each element.

REFERENCES

[1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] I. BABUŠKA, *The finite element method with Lagrangian multiplier*, Numer. Math., 20 (1973), pp. 179–192.

[3] F. BREZZI, *On the existence, uniqueness, and approximation of saddle point problems arising from Lagrangian multipliers*, RAIRO, Anal. Numér., 2 (1974), pp. 129–151.

[4] I. BABUŠKA AND M. ZLÁMAL, *Nonconforming elements in the finite element method with penalty*, SIAM J. Numer. Anal., 10 (1973), pp. 863–875.

[5] F. BASSI AND S. REBAY, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[6] C.E. BAUMANN AND J.T. ODEN, *A discontinuous hp finite element method for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 175 (1999), pp. 311–341.

[7] F. BREZZI, J. DOUGLAS, AND L. MARINI, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.

[8] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Elements*, Springer-Verlag, New York, 1991.

[9] P.G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, New York, 1978.

[10] B. COCKBURN, S. HOU, AND C.-W. SHU, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws* IV: *The multidimensional case*, Math. Comp., 54 (1990), pp. 545–581.

[11] B. COCKBURN, G. KANSCHAT, AND D. SCHOTZAU, *A locally conservative LDG method for the incompressible Navier-Stokes equations*, Math. Comp., 74 (2005), pp. 1067–1095.

[12] B. COCKBURN, G. KANSCHAT, D. SCHÖTZAU, AND C. SCHWAB, *Local discontinuous Galerkin methods for the Stokes system*, SIAM J. Numer. Anal., 40 (2002), pp. 319–343.

[13] B. COCKBURN AND C.-W. SHU, *The local discontinuous Galerkin method for time-dependent convection-diffusion systems*, SIAM J. Numer. Anal., 35 (1998), pp. 2440–2463.

[14] M. COSTABEL AND M. DAUGE, *Crack singularities for general elliptic systems*, Math. Nachr., 235 (2002), pp. 29–49.

[15] M. Dauge, *Elliptic Boundary Value Problems in Corner Domains - Smoothness and Asymptotics of Solutions*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.
[16] M. Dauge, *Stationary Stokes and Navier-Stokes systems on two- or three-dimensional domains with corners. Part* I. *Linearized Equations*, SIAM J. Math. Anal., 20 (1989), pp. 74–97.
[17] J. Douglas and J. Wang, *A new family of spaces in mixed finite element methods for rectangular elements*, Comput. Appl. Math., 12 (1993), pp. 183–197.
[18] V. Girault and J.-L. Lions, *Two-grid finite-element schemes for the steady Navier-Stokes problem in polyhedra*, Port. Math. (N.S.), 58 (2001), pp. 25–57.
[19] V. Girault and P.A. Raviart, *Finite Element Methods for the Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[20] P. Grisvard, *Boundary Value Problems in Non-Smooth Domains*, Pitman, London, 1985.
[21] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows, A Guide to Theory, Practice and Algorithms*, Academic, San Diego, 1989.
[22] P. Hansbo and M. Larson, *Discontinuous Galerkin methods for nearly incompressible elasticity by Nitsche's method*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 1895–1908.
[23] J. R. Kweon, *Corner singularity for compressible Navier-Stokes problem*, in Proceedings of the International Congress of Mathematics, 2002, Higher Education Press, China, 2002.
[24] J. Liu and C. Shu, *A high order discontinuous Galerkin method for* 2D *incompressible flows*, J. Comput. Phys., 160 (2000), pp. 577–596.
[25] P. Raviart and J. Thomas, *A mixed finite element method for second order elliptic problems*, Mathematical Aspects of the Finite Element Method, I. Galligani, E. Magenes, eds., Lectures Notes in Math. 606, Springer-Verlag, New York, 1977.
[26] J. Wang and T. Mathew, *Mixed finite element method over quadrilaterals*, in Proceedings of the Third International Conference on Advances in Numerical Methods and Applications, I. T. Dimov, Bl. Sendov, and P. Vassilevski, eds., World Scientific, River Edge, NJ, 1994, pp. 203–214.
[27] X. Ye, *Discontinuous stable elements for incompressible flow*, Adv. Comput. Math., 20 (2004) pp. 333–345.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1287–1304

CONVERGENCE OF NUMERICAL APPROXIMATIONS OF THE INCOMPRESSIBLE NAVIER–STOKES EQUATIONS WITH VARIABLE DENSITY AND VISCOSITY*

CHUN LIU† AND NOEL J. WALKINGTON‡

Abstract. We consider numerical approximations of incompressible Newtonian fluids having variable, possibly discontinuous, density and viscosity. Since solutions of the equations with variable density and viscosity may not be unique, numerical schemes may not converge. If the solution is unique, then approximate solutions computed using the discontinuous Galerkin method to approximate the convection of the density and stable finite element approximations of the momentum equation converge to the solution. If the solution is not unique, a subsequence of these approximate solutions will converge to a solution.

Key words. Navier–Stokes equations, transport equations, Taylor–Hood approximations

AMS subject classifications. 65M12, 65M60

DOI. 10.1137/050629008

1. Introduction.

We consider numerical approximations of the incompressible Navier–Stokes equations with variable density and viscosity,

$$\begin{aligned} \rho\left(v_t + (v\cdot\nabla)v\right) + \nabla p - \operatorname{div}\left(\mu(\rho)D(v)\right) &= \rho f, \\ \operatorname{div}(v) &= 0, \\ \rho_t + \operatorname{div}(\rho v) &= 0, \end{aligned} \tag{1.1}$$

on a bounded domain $\Omega \subset \mathbb{R}^d$ with initial and boundary conditions

$$v|_{\partial\Omega} = 0, \qquad v|_{t=0} = v_0, \qquad \rho|_{t=0} = \rho_0.$$

These equations model the motion of mixtures of immiscible fluids having different densities and viscosities. The density and viscosity may be discontinuous, so, in general, the solutions will not enjoy any regularity beyond that given by the basic estimates

$$\frac{d}{dt}\int_\Omega (\rho/2)|v|^2\,dx + \int_\Omega \mu(\rho)|D(v)|^2\,dx = \int_\Omega \rho f\cdot v\,dx$$

and $\rho \in L^\infty[0,T;L^\infty(\Omega)]$; in particular, ρ does not have bounded variation. In this situation we can establish convergence of approximate numerical solutions; however, in the absence of additional regularity no rates of convergence can be guaranteed.

The existence of a weak solutions to (1.1) has been established by Lions [15]. Some additional regularity was proven by Antontsev, Kazhikhov, and Monakhov [1] and

*Received by the editors April 11, 2005; accepted for publication (in revised form) November 15, 2006; published electronically June 1, 2007.
http://www.siam.org/journals/sinum/45-3/62900.html

†Department of Mathematics, Pennsylvania State University, State College, PA 16802 (liu@math.psu.edu). This author's work was supported in part by National Science Foundation grant DMS–9972040.

‡Department of Mathematics, Carnegie Mellon University, Pittsburgh, PA 15213 (noelw@cmu.edu). This author's work was supported in part by National Science Foundation grants DMS-0208586 and ITR 0086093. This author's work was also supported by the NSF through the Center for Nonlinear Analysis.

Fujita and Kato [6] when the viscosity is constant and the initial density is bounded from below. In [4], Desjardins provides similar results under weaker assumptions; for instance, when the viscosity does not vary too much with the density. To establish the existence of solutions of the Navier–Stokes equations with discontinuous density and viscosity, sharp results for the convection equation governing the density are required. These were developed by DiPerna and Lions [5], who showed that the weak solutions of convection equations were unique even when the velocity was not Lipschitz, so that characteristics may not exist. They also showed that the solutions would converge strongly when the velocities converged weakly in $L^2[0,T;H^1_0(\Omega)]$. For technical reasons they only considered velocity fields v vanishing[1] on $\partial\Omega$, and for this reason we only consider Dirichlet boundary data for v. Currently, uniqueness of solutions to the coupled system can only be established if the velocity and density satisfy $\nabla v \in L^2[0,T;L^\infty(\Omega)]$, $v_t \in L^2[0,T;L^\infty(\Omega)]$, and $\nabla\rho \in L^2[0,T;L^\infty(\Omega)]$; see [15], so, in general, uniqueness is not expected. In this situation we can only show that subsequences of approximate solutions converge to solutions of the Navier–Stokes equations.

While there is a rich body of literature on numerical approximation of the classical (constant density and viscosity) Navier–Stokes equations, very few results are available for the situation considered here. Algorithms proposed for the approximation of (1.1) include front tracking techniques [7, 8] and level set/phase field methods [2, 16, 17]. Recall that level set methods seek a smooth function ϕ satisfying $\phi_t + \operatorname{div}(\phi v) = 0$ and compute $\rho = H(\phi)$, where $H(.)$ is a suitable translation of the Heaviside graph. Numerical approximations typically approximate the Heaviside graph to give a smooth transition over several grid points. Since ϕ is "smooth," accurate approximations can be computed; however, difficulties arise when attempting to estimate the accuracy of $\rho = H(\phi)$. Indeed, it is difficult to write down the approximate equation satisfied by ρ in this context. For this reason we chose to compute ρ directly using the discontinuous Galerkin method [9, 12]. Below we use the results of Walkington [18], which show that approximations of the density computed using the discontinuous Galerkin method converge strongly in $L^2[0,T;L^2(\Omega)]$. Traditionally the analysis of schemes for hyperbolic equations is based upon the (nonlinear) theory of Kruzkov [10], which requires the coefficients to be C^1. This guarantees that the solutions are regular, in the sense that they have bounded variation, and rates of convergence can be established [11]. This theory fails for the problem considered here since v is not C^1 and ρ does not have bounded variation. This problem was circumvented in [18] by drawing upon the (linear) DiPerna–Lions theory [5]. We refer to [18] for further discussion and references on this topic.

It will be assumed that the viscosity can be determined as a continuous function of the density, $\mu = \mu(\rho)$. Physically each material particle has an associated viscosity, so μ should satisfy the convection equation $\mu_t + v.\nabla\mu = 0$. If $\mu = \mu(\rho)$, then this equation is satisfied when $\rho_t + \operatorname{div}(\rho v) = 0$ and the fluid is incompressible, $\operatorname{div}(v) = 0$. In order to model a mixture of fluids where different components have the same density but different viscosities, the convection equation for μ may be approximated independently. This does not change the analysis below where the major difficulties are due to the coupling between the density and velocity in the convection terms.

1.1. Weak solutions and the energy estimate. Since the solutions of equations (1.1) are not smooth we consider weak solutions. A pair (v,ρ) is a weak solution

[1] Lions and DiPerna also considered the periodic problem and the convection equation on all of $\mathbb{R}^d$.

of (1.1) with initial data $(v_0, \rho_0) \in L^2(\Omega) \times L^\infty(\Omega)$ if

$$v \in V = \{v \in L^\infty[0,T; L^2(\Omega)] \cap L^2[0,T; H_0^1(\Omega)] \mid \operatorname{div}(v) = 0\},$$

$\rho \in \mathcal{R} = L^\infty[0,T; L^\infty(\Omega)]$, and

$$\int_0^T \int_\Omega -\rho v.w_t - (\rho v \otimes v) \cdot \nabla w + \mu(\rho) D(v) \cdot D(w) = \int_\Omega \rho_0 v_0.w(0) + \int_0^T \int_\Omega \rho f.w,$$

$$\int_0^T \int_\Omega \rho(\psi_t + v.\nabla\psi) = \int_\Omega \rho_0 \psi(0), \tag{1.2}$$

for all $w \in \{w \in \mathcal{D}([0,T) \times \Omega) \mid \operatorname{div}(w) = 0\}$ and $\psi \in \mathcal{D}([0,T) \times \Omega)$. DiPerna and Lions [5] and Lions [15] established existence of solutions of this weak problem when ρ_0 is nonnegative. Their weak solutions satisfy the natural energy estimate

$$\frac{d}{dt} \int_\Omega (\rho/2)|v|^2 + \int_\Omega \mu(\rho)|D(v)|^2 \le \int_\Omega \rho f.v, \tag{1.3}$$

which may be derived by formally setting $w = v$ in the weak statement of the momentum equation and $\psi = |v|^2/2$ in the weak statement of the density equation.

1.2. Outline. In the next section we motivate and then state the numerical scheme used to approximate the Navier–Stokes equations with variable density and viscosity (1.1). The requirement of stability, consistency, and nonnegativity of the density, give rise to conflicting requirements. The scheme presented in section 2.4 satisfies these requirements and is subsequently analyzed in section 3.

1.3. Notation. Below, $\Omega \subset \mathbb{R}^d$ will be a bounded domain with unit outward normal n. We will consider a regular family of finite element meshes $\{\mathcal{T}_h\}_{h>0}$, each of which is assumed to triangulate Ω exactly. It is assumed that the finite elements have uniformly bounded aspect ratio, and the parameter $h > 0$ represents the diameter of the largest element in $\mathcal{T}_h$. The space of polynomials of degree k on an element $K \in \mathcal{T}_h$ is denoted $\mathcal{P}_k(K)$. For simplicity we assume that for each $h > 0$ a uniform partition of $[0,T]$ used with $t^n = n\tau$, where $\tau = T/N$, $N \in \mathbb{N}$, is assumed to converge to zero as h tends to zero. We will denote the approximate solutions by (v_h, ρ_h); in particular, the dependence upon τ is implicit. If $a \in \mathbb{R}$, then the positive and negative parts are denoted by $a^\pm$ with $a^+ = \max(a, 0)$ and $a^- = \min(a, 0)$.

Divergences of vectors and matrices are denoted $\operatorname{div}(v) = v_{i,i}$ and $\operatorname{div}(T)_i = T_{ij,j}$, and gradients of vector valued quantities are interpreted as matrices, $(\nabla v)_{ij} = v_{i,j}$. Here indices after the comma represent partial derivatives and the summation convention is used. The symmetric part of the velocity gradient (stretching tensor) is written as $D(v)$. Inner products of vectors v, $w \in \mathbb{R}^d$ are written as $v.w$ and their tensor product $v \otimes w$ is the matrix having components $v_i w_j$. The Frobenius inner product of two matrices A, $B \in \mathbb{R}^{d\times d}$ is denoted by $A \cdot B = \sum_{i,j} A_{ij} B_{ij}$; we frequently use the elementary identities $AB \cdot C = A \cdot CB^T = B \cdot A^T C$.

Standard notation is adopted for the Lebesgue spaces, $L^p(\Omega)$, and the Sobolev spaces, $W^{m,p}(\Omega)$ or $H^m(\Omega)$. The dual exponent to p will be denoted by p', $1/p + 1/p' = 1$. Solutions of the evolution equation will be functions from $[0,T]$ into these spaces, and we adopt the usual notion, $L^2[0,T; H^1(\Omega)]$, $C[0,T; H^1(\Omega)]$, etc. to indicate the temporal regularity of such functions. The space of C^∞ test functions having compact

support in Ω is denoted by $\mathcal{D}(\Omega)$. For vector valued quantities, such as the velocity v, we write $v \in L^2(\Omega)$ to indicate that each component lies in the specified space. The space $H(\mathrm{div};\Omega)$ is the set of vector valued functions in $L^2[0,T;L^2(\Omega)]$ with divergence in $L^2[0,T;L^2(\Omega)]$. Strong convergence of a sequence will be indicated as $\rho_h \to \rho$, weak convergence by $\rho_h \rightharpoonup \rho$, and weak $\star$ convergence by $\rho_h \rightharpoonup^* \rho$.

2. Construction of numerical schemes.

2.1. Overview. Convergence proofs of numerical schemes for linear partial differential equations are almost always a variant of the old adage "stability and consistency imply convergence." For nonlinear problems, some form of compactness is usually also required. Our proof of convergence follows this line of argument; in particular, numerical schemes are constructed so that discrete versions of energy estimate (1.3) (and hence stability) hold.

The low regularity of the solution gives rise to many technical problems. If high order approximations of the density are used, Gibbs phenomena arise, and stable approximations of the momentum equation require the density to be truncated or projected onto a set of strictly positive functions. Since the density has low regularity we cannot establish consistency of such schemes. In this situation we are forced to resort piecewise constant approximations of the density which give rise to monotone schemes. Unfortunately, piecewise constant approximations of the density give rise to a different consistency error; specifically, jump terms arise when the test functions are not continuous.

In the current context the key compactness result is that solutions of the equation for the density ρ will converge strongly in $L^2[0,T;L^2(\Omega)]$ when the velocity converges weakly in $L^2[0,T;H_0^1(\Omega)]$, [5, 15]. The analogous statement for discontinuous Galerkin approximations of the density equation was established by Walkington in [18] and this result will be used below. Again the low regularity of the velocity, which appears as a nonconstant coefficient in the density equation, gives rise to technical problems. Specifically, in order to establish strong convergence of the density the (approximate) velocity fields are required to have average divergence equal to zero on each element [18].

2.2. Stability. The natural energy estimate given in (1.3) was derived assuming that the balance of mass is satisfied *exactly*. Since the balance of mass is only approximately satisfied by numerical approximations, the energy estimate is not automatic. Also, numerical approximations of the density may not be nonnegative, so even if an "energy estimate" holds it may not be useful. One way to circumvent these problems is to observe that if $\rho_t + \mathrm{div}(\rho v) = 0$, then

$$\rho\left(v_t + (v.\nabla)v\right) = \frac{1}{2}\Big(\rho v_t + (\rho v.\nabla)v + (\rho v)_t + \mathrm{div}(\rho v \otimes v)\Big).$$

Taking the dot product of the right-hand side with v vanishing on $\partial\Omega$ and integrating gives $(d/dt)\int(\rho|v|^2/2)$. This identity holds independently of the equation for the balance of mass and also holds if different approximations of the velocity are used as coefficients of the convective terms. This motivates the following weak statement of the momentum equation:

$$\frac{1}{2}\int_\Omega \bar\rho v_t.w + (\rho\bar v.\nabla)v.w + (\bar\rho v)_t.w - (\rho\bar v.\nabla)w.v + \int_\Omega -p\,\mathrm{div}(w) + \mu(\rho)D(v)\cdot D(w) = \int_\Omega \bar\rho f.w. \tag{2.1}$$

In the context of a numerical scheme, ρ is an approximation of the density which may not be positive and $\bar{\rho}$ is a nonnegative projection or truncation of ρ. Similarly, in order to obtain stability of the convection equation, a projection, $\bar{v}$, of v onto a suitable subspace of $H(\mathrm{div};\Omega)$ may be required for the convection terms; see [18]. Selecting $w(t) = v(t)$ in the above equation immediately gives

$$\frac{d}{dt}\int_\Omega (1/2)\bar{\rho}|v|^2 + \int_\Omega \mu(\rho)|D(v)|^2 = \int_\Omega \bar{\rho} f.v.$$

2.3. Consistency. While numerical schemes based upon the weak statement (2.1) will "automatically" be stable, they are not "automatically" consistent. Specifically, in the absence of any estimates on v_t it is necessary to integrate the first term by parts. Then

$$\begin{aligned}\int_0^T\int_\Omega \bar{\rho}v_t.w + (\rho\bar{v}.\nabla)v.w &= \int_0^T\int_\Omega -v.(\bar{\rho}w)_t + (\rho\bar{v}.\nabla)v.w \\ &= \int_0^T\int_\Omega -(\bar{\rho}-\rho)_t\, v.w - \bar{\rho}v.w_t - \left(\rho_t(v.w) - (\rho\bar{v}.\nabla)v.w\right).\end{aligned}$$

(1) If a *high order* approximation of the density equation is used it is possible to select $v.w$ as a test function in the Galerkin approximation of $\rho_t + \mathrm{div}(\rho\bar{v}) = 0$. Then

$$\int_0^T\int_\Omega \bar{\rho}v_t.w + (\rho\bar{v}.\nabla)v.w = \int_0^T\int_\Omega -(\bar{\rho}-\rho)_t\, v.w - \bar{\rho}v.w_t - (\rho\bar{v}\otimes v)\cdot\nabla w$$

and consistency requires the first term to vanish in the limit. For the continuous problem ρ is bounded in $L^\infty[0,T;L^\infty(\Omega)]$ so that the momentum, ρv, is bounded in $L^2[0,T;L^2(\Omega)]$. Since $\rho_t + \mathrm{div}(\rho v) = 0$, it follows that ρ_t is bounded in $L^2[0,T;H^{-1}(\Omega)]$. Unfortunately, L^∞ bounds could not be established for high order approximations of the density, so the analogous estimates could not be established for the time derivative of the discrete density. For this reason we could not construct nonnegative approximations, $\bar{\rho}$, for which $(\bar{\rho}-\rho)_t$ converged to zero in $L^2[0,T;H^{-1}(\Omega)]$. In particular, we could not establish consistency of numerical schemes constructed using high order approximations of the density equation.

(2) If *piecewise constant* approximations of the density are used, then numerical approximations of ρ are nonnegative so it is possible to select $\bar{\rho} = \rho$. The first term in (2.1) then becomes

$$\int_0^T\int_\Omega \rho v_t.w + (\rho\bar{v}.\nabla)v.w = \int_0^T\int_\Omega -\rho v.w_t - \rho_t v.w + (\rho\bar{v}.\nabla)v.w.$$

To establish consistency we would like to multiply the Galerkin approximation of $\rho_t + \mathrm{div}(\rho\bar{v}) = 0$ by $v.w$. When the density is approximated using piecewise constant functions we must first approximate $v.w$ by a (discontinuous) piecewise constant function. This leads to an expression of the form

$$\int_0^T\int_\Omega \rho v_t.w + (\rho\bar{v}.\nabla)v.w = \int_0^T\int_\Omega -\rho v.w_t - (\rho\bar{v}\otimes v)\cdot\nabla w + \text{"jump terms,"}$$

and the scheme is consistent provided the "jump terms" vanish in the limit. In section 3 we show that the jump terms do vanish in the limit, which establishes consistency.

2.4. Scheme. In light of the above discussion we will consider approximations of equations (1.1) where the density is approximated using piecewise constant approximations in space and time, and the momentum equation is approximated using the implicit Euler scheme with velocity-pressure spaces satisfying the Babuska–Brezzi condition. In order to minimize the technicalities it will be assumed that the pressure space contains the (discontinuous) piecewise constant functions on each triangulation. Relaxing this condition is considered in section 4. Since the accuracy of the piecewise constant approximation of the density is formally first order, at each discrete time we can first advance the density and then the velocity and pressure without further loss of accuracy. In this situation the linear systems for the density and velocity/pressure can be decoupled.

Given a triangulation $\mathcal{T}_h$ of Ω and time step $\tau = T/N$, let

$$\mathcal{R}_h = \{\rho \in L^2(\Omega) \mid \rho|_K \in \mathbb{R}\, \forall K \in \mathcal{T}_h\}.$$

If ρ^0 is the projection of $\rho(0)$ onto $\mathcal{R}_h$, then the (piecewise constant) discontinuous Galerkin approximation of $\rho(t^n)$ satisfies $\rho^n \in \mathcal{R}_h$ and

$$\int_K \rho^n \psi^n + \tau \int_{\partial K} \left(\rho^n_-(v^{n-1}.n)^+ + \rho^n_+(v^{n-1}.n)^-\right)\psi^n = \int_K \rho^{n-1}\psi^n, \tag{2.2}$$

for $K \in \mathcal{T}_h$ and $\psi^n \in \mathbb{R}$. Here $v.n = (v.n)^+ + (v.n)^-$ are the positive and negative parts of $v.n$ and $\rho^n_\pm(x) = \lim_{s\searrow 0} \rho^n(x \pm sn)$ so that the middle term gives the "upwind" value of $\rho^n v^{n-1}.n$.

To march the velocity forward, let

$$V_h \subset \{v \in H^1_0(\Omega) \mid v|_K \in \mathcal{P}_k(K),\, K \in \mathcal{T}_h\},$$

and

$$P_h \subset \{p \in L^2(\Omega)/\mathbb{R} \mid p|_K \in \mathcal{P}_\ell(K),\, K \in \mathcal{T}_h\},$$

be a pair of spaces satisfying the Babuska–Brezzi condition and let v^0 be the $L^2(\Omega)$ projection of $v(0)$ onto V_h. Then the approximations, $(v^n, p^n) \in V_h \times P_h$, of $(v(t^n), p(t^n))$ are the solution of

$$\begin{aligned}
&\frac{1}{2}\int_\Omega \left\{ \rho^{n-1}\left(\frac{v^n - v^{n-1}}{\tau}\right).w + (\rho^n v^{n-1}.\nabla)v^n.w \right. \\
&\qquad \left. + \left(\frac{(\rho v)^n - (\rho v)^{n-1}}{\tau}\right).w - (\rho^n v^{n-1}.\nabla)w.v^n \right\} \\
&\qquad + \int_\Omega -p^n\, div(w) + \mu^n D(v^n)\cdot D(w) = \int_\Omega \rho^n f^n.w,
\end{aligned} \tag{2.3}$$

$$\int_\Omega \operatorname{div}(v^n)\, q = 0$$

for all $(w, q) \in V_h \times P_h$. In the above equation, f^n is an approximation of the average of f on $(t^{n-1}, t^n]$ and $\mu^n = \mu(\rho^n)$.

3. Analysis of the numerical scheme.

3.1. Estimates. To establish stability of the scheme (2.2)–(2.3) we first state the natural energy estimate the scheme was designed to satisfy.

Notation: If $\{v_n\}_{n=0}^N \subset V_h$ and $\{\rho^n\}_{n=0}^N \subset \mathcal{R}_h$, then we let $v_h \in L^2[-\tau, T; V_h]$ and $\rho_h \in L^2[-\tau, T; \mathcal{R}_h]$ denote the piecewise constant functions taking values v^n and ρ^n on $(t^{n-1}, t^n]$, respectively.

LEMMA 3.1. *Let (ρ_h, v_h, p_h) be the approximate solution of equations* (1.1) *computed using the scheme* (2.2)–(2.3). *Then*

$$\frac{1}{2}\int_\Omega \rho^n|v^n|^2 + \frac{1}{2}\sum_{i=1}^n \int_\Omega \rho^{n-1}|v^n - v^{n-1}|^2 + \sum_{i=1}^n \tau \int_\Omega \mu^n |D(v^n)|^2 = \frac{1}{2}\int_\Omega \rho^0|v^0|^2 + \sum_{i=1}^n \tau \int_\Omega \rho^n f^n.v^n.$$

Let the pressure space contain the piecewise constant functions. If $0 < c \le \rho(0) \le C$ and $0 < c \le \mu(\rho) \le C$ for constants c, $C \in \mathbb{R}$, $v_0 \in L^2(\Omega)$, and $f \in L^2[0, T; L^2(\Omega)]$, then $\{v_h\}_{h>0}$ is bounded in $L^\infty[0, T; L^2(\Omega)] \cap L^2[0, T; H_0^1(\Omega)]$ and

$$\int_\tau^T \|v_h(t) - v_h(t-\tau)\|^2_{L^2(\Omega)} \le C(v_0, f)\tau.$$

The first estimate follows directly upon substituting $w = v^n$ and $q = p^n$ into equations (2.3). The assumption on the pressure space guarantees that the scheme for the density is monotone [18, Theorem 6.1], so the bounds on the initial data are preserved,

$$\min_\Omega \rho^0 \le \rho^n(x) \le \max_\Omega \rho^0, \qquad x \in \Omega. \tag{3.1}$$

The bounds on $\{v_h\}_{h>0}$ then follow from the energy estimate.

3.2. Consistency of the density equation. To establish compactness of the sequence $\{v_h\}_{h>0}$ in $L^2[0, T; L^2(\Omega)]$, it is necessary to use test functions ψ in the discrete density equation (2.2) which are not piecewise constant. This gives rise to consistency errors which are estimated in this section. The following lemma provides explicit expressions for these errors.

LEMMA 3.2. *Let $\rho_h \in \mathcal{R}_h$ satisfy* (2.2). *If $\psi \in H_0^1(\Omega)$ and $\bar\psi \in \mathcal{R}_h$ is the function taking the average value of ψ on each element $K \in \mathcal{T}_h$, then*

$$\int_\Omega (\rho^n - \rho^{n-1})\psi - \tau\int_\Omega \rho^n v^{n-1}.\nabla\psi = \tau\int_\Omega \rho^n(\psi - \bar\psi)\mathrm{div}(v^{n-1}) + \tau\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho^n](v^{n-1}.n)^-(\psi - \bar\psi),$$

where the value of $\bar\psi$ on ∂K is taken as $\bar\psi|_K$ (that is, the trace from inside K) and $[\rho^n] = \rho^n_+ - \rho^n_-$.

Proof. Select $\psi^n = \bar\psi|_K$ in (2.2) and sum over all of the simplices $K \in \mathcal{T}_h$ to get

$$\int_\Omega \rho^n\psi + \tau\sum_{K\in\mathcal{T}_h}\int_{\partial K}\left(\rho^n_-(v^{n-1}.n)^+ + \rho^n_+(v^{n-1}.n)^-\right)\bar\psi = \int_\Omega \rho^{n-1}\psi. \tag{3.2}$$

In the middle term $\rho^n_\pm(x) = \lim_{s\searrow 0}\rho^n(x \pm n)$ and $\bar\psi|_{\partial K} = \bar\psi|_K$. If $\mathcal{E}_0$ denotes all of the interior edges (faces in $3d$) of the elements, then the middle term may be written as

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\left(\rho^n_-(v^{n-1}.n)^+ + \rho^n_+(v^{n-1}.n)^-\right)\bar\psi$$
$$= \sum_{e\in\mathcal{E}_0}\int_e -\left(\rho^n_-(v^{n-1}.N)^+ + \rho^n_+(v^{n-1}.N)^-\right)[\bar\psi].$$

Here N is one of the normals to e, $\rho^n_\pm(x) = \lim_{s\searrow 0}\rho^n(x \pm sN)$ and $[\bar\psi] = \bar\psi_+ - \bar\psi_-$. Integrals over the edges $e \subset \partial\Omega$ vanish since $\int_e v.n = 0$ on boundary edges. If $\psi : \bar\Omega \to \mathbb{R}$ is continuous and vanishes on $\partial\Omega$, then $[\psi] = 0$ on each edge $e \in \mathcal{E}_0$, so $[\bar\psi] = [\bar\psi - \psi]$. Reversing the above calculation shows

$$\sum_{K\in\mathcal{T}_h}\int_{\partial K}\left(\rho^n_-(v^{n-1}.n)^+ + \rho^n_+(v^{n-1}.n)^-\right)\bar\psi$$
$$= \sum_{K\in\mathcal{T}_h}\int_{\partial K}\left(\rho^n_-(v^{n-1}.n)^+ + \rho^n_+(v^{n-1}.n)^-\right)(\bar\psi - \psi)$$
$$= \sum_{K\in\mathcal{T}_h}\int_{\partial K}\left(\rho^n_-(v^{n-1}.n) + (\rho^n_+ - \rho^n_-)(v^{n-1}.n)^-\right)(\bar\psi - \psi)$$
$$= \sum_{K\in\mathcal{T}_h}\int_K \operatorname{div}\left(\rho^n v^{n-1}(\bar\psi - \psi)\right) + \sum_{K\in\mathcal{T}_h}\int_{\partial K}(\rho^n_+ - \rho^n_-)(v^{n-1}.n)^-(\bar\psi - \psi)$$
$$= -\int_\Omega\left(\rho^n v^{n-1}.\nabla\psi + \rho^n(\psi - \bar\psi)\operatorname{div}(v^{n-1})\right) - \sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho^n](v^{n-1}.n)^-(\psi - \bar\psi).$$

The last step used the property that ρ^n and ψ^n are constant on each element $K \in \mathcal{T}_h$. The lemma follows upon substituting this expression into (3.2). ∎

The following corollary expresses the weak statement satisfied by the discrete density ρ_h in a more convenient form. Given a sequence of functions $\{\psi^n\}_{n=0}^N \subset \mathcal{R}_h$, recall the convention that $\psi_h : (-\tau, T] \to \mathcal{R}_h$ is the function taking values $\psi_h(t) = \psi^n$ for $t \in (n-1)\tau, n\tau]$.

COROLLARY 3.3. *Let $\rho_h \in \mathcal{R}_h$ satisfy (2.2), $\{\psi^n\}_{n=0}^N \subset H^1_0(\Omega)$, and let $\{\bar\psi^n\}_{n=0}^N \subset \mathcal{R}_h$ be the piecewise constant approximations of $\{\psi^n\}_{n=0}^N$. Then*

$$\sum_{j=m+1}^n\int_\Omega(\rho^j - \rho^{j-1})\psi^j - \int_{t^m}^{t^n}\int_\Omega \rho_h v_h(.-\tau).\nabla\psi_h \tag{3.3}$$
$$= \int_{t^m}^{t^n}\int_\Omega \rho_h(\psi_h - \bar\psi_h)\operatorname{div}(v_h(.-\tau)) + \int_{t^m}^{t^n}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho_h](v_h(.-\tau).n)^-(\psi_h - \bar\psi_h),$$

where the value of $\bar\psi_h$ on ∂K is taken as $\bar\psi_h|_K$.

The two terms on the right-hand side represent the consistency error of the piecewise constant DG scheme. The first term is easy to bound, and the following lemma will be used to bound the last one.

LEMMA 3.4. *Let $K \subset \mathbb{R}^d$ be a simplex, $v \in \mathcal{P}_\ell(K)^d$, $\psi \in \mathcal{P}_\ell(K)$ and p, $q \geq 1$. Then there exists a constant C depending only upon d, p, q, ℓ and the aspect ratio of*

K such that

$$\int_{\partial K} |v.n||\psi - \bar{\psi}|^q \le C\|v\|_{L^{p'}(K)} h_K^{q-1} |\psi|^q_{W^{1,pq}(K)},$$

where $\bar{\psi} = (1/|K|)\int_K \psi$ is the average of ψ on K and h_K is the diameter of K.

Proof. Let $\hat{K}$ be the usual reference simplex and $\chi(\xi) = x_0 + B\xi$ be an affine mapping of $\hat{K}$ onto K. We use a hat to denote the natural correspondence between functions defined on K and $\hat{K}$, $\hat{\psi} = \psi \circ \chi$. Writing the integral over the boundary as the sum over the faces $e \subset \partial K$ gives

$$\begin{aligned}
\int_{\partial K} |v.n||\psi - \bar{\psi}|^q &= \sum_{e\subset\partial K} \int_e |v.n||\psi - \bar{\psi}|^q \\
&= \sum_{\hat{e}\subset\partial\hat{K}} \frac{|e|}{|\hat{e}|} \int_{\hat{e}} |\hat{v}.n||\hat{\psi} - \bar{\psi}|^q \\
&\le C \sum_{\hat{e}\subset\partial\hat{K}} |e| \|\hat{v}\|_{L^{p'}(\hat{K})} \|\hat{\psi} - \bar{\psi}\|^q_{L^{qp}(\hat{K})} \\
&\le C \sum_{\hat{e}\subset\partial\hat{K}} |e| \|\hat{v}\|_{L^{p'}(\hat{K})} |\hat{\psi}|^q_{W^{1,pq}(\hat{K})}.
\end{aligned}$$

To obtain the third line the trace theorem was used and the finite dimensionality of $\mathcal{P}_\ell(\hat{K})$ allowed the use of the indicated norms. The last line follows from the Poincaré inequality and the observation that the average of ψ is the average of $\hat{\psi}$.

Since

$$\|\hat{v}\|_{L^{p'}(\hat{K})} = (|\hat{K}|/|K|)^{1/p'} \|v\|_{L^{p'}(K)}, \qquad |\hat{\psi}|_{W^{1,pq}(\hat{K})} \le C(|\hat{K}|/|K|)^{1/pq} h_K |\psi|_{W^{1,pq}(K)},$$

and $|e| \le C|K|/h_K$, where C depends upon the aspect ratio of K, the lemma follows. ∎

3.3. Compactness. The energy estimate shows that $\{v_h\}_{h>0}$ is bounded in $L^\infty[0,T;L^2(\Omega)] \cap L^2[0,T;H_0^1(\Omega)]$. A result of Lions [13] and Lions and Magenes [14] states that compactness of the sequence in $L^2[0,T;L^2(\Omega)]$ will follow if

$$\int_\delta^T \|v_h(t) - v_h(t-\delta)\|_{L^2(\Omega)} \le C\delta^\alpha,$$

for $0 \le \delta \le T$ and some $\alpha > 0$.

We recall Lions' argument [13] which shows that weak solutions of the Navier–Stokes equations with variable density and viscosity satisfy this inequality. This proof carries over to Galerkin approximations with a few modifications which will be considered subsequently.

Lions' compactness argument. Beginning with the weak statement of the momentum equation (cf. (2.3))

$$\begin{aligned}
\int_\Omega \Big\{ &(1/2)\left(\rho v_t + (\rho v)_t\right).w + (1/2)(\rho v.\nabla)v.w - (1/2)(\rho v.\nabla)w.v \\
&- p\,\mathrm{div}(w) + \mu D(v)\cdot D(w) \Big\} = \int_\Omega \rho f.w,
\end{aligned}$$

the identity $\rho v_t = (\rho v)_t - \rho_t v$ is used to obtain

$$\int_\Omega (\rho v)_t.w = \int_\Omega \Big\{\rho f.w + (1/2)\rho_t(v.w) - (1/2)(\rho v.\nabla)v.w \\ + (1/2)(\rho v.\nabla)w.v + p\,\mathrm{div}(w) - \mu D(v)\cdot D(w)\Big\}.$$

The second term on the right-hand side can be eliminated upon writing the weak statement of the balance of mass as

$$\int_\Omega \rho_t \psi = \int_\Omega \rho v.\nabla\psi, \tag{3.4}$$

and selecting $\psi = v.w$, to give

$$\int_\Omega (\rho v)_t.w = \int_\Omega \rho f.w + (\rho v.\nabla)w.v + p\,\mathrm{div}(w) - \mu D(v)\cdot D(w).$$

Integrating this expression with respect to $s \in (t-\delta, t)$ and letting $w = w(t)$ be independent of s gives

$$\int_\Omega \rho v|_{t-\delta}^t .w(t) = \int_{t-\delta}^t \int_\Omega \rho f.w(t) + (\rho v.\nabla)w(t).v + p\,\mathrm{div}(w(t)) - \mu D(v)\cdot D(w(t))\,ds.$$

Integrating the weak statement of the balance of mass (3.4) with respect to $s \in (t-\delta, t)$ and setting $\psi = v(t).w(t)$ shows

$$\int_\Omega \rho|_{t-\delta}^t v(t).w(t) = \int_{t-\delta}^t \int_\Omega \rho v.\nabla(v(t).w(t)).$$

Subtracting this equation from the previous one and observing that

$$\rho v|_{t-\delta}^t .w(t) - \rho|_{t-\delta}^t v(t).w(t) = \rho(t-\delta)(v(t) - v(t-\delta)).w(t)$$

gives

$$\int_\Omega \rho(t-\delta)(v(t)-v(t-\delta)).w(t) = \int_{t-\delta}^t \int_\Omega \Big\{\rho f.w(t) + (\rho v.\nabla)w(t).v \\ + p\,\mathrm{div}(w(t)) - \mu D(v)\cdot D(w(t)) - \rho v.\nabla(v(t).w(t))\Big\}\,ds. \tag{3.5}$$

Upon electing $w(t) = v(t) - v(t-\delta)$ the left-hand side dominates $\|v(t) - v(t-\delta)\|^2_{L^2(\Omega)}$ when ρ is bounded below by $c > 0$. The right-hand side is estimated using the following lemma.

LEMMA 3.5. *Let $\Omega \subset \mathbb{R}^d$ with $d = 2$ or 3 and $v, w \in L^2[0,T; H_0^1(\Omega)] \cap L^\infty[0,T; L^2(\Omega)]$, $\rho, \mu \in L^\infty[0,T; L^\infty(\Omega)]$, and $f \in L^2[0,T; L^2(\Omega)]$. Then there exists a constant $C > 0$ and $\alpha \in (0,1)$ such that*

$$\left| \int_\delta^T \int_{t-\delta}^t \int_\Omega \rho f.w(t) + (\rho v.\nabla)w(t).v - \mu D(v)\cdot D(w(t)) - \rho v.\nabla(v(t).w(t))\,ds\,dt \right| \le C\delta^\alpha,$$

for $0 < \delta < T$. Here C depends only upon d, T, and f, ρ, μ, v, and w through the norms stated in the hypotheses.

This lemma follows from elementary applications of Holder's inequality and the Sobolev embedding theorem, $\|v\|_{L^4(\Omega)} \le \|v\|^\beta_{L^2(\Omega)} \|\nabla v\|^{1-\beta}_{L^2(\Omega)}$, where $\beta = 1/2$ and $\beta = 1/4$ for $d = 2$ and 3, respectively.

Compactness for the discrete problem. The calculations above can be replicated for numerical solutions computed using (2.2)–(2.3) provided the discrete weak statement of the balance of mass (3.3) is used in place of (3.4). This gives rise to four extra terms on the right-hand side of (3.5).

LEMMA 3.6. *Let $\{(\rho_h, v_h)\}_{h>0}$ be numerical approximations of the Navier–Stokes equations with variable density and viscosity computed using (2.2)–(2.3) over a quasi-regular family of triangulations $\{\mathcal{T}_h\}_{h>0}$ of $\Omega \subset \mathbb{R}^d$ with $d = 2$ or $= 3$. Assume the following:*

- *$v^0 \in L^2(\Omega)$, $\rho^0 \in L^\infty(\Omega)$ satisfies $0 < c \le \rho^0(x) \le C$, and $f \in L^2[0,T;L^2(\Omega)]$.*
- *$\mu : \mathbb{R} \to \mathbb{R}^+$ is continuous.*
- *The spaces for the velocity and pressure satisfy the Babuska–Brezzi condition and the pressure space contains the piecewise constant functions.*

Then there exists a constant $C > 0$ independent of h and $\alpha \in (0,1)$ such that

$$\int_\delta^T \|v_h(t) - v_h(t-\delta)\|^2_{L^2(\Omega)} \le C\delta^\alpha,$$

for $0 < \delta < T$.

Proof. Since $\{v_h\}_{h>0}$ are piecewise constant in time it suffices to consider δ a multiple of the time step τ. Writing $(t-\delta, t) = (t^m, t^n)$ and $w(t) = v^n - v^m = w^{mn}$, the discrete analogue of (3.5) is

$$\begin{aligned}
\int_\Omega \rho^m (v^n - v^m).w^{mn} = \int_{t^m}^{t^n} \int_\Omega \Big\{ & \rho_h f.w^{mn} + (\rho_h v_h(.-\tau).\nabla) w^{mn}.v_h \\
& - \mu_h D(v_h) \cdot D(w^{mn}) - \rho_h v_h.\nabla(v^n.w^{mn}) \Big\}\, ds \\
+ \int_{t^m}^{t^n} \int_\Omega \Big\{ & \rho_h (v_h.w^{mn} - \overline{v_h.w^{mn}}) \operatorname{div}(v_h(.-\tau)) \\
& - \rho_h (v^n.w^{mn} - \overline{v^n.w^{mn}}) \operatorname{div}(v_h(.-\tau)) \Big\}\, ds \\
+\tau \sum_{j=m+1}^{n} \sum_{K \in \mathcal{T}_h} \int_{\partial K} \Big\{ & [\rho^j](v^{j-1}.n)^- \quad (v^j.w^{mn} - \overline{v^j.w^{mn}}) \\
& - [\rho^j](v^{j-1}.n)^- (v^n.w^{mn} - \overline{v^n.w^{mn}}) \Big\}.
\end{aligned}$$

The last four terms represent the consistency errors associated with the density equation and the term involving the pressure vanishes since $w^{mn} = v^n - v^m$ is discreetly divergence free. (Recall that $\bar\psi$ is the piecewise constant function having average value of ψ on each element $K \in \mathcal{T}_h$.)

We bound the first term in each of the last two lines since the second is bounded similarly. Since $\|\bar\psi\|_{L^p(\Omega)} \le \|\psi\|_{L^p(\Omega)}$ for any $\psi \in L^p(\Omega)$, the first term on the second to last line may be bounded as

$$\begin{aligned}
& \int_{t^m}^{t^n} \int_\Omega \rho_h (v_h.w^{mn} - \overline{v_h.w^{mn}}) \operatorname{div}(v_h(.-\tau))\, ds \\
& \le 2\|\rho_h\|_{L^\infty[0,T;L^\infty(\Omega)]} \int_{t^m}^{t^n} \|v_h\|_{L^4(\Omega)} \|w^{mn}\|_{L^4(\Omega)} \|\operatorname{div}(v_h(.-\tau))\|_{L^2(\Omega)}\, ds \\
& \le C \int_{t^m}^{t^n} \|\nabla v_h\|^{1-\beta}_{L^2(\Omega)} \|\operatorname{div}(v_h(.-\tau))\|_{L^2(\Omega)}\, ds\, \|\nabla w^{mn}\|^{1-\beta}_{L^2(\Omega)}
\end{aligned}$$

$$\le C\int_{t^m}^{t^n}\|\nabla v_h\|_{L^2(\Omega)}^{1-\beta}\|\nabla v_h(.-\tau)\|_{L^2(\Omega)}\,ds\,\|\nabla w^{mn}\|_{L^2(\Omega)}^{1-\beta}$$
$$\le C\|\nabla v_h\|_{L^2[0,T;L^2(\Omega)]}^{2-\beta}(t^n-t^m)^{\beta/2}\|\nabla w^{mn}\|_{L^2(\Omega)}^{1-\beta}$$

(here $\beta = 1/2$ or $1/4$ for $d = 2$ or 3, respectively). Since v_h, $w^{mn} \in L^\infty[0,T;L^2(\Omega)]$, quantities involving $\|v_h\|_{L^2(\Omega)}$ and $\|w^{mn}\|_{L^2(\Omega)}$ have been absorbed into the constant C. Integrating with respect to $t^n \in (\delta, T)$ and recalling that $t^m = t^n - \delta$ and $w^{mn} = v_h(t^n) - v_h(t^n - \delta)$ shows that this term may be bounded by a constant of the form $C\delta^{\beta/2}$ with C independent of h.

To estimate the first jump term use Lemma 3.4 with $q = 1$ to obtain

$$\tau\sum_{j=m+1}^{n}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho^j](v^{j-1}.n)^-(v^j.w^{mn}-\overline{v^j.w^{mn}})$$
$$\le C\|\rho_h\|_{L^\infty[0,T;L^\infty(\Omega)]}\tau\sum_{j=m+1}^{n}\sum_{K\in\mathcal{T}_h}\int_K\|v^{j-1}\|_{L^{p'}(K)}|v^j.w^{mn}|_{W^{1,p}(K)}$$
$$\le C\|\rho_h\|_{L^\infty[0,T;L^\infty(\Omega)]}\int_{t^m}^{t^n}\|v_h(.-\tau)\|_{L^{p'}(\Omega)}|v_h.w^{mn}|_{W^{1,p}(\Omega)}.$$

When $p < 2$ the terms of the form $\nabla(v.w)$ can be estimated as

$$\begin{aligned}|\nabla(v.w)|_{W^{1,p}(\Omega)} &\le \||v||\nabla w| + |\nabla v||w|\|_{L^p(\Omega)}\\ &\le \||v||\nabla w|\|_{L^p(\Omega)} + \||\nabla v||w|\|_{L^p(\Omega)}\\ &\le \|v\|_{L^{2p/(2-p)}(\Omega)}\|\nabla w\|_{L^2(\Omega)} + \|\nabla v\|_{L^2(\Omega)}\|w\|_{L^{2p/(2-p)}(\Omega)}.\end{aligned}$$

Letting $p = 4/3$ so that $2p/(2-p) = 4$, the first jump term becomes

$$\tau\sum_{j=m+1}^{n}\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho^j](v^{j-1}.n)^-(v^j.w^{mn}-\overline{v^j.w^{mn}})$$
$$\le C\int_{t^m}^{t^n}\|v_h(.-\tau)\|_{L^4(\Omega)}\left(\|v_h\|_{L^4(\Omega)}\|\nabla w^{mn}\|_{L^2(\Omega)} + \|\nabla v_h\|_{L^2(\Omega)}\|w^{mn}\|_{L^4(\Omega)}\right)ds$$
$$\le C\int_{t^m}^{t^n}\|\nabla v_h(.-\tau)\|_{L^2(\Omega)}^{1-\beta}\Big(\|\nabla v_h\|_{L^2(\Omega)}^{1-\beta}\|\nabla w^{mn}\|_{L^2(\Omega)} + \|\nabla v_h\|_{L^2(\Omega)}\|\nabla w^{mn}\|_{L^2(\Omega)}^{1-\beta}\Big)ds$$
$$\le C\Big(\|\nabla v_h\|_{L^2[0,T;L^2(\Omega)]}^{2(1-\beta)}(t^n-t^m)^{\beta}\|\nabla w^{mn}\|_{L^2(\Omega)}$$
$$+\|\nabla v_h\|_{L^2[0,T;L^2(\Omega)]}^{2-\beta}(t^n-t^m)^{\beta/2}\|\nabla w^{mn}\|_{L^2(\Omega)}^{1-\beta}\Big).$$

Integration with respect to $t^n \in (\delta, T)$ bounds this term by a constant of the form $C\delta^{\beta/2}$ with C independent of h. □

3.4. Convergence. The bound on the sequence $\{v_h\}_{h>0}$ and the compactness result of Lions [13] and Lions and Magenes [14] allows passage to a subsequence for which

$$v_h \rightharpoonup^* v \text{ in } L^\infty[0,T;L^2(\Omega)]\cap L^2[0,T;H_0^1(\Omega)] \qquad \text{and} \qquad v_h \to v \text{ in } L^2[0,T;L^2(\Omega)].$$

In this situation, Theorem 5.1 of [18] states that the corresponding densities $\{\rho_h\}_{h>0}$ converge in $L^2[0,T;L^2(\Omega)]$ to a limit which we denote by ρ. We will show that the pair (v,ρ) is a solution of (1.1).

Note that since $\{\rho_h\}_{h>0}$ is bounded in $L^\infty[0,T;L^\infty(\Omega)]$ and converges in $L^2[0,T;L^2(\Omega)]$ it also converges in $L^p[0,T;L^p(\Omega)]$ for any $1 \le p < \infty$. Similarly, since $\{v_h\}_{h>0}$ is bounded in $L^\infty[0,T;L^2(\Omega)] \cap L^2[0,T;H_0^1(\Omega)]$ and converges in $L^2[0,T;L^2(\Omega)]$, the Sobolev embedding theorem and elementary interpolation show that v_h converges in $L^p[0,T;L^q(\Omega)]$ for any pair p, $q \ge 1$ satisfying $1/2 < 1/q + 2/dp$.

THEOREM 3.7. *Let $\Omega \subset \mathbb{R}^d$, $d = 2$ or 3, be a bounded Lipschitz and $\{\mathcal{T}_h\}_{h>0}$ be a regular family of quasi-uniform triangulations of Ω. Let $f \in L^2[0,T;L^2(\Omega)]$, $v_0 \in L^2(\Omega)$, and $\rho_0 \in L^\infty(\Omega)$ satisfy $0 < c \le \rho_0(x) \le C$ for positive constants c and C. Assume that the viscosity, $\mu : \mathbb{R} \to (0,\infty)$, is a continuous nonnegative function of the density.*

Let $\{(v_h,\rho_h)\}_{h>0}$ be the approximate solution of equations (1.1) *computed using the scheme presented in section* 2.4 *with time steps τ converging to zero as $h \to 0$. In particular, assume that the density is computed using the piecewise constant discontinuous Galerkin method, that the velocity-pressure spaces satisfy the Babuska–Brezzi condition, and that the pressure space contains the piecewise constant functions.*

Then, after passing to a subsequence, the densities $\{\rho_h\}$ converge strongly in $L^2[0,T;L^2(\Omega)]$, and the velocities $\{v_h\}$ converge strongly in $L^2[0,T;L^2(\Omega)]$ and weakly star in $L^\infty[0,T;L^2(\Omega)] \cap L^2[0,T;H_0^1(\Omega)]$ to a weak solution of equations (1.1) *with initial data (v_0,ρ_0) and homogeneous Dirichlet boundary data on the velocity. If the solution of equations* (1.1) *is unique, then the whole sequence $\{(v_h,\rho_h)\}_{h>0}$ converges.*

Proof. Notice that the hypotheses of Lemma 3.1 are satisfied since monotonicity of the scheme (2.2) for computing the density guarantees that $0 < c \le \rho_h(x,t) \le C$. Also, $\mu : \mathbb{R} \to (0,\infty)$ is continuous so $\mu_h = \mu(\rho_h)$ satisfies a similar inequality.

Let $w \in \mathcal{D}([0,T) \times \Omega)$ be divergence free and let w^n be the Stokes projection of $w(t^n)$ onto the space

$$\tilde{V}_h = \left\{ v_h \in V_h \mid \int_\Omega \operatorname{div}(v_h)\, q_h = 0 \,\forall q_h \in P_h \right\},$$

and let $w_h \in L^2[0,T;V_h]$ be the piecewise constant function taking values w^n on $(t^{n-1},t^n]$ and $\hat{w}_h \in C[0,T;V_h]$ be the corresponding piecewise linear interpolant. Since the pair (V_h,P_h) satisfies the Babuska–Brezzi condition, w_h and $\hat{w}_h$ converge to w in $L^\infty[0,T,H_0^1(\Omega)]$ and $W^{1,\infty}[0,T;H_0^1(\Omega)]$, respectively. Selecting w^n as the test function in (2.3) and summing over n gives

$$\begin{aligned}\frac{1}{2}\sum_{n=1}^{N}\int_\Omega (\rho^{n-1}-\rho^n)v^n.w^n + \sum_{n=1}^{N}\int_\Omega (\rho v)^{n-1}.(w^{n-1}-w^n)\\ +\frac{\tau}{2}\sum_{n=1}^{N}\int_\Omega (\rho^n v^{n-1}.\nabla)v^n.w^n - (\rho^n v^{n-1}.\nabla)w^n.v^n\\ +\tau\sum_{n=1}^{N}\int_\Omega \mu^n D(v^n)\cdot D(w^n) = \int_\Omega \rho^0 v^0.w^0 + \tau\sum_{n=1}^{N}\int_\Omega \rho^n f^n.w.\end{aligned}$$

To obtain the first line we used the identity

$$\begin{aligned}&\frac{1}{2}\Big(\rho^{n-1}(v^n-v^{n-1}).w^n + \left((\rho v)^n-(\rho v)^{n-1}\right).w^n\Big)\\ &\quad = \frac{1}{2}(\rho^{n-1}-\rho^n)v^n.w^n + \left((\rho v)^n-(\rho v)^{n-1}\right).w^n,\end{aligned}$$

and summed the second term by parts. The upper limit of the summation vanishes since $w \in \mathcal{D}([0,T)\times\Omega)$ implies $w^N = 0$. Recalling the notation that $v_h(t) \in V_h$ is the function taking on value v^n on $(t^{n-1}, t^n]$, we find that

$$\begin{aligned}\frac{1}{2}\sum_{n=1}^{N}\int_\Omega(\rho^{n-1}-\rho^n)v^n.w^n - \int_0^T\int_\Omega \rho_h(.-\tau)v_h(.-\tau)\hat{w}_t \\ +\frac{1}{2}\int_0^T\int_\Omega \left(\rho_h v_h(.-\tau).\nabla\right)v_h.w_h - \left(\rho_h v_h(.-\tau).\nabla\right)w_h.v_h \\ +\int_0^T\int_\Omega \mu_h D(v_h)\cdot D(w_h) = \int_\Omega \rho^0 v^0.w^0 + \int_0^T\int_\Omega \rho_h f_h.w_h.\end{aligned}$$

Selecting $\psi = v_h.w_h$ in Corollary 3.3 shows that the first term can be rewritten as

$$\sum_{n=1}^{N}\int_\Omega(\rho^{n-1}-\rho^n)v^n.w^n = -\int_0^T\int_\Omega \rho_h v_h(.-\tau).\nabla(v_h.w_h) - e_h,$$

where

$$\begin{aligned}e_h = &\int_0^T\int_\Omega \rho_h(v_h.w_h - \overline{v_h.w_h})\mathrm{div}(v_h(.-\tau)) \\ &+\int_0^T\sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho_h](v_h(.-\tau).n)^-(v_h.w_h - \overline{v_h.w_h})\end{aligned}$$

is the consistency error. Then

$$\begin{aligned}-\int_0^T\int_\Omega \rho_h(.-\tau)v_h(.-\tau)\hat{w}_t + \left(\rho_h v_h(.-\tau)\otimes v_h\right)\cdot\nabla w_h \\ +\int_0^T\int_\Omega \mu_h D(v_h)\cdot D(w_h) = \int_\Omega \rho^0 v^0.w^0 + \int_0^T\int_\Omega \rho_h f_h.w_h + e_h.\end{aligned}$$

Now pass to a subsequence along which v_h and ρ_h converge in $L^2[0,T;L^2(\Omega)]$ and $v_h \rightharpoonup v$ in $L^2[0,T;H_0^1(\Omega)]$. Since ρ_h and μ_h are bounded in $L^\infty[0,T;L^\infty(\Omega)]$, they converge in $L^p[0,T;L^p(\Omega)]$ for $1 \le p < \infty$ and v_h converges in $L^p[0,T;L^4(\Omega)]$ for $p < 8/3$. This is sufficient to pass to the limit term-by-term in the above equation; the theorem will then follow provided $e_h \to 0$.

It suffices to show that the consistency error e_h vanishes as $h \to 0$. The first term in e_h is bounded using classical estimates for piecewise constant approximations [3],

$$\begin{aligned}&\int_0^T\int_\Omega \rho_h(v_h.w_h - \overline{v_h.w_h})\mathrm{div}(v_h(.-\tau)) \\ &\le C\|\rho_h\|_{L^\infty[0,T;L^\infty(\Omega)]}\|\mathrm{div}(v_h)\|_{L^2[0,T;L^2(\Omega)]}\|v_h.w_h - \overline{v_h.w_h}\|_{L^2[0,T;L^2(\Omega)]} \\ &\le C\|\rho_h\|_{L^\infty[0,T;L^\infty(\Omega)]}\|\mathrm{div}(v_h)\|_{L^2[0,T;L^2(\Omega)]}|v_h.w_h|_{L^2[0,T;W^{1,p}(\Omega)}h^{1+d(1/2-1/p)}.\end{aligned}$$

As in the proof of Lemma 3.6

$$\begin{aligned}|v_h.w_h|_{W^{1,4/3}(\Omega)} &\le \|v_h\|_{L^4(\Omega)}\|\nabla w_h\|_{L^2(\Omega)} + \|\nabla v_h\|_{L^2(\Omega)}\|w_h\|_{L^4(\Omega)} \\ &\le C\|v_h\|_{H^1(\Omega)}\|w_h\|_{H^1(\Omega)}.\end{aligned}$$

It follows that $|v_h.w_h|_{L^2[0,T;W^{1,4/3}]} \le \|v_h\|_{L^2[0,T;H^1(\Omega)]}\|w_h\|_{L^\infty[0,T;H^1(\Omega)]}$ is bounded so

$$\begin{aligned}
&\int_0^T \int_\Omega \rho_h(v_h.w_h - \overline{v_h.w_h})\mathrm{div}(v_h(.-\tau))\\
&\le C\|\rho_h\|_{L^\infty[0,T;L^\infty(\Omega)]}\|\mathrm{div}(v_h)\|_{L^2[0,T;L^2(\Omega)]}\|v_h\|_{L^2[0,T;H^1(\Omega)]}\|w_h\|_{L^\infty[0,T;H^1(\Omega)]}h^{1-d/4}\\
&\to 0.
\end{aligned}$$

The second term of e_h is bounded using Lemma 3.4 with $q \le 2$,

$$\begin{aligned}
&\int_0^T \sum_{K\in\mathcal{T}_h}\int_{\partial K}[\rho_h](v_h(.-\tau).n)^-(v_h.w_h - \overline{v_h.w_h})\\
&\le \left(\|\rho_h\|^{q'-2}_{L^\infty[0,T;L^\infty(\Omega)]}\int_0^T\sum_{K\in\mathcal{T}_h}\int_{\partial K}|v_h(.-\tau).n|[\rho_h]^2\right)^{1/q'}\\
&\quad\times\left(\int_0^T\sum_{K\in\mathcal{T}_h}\int_{\partial K}|v_h(.-\tau)|(v_h.w_h-\overline{v_h.w_h})^q\right)^{1/q}\\
&\le C\|\rho_h\|^{1-2/q'}_{L^\infty[0,T;L^\infty(\Omega)]}(J_h^N)^{1/q'}\left(\int_0^T\int_\Omega\|v_h(.-\tau)\|_{L^{p'}(\Omega)}|v_h.w_h|^q_{W^{1,pq}(\Omega)}h^{q-1}\right)^{1/q}\\
&\le C\|\rho_h\|^{1-2/q'}_{L^\infty[0,T;L^\infty(\Omega)]}(J_h^N)^{1/q'}\|v_h\|^{1/q}_{L^{r'}[0,T,L^{p'}(\Omega)]}|v_h.w_h|_{L^{rq}[0,T;W^{1,pq}(\Omega)]}h^{1-1/q},
\end{aligned}$$

where $q' \ge 2$ and

$$J_h^N = \sum_{e\in\mathcal{E}_0}\int_0^T\int_e |v_h(.-\tau).n|\,[\rho_h]^2.$$

J_h^N measures the jumps in the density across the interelement boundaries $e \in \mathcal{E}_0$, and it was shown in [18, Theorem 5.1] that, under the hypotheses assumed above, $J_h^N \to 0$ as h (and τ) tend to zero. The parameters p, q, and r are selected so that the norms of v_h and $v_h.w_h$ are bounded. If

$$p = 26/21,\ p' = 26/5, \qquad q = 14/13,\ q' = 14, \qquad r = 13/7,\ r' = 13/6,$$

then $1/2 = 1/p' + 2/dr'$ when $d = 3$, so the terms $\|v_h\|_{L^{r'}[0,T,L^{p'}(\Omega)]}$ and

$$|v_h.w_h|_{L^{rq}[0,T;W^{1,pq}(\Omega)]} = |v_h.w_h|_{L^2[0,T;W^{1,4/3}(\Omega)]}$$

are bounded, and the second term in e_h vanishes as $h \to 0$. ∎

4. Projections of the velocity field. In order to guarantee that the piecewise constant DG scheme is monotone and convergent, the average divergence of the velocity field in (2.2) must vanish on each simplex $K \in \mathcal{T}_h$. Above we assumed space P_h contains the piecewise constant functions so that solution v_h of the approximate momentum equation (2.3) automatically satisfies this condition. In this section projections of the velocity field $v_h \in V_h$ onto a space $\bar{V}_h \subset H(\Omega;\mathrm{div})$ having average divergence on each element equal to zero are considered when P_h does not contain

the piecewise constant functions. In this case the density and velocity/pressure are approximated by $\rho^n \in \mathcal{R}_h$ satisfying

$$\int_K \rho^n \psi^n + \tau \int_{\partial K} \left(\rho^n_-(\bar{v}^{n-1}.n)^+ + \rho^n_+(\bar{v}^{n-1}.n)^- \right) \psi^n = \int_K \rho^{n-1}\psi^n, \tag{4.1}$$

for $K \in \mathcal{T}_h$ and $\psi^n \in \mathbb{R}$, and $(v^n, p^n) \in V_h \times P_h$ satisfying

$$\begin{aligned} \frac{1}{2}\int_\Omega \rho^{n-1}\left(\frac{v^n - v^{n-1}}{\tau}\right).w + (\rho^n \bar{v}^{n-1}.\nabla)v^n.w \\ + \frac{1}{2}\int_\Omega \left(\frac{(\rho v)^n - (\rho v)^{n-1}}{\tau}\right).w - (\rho^n \bar{v}^{n-1}.\nabla)w.v^n \\ + \int_\Omega -p^n \operatorname{div}(w) + \mu^n D(v^n)\cdot D(w) = \int_\Omega \rho^n f^n.w, \end{aligned} \tag{4.2}$$

$$\int_\Omega \operatorname{div}(v^n)\, q = 0,$$

for all $(w, q) \in V_h \times P_h$.

Writing $\bar{v}^{n-1} = P_{\bar{V}_h} v^{n-1}$, where

$$\bar{V}_h \subset \left\{ v_h \in H(\Omega; \operatorname{div}) \mid \int_K \operatorname{div}(v_h) = 0,\ K \in \mathcal{T}_h \right\},$$

examining the proofs shows that the modified scheme will also converge if the projection $P_{\bar{V}_h} : V_h \to \bar{V}_h$ satisfies the following hypotheses.

ASSUMPTION 4.1.

1. *There exists $\ell \in \mathbb{N}$ independent of h such that $\bar{v}_h|_K \in \mathcal{P}_\ell(K)$ for each $K \in \mathcal{T}_h$.*
2. *For each $\bar{v}_h \in \bar{V}_h$*
$$\int_K \operatorname{div}(\bar{v}_h) = 0, \text{and } \int_{\partial K \cap \partial\Omega} \bar{v}_h.n = 0, \qquad K \in \mathcal{T}_h.$$
3. *If $v_h \in V_h$ and $\bar{v}_h = P_{\bar{V}_h} v_h$, then there exists $C > 0$ independent of h such that $\|\operatorname{div}(\bar{v}_h)\|_{L^2(\Omega)} \le C\|v_h\|_{H^1(\Omega)}$.*
4. *If $v_h \in V_h$ and $\bar{v}_h = P_{\bar{V}_h} v_h$, then there exists $C > 0$ independent of h such that $\|\bar{v}_h\|_{L^2(\Omega)} \le C\|v_h\|_{L^2(\Omega)}$ and $\|\bar{v}_h\|_{L^6(\Omega)} \le C\|\nabla v_h\|_{H^1(\Omega)}$.*
5. *Let $\{v_h\}_{h>0}$, $v_h \in V_h$ be bounded in $L^\infty[0,T;L^2(\Omega)] \cap L^2[0,T;H^1_0(\Omega)]$, and $\bar{v}_h = P_{\bar{V}_h} v_h$. If $v_h \to v$ in $L^2[0,T;L^2(\Omega)]$, then $\bar{v}_h \to v$.*

Stokes projections. If $(\bar{V}_h, \bar{P}_h) \subset H^1_0(\Omega)^d \times L^2(\Omega)/\mathbb{R}$ is a family of finite element spaces constructed on $\mathcal{T}_h$ which satisfies the Babuska–Brezzi condition, and if $\bar{P}_h$ contains the piecewise constant functions, then the Stokes projection $P_{\bar{V}_h} : V_h \to \bar{V}_h$ satisfies Assumption 4.1.

The Stokes projection of $v_h \in V_h$ is computed from the unique solution $(\bar{v}_h, \bar{p}_h) \in (\bar{V}_h, \bar{P}_h)$ of

$$a(\bar{v}_h, \bar{w}_h) + b(\bar{p}_h, \bar{w}_h) + b(\bar{q}_h, \bar{v}_h) = a(v_h, \bar{w}_h) \tag{4.3}$$

for all $(\bar{w}_h, \bar{q}_h) \in (\bar{V}_h, \bar{P}_h)$. The bilinear forms $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ and $b : L^2(\Omega) \times H^1(\Omega) \to \mathbb{R}$ are defined by

$$a(v, w) = (v, w)_{H^1(\Omega)}, \qquad b(p, v) = (p, \operatorname{div}(v))_{L^2(\Omega)}.$$

By construction the average of $\operatorname{div}(\bar{v}_h)$ vanishes on each simplex $K \in \mathcal{T}_h$. The next lemma shows that the continuity properties of Assumption 4.1 are also satisfied by this construction.

LEMMA 4.2. *Let $\Omega \subset \mathbb{R}^d$ be sufficiently regular to guarantee $H^2(\Omega)^d \times H^1(\Omega)$ regularity of the Stokes operator, and let $\{\mathcal{T}_h\}_{h>0}$ be a regular quasi-uniform family of triangulations of Ω.*

Let (V_h, P_h) and $(\bar{V}_h, \bar{P}_h) \subset H_0^1(\Omega)^d \times L^2(\Omega)/\mathbb{R}$ be families of finite element spaces constructed on $\mathcal{T}_h$ which satisfy the Babuska–Brezzi condition, and let $(\bar{v}_h, \bar{p}_h) \in (\bar{V}_h, \bar{P}_h)$ be the Stokes projection of a velocity field $v_h \in V_h$ satisfying $b(q_h, v_h) = 0$ for all $q_h \in P_h$. Then

- $\|\bar{v}_h\|_{H^1(\Omega)} \le \|v_h\|_{H^1(\Omega)}$, *and*
- $\|\bar{v}_h - v_h\|_{L^2(\Omega)} \le C\|v_h\|_{H^1(\Omega)} h \le C\|v_h\|_{L^2(\Omega)}$.

The first statement of the lemma follows upon setting $\bar{w}_h = \bar{v}_h$ in (4.3), and the Aubin–Nitsche trick and inverse inequalities are used to establish the second statement. The Sobolev embedding theorem guarantees

$$\|\bar{v}_h\|_{L^6(\Omega)} \le C\|\bar{v}_h\|_{H^1(\Omega)} \le C\|v_h\|_{H^1(\Omega)}.$$

It follows that the Stokes projection $v_h \to \bar{v}_h$ satisfies Assumption 4.1.

Acknowledgment. The authors thank Professor Francisco Guillen-Gonzalez of Universidad de Sevilla, who found a gap in one of the proofs of the original version of this manuscript.

REFERENCES

[1] S. N. ANTONTSEV, A. V. KAZHIKHOV, AND V. N. MONAKHOV, *Boundary Value Problems in Mechanics of Nonhomogeneous Fluids*, Stud. Math. Appl. 22, North-Holland, Amsterdam, 1990. Translated from the Russian.

[2] Y. C. CHANG, T. Y. HOU, B. MERRIMAN, AND S. OSHER, *A level set formulation of Eulerian interface capturing methods for incompressible fluid flows*, J. Comput. Phys., 124 (1996), pp. 449–464.

[3] P. G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[4] B. DESJARDINS, *Regularity results for two-dimensional flows of multiphase viscous fluids*, Arch. Rational Mech. Anal., 137 (1997), pp. 135–158.

[5] R. J. DIPERNA AND P. L. LIONS, *Ordinary differential equations, transport theory, and Sobolev spaces*, Invent. Math., 98 (1989), pp. 511–547.

[6] H. FUJITA AND T. KATO, *On the Navier-Stokes initial value problem.* I, Arch. Rational Mech. Anal., 16 (1964), pp. 269–315.

[7] J. GLIMM, E. ISAACSON, D. MARCHESIN, AND O. MCBRYAN, *Front tracking for hyperbolic systems*, Adv. in Appl. Math., 2 (1981), pp. 91–119.

[8] J. GLIMM, X. L. LI, Y. LIU, AND N. ZHAO, *Conservative front tracking and level set algorithms*, Proc. Natl. Acad. Sci. USA, 98 (2001), pp. 14198–14201.

[9] J. JAFFRÉ, C. JOHNSON, AND A. SZEPESSY, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci., 5 (1995), pp. 367–386.

[10] S. N. KRUZKOV, *First order quasilinear equations with several independent variables*, Math Sb. (N.S.) 81 (1970), pp. 228–255.

[11] N. N. KUZNETSOV, *Accuracy of some approximate methods for computing the weak solutions of a first-order quasi-linear equation*, USSR Comput. Math. Math. Phys., 16 (1976), pp. 105–119.

[12] P. LESAINT AND P.-A. RAVIART, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, (Proceedings of the Symposium, Mathematical Research Center, University of Wisconsin, Madison, WI, 1974), Academic Press, New York, 1974, pp. 89–123.

[13] J. L. Lions, *On some questions in boundary value problems of mathematics physics*, in Contempory Developments in Continuum Mechanics and Partial Differential Equations, North-Holland Math. Stud. 30, North-Holland, Amsterdam, 1978, pp. 283–346.

[14] J. L. Lions and E. Magenes, *Nonhomogeneous Boundary Value Problems and Applications. Vol.* I, Translated from the French by P. Kenneth, Die Grundlehren der Mathematischen Wissenschaften, Band 181, Springer-Verlag, New York, 1972.

[15] P. L. Lions, *Mathematical Topics in Fluid Mechanics, Volume* 1: *Incompressible Models*, Oxford University Press, Oxford, UK, 1996.

[16] B. Merriman, J. Bence, and S. Osher, *Motion of multiple junctions: A level set approach*, J. Comput. Phys., 112 (1994), pp. 334–363.

[17] S. Osher and J. A. Sethian, *Fronts propagating with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations*, J. Comput. Phys., 79 (1988), pp. 12–49.

[18] N. J. Walkington, *Convergence of the discontinuous Galerkin method for discontinuous solutions*, SIAM J. Numer. Anal., 42 (2004), pp. 1801–1817.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1305–1329

UNIFORM ESTIMATES FOR EULERIAN–LAGRANGIAN METHODS FOR SINGULARLY PERTURBED TIME-DEPENDENT PROBLEMS*

HONG WANG† AND KAIXIN WANG‡

Abstract. We prove a priori optimal-order error estimates in a weighted energy norm for several Eulerian–Lagrangian methods for singularly perturbed, time-dependent convection-diffusion equations with full regularity. The estimates depend only on certain Sobolev norms of the initial and right-hand side data, but not on ε or any norm of the true solution, and so hold uniformly with respect to ε. We use the interpolation of spaces and stability estimates to derive an ε-uniform estimate for problems with minimal or intermediate regularity, where the convergence rates are proportional to certain Besov norms of the initial and right-hand side data.

Key words. characteristic methods, convergence analysis, error estimates, Eulerian–Lagrangian methods, interpolation of spaces, uniform error estimates

AMS subject classifications. 65M12, 65M25, 65M60, 76M10, 76S05

DOI. 10.1137/060652816

1. Introduction.

Time-dependent convection-diffusion equations arise in mathematical models of petroleum reservoir simulation, environmental modeling, and other applications [10, 12]. These problems admit solutions with moving fronts and complex structures and present serious mathematical and numerical difficulties. Classical finite difference or finite element methods tend to generate numerical solutions with nonphysical oscillations, while upwind methods often produce excessive numerical diffusion that smears out fronts and generates spurious grid orientation effects [10].

Eulerian–Lagrangian methods combine the convection and capacity terms in the governing equations to carry out the temporal discretization in a Lagrangian coordinate and discretize the diffusion term on a fixed mesh [6, 8, 16, 17, 19]. These methods symmetrize the governing equation and stabilize their numerical approximations. They generate accurate numerical solutions and significantly reduce the numerical diffusion and grid-orientation effect present in upwind methods, even if large time steps and coarse spatial meshes are used. Eulerian–Lagrangian methods were shown to be very competitive in terms of accuracy and efficiency [6, 17, 19].

Optimal-order error estimates were derived for various Eulerian–Lagrangian methods [1, 7, 8, 15, 18]. This type of estimates has drawn debates for two potential problems: The general constant may depend inversely on the parameter ε. Further, the smoothness norms of the true solutions on the right side depend inversely on the parameter ε. Consequently, these estimates could blow up as ε tends to zero.

The goal of the present paper is to derive a priori optimal-order error estimates in an ε-weighted energy norm for Eulerian–Lagrangian methods for singularly perturbed, time-dependent convection-diffusion equations with full regularity. The estimates depend only on certain Sobolev norms of the initial and right-hand side data but not

*Received by the editors February 23, 2006; accepted for publication (in revised form) February 16, 2007; published electronically June 1, 2007.
http://www.siam.org/journals/sinum/45-3/65281.html

†School of Mathematics and System Sciences, Shandong University, Jinan, Shandong 250100, China and Department of Mathematics, University of South Carolina, Columbia, SC 29208 (hwang@math.sc.edu).

‡Corresponding Author. School of Mathematics and System Sciences, Shandong University, Jinan, Shandong 250100, China (kx.wang@mail.sdu.edu.cn).

on ε or any norm of the true solution. The general constant in the estimate does not depend on ε either. Thus, these estimates avoid the problems in the standard estimates. We then use the interpolation of spaces and stability estimates to derive an ε-uniform estimate for problems with minimal or intermediate regularity, where the convergence rates are proportional to certain Besov norms of the initial and right-hand side data.

This paper is organized as follows. In section 2 we recall preliminary results on Sobolev and Besov spaces and interpolation of spaces. In section 3 we revisit the Eulerian-Lagrangian localized adjoint method (ELLAM), the modified method of characteristics (MMOC), and the modified method of characteristics with adjusted advection (MMOCAA). In section 4 we prove an ε-uniform optimal-order error estimate for problems with full regularity. In section 5 we derive ε-uniform error estimates for problems with minimal or intermediate regularity. In section 6 we prove auxiliary lemmas. In section 7 we prove uniform stability of the true solutions in various smoothness norms. Section 8 contains concluding remarks.

2. Model problem and preliminaries. We consider a singularly perturbed, time-dependent convection-diffusion equation in one space dimension

$$\begin{aligned} u_t + (V(x,t)u - \varepsilon D(x,t)u_x)_x &= f(x,t), \quad (x,t) \in (a,b) \times (0,T), \\ u(x,0) &= u_o(x), \qquad x \in [a,b]. \end{aligned} \tag{2.1}$$

Here $V(x,t)$ is a velocity field, $f(x,t)$ accounts for external sources and sinks, $u_o(x)$ is prescribed initial data, and $u(x,t)$ is the ε-dependent unknown function. $D(x,t)$ is a diffusion coefficient with

$$0 < D_{min} \le D(x,t) \le D_{max} < +\infty \qquad \forall (x,t) \in [a,b] \times [0,T].$$

Here $0 < \varepsilon << 1$ is a parameter that scales the diffusion and characterizes the convection dominance of (2.1).

Such Eulerian–Lagrangian methods as the MMOC [8] and the MMOCAA [6, 7] were developed and analyzed for problem (2.1) with periodic boundary conditions. Other methods, such as the ELLAM [3, 18], could handle more general boundary conditions. We analyze these methods in a unified framework and close problem (2.1) with periodic boundary conditions at $x = a$ and $x = b$. This would require that all data functions in the problem are periodic. The assumption of periodicity of the problem may in principle exclude the appearance of boundary layers.

2.1. Sobolev spaces and approximation properties. Let $W_p^k(a,b)$ consist of functions whose weak derivatives up to order-k are pth Lebesgue integrable in (a,b). Let $H^k(a,b) := W_2^k(a,b)$ and $H_E^1(a,b)$ be a subspace of $H^1(a,b)$ with period $b-a$. We introduce an ε-weighted Sobolev norm for any $v \in H^k(a,b)$

$$\|v\|_{H_\varepsilon^k(a,b)} := \left(\|v\|^2_{H^{k-1}(a,b)} + \varepsilon \left\| \frac{d^k v}{dx^k} \right\|^2_{L^2(a,b)} \right)^{1/2}. \tag{2.2}$$

For any Banach space X, we introduce Sobolev spaces involving time [9]

$$W_p^k(t_1,t_2;X) := \left\{ f : \left\| \frac{\partial^\alpha f}{\partial t^\alpha}(\cdot,t) \right\|_X \in L^p(t_1,t_2),\ 0 \le \alpha \le k,\ 1 \le p \le \infty \right\},$$

$$\|f\|_{W_p^k(t_1,t_2;X)} := \begin{cases} \left(\displaystyle\sum_{\alpha=0}^{k} \int_{t_1}^{t_2} \left\| \frac{\partial^\alpha f}{\partial t^\alpha}(\cdot,t) \right\|_X^p dt \right)^{1/p}, & 1 \le p < \infty, \\ \displaystyle\max_{0 \le \alpha \le k} \operatorname*{ess\,sup}_{(t_1,t_2)} \left\| \frac{\partial^\alpha f}{\partial t^\alpha}(\cdot,t) \right\|_X, & p = \infty. \end{cases}$$

We define a uniform space-time partition on $[a,b]\times[0,T]$: $x_i := a+ih$ for $0\le i\le I$, with $h := (b-a)/I$, and $t_n := n\Delta t$ for $0\le n\le N$, with $\Delta t := T/N$. If a function $f(x,t)$ is defined only at discrete time steps t_n, we understand that the function f has been extended by constant to the time interval $(t_{n-1},t_n]$. Thus, the preceding space-time norm reduces to the following equivalent discrete norm:

$$\|f\|_{L^p(0,T;X)} := \begin{cases} \Big(\sum_{n=1}^{N} \|f(\cdot,t_n)\|_X^p \Delta t\Big)^{1/p}, & 1\le p<\infty,\\ \max\limits_{0\le n\le N} \|f(\cdot,t_n)\|_X, & p=\infty. \end{cases}$$

We also introduce the following ε-weighted energy norms:

$$\|f\|_{L_\varepsilon(0,T;H^k(a,b))} := \|f\|_{L^\infty(0,T;H^{k-1}(a,b))} + \sqrt{\varepsilon}\|f\|_{L^2(0,T;H^k(a,b))},$$
$$\|f\|_{L_\varepsilon(0,T;H^1_D(a,b))} := \|f\|_{L^\infty(0,T;L^2(a,b))} + \sqrt{\varepsilon}\|D^{1/2}f_x\|_{L^2(0,T;L^2(a,b))}.$$

Let $S_h(a,b)\subset H^1_E(a,b)$ be the finite element space that consists of continuous and piecewise-linear functions with respect to the spatial partition in $[a,b]$. We let $\Pi_h v\in S_h(a,b)$ be the piecewise-linear interpolation of v for any $v\in H^1_E(a,b)$. The following estimates hold [4, 5]:

$$\begin{aligned} \|\Pi_h v - v\|_{H^k(a,b)} &\le C_1 h^{2-k}\ \|v\|_{H^2(a,b)} && \forall v\in H^2(a,b),\ k=0,1,\\ \|v_h\|_{H^1(a,b)} &\le C_2 h^{-1}\|v_h\|_{L^2(a,b)} && \forall v_h\in S_h(a,b),\\ \|v_h\|_{L^\infty(a,b)} &\le C_2 h^{-1/2}\|v_h\|_{L^2(a,b)} && \forall v_h\in S_h(a,b). \end{aligned} \tag{2.3}$$

2.2. Besov spaces and interpolation of operators. The Besov spaces provide a finer scale and characterization of smoothness of functions than the Sobolev spaces do. We cite the results used in this paper and refer readers to [2, 5] for details.

For $\alpha>0$, $k := \lfloor\alpha\rfloor+1$, and $0<q\le\infty$, the Besov space $B^\alpha_q(L^p(a,b))$ consists of functions $f\in L^p(a,b)$ (for $p<\infty$) or $f\in C[a,b]$, the space of continuous functions on $[a,b]$, (for $p=\infty$) such that

$$\|f\|_{B^\alpha_q(L^p(a,b))} := \begin{cases} \|f\|_{L^p(a,b)} + \Big[\int_0^\infty [\theta^{-\alpha}\omega_k(f,\theta)_p]^q \dfrac{d\theta}{\theta}\Big]^{1/q}, & 0<q<\infty,\\ \|f\|_{L^p(a,b)} + \sup\limits_{\theta>0}\ \theta^{-\alpha}\omega_k(f,\theta)_p, & q=\infty, \end{cases}$$

is finite. Here the kth modulus of smoothness of function f is defined as

$$\omega_k(f,\theta)_p := \sup_{|h|\le\theta} \|\Delta_h^k f\|_{L^p(a+k|h|,b-k|h|)} \quad \text{with } \Delta_h f(x) := f(x+h)-f(x).$$

It is known that $B^\alpha_{q_1}(L^p(a,b))\hookrightarrow B^\alpha_{q_2}(L^p(a,b))$ for $q_1<q_2$ and that $B^\alpha_2(L^2(a,b)) = H^\alpha(a,b)$ with equivalent norms.

Let $X_1\hookrightarrow X_0$ be Banach spaces. We define the K-functional for $f\in X_0$ by

$$K(f,s) := K(f,s;X_0,X_1) := \inf_{g\in X_1}\{\|f-g\|_{X_0} + s\|g\|_{X_1}\},\quad s\ge 0.$$

The interpolation space $[X_0,X_1]_{s,q}$ consists of all functions $f\in X_0$ such that

$$\|f\|_{s,q} := \begin{cases} \Big[\int_0^\infty [\theta^{-s}K(f,\theta)]^q \dfrac{d\theta}{\theta}\Big]^{1/q}, & 0<q<\infty,\\ \sup\limits_{\theta>0}\ \theta^{-s}K(f,\theta), & q=\infty. \end{cases}$$

It is known that $X_1 \hookrightarrow [X_0, X_1]_{s,q} \hookrightarrow X_0$. The following lemmas characterize interpolation spaces.

LEMMA 2.1 (interpolation of Sobolev spaces). *Let m be a positive integer and $1 \le p \le \infty$. For any $0 < s < 1$ and $0 < q \le \infty$, the following relations hold:*

$$[L^2(a,b), H^m(a,b)]_{s,q} = B_q^{sm}(L^2(a,b)).$$

LEMMA 2.2 (interpolation of operators). *Let $X_1 \hookrightarrow X_0$ and Y be Banach spaces. If T is a bounded linear operator from X_i to Y with norm $M_i\,(i = 0, 1)$, then T is a bounded linear operator from the interpolation space $[X_0, X_1]_{s,q}$ to Y with a norm not exceeding $M_0^{1-s} M_1^s$ for any $0 < s < 1$ and $0 < q \le \infty$.*

LEMMA 2.3 (reiteration theorem). *Let $Y_i = [X_0, X_1]_{s_i,q_i}\,(i = 0, 1)$, with $0 < s_0 < s_1 < 1,\ 0 < q_0, q_1 \le \infty$. For any $0 < \beta < 1,\ 0 < r \le \infty$, we have*

$$[Y_0, Y_1]_{\beta,r} = [X_0, X_1]_{\beta',r}, \quad \beta' := (1-\beta)s_0 + \beta s_1,$$

with equivalent norms.

3. Revisit of Eulerian–Lagrangian methods. The ELLAM, MMOC, and MMOCAA schemes use a time-marching approach, so we need only to define these methods at the current time interval $[t_{n-1}, t_n]$.

3.1. The ELLAM. In the ELLAM formulation, the space-time test functions $w(x,t)$ are chosen to be continuous and piecewise smooth and to vanish outside the space-time strip $[a,b] \times (t_{n-1}, t_n]$. In particular, the test functions $w(x,t)$ satisfy $w(x,t_n) = \lim_{t \to t_n - 0} w(x,t)$, but $w(x,t_{n-1}) \ne \lim_{t \to t_{n-1}+0} w(x,t)$ in general. In this case, we use the notation $w(x, t_{n-1}^+) = \lim_{t \to t_{n-1}+0} w(x,t)$ to account for the possible discontinuity of $w(x,t)$ in time at time t_{n-1}.

We multiply (2.1) by test functions w and integrate the resulting equation on $[a,b] \times (t_{n-1}, t_n]$ to obtain a weak formulation

$$\begin{aligned} &\int_a^b u(x,t_n) w(x,t_n)\,dx + \int_{t_{n-1}}^{t_n} \int_a^b \varepsilon D(x,t) u_x(x,t) w_x(x,t)\,dx\,dt \\ &\quad - \int_{t_{n-1}}^{t_n} \int_a^b u(x,t)\Big(w_t(x,t) + V(x,t) w_x(x,t)\Big)\,dx\,dt \\ &\quad = \int_a^b u(x,t_{n-1}) w(x,t_{n-1}^+)\,dx + \int_{t_{n-1}}^{t_n} \int_a^b f(x,t) w(x,t)\,dx\,dt. \end{aligned} \tag{3.1}$$

In the ELLAM framework [3] the test functions w are chosen to satisfy the adjoint equation of the hyperbolic part of (2.1) to define the temporal variation of w

$$w_t + V w_x = 0. \tag{3.2}$$

This implies the test functions w to be constant along the characteristic curve $r(t; x, t_n)$. Here $r(t; \bar{x}, \bar{t})$ refers to the characteristic curve passing $\bar{x}$ at time $\bar{t}$ defined by

$$\frac{dr}{dt} = V(r,t), \quad r(t; \bar{x}, \bar{t})\Big|_{t=\bar{t}} = \bar{x}. \tag{3.3}$$

Thus, once the test functions $w(x,t)$ are specified in $[a,b]$ at time step t_n, they are determined completely in the space-time strip $[a,b] \times (t_{n-1}, t_n]$.

3.1.1. Evaluation of diffusion and source terms. Note that $V(x,t)$ is a periodic function with respect to x of the period $b-a$. Thus, the shifted characteristic curve $r^S(t;b,t_n) := r(t;b,t_n) - (b-a)$ satisfies the initial-value problem

$$\begin{aligned} \frac{dr^S(t;b,t_n)}{dt} &= \frac{dr(t;b,t_n)}{dt} = V(r(t;b,t_n),t) = V(r^S(t;b,t_n),t), \\ r^S(t;b,t_n)\Big|_{t=t_n} &= b-(b-a) = a. \end{aligned}$$

Therefore, both $r^S(t;b,t_n)$ and $r(t;a,t_n)$ are the solutions of the same initial-value problem. The uniqueness of such a problem concludes that

$$r(t;b,t_n) - r(t;a,t_n) = b-a \qquad \forall t \in [t_{n-1},t_n]. \tag{3.4}$$

For clarity of presentation, in the evaluation of source and diffusion terms we reserve x for points in $[a,b]$ at time t_n representing the heads of characteristics. We use the variable y to represent the spatial coordinate of an arbitrary point at time $t \in (t_{n-1},t_n)$. We use the relation (3.4) and the periodicity of problem (2.1) to evaluate the source term by the Euler quadrature as follows:

$$\begin{aligned} &\int_{t_{n-1}}^{t_n}\int_a^b f(y,t)w(y,t)dydt \\ &\quad= \int_{t_{n-1}}^{t_n}\int_{r(t;a,t_n)}^{r(t;b,t_n)} f(y,t)w(y,t)dydt \\ &\quad= \int_a^b\int_{t_{n-1}}^{t_n} f(r(t;x,t_n),t)w(r(t;x,t_n),t)r_x(t;x,t_n)dtdx \\ &\quad= \int_a^b\Big[\int_{t_{n-1}}^{t_n} f(r(t;x,t_n),t)r_x(t;x,t_n)dt\Big]w(x,t_n)dx \\ &\quad= \Delta t\int_a^b f(x,t_n)w(x,t_n)dx + E_1(w). \end{aligned} \tag{3.5}$$

Here $E_1(w)$ is the local truncation error defined by

$$E_1(w) := \int_a^b\int_{t_{n-1}}^{t_n}\Big[f(r(t;x,t_n),t)r_x(t;x,t_n) - f(x,t_n)\Big]dt\ w(x,t_n)dx. \tag{3.6}$$

We evaluate the diffusion term similarly

$$\begin{aligned} &\int_{t_{n-1}}^{t_n}\int_a^b \varepsilon D(y,t)u_y(y,t)w_y(y,t)dydt \\ &\quad= \int_{t_{n-1}}^{t_n}\int_{r(t;a,t_n)}^{r(t;b,t_n)} \varepsilon D(y,t)u_y(y,t)w_y(y,t)dydt \\ &\quad= \int_a^b\int_{t_{n-1}}^{t_n} \varepsilon D(r(t;x,t_n),t)u_y(r(t;x,t_n),t)w_y(x,t_n)r_x(t;x,t_n)dtdx \\ &\quad= \int_a^b\int_{t_{n-1}}^{t_n} \varepsilon D(r(t;x,t_n),t)u_y(r(t;x,t_n),t)w_x(x,t_n)dtdx \\ &\quad= \varepsilon\Delta t\int_a^b D(x,t_n)u_x(x,t_n)w_x(x,t_n)dx + \varepsilon E_2(u,w). \end{aligned} \tag{3.7}$$

Here $E_2(u,w)$ is the local truncation error defined by

$$E_2(u,w) := \int_a^b\int_{t_{n-1}}^{t_n}\Big[(Du_x)(r(t;x,t_n),t) - (Du_x)(x,t_n)\Big]dt\ w_x(x,t_n)dx. \tag{3.8}$$

3.1.2. ELLAM formulation and numerical scheme. We substitute (3.5) and (3.7) into (3.1) to obtain an ELLAM formulation for problem (2.1):

$$\begin{aligned}\int_a^b u(x,t_n)w(x,t_n)dx + \varepsilon\Delta t\int_a^b D(x,t_n)u_x(x,t_n)w_x(x,t_n)dx \\ = \int_a^b u(x^*,t_{n-1})w(x,t_n)r_x(t_{n-1};x,t_n)dx \\ +\Delta t\int_a^b f(x,t_n)w(x,t_n)dx + E_1(w) - \varepsilon E_2(u,w).\end{aligned} \tag{3.9}$$

Here x^* is the foot of the characteristic curve $r(t;x,t_n)$ backtracking from x at time t_n. We also let $\tilde{x}$ be the head of the characteristic curve $r(t;\tilde{x},t_n)$ at time t_n that backtracks to x at time t_{n-1}:

$$x^* = r(t_{n-1};x,t_n), \qquad x = r(t_{n-1};\tilde{x},t_n). \tag{3.10}$$

In (3.9) we have used the periodicity of the problem, a relation similar to (3.4), and the fact that w is constant along the characteristics to rewrite the first integral at time t_{n-1} on the right-hand side of (3.1) as an integral at time t_n in (3.9):

$$\begin{aligned}\int_a^b u(y,t_{n-1})w(y,t_{n-1}^+)dy &= \int_{\tilde{a}}^{\tilde{b}} u(x^*,t_{n-1})w(x,t_n)r_x(t_{n-1};x,t_n)dx \\ &= \int_a^b u(x^*,t_{n-1})w(x,t_n)r_x(t_{n-1};x,t_n)dx.\end{aligned} \tag{3.11}$$

The ELLAM scheme is derived based on (3.9). Note that the characteristics $r(t;x,t_n)$ cannot be tracked exactly, in general, so the test functions w_h in the ELLAM scheme are defined to be constant along the approximate characteristics $r_h(t;x,t_n)$. Here $r_h(t;\bar{x},\bar{t})$ is defined by

$$r_h(t;\bar{x},\bar{t}) = \bar{x} + V(\bar{x},\bar{t})(t-\bar{t}). \tag{3.12}$$

Consequently, the ELLAM scheme states as follows: Find $u_h(x,t_n) \in S_h(a,b)$ for $n = 1,\dots,N$ such that for any $w_h(x,t_n) \in S_h(a,b)$

$$\begin{aligned}\int_a^b u_h(x,t_n)w_h(x,t_n)dx + \varepsilon\Delta t\int_a^b D(x,t_n)u_{h,x}(x,t_n)w_{h,x}(x,t_n)dx \\ = \int_a^b u_h(x_h^*,t_{n-1})w_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx + \Delta t\int_a^b f(x,t_n)w_h(x,t_n)dx.\end{aligned} \tag{3.13}$$

Here x_h^* and $\tilde{x}_h$ are defined by

$$x_h^* = r_h(t_{n-1};x,t_n), \qquad x = r_h(t_{n-1};\tilde{x}_h,t_n). \tag{3.14}$$

The ELLAM, MMOC, MMOCAA, and virtually any other Eulerian–Lagrangian method typically need to impose the following type of constraint on the time step Δt:

$$\|V\|_{L^\infty(0,T;W^1_\infty)}\Delta t < 1. \tag{3.15}$$

This constraint guarantees that the approximate characteristics defined in (3.12), which are extended from different spatial points, do not intersect with each other during the time period $[t_{n-1},t_n]$. In other words, the traceback operator defined by the approximate characteristic tracking is a diffeomorphism. This condition will be used several times in the error estimates in the subsequent sections without being explicitly stated. This constraint can be alleviated if a micro time step Δt_f is used in the characteristic tracking. In this case, the Δt in (3.15) will be replaced by Δt_f.

3.2. The MMOC and MMOCAA. The MMOC and MMOCAA directly apply to a nonconservative analogue of (2.1):

$$u_t + V(x,t)u_x - (\varepsilon D(x,t)u_x)_x + V_x(x,t)u = f(x,t). \tag{3.16}$$

3.2.1. The MMOC. In the MMOC the capacity and convection terms in (3.16) are combined to form a material derivative at time step t_n, which is approximated by a backward difference quotient along the approximate characteristic $r_h(t;x,t_n)$ in the time stepping procedure [8]

$$\begin{aligned} &u_t(x,t_n) + V(x,t_n)u_x(x,t_n) \\ &\quad = \sqrt{1+V(x,t_n)^2}\frac{du}{dt}(x,t_n) = \frac{u(x,t_n) - u(x^*,t_{n-1})}{\Delta t} \\ &\qquad + \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} \sqrt{(r_h(t;x,t_n) - x_h^*)^2 + (t-t_{n-1})^2}\frac{d^2u}{dt^2}(r_h(t;x,t_n),t)dt. \end{aligned} \tag{3.17}$$

We incorporate (3.17) into (3.16) and multiply the equation by any test function $w \in H_E^1(a,b)$. We integrate the resulting equation on the interval (a,b), leading to an MMOC reference equation for problem (3.16): Find $u(x,t_n) \in H_E^1(a,b)$ for $n = 1,\dots,N$ such that for any $w(x) \in H_E^1(a,b)$

$$\begin{aligned} &\int_a^b \frac{u(x,t_n) - u(x_h^*,t_{n-1})}{\Delta t}w(x)dx + \int_a^b \varepsilon D(x,t_n)u_x(x,t_n)w_x(x)dx \\ &\quad + \int_a^b V_x(x,t_n)u(x,t_n)w(x)dx = \int_a^b f(x,t_n)w(x)dx - \frac{1}{\Delta t}E_3(u,w). \end{aligned} \tag{3.18}$$

Here $E_3(u,w)$ is the local truncation error of the MMOC reference equation

$$E_3(u,w) = \int_a^b w(x)\int_{t_{n-1}}^{t_n} \sqrt{(r_h(t;x,t_n) - x_h^*)^2 + (t-t_{n-1})^2}\frac{d^2u}{dt^2}(r_h(t;x,t_n),t)dtdx.$$

The MMOC scheme reads: Find $u_h(x,t_n) \in S_h(a,b)$ for $n = 1,\dots,N$ such that for any $w_h(x) \in S_h(a,b)$

$$\begin{aligned} &\int_a^b \frac{u_h(x,t_n) - u_h(x_h^*,t_{n-1})}{\Delta t}w_h(x)dx + \int_a^b \varepsilon D(x,t_n)u_{h,x}(x,t_n)w_{h,x}(x)dx \\ &\quad + \int_a^b V_x(x,t_n)u_h(x,t_n)w_h(x)dx = \int_a^b f(x,t_n)w_h(x)dx. \end{aligned} \tag{3.19}$$

3.2.2. The MMOCAA. The MMOCAA [6, 7] aims at eliminating the mass balance error in the MMOC [8]. Summing the MMOC scheme (3.19) for all of the test functions yields a mass balance satisfied by the MMOC solution $u_h(x,t_n)$:

$$\int_a^b (1 + \Delta t V_x(x,t_n))u_h(x,t_n)dx = \int_a^b u_h(x_h^*,t_{n-1})dx + \Delta t\int_a^b f(x,t_n)dx. \tag{3.20}$$

If we integrate (3.16) with $u(x,t_{n-1}) = u_h(x,t_{n-1})$ on $(a,b)\times[t_{n-1},t_n]$ and apply Euler quadrature at time t_n to the source term, we obtain a mass balance equation satisfied by the exact solution (up to the order of truncation error)

$$\int_a^b u(x,t_n)dx = \int_a^b u_h(x,t_{n-1})dx + \Delta t\int_a^b f(x,t_n)dx.$$

Let $Q_{n-1} = \int_a^b u_h(x,t_{n-1})dx$ and $Q^*_{n-1} = \int_a^b u_h(x^*_h,t_{n-1})dx - \Delta t \int_a^b V_x(x,t_n)$ $u_h(x,t_n)dx$. The MMOC scheme conserves mass if and only if $Q_{n-1} = Q^*_{n-1}$. To correct the mass balance error of the MMOC when $Q_{n-1} \neq Q^*_{n-1}$, we set for some fixed constant $\kappa > 0$

$$
\begin{aligned}
x^*_{h,+} &= x^*_h + \kappa V(x,t_n)(\Delta t)^2, \qquad x^*_{h,-} = x^*_h - \kappa V(x,t_n)(\Delta t)^2,\\
u^{\#}_h(x^*_h,t_{n-1}) &= \begin{cases} \max\{u_h(x^*_{h,+},t_{n-1}), u_h(x^*_{h,-},t_{n-1})\} \text{ if } Q^*_{n-1} \le Q_{n-1},\\ \min\{u_h(x^*_{h,+},t_{n-1}), u_h(x^*_{h,-},t_{n-1})\} \text{ if } Q^*_{n-1} > Q_{n-1}, \end{cases}\\
Q^{\#}_{n-1} &= \int_a^b u^{\#}_h(x^*_h,t_{n-1})dx.
\end{aligned} \tag{3.21}
$$

If Q_{n-1} can be expressed as a convex combination of Q^*_{n-1} and $Q^{\#}_{n-1}$, then $\check{u}_h(x^*_h,t_{n-1})$ defined by the same convex combination of $u_h(x^*_h,t_{n-1})$ and $u^{\#}_h(x^*_h,t_{n-1})$ will have mass Q_{n-1}. The $u_h(x^*_h,t_{n-1})$ in the MMOC scheme (3.19) is replaced by $\check{u}_h(x^*_h,t_{n-1})$ in the MMOCAA scheme with all other terms unchanged.

4. Error estimates for problem (2.1) with full regularity. We prove a priori optimal-order error estimates for the ELLAM, MMOC, and MMOCAA schemes for problem (2.1), which hold uniformly with respect to ε.

4.1. An optimal-order error estimate for the ELLAM scheme. Let the Courant number $\mathrm{Cr} := \max_{(x,t)\in[0,1]\times[0,T]} |V(x,t)|\Delta t/h$ and $\lambda = 1$ if $\mathrm{Cr} < 1$ or $= 0$ otherwise. The main result is given in the theorem below.

THEOREM 4.1. *Assume $D, V \in L^\infty(0,T;W^{3+\lambda}_\infty(a,b))$, $f \in L^2(0,T;H^{2+\lambda}(a,b))$, and $u_o \in H^{2+\lambda}(a,b)$. Then the following optimal-order error estimate of the ELLAM scheme holds uniformly with respect to ε:*

$$
\begin{aligned}
&\|u_h - u\|_{L_\varepsilon(0,T;H^1_D)}\\
&\quad \le C\Delta t\Big(\|u_o\|_{H^2_\varepsilon} + \Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)} + \|f\|_{L^2(0,T;H^1)}\Big)\\
&\qquad + C(\min\{h,\Delta t\} + h^2)\|u_o\|_{H^2} + C\lambda h^2(\|u_o\|_{H^3} + \|f\|_{L^2(0,T;H^3)}).
\end{aligned} \tag{4.1}
$$

Here the constant C is independent of u and the parameter ε.

Proof. We let $e = u_h - u$ and choose the test function $w(\cdot,t_n)$ in (3.9) to be $w_h(\cdot,t_n) \in S_h(a,b)$. We then subtract (3.13) from the ELLAM reference equation (3.9) to obtain an ELLAM error equation for any $w_h(x,t_n) \in S_h(a,b)$:

$$
\begin{aligned}
&\int_a^b e(x,t_n)w_h(x,t_n)dx + \varepsilon\Delta t\int_a^b D(x,t_n)e_x(x,t_n)w_{h,x}(x,t_n)dx\\
&= \int_a^b (u(x^*_h,t_{n-1})r_{h,x}(t_{n-1};x,t_n) - u(x^*,t_{n-1})r_x(t_{n-1};x,t_n))w_h(x,t_n)dx\\
&\quad + \int_a^b e(x^*_h,t_{n-1})w_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx - E_1(w_h) + \varepsilon E_2(u,w_h).
\end{aligned} \tag{4.2}
$$

Let $\Pi_h u \in S_h(a,b)$ be the interpolation of the true solution u, $\xi_h = u_h - \Pi_h u \in S_h(a,b)$, and $\eta = \Pi_h u - u$. The error estimates for η are given in (2.3), so we need

only to estimate ξ_h. We choose $w_h(x,t_n)=\xi_h(x,t_n)$ in (4.2) and rewrite the error equation in terms of ξ_h and η as follows:

$$
\begin{aligned}
&\int_a^b \xi_h^2(x,t_n)dx+\varepsilon\Delta t\int_a^b D(x,t_n)\xi_{h,x}^2(x,t_n)dx\\
&\quad=\int_a^b \xi_h(x_h^*,t_{n-1})\xi_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx\\
&\qquad+\int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx-\int_a^b \eta(x,t_n)\xi_h(x,t_n)dx\\
&\qquad-\varepsilon\Delta t\int_a^b D(x,t_n)\eta_x(x,t_n)\xi_{h,x}(x,t_n)dx-E_1(\xi_h)+\varepsilon E_2(u,\xi_h)\\
&\qquad+\int_a^b u(x^*,t_{n-1})(r_{h,x}(t_{n-1};x,t_n)-r_x(t_{n-1};x,t_n))\xi_h(x,t_n)dx\\
&\qquad+\int_a^b (u(x_h^*,t_{n-1})-u(x^*,t_{n-1}))r_{h,x}(t_{n-1};x,t_n)\xi_h(x,t_n)dx.
\end{aligned}
\tag{4.3}
$$

We bound the first term on the right-hand side of (4.3) by

$$
\begin{aligned}
&\Big|\int_a^b \xi_h(x_h^*,t_{n-1})\xi_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx\Big|\\
&\quad\le\frac{1+C\Delta t}{2}\int_a^b \xi^2(x,t_n)dx+\frac{1+C\Delta t}{2}\int_a^b \xi^2(x_h^*,t_{n-1})dx\\
&\quad\le\frac{1+C\Delta t}{2}\int_a^b \xi^2(x,t_n)dx+\frac{1+C\Delta t}{2}\int_{a_h^*}^{b_h^*} \xi^2(x_h^*,t_{n-1})\Big|\frac{dx_h^*}{dx}\Big|^{-1}dx_h^*\\
&\quad\le\frac{1+C\Delta t}{2}\|\xi_h(\cdot,t_n)\|_{L^2}^2+\frac{1+C\Delta t}{2}\|\xi_h(\cdot,t_{n-1})\|_{L^2}^2.
\end{aligned}
\tag{4.4}
$$

Here the constant C depends on $\|V\|_{L^\infty(0,T;W^1_\infty)}$. In the second term after the second inequality, we used the substitution of variables from x to x_h^* given by the first equation in (3.14) and changed the limits a and b of the integral to a_h^* and b_h^*, respectively. We also utilized (3.12) and the periodicity of V to conclude that

$$
\begin{aligned}
r_{h,x}(t_{n-1};x,t_n)&=1-V_x(x,t_n)\Delta t,\\
r_{h,x}^{-1}(t_{n-1};x,t_n)&=(1-V_x(x,t_n)\Delta t)^{-1}=1+O(\Delta t),\\
b_h^*-a_h^*&=(b-a)-(V(b,t_n)-V(a,t_n))\Delta t=b-a.
\end{aligned}
\tag{4.5}
$$

The estimate of the second and third terms on the right-hand side of (4.3) presents one of the major difficulties. Standard techniques yield

$$
\begin{aligned}
&\Big|\int_a^b (\eta(x,t_{n-1})-\eta(x_h^*,t_{n-1}))\xi_h(x,t_n)dx\Big|\\
&\quad=\Big|\int_a^b\int_{x_h^*}^x \eta_y(y,t_{n-1})dy\ \xi_h(x,t_n)dx\Big|\\
&\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2+C\Delta t\ h^2\|u\|^2_{L^\infty(0,T;W^2_\infty)},
\end{aligned}
$$

leading to a suboptimal-order error estimate of order $O(h+\Delta t)$ for the ELLAM scheme. This does not coincide with the optimal-order convergence rates observed numerically. A delicate analysis shows an optimal-order error estimate of the second and third terms on the right-hand side of (4.3). For clarity of exposition, the proof is

presented in Lemma 6.2; there we obtain

$$
\begin{aligned}
&\Big|\int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx - \int_a^b \eta(x,t_n)\xi_h(x,t_n)dx\Big| \\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C\Delta t(\min\{h^2,(\Delta t)^2\}+h^4)\|u\|_{L^\infty(0,T;H^2)}^2 \\
&\quad + C(\Delta t)^3\|u\|_{L^\infty(0,T;H^1)}^2 + C\lambda\, h^4(\|u\|_{H^1(t_{n-1},t_n;H^2)}^2 + \Delta t\|u\|_{L^\infty(0,T;H^3)}^2).
\end{aligned} \tag{4.6}
$$

Let $x_{i-1/2}$ be the middle point of the interval $[x_{i-1},x_i]$. Note that $\xi_{h,x}(x,t_n)$ is constant on each interval $[x_{i-1},x_i]$ and that η satisfies $\eta(x_{i-1},t_n)=\eta(x_i,t_n)=0$ for $i=1,\dots,I$. We bound the fourth term on the right-hand side of (4.3):

$$
\begin{aligned}
&\Big|\varepsilon\Delta t\int_a^b D(x,t_n)\eta_x(x,t_n)\xi_{h,x}(x,t_n)dx\Big| \\
&= \Big|\varepsilon\Delta t\sum_{i=1}^I \xi_{h,x}(x_{i-1/2},t_n)\int_{x_{i-1}}^{x_i}(D(x,t_n)-D(x_{i-1/2},t_n))\eta_x(x,t_n)dx\Big| \\
&\le \varepsilon\Delta t\ h\ \|D\|_{L^\infty(0,T;W^1_\infty)}\ \|\xi_{h,x}(\cdot,t_n)\|_{L^2}\ \|\eta_x(\cdot,t_n)\|_{L^2} \\
&\le \frac{1}{4}\varepsilon\Delta t\|\xi_{h,x}(\cdot,t_n)\|_{L_D^2}^2 + C\varepsilon\Delta t\, h^4\|u\|_{L^\infty(0,T;H^2)}^2.
\end{aligned} \tag{4.7}
$$

Here $\|\cdot\|_{L_D^2} = \|D^{1/2}\cdot\|_{L^2}$.

We use the estimate (6.1) to bound the fifth term on the right side of (4.3):

$$
\begin{aligned}
&\Big|\int_a^b\int_{t_{n-1}}^{t_n}\Big[f(r(t;x,t_n),t)r_x(t;x,t_n)-f(x,t_n)\Big]dt\,\xi_h(x,t_n)dx\Big| \\
&\le \int_a^b\int_{t_{n-1}}^{t_n}|f(x,t_n)-f(r(t;x,t_n),t)|dt\ |\xi_h(x,t_n)|dx \\
&\quad + \int_a^b\int_{t_{n-1}}^{t_n}|f(r(t;x,t_n),t)|\ |1-r_x(t;x,t_n)|dt\ |\xi_h(x,t_n)|dx \\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C(\Delta t)^2\Big(\Big\|\frac{df}{dt}\Big\|_{L^2(t_{n-1},t_n;L^2)}^2 + \|f\|_{L^2(t_{n-1},t_n;L^2)}^2\Big).
\end{aligned} \tag{4.8}
$$

We similarly bound the sixth term on the right-hand side of (4.3) by

$$
\begin{aligned}
&\Big|\varepsilon\int_a^b\int_{t_{n-1}}^{t_n}[(D\ u_x)(x,t_n)-(D\ u_x)(r(t;x,t_n),t)]dt\ \xi_{h,x}(x,t_n)dx\Big| \\
&= \Big|\varepsilon\int_a^b \xi_{h,x}(x,t_n)\Big[\int_{t_{n-1}}^{t_n}\int_t^{t_n}\frac{d}{d\theta}(D\ u_x)(r(\theta;x,t_n),\theta)d\theta dt\Big]dx\Big| \\
&\le \frac{1}{4}\varepsilon\Delta t\|\xi_{h,x}(\cdot,t_n)\|_{L_D^2}^2 + C\varepsilon(\Delta t)^2\Big(\Big\|\frac{du}{dt}\Big\|_{L^2(t_{n-1},t_n;H^1)}^2 + \|u\|_{L^2(t_{n-1},t_n;H^1)}^2\Big).
\end{aligned} \tag{4.9}
$$

We use the estimate (6.2) to bound the seventh term on the right side of (4.3) in a similar way to the estimate (4.4) to get

$$
\begin{aligned}
&\Big|\int_a^b u(x^*,t_{n-1})(r_{h,x}(t_{n-1};x,t_n)-r_x(t_{n-1};x,t_n))\xi_h(x,t_n)dx\Big| \\
&\le C(\Delta t)^2\|\xi_h(\cdot,t_n)\|_{L^2}\Big(\int_a^b u^2(x^*,t_{n-1})dx\Big)^{1/2} \\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C(\Delta t)^3\|u\|_{L^\infty(0,T;L^2)}^2.
\end{aligned} \tag{4.10}
$$

Let $\chi_{(\alpha,\beta)}$ be the indicator function of the interval (α,β), which is 1 on (α,β) or 0 elsewhere. We use the estimate (6.1) to bound the last term on the right-hand side

of (4.3):

$$
\begin{aligned}
&\Big|\int_a^b (u(x_h^*,t_{n-1})-u(x^*,t_{n-1}))r_{h,x}(t_{n-1};x,t_n)\xi_h(x,t_n)dx\Big|\\
&\quad\le C\int_a^b\Big|\int_{x^*}^{x_h^*}|u_y(y,t_{n-1})|dy\Big|\,|\xi_h(x,t_n)|dx\\
&\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}\Big(\int_a^b\int_a^b \chi(y)_{(x^*-C(\Delta t)^2,x^*+C(\Delta t)^2)}u_y^2(y,t_{n-1})dydx\Big)^{1/2}\\
&\quad= C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}\Big(\int_a^b\int_a^b \chi(x)_{(y-C(\Delta t)^2,y+C(\Delta t)^2)}dx\ u_y^2(y,t_{n-1})dy\Big)^{1/2}\\
&\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2+C(\Delta t)^3\|u\|^2_{L^\infty(0,T;H^1)}.
\end{aligned}
\tag{4.11}
$$

We substitute estimates (4.4)–(4.11) for the corresponding terms in (4.3) to obtain the following estimate:

$$
\begin{aligned}
&\|\xi_h(\cdot,t_n)\|_{L^2}^2+\varepsilon\Delta t\|\xi_{h,x}(\cdot,t_n)\|^2_{L^2_D}\\
&\quad\le\frac{1+C\Delta t}{2}\Big(\|\xi_h(\cdot,t_n)\|_{L^2}^2+\|\xi_h(\cdot,t_{n-1})\|_{L^2}^2\Big)+\frac{1}{2}\varepsilon\Delta t\|\xi_{h,x}(\cdot,t_n)\|^2_{L^2_D}\\
&\qquad+C(\Delta t)^2\Big(\varepsilon\Big\|\frac{du}{dt}\Big\|^2_{L^2(t_{n-1},t_n;H^1)}+\Delta t\|u\|^2_{L^\infty(0,T;H^1)}+\Big\|\frac{df}{dt}\Big\|^2_{L^2(t_{n-1},t_n;L^2)}\\
&\qquad\quad+\|f\|^2_{L^2(t_{n-1},t_n;L^2)}\Big)+C\Delta t\min\{h^2,(\Delta t)^2\}\|u\|^2_{L^\infty(0,T;H^2)}\\
&\qquad+Ch^4\Big(\Delta t\|u\|^2_{L^\infty(0,T;H^2)}+\lambda(\|u\|^2_{H^1(t_{n-1},t_n;H^2)}+\Delta t\|u\|^2_{L^\infty(0,T;H^3)})\Big).
\end{aligned}
$$

We sum the estimate for $n=1,\dots,N_1(\le N)$ and cancel like terms to obtain

$$
\begin{aligned}
&\|\xi_h(\cdot,t_{N_1})\|_{L^2}^2+\varepsilon\Delta t\sum_{n=1}^{N_1}\|\xi_{h,x}(\cdot,t_n)\|^2_{L^2_D}\\
&\quad\le C\Delta t\sum_{n=0}^{N_1-1}\|\xi_h(\cdot,t_n)\|_{L^2}^2+C(\Delta t)^2\Big(\varepsilon\Big\|\frac{du}{dt}\Big\|^2_{L^2(0,T;H^1)}+\|u\|^2_{L^\infty(0,T;H^1)}\\
&\qquad+\Big\|\frac{df}{dt}\Big\|^2_{L^2(0,T;L^2)}+\|f\|^2_{L^2(0,T;L^2)}\Big)+C(\min\{h^2,(\Delta t)^2\}+h^4)\|u\|^2_{L^\infty(0,T;H^2)}\\
&\qquad+C\lambda h^4(\|u\|^2_{H^1(0,T;H^2)}+\|u\|^2_{L^\infty(0,T;H^3)}).
\end{aligned}
$$

We then apply the Gronwall inequality to conclude

$$
\begin{aligned}
&\|\xi_h\|_{L_\varepsilon(0,T;H^1_D(a,b))}\\
&\quad\le C\Delta t\Big(\sqrt{\varepsilon}\Big\|\frac{du}{dt}\Big\|_{L^2(0,T;H^1)}+\|u\|_{L^\infty(0,T;H^1)}+\Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)}+\|f\|_{L^2(0,T;L^2)}\Big)\\
&\qquad+C(\min\{h,\Delta t\}+h^2)\|u\|_{L^\infty(0,T;H^2)}+C\lambda h^2(\|u\|_{H^1(0,T;H^2)}+\|u\|_{L^\infty(0,T;H^3)}).
\end{aligned}
$$

The general constant C depends exponentially on the final time T in problem (2.1), due to the application of the Gronwall inequality, but does not depend on the parameter ε. We combine this estimate with (2.3) and the stability estimate of the true solution in Theorem 7.2 to finish the proof. ∎

4.2. The optimal-order error estimate for the MMOC and MMOCAA schemes. We prove an optimal-order error estimate for the MMOC scheme (3.19) and outline a similar estimate for the MMOCAA scheme.

THEOREM 4.2. *Assume $D, V \in L^\infty(0,T;W^4_\infty(a,b))$, $f \in L^2(0,T;H^3(a,b))$, and $u_o \in H^3(a,b)$. Then the following optimal-order error estimate of the MMOC scheme holds uniformly with respect to ε:*

$$\begin{aligned}
&\|u_h - u\|_{L_\varepsilon(0,T;H^1_D)} \\
&\quad \le C\Delta t\Big(\|u_o\|_{H^2_\varepsilon} + \varepsilon\|u_o\|_{H^3_\varepsilon} + \Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)} + \|f\|_{L^2(0,T;H^2)}\Big) \\
&\qquad + C(\min\{h,\Delta t\} + h^2)\|u_o\|_{H^2} + C\lambda h^2(\|u_o\|_{H^3} + \|f\|_{L^2(0,T;H^3)}).
\end{aligned} \tag{4.12}$$

Here the constant C is independent of u and the parameter ε.

Proof. We let $e(x,t_n)$, $\xi_h(x,t_n)$, and $\eta(x,t_n)$ be defined as in section 4.1 and choose $w_h(x) = \xi_h(x,t_n)$ in the MMOC reference equation (3.18) and the MMOC scheme (3.19). We subtract the latter from the former and rewrite the equation in terms of ξ and η as follows:

$$\begin{aligned}
&\int_a^b \xi_h^2(x,t_n)dx + \varepsilon\Delta t\int_a^b D(x,t_n)\xi_{h,x}^2(x,t_n)dx + \Delta t\int_a^b V_x(x,t_n)\xi_h^2(x,t_n)dx \\
&\quad = \int_a^b \xi_h(x_h^*,t_{n-1})\xi_h(x,t_n)dx + \int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n)dx \\
&\qquad - \int_a^b \eta(x,t_n)\xi_h(x,t_n)dx - \varepsilon\Delta t\int_a^b D(x,t_n)\eta_x(x,t_n)\xi_{h,x}(x,t_n)dx \\
&\qquad - \Delta t\int_a^b V_x(x,t_n)\eta(x,t_n)\xi_h(x,t_n)dx + E_3(u,\xi_h).
\end{aligned} \tag{4.13}$$

The first through fourth terms on the right-hand side of (4.13) were already bounded in (4.4)–(4.7). We need only to bound the remaining two terms on the right-hand side. The fifth term on the right-hand side is bounded by

$$\begin{aligned}
\Big|\Delta t\int_a^b V_x(x,t_n)\eta(x,t_n)\xi_h(x,t_n)dx\Big| &\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}\ \|\eta(\cdot,t_n)\|_{L^2} \\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C\Delta t\ h^4\|u\|^2_{L^\infty(0,T;H^2)}.
\end{aligned}$$

We use the expression of $E_3(u,\xi_h)$ (below (3.18)) to bound this term by

$$\begin{aligned}
E_3(u,\xi_h) &\le C(\Delta t)^{3/2}\|\xi_h(\cdot,t_n)\|_{L^2}\Big\|\frac{d^2u}{dt^2}\Big\|_{L^2(t_{n-1},t_n;L^2)} \\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C(\Delta t)^2\Big\|\frac{d^2u}{dt^2}\Big\|^2_{L^2(t_{n-1},t_n;L^2)}.
\end{aligned}$$

We combine these estimates and the estimates (4.4)–(4.7) to get

$$\begin{aligned}
&\|\xi_h(\cdot,t_n)\|_{L^2}^2 + \varepsilon\Delta t\|\xi_{h,x}(\cdot,t_n)\|^2_{L^2_D} \\
&\quad \le \frac{1+C\Delta t}{2}(\|\xi_h(\cdot,t_n)\|_{L^2}^2 + \|\xi_h(\cdot,t_{n-1})\|_{L^2}^2) + \frac{1}{4}\varepsilon\Delta t\|\xi_{h,x}(\cdot,t_n)\|^2_{L^2_D} \\
&\qquad + C(\Delta t)^2\Big\|\frac{d^2u}{dt^2}\Big\|^2_{L^2(t_{n-1},t_n;L^2)} + C\Delta t\min\{h^2,(\Delta t)^2\}\|u\|^2_{L^\infty(0,T;H^2)} \\
&\qquad + Ch^4(\Delta t\|u\|^2_{L^\infty(0,T;H^2)} + \lambda(\|u\|^2_{H^1(t_{n-1},t_n;H^2)} + \Delta t\|u\|^2_{L^\infty(0,T;H^3)})).
\end{aligned}$$

The rest of the proof is the same as that in Theorem 4.1. ∎

The MMOCAA scheme corrects the MMOC scheme by replacing $u_h(x_h^*, t_{n-1})$ by $\check{u}_h(x_h^*, t_{n-1}) = u_h(x_h^{**}, t_{n-1})$, where x_h^{**} is an order $O((\Delta t)^2)$ perturbation to x_h^*. This does not affect the order of each term in the error analysis but introduces extra differences of the same term at x_h^* and x_h^{**} and slightly complicates the analysis.

5. Error estimates for problem (2.1) with minimal or intermediate regularity. We prove a uniform stability estimate for the ELLAM, MMOC, and MMOCAA schemes, assuming minimal regularity of problem (2.1). We then use the theory of interpolation of operators to derive a priori error estimates, which hold uniformly with respect to ε, for problem (2.1) with minimal or intermediate regularity.

5.1. A uniform stability estimate. In this subsection we prove a uniform stability estimate for the ELLAM, MMOC, and MMOCAA schemes.

THEOREM 5.1. *Assume $V \in L^\infty(0,T;W^1_\infty(a,b))$, $D \in L^\infty(0,T;L^\infty(a,b))$, $u_o \in L^2(a,b)$, and $f \in L^2(0,T;L^2(a,b))$. Let $u_h(x,0)$ in the ELLAM scheme* (3.13)*, the MMOC scheme* (3.19)*, or the MMOCAA scheme be the L^2 projection of $u_o(x)$. Then an uniform stability estimate holds:*

$$\|u_h\|_{L_\varepsilon(0,T;H^1_D)} \le C(\|u_o\|_{L^2} + \|f\|_{L^2(0,T;L^2)}). \tag{5.1}$$

Proof. We choose $w_h(x,t_n)$ in the ELLAM scheme (3.13) to be $u_h(x,t_n)$ to get

$$\begin{aligned} &\int_a^b u_h^2(x,t_n)dx + \varepsilon\Delta t\int_a^b D(x,t_n)u_{h,x}^2(x,t_n)dx \\ &\quad = \int_a^b u_h(x_h^*,t_{n-1})u_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx + \Delta t\int_a^b f(x,t_n)u_h(x,t_n)dx. \end{aligned} \tag{5.2}$$

We bound the first term on the right-hand side similarly to estimate (4.4) and incorporate the estimate into (5.2). We cancel like terms and sum the resulting inequalities for $n = 1,\dots,N_1(\le N)$ to get

$$\begin{aligned} &\|u_h(\cdot,t_{N_1})\|_{L^2}^2 + \varepsilon\sum_{n=1}^{N_1}\Delta t\|u_{h,x}(\cdot,t_n)\|_{L^2_D}^2 \\ &\quad \le C\Delta t\sum_{n=0}^{N_1}(\|u_h(\cdot,t_n)\|_{L^2}^2 + \|f(\cdot,t_n)\|_{L^2(0,T;L^2)}^2) + \|u_o\|_{L^2}^2. \end{aligned}$$

We choose $C\Delta t \le 1/2$ and apply the Gronwall inequality to finish the proof of (5.1) in the context of the ELLAM scheme.

We similarly choose $w_h(x) = u_h(x,t_n)$ in the MMOC scheme (3.19) to get

$$\begin{aligned} &\int_a^b u_h^2(x,t_n)dx + \varepsilon\Delta t\int_a^b D(x,t_n)u_{h,x}^2(x,t_n)dx \\ &\quad = \int_a^b u_h(x_h^*,t_{n-1})u_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx \\ &\qquad -\Delta t\int_a^b V_x(x,t_n)u_h^2(x,t_n)dx + \Delta t\int_a^b f(x,t_n)u_h(x,t_n)dx. \end{aligned} \tag{5.3}$$

Compared with (5.2), the only extra term is the second term on the right-hand side of (5.3) that can be bounded by

$$\left|\Delta t\int_a^b V_x(x,t_n)u_h^2(x,t_n)dx\right| \le C\Delta t\|u_h(\cdot,t_n)\|_{L^2}^2.$$

Thus, we can prove the stability estimate (5.1) for the MMOC as we did for the ELLAM. As for the MMOCAA, the only difference is the accumulation term at the time step t_{n-1}, in which the foot of the approximate characteristics is a perturbation of order $O((\Delta t)^2)$ to the Euler tracking. Therefore, the estimate (5.1) is still true. ∎

5.2. Error estimates for problems with minimal or intermediate regularity. We apply the theory of interpolation of spaces to derive a priori error estimates for the ELLAM, MMOC, and MMOCAA schemes, which hold uniformly with respect to ε, for problem (2.1) with minimal or intermediate regularity.

THEOREM 5.2. *Let u be the true solution to problem* (2.1) *and u_h be the numerical solution of the ELLAM scheme* (3.13). *Then the following error estimate holds uniformly with respect to ε for $0 < s < 1$:*

$$\begin{aligned}\|u_h - u\|_{L_\varepsilon(0,T;H^1_D)} \le{}& C(\Delta t)^s \Big(\|u_o\|_{B^s_q(L^2)} + \sqrt{\varepsilon}\|u_o\|_{B^{2s}_q(L^2)} + \Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)} \\ &+ \|f\|_{L^2(0,T;B^s_q(L^2))} \Big) + C\min\{h^s, (\Delta t)^s\}\|u_o\|_{B^{2s}_q(L^2)} \\ &+ Ch^{2s}(\|u_o\|_{B^{3s}_q(L^2)} + \lambda\|f\|_{L^2(0,T;B^{3s}_q(L^2))}).\end{aligned} \tag{5.4}$$

Proof. We split $u_h - u = (u_h^{(1)} - u^{(1)}) + (u_h^{(2)} - u^{(2)})$. Here $u_h^{(1)}$ and $u^{(1)}$ are the numerical and true solutions, respectively, to a homogeneous version of problem (2.1), whereas $u_h^{(2)}$ and $u^{(2)}$ are, respectively, the numerical and true solutions to problem (2.1) with zero initial data.

We combine the stability of the numerical solution $u_h^{(1)}$ and the true solution $u^{(1)}$, which is proved in Theorem 7.1, to obtain

$$\|u_h^{(1)} - u^{(1)}\|_{L_\varepsilon(0,T;H^1_D)} \le C\|u_o\|_{L^2}. \tag{5.5}$$

Theorem 4.1 gives an optimal-order error estimate assuming full regularity:

$$\begin{aligned}&\|u_h^{(1)} - u^{(1)}\|_{L_\varepsilon(0,T;H^1_D)} \\ &\quad \le C\Delta t\|u_o\|_{H^2_\varepsilon} + C(\min\{h, \Delta t\} + h^2)\|u_o\|_{H^2} + C\lambda h^2\|u_o\|_{H^3}.\end{aligned} \tag{5.6}$$

We use Lemma 2.1 with $m = 3$ to interpolate $L^2(a,b)$ and the Sobolev space $H^3(a,b)$:

$$[L^2(a,b), H^3(a,b)]_{s,q} = B^{3s}_q(L^2(a,b)), \quad 0 < s < 1.$$

Note that $[L^2(a,b), H^3(a,b)]_{k/3,1} \subset H^k(a,b) \subset [L^2(a,b), H^3(a,b)]_{k/3,\infty}$ for $k = 1, 2$. We apply Lemma 2.3 to reiterate the interpolation process on both ends to get

$$[L^2(a,b), H^k(a,b)]_{s,q} = [L^2(a,b), H^3(a,b)]_{ks/3,q} = B^{ks}_q(L^2(a,b)),\ 0 < s < 1,\ k = 1, 2.$$

We apply Lemma 2.2 to the estimates (5.5) and (5.6) to conclude

$$\begin{aligned}&\|u_h^{(1)} - u^{(1)}\|_{L_\varepsilon(0,T;H^1_D)} \\ &\quad \le C\lambda h^{2s}\|u_o\|_{B^{3s}_q(L^2)} + C(\min\{h^s, (\Delta t)^s\} + h^{2s})\|u_o\|_{B^{2s}_q(L^2)} \\ &\qquad + C(\Delta t)^s(\|u_o\|_{B^s_q(L^2)} + \sqrt{\varepsilon}\|u_o\|_{B^{2s}_q(L^2)}), \quad 0 < s < 1.\end{aligned} \tag{5.7}$$

We use (5.1), (7.2), and (4.1) to bound $u_h^{(2)} - u^{(2)}$ by

$$\|u_h^{(2)} - u^{(2)}\|_{L_\varepsilon(0,T;H^1_D)} \le C\|f\|_{L^2(0,T;L^2)} \tag{5.8}$$

and

$$\|u_h^{(2)} - u^{(2)}\|_{L_\varepsilon(0,T;H_D^1)} \le C\Delta t\Big(\Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)} + \|f\|_{L^2(0,T;H^1)}\Big) + C\lambda h^2\|f\|_{L^2(0,T;H^3)}. \tag{5.9}$$

We then use the interpolation to derive the following estimate:

$$\|u_h^{(2)} - u^{(2)}\|_{L_\varepsilon(0,T;H_D^1)} \le C(\Delta t)^s\Big(\Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)} + \|f\|_{L^2(0,T;B_q^s(L^2))}\Big) + C\lambda h^{2s}\|f\|_{L^2(0,T;B_q^{3s}(L^2))}. \tag{5.10}$$

We combine the estimates (5.7) and (5.10) to finish the proof of (5.4). ∎

In the theorem, the coefficients are required to be in the appropriate interpolation spaces with proper fine tuning. An analogue can be proved for the MMOC and MMOCAA schemes in a similar manner.

THEOREM 5.3. *Let u be the true solution to problem* (2.1) *and u_h be the numerical solution of the MMOC scheme* (3.19) *or the MMOCAA scheme. Then, for $0 < s < 1$, the following error estimate holds uniformly with respect to ε:*

$$\begin{aligned}
&\|u_h - u\|_{L_\varepsilon(0,T;H_D^1)}\\
&\quad\le C(\Delta t)^s\Big(\|u_o\|_{B_q^s(L^2)} + \sqrt{\varepsilon}\|u_o\|_{B_q^{2s}(L^2)} + \varepsilon^{3/2}\|u_o\|_{B_q^{3s}(L^2)}\\
&\qquad + \Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)} + \|f\|_{L^2(0,T;B_q^{2s}(L^2))}\Big) + C\min\{h^s,(\Delta t)^s\}\|u_o\|_{B_q^{2s}(L^2)}\\
&\qquad + Ch^{2s}(\|u_o\|_{B_q^{3s}(L^2)} + \lambda\|f\|_{L^2(0,T;B_q^{3s}(L^2))}).
\end{aligned} \tag{5.11}$$

6. Auxiliary lemmas. We prove two auxiliary lemmas in this section. The first lemma addresses error bounds on the approximate characteristics to the true characteristics. The second lemma proves the optimal-order error bound in (4.6).

6.1. Estimates on approximations to characteristics. We prove several bounds on the differences between the approximate and true characteristics.

LEMMA 6.1. *Let $r(t;x,t_n)$ and $r_h(t;x,t_n)$ be the true and approximate characteristics defined in* (3.3) *and* (3.12)*, respectively. Assume that V, $\frac{dV}{dt} \in L^\infty(0,T;W_\infty^1(a,b))$. Then the following estimates hold:*

$$\begin{aligned}
&|x_h^* - x^*| \le \frac{(\Delta t)^2}{2}\Big\|\frac{dV}{dt}\Big\|_{L^\infty(0,T;L^\infty)} = O((\Delta t)^2),\\
&|r_x(t;x,t_n) - 1| \le \int_t^{t_n} |V_x(r(\theta;x,t_n),\theta)|d\theta e^{\int_t^{t_n}|V_x(r(\theta;x,t_n),\theta)|d\theta} = O(t_n - t),
\end{aligned} \tag{6.1}$$

and

$$\begin{aligned}
&|r_{h,x}(t_{n-1};x,t_n) - r_x(t_{n-1};x,t_n)|\\
&\quad\le \frac{(\Delta t)^2}{2}\Big(\Big\|\frac{dV}{dt}\Big\|_{L^\infty(0,T;W_\infty^1)} + \|V\|^2_{L^\infty(0,T;W_\infty^1)}\Big)e^{\int_{t_{n-1}}^{t_n}|V_x(r(\theta;x,t_n),\theta)|d\theta}\\
&\quad= O((\Delta t)^2).
\end{aligned} \tag{6.2}$$

Proof. The definition (3.3) directly yields

$$r(t;x,t_n) = x - \int_t^{t_n} V(r(\theta;x,t_n),\theta)d\theta. \tag{6.3}$$

We subtract this equation from (3.12) to get the following, which directly leads to the first inequality in (6.1):

$$\begin{aligned}|x_h^* - x^*| &= \Big|\int_{t_{n-1}}^{t_n} V(r(t;x,t_n),t)dt - V(x,t_n)\Delta t\Big| \\ &= \Big|\int_{t_{n-1}}^{t_n} [V(r(t;x,t_n),t) - V(x,t_n)]dt\Big| \\ &\le \int_{t_{n-1}}^{t_n}\int_{\theta}^{t_n} \Big|\frac{dV(r(\theta;x,t_n),\theta)}{d\theta}\Big| d\theta dt.\end{aligned}$$

Differentiating (6.3) leads to

$$\begin{aligned}r_x(t;x,t_n) - 1 &= -\int_t^{t_n} V_x(r(\theta;x,t_n),\theta) r_x(\theta;x,t_n)d\theta \\ &= -\int_t^{t_n} V_x(r(\theta;x,t_n),\theta)d\theta \\ &\quad -\int_t^{t_n} V_x(r(\theta;x,t_n),\theta)\ (r_x(\theta;x,t_n)-1)d\theta.\end{aligned} \tag{6.4}$$

Application of the Gronwall inequality leads to the second estimate in (6.1).

We differentiate (3.12) and use (6.4) to get

$$\begin{aligned}&r_{h,x}(t_{n-1};x,t_n) - r_x(t_{n-1};x,t_n) \\ &= \int_{t_{n-1}}^{t_n} [V_x(r(t;x,t_n),t) - V_x(x,t_n)]dt \\ &\quad -\int_{t_{n-1}}^{t_n} V_x(r(t;x,t_n),t)[r_{h,x}(t;x,t_n) - r_x(t;x,t_n) + V_x(x,t_n)(t_n - t)]dt.\end{aligned}$$

Application of the Gronwall inequality to the following proves (6.2):

$$\begin{aligned}&|r_{h,x}(t_{n-1};x,t_n) - r_x(t_{n-1};x,t_n)| \\ &\quad \le \frac{(\Delta t)^2}{2}\Big(\Big\|\frac{dV}{dt}\Big\|_{L^\infty(0,T;W^1_\infty)} + \|V\|^2_{L^\infty(0,T;W^1_\infty)}\Big) \\ &\qquad + \int_{t_{n-1}}^{t_n} |V_x(r(t;x,t_n),t)|\ |r_{h,x}(t;x,t_n) - r_x(t;x,t_n)|dt. \qquad \square\end{aligned}$$

6.2. A superconvergent estimate on interpolation. We prove the following superconvergence estimate on the interpolation error.

LEMMA 6.2. *Assume $u \in L^\infty(0,T;H^3(a,b)) \cap H^1(0,T;H^2(a,b))$. Let $\Pi_h u \in S_h(a,b)$ be the interpolation of u and $\eta = \Pi_h u - u$. Let λ be the parameter defined in Theorem* 4.1. *Then the following superconvergence estimate holds:*

$$\begin{aligned}&\Big|\int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n) r_{h,x}(t_{n-1};x,t_n)dx - \int_a^b \eta(x,t_n)\xi_h(x,t_n)dx\Big| \\ &\quad \le C\Delta t\|\xi_h(\cdot,t_n)\|^2_{L^2} + C\Delta t(\min\{h^2,(\Delta t)^2\} + h^4)\|u\|^2_{L^\infty(0,T;H^2)} \\ &\qquad + C(\Delta t)^3\|u\|^2_{L^\infty(0,T;H^1)} + C\lambda\, h^4(\|u\|^2_{H^1(t_{n-1},t_n;H^2)} + \Delta t\|u\|^2_{L^\infty(0,T;H^3)}).\end{aligned} \tag{6.5}$$

Proof. We rewrite the left-hand side of (6.5) as

$$\begin{aligned}&\int_a^b \eta(x,t_n)\xi_h(x,t_n)dx - \int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n)r_{h,x}(t_{n-1};x,t_n)dx\\ &\quad=\int_a^b (\eta(x,t_n)-\eta(x_h^*,t_{n-1}))\xi_h(x,t_n)dx\\ &\qquad+\Delta t\int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n)V_x(x,t_n)dx.\end{aligned}\tag{6.6}$$

We bound the second term on the right-hand side in a similar way to (4.4):

$$\begin{aligned}&\Big|\Delta t\int_a^b \eta(x_h^*,t_{n-1})\xi_h(x,t_n)V_x(x,t_n)dx\Big|\\ &\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}\ \|\eta(x_h^*,t_{n-1})\|_{L^2(a,b)}\\ &\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}\ \|\eta(\cdot,t_{n-1})\|_{L^2}\\ &\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C\Delta t\ h^4\|u\|^2_{L^\infty(0,T;H^2)}.\end{aligned}\tag{6.7}$$

When $\mathrm{Cr}\ge 1$ that implies $h\le C\Delta t$, we bound the first term in (6.6) by

$$\begin{aligned}&\Big|\int_a^b (\eta(x,t_n)-\eta(x_h^*,t_{n-1}))\xi_h(x,t_n)dx\Big|\\ &\quad\le C\|\xi_h(\cdot,t_n)\|_{L^2}(\|\eta(\cdot,t_n)\|_{L^2}+\|\eta(\cdot,t_{n-1})\|_{L^2})\\ &\quad\le Ch^2\|\xi_h(\cdot,t_n)\|_{L^2}\ \|u\|_{L^\infty(0,T;H^2)}\\ &\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + C\Delta t\min\{h^2,(\Delta t)^2\}\|u\|^2_{L^\infty(0,T;H^2)}.\end{aligned}\tag{6.8}$$

For $\mathrm{Cr}<1$, we decompose this term as follows:

$$\begin{aligned}&\int_a^b (\eta(x,t_n)-\eta(x_h^*,t_{n-1}))\xi_h(x,t_n)dx\\ &\quad=\int_a^b\int_{t_{n-1}}^{t_n}\eta_t(x,t)dt\ \xi_h(x,t_n)dx\\ &\qquad+\int_a^b (\eta(x,t_{n-1})-\eta(x_h^*,t_{n-1}))\xi_h(x,t_n)dx.\end{aligned}\tag{6.9}$$

The first term on the right-hand side is bounded by

$$\begin{aligned}&\Big|\int_a^b\int_{t_{n-1}}^{t_n}\eta_t(x,t)dt\ \xi_h(x,t_n)dx\Big|\\ &\quad\le(\Delta t)^{1/2}\|\xi_h(\cdot,t_n)\|_{L^2}\ \|\eta\|_{H^1(t_{n-1},t_n;L^2)}\\ &\quad\le\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2 + Ch^4\|u\|^2_{H^1(t_{n-1},t_n;H^2)}.\end{aligned}\tag{6.10}$$

We use the following expressions in the second term on the right side of (6.9):

$$\begin{aligned}\eta(x,t_{n-1})-\eta(x_h^*,t_{n-1}) &= \int_0^1 \frac{d}{d\theta}\eta(x_h^*+\theta(x-x_h^*),t_{n-1})d\theta\\ &= \int_0^1 \eta_x(x_h^*+\theta(x-x_h^*),t_{n-1})(x-x_h^*)d\theta,\\ \frac{\partial}{\partial x}\Big(\eta(x_h^*+\theta(x-x_h^*),t_{n-1})\Big) &= \eta_x(x_h^*+\theta(x-x_h^*),t_{n-1})(x_{h,x}^*+\theta(1-x_{h,x}^*))\\ &= \eta_x(x_h^*+\theta(x-x_h^*),t_{n-1})(1-(1-\theta)V_x(x,t_n)\Delta t)\end{aligned}$$

and then integrate the resulting term by parts to yield

$$
\begin{aligned}
&\int_a^b (\eta(x,t_{n-1})-\eta(x_h^*,t_{n-1}))\xi_h(x,t_n)dx\\
&\quad=\int_a^b\int_0^1 \eta_x(x_h^*+\theta(x-x_h^*),t_{n-1})(x-x_h^*)d\theta\ \xi_h(x,t_n)dx\\
&\quad=\int_a^b\int_0^1 \frac{\partial}{\partial x}\Big(\eta(x_h^*+\theta(x-x_h^*),t_{n-1})\Big)\\
&\qquad\qquad (1-(1-\theta)V_x(x,t_n)\Delta t)^{-1}(x-x_h^*)d\theta\ \xi_h(x,t_n)dx\\
&\quad=\int_a^b\int_0^1 \frac{\partial}{\partial x}\Big(\eta(x_h^*+\theta(x-x_h^*),t_{n-1})\Big)(x-x_h^*)d\theta\ \xi_h(x,t_n)dx\\
&\qquad+\int_a^b\int_0^1 \frac{\partial}{\partial x}\Big(\eta(x_h^*+\theta(x-x_h^*),t_{n-1})\Big)O((\Delta t)^2)d\theta\ \xi_h(x,t_n)dx\\
&\quad=-\int_0^1\int_a^b \eta(x_h^*+\theta(x-x_h^*),t_{n-1})(x-x_h^*)_x\xi_h(x,t_n)dxd\theta\\
&\qquad-\int_0^1\int_a^b \eta(x_h^*+\theta(x-x_h^*),t_{n-1})(x-x_h^*)\xi_{h,x}(x,t_n)dxd\theta\\
&\qquad+\int_a^b\int_0^1 \frac{\partial}{\partial x}\Big(\eta(x_h^*+\theta(x-x_h^*),t_{n-1})\Big)O((\Delta t)^2)d\theta\ \xi_h(x,t_n)dx.
\end{aligned}
\tag{6.11}
$$

Let $y=x_h^*+\theta(x-x_h^*)$. We bound the first and third terms on the right side by

$$
\begin{aligned}
&\Big|\int_0^1\int_a^b \eta(x_h^*+\theta(x-x_h^*),t_{n-1})(x-x_h^*)_x\xi_h(x,t_n)dxd\theta\\
&\qquad+\int_a^b\int_0^1 \frac{\partial}{\partial x}\Big(\eta(x_h^*+\theta(x-x_h^*),t_{n-1})\Big)O((\Delta t)^2)d\theta\ \xi_h(x,t_n)dx\Big|\\
&\le \Delta t\int_0^1\int_a^b |V_x(x,t_n)||\eta(x_h^*+\theta(x-x_h^*),t_{n-1})||\xi_h(x,t_n)|dxd\theta\\
&\qquad+C(\Delta t)^2\int_0^1\int_a^b |\eta_x(x_h^*+\theta(x-x_h^*),t_{n-1})||\xi_h(x,t_n)|dxd\theta\\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}\Big[\Big(\int_0^1\int_a^b \eta^2(y,t_{n-1})(\theta+(1-\theta)x_{h,x}^*)^{-1}dyd\theta\Big)^{1/2}\\
&\qquad+\Delta t\Big(\int_0^1\int_a^b \eta_x^2(y,t_{n-1})(\theta+(1-\theta)x_{h,x}^*)^{-1}dyd\theta\Big)^{1/2}\Big]\\
&\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2+C\Delta t h^4\|u\|_{L^\infty(0,T;H^2)}^2+C(\Delta t)^3\|u\|_{L^\infty(0,T;H^1)}^2.
\end{aligned}
\tag{6.12}
$$

We decompose the second term on the right-hand side of (6.11) as

$$
\begin{aligned}
&\int_0^1\int_a^b \eta(x_h^*+\theta(x-x_h^*),t_{n-1})(x-x_h^*)\xi_{h,x}(x,t_n)dxd\theta\\
&\quad=\Delta t\int_a^b V(x,t_n)\xi_{h,x}(x,t_n)\Big(\eta(x,t_{n-1})\\
&\qquad+\int_0^1\int_0^1 \frac{d}{d\gamma}\eta(x+\gamma(1-\theta)(x_h^*-x),t_{n-1})d\gamma d\theta\Big)dx\\
&\quad=\Delta t\int_a^b V(x,t_n)\xi_{h,x}(x,t_n)\eta(x,t_{n-1})dx\\
&\qquad+\Delta t\int_0^1\int_0^1\int_a^b V(x,t_n)\xi_{h,x}(x,t_n)(1-\theta)(x_h^*-x)\\
&\qquad\quad\times\eta_x(x+\gamma(1-\theta)(x_h^*-x),t_{n-1})dxd\gamma d\theta.
\end{aligned}
\tag{6.13}
$$

We use the inverse inequality (2.3) to bound the second term by

$$\begin{aligned}(6.14)\qquad &\Big|\Delta t\int_0^1\int_0^1\int_a^b V(x,t_n)\xi_{h,x}(x,t_n)(1-\theta)(x_h^*-x)\\ &\qquad\times\eta_x(x+\gamma(1-\theta)(x_h^*-x),t_{n-1})dxd\gamma d\theta\Big|\\ &\le C(\Delta t)^2h\|\xi_{h,x}(\cdot,t_n)\|_{L^2}\|u(\cdot,t_{n-1})\|_{H^2}\\ &\le C\Delta t\ \|\xi_h(\cdot,t_n)\|_{L^2}^2+C(\Delta t)^3\ \|u\|_{L^\infty(0,T;H^2)}^2.\end{aligned}$$

A standard estimate of the first term on the right-hand side of (6.13) yields

$$\begin{aligned}\Big|\Delta t\int_a^b V(x,t_n)\eta(x,t_{n-1})\xi_{h,x}(x,t_n)dx\Big| &\le C\Delta t\|\xi_{h,x}(\cdot,t_n)\|_{L^2}\|\eta(\cdot,t_{n-1})\|_{L^2}\\ &\le C\Delta t\,h^2\|\xi_{h,x}(\cdot,t_n)\|_{L^2}\|u(\cdot,t_{n-1})\|_{H^2}\\ &\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2+C\Delta t\,h^2\|u(\cdot,t_{n-1})\|_{H^2}.\end{aligned}$$

This will result in a suboptimal-order estimate of order $O(h+\Delta t)$ for the ELLAM scheme. To derive an optimal-order estimate, we sum this term by parts to obtain

$$\begin{aligned}(6.15)\qquad &\Delta t\int_a^b V(x,t_n)\eta(x,t_{n-1})\xi_{h,x}(x,t_n)dx\\ &=-\frac{\Delta t}{h}\sum_{i=1}^I\int_{x_{i-1}}^{x_i}(V(x+h,t_n)-V(x,t_n))\eta(x,t_{n-1})\xi_h(x_i,t_n)dx\\ &\quad-\frac{\Delta t}{h}\sum_{i=1}^I\int_{x_{i-1}}^{x_i}(\eta(x+h,t_{n-1})-\eta(x,t_{n-1}))V(x+h,t_n)\xi_h(x_i,t_n)dx.\end{aligned}$$

We bound the first term on the right-hand side of (6.15) by

$$\begin{aligned}(6.16)\qquad &\Big|\frac{\Delta t}{h}\sum_{i=1}^I\int_{x_{i-1}}^{x_i}(V(x+h,t_n)-V(x,t_n))\eta(x,t_{n-1})\xi_h(x_i,t_n)dx\Big|\\ &\le C\Delta t\ \|\xi_h(\cdot,t_n)\|_{L^2}\ \|\eta(\cdot,t_{n-1})\|_{L^2}\\ &\le C\Delta t\ \|\xi_h(\cdot,t_n)\|_{L^2}^2+C\Delta t\ h^4\ \|u\|_{L^\infty(0,T;H^2)}^2,\end{aligned}$$

where we have used the equivalence between the discrete and continuous L^2 norms.

However, if we similarly bound the second term on the right side of (6.15), we can obtain only a suboptimal-order estimate. To derive an optimal-order estimate, we introduce an auxiliary function $\psi(x,t)$ by

$$\psi(x,t)=u(x+h,t)-u(x,t)=\int_0^h u_\alpha(\alpha+x,t)d\alpha.$$

Because $\eta(x+h,t_{n-1})$ is a shift of $\eta(x,t_{n-1})$ by one grid point, so the forward difference operator and the shift operator are commutative:

$$\begin{aligned}\eta(x+h,t_{n-1})-\eta(x,t_{n-1}) &=(\Pi_h-\mathrm{I})u(x+h,t_{n-1})-(\Pi_h-\mathrm{I})u(x,t_{n-1})\\ &=(\Pi_h-\mathrm{I})(u(x+h,t_{n-1})-u(x,t_{n-1}))\\ &=(\Pi_h-\mathrm{I})\psi(x,t_{n-1}).\end{aligned}$$

Inserting this identity into the second term on the right-hand side of (6.15) gives

$$\begin{aligned}&\Big|\frac{\Delta t}{h}\sum_{i=1}^{I}\int_{x_{i-1}}^{x_i}(\eta(x+h,t_n)-\eta(x,t_n))V(x+h,t_n)\xi_h(x_i,t_n)dx\Big|\\&\quad\le\frac{C\Delta t}{h}\|\xi_h(\cdot,t_n)\|_{L^2}\ \|(\Pi_h-\mathrm{I})\psi(\cdot,t_{n-1})\|_{L^2}\le C\Delta t\ h\|\xi_h(\cdot,t_n)\|_{L^2}\|\psi(\cdot,t_{n-1})\|_{H^2}\\&\quad\le C\Delta t\|\xi_h(\cdot,t_n)\|_{L^2}^2+C\Delta t\ h^4\|u\|^2_{L^\infty(0,T;H^3)}.\end{aligned}\tag{6.17}$$

Combining all of these estimates, we have proved (6.5). ∎

7. Uniform stability estimates of the exact solutions. The existence, uniqueness, and stability estimates for problem (2.1) can be found, e.g., in [9]. We derive a priori stability estimates for problem (2.1) in ε-weighted norms, which hold uniformly with respect to ε, under different regularity assumptions.

7.1. A generic stability estimate with minimal regularity assumption. We prove a priori stability estimates for a slightly more general initial-boundary value problem than problem (2.1) with minimal regularity assumption:

$$\begin{aligned}&z_t+(\bar V(x,t)z-\varepsilon\bar D(x,t)z_x)_x+\bar R(x,t)z=\bar f(x,t),\quad (x,t)\in(a,b)\times(0,T],\\&\qquad\qquad z(x,0)=z_o(x),\end{aligned}\tag{7.1}$$

which is closed by a periodic boundary condition at $x=a$ and $x=b$.

THEOREM 7.1. *Assume $\bar D\in L^\infty(0,T;L^\infty)$, $\bar R\in L^1(0,T;L^\infty)$, $\bar V\in L^1(0,T;W^1_\infty)$, $z_o\in L^2(a,b)$, and $\bar f\in L^2(0,T;L^2)$. Then the following estimate holds:*

$$\begin{aligned}\|z\|_{L_\varepsilon(0,T;H^1_D)}&\le 2(\|z_o\|_{L^2}+\|\bar f\|_{L^2(0,T;L^2)})e^{\frac12\|(1-2\bar R-\bar V_x)^+\|_{L^1(0,T;L^\infty)}}\\&\le C(\|z_o\|_{L^2}+\|\bar f\|_{L^2(0,T;L^2)}).\end{aligned}\tag{7.2}$$

Here C depends on $\|(1-2\bar R-\bar V_x)^+\|_{L^1(0,T;L^\infty)}$ but not on ε, $g^+(x,t)=\max\{g(x,t),0\}$.

Proof. We combine $z_t+\bar V z_x$ in (7.1) to form a material derivative $\frac{dz}{dt}$ along the characteristics $x=r(t;\bar x,\bar t)$ defined by (3.3) with V being replaced by $\bar V$. We multiply the equation by $z(x,t)$ and integrate the resulting equation to get

$$\int_a^b\frac{dz}{dt}zdx+\varepsilon\int_a^b\bar D(x,t)z_x^2dx+\int_a^b(\bar R(x,t)+\bar V_x(x,t))z^2dx=\int_a^b\bar f(x,t)zdx.$$

We apply the Reynolds transport theorem

$$\frac{d}{dt}\int_{r(t;a,\bar t)}^{r(t;b,\bar t)}gdx=\int_{r(t;a,\bar t)}^{r(t;b,\bar t)}\Big(\frac{dg}{dt}+\bar V_xg\Big)dx$$

to $g=z^2(x,t)$ at $t=\bar t$ to rewrite the weak formulation as

$$\frac12\frac{d}{dt}\int_a^bz^2dx+\varepsilon\int_a^b\bar Dz_x^2dx+\int_a^b\Big(\bar R+\frac12\bar V_x\Big)z^2dx=\int_a^bfzdx.$$

Integrating this equation from $t = 0$ to $t = \bar{t}$ yields

$$\begin{aligned}
&\int_a^b z^2(x,\bar{t})dx + 2\varepsilon \int_0^{\bar{t}} \int_a^b \bar{D}(x,t) z_x^2(x,t) dx dt \\
&\quad = \int_a^b z^2(x,\bar{t})dx + 2\varepsilon \int_0^{\bar{t}} \int_{r(t;a,\bar{t})}^{r(t;b,\bar{t})} \bar{D}(x,t) z_x^2(x,t) dx dt \\
&\quad \le \int_{r(0;a,\bar{t})}^{r(0;b,\bar{t})} z_o^2(x) dx + \int_0^{\bar{t}} \int_{r(t;a,\bar{t})}^{r(t;b,\bar{t})} \bar{f}^2(x,t) dx dt \\
&\qquad + \int_0^{\bar{t}} \|(1 - 2\bar{R} - \bar{V}_x)^+(\cdot,t)\|_{L^\infty(a,b)} \int_{r(t;a,\bar{t})}^{r(t;b,\bar{t})} z^2(x,t) dx dt \\
&\quad = \|z_o\|_{L^2}^2 + \int_0^{\bar{t}} \|\bar{f}(\cdot,t)\|_{L^2}^2 dt + \int_0^{\bar{t}} \|(1 - 2\bar{R} - \bar{V}_x)^+(\cdot,t)\|_{L^\infty(a,b)} \, \|z(\cdot,t)\|_{L^2}^2 dt.
\end{aligned}$$

Here we have used the fact that $\int_{r(t;a,\bar{t})}^{r(t;b,\bar{t})} g(x)dx = \int_a^b g(x)dx$. Applying the Gronwall inequality finishes the proof of (7.2). ∎

7.2. Uniform stability estimates for problem (2.1). In this subsection we prove a priori stability estimates for problem (2.1) in different Sobolev norms.

THEOREM 7.2. *Assume $D, V \in L^\infty(0,T; W_\infty^{k+1}(a,b))$, $u_o \in H^k(a,b)$, and $f \in L^2(0,T; H^k(a,b))$. Then the following stability estimate holds for problem* (2.1)*:*

$$\|u\|_{L_\varepsilon(0,T;H^{k+1})} \le C(\|u_o\|_{H^k} + \|f\|_{L^2(0,T;H^k)}), \tag{7.3}$$

where $C = C(\|D\|_{L^\infty(0,T;W_\infty^{k+1})}, \|V\|_{L^\infty(0,T;W_\infty^{k+1})})$, but not on ε. Further, if $D, V \in L^\infty(0,T; W_\infty^4(a,b))$, $u_o \in H^3(a,b)$, and $f \in L^2(0,T; H^3(a,b))$, we have

$$\|u\|_{H^1(0,T;H^2)} \le C(\|u_o\|_{H^3} + \|f\|_{L^2(0,T;H^3)}) \tag{7.4}$$

and

$$\left\|\frac{d^2u}{dt^2}\right\|_{L^2(0,T;L^2)} \le C\Big(\varepsilon\|u_o\|_{H_\varepsilon^3} + \|u_o\|_{H_\varepsilon^1} + \|f\|_{L^2(0,T;H^2)} + \left\|\frac{df}{dt}\right\|_{L^2(0,T;L^2)}\Big). \tag{7.5}$$

Here $C = C(\|D\|_{L^\infty(0,T;W_\infty^4)}, \|V\|_{L^\infty(0,T;W_\infty^4)})$, but not on ε. Finally, assume $D, V \in L^\infty(0,T; W_\infty^{l+1}(a,b))$, $u_o \in H^l(a,b)$, and $f \in L^2(0,T; H^l(a,b))$ for $l = 0, 1$. We have

$$\left\|\frac{du}{dt}\right\|_{L^2(0,T;H^l)} \le C(\|u_o\|_{H_\varepsilon^{l+1}} + \|f\|_{L^2(0,T;H^l)}). \tag{7.6}$$

Proof. (7.3) for $k = 0$ is a direct consequence of (7.2). To prove (7.3) for $k = 1$, we differentiate problem (2.1) with respect to x. The resulting equation corresponds to (7.1) with $z = u_x$, $\bar{D} = D$, $\bar{V} = V - \varepsilon D_x$, $\bar{R} = V_x$, and $\bar{f} = f_x - V_{xx}$. The estimate (7.2) yields (7.3) with $k = 1$. We prove (7.3) for $k \ge 2$ by induction.

To prove the estimate (7.4) we use the governing equation in (2.1) to express u_t in terms of spatial derivatives and bound these spatial derivatives by the estimate (7.3). We similarly bound $\frac{du}{dt}$ in terms of spatial derivatives as follows:

$$\left\|\frac{du}{dt}\right\|_{L^2(0,T;H^l)} \le C(\varepsilon\|(Du_x)_x\|_{L^2(0,T;H^l)} + \|V_x u\|_{L^2(0,T;H^l)} + \|f\|_{L^2(0,T;H^l)}).$$

This together with (7.3) proves the estimate of $\frac{du}{dt}$. We similarly express $\frac{d^2u}{dt^2}$ as

$$\begin{aligned}
&\Big\|\frac{d^2u}{dt^2}\Big\|_{L^2(0,T;L^2)}\\
&\quad=\Big\|\varepsilon[D(\varepsilon(Du_x)_x-V_xu+f)_x]_x+\varepsilon\Big(\frac{dD}{dt}u_x\Big)_x-V_x\frac{du}{dt}-\frac{dV_x}{dt}u+\frac{df}{dt}\Big\|_{L^2(0,T;L^2)}\\
&\quad\le\varepsilon^2\|u\|_{L^2(0,T;H^4)}+\varepsilon\|u\|_{L^2(0,T;H^2)}+\|u\|_{L^2(0,T;L^2)}+\|f\|_{L^2(0,T;H^2)}\\
&\qquad+\Big\|\frac{du}{dt}\Big\|_{L^2(0,T;L^2)}+\Big\|\frac{df}{dt}\Big\|_{L^2(0,T;L^2)}.
\end{aligned}$$

We combine this inequality with estimates (7.3) and (7.6) to finish the proof. ∎

8. Concluding remarks. In this section we summarize the main results in this paper and address several related issues. We also carry out numerical example runs to verify the theoretical estimates numerically. We conclude the paper by briefly discussing the directions of future work.

8.1. The ε-weighted energy norm and L^∞ norm. In the context of stationary convection-diffusion equations, the location of internal and boundary layers is known a priori. A piecewise-uniform mesh was proposed and analyzed by Shishkin to resolve the boundary and internal layers. Moreover, an ε-uniform L^∞ error estimate was proved for numerical methods with Shishkin mesh [11, 14]. However, in the context of transient convection-diffusion equations, the fronts are dynamic and do not always coincide with the spatial mesh. Thus, although an ε-uniform error estimate in the L^∞ norm is ideal, it is generally impossible especially in the context of multiple space dimensions and in the limiting case of $\varepsilon=0$. This is why the L^∞ norm is not used in the numerical methods for hyperbolic conservation laws [13].

In this paper we derived ε-uniform error estimates in the ε-weighted energy norm $\|\cdot\|_{H^1_\varepsilon}$. We now discuss the relation between the error estimates measured in $\|\cdot\|_{L^\infty}$ and in $\|\cdot\|_{H^1_\varepsilon}$. In the context of an exponential layer (say, located at $x=1$), the approximation error e is expected to be of the form [11, 14]

$$e=\exp\Big(\frac{-(1-x)}{\varepsilon}\Big),\qquad 0\le x\le 1. \tag{8.1}$$

$\|e\|_{L^\infty}=O(1)$ and $\|e\|_{H^1_\varepsilon(0,1)}=\|e\|_{L^2(0,1)}+\sqrt{\varepsilon}\|e_x\|_{L^2(0,1)}=\sqrt{\varepsilon}+\sqrt{\varepsilon}\cdot(1/\sqrt{\varepsilon})=O(1)$. Thus, $\|e\|_{L^\infty}$ is comparable to $\|e\|_{H^1_\varepsilon(0,1)}$, and both norms recognize the exponential layer. When problem (2.1) has a smooth solution, $\|e\|_{L^\infty}=O(h^2)$ and $\|e\|_{H^1_\varepsilon(0,1)}=\|e\|_{L^2(0,1)}+\sqrt{\varepsilon}\|e_x\|_{L^2(0,1)}=O(h^2+\sqrt{\varepsilon}h)=O(h^2)$ for $\varepsilon<h^2$. Thus, $\|e\|_{L^\infty}$ is still comparable to $\|e\|_{H^1_\varepsilon(0,1)}$.

In the context of a parabolic layer (say, located at $x=1$), the approximation error e is expected to be of the form [11, 14]

$$e=\exp\Big(\frac{-(1-x)}{\sqrt{\varepsilon}}\Big),\qquad 0\le x\le 1. \tag{8.2}$$

For a parabolic layer, $\|e\|_{L^\infty}=O(1)$ and $\|e\|_{H^1_\varepsilon(0,1)}=\|e\|_{L^2(0,1)}+\sqrt{\varepsilon}\|e_x\|_{L^2(0,1)}=\varepsilon^{1/4}+\sqrt{\varepsilon}\cdot\varepsilon^{-1/4}=O(\varepsilon^{1/4})$. In this case, the L^∞ norm still recognizes the parabolic layer, but the ε-weighted norm does not recognize the layer.

In summary, the ε-weighted norm is comparable to the L^∞ norm in the cases of a smooth solution or an exponential layer. An exponential layer exhibits the strongest layer behavior and is of the major concern from a numerical and analysis viewpoint.

On the other hand, in the context of a parabolic layer, $\|\cdot\|_{H^1_\varepsilon}$ is somewhat weaker than $\|\cdot\|_{L^\infty}$. Thus, the ε-weighted norm is the most feasible measure for transient convection-diffusion equations and is closely related to the L^∞ norm. The L^∞ norm is an ideal but impossible measure in this context.

8.2. Measurements in Besov spaces. In Theorems 5.2 and 5.3, we used the interpolation theory and stability estimates to derive an ε-uniform estimate for problem (2.1) with minimal or intermediate regularity, where the initial and right-hand side data were measured in Besov spaces that generate a refined scale of smoothness and convergence rate. As an example, we consider an initial configuration that contains interior layers. Note that the indicator function $\chi_{(0.4,0.8)}$ (introduced before estimate (4.11)) satisfies $\chi_{(0.4,0.8)} \in H^{1/2-\delta}(0,1)$ for any $0 < \delta \le 1/2$ but $\notin H^{1/2}(0,1)$. Direct calculation shows that

$$\|\chi_{(0.4,0.8)}\|_{H^{1/2-\delta}(0,1)} = O(\delta^{-1/2}) \to \infty \quad \text{as } \delta \to 0. \tag{8.3}$$

Consequently, the convergence rate in the estimates (5.4) and (5.11) will be $s = 1/4 - \delta/2$ in the case of $\mathrm{Cr} \ge 1$, and the norm will blow up as $\delta \to 0$, if the initial configuration is measured in the Sobolev spaces. On the other hand, it can be verified that $\chi_{(0.4,0.8)} \in B^{1/2}_\infty(L^2(0,1))$ but $\notin B^{1/2}_q(L^2(0,1))$ for $q < \infty$. With this measure of the same initial data, the estimates (5.4) and (5.11) yield a sharp convergence rate of order $s = 1/4$ in the case of $\mathrm{Cr} \ge 1$.

8.3. Numerical experiments. Various numerical experiments were reported in the literature, which confirmed spatial and temporal convergence rates of Eulerian–Lagrangian methods (see, e.g., [7, 15, 18]). In this section we conduct numerical experiments to observe the convergence behavior of the truncation error $u_h - u$ and its dependence on ε. We simulate the transport of a Gaussian pulse subject to (2.1) with the initial configuration being given by

$$u_o(x) = \exp\Big(-\frac{(x - x_c)^2}{2\sigma^2}\Big), \tag{8.4}$$

where x_c and σ are the centered and standard deviations, respectively, of the Gaussian pulse. The analytical solution $u(x,t)$ for a homogeneous equation (2.1) is given by

$$u(x,t) = \frac{\sqrt{2}\sigma}{\sqrt{2\sigma^2 + 4\varepsilon t}} \exp\Big(-\frac{(x - x_c - Vt)^2}{2\sigma^2 + 4\varepsilon t}\Big). \tag{8.5}$$

In the numerical example runs, the spatial domain is $(a,b) = (0,3)$, and the time interval is $(0,T) = (0,1)$. We select $V = 1$ and $D = 1$ so the convection dominance is controlled by the magnitude of ε, which is chosen to be 0.001, 0.0001, and 0. We also choose $x_c = 0.5$, and $\sigma = 0.1$. We fix a small time step Δt and use a linear regression to fit the convergence rates and the associated constants in the weighted energy norm

$$\|u_h - u\| \le C_\alpha h^\alpha. \tag{8.6}$$

The results are presented in Table 8.1, which shows that the ELLAM scheme maintains second-order accuracy in space. Moreover, these convergence rates hold uniformly as ε tends to zero, even in the limiting case of $\varepsilon = 0$, as predicted by the theorems proved in this paper.

TABLE 8.1
Spatial convergence rates in the ε-weighted energy norm with $\Delta t = 1/100$.

h	$\varepsilon = 0.001$	$\varepsilon = 0.0001$	$\varepsilon = 0$
1/10	5.74×10^{-2}	7.52×10^{-2}	7.76×10^{-2}
1/20	7.99×10^{-3}	1.02×10^{-2}	1.05×10^{-2}
1/30	2.68×10^{-3}	3.10×10^{-3}	3.15×10^{-3}
1/40	1.24×10^{-3}	1.36×10^{-3}	1.37×10^{-3}
1/60	4.01×10^{-4}	4.31×10^{-4}	4.32×10^{-4}
1/80	1.86×10^{-4}	2.12×10^{-4}	2.15×10^{-4}
	$C_\alpha = 32$, $\alpha = 2.75$	$C_\alpha = 51$, $\alpha = 2.84$	$C_\alpha = 54$, $\alpha = 2.86$

8.4. Summary and discussions. In this paper we proved ε-uniform error estimates in the ε-weighted energy norm for the ELLAM, MMOC, and MMOCAA schemes for one-dimensional singularly perturbed, time-dependent convection-diffusion equations with periodic boundary conditions. The estimates were derived on a uniform space-time partition with no upstream weighting or local grid refinement or any other special arrangements of the grid, so these estimates justify the strength of Eulerian–Lagrangian methods. The analysis fully utilizes the simplicity of the one space dimension and the periodic boundary conditions.

However, a multidimensional analogue of problem (2.1) with general boundary conditions presents much more severe challenges, due to the complication of multiple space dimensions, the solution structures, and the appearance of boundary and interior layers. These issues will be investigated in the near future.

Acknowledgments. This work was done during the visit of the first author to Shandong University. The authors express their sincere thanks to the referees for their very helpful comments and suggestions, which greatly improved the quality of this paper.

REFERENCES

[1] T. ARBOGAST AND M.F. WHEELER, *A characteristics-mixed finite element method for advection-dominated transport problems*, SIAM J. Numer. Anal., 32 (1995), pp. 404–424.
[2] C. BENNETT AND R.C. SHARPLEY, *Interpolation of Operators*, Academic, New York, 1988.
[3] M.A. CELIA, T.F. RUSSELL, I. HERRERA, AND R.E. EWING, *An Eulerian–Lagrangian localized adjoint method for the advection-diffusion equation*, Adv. Water Resources, 13 (1990), pp. 187–206.
[4] P.G. CIARLET, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[5] R.A. DEVORE AND G.G. LORENTZ, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.
[6] J. DOUGLAS, JR., F. FURTADO, AND F. PEREIRA, *On the numerical simulation of waterflooding of heterogeneous petroleum reservoirs*, Comput. Geosci., 1 (1997), pp. 155–190.
[7] J. DOUGLAS, JR., C.-S. HUANG, AND F. PEREIRA, *The modified method of characteristics with adjusted advection*, Numer. Math., 83 (1999), pp. 353–369.
[8] J. DOUGLAS, JR., AND T.F. RUSSELL, *Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures*, SIAM J. Numer. Anal., 19 (1982), pp. 871–885.
[9] L.C. EVANS, *Partial Differential Equations*, Grad. Stud. Math. 19, American Mathematical Society, Providence, RI, 1998.
[10] R.E. EWING, ED., *The Mathematics of Reservoir Simulation*, Frontiers Appl. Math. 1, SIAM, Philadelphia, 1984.
[11] P.A. FARRELL, A.F. HEGARTY, J.J.H. MILLER, E. O'RIORDAN, AND G.I. SHISHKIN, *Robust Computational Techniques for Boundary Layers*, Appl. Math. 16, Chapman and Hall/CRC, Boca Raton, FL, 2000.

[12] R. Helmig, *Multiphase Flow and Transport Processes in the Subsurface*, Springer-Verlag, Berlin, 1997.

[13] R.J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge Texts Appl. Math., Cambridge University Press, Cambridge, 2002.

[14] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.

[15] H. Wang, *An optimal-order error estimate for an ELLAM scheme for two-dimensional linear advection-diffusion equations*, SIAM J. Numer. Anal., 37 (2000), pp. 1338–1368.

[16] H. Wang, H.K. Dahle, R.E. Ewing, M.S. Espedal, R.C. Sharpley, and S. Man, *An ELLAM Scheme for advection-diffusion equations in two dimensions*, SIAM J. Sci. Comput., 20 (1999), pp. 2160–2194.

[17] H. Wang, R.E. Ewing, G. Qin, S.L. Lyons, M. Al-Lawatia, and S. Man, *A family of Eulerian-Lagrangian localized adjoint methods for multi-dimensional advection-reaction equations*, J. Comput. Phys., 152 (1999), pp. 120–163.

[18] H. Wang, R.E. Ewing, and T.F. Russell, *Eulerian-Lagrangian localized methods for convection-diffusion equations and their convergence analysis*, IMA J. Numer. Anal., 15 (1995), pp. 405–459.

[19] H. Wang, D. Liang, R.E. Ewing, S.L. Lyons, and G. Qin, *An ELLAM-MFEM solution technique for compressible fluid flows in porous media with point sources and sinks*, J. Comput. Phys., 159 (2000), pp. 344–376.

SIAM J. NUMER. ANAL.
Vol. 45, No. 3, pp. 1330–1347

UNIFIED CONVERGENCE RESULTS ON A MINIMAX ALGORITHM FOR FINDING MULTIPLE CRITICAL POINTS IN BANACH SPACES*

XUDONG YAO† AND JIANXIN ZHOU‡

Abstract. A minimax method for finding multiple critical points in Banach spaces is successfully developed in [X. Yao and J. Zhou, *SIAM J. Sci. Comput.*, 26 (2005), pp. 1796–1809] by using a projected pseudogradient as a search direction. Since several different techniques can be used to compute a projected pseudogradient, the uniform stepsize and the continuity of a search direction, two key properties for the convergence results in [Y. Li and J. Zhou, *SIAM J. Sci. Comput.*, 24 (2002), pp. 865–885], get lost. In this paper, instead of proving convergence results of the algorithm for each technique, unified convergence results are obtained with a weaker stepsize assumption. An abstract existence-convergence result is also established. It is independent of the algorithm and explains why function values always converge faster than their gradients do. The weaker stepsize assumption is then verified for several different cases. As an illustration to the new results, the Banach space $W_0^{1,p}(\Omega)$ is considered and the conditions posed in the new results are verified for a quasi-linear elliptic PDE.

Key words. multiple critical points, Banach space, pseudogradient, convergence, min-orthogonal characterization

AMS subject classifications. 58E05, 58E30, 35A40, 35A65

DOI. 10.1137/050627320

1. Introduction. Let B be a Banach space, B^* its topological dual, $\langle , \rangle$ the dual relation, and $\|\cdot\|$ the norm in B. Let $J \in C^1(B, R)$ and $\nabla J \in B^*$ be its (Fréchet) gradient. A point $u^* \in B$ is a critical point of J if u^* solves the Euler–Lagrange equation $\nabla J(u^*) = 0$. The first candidates for a critical point are the local extrema. Traditional numerical algorithms focus on finding such *stable solutions*. Critical points that are not local extrema are *unstable* and called *saddle points*. In physical systems, saddle points appear as *unstable equilibria* or *transient excited states*. Multiple critical points exist in many nonlinear problems in applications (see, e.g., [3, 6, 7, 8, 10, 11]). Choi and McKenna [1] in 1993 and Ding, Costa, and Chen [2] in 1999 devised two algorithms for finding critical points of (the Morse index) MI = 1 and MI = 2, respectively. But no mathematical justification or convergence of the algorithms is established. Based on a local minimax characterization of saddle points, Li and Zhou [4] developed a local minimax algorithm (LMM) for finding critical points of MI $= 1, 2, \ldots, n$ and proved its convergence in [5]. All those algorithms are formulated in Hilbert spaces, where the gradient $\nabla J(u)$ played a key role in constructing a search direction. In order to find multiple solutions in Banach spaces [3, 10], Yao and Zhou successfully developed the first LMM in Banach spaces and solved several quasi-linear elliptic PDEs for multiple solutions [12]. The method is also modified to solve the nonlinear p-Laplacian operator for multiple eigenpairs [13]. The key to the success of Yao and Zhou's algorithm is to replace the gradient with

*Received by the editors March 22, 2005; accepted for publication (in revised form) January 5, 2007; published electronically June 7, 2007.
http://www.siam.org/journals/sinum/45-3/62732.html
†Department of Mathematics, University of Connecticut, Storrs, CT 06269 (xudong@math.uconn.edu).
‡Department of Mathematics, Texas A&M University, College Station, TX 77843 (jzhou@math.tamu.edu). This author's research was supported in part by NSF DMS-0311905.

a projected pseudogradient (PPG). The purpose of this paper is to establish some convergence results for the algorithm.

Compared to those results in Hilbert spaces [5], there are several significant differences. When B is a Hilbert space, the gradient $\nabla J(u)$, which played the key role in constructing a search direction in the LMM in [5], is uniquely determined in B and naturally continuous if J is C^1 and $B = L \oplus L^\perp$ holds for any closed subspace L. When B is a Banach space, however, the gradient $\nabla J(u)$ is in B^*, not B, and cannot be directly used as a search direction in B. Thus a PPG is introduced to the LMM. Although theoretically a Lipschitz continuous PPG flow exists, for most cases no explicit formula is available. On the other hand, there are many different ways to select a PPG. When PPGs are numerically computed in an implementation, they may belong to different PPG flows. We lost the uniform stepsize property and the continuity of a search direction, two key conditions in proving the convergence results in [5]. To make up the first loss, we design a weaker stepsize condition, called Assumption (H), to replace the old uniform stepsize property; to make up the second loss, we generalize the notion of a peak selection to that of an L-$\perp$ selection with which its continuity or smoothness can be verified. Thus corresponding modifications in the LMM [12] have to be made.

To simplify our approach, in this paper we assume $B = L \oplus L'$. When L is finite-dimensional, such an L' always exists. In particular, for the commonly used Banach space $W_0^{1,p}(\Omega)$, we present an explicit formula for obtaining L' and a practical technique to compute a PPG.

Instead of proving convergence results of the algorithm for each of the techniques used to compute a PPG, in this paper we establish some unified convergence results. To do so, in section 2 we generalize a peak selection to an L-$\perp$ selection and prove the existence of a PPG at a value of an L-$\perp$ selection. A new LMM and its mathematical foundation are also presented there. Section 3 is devoted to proving unified convergence results. We introduce a new stepsize assumption, (H), and then prove a subsequence convergence result, Theorem 3.4, under some very reasonable assumptions. An abstract existence-convergence result, Theorem 3.5, is then established. This result is actually independent of any algorithms. It also explains why in our LMM, function values always converge faster than their gradients do. Based on this abstract result, another convergence result, Corollary 3.6, is proved to show that under certain conditions, a convergent subsequence implies a point-to-set convergence. Assumption (H) is then verified for several different cases, in particular, for the commonly used Banach space $W_0^{1,p}(\Omega)$. In the last section, we discuss how to check other conditions we posed in the convergence results. In particular, we present a quasi-linear elliptic PDE and verify those conditions.

2. A min-$\perp$ method. Let L be a closed subspace of B. For a subspace $A \subseteq B$, denote $S_A = \{v \in A \mid \|v\| = 1\}$. For a point $v \in S_B$, let $[L, v] = \{tv + w \mid w \in L, t \in \mathbb{R}\}$. Since $\nabla J(u) \in B^*$, not B, it cannot be used as a search direction in B. Thus a pseudogradient is used instead.

DEFINITION 2.1. *Let $J : B \to \mathbb{R}$ be Fréchet differentiable at $u \in B$ with $\nabla J(u) \neq 0$ and let $0 < \theta \leq 1$ be given. A point $\Psi(u) \in B$ is a* pseudogradient *of J at u w.r.t. θ if*

$$\|\Psi(u)\| \leq 1, \quad \langle \nabla J(u), \Psi(u) \rangle \geq \theta \|\nabla J(u)\|. \tag{2.1}$$

A pseudogradient flow *of J w.r.t. θ is a continuous mapping $\mathcal{F} : B \to B$ such that $\forall u \in B$ with $\nabla J(u) \neq 0$, $\mathcal{F}(u)$ is a pseudogradient of J at u w.r.t. θ.*

In Definition 2.1, the condition $\|\Psi(u)\| \le 1$ is not essential. It can be replaced with any bound $M \ge 1$, since after a normalization, θ can always be replaced with $\frac{\theta}{M}$. It is known [9] that a C^1 functional has a locally Lipschitz continuous pseudogradient flow. Pseudogradients have been used in the literature to find a minimum of a functional in Banach spaces. However, as saddle points are concerned, such pseudogradients do not help much, since they lead to a local minimum point. Thus we introduce a new notion, called a projected pseudogradient (PPG), which plays a key role in the success of our LMM in Banach spaces.

DEFINITION 2.2. *An L'-PPG $G(u)$ of J is a pseudogradient of J at u such that $G(u) \in L'$.*

The motivation for defining a PPG is twofold. First, as a pseudogradient, it provides a proper searching termination criterion, i.e., with (2.1), $G(u) = 0$ implies $\nabla J(u) = 0$. Second, the condition $G(u) \in L'$ is meant to prevent a pseudogradient search from entering the subspace L, which is spanned by previously found critical points. The existence of such an L'-PPG of J at $u = \mathcal{P}(v)$ is established by Lemma 2.1 in [12], where $\mathcal{P}$ is a peak selection defined below.

DEFINITION 2.3 (see [12]). *A set-valued mapping $P : S_{L'} \to 2^B$ is called the* peak mapping *of J w.r.t. L if*

$$P(v) = \Big\{ w = \arg \textit{local-} \max_{u \in [L,v]} J(u) \Big\} \quad \forall v \in S_{L'}.$$

A mapping $\mathcal{P} : S_{L'} \to B$ is called a peak selection *of J w.r.t. L if $\mathcal{P}(v) \in P(v)$ $\forall v \in S_{L'}$. If a* peak selection $\mathcal{P}$ *is* locally *defined near a point $v \in S_{L'}$, we say that J has a* local peak selection $\mathcal{P}$ *at v.*

By using a peak selection and a PPG, an LMM is successfully developed in [12] for computing multiple saddle points in Banach spaces. However, as convergence analysis is concerned, such an algorithm has an ill-condition; i.e., the graph defined by a peak selection is not closed. In other words, a limit of a sequence of local maxima is not necessarily a local maximum point. Consequently, we cannot talk about a limit, or continuity, or do convergence analysis within the context of a peak selection. We introduce a generalized notion.

DEFINITION 2.4. *A set-valued mapping $P : S_{L'} \to 2^B$ is called the L-$\perp$* mapping *of J if*

$$P(v) = \{ u \in [L, v] : \langle \nabla J(u), w \rangle = 0 \ \forall w \in [L, v] \} \quad \forall v \in S_{L'}.$$

A mapping $\mathcal{P} : S_{L'} \to B$ is called an L-$\perp$ selection *of J if $\mathcal{P}(v) \in P(v)$ $\forall v \in S_{L'}$. If an L-$\perp$* selection $\mathcal{P}$ *is* locally *defined near a given $v \in S_{L'}$, we say that J has a* local L-$\perp$ selection $\mathcal{P}$ *at v.*

LEMMA 2.5. *If J is C^1, then the graph $Gr = \{(u, v) : v \in S_{L'},\ u \in P(v) \ne \emptyset\}$ is closed.*

Proof. Let $(u_n, v_n) \in Gr$ and $(u_n, v_n) \to (u_0, v_0)$. We have $u_n \in [L, v_n]$, $\nabla J(u_n) \perp [L, v_n]$. Since $v_n \to v_0 \in S_{L'}$, for each $w \in [L, v_0]$ there are $w_n \in [L, v_n]$ such that $w_n \to w$. Thus $\nabla J(u_n) \perp w_n$. But J is C^1, $u_n \to u_0$ and $w_n \to w$ lead to $\nabla J(u_0) \perp w$, i.e., $\nabla J(u_0) \perp [L, v_0]$. Thus $v_0 \in S_{L'}$ and $u_0 \in P(v_0)$, i.e., $(u_0, v_0) \in Gr$. ∎

It is clear that if $\mathcal{P}$ is a peak selection of J w.r.t. L, then $\mathcal{P}$ is an L-$\perp$ selection of J. The generalization not only removes the ill-condition and makes it possible to check the continuity of $\mathcal{P}$ but also exceeds the scope of a minimax principle, the most popular approach in critical point theory. It enables us to treat nonminimax-type

saddle points, such as the monkey saddles, or a problem without a mountain pass structure; see Example 2.1 in [14]. By a similar argument as in Lemma 2.1 of [12] we can prove the following existence of an L'-PPG.

LEMMA 2.6. *Assume $B = L \oplus L'$ and let $0 < \theta < 1$ be given. For $v_0 \in S_{L'}$, if $\mathcal{P}$ is a local L-$\perp$ selection of J at v_0 such that $\nabla J(\mathcal{P}(v_0)) \neq 0$ and $\Psi(\mathcal{P}(v_0)) \in B$ is a pseudogradient of J at $\mathcal{P}(v_0)$ w.r.t. θ, then there exists an L'-PPG $G(\mathcal{P}(v_0))$ of J at $\mathcal{P}(v_0)$ w.r.t. θ s.t.*

(a) *$G(\mathcal{P}(v_0)) \in L'$, $0 < \|G(\mathcal{P}(v_0))\| \leq M := \|\mathbb{P}\|$, where $\mathbb{P} : B \to L'$ is the linear projection.*
(b) *$\langle \nabla J(\mathcal{P}(v_0)), G(\mathcal{P}(v_0))\rangle = \langle \nabla J(\mathcal{P}(v_0)), \Psi(\mathcal{P}(v_0))\rangle$.*
(c) *If $\Psi(\mathcal{P}(v_0))$ is the value of a pseudogradient flow $\Psi(\cdot)$ of J at $\mathcal{P}(v_0)$, then $G(\cdot)$ is continuous and $G(\mathcal{P}(v_0))$ is called the value of an L'-PPG flow of J at $\mathcal{P}(v_0)$.*

We now establish some mathematical foundations for our new algorithm.

LEMMA 2.7 (see [12]).

$$\left\| v - \frac{v-w}{\|v-w\|} \right\| \leq \frac{2\|w\|}{\|v-w\|} \quad \forall v \in B,\ \|v\| = 1,\ \forall w \in B.$$

LEMMA 2.8. *For $v_0 \in S_{L'}$, if J has a local L-$\perp$ selection $\mathcal{P}$ at v_0 satisfying (1) $\mathcal{P}$ is continuous at v_0, (2) $d(\mathcal{P}(v_0), L) > 0$, and (3) $\nabla J(\mathcal{P}(v_0)) \neq 0$, then there exists $s_0 > 0$ such that for $0 < s < s_0$*

$$J(\mathcal{P}(v_0(s))) - J(\mathcal{P}(v_0)) < -\frac{\theta s}{4}|t_0|\|\nabla J(\mathcal{P}(v_0))\|, \tag{2.2}$$

where $\mathcal{P}(v_0) = t_0 v_0 + w_0$ for some

$$t_0 \in \mathbb{R}, \quad w_0 \in L, \quad v_0(s) = \frac{v_0 - sign(t_0) s G(\mathcal{P}(v_0))}{\|v_0 - sign(t_0) s G(\mathcal{P}(v_0))\|},$$

and $G(\mathcal{P}(v_0))$ is an L'-PPG of J w.r.t. θ at $\mathcal{P}(v_0)$.

The proof of Lemma 2.8 can follow a similar argument of Lemma 2.4 in [12]. The inequality (2.2) will be used to define a stepsize rule for the algorithm and establish convergence results. With Lemma 2.8, the following characterization of saddle points is clear.

THEOREM 2.9. *Let $v_0 \in S_{L'}$. Assume that J has a local L-$\perp$ selection $\mathcal{P}$ at v_0 such that (1) $\mathcal{P}$ is continuous at v_0, (2) $d(\mathcal{P}(v_0), L) > 0$, and (3) v_0 is a local minimum point of $J(\mathcal{P}(v))$. Then $\mathcal{P}(v_0)$ is a critical point of J.*

2.1. A min-orthogonal algorithm.

DEFINITION 2.10. *Let $v_0 \in S_{L'}$ and $\mathcal{P}$ be a local L-$\perp$ selection of J at v_0 with $\nabla J(\mathcal{P}(v_0)) \neq 0$. A point $w \in L'$ is a* descent direction *of $J(\mathcal{P}(\cdot))$ at v_0 if there is $s_0 > 0$ such that*

$$J(\mathcal{P}(v_0(s))) < J(\mathcal{P}(v_0)) \quad \forall\, 0 < s < s_0, \quad \text{where} \quad v_0(s) = \frac{v_0 + sw}{\|v_0 + sw\|}.$$

By Theorem 2.9, a descent direction method to approximate a local minimum of $J(\mathcal{P}(v))$ leads to the following min-$\perp$ algorithm.

Assume that $L = [u^1, u^2, \ldots, u^{n-1}]$, where $u^1, u^2, \ldots, u^{n-1}$ are $n-1$ previously found critical points of J and L' is a subspace of B such that $B = L \oplus L'$. For given positive numbers $\lambda, \theta \in (0,1)$ and ε.

Step 1: Let $v_1 \in S_{L'}$ be an ascent-descent direction at u^{n-1}.

Step 2: Set $k = 1$. Solve for $u_k \equiv \mathcal{P}(v_k) \equiv t_0^k v_k + t_1^k u^1 + \cdots + t_{n-1}^k u^{n-1}$ such that $t_0^k > 0$,

$$\langle \nabla J(\mathcal{P}(v_k)), v_k \rangle = 0 \quad \text{and} \quad \langle \nabla J(\mathcal{P}(v_k)), u^i \rangle = 0, \quad i = 1, 2, \ldots, n-1.$$

Step 3: Find a descent direction $w_k \in L'$ of $J(\mathcal{P}(\cdot))$ at v_k.

Step 4: If $\|\nabla J(u_k)\| \le \varepsilon$, then output $u_k = \mathcal{P}(v_k)$, stop. Otherwise, do Step 5.

Step 5: For each $s > 0$, denote

$$v_k(s) = \frac{v_k + s w_k}{\|v_k + s w_k\|}$$

and set $v_{k+1} = v_k(s_k)$, where

$$s_k = \max_{m \in \mathbb{N}} \left\{ \frac{\lambda}{2^m} \;\middle|\; 2^m > \|w_k\|,\ J\left(\mathcal{P}\left(v_k\left(\frac{\lambda}{2^m}\right)\right)\right) - J(u_k) < -\frac{\theta |t_0^k|}{4}\left(\frac{\lambda}{2^m}\right)\|\nabla J(u_k)\| \right\}.$$

Step 6: Update $k = k + 1$ and go to Step 3.

Remark 2.1.

(1) The constant $\lambda \in (0, 1)$ is used to prevent the stepsize from being too large to lose search stability. From now on we always assume that λ is such a constant.

(2) In Step 2, one way to solve the equations while satisfying the nondegenerate condition $t_0^k > 0$ is to find a local maximum point u_k of J in the subspace $[L, v_k]$, i.e., $u_k = \mathcal{P}(v_k)$ and $\mathcal{P}$ is a peak selection of J w.r.t. L.

(3) In Step 3, we may assume $\|w_k\| \le M$ for some $M \ge 1$. There are many different ways to select a descent direction w_k. However, when a descent direction is selected, a corresponding stepsize rule in Step 5 has to be designed so that it can be achieved and lead to a convergence. For example, when a negative L'-PPG flow, or a negative L'-PPG, is used as a descent direction, we have $v_k \in S_{L'}$, and a positive stepsize s_k for the current stepsize rule in Step 5 can always be obtained. In some cases, when $-\nabla J(\mathcal{P}(v_k))$ is used to construct a descent direction, the stepsize rule in Step 5 has to be modified as in Case 3 below.

3. Unified convergence results.

DEFINITION 3.1. *For each $v \in S_{L'}$ with $\|\nabla J(\mathcal{P}(v))\| \ne 0$, write $\mathcal{P}(v) = t_v v + v_L$ for some $v_L \in L$ and define the stepsize $s(v)$ at v as*

$$s(v) = \max_{m \in \mathbb{N}} \left\{ s = \frac{\lambda}{2^m} \;\middle|\; 2^m > \|w\|,\ J(\mathcal{P}(v(s))) - J(\mathcal{P}(v)) < -\frac{1}{4}\theta |t_v| s \|\nabla J(\mathcal{P}(v))\| \right\},$$

where w is a descent direction J at $\mathcal{P}(v)$ and

$$v(s) = \frac{v + sw}{\|v + sw\|}.$$

Let $\{u_k\}$ be the sequence generated by the algorithm, where $u_k = \mathcal{P}(v_k)$. Since a PPG can be computed many different ways, we lost the uniform stepsize, one of the key conditions in [5]. Here we design a new stepsize assumption, (H), to establish unified convergence results. This condition is weaker than the uniform stepsize assumption in [5] and will be verified for several different cases.

Assumption (H). If $v_0 \in S_{L'}$ with $\nabla J(\mathcal{P}(v_0)) \neq 0$ and $v_k \to v_0$, then there is $s_0 > 0$ such that $s_k = s(v_k) \geq s_0$ when k is large.

We need the following Palais–Smale (PS) condition and Ekeland's variational principle [10].

DEFINITION 3.2. *A function $J \in C^1(B, \mathbb{R})$ is said to satisfy the PS condition if any sequence $\{u_i\} \subset B$ such that $\{J(u_i)\}$ is bounded and $\nabla J(u_i) \to 0$ possesses a convergent subsequence.*

LEMMA 3.3 (Ekeland's variational principle). *Let X be a complete metric space and $J : X \to \mathbb{R} \cup \{+\infty\}$ be a lower semicontinuous function bounded from below. Then, for any $\varepsilon > 0$ and $x_0 \in X$ with $J(x_0) < +\infty$, there is $\bar{x} \in X$ such that*

$$J(\bar{x}) + \varepsilon d(x_0, \bar{x}) \leq J(x_0) \text{ and } J(x) + \varepsilon d(x, \bar{x}) > J(\bar{x}) \quad \forall x \in X \text{ and } x \neq \bar{x}.$$

First, we prove a subsequence convergence result, the conditions of which will be verified with an application problem in section 4.

THEOREM 3.4. *Let $J \in C^1(B, \mathbb{R})$ satisfy the PS condition. If an L-$\perp$ selection $\mathcal{P}$ of J satisfies that*

(1) *$\mathcal{P}$ is continuous on $S_{L'}$,*
(2) *$d(\mathcal{P}(v_k), L) \geq \alpha > 0$ $\forall k = 1, 2, \ldots,$*
(3) *$\inf_{1 \leq k < \infty} J(\mathcal{P}(v_k)) > -\infty$,*
(4) *Assumption* (H) *is satisfied,*

then

(a) *$\{v_k\}$ has a subsequence $\{v_{k_i}\}$ such that $u_{k_i} = \mathcal{P}(v_{k_i})$ converges to a critical point of J;*
(b) *if a subsequence $v_{k_i} \to v_0$ as $i \to \infty$, then $u_0 = \mathcal{P}(v_0)$ is a critical point of J.*

Proof. (a) By the stepsize rule and Lemma 2.7, for $k = 1, 2, \ldots,$ we have

$$\begin{aligned} J(u_{k+1}) - J(u_k) &\leq -\frac{1}{4}\theta\alpha s_k \|\nabla J(\mathcal{P}(v_k))\| \\ &\leq -\frac{|1-\lambda|}{4M}\theta\alpha \|v_{k+1} - v_k\| \|\nabla J(\mathcal{P}(v_k))\|. \end{aligned} \tag{3.1}$$

Suppose there is $\delta > 0$ such that $\|\nabla J(\mathcal{P}(v_k))\| \geq \delta$ for any k. From (3.1), we have

$$J(u_{k+1}) - J(u_k) \leq -\frac{|1-\lambda|}{4M}\theta\alpha\delta \|v_{k+1} - v_k\| \quad \forall k = 0, 1, 2, \ldots. \tag{3.2}$$

Adding up the two sides of (3.2) gives

$$\lim_{k\to\infty} J(u_k) - J(u_0) = \sum_{k=0}^{\infty} [J(u_{k+1}) - J(u_k)] \leq -\frac{|1-\lambda|}{4M}\theta\alpha\delta \sum_{k=0}^{\infty} \|v_{k+1} - v_k\|; \tag{3.3}$$

i.e., $\{v_k\}$ is a Cauchy sequence. Thus $v_k \to \hat{v} \in S_{L'}$. By the continuity of $\mathcal{P}$, $\|\nabla J(\mathcal{P}(\hat{v}))\| \geq \delta > 0$. On the other hand, adding up the two sides of (3.1) gives

$$\lim_{k\to\infty} J(u_k) - J(u_0) \leq -\frac{1}{4}\theta\alpha \sum_{k=0}^{\infty} s_k \|\nabla J(\mathcal{P}(v_k))\| \leq -\frac{1}{4}\theta\alpha\delta \sum_{k=0}^{\infty} s_k,$$

or $s_k \to 0$ as $k \to \infty$. It contradicts Assumption (H). Therefore, there is a subsequence $\{v_{k_i}\}$ such that $\|\nabla J(\mathcal{P}(v_{k_i}))\| \to 0$ as $i \to \infty$ and $\{J(\mathcal{P}(v_{k_i}))\}$ is convergent. By the PS condition, $\{\mathcal{P}(v_{k_i})\}$ has a subsequence that converges to a critical point u_0.

(b) Suppose $u_0 = \mathcal{P}(v_0)$ is not a critical point. Then there is $\delta > 0$ such that $\|\nabla J(u_{k_i})\| > \delta$, $i = 1, 2, \ldots$. Similar to (3.1), we have

$$J(u_{k_i+1}) - J(u_{k_i}) \le -\frac{1}{4}\theta\alpha s_{k_i}\|\nabla J(u_{k_i})\| < -\frac{1}{4}\theta\alpha\delta s_{k_i}.$$

Since $\sum_{k=0}^{\infty}[J(u_{k+1}) - J(u_k)] = \lim_{k\to\infty} J(u_k) - J(u_0)$, it leads to $\lim_{i\to\infty}(J(u_{k_i+1}) - J(u_{k_i})) = 0$. Hence $\lim_{i\to\infty} s_{k_i} = 0$. It contradicts Assumption (H). Thus u_0 is a critical point. ∎

Next we prove an abstract existence-convergence result that is actually independent of the algorithm and also explains why function values always converge faster than the gradients do. Denote $K_c = \{u \in B \mid \nabla J(u) = 0,\ J(u) = c\}$. By the PS condition, K_c is compact.

THEOREM 3.5. *Let $B = L \oplus L'$, $V \subset B$ be open, and $U = V \cap S_{L'} \neq \emptyset$. Assume that $J \in C^1(B, \mathbb{R})$ satisfies the PS condition and*

(1) *$\mathcal{P}$ is a continuous L-$\perp$ selection of J in $\bar{U}$, where $\bar{U}$ is the closure of U on $S_{L'}$.*
(2) *$\inf_{v\in U} d(\mathcal{P}(v), L) > \alpha > 0$.*
(3) *$\inf_{v\in\partial\bar{U}} J(\mathcal{P}(v)) > c = \inf_{v\in U} J(\mathcal{P}(v)) > -\infty$, where $\partial\bar{U}$ is the boundary of $\bar{U}$ on $S_{L'}$.*

Then $K_c^p = \mathcal{P}(U) \cap K_c \neq \emptyset$, and for any $\{v_k\} \subset U$ with $J(u_k) \to c$, where $u_k = \mathcal{P}(v_k)$,

(a) *$\forall \varepsilon > 0$, there is $\bar{k} > 0$ such that $d(K_c^p, u_k) < \varepsilon$ $\forall k > \bar{k}$.*
(b) *if in addition, $\nabla J(\mathcal{P}(\cdot))$ is Lipschitz continuous in U, then there is a constant C such that $\|\nabla J(u_k)\| \le C(J(u_k) - c)^{\frac{1}{2}}$.*

Proof. Define

$$\hat{J}(\mathcal{P}(v)) = \begin{cases} J(\mathcal{P}(v)), & v \in \bar{U}, \\ +\infty, & v \notin \bar{U}. \end{cases}$$

Then, $\hat{J}(\mathcal{P}(\cdot))$ is lower semicontinuous and bounded from below on the complete metric space $S_{L'}$. Let $\{v_k\} \subset U$ be any sequence such that $J(\mathcal{P}(v_k)) \to c$. By our assumption (3), such a sequence always exists. Denote $u_k = \mathcal{P}(v_k)$. Applying Ekeland's variational principle to $\hat{J}(\mathcal{P}(\cdot))$, for every $v_k \in U$ and $\delta_k = (J(u_k) - c)^{\frac{1}{2}}$ there is $\bar{v}_k \in S_{L'}$ such that

$$\hat{J}(\mathcal{P}(\bar{v}_k)) - \hat{J}(\mathcal{P}(v)) \le \delta_k\|\bar{v}_k - v\| \qquad \forall v \in S_{L'}, \tag{3.4}$$

$$\hat{J}(\mathcal{P}(\bar{v}_k)) - \hat{J}(\mathcal{P}(v_k)) \le -\delta_k\|\bar{v}_k - v_k\|. \tag{3.5}$$

By the definition of $\hat{J}(\mathcal{P}(\cdot))$ and assumptions on $\mathcal{P}$, we have $\bar{v}_k \in \bar{U}$,

$$J(\mathcal{P}(\bar{v}_k)) - J(\mathcal{P}(v)) \le \delta_k\|\bar{v}_k - v\| \qquad \forall v \in U, \tag{3.6}$$

$$J(\mathcal{P}(\bar{v}_k)) - J(\mathcal{P}(v_k)) \le -\delta_k\|\bar{v}_k - v_k\|. \tag{3.7}$$

It follows that $c \le J(\mathcal{P}(\bar{v_k})) \le J(u_k) - \delta_k\|\bar{v_k} - v_k\|$, or

$$\|\bar{v}_k - v_k\| \le \delta_k, \tag{3.8}$$

and $d(L, \mathcal{P}(\bar{v}_k)) > \alpha$ when k is large. Then $J(\mathcal{P}(v_k)) \to c$ implies $J(\mathcal{P}(\bar{v}_k)) \to c$. By assumption (3), we have $\bar{v}_k \in U$ for large k. For those large k, if $\nabla J(\mathcal{P}(\bar{v}_k)) \neq 0$, by applying Lemma 2.8 and then Lemma 2.7, when s is small, we have

$$J(\mathcal{P}(\bar{v}_k(s))) - J(\mathcal{P}(\bar{v}_k)) \le -\frac{\theta s}{4}|t_0^k|\|\nabla J(\mathcal{P}(\bar{v}_k))\| \le -\frac{\alpha\theta}{8M}\|\nabla J(\mathcal{P}(\bar{v}_k))\|\|\bar{v}_k(s) - \bar{v}_k\|,$$

where $\bar{v}_k(s) = \frac{\bar{v}_k + s\bar{w}_k}{\|\bar{v}_k + s\bar{w}_k\|} \in U$, $\bar{w}_k = -\,sign(t_0^k) G(\mathcal{P}(\bar{v}_k))$, $\mathcal{P}(\bar{v}_k) = t_0^k \bar{v}_k + u_L^k$ for some $u_L^k \in L$, $G(\mathcal{P}(\bar{v}_k))$ is an L'-PPG of J at $\mathcal{P}(\bar{v}_k)$ with $\|G(\mathcal{P}(\bar{v}_k))\| \le M$; see Lemma 2.6 and $|t_0^k| > \alpha$ by our assumption (2). Hence by (3.6), we get

$$\|\nabla J(\mathcal{P}(\bar{v}_k))\| \le \frac{8M}{\alpha\theta}\delta_k, \tag{3.9}$$

which implies $\nabla J(\mathcal{P}(\bar{v}_k)) \to 0$ and then $\nabla J(\mathcal{P}(v_k)) \to 0$ by (3.8). $\{J(\mathcal{P}(v_k))\}$ is already bounded. By the PS condition, $\{u_k\}$ has a subsequence that converges to a critical point u^*. By assumptions (3) and (1), it is clear that $u^* \in K_c^p \neq \emptyset$. Let β be any limit point of $\{d(K_c^p, u_k)\}$ and $u_{k_i} = \mathcal{P}(v_{k_i}) \in \{u_k\}$ such that $\lim_{i\to\infty} d(K_c^p, u_{k_i}) = \beta$. By the PS condition, $\{\mathcal{P}(v_{k_i})\}$ has a subsequence that converges to a critical point $\bar{u}$. Again $\bar{u} \in K_c^p$, i.e., $\beta = 0$. Thus conclusion (a) holds.

If in addition, $\nabla J(\mathcal{P}(\cdot))$ is Lipschitz continuous in U with a Lipschitz constant ℓ_1, then by (3.8) and (3.9), conclusion (b) follows from

$$\begin{aligned}\|\nabla J(\mathcal{P}(v_k))\| &\le \|\nabla J(\mathcal{P}(\bar{v}_k))\| + \|\nabla J(\mathcal{P}(v_k)) - \nabla J(\mathcal{P}(\bar{v}_k))\| \\ &\le \frac{16M}{\alpha\theta}\delta_k + \ell_1\|\bar{v}_k - v_k\| \le \left(\frac{16M}{\alpha\theta} + \ell_1\right)(J(u_k) - c)^{\frac{1}{2}}. \qquad \square\end{aligned}$$

COROLLARY 3.6. *Let $J \in C^1(B, \mathbb{R})$ satisfy the PS condition, and let V_1 and V_2 be open in L' with $\emptyset \neq U_2 \equiv V_2 \cap S_{L'} \subset V_1 \cap S_{L'} \equiv U_1$. If $\mathcal{P}$ is a continuous L-$\perp$ selection of J in U_1 with*

(1) $\inf_{v\in U_1} d(\mathcal{P}(v), L) \ge \alpha > 0$, $c = \inf_{v\in U_1} J(\mathcal{P}(v)) > -\infty$, *and* $K_c^p = \mathcal{P}(U_1) \cap K \subset K_c$,
(2) *a* $d > 0$ *with* $\inf_{v\in U_1}\{J(\mathcal{P}(v)) \mid d(v, \partial U_1) \le d\} = a > b = \sup_{v\in U_2}\{J(\mathcal{P}(v))\}$,
(3) *given* $\{v_k\}$ *such that* $v_1 \in U_2$, $\|v_{k+1} - v_k\| < d$, $J(u_{k+1}) < J(u_k)$ *and* $\{u_k\}$ *has a subsequence that converges to a critical point* u_0, *where* $u_k = \mathcal{P}(v_k)$,

then

(a) $\forall \varepsilon > 0$, *there is* $\bar{k} > 0$ *such that* $d(K_c^p, u_k) < \varepsilon$ $\forall k > \bar{k}$,
(b) *if in addition,* $\nabla J(\mathcal{P}(\cdot))$ *is Lipschitz continuous in* U_1, *then there is a constant* C *such that* $\|\nabla J(u_k)\| \le C(J(u_k) - c)^{\frac{1}{2}}$.

Proof. First, we prove that $v_k \in U_1$ and $d(v_k, \partial U_1) > d$, $k = 1, 2, \ldots$. In fact, if $v_k \in U_1$, $d(v_k, \partial U_1) > d$, and $J(u_k) \le b$, then $v_{k+1} \in U_1$ and $J(u_{k+1}) < b$, i.e., $v_{k+1} \in U_1$ and $d(v_{k+1}, \partial U_1) > d$. Thus, for $v_1 \in U_2$, $v_k \in U_1$ and $d(v_k, \partial U_1) > d$, $k = 1, 2, \ldots$. Since $K_c^p = \mathcal{P}(U_1) \cap K \subset K_c$ and $\{u_k\}$ has a subsequence that converges to a critical point u_0, we have $u_0 \in K_c^p \neq \emptyset$. Denote $U = \{v \in U_1 \mid d(v, \partial U_1) > d\}$. Then by the monotonicity of $\{J(u_k)\}$, we have $J(u_k) \to c = \inf_{v\in U} J(\mathcal{P}(v))$ as $k \to \infty$, and

$$\inf_{v\in\partial\bar{U}} J(\mathcal{P}(v)) \ge a > b \ge J(\mathcal{P}(v_1)) \ge c = \inf_{v\in U} J(\mathcal{P}(v)).$$

Thus all the assumptions of Theorem 3.5 are satisfied and the conclusions follow. $\square$

Remark 3.1. There are two types of conditions posed in our convergence results. One is used in the literature to guarantee the existence of multiple solutions. The other is what we posed to ensure a convergence for the algorithm. We will focus on verification of the latter.

(1) In Theorem 3.5, condition (3) can be simplified as $c = \inf_{v\in S_{L'}} J(\mathcal{P}(v)) > -\infty$ if $U = S_{L'}$.

(2) In Corollary 3.6, condition (2) is, or its variants are, frequently used in the literature to form a topological linking for applying a deformation lemma to prove an existence of multiple solutions. It is clear that condition (3) in Theorem 3.5 is much weaker. It is used to trap a descending flow away from critical points at other levels. Condition (3) in Corollary 3.6 is designed for our algorithm to cover several different cases in Banach spaces and is guaranteed by our Assumption (H) and Theorem 3.4.

(3) Note that when K_c^p contains only one point, Theorem 3.5 can be easily stated as a point-to-point convergence result. Theorem 3.5, together with its Corollary 3.6, improves a convergence result, Theorem 3.3 in [5], in Hilbert space in several directions, which (a) cover several different cases in Banach spaces, (b) do not require $\mathcal{P}$ to be a homeomorphism, and (c) contain a new result on relative convergence rate, i.e., inequality (2), which explains why in our numerical computations, $J(\mathcal{P}(v_n^k))$ always converges much faster than $\|\nabla J(\mathcal{P}(v_n^k))\| \to 0$.

Next we verify Assumption (H) for several different cases. This is done in Lemmas 3.8, 3.9, 3.11, and 3.12 below. Cases 1 and 2 are general, so we assume $B = L \oplus L'$, where L' is a closed subspace of B and $\mathbb{P} : B \to L'$ is the corresponding projection. In Step 3 of the algorithm, we choose $w_k = -\,sign(t_0)G(\mathcal{P}(v_k))$, where G is either an L'-PPG of J or the value of an L'-PPG flow of J. Then $\|w_k\| \le M = \|\mathbb{P}\|$. By Lemma 2.8 we obtain the following.

LEMMA 3.7. *If $\mathcal{P}$ is a local L-$\perp$ selection of J at $v \in S_{L'}$ such that* (1) *$\mathcal{P}$ is continuous at v,* (2) *$d(\mathcal{P}(v), L) > 0$, and* (3) *$\nabla J(\mathcal{P}(v)) \neq 0$, then $s(v) > 0$.*

Case 1. Use the value of a negative PPG flow G as a descent direction.

Here $G(\mathcal{P}(v_k))$ is the value of an L'-PPG flow of J at $\mathcal{P}(v_k) = t_0^k v_k + v_k^L$ for some $v_k^L \in L$.

LEMMA 3.8. *If $\mathcal{P}$ is a local L-$\perp$ selection of J at $v_0 \in S_{L'}$ such that* (1) *$\mathcal{P}$ is continuous at v_0,* (2) *$d(\mathcal{P}(v_0), L) > 0$, and* (3) *$\nabla J(\mathcal{P}(v_0)) \neq 0$, then there exist $\varepsilon > 0$ and $s_0 = \frac{\lambda}{2^m}$ for some integer m such that for each $v \in S_{L'}$ with $\|v - v_0\| < \varepsilon$, $\lambda \ge s_0\|G(\mathcal{P}(v))\|$ and*

$$J(\mathcal{P}(v(s_0))) - J(\mathcal{P}(v)) < -\frac{s_0\theta|t_v|}{4}\|\nabla J(\mathcal{P}(v))\|, \quad v(s_0) = \frac{v + sign(t_v)s_0 G(\mathcal{P}(v))}{\|v + sign(t_v)s_0 G(\mathcal{P}(v))\|},$$

$\mathcal{P}(v) = t_v v + w_v$ for some $w_v \in L$, and $G(\mathcal{P}(v))$ is the value of an L'-PPG flow of J at $\mathcal{P}(v)$ w.r.t. the constant θ.

Proof. By Lemma 2.8, there is $\bar{s} > 0$ such that as $0 < s < \bar{s}$,

$$J(\mathcal{P}(v_0(s))) - J(\mathcal{P}(v_0)) < -\frac{s\theta|t_0|}{4}\|\nabla J(\mathcal{P}(v_0))\|, \tag{3.10}$$

where

$$v_0(s) = \frac{v_0 - sign(t_0)sG(\mathcal{P}(v_0))}{\|v_0 - sign(t_0)sG(\mathcal{P}(v_0))\|}, \quad \mathcal{P}(v_0) = t_0 v_0 + w_0$$

for some $w_0 \in L$. Actually, for each fixed s, the two sides of (3.10) are continuous in v_0. Thus, there are $\varepsilon > 0$, $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda \ge s_0\|G(\mathcal{P}(v))\|$ and

$$J(\mathcal{P}(v(s_0))) - J(\mathcal{P}(v)) < -\frac{s_0\theta|t_v|}{4}\|\nabla J(\mathcal{P}(v))\|$$

$\forall v \in S_{L'}$ with $\|v - v_0\| \le \varepsilon$. ∎

Case 2. Use a negative PPG G as a descent direction.

Here $G(\mathcal{P}(v_k))$ is an L'-PPG of J at $\mathcal{P}(v_k) = t_0^k v_k + v_k^L$ for some $v_k^L \in L$. Since an L'-PPG may be chosen from different L'-PPG flows, we lost the continuity. To compensate for the loss, we assume that an L-$\perp$ selection $\mathcal{P}$ of J is locally Lipschitz continuous.

LEMMA 3.9. *Let $\mathcal{P}$ be a local L-$\perp$ selection of J at $v_0 \in S_{L'}$. If (1) $\mathcal{P}$ is Lipschitz continuous in a neighborhood of v_0, (2) $d(\mathcal{P}(v_0), L) > 0$, and (3) $\nabla J(\mathcal{P}(v_0)) \neq 0$, then there are $\varepsilon > 0$ and $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda \ge s_0 \|G(\mathcal{P}(v))\|$ and*

$$J(\mathcal{P}(v(s_0))) - J(\mathcal{P}(v)) < -\frac{1}{4} s_0 \theta |t_v| \|\nabla J(\mathcal{P}(v))\|$$

$\forall v \in S_{L'}$ with $\|v - v_0\| < \varepsilon$, where

$$v(s) = \frac{v - sign(t_v) s G(\mathcal{P}(v))}{\|v - sign(t_v) s G(\mathcal{P}(v))\|},\ s > 0, \quad \mathcal{P}(v) = t_v v + w_v \text{ for some } w_v \in L,$$

and $G(\mathcal{P}(v))$ is an L'-PPG of J at $\mathcal{P}(v)$ w.r.t. the constant θ.

Proof. First, denote $\mathcal{P}(v(s)) = t_v^s v(s) + w_v(s)$ for some $w_v(s) \in L$. We have

$$\begin{aligned} &J(\mathcal{P}(v(s))) - J(\mathcal{P}(v)) \\ &\quad = \langle \nabla J(\mathcal{P}(v)) + (\nabla J(\zeta(v,s)) - \nabla J(\mathcal{P}(v))),\, \mathcal{P}(v(s)) - \mathcal{P}(v) \rangle, \end{aligned} \tag{3.11}$$

where $\zeta(v,s) = (1-\eta)\mathcal{P}(v) + \eta \mathcal{P}(v(s))$ for some $\eta \in [0,1]$. By assumption (1) and Lemma 2.7, as s is small and for any v close to v_0,

$$\|\mathcal{P}(v(s)) - \mathcal{P}(v)\| \le \ell \|v(s) - v\| \le \frac{2\ell s \|G(\mathcal{P}(v))\|}{\|v - sign(t_v) s G(\mathcal{P}(v))\|} \le 4\ell M s. \tag{3.12}$$

On the other hand, by the definition of an L-$\perp$ selection of J, we have

$$\begin{aligned} \langle \nabla J(\mathcal{P}(v)), \mathcal{P}(v(s)) - \mathcal{P}(v) \rangle &= -\frac{sign(t_v) t_v^s s \langle \nabla J(\mathcal{P}(v)), G(\mathcal{P}(v)) \rangle}{\|v - sign(t_v) s G(\mathcal{P}(v))\|} \\ &= -\frac{|t_v^s| s \langle \nabla J(\mathcal{P}(v)), \Psi(\mathcal{P}(v)) \rangle}{\|v - sign(t_v) s G(\mathcal{P}(v))\|} \le -\frac{s\theta |t_v| \|\nabla J(\mathcal{P}(v))\|}{2} < 0 \end{aligned} \tag{3.13}$$

and

$$\begin{aligned} &|\langle \nabla J(\zeta(v,s)) - \nabla J(\mathcal{P}(v)),\, \mathcal{P}(v(s)) - \mathcal{P}(v) \rangle| \\ &\quad \le \|\nabla J(\zeta(v,s)) - \nabla J(\mathcal{P}(v))\|\, \|\mathcal{P}(v(s)) - \mathcal{P}(v)\| \le \frac{s\theta |t_v| \|\nabla J(\mathcal{P}(v))\|}{4}, \end{aligned} \tag{3.14}$$

where in the last inequality, since J is C^1 and by assumptions (2) and (3), we have

$$\|\nabla J(\zeta(v,s)) - \nabla J(\mathcal{P}(v))\| \le \frac{\theta |t_v| \|\nabla J(\mathcal{P}(v))\|}{16 \ell M}. \tag{3.15}$$

By (3.11) and the boundedness of L'-PPGs, there exist $\varepsilon > 0$ and $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda \ge s_0 \|G(\mathcal{P}(v))\|$ and

$$J(\mathcal{P}(v(s_0))) - J(\mathcal{P}(v)) \le -\frac{s_0 \theta |t_v| \|\nabla J(\mathcal{P}(v))\|}{4} \quad \forall v \in S_{L'} \text{ with } \|v - v_0\| < \varepsilon. \quad ∎$$

Case 3. Use a practical technique for a descent direction in $B = W_0^{1,p}(\Omega)$.

Let $B = W_0^{1,p}(\Omega)$ where $\Omega \subset \mathbb{R}^n$ is open and bounded, $p > 1$, $B^* = W^{-1,q}(\Omega)$ with $\frac{1}{p} + \frac{1}{q} = 1$. The usual gradient $\delta J(u) \in B^* = W^{-1,q}(\Omega)$ cannot be used as a search direction. Thus $d = \Delta_p^{-1}(\delta J(u)) \in B$ has been used in the literature as a descent direction to find a local minimum of $J : B \to \mathbb{R}$, where Δ_p is the p-Laplacian operator defined in (4.1) and Δ_p^{-1} is its inverse. It leads to solving a sequence of quasi-linear elliptic equations $\Delta_p d_k = \delta J(u_k)$. But such a d is not a PPG and does not help much in finding a saddle point. A practical technique is used in [12] for numerical implementation to compute a PPG. The results are very promising. Here we wish to provide some mathematical justification. This technique is based on the understanding that when a nice smooth initial guess v_0 is used, we may expect that "nice" functions are actually used to approximate a critical point. Let $\mathcal{P}$ be an L-$\perp$ selection of J. For $v \in S_{L'}$, $u = \mathcal{P}(v)$, by the definition of $\mathcal{P}$, $\delta J(u) \perp L$. But $\delta J(u) \in W^{-1,q}(\Omega)$, and its smoothness is poor. We first lift its smoothness by computing $d := \nabla J(u) = \Delta^{-1}(-\delta J(u)) \in W_0^{1,q}(\Omega)$, i.e., $d_k = \nabla J(u_k)$ is solved from

$$\Delta d_k(x) = -\delta J(u_k)(x),\ x \in \Omega, \quad d_k(x)|_{\partial\Omega} = 0.$$

Observe that notationally for any $w \in B$,

$$\begin{aligned}\langle d, w\rangle_{W_0^{1,q}\times W_0^{1,p}} &\equiv \langle \nabla d, \nabla w\rangle_{L^q\times L^p} \equiv \int_\Omega \nabla d(x)\cdot\nabla w(x)\,dx \\ &= \int_\Omega -\Delta d(x) w(x)\,dx = \int_\Omega \delta J(u)(x) w(x)\,dx \equiv \langle \delta J(u), w\rangle_{W^{-1,q}\times W_0^{1,p}}.\end{aligned}$$

This suggests that $d = \nabla J(u)$ be used as a gradient of J at u. In particular when $u = \mathcal{P}(v)$,

$$\langle \nabla J(u), w\rangle_{W_0^{1,q}\times W_0^{1,p}} = \langle \delta J(u), w\rangle_{W^{-1,q}\times W_0^{1,p}} = 0 \quad \forall w \in [L, v], \tag{3.16}$$

or $\nabla J(u) \perp [L, v]$. This suggests a natural way to choose L'. We will discuss it later. Since

$$\begin{aligned}\|\delta J(u)\|_{W^{-1,q}} &= \sup_{\|w\|_{W_0^{1,p}}=1} \langle \delta J(u), w\rangle_{W^{-1,q}\times W_0^{1,p}} \\ &= \sup_{\|w\|_{W_0^{1,p}}=1} \langle d, w\rangle_{W_0^{1,q}\times W_0^{1,p}} = \sup_{\|\nabla w\|_{L^p}=1} |\langle \nabla d, \nabla w\rangle_{L^q\times L^p}| \le \|d\|_{W_0^{1,q}},\end{aligned} \tag{3.17}$$

the PS condition of J in terms of δJ implies the PS condition of J in terms of ∇J. In our convergence analysis of the algorithm, the first order approximation contains a term

$$\begin{aligned}\langle \delta J(v_0), \nabla J(v_0)\rangle_{W^{-1,q}\times W_0^{1,p}} &= \int_\Omega \delta J(v_0)\nabla J(v_0)\,dx \\ &= \int_\Omega -\Delta(\nabla J(v_0))\nabla J(v_0)\,dx = \int_\Omega \nabla(\nabla J(v_0))\cdot\nabla(\nabla J(v_0))\,dx = \|\nabla J(v_0)\|_{W^{1,2}}^2,\end{aligned}$$

which will be used to design a new stepsize rule. Next we let $u_k = \mathcal{P}(v_k)$ and check the ratio

$$1 \ge \theta_k \equiv \frac{\|\nabla J(u_k)\|_2^2}{\|\nabla J(u_k)\|_q \|\nabla J(u_k)\|_p} \ge \theta > 0 \quad \forall k = 1, 2, \ldots, \tag{3.18}$$

where $\|\cdot\|_r$ is the norm in $W_0^{1,r}(\Omega)$ with $r > 1$. Let

$$G(u_k) = \frac{\nabla J(u_k)}{\|\nabla J(u_k)\|_p}.$$

We have $\|G(u_k)\|_p \le M = 1$ and

$$\begin{aligned}\langle \delta J(u_k), G(u_k)\rangle &= \frac{\|\nabla J(u_k)\|_2^2}{\|\nabla J(u_k)\|_p} = \theta_k \|\nabla J(u_k)\|_q \\ &\ge \theta_k \|\delta J(u_k)\|_{W^{-1,q}} \ge \theta \|\delta J(u_k)\|_{W^{-1,q}},\end{aligned}$$

where the last inequality holds if (3.18) is satisfied, i.e., $G(u_k)$ is a pseudogradient of J at u_k. Then (3.16) suggests that $G(u_k)$ is also an L'-PPG of J at $u_k = \mathcal{P}(v_k)$ if $L' = L^\perp$, where $L^\perp$ is given in (3.19), $L = [w_1, \dots, w_{n-1}]$, and $w_1, \dots, w_{n-1}$ are linearly independent. To show $B = L \oplus L^\perp$, we need further assume that when $1 < p < 2$, $w_1, \dots, w_{n-1}$ are $n-1$ previously found nice critical points, or at least they are nice approximations of some exact critical points such that $L \subset W_0^{1,q}(\Omega)$. Such an assumption holds automatically when $2 \le p$. Thus

$$L' := L^\perp = \left\{ u \in B \;\middle|\; \int_\Omega \nabla u(x) \cdot \nabla v(x)\, dx = 0 \ \forall v \in L \right\} \tag{3.19}$$

is well defined and $L \cap L' = \{0\}$ holds. For any $w \in B$, we compute

$$w_L := \sum_{i=1}^{n-1} \alpha_i w_i$$

from

$$\int_\Omega \nabla w_L(x) \cdot \nabla w_j(x)\, dx = \int_\Omega \nabla w(x) \cdot \nabla w_j(x)\, dx, \quad j = 1, \dots, n-1.$$

Thus $w_L \in L$ and $w - w_L \in L^\perp$, i.e., $B = L \oplus L'$.

But we cannot assume that such a $G(u_k)$ is the value of a PPG flow of J at $u_k = \mathcal{P}(v_k)$, because we do not know the ratio at other points. In all our numerical examples, (3.18) is satisfied. But we note that the ratio is stable for $1 < p \le 2$ and gets closer to 0 as $p \to +\infty$. Thus we treat those two cases differently in our convergence analysis. For $1 < p \le 2$, we assume that (3.18) is satisfied. But for $p > 2$, we only assume $\|\nabla J(u_k)\|_p \le M$ for some $M > 0$. Either one of the assumptions implies $\nabla J(u_k) \in L^\perp \subset B$. By comparing $G(u_k)$ and $\nabla J(u_k)$, Step 3 and the stepsize rule in Step 5 need to be modified as below.

Step 3: Find a descent direction w_k of J at $u_k = \mathcal{P}(v_k)$, $w_k = -\,sign(t_0^k)\nabla J(u_k)$. Compute the ratio

$$\theta_k = \frac{\|w_k\|_2^2}{\|w_k\|_p \|w_k\|_q} > 0.$$

Since $\langle \delta J(u_k), \nabla J(u_k)\rangle = \|\nabla J(u_k)\|_2^2$, the stepsize rule in Step 5 has to be changed to

$$s_k = \max_{m \in \mathbb{N}} \left\{ s = \frac{\lambda}{2^m} \;\middle|\; 2^m > \|w_k\|,\ J(\mathcal{P}(v_k(s))) - J(u_k) \le \frac{|t_0^k| s}{-4} \|\nabla J(u_k)\|_2^2 \right\}, \tag{3.20}$$

where $0 < \lambda < 1$. Note that if $\theta_k > \theta > 0$, theoretically the term $\|\nabla J(u_k)\|_2^2$ in the above stepsize rule can be replaced with $\theta\|\nabla J(u_k)\|_q$, i.e., we use

$$G(\mathcal{P}(v_k)) = \frac{\nabla J(u_k)}{\|\nabla J(u_k)\|_p}$$

as an L'-PPG of J at $u_k = \mathcal{P}(v_k)$. Then this case can be covered by Case 2. But in implementation, the lower bound θ of the ratio is usually not known beforehand. In particular, we do not known whether or not the ratio is satisfied at a limit point of the sequence. Thus the current stepsize rule (3.20) has to be used in implementation. First, we show that if $0 < \|\nabla J(\mathcal{P}(v_0))\|_p < +\infty$, then a positive stepsize can always be attained.

LEMMA 3.10. *For $v_0 \in S_{L'}$, if J has a local L-$\perp$ selection $\mathcal{P}$ at v_0 satisfying (1) $\mathcal{P}$ is continuous at v_0, (2) $d(\mathcal{P}(v_0), L) > 0$, and (3) $0 < \|\nabla J(\mathcal{P}(v_0))\|_2 < +\infty$, then there exists $s_0 > 0$ such that as $0 < s < s_0$*

$$J(\mathcal{P}(v_0(s))) - J(\mathcal{P}(v_0)) < -\frac{|t_0|s}{4}\|\nabla J(\mathcal{P}(v_0))\|_2^2, \tag{3.21}$$

where

$$v_0(s) = \frac{v_0 - sign(t_0)s\nabla J(\mathcal{P}(v_0))}{\|v_0 - sign(t_0)s\nabla J(\mathcal{P}(v_0))\|}$$

and $\mathcal{P}(v_0) = t_0 v_0 + w_0$ for some $t_0 \in \mathbb{R}$, $w_0 \in L$.

Proof. Since $\|\mathcal{P}(v_0(s)) - \mathcal{P}(v_0)\| \to 0$ as $s \to 0$, we have

$$\begin{aligned}
&J(\mathcal{P}(v_0(s)) - J(\mathcal{P}(v_0)) \\
&\quad = \langle \delta J(\mathcal{P}(v_0)), \mathcal{P}(v_0(s)) - \mathcal{P}(v_0)\rangle + o(\|\mathcal{P}(v_0(s)) - \mathcal{P}(v_0)\|) \\
&\quad = -\frac{sign(t_0)t_0^s s\|\nabla J(\mathcal{P}(v_0))\|_2^2}{\|v_0 - sign(t_0)s\nabla J(\mathcal{P}(v_0))\|} + o(\|\mathcal{P}(v_0(s)) - \mathcal{P}(v_0)\|) < -\frac{|t_0|s}{4}\|\nabla J(\mathcal{P}(v_0))\|_2^2,
\end{aligned}$$

where $\mathcal{P}(v_0(s)) = t_0^s v_0(s) + w_0^s$ and $w_0^s \in L$, when $s > 0$ is very small. ∎

Next we verify Assumption (H).

Subcase $p < 2$ (we assume (3.18) holds). We have $\nabla J(u_k) \in W_0^{1,q}(\Omega) \subset B$. The conclusion in the next lemma is actually stronger than Assumption (H).

LEMMA 3.11. *Let $J \in C^1(B, \mathbb{R})$ and $v_0 \in S_{L^\perp}$. Let $\mathcal{P}$ be a local L-$\perp$ selection of J at v_0 such that $\mathcal{P}$ is continuous at v_0 and $d(\mathcal{P}(v_0), L) > 0$. If $\delta J(\mathcal{P}(v_0)) \neq 0$, then there are $\varepsilon > 0$ and $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda \geq s_0\|\nabla J(\mathcal{P}(v))\|_p$ and*

$$J(\mathcal{P}(v(s_0))) - J(\mathcal{P}(v)) < -\frac{|t_v|s_0}{4}\|\nabla J(\mathcal{P}(v))\|_2^2 \quad \forall v \in S_{L'},\ \|v - v_0\| < \varepsilon,$$

where $\mathcal{P}(v) = t_v v + w$, $w \in L$, and

$$v(s_0) = \frac{v - sign(t_v)s_0\nabla J(\mathcal{P}(v))}{\|v - sign(t_v)s_0\nabla J(\mathcal{P}(v))\|}.$$

Proof. By Lemma 3.10, we have

$$J(\mathcal{P}(v_0(s))) - J(\mathcal{P}(v_0)) < -\frac{|t_0|s}{4}\|\nabla J(\mathcal{P}(v_0))\|_2^2. \tag{3.22}$$

When $p < 2$, we have $q > 2$. J is C^1 implies that ∇J is continuous in the $\|\cdot\|_2$-norm. For fixed s, all the terms in (3.22) are continuous in v_0. Thus there exist $\varepsilon > 0$ and $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda \geq s_0\|\nabla J(\mathcal{P}(v))\|_p$ since J is C^1 and

$$J(\mathcal{P}(v(s_0))) - J(\mathcal{P}(v)) < -\frac{|t_v|s_0}{4}\|\nabla J(\mathcal{P}(v))\|_2^2 \quad \forall v \in S_{L'},\ \|v - v_0\| < \varepsilon. \qquad \square$$

With the new stepsize rule and the uniform stepsize result, Lemma 3.11, if $\theta_k > \theta > 0$ holds in Step 3. We can verify Theorem 3.4. The proof is similar. We need only replace (3.1) with

$$\begin{aligned} J(u_{k+1}) - J(u_k) &< -\frac{\alpha s_k}{4}\|\nabla J(u_k)\|_2^2 \quad \text{(by (3.18))} \\ &< -\frac{\alpha\theta s_k}{4}\|\nabla J(u_k)\|_p\|\nabla J(u_k)\|_q < -\frac{\alpha\theta|1-\lambda|}{4M}\|v_{k+1} - v_k\|\|\nabla J(u_k)\|_q, \end{aligned} \tag{3.23}$$

where $0 < \lambda < 1$ is given in (3.20), and then follow the proof.

Subcase $p > 2$ (we assume only $\|\nabla J(u_k)\|_p \leq M$ for some $M > 0$). We have $B = W_0^{1,p}(\Omega) \subset W_0^{1,2}(\Omega)$. To verify Assumption (H) and prove the convergence of the algorithm, we note that in this case, $\nabla J(u_k) \in L' = L^\perp$ still holds; i.e., $-\nabla J(u_k)$ can be used as a search direction. But "J is C^1" means that δJ is continuous in the $\|\cdot\|_{(-1,q)}$-norm and ∇J is continuous in the $\|\cdot\|_q$-norm, but not necessarily in the $\|\cdot\|_2$-norm. Thus we need an L-$\perp$ selection $\mathcal{P}$ to be locally Lipschitz continuous.

LEMMA 3.12. *Let $J \in C^1(B, \mathbb{R})$ and $v_0 \in S_{L'}$. Assume that $\mathcal{P}$ is a local L-$\perp$ selection of J at v_0 such that* (1) *$\mathcal{P}$ is locally Lipschitz continuous,* (2) *$d(\mathcal{P}(v_0), L) > 0$, and* (3) *$\delta J(\mathcal{P}(v_0)) \neq 0$. Then for any $v_k \in S_{L'}$ with $\lim_{k\to\infty} v_k = v_0$ and $\|\nabla J(\mathcal{P}(v_k))\|_p \leq M$, there are $\bar{k}$, $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda > s_0\|\nabla J(\mathcal{P}(v_k))\|_p$ and*

$$J(\mathcal{P}(v_k(s_0))) - J(\mathcal{P}(v_k)) \leq -\frac{s_0|t_k|\|\nabla J(\mathcal{P}(v_k))\|_2^2}{4} \quad \forall k > \bar{k},$$

where

$$v_k(s) = \frac{v_k - sign(t_k)s\nabla J(\mathcal{P}(v_k))}{\|v_k - sign(t_k)s\nabla J(\mathcal{P}(v_k))\|}$$

and $\mathcal{P}(v_k) = t_k v_k + v_k^L$, $v_k^L \in L$.

Proof. Denote $\mathcal{P}(v_k(s)) = t_k^s v_k(s) + v_k^L(s)$ for some $v_k^L(s) \in L$. Then, by the mean value theorem, we have

$$\begin{aligned} &J(\mathcal{P}(v_k(s))) - J(\mathcal{P}(v_k)) \\ &\quad = \langle \delta J(\mathcal{P}(v_k)) + (\delta J(\zeta(v_k, s)) - \delta J(\mathcal{P}(v_k))),\ \mathcal{P}(v_k(s)) - \mathcal{P}(v_k)\rangle, \end{aligned} \tag{3.24}$$

where $\zeta(v_k, s) = (1 - \lambda_k)\mathcal{P}(v_k) + \lambda_k\mathcal{P}(v_k(s))$ for some $\lambda_k \in [0, 1]$. By assumption (1) and Lemma 2.7,

$$\|\mathcal{P}(v_k(s)) - \mathcal{P}(v_k)\| \leq \ell\|v_k(s) - v_k\| \leq \frac{2\ell s\|\nabla J(\mathcal{P}(v_k))\|_p}{\|v_k - sign(t_k)s\nabla J(\mathcal{P}(v_k))\|}.$$

On the other hand, by the definition of an L-$\perp$ selection of J, as $s > 0$ is small and k is large, we have

$$\begin{aligned} \langle \delta J(\mathcal{P}(v_k)), \mathcal{P}(v_k(s)) - \mathcal{P}(v_k)\rangle &= -\frac{sign(t_k)t_k^s s\|\nabla J(\mathcal{P}(v_k))\|_2^2}{\|v_k - sign(t_k)s\nabla J(\mathcal{P}(v_k))\|} \\ &\leq -\frac{s|t_k|}{2}\|\nabla J(\mathcal{P}(v_k))\|_2^2 < 0. \end{aligned} \tag{3.25}$$

Since J is C^1 and $1 < q < 2$ in this case, by assumptions (2) and (3) and applying inequality (3.17), there exists $\delta > 0$ such that when s is small and k is large,

$$\frac{|t_k|\|v_k - sign(t_k)s\nabla J(\mathcal{P}(v_k))\|\|\nabla J(\mathcal{P}(v_k))\|_2^2}{8\ell\|\nabla J(\mathcal{P}(v_k))\|_p} > \delta > 0.$$

Thus we can choose $s > 0$ small and k large such that

$$\|\delta J(\zeta(v_k, s)) - \delta J(\mathcal{P}(v_k))\| \le \frac{|t_k|\|v_k - sign(t_k)s\nabla J(\mathcal{P}(v_k))\|\|\nabla J(\mathcal{P}(v_k))\|_2^2}{8\ell\|\nabla J(\mathcal{P}(v_k))\|_p}.$$

Hence

$$\begin{aligned} &|\langle \delta J(\zeta(v_k, s)) - \delta J(\mathcal{P}(v_k)),\, \mathcal{P}(v_k(s)) - \mathcal{P}(v_k)\rangle| \\ &\le \|\delta J(\zeta(v_k, s)) - \delta J(\mathcal{P}(v_k))\|\|\mathcal{P}(v_k(s)) - \mathcal{P}(v_k)\| \le \frac{s|t_k|\|\nabla J(\mathcal{P}(v_k))\|_2^2}{4}. \end{aligned} \tag{3.26}$$

Applying inequalities (3.25) and (3.26) to (3.24), we see that there exist $\bar{k}$, $s_0 = \frac{\lambda}{2^m}$ for some integer m such that $\lambda > s_0\|\nabla J(\mathcal{P}(v_k))\|_p$ and

$$J(\mathcal{P}(v_k(s_0))) - J(\mathcal{P}(v_k)) \le -\frac{s_0|t_k|\|\nabla J(\mathcal{P}(v_k))\|_2^2}{4} \quad \forall k > \bar{k}. \qquad \square$$

With the new stepsize rule (3.20) and the assumption $\|\nabla J(u_k)\|_p < M$, the conclusion of Lemma 3.12 implies Assumption (H), i.e., $s(v_k) \ge s_0$. Then the convergence result, Theorem 3.4, can be verified. The proof is similar. Note that when $\|\nabla J(u_k)\|_q > \delta_0$ for some $\delta_0 > 0$, $\|\nabla J(u_k)\|_2 > \delta$ for some $\delta > 0$. We need only replace (3.1) and (3.2) with

$$\begin{aligned} J(u_{k+1}) - J(u_k) &< -\frac{\alpha s_k}{4}\|\nabla J(u_k)\|_2^2 \le -\frac{\alpha s_k}{4}\delta^2 \\ &\le -\frac{\alpha s_k \delta^2}{4M}\|\nabla J(u_k)\|_p \le -\frac{\alpha\delta^2|1-\lambda|}{4M}\|v_{k+1} - v_k\|_p, \end{aligned}$$

where $0 < \lambda < 1$ is given in (3.20) and the last inequality follows from Lemma 2.7. Then following the proof, the unified convergence result, Theorem 3.4, is also obtained.

4. An application to the nonlinear p-Laplacian equation. As an application, let us consider the following quasi-linear elliptic boundary-value problem on a bounded smooth domain $\Omega \subset \mathbb{R}^n$:

$$\begin{cases} \Delta_p u(x) + f(x, u(x)) = 0, & x \in \Omega, \\ u(x) = 0, & x \in \partial\Omega, \end{cases} \qquad u \in B \equiv W^{1,p}(\Omega),\ p > 1, \tag{4.1}$$

where Δ_p defined by $\Delta_p u(x) = \mathrm{div}(|\nabla u(x)|^{p-2}\nabla u(x))$ is the p-Laplacian operator which has a variety of applications in physical fields, such as in fluid dynamics when the shear stress and the velocity gradient are related in a certain manner, where $p = 2$, $p < 2$, $p > 2$ if the fluid is Newtonian, pseudoplastic, or dilatant, respectively. The p-Laplacian operator also appears in the study of flow in porous media ($p = \frac{3}{2}$), nonlinear elasticity ($p > 2$), and glaciology ($p \in (1, \frac{4}{3})$). Under certain standard conditions on f, it can be shown that a point $u^* \in W_0^{1,p}(\Omega)$ is a weak solution of (4.1) if and only if u^* is a critical point of the functional

$$J(u) = \frac{1}{p}\int_\Omega |\nabla u(x)|^p\, dx - \int_\Omega F(x, u(x))\, dx, \quad \text{where} \quad F(x,t) = \int_0^t f(x,s)\, ds. \tag{4.2}$$

Many multiple solutions to the above quasi-linear elliptic PDE have been numerically computed in [12, 13] for $p < 2$ and $p > 2$. Convergence results obtained in section 3 can be applied; see Case 3. Since conditions (1), (2), and (3) in Theorem 3.4 are basic assumptions in our results and new in the literature, we focus on verifying them in this section. Other conditions, such as the PS condition, have been studied in the literature and therefore will not be discussed here. Let us assume some of the standard growth and regularity conditions in the literature. Denote the Sobolev exponent $p^* = \frac{np}{n-p}$ for $p < n$ and $p^* = \infty$ for $p \geq n$. Assume

(a) $f \in C^1(\bar{\Omega} \times \mathbb{R}, \mathbb{R})$, $f(x,0) = 0$, $\frac{f(x,t\xi)}{|t\xi|^{p-2}t\xi}$ monotonically increases to $+\infty$ in t.

(b) For each $\varepsilon > 0$, there is $c_1 = c_1(\varepsilon) > 0$ such that $f(x,t)t < \varepsilon|t|^p + c_1|t|^{p^*}$ $\forall t \in \mathbb{R}$, $x \in \Omega$.

It is clear that $u = 0$ is a critical point of the least critical value of J and $f(x,u) = |u|^{q-2}u$ for $q > p$ satisfies condition (a). For each $v \in B$ with $\|v\| = 1$ and $t > 0$, let $g(t) = J(tv)$. We have

$$\begin{aligned} g'(t) = \langle \nabla J(tv), v\rangle &= \int_\Omega \left(t^{p-1}|\nabla v(x)|^p - f(x, tv(x))v(x)\right) dx \\ &= t^{p-1}\left(1 - \int_\Omega \frac{f(x,tv(x))|v(x)|^p}{|tv(x)|^{p-2}tv(x)}\right) dx. \end{aligned}$$

Thus, by condition (a), there is a unique $t_v > 0$ such that $g'(t_v) = 0$; i.e., for $L = \{0\}$ and each $v \in S_B$, the L-$\perp$ selection (actually a peak selection) $\mathcal{P}(v) = t_v v$ is uniquely determined with $J(\mathcal{P}(v)) > 0$. By taking a derivative of condition (a) w.r.t. t, we have

$$\begin{aligned} g''(t) &= (p-1)t^{(p-2)} - \int_\Omega f'_u(x, tv(x))v^2(x)\, dx \\ &< (p-1)t^{(p-2)} - \int_\Omega \frac{(p-1)}{t} f(x, tv(x))v(x)\, dx = \frac{p-1}{t} g'(t). \end{aligned} \tag{4.3}$$

Thus condition (3) in Theorem 3.4 is always satisfied for any L. Next let us recall that when $L = [u^1, u^2, \ldots, u^{n-1}]$, by the definition of an L-$\perp$ selection, $\mathcal{P}(v) = t_0 v + t_1 u^1 + \cdots + t_{n-1}u^{n-1}$ is solved from

$$\begin{aligned} &\frac{\partial}{\partial t_0} g(t_0, \ldots, t_{n-1}) = \langle \nabla J(t_0 v + t_1 u^1 + \cdots + t_{n-1}u^{n-1}), v\rangle = 0, \\ &\frac{\partial}{\partial t_i} g(t_0, \ldots, t_{n-1}) = \langle \nabla J(t_0 v + t_1 u^1 + \cdots + t_{n-1}u^{n-1}), u^i\rangle = 0, \quad i = 1, \ldots, n-1, \end{aligned} \tag{4.4}$$

where $g(t_0, \ldots, t_{n-1}) = J(t_0 v + t_1 u^1 + \cdots + t_{n-1}u^{n-1})$. If $u = \mathcal{P}(v) = t_0 v + t_1 u^1 + \cdots + t_{n-1}u^{n-1}$ satisfies (4.4) and at u, the $n \times n$ matrix

$$\begin{aligned} Q &= \left[\frac{\partial^2}{\partial t_i \partial t_j} g(t_0, \ldots, t_{n-1}) \right] \\ &= \begin{bmatrix} \langle J''(u)v, v\rangle & \langle J''(u)u^1, v\rangle & \cdots & \langle J''(u)u^{n-1}, v\rangle \\ \langle J''(u)v, u^1\rangle & \langle J''(u)u^1, u^1\rangle & \cdots & \langle J''(u)u^{n-1}, u^{n-1}\rangle \\ \cdots & & \cdots & \\ \langle J''(u)v, u^{n-1}\rangle & \langle J''(u)u^1, u^{n-1}\rangle & \cdots & \langle J''(u)u^{n-1}, u^{n-1}\rangle \end{bmatrix} \end{aligned}$$

is invertible, i.e., $|Q| \neq 0$, then by the implicit function theorem, around u, the L-$\perp$ selection $\mathcal{P}$ is well-defined and continuously differentiable. The condition $|Q| \neq 0$ can be easily and numerically checked. For the current case $L = \{0\}$, by (4.3), we have $Q = g''(t_v) < 0$. Thus the L-$\perp$ selection $\mathcal{P}$ is C^1. To show that $d(\mathcal{P}(v), L) > \alpha > 0$ for all $v \in S_B$, by (b), for any $\varepsilon > 0$, there is $c_1 = c_1(\varepsilon)$ such that $f(x, v(x))v(x) < \varepsilon|v(x)|^p + c_1|v(x)|^{p^*}$. It follows that

$$\begin{aligned}\int_\Omega f(x, v(x))v(x)\,dx &< \varepsilon \int_\Omega |v(x)|^p\,dx + c_1 \int_\Omega |v(x)|^{p^*}\,dx \\ &\qquad \text{(by the Poincaré and Sobolev inequalities)} \\ &\leq \varepsilon c_0(\Omega) \int_\Omega |\nabla v(x)|^p\,dx + c_1 c_2(\Omega) \left(\int_\Omega |\nabla v(x)|^p\,dx \right)^{\frac{p^*}{p}} \\ &= \left[\varepsilon c_0(\Omega) + c_1 c_2(\Omega) \left(\int_\Omega |\nabla v(x)|^p\,dx \right)^{\frac{p^*}{p}-1} \right] \int_\Omega |\nabla v(x)|^p\,dx.\end{aligned}$$

Thus

$$\begin{aligned}\langle \nabla J(v), v \rangle &\geq \left[1 - \varepsilon c_0(\Omega) - c_1 c_2(\Omega) \left(\int_\Omega |\nabla v(x)|^p\,dx \right)^{\frac{p^*}{p}-1} \right] \int_\Omega |\nabla v(x)|^p\,dx \\ &= \left[1 - \varepsilon c_0(\Omega) - c_1 c_2(\Omega) \|v\|^{p^*-p} \right] \|v\|^p.\end{aligned}$$

It follows that for any small $\varepsilon > 0$, c_1, $c_0(\Omega)$, and $c_2(\Omega)$, there is a $t_0 > 0$ such that when $0 < \|v\| = t < t_0$, we have $\langle \nabla J(v), v \rangle \geq \left[1 - \varepsilon c_0(\Omega) - c_1 c_2(\Omega) t^{p^*-p} \right] t^p > 0$. Therefore the L-$\perp$ selection $\mathcal{P}(v)$ satisfies $\|\mathcal{P}(v)\| > t_0$ or $d(\mathcal{P}(v), L) > t_0 > 0\ \forall v \in S_B$, where $L = \{0\}$.

Acknowledgment. The authors would like to thank two anonymous referees for their helpful suggestions.

REFERENCES

[1] Y. S. Choi and P. J. McKenna, *A mountain pass method for the numerical solution of semilinear elliptic problems*, Nonlinear Anal., 20 (1993), pp. 417–437.
[2] Z. Ding, D. Costa, and G. Chen, *A high linking method for sign changing solutions for semilinear elliptic equations*, Nonlinear Anal., 38 (1999), pp. 151–172.
[3] P. K. Kythe, *Differential Operators and Applications*, Birkhäuser, Boston, 1996.
[4] Y. Li and J. Zhou, *A minimax method for finding multiple critical points and its applications to semilinear PDEs*, SIAM J. Sci. Comput., 23 (2001), pp. 840–865.
[5] Y. Li and J. Zhou, *Convergence results of a local minimax method for finding multiple critical points*, SIAM J. Sci. Comput., 24 (2002), pp. 865–885.
[6] F. Lin and T. Lin, *Minimax solutions of the Ginzburg-Landau equations*, Selecta Math. (N.S.), 3 (1997), pp. 99–113.
[7] J. Mawhin and M. Willem, *Critical Point Theory and Hamiltonian Systems*, Springer-Verlag, New York, 1989.
[8] P. Rabinowitz, *Minimax Methods in Critical Point Theory with Applications to Differential Equations*, CBMS Regional Conf. Ser. in Math. 65, AMS, Providence, RI, 1986.
[9] M. Schechter, *Linking Methods in Critical Point Theory*, Birkhäuser, Boston, 1999.
[10] M. Struwe, *Variational Methods*, Springer-Verlag, Berlin, 1996.
[11] M. Willem, *Minimax Theorems*, Birkhäuser, Boston, 1996.

[12] X. Yao and J. Zhou, *A minimax method for finding multiple critical points in Banach spaces and its application to quasi-linear elliptic PDE*, SIAM J. Sci. Comput., 26 (2005), pp. 1796–1809.

[13] X. Yao and J. Zhou, *Numerical methods for computing nonlinear eigenpairs: Part* I. *Iso-homogeneous cases*, SIAM J. Sci. Comput., to appear.

[14] J. Zhou, *A local min-orthogonal method for finding multiple saddle points*, J. Math. Anal. Appl., 291 (2004), pp. 66–81.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1349–1369

ERROR ANALYSIS OF A CONTINUOUS-DISCONTINUOUS GALERKIN FINITE ELEMENT METHOD FOR GENERALIZED 2D VORTICITY DYNAMICS*

JAAP J. W. VAN DER VEGT†, FERENC IZSÁK‡, AND ONNO BOKHOVE†

Abstract. A detailed a priori error estimate is provided for a continuous-discontinuous Galerkin finite element method for the generalized two-dimensional vorticity dynamics equations. These equations describe several types of geophysical flows, including the Euler equations. The algorithm consists of a continuous Galerkin finite element method for the stream function and a discontinuous Galerkin finite element method for the (potential) vorticity. Since this algorithm satisfies a number of invariants, such as energy and enstrophy conservation, it is possible to provide detailed error estimates for this nonlinear problem. The main result of the analysis is a reduction in the smoothness requirements on the vorticity field from $H^2(\Omega)$, obtained in a previous analysis, to $W_p^r(\Omega)$ with $r > \frac{1}{p}$ and $p > 2$. In addition, sharper estimates for the dependence of the error on time and numerical examples on a model problem are provided.

Key words. generalized vorticity dynamics, continuous-discontinuous Galerkin finite element methods, a priori error estimates

AMS subject classifications. 65M15, 65M12, 65M60

DOI. 10.1137/050633202

1. Introduction.

The accurate numerical simulation of geophysical flows over long periods of time frequently requires the preservation of invariants of the flow field, such as energy and enstrophy conservation. These requirements are nontrivial and an active area of research in numerical analysis. A promising new technique for geophysical flows is provided by the recently developed continuous-discontinuous Galerkin (CDG) finite element method [3, 9, 10], which is capable of preserving important invariants of the flow field also at the discrete level. The accuracy of the CDG finite element method is discussed in [3, 9], but the a priori error analysis requires quite strong smoothness assumptions, which often are not realistic. The main objective of this paper is to analyze the CDG finite element method for geophysical flows under only very weak smoothness assumptions. In addition, we aim at obtaining sharper estimates for the dependence of the error on time than obtained in [3, 9]. These results will significantly extend the range of applications covered by the error estimates.

The geophysical problems studied in this paper can be described by a hyperbolic equation for the (potential) vorticity ξ and an elliptic equation for the stream function ψ. The coupled set of equations in a simply connected bounded domain $\Omega \times (t_0, T) \subset$

*Received by the editors June 7, 2005; accepted for publication (in revised form) December 13, 2006; published electronically June 21, 2007.

http://www.siam.org/journals/sinum/45-4/63320.html

†Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands (j.j.w.vandervegt@math.utwente.nl, o.bokhove@math.utwente.nl). The research of the first author was partly supported by the Dutch government through the national program BSIK: knowledge and research capacity, in the ICT project BRICKS (http://www.bsik-bricks.nl), theme MSV1. The research of the third author was supported by a fellowship from the Royal Netherlands Academy of Arts and Sciences (KNAW).

‡Department of Applied Mathematics, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands, and Department of Applied Analysis, Loránd Eötvös University, P.O. Box 120, 1518 Budapest, Hungary (izsakf@cs.elte.hu). The research of this author was partly supported by the Dutch government through the national program BSIK: knowledge and research capacity, in the ICT project BRICKS (http://www.bsik-bricks.nl), theme MSVI.

$\mathbb{R}^2 \times (t_0, T)$ is defined as

$$(1/A)\partial_t \xi + \boldsymbol{\nabla} \cdot (\xi \mathbf{U}) = 0, \tag{1.1a}$$

$$\mathbf{U} = \boldsymbol{\nabla}^{\perp} \psi, \tag{1.1b}$$

$$\boldsymbol{\nabla} \cdot (A \boldsymbol{\nabla} \psi) - B\psi + D = (1/A)\, \xi, \tag{1.1c}$$

with t representing time, where t_0 and T denote the initial and final time, respectively, and $A, B, D : \Omega \to \mathbb{R}$ given functions satisfying $0 \leq B(x)$ and $0 < A_0 \leq A(x) \leq A_1 < \infty$ for some finite $A_0, A_1 \in \mathbb{R}^+$. The gradient operator is given by $\boldsymbol{\nabla} = (\partial_x, \partial_y)^T$ and the two-dimensional curl operator by $\boldsymbol{\nabla}^{\perp} = (-\partial_y, \partial_x)^T$. At the domain boundary $\partial\Omega$, a slip flow condition is imposed at $\partial\Omega_D \subseteq \partial\Omega$,

$$\mathbf{U} \cdot \mathbf{n} = 0, \tag{1.2}$$

with $\mathbf{n} = (n_x, n_y)^T$ the outward unit vector normal to the boundary $\partial\Omega$, and periodic boundary conditions at $\partial\Omega \backslash \partial\Omega_D$. The slip flow boundary condition (1.2) is equivalent with the condition $\frac{\partial \psi}{\partial \tau} = 0$ at $\partial\Omega_D$, with $\tau = (-n_y, n_x)^T$ the tangential vector at the domain boundary. This implies that $\psi = c_D(t)$ at $\partial\Omega_D$ for a smooth ψ, with c_D a function which can depend only on time. The boundary conditions for ψ are completed by setting $c_D = 0$ when $B = 0$, almost everywhere. In case $B \neq 0$ for all $x \in \Omega_0 \subseteq \Omega$, where Ω_0 has a nonzero measure, then the circulation $\mathcal{C}$ around the boundary of the domain is imposed as an additional boundary condition, which is defined as

$$\mathcal{C} = \int_{\partial\Omega} A\mathbf{U} \cdot \tau d\Gamma = \int_{\partial\Omega} A \boldsymbol{\nabla} \psi \cdot \mathbf{n} d\Gamma. \tag{1.3}$$

Initial conditions are provided by specifying the initial vorticity field ξ_0. The generalized system (1.1) serves as a model for several fluid flow problems by choosing A, B, and D to yield the incompressible two-dimensional (2D) Euler equations [7], the quasi-geostrophic equations [12], and the rigid-lid equations [8], often used in atmosphere and ocean dynamics. An overview of the specific values of the coefficients A, B, D for these different problems can be found in [3]. In all these cases, ξ represents the (potential) vorticity of the fluid, $\mathbf{u} = A\mathbf{U}$ the velocity, and $\mathbf{U}$ the (mass transport) velocity of the fluid.

The CDG finite element method was introduced in [9, 10] for the 2D Euler equations and extended in [3] to the generalized 2D vorticity dynamics equations (1.1) in multiple connected domains. Apart from detailed numerical experiments, an a priori error analysis was also given in [3] for the CDG finite element discretization of (1.1) under slightly less restrictive assumptions than the analysis for the 2D Euler equations discussed in [9]. The analysis requires, however, that the vorticity field belongs at least to $H^2(\Omega)$, which is frequently not valid. Also, the analyses in [3, 9], which both use Gronwall's inequality, provide only an exponentially growing upper bound for the error in time.

In order to alleviate these restrictions, we provide in this paper an error analysis which imposes only weak smoothness requirements on the (generalized) vorticity field. This analysis extends the work in [10], which proves convergence of the CDG finite element algorithm for the 2D Euler equations when the initial vorticity field is in $L^2(\Omega)$, in two ways. In the first place, we consider the generalized 2D vorticity dynamics equations (1.1) instead of the 2D Euler equations, and in the second place, we provide detailed a priori error estimates in terms of the mesh size and initial vorticity. This result shows that the CDG finite element method in [3] is applicable

to a wide range of geophysical problems, with only weak smoothness requirements on the vorticity.

The outline of the paper is as follows. After some preliminaries in section 2, we briefly state in section 3 the CDG finite element method for (1.1). Next, we recall in section 4 some important properties of the CDG finite element method regarding energy conservation and enstrophy stability. These properties are essential in order to prove the main result of this paper, an a priori error estimate for the CDG finite element method with limited smoothness requirements, which we discuss in section 5. This analysis also requires an upper bound on the stream function and its trace at the boundary, which we provide in a separate lemma. In section 6, we discuss the evolution of a concentrated patch of vorticity as an example of a model problem in geophysical flows where the vorticity field has limited smoothness. Finally, in Appendix A, we analyze the smoothness of this vorticity field.

2. Preliminaries.

2.1. Function spaces. We denote with Ω a simply connected bounded domain $\Omega \subset \mathbb{R}^2$ with Lipschitz boundary $\partial\Omega$. For any nonnegative integer k, $C^k(\Omega)$ denotes the space of all functions w which, together with all their partial derivatives $D^\alpha w$ of order $|\alpha| \leq k$, are continuous in Ω. For $0 < \lambda \leq 1$, we define $C^{k,\lambda}(\overline{\Omega})$ to be the subspace of $C^k(\overline{\Omega})$, with $\overline{\Omega}$ the closure of Ω, consisting of those functions w for which for $0 \leq |\alpha| \leq k$, $D^\alpha w$ satisfies in Ω a Hölder condition of exponent λ. That is, there exists a constant C, such that

$$|D^\alpha w(x) - D^\alpha w(y)| \leq C|x-y|^\lambda \qquad \forall x, y \in \Omega.$$

The Lebesgue measure is also denoted with $|\cdot|$, while for the Lebesgue spaces we use the notation $L^p(\Omega)$, for $1 \leq p \leq \infty$. These spaces are equipped with the norm $\|w\|_{p,\Omega} = (\int_\Omega |w|^p d\Omega)^{\frac{1}{p}}$ for $1 \leq p < \infty$ and $\|w\|_{\infty,\Omega} = \operatorname{ess\,sup}_{x\in\Omega}|w(x)|$. In addition, we define the Sobolev spaces $W_p^s(\Omega)$, with the norm indicated as $\|w\|_{s,p,\Omega}$. For s integer, $s \geq 0$, and $1 \leq p < \infty$, the Sobolev norm is defined as $\|w\|_{s,p,\Omega} = (\sum_{|\alpha|\leq s} \|D^\alpha w\|_{p,\Omega}^p)^{\frac{1}{p}}$ and the seminorm as $|w|_{s,p,\Omega} = (\sum_{|\alpha|=s} \|D^\alpha w\|_{p,\Omega}^p)^{\frac{1}{p}}$, whereas for s integer, $s \geq 0$, and $p = \infty$, we have $\|w\|_{s,\infty,\Omega} = \max_{|\alpha|\leq s} \|D^\alpha w\|_{\infty,\Omega}$, with the usual modification for the seminorm. For noninteger values $s > 0$, we use Banach space interpolation to define the fractional order Sobolev spaces $W_p^s(\Omega)$. For a detailed discussion, we refer the reader to [4, Chapter 14.2]. For $s = 0$, we have $W_p^0(\Omega) = L^p(\Omega)$. The $L^2(\Omega)$ and $(L^2(\Omega))^2$ spaces are equipped with the inner product $(u,v)_\Omega = \int_\Omega uv\, d\Omega$ and $(u,v)_\Omega = \int_\Omega u \cdot v\, d\Omega$, respectively, and we use the shorthand notation $\|w\|_\Omega$ for the $L^2(\Omega)$ norm $\|w\|_{0,2,\Omega}$. We also use the notation $H^s(\Omega)$ for $W_2^s(\Omega)$, with $s \in \mathbb{R}$. For $1 < p < \infty$ and $s \in \mathbb{R}$, $s \geq 0$, the Sobolev spaces with a negative index $W_p^{-s}(\Omega)$ are defined as the dual spaces of $W_p^s(\Omega)$, see [1, p. 65], equipped with the norm $\|w\|_{-s,p,\Omega} = \sup_{0\neq v\in W_{p'}^s(\Omega)} \frac{(w,v)_\Omega}{\|v\|_{s,p',\Omega}}$, where $1/p + 1/p' = 1$, and $(w,v)_\Omega$ denotes the duality pairing between w and v with $L^2(\Omega)$ as pivot space. All the above definitions also apply with the domain Ω replaced by the element K or its boundary ∂K. Finally, we define the broken Sobolev spaces $W_p^s(\mathcal{T}_h)$ for all $s \in \mathbb{R}$, with $\mathcal{T}_h$ a tessellation covering Ω, as the space of functions such that their restriction to each $K \in \mathcal{T}_h$ belongs to $W_p^s(K)$.

2.2. Interpolation operators. In the error analysis, we assume a quasi-uniform mesh (see [4, section 4.4]) with tessellation $\mathcal{T}_h$. Since the vorticity field is not necessarily continuous, we use the more general Clément-type interpolation operator defined

in [2]. Let $\mathcal{T}_h$ be a regular partition of Ω into triangles or quadrilaterals; then the macroelement $\tilde{K}$, associated with element $K \in \mathcal{T}_h$, consists of those elements which share at least one vertex with K, and hence $\tilde{K} = \text{int}\left\{\cup K', K' \in \mathcal{T}_h \,|\, \overline{K'} \cap \overline{K} \neq \emptyset\right\}$, where the overbar denotes the closure of a set and int the interior. The n elements $K \in \mathcal{T}_h$ constituting the macroelement $\tilde{K}$ are denoted $\tilde{K}_i$, with $1 \leq i \leq n$. With each element $\tilde{K}_i$ we associate a reference element $\widehat{\tilde{K}}_i$ using the continuous and invertible mapping $F_{\tilde{K}_i} : \widehat{\tilde{K}}_i \rightarrow \tilde{K}_i$. We denote with $\widehat{\tilde{K}} = \cup_{i=1}^n \widehat{\tilde{K}}_i$ the reference macroelement associated with $\tilde{K}$. Both elements are connected with the mapping $F_{\tilde{K}} : \widehat{\tilde{K}} \rightarrow \tilde{K}$, which consists of the individual mappings $F_{\tilde{K}_i}$, with $1 \leq i \leq n$. Next, we define the local finite element spaces on the macroelement $\tilde{K}$ and its reference macroelement $\widehat{\tilde{K}}$ as follows:

$$\Theta_h^k\left(\widehat{\tilde{K}}\right) = \left\{\hat{v} \in C^0\left(\widehat{\tilde{K}}\right) \,|\, \forall \widehat{\tilde{K}}_i \subset \widehat{\tilde{K}},\, \hat{v}_{|\widehat{\tilde{K}}_i} \in \mathcal{P}_k\right\},$$
$$\Theta_h^k(\tilde{K}) = \left\{v \in C^0(\tilde{K}) \,|\, \forall \tilde{K}_i \subset \tilde{K},\, v_{|\tilde{K}_i} = \hat{v}_{|\widehat{\tilde{K}}_i} \circ F_{\tilde{K}_i}^{-1},\, \hat{v}_{|\widehat{\tilde{K}}_i} \in \mathcal{P}_k\right\},$$

with $\mathcal{P}_k$ polynomials of total degree less than or equal to k on triangles and of degree less than or equal to k in each coordinate direction on quadrilaterals. For any function $\hat{u}$ in $L^1(\widehat{\tilde{K}})$, we define the projection $\widehat{\mathcal{P}}_{\widehat{\tilde{K}}}\hat{u}$ in $\Theta_h^k(\widehat{\tilde{K}})$ by

$$\int_{\widehat{\tilde{K}}} \left(\widehat{\mathcal{P}}_{\widehat{\tilde{K}}}\hat{u} - \hat{u}\right)\hat{v}\, d\hat{x} = 0 \qquad \forall \hat{v} \in \Theta_h^k\left(\widehat{\tilde{K}}\right), \tag{2.1}$$

and for any function u in $L^1(\tilde{K})$, we define the projection $\mathcal{P}_{\tilde{K}}u$ on $\Theta_h^k(\tilde{K})$ by

$$\mathcal{P}_{\tilde{K}}u \circ F_{\tilde{K}} = \widehat{\mathcal{P}}_{\widehat{\tilde{K}}}(u \circ F_{\tilde{K}}). \tag{2.2}$$

The following lemma on the interpolation error is proved in Theorem 2.2 and Remark 4 in [2] for triangles and Theorem 3.5 and Remark 5 in [2] for quadrilaterals.

LEMMA 2.1. *For any integers $k \geq 1$ and any real number t, with $0 \leq t \leq 1$, provided the function u belongs to $W_p^s(\tilde{K})$, there exists a constant C, depending only on the regularity of $\mathcal{T}_h$, such that for any element $\tilde{K}_i \subset \tilde{K}$*

$$|u - \mathcal{P}_{\tilde{K}}u|_{t,p,\tilde{K}_i} \leq C h_K^{s-t} |u|_{s,p,\tilde{K}}, \qquad \text{with } t \leq s \leq k+1,\ 1 \leq p \leq \infty, \tag{2.3}$$

and

$$|u - \mathcal{P}_{\tilde{K}}u|_{t,p,\partial\tilde{K}_i} \leq C h_K^{s-t-\frac{1}{p}} |u|_{s,p,\tilde{K}}, \quad \text{with } t + \frac{1}{p} < s \leq k+1,\ 1 \leq p < \infty, \tag{2.4}$$

with h_K the diameter of elements $K \in \mathcal{T}_h$.

3. Finite element method. The potential vorticity equation (1.1a) and the generalized stream function equation (1.1c) are discretized with a discontinuous and a continuous Galerkin finite element method, respectively. The key benefit of this approach is that we can ensure under certain conditions that the discretization satisfies important constraints, such as energy conservation and enstrophy stability, which are summarized in section 4. In this section, we summarize the CDG finite element method for the system (1.1). More details can be found in [3].

3.1. Continuous Galerkin discretization for generalized stream function. The definition of the weak formulation for (1.1c) requires the treatment of the nonstandard boundary conditions for ψ given by (1.2)–(1.3), which are a mixture of essential and natural boundary conditions. The slip flow boundary condition (1.2) is equivalent with the condition $\frac{\partial \psi}{\partial \tau} = 0$ at $\partial\Omega_D$, which implies that $\gamma(\psi)|_{\partial\Omega_D} = c_D(t)$, with $\gamma : H^1(\Omega) \to H^{\frac{1}{2}}(\partial\Omega)$ the trace operator. The value of c_D, however, still depends on time and is determined implicitly by the circulation condition (1.3), which arises as a natural boundary condition in the weak formulation (3.3). Introduce now the space

$$H^1_{0,D}(\Omega) := \left\{ w \in H^1(\Omega) \,\middle|\, \gamma(w)|_{\partial\Omega_D} = 0 \right\}, \tag{3.1}$$

and split ψ into

$$\psi = \psi_0 + c_D \cdot \mathbf{1}, \tag{3.2}$$

where $\mathbf{1} : \Omega \to 1$ is the constant function. The weak formulation for (1.1c) at time t can then be straightforwardly derived and is equal to the following:

Find a $\psi_0 \in H^1_{0,D}(\Omega)$ such that for all $w_0 \in H^1_{0,D}(\Omega)$ the following relation is satisfied:

$$L(\psi_0, w_0) = F_\xi(w_0) - L(c_D, w_0), \tag{3.3}$$

with the operators $L : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ and $F_\xi : H^{-1}(\Omega) \to \mathbb{R}$ defined as

$$L(\psi, w) := \left(\sqrt{A}\boldsymbol{\nabla}\psi, \sqrt{A}\boldsymbol{\nabla}w\right)_\Omega + \left(\sqrt{B}\psi, \sqrt{B}w\right)_\Omega, \tag{3.4}$$

$$F_\xi(w) := -\left(\xi/A, w\right)_\Omega + (D, w)_\Omega + \gamma(w)\mathcal{C}, \tag{3.5}$$

and $\mathcal{C}$ given by (1.3). Choosing $w = \mathbf{1}$ in (3.3), we obtain that

$$c_D = \frac{F_\xi(\mathbf{1}) - (\psi_0\sqrt{B}, \sqrt{B})_\Omega}{\|\sqrt{B}\|^2_\Omega}. \tag{3.6}$$

For the finite element discretization, we introduce the spaces

$$X^k_h := \left\{ w \in C^0(\Omega) \mid w|_K \in \mathcal{P}_k(K) \, \forall K \in \mathcal{T}_h \right\}$$

and $\mathsf{W}^k_h = H^1_{0,D}(\Omega) \cap X^k_h$. By restricting ψ and w to the finite-dimensional space W^k_h, we obtain the following continuous finite element formulation:

Find a $\psi_h \in \mathsf{W}^k_h$ such that for all $w \in \mathsf{W}^k_h$ the following relation is satisfied:

$$L(\psi_h, w) = F_{\xi_h}(w) - L(c_D, w), \tag{3.7}$$

with ξ_h the discrete vorticity field computed with the discontinuous Galerkin finite element discretization specified in the next section.

3.2. Discontinuous Galerkin space discretization. The weak formulation (in space) for the (potential) vorticity equation (1.1a), with U replaced by (1.1b), can be defined straightforwardly as follows:

Find a $\xi \in L^1((t_0, T), W^r_p(\Omega))$, with $r > \frac{1}{p}$ and $p > 2$, such that for all $v \in H^1(\Omega)$ the following relation is satisfied:

$$\frac{d}{dt}\left(\frac{\xi}{A}, v\right)_\Omega = \left(\xi\boldsymbol{\nabla}^\perp\psi, \boldsymbol{\nabla}v\right)_\Omega - \left(\xi\boldsymbol{\nabla}^\perp\psi \cdot \mathbf{n}, v\right)_{\partial\Omega}, \tag{3.8}$$

where the boundary contribution should be interpreted in the sense of traces and ψ is given by (3.2). For the discontinuous Galerkin (DG) discretization, we define the following space of discontinuous functions:

$$\mathsf{V}_h^k := \left\{ v_h \in L^2(\Omega) \,\middle|\, v_h|_K \in \mathcal{P}_k(K) \quad \forall K \in \mathcal{T}_h \right\}. \tag{3.9}$$

Note that conservation of energy of the numerical solution requires $\mathsf{W}_h^k \subset \mathsf{V}_h^k$ (see [3]). The DG weak formulation is now equal to the following:

Find a $\xi_h \in \mathsf{V}_h^k$ such that for all $v \in \mathsf{V}_h^k$ the following relation holds:

$$\sum_{K\in\mathcal{T}_h} \frac{d}{dt}\Big(\frac{\xi_h}{A}, v\Big)_K = \sum_{K\in\mathcal{T}_h} R_K(\xi_h, \psi_h, v), \tag{3.10}$$

with the operator $R_K : \mathsf{V}_h^k \times \mathsf{W}_h^k \times \mathsf{V}_h^k \to \mathbb{R}$ defined by

$$R_K(\xi_h, \psi_h, v) = \Big(\xi_h \boldsymbol{\nabla}^\perp \psi_h, \boldsymbol{\nabla} v\Big)_K - \int_{\partial K} v^- \hat{f}(\xi_h^+, \xi_h^-, \boldsymbol{\nabla}^\perp \psi_h \cdot \mathbf{n})\, d\Gamma, \tag{3.11}$$

where the superscripts $^-$ and $^+$ denote the trace values at the boundary point ∂K taken from the inside and outside of the element, respectively. Here, $\hat{f}$ denotes the numerical flux which is necessary to account for the discontinuity in the DG basis functions at the element boundaries. The numerical flux is defined as

$$\text{central} \quad \hat{f}(\xi^+, \xi^-, U_n) = \frac{\xi^+ + \xi^-}{2} U_n, \tag{3.12a}$$

$$\text{upwind} \quad \hat{f}(\xi^+, \xi^-, U_n) = U_n \begin{cases} \xi^+ & \text{if } U_n < 0, \\ \xi^- & \text{if } U_n \geq 0, \end{cases} \tag{3.12b}$$

$$\text{Lax–Friedrichs} \quad \hat{f}(\xi^+, \xi^-, U_n) = \frac{1}{2}\left(U_n(\xi^+ + \xi^-) - \alpha_{LF}(\xi^+ - \xi^-)\right), \tag{3.12c}$$

with $U_n = \mathbf{U}_h \cdot \mathbf{n} = \boldsymbol{\nabla}^\perp \psi_h \cdot \mathbf{n}$ and $\alpha_{LF} \geq 0$. For ease of notation, we also write the numerical flux as

$$\hat{\xi}_h = \hat{f}(\xi^+, \xi^-, U_n)/U_n.$$

A common choice for α_{LF} is $\alpha_{LF} = \max|U_n|$ with a local or global maximum. For $\alpha_{LF} = 0$ and $\alpha_{LF} = |U_n|$, we obtain the central and upwind flux, respectively.

4. Conservation of energy and enstrophy. The equations for generalized 2D vorticity dynamics (1.1) and the numerical discretization given by (3.7) and (3.10) satisfy a number of important invariants and constraints, including energy and enstrophy conservation. These invariants are essential for the error analysis discussed in section 5, and we summarize them here for completeness. Define for $A(x, y) > 0$ and $B(x, y) \geq 0$ the total energy E and enstrophy S of the flow field as

$$E(\psi, t) = \frac{1}{2}\|\sqrt{A}\boldsymbol{\nabla}\psi(\cdot, t)\|_\Omega^2 + \frac{1}{2}\|\sqrt{B}\psi(\cdot, t)\|_\Omega^2, \tag{4.1}$$

$$S(\xi, t) = \frac{1}{2}\left\|\frac{\xi(\cdot, t)}{\sqrt{A}}\right\|_\Omega^2. \tag{4.2}$$

The system of equations for generalized 2D vorticity dynamics (1.1) satisfies the following invariants.

LEMMA 4.1. *Assume that $A, B \in L^\infty(\Omega)$, with $A > 0$ and $B \geq 0$; then the energy E and enstrophy S of system* (1.1) *subject to slip flow boundary conditions* (1.2) *and constant circulation* (1.3) *is conserved:*

$$\frac{dE(\psi,t)}{dt} = 0, \qquad \frac{dS(\xi,t)}{dt} = 0. \tag{4.3}$$

For a proof, see, e.g., [3]. The numerical solution of the generalized 2D vorticity dynamics equations (1.1) obtained with the CDG method (partly) satisfies the same invariants.

LEMMA 4.2. *Consider the solution of* (3.7) *and* (3.10) *subject to slip flow boundary conditions* (1.2). *The energy E associated with this numerical solution is a conserved quantity, and the enstrophy S is bounded:*

$$\frac{dE(\psi_h,t)}{dt} = 0, \qquad \frac{dS(\xi_h,t)}{dt} \leq 0. \tag{4.4}$$

For the central flux (3.12a), *the relation for the enstrophy becomes an equality.*

For a proof, see [3].

5. Error analysis. In this section, we will prove an a priori error estimate for the CDG finite element discretization (3.7) and (3.10) for the generalized 2D vorticity dynamics equations (1.1). The key objective is to minimize the requirements imposed on the smoothness of the vorticity field ξ in order to extend the validity of the analysis and the numerical method to a much larger range of physically relevant problems. The error estimate provides a significant extension of the error estimate Theorem 11 in [3], which required the vorticity field ξ to belong to $H^2(\Omega)$ and provided only an exponentially growing upper bound for the growth of the error in time.

The main result of this paper is provided by the following error estimate for the velocity field $\mathbf{u}_h$ and the vorticity field ξ_h computed with the CDG method.

THEOREM 5.1. *Assume that Ω is a polyhedral simply connected bounded domain and that the vorticity field satisfies the condition $\xi \in L^1((t_0,T), W_p^r(\Omega))$, with $r \in \mathbb{R}$, $r > \frac{1}{p}$, and $p > 2$. In addition, we assume that the coefficients in* (1.1c) *satisfy $A, B \in C^{r,1}(\overline{\Omega})$, $D \in L^\infty(\Omega)$, with $0 \leq B(x)$ and $0 < A_0 \leq A(x) \leq A_1 < \infty$ for some finite $A_0, A_1 \in \mathbb{R}^+$; then the error in the CDG finite element discretization* (3.7) *and* (3.10) *on a quasi-uniform mesh $\mathcal{T}_h$, with $h < 1$, can be estimated as*

$$\|(\mathbf{u} - \mathbf{u}_h)/A\|_{0,q,\Omega} + \|(\xi - \xi_h)/A\|_{-1,q',\mathcal{T}_h} \leq Ch^s \|\xi_0\|_\Omega \int_{t_0}^{T} \|\xi(\cdot,t)\|_{r,p,\Omega} dt,$$

where ξ_0 denotes the initial vorticity field, $q = \frac{2p}{p-2}$, $q' = \frac{q}{q-1}$, $\frac{1}{p} < s \leq \min(k-1, r + \frac{2}{p} - \epsilon_0)$, with k the order of the polynomial basis functions, $\epsilon_0 > 0$ an arbitrary small constant, and C a positive constant, independent of h, $\mathbf{u}$, and ξ.

The proof of Theorem 5.1 is split into several parts. Since the weak formulation (3.3) has nonstandard boundary conditions, we first need to show that it has a unique solution. Moreover, we need an upper bound for the stream function ψ and $c_D(t)$, the trace of ψ at $\partial\Omega_D$, for the regularity estimate in the second part of the proof.

LEMMA 5.2. *The variational problem* (3.3) *is well posed in $H^1_{0,D}(\Omega) \oplus \mathbb{R}$, and the*

$H^1(\Omega)$ norm of the stream function can be estimated as

$$\|\psi(\cdot,t)\|_{1,2,\Omega} \leq \left(\gamma + \frac{\sqrt{|\Omega|}}{\|B\|_{1,\Omega}}\left(\sqrt{|\Omega|} + \gamma\|B\|_\Omega\right)\right)\left(\frac{1}{\sqrt{A_0}}\left\|\frac{\xi_0}{\sqrt{A}}\right\|_\Omega + \|D\|_\Omega\right)$$
$$+ \frac{1}{\|B\|_{1,\Omega}}\left(\sqrt{|\Omega|} + \gamma\|B\|_\Omega\right)|\mathcal{C}|, \tag{5.1}$$

with ξ_0 the initial vorticity, $\gamma = \frac{1}{\alpha}(1 + \frac{\sqrt{|\Omega|}\|B\|_\Omega}{\|B\|_{1,\Omega}})$, and $\alpha = \frac{1}{2}\min(A_0/C_p, A_0)$, with C_p the Poincaré constant of Ω. For the case $\|B\|_{1,\Omega} = 0$, this can be simplified into

$$\|\psi(\cdot,t)\|_{1,2,\Omega} \leq \frac{1}{\alpha}\left(\frac{1}{\sqrt{A_0}}\|\xi_0/\sqrt{A}\|_\Omega + \|D\|_\Omega\right). \tag{5.2}$$

Proof. First, assume $\|B\|_{1,\Omega} > 0$. If we introduce the relation for c_D (3.6) into (3.3), then we obtain the following weak formulation:

Find a $\psi_0 \in H^1_{0,D}(\Omega)$ such that for all $w_0 \in H^1_{0,D}(\Omega)$ the following relation is satisfied:

$$\tilde{L}(\psi_0, w_0) = \tilde{F}_\xi(w_0), \tag{5.3}$$

with the operators $\tilde{L} : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ and $\tilde{F}_\xi : H^{-1}(\Omega) \to \mathbb{R}$ defined as

$$\tilde{L}(\psi_0, w_0) := (\sqrt{A}\boldsymbol{\nabla}\psi_0, \sqrt{A}\boldsymbol{\nabla}w_0)_\Omega + (\sqrt{B}\psi_0, \sqrt{B}w_0)_\Omega$$
$$- \frac{1}{\|\sqrt{B}\|^2_\Omega}(\sqrt{B}\psi_0, \sqrt{B})_\Omega(\sqrt{B}w_0, \sqrt{B})_\Omega, \tag{5.4}$$

$$\tilde{F}_\xi(w_0) := F_\xi(w_0) - \frac{1}{\|\sqrt{B}\|^2_\Omega}(\sqrt{B}F_\xi(\mathbf{1}), \sqrt{B}w_0)_\Omega. \tag{5.5}$$

Next, we consider the coercivity of $\tilde{L}$ in $H^1_{0,D}(\Omega)$. Taking $w_0 = \psi_0$ in (5.4) and using the Schwarz inequality $(\sqrt{B}\psi_0, \sqrt{B})_\Omega \leq \|\sqrt{B}\psi_0\|_\Omega\|\sqrt{B}\|_\Omega$, we obtain for all $\psi_0 \in H^1_{0,D}(\Omega)$

$$\tilde{L}(\psi_0,\psi_0) = (\sqrt{A}\boldsymbol{\nabla}\psi_0, \sqrt{A}\boldsymbol{\nabla}\psi_0)_\Omega + (\sqrt{B}\psi_0, \sqrt{B}\psi_0)_\Omega - \frac{1}{\|\sqrt{B}\|^2_\Omega}(\psi_0\sqrt{B}, \sqrt{B})^2_\Omega$$
$$\geq A_0\|\boldsymbol{\nabla}\psi_0\|^2_\Omega.$$

Finally, using the Poincaré inequality $\|\psi_0\|^2_\Omega \leq C_p\|\boldsymbol{\nabla}\psi_0\|^2_\Omega$ for all $\psi_0 \in H^1_{0,D}(\Omega)$, we obtain the coercivity estimate:

$$\tilde{L}(\psi_0,\psi_0) \geq \alpha\|\psi_0\|^2_{1,2,\Omega} \qquad \forall\psi_0 \in H^1_{0,D}(\Omega),$$

with $\alpha = \frac{1}{2}\min(\frac{A_0}{C_p}, A_0)$. The boundedness of $\tilde{L}$ in $H^1_{0,D}(\Omega)$ follows from a straightforward application of the Schwarz inequality:

$$\tilde{L}(\psi_0, w_0) \leq (A_1 + 2\|\sqrt{B}\|^2_{\infty,\Omega})\|\psi_0\|_{1,2,\Omega}\|w_0\|_{1,2,\Omega} \qquad \forall\psi_0, w_0 \in H^1_{0,D}(\Omega).$$

Since $\tilde{L}$ is coercive and bounded on $H^1_{0,D}(\Omega)$, the Lax–Milgram theorem (see, e.g., [4]) states that (5.3) has a unique solution $\psi_0 \in H^1_{0,D}(\Omega)$, provided $\tilde{F}_\xi \in H^{-1}(\Omega)$ and we have the following upper bound for ψ_0:

$$\|\psi_0\|_{1,2,\Omega} \leq \frac{1}{\alpha}\|\tilde{F}_\xi\|_{-1,2,\Omega} \qquad \forall\psi_0 \in H^1_{0,D}(\Omega). \tag{5.6}$$

The right-hand side of (5.6) can be estimated directly using (5.5) and the definition of the dual norm

$$\|\tilde{F}_\xi\|_{-1,2,\Omega} \leq \frac{1}{\sqrt{A_0}} \left\| \frac{\xi}{\sqrt{A}} \right\|_\Omega + \|D\|_\Omega + \frac{\|B\|_\Omega |F_\xi(\mathbf{1})|}{\|B\|_{1,\Omega}}, \tag{5.7}$$

where we used the relations $\|\sqrt{B}\|_\Omega^2 = \|B\|_{1,\Omega}$ and $\|\xi/A\|_\Omega \leq \frac{1}{\sqrt{A_0}} \|\xi/\sqrt{A}\|_\Omega$. The contribution $F_\xi(\mathbf{1})$ can be estimated from (3.5):

$$F_\xi(\mathbf{1}) \leq \sqrt{|\Omega|} \left(\frac{1}{\sqrt{A_0}} \left\| \frac{\xi}{\sqrt{A}} \right\|_\Omega + \|D\|_\Omega \right) + |\mathcal{C}|. \tag{5.8}$$

Combining (5.6)–(5.8) yields the following upper bound for ψ_0:

$$\|\psi_0\|_{1,2,\Omega} \leq \gamma \left(\frac{1}{\sqrt{A_0}} \left\| \frac{\xi}{\sqrt{A}} \right\|_\Omega + \|D\|_\Omega \right) + \frac{\|B\|_\Omega}{\alpha \|B\|_{1,\Omega}} |\mathcal{C}|. \tag{5.9}$$

The coefficient c_D in (3.6) can be estimated using (5.8)–(5.9), yielding

$$\begin{aligned} \|c_D \cdot \mathbf{1}\|_\Omega \leq & \frac{\sqrt{|\Omega|}}{\|B\|_{1,\Omega}} \left(\sqrt{|\Omega|} + \gamma \|B\|_\Omega \right) \left(\frac{1}{\sqrt{A_0}} \left\| \frac{\xi}{\sqrt{A}} \right\|_\Omega + \|D\|_\Omega \right) \\ & + \frac{\sqrt{|\Omega|}}{\|B\|_{1,\Omega}} \left(1 + \frac{\|B\|_\Omega^2}{\alpha \|B\|_{1,\Omega}} \right) |\mathcal{C}|. \end{aligned} \tag{5.10}$$

Combining the estimates for ψ_0 and c_D, given by (5.9)–(5.10), respectively, and using the triangle inequality and the enstrophy conservation (4.3) gives after some algebraic manipulations the estimate (5.1) for the stream function in the $H^1(\Omega)$ norm.

Finally, for the case $\|B\|_{1,\Omega} = 0$, a straightforward calculation gives (5.2). ∎

As a direct consequence of (5.10), we have the following corollary.

COROLLARY 5.3. *The trace $\gamma(\psi)|_{\partial\Omega_D} = c_D(t)$ of $\psi = \psi_0 + c_D$, with ψ_0 satisfying* (3.3)*, is bounded by*

$$\begin{aligned} \|c_D\|_{\infty,(t_0,T)} \leq & \frac{1}{\|B\|_{1,\Omega}} \left(\sqrt{|\Omega|} + \gamma \|B\|_\Omega \right) \left(\frac{1}{\sqrt{A_0}} \left\| \frac{\xi_0}{\sqrt{A}} \right\|_\Omega + \|D\|_\Omega \right) \\ & + \frac{1}{\|B\|_{1,\Omega}} \left(1 + \frac{\|B\|_\Omega^2}{\alpha \|B\|_{1,\Omega}} \right) |\mathcal{C}|. \end{aligned} \tag{5.11}$$

For the error analysis, we now introduce the errors $\epsilon = \xi - \xi_h$ and $\delta = \psi - \psi_h$ in the vorticity field and stream function, respectively, and their $\mathcal{P}_{\tilde{K}}$ projection

$$\epsilon_h = \mathcal{P}_{\tilde{K}} \epsilon = \mathcal{P}_{\tilde{K}} \xi - \xi_h, \qquad \delta_h = \mathcal{P}_{\tilde{K}} \delta = \mathcal{P}_{\tilde{K}} \psi - \psi_h. \tag{5.12}$$

Proof of Theorem 5.1. We start with an analysis of the error in the DG finite element discretization for the vorticity equation, where we initially assume that $\psi \in C^0(\Omega) \cap H^1(\Omega)$, which will be confirmed in the second part of the proof by a regularity estimate. This also implies that $\boldsymbol{\nabla}^\perp \psi \cdot \mathbf{n} = -\frac{\partial \psi}{\partial \tau}$, with τ the tangential vector at ∂K, has the same trace when taking the limit either from the inside or the outside of element K.

First, we consider the error at a fixed time t. Subtracting the discretized weak formulation (3.10) from (3.8), with Ω restricted to K and v to V_h^k, yields the error

equation

$$\begin{aligned}\frac{d}{dt}\Big(\frac{\epsilon}{A},v\Big)_K &= \Big(\xi\boldsymbol{\nabla}^\perp\psi-\xi_h\boldsymbol{\nabla}^\perp\psi_h,\boldsymbol{\nabla}v\Big)_K\\ &\quad-\Big(\xi^-\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n}-\hat{\xi}_h\boldsymbol{\nabla}^\perp\psi_h\cdot\mathbf{n},v^-\Big)_{\partial K}\qquad\forall v\in\mathsf{V}_h^k,\forall K\in\mathcal{T}_h.\end{aligned}\tag{5.13}$$

Adding and subtracting the projections $\mathcal{P}_{\tilde{K}}\psi$ and $\mathcal{P}_{\tilde{K}}\xi$ yields the following relation for the error in the vorticity field:

$$\begin{aligned}\frac{d}{dt}\Big(\frac{\epsilon}{A},v\Big)_K =&\big((\xi-\mathcal{P}_{\tilde{K}}\xi)\boldsymbol{\nabla}^\perp\psi+(\mathcal{P}_{\tilde{K}}\xi-\xi_h)\boldsymbol{\nabla}^\perp\psi\\ &+\ \xi_h\boldsymbol{\nabla}^\perp(\psi-\mathcal{P}_{\tilde{K}}\psi)+\xi_h\boldsymbol{\nabla}^\perp(\mathcal{P}_{\tilde{K}}\psi-\psi_h),\boldsymbol{\nabla}v\big)_K\\ &-\big((\xi^--\mathcal{P}_{\tilde{K}}\xi)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n}+(\mathcal{P}_{\tilde{K}}\xi-\hat{\xi}_h)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n}\\ &+\ \hat{\xi}_h\boldsymbol{\nabla}^\perp(\psi-\mathcal{P}_{\tilde{K}}\psi)\cdot\mathbf{n}+\hat{\xi}_h\boldsymbol{\nabla}^\perp(\mathcal{P}_{\tilde{K}}\psi-\psi_h)\cdot\mathbf{n},v^-\big)_{\partial K}.\end{aligned}\tag{5.14}$$

Note that, since the projection $\mathcal{P}_{\tilde{K}}\xi$ is continuous on the macroelement $\tilde{K}$ and $K\subset\mathrm{int}(\tilde{K})$, we have $\mathcal{P}_{\tilde{K}}\xi^-=\mathcal{P}_{\tilde{K}}\xi^+=\mathcal{P}_{\tilde{K}}\xi$ at ∂K. Introducing the definitions for ϵ_h and δ_h given by (5.12) results in the error equation

$$\begin{aligned}\frac{d}{dt}\Big(\frac{\epsilon}{A},v\Big)_K =&\big((\xi-\mathcal{P}_{\tilde{K}}\xi)\boldsymbol{\nabla}^\perp\psi+\epsilon_h\boldsymbol{\nabla}^\perp\psi+\xi_h\boldsymbol{\nabla}^\perp(\psi-\mathcal{P}_{\tilde{K}}\psi)\\ &+\xi_h\boldsymbol{\nabla}^\perp\delta_h,\boldsymbol{\nabla}v\big)_K-\big((\xi^--\mathcal{P}_{\tilde{K}}\xi)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n}\\ &+(\mathcal{P}_{\tilde{K}}\xi-\hat{\xi}_h)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n}+\hat{\xi}_h\boldsymbol{\nabla}^\perp(\psi-\mathcal{P}_{\tilde{K}}\psi)\cdot\mathbf{n}\\ &+\hat{\xi}_h\boldsymbol{\nabla}^\perp\delta_h\cdot\mathbf{n},v^-\big)_{\partial K}.\end{aligned}\tag{5.15}$$

We can simplify (5.15) for the central flux $\hat{\xi}_h=\frac{1}{2}(\xi_h^-+\xi_h^+)$ using the following relations:

$$\begin{aligned}&(\epsilon_h\boldsymbol{\nabla}^\perp\psi,\boldsymbol{\nabla}v)_K-((\mathcal{P}_{\tilde{K}}\xi-\hat{\xi}_h)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n},v^-)_{\partial K}\\ &\quad=-(\boldsymbol{\nabla}\cdot(\epsilon_h\boldsymbol{\nabla}^\perp\psi),v)_K+(\epsilon_h^-\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n},v^-)_{\partial K}\\ &\qquad-((\epsilon_h^-+\xi_h^--\hat{\xi}_h)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n},v^-)_{\partial K}\\ &\quad=-(\boldsymbol{\nabla}\epsilon_h\cdot\boldsymbol{\nabla}^\perp\psi,v)_K-\frac{1}{2}((\xi_h^--\xi_h^+)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n},v^-)_{\partial K},\end{aligned}\tag{5.16}$$

where in the first step we integrated by parts and introduced $\mathcal{P}_{\tilde{K}}\xi=\epsilon_h^-+\xi_h^-$ and in the second step we used $\boldsymbol{\nabla}\cdot(\epsilon_h\boldsymbol{\nabla}^\perp\psi)=\boldsymbol{\nabla}\epsilon_h\cdot\boldsymbol{\nabla}^\perp\psi$, since $\boldsymbol{\nabla}\cdot\boldsymbol{\nabla}^\perp\psi=0$. Similarly, we obtain

$$\begin{aligned}&(\xi_h\boldsymbol{\nabla}^\perp\delta_h,\boldsymbol{\nabla}v)_K-(\hat{\xi}_h\boldsymbol{\nabla}^\perp\delta_h\cdot\mathbf{n},v^-)_{\partial K}\\ &\quad=-(\boldsymbol{\nabla}\cdot(\xi_h\boldsymbol{\nabla}^\perp\delta_h),v)_K+(\xi_h^-\boldsymbol{\nabla}^\perp\delta_h\cdot\mathbf{n},v^-)_{\partial K}\\ &\qquad-(\hat{\xi}_h\boldsymbol{\nabla}^\perp\delta_h\cdot\mathbf{n},v^-)_{\partial K}\\ &\quad=-(\boldsymbol{\nabla}\xi_h\cdot\boldsymbol{\nabla}^\perp\delta_h,v)_K+\frac{1}{2}((\xi_h^--\xi_h^+)\boldsymbol{\nabla}^\perp\delta_h\cdot\mathbf{n},v^-)_{\partial K}.\end{aligned}\tag{5.17}$$

For the upwind flux, we obtain a similar relation with $\frac{1}{2}(\xi_h^--\xi_h^+)$ replaced by $\frac{1}{2}(\xi_h^--\xi_h^+)\big(1-\mathrm{sign}(\boldsymbol{\nabla}^\perp\psi_h\cdot\mathbf{n})\big)$ in (5.16) and (5.17), where sign is the sign function, with

$\operatorname{sign}(x) = -1$ if $x < 0$ and $\operatorname{sign}(x) = 1$ if $x \geq 0$. If we introduce (5.16) and (5.17) into (5.15) and use the relation

$$-\frac{1}{2}(\xi_h^- - \xi_h^+)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n} + \frac{1}{2}(\xi_h^- - \xi_h^+)\boldsymbol{\nabla}^\perp\delta_h\cdot\mathbf{n} \\ = -\frac{1}{2}(\xi_h^- - \xi_h^+)\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi)\cdot\mathbf{n} - \frac{1}{2}(\xi_h^- - \xi_h^+)\boldsymbol{\nabla}^\perp\psi_h\cdot\mathbf{n},$$

then we obtain the following relation for ϵ for all $v \in V_h^k$:

$$\begin{aligned}\frac{d}{dt}\Big(\frac{\epsilon}{A}, v\Big)_K &+ (\boldsymbol{\nabla}\epsilon_h\cdot\boldsymbol{\nabla}^\perp\psi, v)_K + (\boldsymbol{\nabla}\xi_h\cdot\boldsymbol{\nabla}^\perp\delta_h, v)_K \\ &+\frac{1}{2}((\epsilon_h^+ - \epsilon_h^-)\boldsymbol{\nabla}^\perp\psi_h\cdot\mathbf{n}, v^-)_{\partial K} \\ &=((\xi - \mathcal{P}_{\tilde{K}}\xi)\boldsymbol{\nabla}^\perp\psi + \xi_h\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi), \boldsymbol{\nabla}v)_K \\ &-\Bigg((\xi^- - \mathcal{P}_{\tilde{K}}\xi)\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n} \\ &+\left(\hat{\xi}_h + \frac{1}{2}(\xi_h^- - \xi_h^+)\right)\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi)\cdot\mathbf{n}, v^-\Bigg)_{\partial K} \\ &=: \mathcal{Q}_K(\xi, \xi_h, \psi; v), \end{aligned} \tag{5.18}$$

where we used for the fourth term on the left-hand side $\xi_h^- - \xi_h^+ = \mathcal{P}_{\tilde{K}}\xi - \xi_h^+ - (\mathcal{P}_{\tilde{K}}\xi - \xi_h^-) = \epsilon_h^+ - \epsilon_h^-$. Note that the right-hand side depends only on ξ_h, ψ, and the interpolation errors $\xi - \mathcal{P}_{\tilde{K}}\xi$ and $\psi - \mathcal{P}_{\tilde{K}}\psi$ but is independent of ϵ, ϵ_h, and δ_h. Using the generalized Hölder inequality (see [1, Cor. 2.6, p. 25]), we can estimate $\mathcal{Q}_K$ as

$$\begin{aligned}\mathcal{Q}_K(\xi, \xi_h, \psi; v) \leq &\|\xi - \mathcal{P}_{\tilde{K}}\xi\|_{p,K}\|\boldsymbol{\nabla}^\perp\psi\|_K\|\boldsymbol{\nabla}v\|_{q,K} \\ &+\|\xi_h\|_K\|\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi)\|_{p,K}\|\boldsymbol{\nabla}v\|_{q,K} \\ &+\|\xi^- - \mathcal{P}_{\tilde{K}}\xi\|_{p,\partial K}\|\boldsymbol{\nabla}^\perp\psi\cdot\mathbf{n}\|_{p',\partial K}\|v^-\|_{\infty,\partial K} \\ &+\left\|\hat{\xi}_h + \frac{1}{2}(\xi_h^- - \xi_h^+)\right\|_{\partial K}\|\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi)\cdot\mathbf{n}\|_{\partial K}\|v^-\|_{\infty,\partial K}, \end{aligned} \tag{5.19}$$

with $p > 2$, $q = \frac{2p}{p-2}$, and $p' = \frac{p}{p-1}$. Using the interpolation estimate (2.3) given by Lemma 2.1, we can directly estimate the interpolation errors in (5.19):

$$\|\xi - \mathcal{P}_{\tilde{K}}\xi\|_{p,K} = |\xi - \mathcal{P}_{\tilde{K}}\xi|_{0,p,K} \leq Ch_K^s|\xi|_{s,p,\tilde{K}}, \tag{5.20}$$

$$\|\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi)\|_{p,K} = |\psi - \mathcal{P}_{\tilde{K}}\psi|_{1,p,K} \leq Ch_K^t|\psi|_{t+1,p,\tilde{K}}, \tag{5.21}$$

with $0 \leq s \leq k+1$, $0 \leq t \leq k$, and $p > 2$. Similarly, using (2.4), together with the fact that the normal vector $\mathbf{n}$ has length one and the triangle inequality, we can estimate the boundary contributions in (5.19):

$$\|\xi^- - \mathcal{P}_{\tilde{K}}\xi\|_{p,\partial K} \leq Ch_K^{s-\frac{1}{p}}|\xi|_{s,p,\tilde{K}}, \tag{5.22}$$

$$\|\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi)\cdot\mathbf{n}\|_{\partial K} \leq Ch_K^t|\psi|_{t+\frac{3}{2},2,\tilde{K}}, \tag{5.23}$$

with $0 < \frac{1}{p} < s \leq k+1$ and $0 < t \leq k - \frac{1}{2}$. Since $\xi_h \in \mathsf{V}_h^k$, which is a finite-dimensional space, we can directly use the equivalence of norms in a finite-dimensional space to obtain the inverse estimate (see, e.g., [5, p. 137])

$$\|\xi_h\|_{\partial K} \leq \frac{C}{\sqrt{h}}\|\xi_h\|_K,$$

which, combined with the triangle inequality, gives for the central flux

$$\left\| \hat{\xi}_h + \frac{1}{2}(\xi_h^- - \xi_h^+) \right\|_{\partial K} \leq \frac{C}{\sqrt{h}} \|\xi_h\|_{\tilde{K}}, \tag{5.24}$$

and a similar relation in the case of the upwind flux. If we choose $t = s + \frac{1}{2}$ in (5.23) and use (5.24), then we obtain

$$\left\| \hat{\xi}_h + \frac{1}{2}(\xi_h^- - \xi_h^+) \right\|_{\partial K} \|\boldsymbol{\nabla}^\perp(\psi - \mathcal{P}_{\tilde{K}}\psi) \cdot \mathbf{n}\|_{\partial K} \leq C h_K^s \|\xi_h\|_{\tilde{K}} |\psi|_{s+2,2,\tilde{K}}, \tag{5.25}$$

with $-\frac{1}{2} < s \leq k-1$. Next, we provide an upper bound for $\|\boldsymbol{\nabla}^\perp\psi \cdot \mathbf{n}\|_{p',\partial K}$ appearing in the third term of (5.19). For this, in the first line of (5.26), we use the Hölder inequality (see [1, Cor. 2.6, p. 25]) with $\frac{1}{p'} = \frac{1}{2} + \frac{1}{w}$. In the second line, we use $\|\mathbf{1}\|_{w,\partial K} \leq C h_K^{\frac{1}{w}}$, with h_K the element diameter, and the trace is estimated for $\frac{1}{2} < t < \frac{3}{2}$ using Theorem 3.38 in [11]:

$$\begin{aligned} \|\boldsymbol{\nabla}^\perp\psi \cdot \mathbf{n}\|_{p',\partial K} &\leq \|\boldsymbol{\nabla}^\perp\psi\|_{p',\partial K} \leq \|\boldsymbol{\nabla}^\perp\psi\|_{2,\partial K} \|\mathbf{1}\|_{w,\partial K} \\ &\leq C h_K^{\frac{1}{w}} \|\boldsymbol{\nabla}^\perp\psi\|_{t-\frac{1}{2},2,\partial K} \leq C h_K^{\frac{1}{w}} \|\boldsymbol{\nabla}^\perp\psi\|_{t,2,K} \leq C h_K^{\frac{1}{w}} \|\psi\|_{t+1,2,K}. \end{aligned} \tag{5.26}$$

We can further sharpen this estimate by introducing $\epsilon = t - \frac{1}{2}$, with $0 < \epsilon \leq \frac{1}{2}$, in the last term of (5.26) and using the imbedding $W^2_{\frac{4}{3-2\epsilon}}(\Omega) \to W_2^{\frac{3}{2}+\epsilon}(\Omega)$, with Ω a Lipschitz domain (see [6, Theorem 1.4.4.1, p. 27]):

$$\|\boldsymbol{\nabla}^\perp\psi \cdot \mathbf{n}\|_{p',\partial K} \leq C h_K^{\frac{1}{w}} \|\psi\|_{\frac{3}{2}+\epsilon,2,K} \leq C h_K^{\frac{1}{w}} \|\psi\|_{2,\frac{4}{3-2\epsilon},K}.$$

This relation can be further evaluated using again the Hölder inequality with $\frac{1}{\frac{4}{3-2\epsilon}} = \frac{1}{2} + \frac{1-2\epsilon}{4}$ and the inequality $(\sum_{i=1}^m a_i^\beta)^{\frac{1}{\beta}} \leq \sum_{i=1}^m a_i$ for $a_i \geq 0$ and $\beta \geq 1$, which give

$$\begin{aligned} \|\boldsymbol{\nabla}^\perp\psi \cdot \mathbf{n}\|_{p',\partial K} &\leq C h_K^{\frac{1}{w}} \left(\sum_{|\alpha|\leq 2} \|D^\alpha\psi\|^{\frac{4}{3-2\epsilon}}_{\frac{4}{3-2\epsilon},K} \right)^{\frac{3-2\epsilon}{4}} \\ &\leq C h_K^{\frac{1}{w}} \left[\sum_{|\alpha|\leq 2} (\|D^\alpha\psi\|_{2,K} \|\mathbf{1}\|_{\frac{4}{1-2\epsilon},K})^{\frac{4}{3-2\epsilon}} \right]^{\frac{3-2\epsilon}{4}} \\ &\leq C h_K^{\frac{1}{w}+\frac{1-2\epsilon}{2}} \left[\sum_{|\alpha|\leq 2} \|D^\alpha\psi\|^{\frac{4}{3-2\epsilon}}_{2,K} \right]^{\frac{3-2\epsilon}{4}} \leq C h_K^{\frac{1}{w}+\frac{1}{2}-\epsilon} \|\psi\|_{2,2,K}. \end{aligned} \tag{5.27}$$

If we introduce (5.20)–(5.22), (5.25), and (5.27) into (5.19) and choose $0 < \epsilon \leq \frac{1}{2}$ such that the relation $s - \frac{1}{p} + \frac{1}{w} + \frac{1}{2} - \epsilon = s - \frac{1}{p} + \frac{1}{p'} - \epsilon = s + \frac{p-2}{p} - \epsilon \geq s$ holds for $p > 2$, then we obtain

$$\begin{aligned} \mathcal{Q}_K(\xi, \xi_h, \psi; v) \leq C h_K^s \big(&\big(|\xi|_{s,p,\tilde{K}} \|\boldsymbol{\nabla}^\perp\psi\|_K + \|\xi_h\|_K |\psi|_{s+1,p,\tilde{K}}\big) \|\boldsymbol{\nabla} v\|_{q,K} \\ &+ \big(|\xi|_{s,p,\tilde{K}} \|\psi\|_{2,2,K} + \|\xi_h\|_{\tilde{K}} |\psi|_{s+2,2,\tilde{K}}\big) \|v^-\|_{\infty,\partial K} \big), \end{aligned} \tag{5.28}$$

with $q = \frac{2p}{p-2}$ and $\frac{1}{p} < s \leq k-1$ determined by the smoothness of ξ and the order k of the polynomial basis functions. Note that, due to the condition $\mathsf{W}_h^k \subset \mathsf{V}_h^k$, which is necessary to ensure conservation of energy as stated in Lemma 4.2, we have to use at least the same polynomial order for ξ_h as for ψ_h. This condition is, however, not essential for the present proof, which requires only polynomials of order $k-2$ for the vorticity field ξ_h when polynomials of order k are used for the stream function ψ_h, but then the discrete energy is no longer conserved.

Next, we further estimate the right-hand side of (5.28). Using Lemmas 4.1 and 4.2, we can also bound the L^2 norm of the vorticity by the initial vorticity

$$\|\xi(\cdot,t)/\sqrt{A}\|_\Omega = \|\xi(\cdot,0)/\sqrt{A}\|_\Omega = \|\xi_0/\sqrt{A}\|_\Omega, \tag{5.29}$$

$$\|\xi_h(\cdot,t)/\sqrt{A}\|_\Omega \leq \|\xi_h(\cdot,0)/\sqrt{A}\|_\Omega = \|\xi_0/\sqrt{A}\|_\Omega, \tag{5.30}$$

with ξ_0 the initial vorticity. An upper bound for the norms of ψ is obtained by first considering ψ_0 defined in (3.2). As shown in the proof of Lemma 5.2, the operator $\tilde{L}$ in (5.4) is coercive and bounded in $H^1_{0,D}(\Omega)$, and together with the condition $A, B \in C^{s,1}(\overline{\Omega})$, the regularity estimate Theorem 4.18(i) on pages 137–138 in [11] implies that for ψ_0 satisfying (5.3) we have the estimate

$$\|\psi_0\|_{s+2,2,\Omega_1} \leq C\big(\|\psi_0\|_{1,2,\Omega_2} + \|\xi\|_{s,2,\Omega_2}\big), \qquad s \geq 0, \tag{5.31}$$

where $\Omega_1 = G_1 \cap \Omega$, $\Omega_2 = G_2 \cap \Omega$, with G_1 and G_2 open subsets of $\mathbb{R}^2$ such that $\overline{G}_1$ is a compact subset of G_2, and $\Gamma_2 = G_2 \cap \partial\Omega$ with Γ_2 a $C^{s+1,1}$ function. Here, G_1 intersects the boundary of Ω, and G_2 has a smooth boundary not necessarily completely contained in Ω. Note that the constant C in (5.31) does not depend on ξ. An estimate for $\|\psi_0\|_{1,2,\Omega}$ satisfying (5.3) is given by (5.9). Combining this result with (5.31) and the estimate for $\|c_D \cdot \mathbf{1}\|_\Omega$ given by (5.10), we obtain

$$\|\psi\|_{s+2,2,\Omega} \leq C\|\xi\|_{s,2,\Omega}, \qquad s \geq 0, \tag{5.32}$$

where we used that Ω can be covered by a finite number of sets Ω_1. Note that the regularity estimate (5.32) also applies for $K \in \mathcal{T}_h$, which follows directly if we set $\Omega_1 = K$. Since $W_2^{s+2}(\Omega)$, $s \geq 0$, is embedded in $C^0(\overline{\Omega})$ (see [1, Theorem 4.12, p. 85]), this also confirms the assumption made at the start of the proof, namely, that ψ is continuous.

In order to estimate $\|\boldsymbol{\nabla}^\perp \psi(\cdot,t)\|_K$, we first use the Hölder inequality with $\frac{1}{2} = \frac{1}{2+\epsilon} + \frac{\epsilon}{2(2+\epsilon)}$ ($\epsilon > 0$), the imbedding Theorem 4.12, Case B in [1] with $j = m = 1$, $p = n = 2$, $q = \frac{2(2+\epsilon)}{\epsilon}$, and (5.32) together with (5.29) to obtain the estimate

$$\begin{aligned}
\|\boldsymbol{\nabla}^\perp \psi(\cdot,t)\|_K &\leq \|\boldsymbol{\nabla}^\perp \psi(\cdot,t)\|_{2(2+\epsilon)/\epsilon,K} \|\mathbf{1}\|_{2+\epsilon,K} \\
&\leq C h_K^{\frac{2}{2+\epsilon}} \|\psi(\cdot,t)\|_{1,2(2+\epsilon)/\epsilon,K} \leq C h_K^{\frac{2}{2+\epsilon}} \|\psi(\cdot,t)\|_{2,2,K} \\
&\leq C h_K^{\frac{2}{2+\epsilon}} \|\xi(\cdot,t)\|_{2,\Omega} \leq C h_K^{\frac{2}{2+\epsilon}} \|\xi_0\|_\Omega.
\end{aligned} \tag{5.33}$$

In a similar way, using the Hölder inequality with $\frac{1}{p} = \frac{1}{p+\epsilon} + \frac{\epsilon}{p(p+\epsilon)}$ ($\epsilon > 0$) and the imbedding $W_2^{s+2}(\Omega) \to W^{s+1}_{p(p+\epsilon)/\epsilon}(\Omega)$ (see Theorem 1.4.4.1 in [6]), we obtain

$$\begin{aligned}
|\psi(\cdot,t)|_{s+1,p,\tilde{K}} &\leq |\psi(\cdot,t)|_{s+1,p(p+\epsilon)/\epsilon,\tilde{K}} \|\mathbf{1}\|_{p+\epsilon,\tilde{K}} \\
&\leq C h_K^{\frac{2}{p+\epsilon}} |\psi(\cdot,t)|_{s+1,p(p+\epsilon)/\epsilon,\tilde{K}} \\
&\leq C h_K^{\frac{2}{p+\epsilon}} \|\psi(\cdot,t)\|_{s+2,2,\Omega} \leq C h_K^{\frac{2}{p+\epsilon}} \|\xi(\cdot,t)\|_{s,2,\Omega}.
\end{aligned} \tag{5.34}$$

We also use the imbedding $W^1_{2+\epsilon}(K) \to L_\infty(K)$ for all $\epsilon > 0$, which is a consequence of Theorem 4.12, Case A in [1]. Applying this for the partial derivatives of v together with the Hölder inequality for $0 < \epsilon < q-2$ with $\frac{1}{2+\epsilon} = \frac{1}{q} + \frac{q-2-\epsilon}{q(2+\epsilon)}$, we obtain

$$\begin{aligned}\|v^-\|_{\infty,\partial K} &\leq C\|v\|_{\infty,K} \leq C\|v\|_{1,2+\epsilon,K} \leq C\|v\|_{1,q,K}\|1\|_{\frac{q(2+\epsilon)}{q-2-\epsilon},K} \\ &\leq Ch_K^{\frac{2(q-2-\epsilon)}{q(2+\epsilon)}}\|v\|_{1,q,K}. \qquad (5.35)\end{aligned}$$

If we introduce (5.30) and (5.32)–(5.35) into (5.28) and use the relation

$$\lim_{\epsilon\to 0} \min\left\{\frac{2(q-2-\epsilon)}{q(2+\epsilon)}, \frac{2}{p+\epsilon}\right\} = \min\left\{\frac{q-2}{q}, \frac{2}{p}\right\} = \frac{2}{p},$$

since $q = \frac{2p}{p-2}$, then we obtain, when $h_K \leq 1$ and for $\epsilon_0 > 0$ arbitrary, the following estimate for $\mathcal{Q}_K$ only in terms of the vorticity field:

$$\begin{aligned}\mathcal{Q}_K(\xi,\xi_h,\psi;v) &\leq Ch_K^{s+\frac{2}{p}-\epsilon_0}\Big(|\xi|_{s,p,\tilde{K}}\|\xi_0\|_\Omega + \|\xi_0\|_\Omega\|\xi\|_{s,2,\Omega} + \|\xi\|_{s,p,\tilde{K}}\|\xi_0\|_\Omega \\ &\quad + \|\xi_0\|_\Omega\|\xi\|_{s,2,\Omega}\Big)\|v\|_{1,q,K} \\ &\leq Ch_K^{s+\frac{2}{p}-\epsilon_0}\|\xi_0\|_\Omega\|\xi\|_{s,p,\Omega}\|v\|_{1,q,K}, \qquad (5.36)\end{aligned}$$

with $\frac{1}{p} < s \leq k-1$ and $q > 2$, which follows directly from the relation $q = \frac{2p}{p-2}$ and the condition $p > 2$.

We now introduce (5.36) into (5.18) and divide this expression by $\|v\|_{1,q,K}$. Finally, we take the supremum over all $v \in W^1_q(K)\backslash\{0\}$, sum over all elements $K \in \mathcal{T}_h$, integrate in time, and obtain an expression for ϵ in the norm of the $W^{-1}_{q'}(\mathcal{T}_h)$ Sobolev space with $\frac{1}{q} + \frac{1}{q'} = 1$, which is dual to $W^1_q(\mathcal{T}_h)$:

$$\begin{aligned}\left\|\frac{\epsilon}{A}\right\|_{-1,q',\mathcal{T}_h} &+ \int_{t_0}^T \Bigg(\|\boldsymbol{\nabla}\epsilon_h\cdot\boldsymbol{\nabla}^\perp\psi\|_{-1,q',\mathcal{T}_h} + \|\boldsymbol{\nabla}\xi_h\cdot\boldsymbol{\nabla}^\perp\delta_h\|_{-1,q',\mathcal{T}_h} \\ &+ \frac{1}{2}\sum_{K\in\mathcal{T}_h}\sup_{0\neq v\in W^1_q(K)}\frac{((\epsilon_h^+ - \epsilon_h^-)\boldsymbol{\nabla}^\perp\psi_h\cdot\mathbf{n}, v^-)_{\partial K}}{\|v\|_{1,q,K}}\Bigg)dt \\ &\leq Ch^{s+\frac{2}{p}-\epsilon_0}\|\xi_0\|_\Omega\int_{t_0}^T\|\xi(\cdot,t)\|_{s,p,\Omega}dt, \qquad (5.37)\end{aligned}$$

with $h = \max_{K\in\mathcal{T}_h} h_K$, $\frac{1}{q'} + \frac{1}{q} = 1$. Note that $1 < q' < 2$, since $q > 2$, and $\frac{1}{p} < s \leq k-1$, with k the order of the polynomial basis functions. Since all terms are positive on the left-hand side of (5.37), we obtain the following negative index Sobolev norm estimate for ϵ/A:

$$\|\epsilon/A\|_{-1,q',\mathcal{T}_h} \leq Ch^{s+\frac{2}{p}-\epsilon_0}\|\xi_0\|_\Omega\int_{t_0}^T\|\xi(\cdot,t)\|_{s,p,\Omega}dt. \qquad (5.38)$$

The error equation for ψ is obtained by subtracting (3.7) from (3.3) and restricting w to W^k_h:

$$L(\delta,w) = (A\boldsymbol{\nabla}\delta,\boldsymbol{\nabla}w)_\Omega + (B\delta,w)_\Omega = -\sum_{K\in\mathcal{T}_h}(\epsilon/A,w)_K. \qquad (5.39)$$

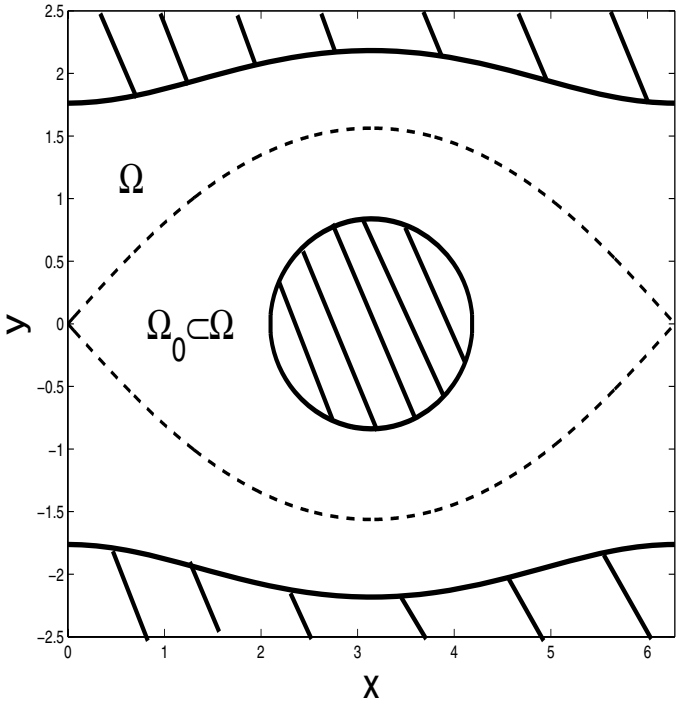


FIG. 1. *Domain used for model problem.*

The operator $L : W_q^1(\Omega) \times W_q^1(\Omega) \to \mathbb{R}$ is coercive on $W_q^1(\Omega)$ with $q > 2$, which follows directly from Theorem 5.3.3 in [4], which also applies to the case $q \neq 2$ (see the remark at the end of page 135 in [4], and the fact that (1.2)–(1.3) are equivalent to a Dirichlet boundary condition on ψ. Using the coercivity of L in $W_q^1(\Omega)$ and (5.39), we obtain the following relation:

$$\begin{aligned}\|\delta\|_{1,q,\Omega} &\leq \frac{1}{\alpha_\delta} \sup_{0\neq w\in W_q^1(\Omega)} \frac{L(\delta, w)}{\|w\|_{1,q,\Omega}} \\ &\leq \frac{1}{\alpha_\delta} \sum_{K\in\mathcal{T}_h} \sup_{0\neq w\in W_q^1(K)} \frac{(\epsilon/A, w)_K}{\|w\|_{1,q,K}} \\ &= \frac{1}{\alpha_\delta} \left\| \frac{\epsilon}{A} \right\|_{-1,q',\mathcal{T}_h}, \end{aligned} \tag{5.40}$$

with α_δ the coercivity constant.

The proof is completed by combining (5.38) and (5.40) and using the relation $\mathbf{u} = A\mathbf{U} = A\boldsymbol{\nabla}^\perp\psi$, and hence $\|\mathbf{u} - \mathbf{u}_h\|_{0,q,\Omega} \leq C\|\psi - \psi_h\|_{1,q,\Omega} = C\|\delta\|_{1,q,\Omega}$. ∎

6. Application to a geophysical model problem. In this section, we consider the flow field generated by a concentrated patch of vorticity in a 2D geophysical flow. The flow field is described by (1.1) with $A = 1$, $B = D = 0$. The flow domain Ω and the subdomain Ω_0 are shown in Figure 1. The domain Ω has an island in the middle, and periodic boundary conditions are applied at $x = 0$ and $x = 2\pi$. Solid walls are hatched and the slip flow boundary condition (1.2) is applied at these

surfaces. We define in Ω the initial vorticity ξ_0 as

$$\xi_0 = \chi_{\Omega_0},$$

with χ_{Ω_0} the characteristic function of Ω_0, which is *one* inside of Ω_0 and *zero* elsewhere. In Appendix A, we prove that $\xi_0 \in W_p^{\frac{1}{p}-\epsilon}(\Omega)$, with $1 \leq p < \infty$ and ϵ an arbitrary positive number, but $\xi_0 \notin H^2(\Omega)$, as is required to apply the error estimates discussed in [3]. The present test case is thus an excellent example of a nonsmooth flow which has the minimal amount of smoothness required by the theory developed in this article.

At first sight, the initial vorticity field appears, however, to be just outside the range of validity for Theorem 5.1, but using the continuity of the Clément interpolation in Lemma 2.1, we can assume that the initial vorticity field $\xi_0 \in W_p^{\frac{1}{p}-\epsilon}$ is equivalent to a slightly smoother field $\tilde{\xi}_0 \in W_p^{\frac{1}{p}+\epsilon}$ which has been taken such that for all macroelements $\tilde{K}$

$$\mathcal{P}_{\tilde{K}}\tilde{\xi_0} = \mathcal{P}_{\tilde{K}}\xi_0 \quad \text{and} \quad \lim_{\epsilon\to 0} \|\xi_0\|_{\frac{1}{p}-2\epsilon,p,\tilde{K}} = \lim_{\epsilon\to 0} \|\tilde{\xi}_0\|_{\frac{1}{p}+\epsilon,p,\tilde{K}}.$$

Since in Theorem 5.1 we must have $p > 2$, we choose $p = 2+\epsilon$, $r = \frac{1}{2+\epsilon} - \epsilon$, which gives that $q = \frac{2(2+\epsilon)}{\epsilon}$, $q' = \frac{2(2+\epsilon)}{2(2+\epsilon)-\epsilon}$, and therefore $s \leq \min\{k-1, \frac{3}{2+\epsilon} - \epsilon\}$. Taking the limits $\epsilon \to 0$, we should then investigate the $L^\infty(\Omega)$ norm for the error in the velocity field and the $W_1^{-1}(\mathcal{T}_h)$ norm for the vorticity field.

In the computations, we used second order polynomials ($k = 2$), and therefore we expect based on Theorem 5.1 a first order convergence in the L^∞ norm for the velocity and in the $W_1^{-1}(\mathcal{T}_h)$ norm for the vorticity. The time integration is conducted with a third order Runge–Kutta method with such a small time step that it does not influence the spatial accuracy. Since we do not know the exact solution, we use a sequence of unstructured uniformly refined meshes, see Figure 2, to obtain an accurate estimate for the norm of the error and the rate of convergence s. The finest mesh is denoted as reference mesh, with mesh size h_{ref} and solution u_{ref}. The meshes satisfy $h_{\text{ref}} < h_3 < h_2 < h_1$, with the mesh size approximately doubling for each mesh; see Table 1. Assuming the following asymptotic behavior of the error,

$$\|u - u_h\| \leq \|u - u_{\text{ref}}\| + \|u_{\text{ref}} - u_h\| \cong C_r h^s, \tag{6.1}$$

we can eliminate the unknown contribution $\|u - u_{\text{ref}}\|$ by considering (6.1) for three meshes with mesh sizes h_1, h_2, and h_3, respectively. This results in the following relation for the rate of convergence s:

$$\frac{h_1^s - h_2^s}{h_2^s - h_3^s} = \frac{\|u_{\text{ref}} - u_{h_1}\| - \|u_{\text{ref}} - u_{h_2}\|}{\|u_{\text{ref}} - u_{h_2}\| - \|u_{\text{ref}} - u_{h_3}\|}, \tag{6.2}$$

which is solved numerically. The constant C_r can be computed easily from

$$C_r = \frac{1}{h_2^s - h_3^s}\Big(\|u_{\text{ref}} - u_{h_2}\| - \|u_{\text{ref}} - u_{h_3}\|\Big), \tag{6.3}$$

and (6.1) then provides an estimate the norm of the error.

In Table 1, we present the estimated L_2-error $\|\xi_0 - \xi_{0h}\|_{\mathcal{T}_h}$ and convergence rate $\hat{s}_\xi$ for the initial vorticity field for different numbers of elements and mesh sizes, obtained

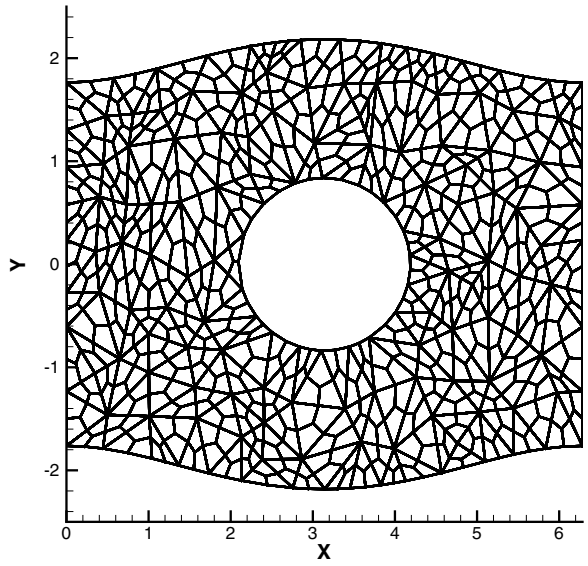


FIG. 2. *Unstructured mesh in the domain* Ω.

TABLE 1
Estimated L_2*-error and rate of convergence in the initial vorticity field.*

Number of elements	h	$\|\xi_0 - \xi_{0h}\|_{\mathcal{T}_h}$	$\hat{s}_\xi$
782	0.0850288	0.5799913	-
3004	0.0434386	0.3982912	0.5596
11768	0.0219504	0.2886455	0.4717
46576	0.0110329	0.2029058	0.5124
185312	0.0055309	0.1433234	0.5034

from (6.1)–(6.3). For $p = 2 + \epsilon$, with $\epsilon > 0$ arbitrary small, the initial vorticity field satisfies $\xi_0 \in W_{2+\epsilon}^{\frac{1}{2+\epsilon}-\epsilon}$, and we expect a convergence rate $\hat{s}_\xi \cong 0.5$, which is closely confirmed by the results in Table 1.

In the time dependent computations, the initial vorticity field evolves into a complex pattern of continuously thinner shear layers which wrap around each other, but since there is no viscosity in the fluid, the smoothness of the vorticity field does not increase. The vorticity field at several instances in time is shown in Figures 3–5. Note that, since the element boundaries are not aligned with the jumps in vorticity, we see a slight distortion in the plots near these discontinuities.

In Table 2, the estimated L_∞-error in the velocity field $\|\mathbf{u} - \mathbf{u}_h\|_{\infty,\Omega}$ and the estimated convergence rate $\hat{s}_\mathbf{u}$ as a function of the number of elements and mesh size are presented. These results show that the convergence rate for the velocity field is close to one, as was predicted by Theorem 5.1. For the vorticity, we would have to compute the $W_1^{-1}(\mathcal{T}_h)$ norm of the error, but unfortunately it was not possible to obtain accurate numerical estimates for this norm, since this requires meshes beyond our computational means.

Considering the results on the order of accuracy of the CDG scheme summarized in Table 2, we can conclude that the analysis is sharp for this type of nonsmooth flow and also that the CDG scheme behaves properly for this nonsmooth flow field. More

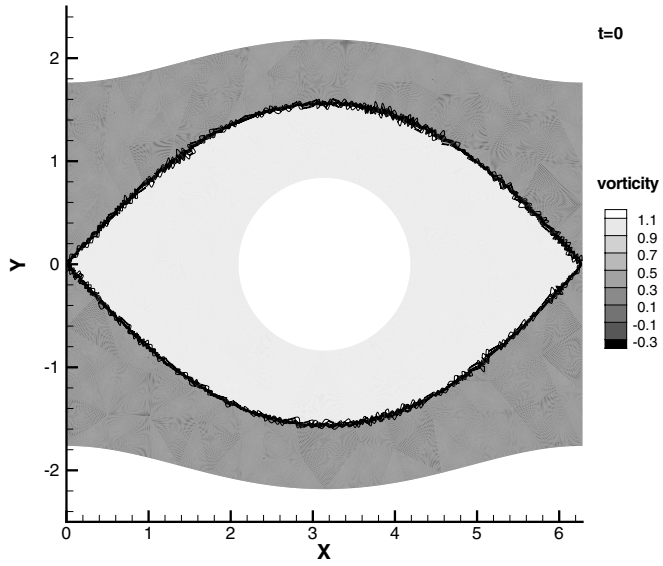


FIG. 3. *Vorticity field at* $t = 0$.

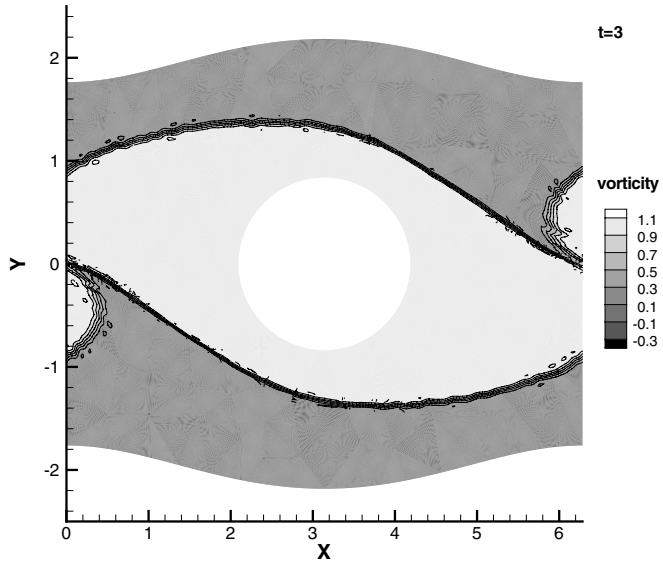


FIG. 4. *Vorticity field at* $t = 0.3$.

details on the accuracy of the CDG scheme for smooth flows can be found in [3].

Appendix A. Regularity of characteristic function χ_{Ω_0} in 2D domains. In this section, we investigate which fractional order Sobolev spaces contain the characteristic function χ_{Ω_0}.

LEMMA A.1. *Let* $\Omega \subset \mathbb{R}^2$ *denote a simply connected domain with boundary* $\partial\Omega$. *Let* $\Omega_0 \subset \Omega$ *denote a subdomain with* $\bar{\Omega}_0 \subset \Omega$ *and Lipschitz boundary* $\partial\Omega_0$. *Then*

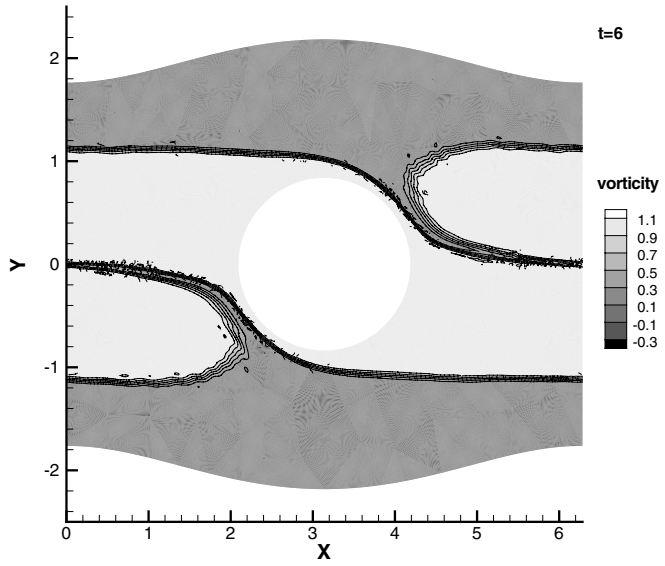


FIG. 5. *Vorticity field at* $t = 0.6$.

TABLE 2
Estimated L_∞*-error and rate of convergence for the velocity field.*

Number of elements	h	$\|\mathbf{u} - \mathbf{u}_h\|_{\infty,\Omega}$	$\hat{s}_{\mathbf{u}}$
782	0.0850288	0.0603981	-
3004	0.0434386	0.0286827	1.1085
11768	0.0219504	0.0157522	0.8780
46576	0.0110329	0.0080533	0.9753

the characteristic function of Ω_0, $\chi_{\Omega_0} : \Omega \to \mathbb{R}$ *(which is* one *inside of* Ω_0 *and* zero *elsewhere) belongs to the space* $W_p^{\frac{1}{p}-\epsilon}(\Omega)$, *with* $1 \le p < \infty$ *and* $\epsilon > 0$.

Proof. We use the definition of fractional order Sobolev spaces $W_p^s(\Omega)$ in [4] (with $n = 2$ and $1 \le p < \infty$), and, accordingly, we have to determine s such that

$$\int_\Omega \int_\Omega \frac{|\chi_{\Omega_0}(x) - \chi_{\Omega_0}(y)|^p}{|x-y|^{2+ps}} \,\mathrm{d}x\,\mathrm{d}y \tag{A.1}$$

is finite. In the consecutive estimates, d_Ω denotes the diameter of Ω. The distance of two sets $S_1, S_2 \subset \mathbb{R}^2$ is denoted with $d(S_1, S_2) = \inf\{|x_1 - x_2| : x_1 \in S_1, x_2 \in S_2\}$, while $B(x, r)$ yields the open ball $\{y \in \mathbb{R}^2 : |x - y| < r\}$. Using the definition of χ_{Ω_0}, it is clear that $\chi_{\Omega_0}(x) - \chi_{\Omega_0}(y) = 0$ whenever $x, y \in \Omega_0$ or $x, y \in \Omega \setminus \Omega_0$. Therefore, we can estimate the double integral in (A.1) as follows:

$$\begin{aligned}
\int_\Omega \int_\Omega \frac{|\chi_{\Omega_0}(x) - \chi_{\Omega_0}(y)|^p}{|x-y|^{2+ps}} \,\mathrm{d}x\,\mathrm{d}y &= 2\int_{\Omega_0}\int_{\Omega\setminus\Omega_0} \frac{1}{|x-y|^{2+ps}} \,\mathrm{d}x\,\mathrm{d}y \\
&\le 2\int_{\Omega_0}\int_{B(y,d_\Omega)\setminus B(y,d(y,\partial\Omega_0))} \frac{1}{|x-y|^{2+ps}} \,\mathrm{d}x\,\mathrm{d}y \\
&= -\frac{4\pi}{ps} d_\Omega^{-ps}|\Omega_0| + \frac{4\pi}{ps}\int_{\Omega_0} d(y,\partial\Omega_0)^{-ps} \,\mathrm{d}y, \qquad \text{(A.2)}
\end{aligned}$$

with $|\Omega_0|$ the Lebesgue measure of Ω_0. For an upper bound of the integral in (A.2), we use the definition of the Lipschitz domain as given in [11, Definition 3.28] and the consecutive analysis. According to this, there is a finite cover $\cup W_j$ of $\partial\Omega_0$ with open sets W_j and a finite collection of open sets Ω_j such that $W_j \cap \Omega_0 = W_j \cap \Omega_j$ for all indices j and Ω_j can be considered as a Lipschitz hypograph (applying a measure preserving transformation).

Since $\cup W_j$ is an open set containing $\partial\Omega_0$, there is a positive number ϵ such that for all $y \in \Omega_0 \setminus \cup_j W_j$ the relation $d(y, \partial\Omega_0) > \epsilon$ holds. Accordingly, we split the integral in (A.2) as

$$\begin{aligned}\int_{\Omega_0} d(y,\partial\Omega_0)^{-ps}\,\mathrm{d}y &= \int_{\Omega_0\setminus\cup_j W_j} d(y,\partial\Omega_0)^{-ps}\,\mathrm{d}y + \int_{\Omega_0\cap\cup_j W_j} d(y,\partial\Omega_0)^{-ps}\,\mathrm{d}y \\ &\le \int_{\Omega_0\setminus\cup_j W_j} \epsilon^{-ps}\,\mathrm{d}y + \sum_j \int_{\Omega_j\cap W_j} d(y,\partial\Omega_0)^{-ps}\,\mathrm{d}y. \qquad \text{(A.3)}\end{aligned}$$

The first integral in (A.3) is obviously finite; therefore, in the consecutive analysis we give an upper estimate for an individual term in the sum.

Since Ω_j is a Lipschitz hypograph and W_j is bounded, we get the inclusion

$$\Omega_j \cap W_j \subset \{y = (y', y_n) \in \mathbb{R}^{n-1} \times \mathbb{R}^+ : y' \in \Omega_j^*, y_n \le f_j(y')\} \qquad \text{(A.4)}$$

in an appropriate coordinate system with the Lipschitz function f_j, where $\Omega_j^* \subset \mathbb{R}^{n-1}$ such that $\operatorname{diam}\Omega_j^* \le \operatorname{diam}\Omega_0$. For an arbitrary point $z = (z', f_j(z')) = (z', z_n)$ on $\partial\Omega_j \cap \partial\Omega_0$, we have

$$\begin{aligned}|f_j(y') - y_n| &= |z_n - y_n + f_j(y') - f_j(z')| \le |z_n - y_n| + M_j|z' - y'| \\ &\le |z_n - y_n| + M_j|z' - y'| \le \sqrt{1 + M_j^2}|z - y|, \qquad \text{(A.5)}\end{aligned}$$

with the Lipschitz constant M_j of f_j. In this way, for any j and any $y \in \Omega_j \cap W_j$, taking the infimum over all $z \in \partial\Omega_j \cap \partial\Omega_0$, the estimate $d(y, \partial\Omega_0) \ge \frac{f_j(y') - y_n}{\sqrt{1+M^2}}$ holds, where $M = \max_j M_j$. Accordingly, using (A.4) we obtain for all indices j that

$$\begin{aligned}\int_{\Omega_j\cap W_j} d(y,\partial\Omega_0)^{-ps}\,\mathrm{d}y &\le (1+M^2)^{\frac{ps}{2}} \int_{\Omega_j\cap W_j} (f_j(y') - y_n)^{-ps}\,\mathrm{d}y \\ &= (1+M^2)^{\frac{ps}{2}} \int_{\Omega_j^*}\int_0^{f_j(y')} (f_j(y') - y_n)^{-ps}\,\mathrm{d}y_n\,\mathrm{d}y',\end{aligned}$$

which is finite when $ps < 1$. ∎

Acknowledgment. It is a pleasure to thank M.Sc. Erik Bernsen for getting JvdV involved with the error analysis of CDG methods, pointing out reference [10], and for providing assistance with the computations.

REFERENCES

[1] R. A. Adams and J. J. F. Fournier, *Sobolev Spaces*, 2nd ed., Academic Press, New York, 2003.

[2] C. Bernardi and V. Girault, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.

[3] E. Bernsen, O. Bokhove, and J. J. W. van der Vegt, *A (dis)continuous finite element model for generalized 2D vorticity dynamics*, J. Comput. Phys., 211 (2006), pp. 719–747.

[4] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer, New York, 2002.
[5] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, P. G. Ciarlet and J. L. Lions, eds., Vol. II, North–Holland, Amsterdam, 1991, pp. 17–351.
[6] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, MA, 1985.
[7] L. D. Landau and E. M. Lifschitz, *Fluid Mechanics*, Pergamon Press, London, Paris, Frankfurt, 1959.
[8] P. H. LeBlond and L. A. Mysak, *Waves in the Ocean*, Elsevier, New York, 1978.
[9] J. Liu and C.-W. Shu, *A high-order discontinuous Galerkin method for* 2*d incompressible flows*, J. Comput. Phys., 160 (2000), pp. 577–596.
[10] J.-G. Liu and Z. Xin, *Convergence of a Galerkin method for* 2*-D discontinuous Euler flows*, Comm. Pure Appl. Math., 53 (2000), pp. 786–798.
[11] W. C. H. McLean, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambridge University Press, Cambridge, UK, 2000.
[12] J. Pedlosky, *Geophysical Fluid Dynamics*, Springer, Berlin, 1979.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1370–1397

DISCONTINUOUS GALERKIN FINITE ELEMENT APPROXIMATION OF NONLINEAR SECOND-ORDER ELLIPTIC AND HYPERBOLIC SYSTEMS*

CHRISTOPH ORTNER† AND ENDRE SÜLI†

Abstract. We develop the convergence analysis of discontinuous Galerkin finite element approximations to symmetric second-order quasi-linear elliptic and hyperbolic systems of partial differential equations in divergence form in a bounded spatial domain in $\mathbb{R}^d$, subject to mixed Dirichlet–Neumann boundary conditions. Optimal-order asymptotic bounds are derived on the discretization error in each case without requiring the global Lipschitz continuity or uniform monotonicity of the stress tensor. Instead, only local smoothness and a Gårding inequality are used in the analysis.

Key words. nonlinear elliptic and hyperbolic systems of partial differential equations, discontinuous Galerkin methods, Legendre–Hadamard condition, broken Gårding inequality

AMS subject classifications. 65M60, 74S05, 74H15

DOI. 10.1137/06067119X

1. Introduction.

Second-order nonlinear elliptic and hyperbolic systems of partial differential equations arise in numerous applications, and a substantial body of research has been devoted to their analytical and computational study. This paper is concerned with the construction and convergence analysis of a class of numerical algorithms—discontinuous Galerkin finite element methods—for the approximate solution of quasi-linear elliptic and hyperbolic systems. Nonlinear elasticity is a particularly fertile source of equations of this type, and our results are phrased with this particular application area in mind, although the ideas and techniques developed are valid generally, provided the structural hypotheses on the nonlinearity assumed herein are satisfied.

In order to motivate the discussion that will follow, we begin by formulating a static problem from nonlinear elasticity which results in a mixed Dirichlet–Neumann boundary-value problem for a system of second-order quasi-linear elliptic partial differential equations. We shall then state the corresponding dynamic problem, which is a mixed initial-boundary-value problem for a second-order quasi-linear hyperbolic system.

Suppose that Ω is a bounded open set in $\mathbb{R}^d$, $d \in \{2, 3\}$, with Lipschitz continuous boundary $\partial\Omega$. We shall seek a displacement field $u : \overline{\Omega} \to \mathbb{R}^d$ such that u is a stationary point of the energy functional

$$J : v \mapsto J(v) := \int_\Omega [W(\nabla v(x)) - f(x) \cdot v(x)] \, \mathrm{d}x - \int_{\Gamma_\mathrm{N}} g_\mathrm{N}(s) \cdot v(s) \, \mathrm{d}s, \tag{1.1}$$

defined over the set of all (sufficiently smooth) d-component vector functions v on $\overline{\Omega}$ satisfying the boundary condition $v = g_\mathrm{D}$ on Γ_D, where $\Gamma_\mathrm{D} \subset \Gamma = \partial\Omega$ has positive

*Received by the editors October 1, 2006; accepted for publication (in revised form) March 22, 2007; published electronically July 11, 2007. The authors acknowledge the financial support received from the European research project HPRN-CT-2002-00284: *New Materials, Adaptive Systems and their Nonlinearities. Modelling, Control and Numerical Simulation* and the kind hospitality of Carlo Lovadina and Matteo Negri (University of Pavia).

http://www.siam.org/journals/sinum/45-4/67119.html

†University of Oxford, Computing Laboratory, Wolfson Building, Parks Road, Oxford OX1 3QD, United Kingdom (Christoph.Ortner@comlab.ox.ac.uk, Endre.Suli@comlab.ox.ac.uk).

$(d-1)$-dimensional surface measure $\mathcal{H}^{d-1}(\Gamma_{\rm D})$, $\Gamma_{\rm N} = \Gamma \setminus \Gamma_{\rm D}$, $W \in {\rm C}^4(\mathbb{R}^{d\times d};\mathbb{R})$ is the stored energy function, $f \in {\rm L}^2(\Omega)^d$ is a given body force, and $g_{\rm N} \in {\rm L}^2(\Gamma_{\rm N})^d$. Let us define the Piola–Kirchhoff stress tensor S as the gradient of W, that is,

$$S_{i\alpha}(\eta) := \frac{\partial}{\partial \eta_{i\alpha}} W(\eta), \qquad \eta \in \mathbb{R}^{d\times d},$$

and let

$$A_{i\alpha j\beta}(\eta) := \frac{\partial}{\partial \eta_{j\beta}} S_{i\alpha}(\eta) = \frac{\partial^2}{\partial \eta_{i\alpha}\partial \eta_{j\beta}} W(\eta), \qquad \eta \in \mathbb{R}^{d\times d}.$$

Clearly, $A_{i\alpha j\beta}(\eta) = A_{j\beta i\alpha}(\eta)$ for all $\eta \in \mathbb{R}^{d\times d}$ and $i,\alpha,j,\beta = 1,\dots,d$.

Formal calculations show that sufficiently smooth stationary points $u = u(x)$ of the functional J satisfy the following Euler–Lagrange equation:

$$-\sum_{\alpha=1}^{d} \partial_{x_\alpha} S_{i\alpha}(\nabla u(x)) = f_i(x), \qquad i = 1,\dots,d, \qquad x \in \Omega, \tag{1.2}$$

subject to the boundary conditions

$$u = g_{\rm D} \quad \text{on } \Gamma_{\rm D} \qquad \text{and} \qquad S(\nabla u)\nu = g_{\rm N} \quad \text{on } \Gamma_{\rm N}, \tag{1.3}$$

on the Dirichlet and Neumann parts $\Gamma_{\rm D}$ and $\Gamma_{\rm N}$ of the boundary Γ, respectively. Here ν is the unit outward normal vector to Γ, and $\partial_{x_\alpha} = \partial/\partial x_\alpha$. We note that, except in section 7, we do not use the fact that (1.2) is an Euler–Lagrange equation but only require the symmetry of the tensor $A_{i\alpha j\beta}$.

The weak formulation of the boundary-value problem (1.2), (1.3) is posed as follows: Find the function $u \in {\rm H}^1_{{\rm D},g_{\rm D}}(\Omega)^d = \{v \in {\rm H}^1(\Omega)^d \,:\, v|_{\Gamma_{\rm D}} = g_{\rm D}\}$ such that

$$\int_\Omega S(\nabla u) : \nabla v \,{\rm d}x = \int_\Omega f \cdot v \,{\rm d}x + \int_{\Gamma_{\rm N}} g_{\rm N} \cdot v \,{\rm d}s \qquad \forall v \in {\rm H}^1_{{\rm D},0}(\Omega)^d.$$

We shall assume that this problem has a solution $u \in {\rm H}^{m+1}(\Omega)^d \cap {\rm H}^1_{{\rm D},g_{\rm D}}(\Omega)^d$, with $m > d/2$. By the Sobolev embedding theorem u is then, in fact, contained in ${\rm C}^{1,\hat\alpha}(\overline{\Omega})^d$ for some $\hat\alpha \in (0,1)$.

For future reference we also define the bilinear form $a(\Phi;\cdot,\cdot)$, $\Phi \in {\rm L}^\infty(\Omega)^{d\times d}$, by

$$a(\Phi;v,w) := \sum_{i,\alpha,j,\beta=1}^{d} \int_\Omega A_{i\alpha j\beta}(\Phi)\partial_{x_\alpha} v_i \partial_{x_\beta} w_j \,{\rm d}x \qquad \forall v,w \in {\rm H}^1_{{\rm D},0}(\Omega)^d. \tag{1.4}$$

Formally at least, $a(\nabla u;\cdot,\cdot)$ defines the hessian of J at u; more generally, we shall consider $a(\Phi;\cdot,\cdot)$ for Φ in a certain neighborhood of ∇u which we shall now define.

For $\delta > 0$, let

$$\mathcal{Z}_\delta := \big\{\Phi \in {\rm C}_{\rm pw}(\overline{\Omega})^{d\times d} \,:\, \|\Phi - \nabla u\|_{{\rm L}^\infty(\Omega)} \le \delta\big\}, \tag{1.5}$$

where ${\rm C}_{\rm pw}(\overline{\Omega})$ denotes the set of bounded piecewise continuous functions defined on $\overline{\Omega}$. The set $\mathcal{Z}_\delta$ will be required in the convergence analysis of the finite element method: We will show that, for sufficiently small h, it contains the *piecewise gradients* (relative to the finite element subdivision $\mathcal{T}_h$ of the computational domain Ω) of discontinuous

Galerkin finite element approximations to u. Their point values must therefore be contained in the set

$$\mathcal{M}_\delta := \mathrm{conv}\big\{\eta \in \mathbb{R}^{d\times d} : \inf_{x\in\Omega} |\eta - \nabla u(x)| \le \delta\big\},$$

where $|\cdot|$ denotes the Frobenius norm on $\mathbb{R}^{d\times d}$ defined, for $\eta \in \mathbb{R}^{d\times d}$, by $|\eta| = (\eta : \eta)^{1/2}$. Clearly, as it is the convex hull of a closed and bounded set, $\mathcal{M}_\delta$ is itself closed, bounded, and, of course, convex.

We note here that we do not require S to be globally Lipschitz continuous, but we will use the local Lipschitz constant of S in $\mathcal{M}_\delta$, defined by

$$K_\delta := \sup_{\eta\in\mathcal{M}_\delta} \left(\sum_{i,\alpha,j,\beta=1}^{d} |A_{i\alpha j\beta}(\eta)|^2 \right)^{1/2}, \tag{1.6}$$

and the local Lipschitz constant of the fourth-order elasticity tensor $A = \nabla S$, defined by

$$L_\delta := \sup_{\eta,\sigma\in\mathcal{M}_\delta,\eta\neq\sigma} |\eta-\sigma|^{-1} \left(\sum_{i,\alpha,j,\beta=1}^{d} |A_{i\alpha j\beta}(\eta) - A_{i\alpha j\beta}(\sigma)|^2 \right)^{1/2}. \tag{1.7}$$

Since, for every $\delta > 0$, the set $\mathcal{M}_\delta$ is compact in $\mathbb{R}^{d\times d}$ and $A \in \mathrm{C}^2(\mathcal{M}_\delta)^{d\times d\times d\times d}$, it follows that K_δ and L_δ are finite.

We shall also consider the dynamic counterpart of the boundary-value problem (1.2), (1.3)—the initial-boundary-value problem for the second-order nonlinear evolution equation

$$\partial_t^2 u_i - \sum_{\alpha=1}^{d} \partial_{x_\alpha} S_{i\alpha}(\nabla u) = f_i(t,x), \qquad i = 1,\dots,d, \qquad x \in \Omega, \quad t \in (0,T], \tag{1.8}$$

subject to the initial conditions $u(0,x) = u_0(x)$, $\partial_t(0,x) = u_1(x)$, $x \in \Omega$, and the same boundary conditions as in the static problem above. Here $\partial_t^2 u = \frac{\partial^2 u}{\partial t^2}$; we shall also write $\ddot{u}$ instead of $\partial_t^2 u$ and $\dot{u}$ instead of $\partial_t u = \frac{\partial u}{\partial t}$. For a detailed discussion concerning the physical background to these equations in the field of nonlinear elasticity, we refer to [11, 1], for example. Suitable generalizations of the sets $\mathcal{M}_\delta$ and $\mathcal{Z}_\delta$ for the hyperbolic case are given in section 6.

We now formulate our structual hypotheses on the stress tensor S. For most constitutive laws in solid mechanics and many other applications, the mapping $\eta \mapsto S(\eta)$ satisfies the *axiom of frame indifference*, that is,

$$S(F - \mathrm{id}) = S(QF - \mathrm{id}) \qquad \forall Q \in SO(d), \quad \forall F \in \mathbb{R}^{d\times d}, \tag{1.9}$$

where id is the $d \times d$ identity matrix and $SO(d)$ is the group of special orthogonal $d \times d$ matrices. Note that the form of (1.9) is slighly nonstandard, as our partial differential equation is formulated in terms of displacement rather than deformation. If S satisfies (1.9), then, except in trivial cases, S cannot be monotone; for a detailed discussion of this point, we refer to pages 490–491 in the monograph of Antman [1]. Hence, the *uniform monotonicity* condition which hypothesizes the existence of a real number $M_1 > 0$ such that

$$(S(F) - S(G)) : (F - G) \ge M_1 |F - G|^2 \qquad \forall F, G \in \mathbb{R}^{d\times d}, \tag{1.10}$$

which is commonly assumed in the analysis of finite element approximations to quasi-linear elliptic problems, is inappropriate in the context of nonlinear elasticity and needs to be relaxed in order to cover physically meaningful problems.

In fact, the condition (1.10) can be relaxed in several ways in order to capture the physics while still recovering some of the theory available in the uniformly elliptic setting which stems from the uniform monotonicity condition (1.10). It is reasonable, for example, to assume that a metastable state of the elastic energy functional (1.1) is not merely a critical point satisfying the Euler–Lagrange equation but that the hessian of J is positive definite at this point. Thus, in the static case, we shall replace (1.10) by the following condition, which requires the existence of a real number $M_1 = M_1(u) > 0$ such that

$$a(\nabla u; v, v) \geq M_1 \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall v \in \mathrm{H}^1_{\mathrm{D},0}(\Omega)^d. \tag{1.11}$$

Similarly, for the dynamic case, it was shown in [8] that, if S satisfies the strong Legendre–Hadamard condition

$$\sum_{i,\alpha,j,\beta=1}^{d} A_{i\alpha j\beta}(\eta)\zeta_i\zeta_j\xi_\alpha\xi_\beta \geq M_1|\zeta|^2|\xi|^2 \qquad \forall \zeta, \xi \in \mathbb{R}^d, \quad \forall \eta \in \mathbb{R}^{d\times d}, \tag{1.12}$$

for some constant $M_1 > 0$, then a smooth solution to (1.8) is guaranteed to exist locally in time, subject to given initial conditions and the same boundary conditions as in the static case (at least when $\Gamma_{\mathrm{N}} = \emptyset$ and $g_{\mathrm{D}} = 0$). Condition (1.12) is satisfied by most constitutive laws for elastic materials. In this case, the semilinear form a defined in (1.4) satisfies the following Gårding inequality: For any $\varphi \in \mathrm{C}^1(\overline{\Omega})^d$, there exists $M_0 = M_0(\varphi) \geq 0$ such that

$$a(\nabla\varphi; v, v) \geq \tfrac{1}{2}M_1\|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - M_0(\varphi)\|v\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall v \in \mathrm{H}^1_0(\Omega)^d; \tag{1.13}$$

cf. Theorem 6.5.1 on p. 253 in [16]. Even this weaker inequality is, to the best of our knowledge, known only for $v \in \mathrm{H}^1_0(\Omega)^d$. As we shall see, (1.13) is sufficient for the convergence analysis in the dynamic case.

In the case of classical conforming finite element methods based on finite-dimensional subspaces of $\mathrm{H}^1_{\mathrm{D},0}(\Omega)^d$ or $\mathrm{H}^1_0(\Omega)^d$, as the case may be, consisting of continuous piecewise polynomial functions of degree $p \geq 1$ defined over a family of subdivisions $\{\mathcal{T}_h\}_{h>0}$ of the computational domain Ω, the inequalities (1.11) and (1.13) will automatically hold in such subspaces. Discontinuous Galerkin finite element methods which are the focus of this paper are, however, built over finite-dimensional spaces consisting of discontinuous piecewise polynomial functions defined on Ω, which are, clearly, *not* contained in $\mathrm{H}^1(\Omega)^d$, let alone $\mathrm{H}^1_{\mathrm{D},0}(\Omega)^d$ or $\mathrm{H}^1_0(\Omega)^d$. As a matter of fact, both (1.11) and (1.13) are global conditions and, unlike uniform monotonicity (1.10), do not automatically translate to the space $S^p(\Omega, \mathcal{T}_h, \mathbf{F})$, defined in section 2, of discontinuous piecewise polynomial functions of degree p on $\mathcal{T}_h$. Thus, in section 3, we shall derive the "broken" versions of these inequalities which hold over $S^p(\Omega, \mathcal{T}_h, \mathbf{F})$. To the best of our knowledge, the analysis of discontinuous Galerkin finite element approximations to second-order quasi-linear systems of partial differential equations has not been previously considered in the literature under such weak structural assumptions.

In recent years there has been considerable interest in discontinuous Galerkin finite element methods for the numerical solution of a wide range of partial differential equations which arise from continuum mechanics. We shall not attempt to

give a detailed review of this area of research: The reader is referred to [7] for a comprehensive historical survey of the field and [2, 13] for convergence analyses of the method for second-order linear elliptic problems and partial differential equations with nonnegative characteristic form. Discontinuous Galerkin finite element methods were introduced in the early 1970s for the numerical solution of first-order hyperbolic problems. Simultaneously, but quite independently, they were proposed as nonstandard schemes for the approximation of second-order elliptic equations. The recent upsurge of interest in this class of techniques has been stimulated by the computational convenience of discontinuous Galerkin methods due to their high degree of locality and the presence of associated local conservation properties, as well as the need to accommodate high-order *hp* and spectral element discretizations on irregular finite element meshes. The present work has been stimulated by our ongoing research on discontinuous Galerkin methods in the field of fracture mechanics.

The paper is structured as follows. The next section is devoted to the construction of the discontinuous Galerkin method for the nonlinear elliptic boundary-value problem (1.2), (1.3). In section 3, we derive broken Gårding inequalities to aid us in our subsequent analysis. In section 4 we develop the linearization of the semilinear form appearing in the definition of the finite element method. In section 5 we perform the convergence analysis of the discontinuous Galerkin finite element approximation of the elliptic boundary-value problem (1.2), (1.3) under hypothesis (1.11). We note, in particular, that our analysis does not assume the global Lipschitz continuity of the functions $S_{i\alpha}$, $i, \alpha = 1, \ldots, d$, with respect to ∇u, nor do we explicitly require the uniform monotonicity condition (1.10). Building on the work of Makridakis [15] for classical conforming methods, in section 6 we develop the convergence analysis of semidiscrete discontinuous Galerkin finite element approximations of mixed Dirichlet–Neumann initial-boundary-value problems for systems of second-order quasi-linear hyperbolic equations of the form (1.8). This analysis requires a nonlinear projection operator whose approximation properties are analyzed, closely following section 5, in Appendix A. Extensions of our analysis to fully discrete approximations of the hyperbolic problem would proceed along the same lines as in [15] in the case of conforming methods; thus, we do not consider these here. In section 7 we show how our framework can be used to derive optimal error estimates for discontinuous Galerkin finite element methods other than the formulation which we have adopted in this paper.

2. Finite element spaces. For $h \in (0, 1]$, let $\mathcal{T}_h$ be a subdivision of Ω into disjoint open *element domains* (or, simply, *elements*) κ such that $\overline{\Omega} = \cup_{\kappa \in \mathcal{T}_h} \overline{\kappa}$. Here $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$, where $h_\kappa = \mathrm{diam}(\kappa)$. Each $\kappa \in \mathcal{T}_h$ is assumed to be the image of the open reference simplex under a bijective affine mapping or of the open unit hypercube under a bilinear mapping, denoted by F_κ. We shall denote either master element by $\hat{\kappa}$.

For a nonnegative integer k, we denote by $\mathcal{P}_k(\hat{\kappa})$ the set of polynomials of total degree k on $\hat{\kappa}$. When $\hat{\kappa}$ is the unit hypercube, we also consider $\mathcal{Q}_k(\hat{\kappa})$, the set of all tensor-product polynomials on $\hat{\kappa}$ of degree k in each coordinate direction. We collect the F_κ in the vector $\mathbf{F} = \{F_\kappa : \kappa \in \mathcal{T}_h\}$ and consider, for $p \geq 1$, the finite element space

$$S^p(\Omega, \mathcal{T}_h, \mathbf{F}) := \{v \in \mathrm{L}^2(\Omega)^d : \ v|_\kappa \circ F_\kappa \in \mathcal{R}_p(\hat{\kappa})^d \quad \forall \kappa \in \mathcal{T}_h\},$$

where $\mathcal{R}$ is either $\mathcal{P}$ or $\mathcal{Q}$.

Let us consider the set $\mathcal{E}$ of all $(d-1)$-dimensional open faces—or, simply, *faces*—of all elements $\kappa \in \mathcal{T}_h$. Since hanging nodes are permitted (cf. Figure 2.1), $\mathcal{T}_h$ may be

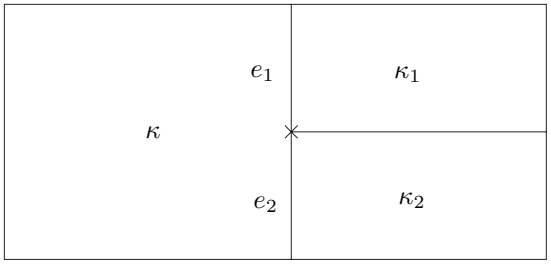


FIG. 2.1. *Hanging node* $\times$ *and faces* $e_1, e_2 \in \mathcal{E}_{\rm int}$.

irregular, and therefore $\mathcal{E}$ will be understood to contain the smallest common $(d-1)$-dimensional open faces of neighboring elements. Further, we denote by $\mathcal{E}_{\rm int}$ the set of all e in $\mathcal{E}$ that are contained in Ω, we let $\Gamma_{\rm int} = \{x \in \Omega : x \in e \text{ for some } e \in \mathcal{E}_{\rm int}\}$, and we introduce the set $\mathcal{E}_{\rm D}$ of $(d-1)$-dimensional boundary faces contained in the subset $\Gamma_{\rm D}$ of Γ. Implicit in these definitions is the assumption that $\mathcal{T}_h$ respects the decomposition of Γ in the sense that each $e \in \mathcal{E}$ that lies on Γ belongs to the interior of exactly one of $\Gamma_{\rm D}$ or $\Gamma_{\rm N}$. Given $e \in \mathcal{E}$, we define $h_e := \mathrm{diam}(e)$.

In the convergence analyses of the discontinuous Galerkin finite element approximations to the partial differential equations considered here, we shall adopt the following hypotheses on the family $\{\mathcal{T}_h\}_{h>0}$, the first of which controls the number of hanging nodes which any one element may have, the second is the standard quasi-uniformity assumption, while the third is a technical condition on the lowest polynomial degree which our analysis admits. **H2** and **H3** are required in order to deduce, by the use of inverse inequalities from bounds in a broken H^1 norm, that the element-wise gradient of the numerical solution lies in $\mathcal{Z}_\delta$. Finally, the fourth hypothesis is required for the definition of the *continuous reconstruction operator* in section 3. We assume that the assumptions **H1**–**H4** hold throughout the remainder of the article.

H1. The family of subdivisions $\{\mathcal{T}_h\}_{h>0}$ is *contact regular*; i.e., there exist positive constants c_d and c_e independent of h such that, for each $\kappa \in \mathcal{T}_h$,

$$\sharp\{\kappa' \in \mathcal{T}_h : \kappa' \neq \kappa,\ \mathcal{H}^{d-1}(\overline{\kappa'} \cap \overline{\kappa}) > 0\} \leq c_d, \quad \text{and} \quad c_e h_\kappa \leq h_e \quad \text{for every face } e \text{ of } \kappa.$$

H2. The family of subdivisions $\{\mathcal{T}_h\}_{h>0}$ is *quasi-uniform*; i.e., there exist positive constants c_0 and c_1, independent of h, such that for each $\kappa \in \mathcal{T}_h$ there exist open balls $B(x_0, c_0 h)$ and $B(x_1, c_1 h)$ such that $B(x_0, c_0 h) \subset \kappa \subset B(x_1, c_1 h)$.

H3. In the case of the elliptic problem (1.2) the polynomial degree $p > d/2$, and in the case of the hyperbolic problem (1.8) the polynomial degree $p > (d/2) + 1$ (viz. $p \geq 2$ for $d = 2, 3$, and $p \geq 3$ for $d = 2, 3$, respectively).

H4. The family of subdivisions $\{\mathcal{T}_h\}_{h>0}$ is *uniformly simplicially reducible*; i.e., for each $h > 0$ there exists a regular (no hanging nodes) simplicial mesh $\tilde{\mathcal{T}}_h$ such that the closure of each element in $\mathcal{T}_h$ is a union of closures of elements of $\tilde{\mathcal{T}}_h$ and such that there exist positive constants θ and C, independent of h, such that the smallest angle between any two edges in $\tilde{\mathcal{T}}_h$ is greater than or equal to θ and $h/\min_{\kappa \in \tilde{\mathcal{T}}_h} h_\kappa \leq C$.

Suppose that e is a $(d-1)$-dimensional open face of an element $\kappa \in \mathcal{T}_h$, and recall the notation introduced above: $h_\kappa = \mathrm{diam}(\kappa)$ and $h_e = \mathrm{diam}(e)$. The following *inverse inequalities* hold: There exists a positive constant C_3, independent of the

discretization parameter h, such that

$$\begin{aligned} \|\nabla w\|_{\mathrm{L}^\infty(\kappa)} &\leq \frac{C_3}{h_\kappa^{d/2}}\|\nabla w\|_{\mathrm{L}^2(\kappa)}, \quad \|\nabla w\|^2_{\mathrm{L}^2(\kappa)} \leq \frac{C_3}{h_\kappa^2}\|w\|^2_{\mathrm{L}^2(\kappa)}, \\ \|w\|^2_{\mathrm{L}^2(e)} &\leq \frac{C_3}{h_e}\|w\|^2_{\mathrm{L}^2(\kappa)}, \qquad \|\nabla w\|^2_{\mathrm{L}^2(e)} \leq \frac{C_3}{h_e}\|\nabla w\|^2_{\mathrm{L}^2(\kappa)}, \end{aligned} \tag{2.1}$$

for all $w \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$. In the case of the first two inverse inequalities C_3 depends only on the shape-regularity parameters of $\mathcal{T}_h$, while in the case of the other two inequalities it also depends on the contact-regularity parameter c_e. In fact, h_e in the last two inequalities can be replaced by h_κ at the expense of possibly altering the value of the constant C_3.

In the discussion that follows, we shall frequently need to consider the elementwise weak derivative (called the broken derivative) and the elementwise weak gradient (called the broken gradient) of a function that belongs to a broken Sobolev space. In order to simplify the presentation, our notation will not distinguish these from weak derivatives and weak gradients; the implied meaning of the notation will always be clear from the context. Thus, we adopt the following definition.

DEFINITION 1. *Let the* broken Sobolev space $\mathrm{H}^1(\Omega, \mathcal{T}_h)$ *be defined by*

$$\mathrm{H}^1(\Omega, \mathcal{T}_h) := \left\{ v \in \mathrm{L}^2(\Omega) : v|_\kappa \in \mathrm{H}^1(\kappa) \quad \forall \kappa \in \mathcal{T}_h \right\}.$$

For $v \in \mathrm{H}^1(\Omega, \mathcal{T}_h)$, we use ∇v to denote the piecewise weak gradient of v (relative to $\mathcal{T}_h$), i.e.,

$$\nabla v(x) := \nabla v|_\kappa(x) \qquad \forall x \in \kappa, \quad \forall \kappa \in \mathcal{T}_h,$$

where, on the right-hand side, $\nabla v|_\kappa$ denotes the weak gradient of $v|_\kappa \in \mathrm{H}^1(\kappa)$. The broken partial derivative $\partial_{x_j} v_i = \partial v_i / \partial x_j$ of $v \in \mathrm{H}^1(\Omega, \mathcal{T}_h)^d$ is the (i, j) component of its broken gradient ∇v.

For each $e \in \mathcal{E}_{\mathrm{int}}$ there exist indices i and j such that $i > j$ and κ_i and κ_j share the face e; we define the (element-numbering-dependent) *jump* of $v \in \mathrm{H}^1(\Omega, \mathcal{T}_h)^d$ across e and the *mean value* of v on e by

$$[\![v]\!]_e := v|_{\partial\kappa_i \cap e} - v|_{\partial\kappa_j \cap e} \quad \text{and} \quad \langle v \rangle_e := \tfrac{1}{2}\left(v|_{\partial\kappa_i \cap e} + v|_{\partial\kappa_j \cap e} \right),$$

respectively. If $e \in \mathcal{E}_{\mathrm{D}}$ is a face on the Dirichlet boundary, contained in the boundary $\partial\kappa$ of an element $\kappa \in \mathcal{T}_h$, it is also customary to define

$$[\![v]\!]_e := v|_{\partial\kappa \cap e} \quad \text{and} \quad \langle v \rangle_e := v|_{\partial\kappa \cap e}.$$

These definitions will enable us to condense our notation. For the sake of simplicity, the subscript e will be suppressed, and we shall simply write $[\![v]\!]$ and $\langle v \rangle$; the implied choice of e will be clear from the context. In addition, we associate with the face e the unit normal vector ν which points from κ_i to κ_j, $i > j$.

Suppose that σ is a positive, piecewise constant function defined on $\Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{int}}$ (to be defined below). We equip the space $\mathrm{H}^1(\Omega, \mathcal{T}_h)$ with the *broken Sobolev norm* $\|\cdot\|_{1,h}$ defined by

$$\|v\|_{1,h} := \left(\int_\Omega |\nabla v|^2 \, \mathrm{d}x + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma |[\![v]\!]|^2 \, \mathrm{d}s \right)^{1/2}.$$

For the definition of the discontinuous Galerkin method, we introduce the semilinear form

$$B(w,v) := \int_\Omega S(\nabla w) : \nabla v \,\mathrm{d}x - \int_{\Gamma_\mathrm{D}} S(\nabla w)\nu \cdot v \,\mathrm{d}s - \int_{\Gamma_\mathrm{int}} \langle S(\nabla w)\nu\rangle \cdot [\![v]\!] \,\mathrm{d}s$$
$$+ \int_{\Gamma_\mathrm{int}\cup\Gamma_\mathrm{D}} \sigma [\![w]\!]\cdot[\![v]\!]\,\mathrm{d}s, \qquad w \in \mathrm{C}^1(\overline{\Omega},\mathcal{T}_h)^d, \quad v \in \mathrm{H}^1(\Omega,\mathcal{T}_h)^d, \tag{2.2}$$

and the linear functional

$$\ell(v) := \int_\Omega f\cdot v\,\mathrm{d}x + \int_{\Gamma_\mathrm{D}} \sigma g_\mathrm{D}\cdot v\,\mathrm{d}s + \int_{\Gamma_\mathrm{N}} g_\mathrm{N}\cdot v\,\mathrm{d}s, \qquad v\in \mathrm{H}^1(\Omega,\mathcal{T}_h)^d. \tag{2.3}$$

Here $h^{-1}|_e = h_e^{-1}$ for all $e \subset \Gamma_\mathrm{int}\cup\Gamma_\mathrm{D}$. Let $\kappa \in \mathcal{T}_h$, and let e be a $(d-1)$-dimensional face of $\partial\kappa$. The function σ, referred to as the *discontinuity penalization parameter*, featured in $B(\cdot,\cdot)$ and $\ell(\cdot)$ above, is defined by

$$\sigma|_e := \sigma_e = \frac{\alpha}{h_e} \qquad \text{for } e \subset \Gamma_\mathrm{int}\cup\Gamma_\mathrm{D}. \tag{2.4}$$

Here α is a positive constant whose size will be fixed later.

The discontinuous Galerkin finite element approximation of problem (1.2), (1.3) is posed as follows: Find $u_\mathrm{DG} \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$B(u_\mathrm{DG},v) = \ell(v) \qquad \forall v \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}). \tag{2.5}$$

If the problem were linear, our discretization would correspond to the incomplete interior penalty method (see, for example, [9, 18]).

3. Broken Gårding inequality. The proofs of the broken versions of the Gårding inequalities (1.11) and (1.13) rely on the construction of a recovery operator, which connects each discontinuous piecewise polynomial function from $S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ to a continuous relative. Such an operator has been used previously in similar contexts, for example, by Karakashian and Pascal [14] for deriving residual-based a posteriori error estimates and by Brenner [4] for the proof of broken Korn inequalities.

Here we follow the construction used by Karakashian and Pascal [14], though we will slightly reformulate their result. By our hypothesis **H4**, the family of meshes $(\mathcal{T}_h)_{h>0}$ is uniformly simplicially reducible, meaning that, for each h there exists a regular simplicial mesh $\tilde{\mathcal{T}}_h$ which refines $\mathcal{T}_h$. For example, quasi-uniform families of 1-irregular meshes in two dimensions satisfy this property (cf. Figure 3.1 and Proposition 2 in [17]). Another important class are quasiuniform quadrilateral meshes obtained by hierarchical refinement (cf. Proposition 3 in [17]). For such families of meshes, we have the following result. For a proof we refer to Theorems 2.2 and 2.3 in [14] or section 7.1 in [17].

LEMMA 3.1. *There exists a constant C_r, independent of h, and a linear operator $\mathcal{R}\colon S^p(\Omega,\mathcal{T}_h,\mathbf{F}) \to \mathrm{H}^1_{\mathrm{D},0}(\Omega)^d$ such that, for all $u \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ and $k\in\{0,1\}$,*

$$\|\nabla^k(u-\mathcal{R}u)\|_{\mathrm{L}^2(\Omega)} \le C_r \int_{\Gamma_\mathrm{int}\cup\Gamma_\mathrm{D}} h^{1-2k}|[\![u]\!]|^2 \,\mathrm{d}s, \tag{3.1}$$

where $\nabla^0 = \mathrm{id}$ and $\nabla^1 = \nabla$.

Lemma 3.1 provides a link between discontinuous piecewise polynomial functions and functions in $\mathrm{H}^1_{\mathrm{D},0}(\Omega)^d$. Thus, to establish a broken Gårding inequality, we replace the test function v by its continuous representative $\mathcal{R}v$ and estimate the error

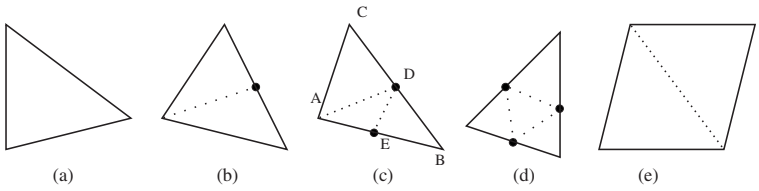


FIG. 3.1. *Refinement of triangular elements in the presence of hanging nodes in order to obtain the mesh $\tilde{\mathcal{T}}_h$ featured in hypothesis* **H4**.

committed in doing so in terms of the jumps of v. This procedure yields the following result.

LEMMA 3.2. *Let $u \in \mathrm{C}^1(\overline{\Omega})^d$ be such that the following Gårding inequality holds:*

$$a(\nabla u; v, v) \geq M_1 \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall v \in \mathrm{H}^1_{\mathrm{D},0}(\Omega)^d, \tag{3.2}$$

where $M_1 > 0$ and $M_0 \geq 0$. Assume furthermore that $\delta \leq M_1/(4L_\delta)$. Then, for all $\Phi \in \mathcal{Z}_\delta$ and $h \leq 1$, the following broken Gårding inequality holds:

$$\begin{aligned} a(\Phi; v, v) \geq{}& \tfrac{1}{2} M_1 \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} \\ & - C_1 \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} h^{-1} |[\![v]\!]|^2 \, \mathrm{d}s \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}), \end{aligned} \tag{3.3}$$

where $C_1 = C_1(M_0, M_1, K_\delta, C_r)$ is independent of h.

Proof. Note that the definition (1.6) of K_δ implies that

$$a(\nabla u; v, w) \leq K_\delta \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\nabla w\|_{\mathrm{L}^2(\Omega)} \qquad \forall v, w \in \mathrm{H}^1(\Omega, \mathcal{T}_h)^d.$$

Step 1. We begin by assuming that $\Phi = \nabla u$. In this case, we then have that

$$\begin{aligned} a(\nabla u; v, v) &= a(\nabla u; \mathcal{R}v, \mathcal{R}v) + a(\nabla u; v - \mathcal{R}v, v - \mathcal{R}v) + 2a(\nabla u; v - \mathcal{R}v, \mathcal{R}v) \\ &\geq M_1 \|\nabla \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)} - M_0 \|\mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)} - K_\delta \|\nabla v - \nabla \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)} \\ &\quad - 2K_\delta \|\nabla v - \nabla \mathcal{R}v\|_{\mathrm{L}^2(\Omega)} \|\nabla \mathcal{R}v\|_{\mathrm{L}^2(\Omega)} \\ &\geq M_1 \|\nabla v + (\nabla \mathcal{R}v - \nabla v)\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|\mathcal{R}v - v\|^2_{\mathrm{L}^2(\Omega)} \\ &\quad - 3K_\delta \|\nabla v - \nabla \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)} - 2K_\delta \|\nabla v - \nabla \mathcal{R}v\|_{\mathrm{L}^2(\Omega)} \|\nabla v\|_{\mathrm{L}^2(\Omega)}. \end{aligned}$$

Using the inverse triangle inequality in the first term on the right-hand side of the last inequality gives

$$\begin{aligned} a(\nabla u; v, v) \geq{}& M_1 (\|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - 2\|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\nabla v - \nabla \mathcal{R}v\|_{\mathrm{L}^2(\Omega)} + \|\nabla v - \nabla \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)}) \\ & - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|\mathcal{R}v - v\|^2_{\mathrm{L}^2(\Omega)} \\ & - 3K_\delta \|\nabla v - \nabla \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)} - 2K_\delta \|\nabla v - \nabla \mathcal{R}v\|_{\mathrm{L}^2(\Omega)} \|\nabla v\|_{\mathrm{L}^2(\Omega)}. \end{aligned}$$

We use the ε-inequality, $ab \leq \frac{\varepsilon}{2} a^2 + \frac{1}{2\varepsilon} b^2$, $\varepsilon > 0$, twice, with $\varepsilon = \varepsilon_1 > 0$ and $\varepsilon = \varepsilon_2 > 0$, to obtain

$$\begin{aligned} a(\nabla u; v, v) \geq{}& (M_1 - \varepsilon_1 M_1 - \varepsilon_2 K_\delta) \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|v - \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)} \\ & - (3K_\delta - M_1 + \varepsilon_1^{-1} M_1 + \varepsilon_2^{-1} K_\delta) \|\nabla v - \nabla \mathcal{R}v\|^2_{\mathrm{L}^2(\Omega)}. \end{aligned} \tag{3.4}$$

Step 2. Next, for each $\Phi \in \mathcal{Z}$, and $v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$, we can use the Lipschitz condition (1.7), which immediately implies that

$$a(\Phi; v, v) \geq a(\nabla u; v, v) - L_\delta \|\nabla u - \Phi\|_{\mathrm{L}^\infty(\Omega)} \|\nabla v\|^2_{\mathrm{L}^2(\Omega)}.$$

As, by hypothesis, $\|\nabla u - \Phi\|_{\mathrm{L}^\infty(\Omega)} \leq \delta$, with $\delta \leq M_1/(4L_\delta)$, it is straightforward to choose ε_1 and ε_2 in (3.4) and to apply (3.1) in order to obtain (3.3). ☐

4. Linearization. Before embarking on the analysis of the discontinuous Galerkin finite element method (2.5), we prove some auxiliary results about its linearization. We begin by noting that for any $\eta, \zeta \in \mathbb{R}^{d\times d}$ we have that

$$\begin{aligned} S_{i\alpha}(\eta) - S_{i\alpha}(\zeta) &= \sum_{j,\beta=1}^{d} (\eta_{j\beta} - \zeta_{j\beta}) \int_0^1 \frac{\partial S_{i\alpha}}{\partial \eta_{j\beta}}(\tau\eta + (1-\tau)\zeta)\,\mathrm{d}\tau \\ &= \sum_{j,\beta=1}^{d} (\eta_{j\beta} - \zeta_{j\beta}) \int_0^1 A_{i\alpha j\beta}(\tau\eta + (1-\tau)\zeta)\,\mathrm{d}\tau. \end{aligned} \tag{4.1}$$

Let $\mathrm{C}^1(\overline{\Omega}, \mathcal{T}_h)^d$ denote the space of all d-component piecewise C^1 functions, relative to the subdivision $\mathcal{T}_h$, defined on $\overline{\Omega}$. Taking (4.1) as a starting point, a straightforward computation shows that for any $w_i \in \mathrm{C}^1(\overline{\Omega}, \mathcal{T}_h)^d$, $i = 1, 2$, we have that

$$B(w_1, v) - B(w_2, v) = \int_0^1 \tilde{b}(w_2 + \tau(w_1 - w_2); w_1 - w_2, v)\,\mathrm{d}\tau \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}), \tag{4.2}$$

where, for $\varphi \in \mathrm{C}^1(\overline{\Omega}, \mathcal{T}_h)^d$, $\tilde{b}(\varphi; \cdot, \cdot)$ is the bilinear form defined by

$$\begin{aligned} \tilde{b}(\varphi; v, w) := &\int_\Omega \sum_{i,\alpha,j,\beta=1}^{d} A_{i\alpha j\beta}(\nabla\varphi) \frac{\partial w_i}{\partial x_\alpha} \frac{\partial v_j}{\partial x_\beta}\,\mathrm{d}x - \int_{\Gamma_{\mathrm{D}}} \sum_{i,\alpha,j,\beta=1}^{d} A_{i\alpha j\beta}(\nabla\varphi) w_i \nu_\alpha \frac{\partial v_j}{\partial x_\beta}\,\mathrm{d}s \\ &- \int_{\Gamma_{\mathrm{int}}} \sum_{i,\alpha,j,\beta=1}^{d} \left\langle A_{i\alpha j\beta}(\nabla\varphi)\nu_\alpha \frac{\partial v_j}{\partial x_\beta} \right\rangle [\![w_i]\!]\,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma [\![v]\!] \cdot [\![w]\!]\,\mathrm{d}s. \end{aligned}$$

In the next section, we shall use $\tilde{b}$ to perform a convergence analysis of the method (2.5), where the Gårding inequality (3.2) and the local Lipschitz continuity of $\tilde{b}$ w.r.t. its first argument are crucial. We prove these three results in the following three lemmas.

LEMMA 4.1. *Suppose that $u \in \mathrm{C}^1(\overline{\Omega})^d$ satisfies the Gårding inequality* (3.2). *Then there exists $\alpha_0 > 0$, independent of h, such that for all $\alpha \geq \alpha_0$, for all $h \in (0, 1]$, and for all $\varphi \in \mathrm{C}^1(\overline{\Omega}, \mathcal{T}_h)^d$, with $\nabla\varphi \in \mathcal{Z}_\delta$,*

$$\tilde{b}(\varphi; v, v) \geq \tilde{M}_1 \|v\|^2_{1,h} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}), \tag{4.3}$$

where $\tilde{M}_1 := \frac{1}{4}\min(1, M_1)$.

Proof. For $\nabla\varphi \in \mathcal{Z}_\delta$ fixed and $v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ we consider

$$\begin{aligned} \tilde{b}(\varphi; v, v) = &\int_\Omega \sum_{i,\alpha,j,\beta=1}^{d} A_{i\alpha j\beta}(\nabla\varphi) \frac{\partial v_i}{\partial x_\alpha} \frac{\partial v_j}{\partial x_\beta}\,\mathrm{d}x - \int_{\Gamma_{\mathrm{D}}} \sum_{i,\alpha,j,\beta=1}^{d} A_{i\alpha j\beta}(\nabla\varphi) v_i \nu_\alpha \frac{\partial v_j}{\partial x_\beta}\,\mathrm{d}s \\ &- \int_{\Gamma_{\mathrm{int}}} \sum_{i,\alpha,j,\beta=1}^{d} \left\langle A_{i\alpha j\beta}(\nabla\varphi)\nu_\alpha \frac{\partial v_j}{\partial x_\beta} \right\rangle [\![v_i]\!]\,\mathrm{d}s + \int_{\Gamma_{\mathrm{D}}} \sigma |v|^2\,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}} \sigma |[\![v]\!]|^2\,\mathrm{d}s \\ \equiv\ & T_1 + T_2 + T_3 + T_4 + T_5. \end{aligned}$$

Lemma 3.2 implies that

$$T_1 \geq \tfrac{1}{2} M_1 \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} - C_1 \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} h^{-1} |[\![v]\!]|^2 \, \mathrm{d}s,$$

where C_1 is independent of h and φ.

Next, we bound T_2. Since we assumed that $\nabla\varphi \in \mathcal{Z}_\delta$, it follows that $\nabla\varphi(x) \in \mathcal{M}_\delta$ for a.e. $x \in \Omega$. Hence,

$$\begin{aligned} |\mathrm{T}_2| &\leq K_\delta \int_{\Gamma_{\mathrm{D}}} \left(\sum_{i,\alpha,j,\beta=1}^{d} |v_i|^2 |\nu_\alpha|^2 \left| \frac{\partial v_j}{\partial x_\beta} \right|^2 \right)^{1/2} \mathrm{d}s \\ &\leq K_\delta \left(\int_{\Gamma_{\mathrm{D}}} \sigma^{-1} |\nabla v|^2 \, \mathrm{d}s \right)^{1/2} \left(\int_{\Gamma_{\mathrm{D}}} \sigma |v|^2 \, \mathrm{d}s \right)^{1/2}, \end{aligned}$$

where K_δ is defined in (1.6). Using the third of the inverse inequalities (2.1) and recalling the definition of the penalty parameter σ_e on $e \subset \Gamma_{\mathrm{D}}$, we have that

$$|\mathrm{T}_2| \leq K_\delta (C_3 \alpha^{-1} 2d)^{1/2} \left(\int_\Omega |\nabla v|^2 \, \mathrm{d}x \right)^{1/2} \left(\int_{\Gamma_{\mathrm{D}}} \sigma |v|^2 \, \mathrm{d}s \right)^{1/2},$$

where $2d$ stands for the maximum number of faces any one element may have on Γ_{D}.

Analogously,

$$|\mathrm{T}_3| \leq K_\delta \int_{\Gamma_{\mathrm{int}}} \langle |\nabla v| \rangle \, |[\![v]\!]| \, \mathrm{d}s \leq K_\delta \left(\int_{\Gamma_{\mathrm{int}}} \sigma^{-1} \langle |\nabla v| \rangle^2 \, \mathrm{d}s \right)^{1/2} \left(\int_{\Gamma_{\mathrm{int}}} \sigma [\![v]\!]^2 \, \mathrm{d}s \right)^{1/2}.$$

Let us note that

$$\int_{\Gamma_{\mathrm{int}}} \sigma^{-1} \langle |\nabla v| \rangle^2 \, \mathrm{d}s = \sum_{e \in \mathcal{E}_{\mathrm{int}}} \sigma_e^{-1} \int_e \langle |\nabla v| \rangle^2 \, \mathrm{d}s,$$

and, for $e \in \mathcal{E}_{\mathrm{int}}$, let κ and κ' be the two elements that share e. Then

$$\begin{aligned} \int_e \langle |\nabla v| \rangle^2 \, \mathrm{d}s &\leq \frac{1}{2} \int_e |\nabla v|_\kappa|^2 \, \mathrm{d}s + \frac{1}{2} \int_e |\nabla v|_{\kappa'}|^2 \, \mathrm{d}s \\ &\leq \frac{C_3}{2h_e} \int_\kappa |\nabla v|^2 \, \mathrm{d}x + \frac{C_3}{2h_e} \int_{\kappa'} |\nabla v|^2 \, \mathrm{d}x \\ &\leq \frac{C_3}{h_e} \max \left\{ \int_\kappa |\nabla v|^2 \, \mathrm{d}x, \int_{\kappa'} |\nabla v|^2 \, \mathrm{d}x \right\}. \end{aligned}$$

On recalling from the definition of σ that $\sigma_e = \alpha / h_e$ for $e \in \mathcal{E}_{\mathrm{int}}$, we have that

$$\sum_{e \in \mathcal{E}_{\mathrm{int}}} \sigma_e^{-1} \int_e \langle |\nabla v| \rangle^2 \, \mathrm{d}s \leq C_3 \alpha^{-1} \sum_{e \in \mathcal{E}_{\mathrm{int}}} \max_{\{\kappa \,:\, e \subset \partial\kappa\}} \int_\kappa |\nabla v|^2 \, \mathrm{d}x.$$

Thanks to our assumption **H1** of contact regularity, it follows that no element κ can have more than c_d faces, where c_d is a finite number independent of h. We have that

$$\sum_{e \in \mathcal{E}_{\mathrm{int}}} \sigma_e^{-1} \int_e \langle |\nabla v| \rangle^2 \, \mathrm{d}s \leq C_3 \alpha^{-1} c_d \sum_{\kappa \in \mathcal{T}_h} \int_\kappa |\nabla v|^2 \, \mathrm{d}x,$$

and therefore

$$|\mathrm{T}_3| \le K_\delta (C_3 \alpha^{-1} c_d)^{1/2} \left(\int_\Omega |\nabla v|^2 \,\mathrm{d}x \right)^{1/2} \left(\int_{\Gamma_{\mathrm{int}}} \sigma [\![v]\!]^2 \,\mathrm{d}s \right)^{1/2}. \tag{4.4}$$

Using the lower bound on T_1 and the upper bounds on T_2 and T_3, we thus deduce that

$$\begin{aligned} \int_0^1 \tilde{b}(\varphi; v, v) \,\mathrm{d}\tau \;\ge\; & \tfrac{1}{2} M_1 \|\nabla v\|^2_{\mathrm{L}^2(\Omega)} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\mathrm{D}}} \sigma |v|^2 \,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}} \sigma |[\![v]\!]|^2 \,\mathrm{d}s \\ & - K_\delta (C_3 \alpha^{-1} 2d)^{1/2} \left(\int_\Omega |\nabla v|^2 \,\mathrm{d}x \right)^{1/2} \left(\int_{\Gamma_{\mathrm{D}}} \sigma |v|^2 \,\mathrm{d}s \right)^{1/2} \\ & - K_\delta (C_3 \alpha^{-1} c_d)^{1/2} \left(\int_\Omega |\nabla v|^2 \,\mathrm{d}x \right)^{1/2} \left(\int_{\Gamma_{\mathrm{int}}} \sigma |[\![v]\!]|^2 \,\mathrm{d}s \right)^{1/2}. \end{aligned}$$

Applying Cauchy's inequality $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$ to the last two terms on the right-hand side and defining $C_d = c_d + 2d$, we have that

$$\int_0^1 \tilde{b}(\varphi; v, v) \,\mathrm{d}\tau \ge \frac{M_1}{2} \left(1 - \frac{K_\delta^2 C_3 C_d}{2 M_1 \alpha} \right) \int_\kappa |\nabla v|^2 \,\mathrm{d}x + \frac{1}{2} \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma [\![v]\!]^2 \,\mathrm{d}s - 2M_0 \|v\|^2_{\mathrm{L}^2}.$$

On selecting α such that $\alpha \ge K_\delta^2 M_1^{-1} C_3 C_d \equiv \alpha_0$, we deduce that, for all $h \in (0, 1]$, (4.3) holds. ∎

LEMMA 4.2. *For each $\delta > 0$ there exists a constant $\tilde{K}_\delta$ depending only on K_δ, C_3, and c_d such that, for all $\varphi \in \mathrm{C}^1(\Omega, \mathcal{T}_h)^d$ with $\nabla \varphi \in \mathcal{Z}_\delta$,*

$$\left| \tilde{b}(\varphi; v, w) \right| \le \tilde{K}_\delta \|v\|_{1,h} \|w\|_{1,h} \qquad \forall v, w \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).$$

Proof. By the definition of K_δ, for $\nabla \varphi \in \mathcal{Z}_\delta$, we have that

$$\sum_{i,\alpha,j,\beta=1}^{d} \int_\Omega |A_{i\alpha j\beta}(\nabla \varphi)| |\partial_{x_\alpha} v_i| |\partial_{x_\beta} w_j| \,\mathrm{d}x \le K_\delta \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\nabla w\|_{\mathrm{L}^2(\Omega)},$$

and, using also the fourth inverse inequality from (2.1),

$$\begin{aligned} \sum_{i,\alpha,j,\beta=1}^{d} \int_{\Gamma_{\mathrm{D}}} |A_{i\alpha j\beta}(\nabla \varphi)| |w_i| |\nu_\alpha| |\partial_{x_\beta} v_j| \,\mathrm{d}s &\le K_\delta \int_{\Gamma_{\mathrm{D}}} |w| |\nabla v| \,\mathrm{d}s \\ &\le K_\delta \left[\int_{\Gamma_{\mathrm{D}}} \sigma |w|^2 \,\mathrm{d}s \right]^{1/2} \left[\int_{\Gamma_{\mathrm{D}}} \sigma^{-1} |\nabla v|^2 \,\mathrm{d}s \right]^{1/2} \\ &\le K_\delta C(C_3, c_d) \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\sigma^{1/2} w\|_{\mathrm{L}^2(\Gamma_{\mathrm{D}})}. \end{aligned}$$

Using a similar argument, we can deduce that

$$\sum_{i,\alpha,j,\beta=1}^{d} \int_{\Gamma_{\mathrm{int}}} \left| \left\langle A_{i\alpha j\beta}(\nabla \varphi) \nu_\alpha \partial_{x_\beta} v \right\rangle \right| \left| [\![w_i]\!] \right| \,\mathrm{d}s \le K_\delta C(C_3, c_d) \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\sigma^{1/2} [\![w]\!]\|_{\mathrm{L}^2(\Gamma_{\mathrm{int}})}.$$

The result follows by inserting these three estimates into the definition of $\tilde{b}$. ∎

LEMMA 4.3. *For every $\delta > 0$ there exists a constant $\tilde{L}_\delta$, depending only on L_δ, C_3, and c_d such that, for all $\varphi, \psi \in \mathrm{C}^1(\Omega, \mathcal{T}_h)^d$ with $\nabla\varphi, \nabla\psi \in \mathcal{Z}_\delta$,*

$$\left|\tilde{b}(\varphi; v, w) - \tilde{b}(\psi; v, w)\right| \leq \tilde{L}_\delta \|\nabla\varphi - \nabla\psi\|_{\mathrm{L}^\infty(\Omega)} \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|w\|_{1,h} \quad \forall v, w \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).$$

Proof. The proof follows precisely that of Lemma 4.2. Using the fact that the integrands in $\tilde{b}$ are linear in the tensor, we can replace $A_{i\alpha j\beta}(\nabla\varphi)$ by $(A_{i\alpha j\beta}(\nabla\varphi) - A_{i\alpha j\beta}(\nabla\psi))$ and use the Lipschitz condition (1.7) instead of the bound (1.6). Furthermore, the penalty terms cancel each other out, which gives $\|\nabla v\|_{\mathrm{L}^2(\Omega)}$ instead of $\|v\|_{1,h}$; see [17] for additional details. ∎

5. The elliptic case. Throughout this section, we assume that $u \in \mathrm{H}^{m+1}(\Omega)^d$, with $m > d/2$, is a solution of (1.2), (1.3), satisfying the Gårding inequality (1.11); in our analysis of the discontinuous Galerkin finite element approximation to the corresponding hyperbolic problem (1.8), we shall suppose that the weaker inequality (3.2) holds.

The convergence analysis will be based on Banach's fixed point theorem. We begin by constructing a nonlinear mapping whose unique fixed point in a neighborhood of u is the numerical solution u_{DG}. For this purpose, let $\Pi_h u$ denote the finite element interpolant, from $S^p(\Omega, \mathcal{T}_h, \mathbf{F})$, of the analytical solution u, defined by $(\Pi_h u)|_\kappa := \Pi_p^{\hat{\kappa}}(u|_\kappa \circ F_\kappa) \in \mathcal{R}_p(\hat{\kappa})$, where $\Pi_p^{\hat{\kappa}}(u|_\kappa \circ F_\kappa)$ is the classical finite element interpolant of $u|_\kappa \circ F_\kappa$ from $\mathcal{R}_p(\hat{\kappa})$. We can take $w_1 = u_{\mathrm{DG}}$ and $w_2 = \Pi_h u$ in the identity (4.2) above, which gives

$$B(u_{\mathrm{DG}}, v) - B(\Pi_h u, v) = \int_0^1 \tilde{b}(\Pi_h u + \tau(u_{\mathrm{DG}} - \Pi_h u); u_{\mathrm{DG}} - \Pi_h u, v)\, \mathrm{d}\tau \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).$$

Let us write

$$u - u_{\mathrm{DG}} = (u - \Pi_h u) - (u_{\mathrm{DG}} - \Pi_h u) \equiv \eta - \xi.$$

Note that since $u \in \mathrm{C}^1(\overline{\Omega})^d \cap \mathrm{H}^2(\Omega)^d$, we have that $B(u, v) = \ell(v)$ for all v in $S^p(\Omega, \mathcal{T}_h, \mathbf{F})$. Hence,

$$B(u_{\mathrm{DG}}, v) - B(\Pi_h u, v) = \ell(v) - B(\Pi_h u, v) = B(u, v) - B(\Pi_h u, v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}),$$

and therefore, for all $v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$,

$$\int_0^1 \tilde{b}(\Pi_h u + \tau(u_{\mathrm{DG}} - \Pi_h u); u_{\mathrm{DG}} - \Pi_h u, v)\, \mathrm{d}\tau = \int_0^1 \tilde{b}(\Pi_h u + \tau(u - \Pi_h u); u - \Pi_h u, v)\, \mathrm{d}\tau. \tag{5.1}$$

Upon defining the bilinear form $\tilde{B}(\varphi; \cdot, \cdot)$ by

$$\tilde{B}(\varphi; v, w) := \int_0^1 \tilde{b}(\Pi_h u + \tau(\varphi - \Pi_h u); v, w)\, \mathrm{d}\tau,$$

we may rewrite (5.1) as

$$\tilde{B}(u_{\mathrm{DG}}; u_{\mathrm{DG}} - \Pi_h u, v) = \tilde{B}(u; u - \Pi_h u, v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{5.2}$$

Lemmas 4.1–4.3 immediately imply that

$$\tilde{B}(\varphi; v, v) \geq \tilde{M}_1 \|v\|_{1,h}^2, \tag{5.3}$$

$$\left|\tilde{B}(\varphi; v, w)\right| \leq \tilde{K}_\delta \|v\|_{1,h} \|w\|_{1,h}, \quad \text{and} \tag{5.4}$$

$$\left|\tilde{B}(\varphi; v, w) - \tilde{B}(\psi; v, w)\right| \leq \tilde{L}_\delta \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|w\|_{1,h} \tag{5.5}$$

for all $v, w \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ and $\varphi, \psi \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ such that $\nabla\varphi, \nabla\psi \in \mathcal{Z}_\delta$.

Let us recall our hypotheses that $u \in \mathrm{H}^{m+1}(\Omega)^d$, with $m > d/2$, and that the polynomial degree $p > d/2$. Let $d/2 < r \leq \min(m, p)$, and define the following subset of the broken Sobolev space $\mathrm{H}^1(\Omega, \mathcal{T}_h)^d$:

$$\mathcal{J} := \left\{ \varphi \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}) \,:\, \|\varphi - \Pi_h u\|_{1,h} \leq C_* h^r \|u\|_{\mathrm{H}^{r+1}(\Omega)} \right\},$$

where C_* is a fixed positive constant, independent of h, whose actual value will be fixed below. We note that, since $\Pi_h u \in \mathcal{J}$, the set $\mathcal{J}$ is nonempty. Further, $\mathcal{J}$ is a closed, convex subset of $\mathrm{H}^1(\Omega, \mathcal{T}_h)^d$ in the topology induced by the norm $\|\cdot\|_{1,h}$. Finally, we note that for each $v \in \mathcal{J}$, using the first inverse inequality in (2.1) and the approximation properties of Π_h (see, for example, [6]), we have that

$$\begin{aligned}
\|\nabla v - \nabla u\|_{\mathrm{L}^\infty(\Omega)} &\leq \|\nabla v - \nabla \Pi_h u\|_{\mathrm{L}^\infty(\Omega)} + \|\nabla \Pi_h u - \nabla u\|_{\mathrm{L}^\infty(\Omega)} \\
&\leq C_* C_3 h^{r-d/2} \|u\|_{\mathrm{H}^{r+1}(\Omega)} + \|\nabla \Pi_h u - \nabla u\|_{\mathrm{L}^\infty(\Omega)} \\
&\leq C_* C_3 h^{r-d/2} \|u\|_{\mathrm{H}^{r+1}(\Omega)} + C_5 h^{r-d/2} \|u\|_{\mathrm{H}^{r+1}(\Omega)}.
\end{aligned}$$

Since $r > d/2$ by hypothesis, given $\delta > 0$, there exists $h_0 \in (0, 1]$ such that, for all $h \in (0, h_0]$,

$$\varphi \in \mathcal{J} \;\Rightarrow\; \nabla \varphi \in \mathcal{Z}_\delta. \tag{5.6}$$

Motivated by the form of (5.2), we define the fixed point mapping $\mathcal{N}$ on $\mathcal{J}$ as follows. Given $\varphi \in \mathcal{J}$, we denote by $\mathcal{N}(\varphi) \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ the solution to the following linear variational problem: Find $\mathcal{N}(\varphi) \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$\tilde{B}(\varphi; \mathcal{N}(\varphi) - \Pi_h u, v) = \tilde{B}(u; u - \Pi_h u, v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{5.7}$$

Equivalently, we can restate this as follows: Find $\mathcal{N}(\varphi) \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$\tilde{B}(\varphi; \mathcal{N}(\varphi), v) = \tilde{B}(u; u - \Pi_h u, v) + \tilde{B}(\varphi; \Pi_h u, v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).$$

Since $S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ is a finite-dimensional linear space, the existence and uniqueness of a solution $\mathcal{N}(\varphi) \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ to problem (5.7) follows immediately from (5.3).

To prove that $\mathcal{N}$ maps $\mathcal{J}$ into itself, we test (5.7) with $v = \mathcal{N}(\varphi) - \Pi_h u$ and use (5.3) and (5.4) to obtain

$$\begin{aligned}
\tilde{M}_1 \|\mathcal{N}(\varphi) - \Pi_h u\|_{1,h}^2 &\leq \tilde{B}(\varphi; \mathcal{N}(\varphi) - \Pi_h u, \mathcal{N}(\varphi) - \Pi_h u) \\
&= \tilde{B}(u; u - \Pi_h u, \mathcal{N}(\varphi) - \Pi_h u) \\
&\leq \tilde{K}_\delta \|u - \Pi_h u\|_{1,h} \|\mathcal{N}(\varphi) - \Pi_h u\|_{1,h}.
\end{aligned}$$

Using the approximation properties of the projector $\Pi_h u$, we deduce that

$$\|\mathcal{N}(\varphi) - \Pi_h u\|_{1,h} \leq \tilde{M}_1^{-1} \tilde{K}_\delta C_6 h^r \|u\|_{\mathrm{H}^{r+1}(\Omega)}.$$

If we define $C_* = \tilde{M}_1^{-1} \tilde{C}_\delta C_6$, then $\mathcal{N}$ indeed maps $\mathcal{J}$ into itself. Note that, while h_0 depends on C_*, the constant C_* does not depend on h_0, and hence this seemingly implicit construction of C_* is correct.

It remains to show that $\mathcal{N}$ is a contraction of $\mathcal{J}$ in the norm $\|\cdot\|_{1,h}$. To do so, let us suppose that φ and ψ belong to $\mathcal{J}$. Then

$$\begin{aligned}
&\tilde{B}(\varphi; \mathcal{N}(\varphi) - \Pi_h u, v) = \tilde{B}(u; u - \Pi_h u, v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}) \;\text{ and} \\
&\tilde{B}(\psi; \mathcal{N}(\psi) - \Pi_h u, v) = \tilde{B}(u; u - \Pi_h u, v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).
\end{aligned}$$

Upon subtracting the second line from the first, choosing $v = \mathcal{N}(\varphi) - \mathcal{N}(\psi)$, and using (5.3) and (5.5), we deduce that

$$\begin{aligned}\tilde{M}_1\|\mathcal{N}(\varphi)-\mathcal{N}(\psi)\|_{1,h}^2 &\le \tilde{B}(\varphi;\mathcal{N}(\varphi)-\mathcal{N}(\psi),\mathcal{N}(\varphi)-\mathcal{N}(\psi))\\ &= \tilde{B}(\psi;\mathcal{N}(\psi)-\Pi_h u,\mathcal{N}(\varphi)-\mathcal{N}(\psi)) - \tilde{B}(\varphi;\mathcal{N}(\psi)-\Pi_h u,\mathcal{N}(\varphi)-\mathcal{N}(\psi))\\ &\le \tilde{L}_\delta\|\nabla\psi-\nabla\varphi\|_{\mathrm{L}^\infty(\Omega)}\|\mathcal{N}(\psi)-\Pi_h u\|_{1,h}\|\mathcal{N}(\varphi)-\mathcal{N}(\psi)\|_{1,h}.\end{aligned}$$

Using the first inverse inequality in (2.1), and the fact that $\mathcal{N}(\psi) \in \mathcal{J}$, we have that

$$\|\mathcal{N}(\varphi)-\mathcal{N}(\psi)\|_{1,h} \le \tilde{M}_1^{-1}\tilde{L}_\delta C_3 C_* h^{r-d/2}\|\nabla\varphi-\nabla\psi\|_{\mathrm{L}^2(\Omega)} \le c(h)\|\varphi-\psi\|_{1,h},$$

where $c(h) = \tilde{M}^{-1}\tilde{L}_\delta C_3 C_* h^{r-d/2}$. Since $r > d/2$ by hypothesis **H3**, there exists a positive constant $h_1 \in (0,1]$ such that $c(h) < 1$. Thus, for $h \in (0, \min(h_0, h_1)]$, the mapping $\mathcal{N}$ is a contraction in the norm $\|\cdot\|_{1,h}$ of the closed set $\mathcal{J}$. By Banach's fixed point theorem, $\mathcal{N}$ has a unique fixed point u_{DG} in $\mathcal{J}$; in particular, by the definition of the set $\mathcal{J}$, the finite element approximation u_{DG} of u satisfies the bound

$$\|u_{\mathrm{DG}} - \Pi_h u\|_{1,h} \le C_* h^r \|u\|_{\mathrm{H}^{r+1}(\Omega)}, \qquad d/2 < r \le \min(m,p); \tag{5.8}$$

furthermore, $\nabla u_{\mathrm{DG}} \in \mathcal{Z}_\delta$ for all $h \in (0, \min(h_0, h_1)]$.

Let us write $a \lesssim b$ to express the fact that, for real numbers a and b, there exists a positive constant C, depending on the analytical solution u but *independent* of the discretization parameter h, such that $a \le Cb$ for all h in a closed subinterval of $[0,1]$ containing 0. We shall write $a \approx b$ if and only if $a \lesssim b$ and $b \lesssim a$. Since

$$\|u - \Pi_h u\|_{1,h} \le C_6 h^r \|u\|_{\mathrm{H}^{r+1}(\Omega)}, \qquad d/2 < r \le \min(m,p), \tag{5.9}$$

we deduce from (5.8) and (5.9) via the triangle inequality that, for all $h \in (0, \min(h_0, h_1)]$,

$$\|u - u_{\mathrm{DG}}\|_{1,h} \lesssim h^r \|u\|_{\mathrm{H}^{r+1}(\Omega)}, \qquad d/2 < r \le \min(m,p), \tag{5.10}$$

which is the required optimal bound on the error in the discontinuous Galerkin finite element method.

6. The hyperbolic problem. Now consider the hyperbolic problem

$$\partial_t^2 u_i - \sum_{\alpha=1}^{d} \partial_{x_\alpha}(S_{i\alpha}(\nabla u)) = f_i(t,x), \qquad i = 1,\dots,d, \qquad t \in (0,T], \quad x \in \Omega,$$

subject to the pair of initial conditions $u(0,x) = u_0(x)$, $\partial_t u(0,x) = u_1(x)$, $x \in \Omega$, where $u_0, u_1 \in \mathrm{H}^{m+1}(\Omega)^d$, and analogous boundary conditions as in the case of the static problem considered earlier; that is,

$$u(t,x) = g_{\mathrm{D}}(t,x) \quad \text{on } (0,T]\times\Gamma_{\mathrm{D}} \qquad \text{and} \qquad S(\nabla u(t,x))\nu = g_{\mathrm{N}}(t,x) \quad \text{on } (0,T]\times\Gamma_{\mathrm{N}}. \tag{6.1}$$

Since g_{D} and g_{N} now depend on t, so does the linear functional, which we denote by $\ell(t,\cdot)$ and is otherwise defined as in (2.3).

We refer to [8, 5] for theoretical results concerning the existence of a unique local (in time) solution to (6.1), subject to the given initial conditions, in the special case of a homogeneous Dirichlet boundary condition on Γ.

It will be assumed throughout that

$$u \in \mathrm{C}^2([0,T];\mathrm{H}^{m+1}(\Omega)^d), \qquad m > (d/2) + 1.$$

For simplicity, when there is no danger of confusion, we shall suppress the x-dependence in our notation and write $u(t)$, $v(t)$, etc., instead of $u(t,x)$, $v(t,x)$, etc.; we shall, on occasion, suppress both the x- and the t-dependence and write u, v, and so on. We shall further suppose that, for all $t \in [0,T]$, $u(t,\cdot)$ satisfies the Gårding inequality (3.2) for some $M_0 \geq 0$ and $M_1 > 0$, both independent of t. If one assumes the uniform monotonicity condition (1.10), then this is always true with $M_0 = 0$. If, on the other hand, one adopts the (considerably weaker) strong Legendre–Hadamard condition (1.12) and $\Gamma_{\rm D} = \Gamma$, then the Gårding inequality (3.2) holds with $M_1 > 0$ for some $M_0 \geq 0$ which may depend on $u(t)$; however, since $u \in {\rm C}^2([0,T] \times \overline{\Omega})$ by the Sobolev embedding theorem, M_0 can be chosen independent of t; cf. Theorem 6.5.1 on p. 253 of Morrey [16].

As in the elliptic case, let $\mathcal{M}_\delta$ be defined by

$$\mathcal{M}_\delta := {\rm conv}\big\{\eta \in \mathbb{R}^{d\times d} : \inf_{x\in\Omega, t\in[0,T]} |\eta - \nabla u(t,x)| \leq \delta\big\},$$

and define the constants K_δ and L_δ by the formulas (1.6) and (1.7). The set $\mathcal{Z}_\delta$ is now given by

$$\mathcal{Z}_\delta := \big\{\Phi \in {\rm C}_{\rm pw}(\overline{\Omega})^{d\times d} : \min_{t\in[0,T]} \|\Phi - \nabla u(t)\|_{{\rm L}^\infty(\Omega)} \leq \delta\big\}.$$

Let us consider, for $t \in [0,T]$ and $p > (d/2)+1$, the (semidiscrete) discontinuous Galerkin finite element approximation $u_{\rm DG}(t,\cdot) \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ to $u(t,\cdot)$ such that

$$(\ddot{u}_{\rm DG}, v) + B(u_{\rm DG}, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{u}_{\rm DG}]\!]\cdot[\![v]\!]\,{\rm d}s = \ell(t,v) + \int_{\Gamma_{\rm D}} \sigma\dot{g}_{\rm DG}\cdot v\,{\rm d}s \tag{6.2}$$

for all for $v \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ and all $t \in (0,T]$, and

$$u_{\rm DG}(0,x) = u^0_{\rm DG}(x), \qquad \dot{u}_{\rm DG}(0,x) = u^1_{\rm DG}(x), \qquad x \in \Omega,$$

with $u^0_{\rm DG}$ and $u^1_{\rm DG}$ in $S^p(\Omega,\mathcal{T}_h,\mathbf{F})$.

We highlight the presence of the last term on the left-hand side and the second term on the right-hand side of (6.2) which did not feature in the definition of our discontinuous Galerkin approximation of the elliptic problem considered in the earlier sections. The inclusion of these terms does not affect the consistency of the method. On the other hand, they play a crucial role in ensuring the validity of energy estimates in sufficiently strong norms. In order to highlight this point further, note that, in an energy analysis of the discontinuous Galerkin approximation (2.5) to the elliptic problem (1.2), (1.3), the natural choice of test function is $v = u_{\rm DG}$, while in the case of (6.2) it is $v = \dot{u}_{\rm DG}$, which, in turn, motivates the inclusion of the additional terms in (6.2) compared to the elliptic case.

Let $M_0 \geq 0$ be the constant from (3.2). We denote by $W(t) \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ the nonlinear projection of $u(t)$ defined by

$$B(W(t),v) + 2M_0(W(t),v) = B(u(t),v) + 2M_0(u(t),v) \qquad \forall v \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}), \qquad 0 \leq t \leq T,$$

and we select $u^0_{\rm DG}$ and $u^1_{\rm DG}$ in $S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$\|u^0_{\rm DG} - W(0)\|_{1,h} + \|u^1_{\rm DG} - \dot{W}(0)\|_{{\rm L}^2(\Omega)} \lesssim h^r, \qquad (d/2)+1 < r \leq \min(m,p).$$

The existence, uniqueness, approximation properties, and differentiability with respect to t of $W(t)$ are established in Appendix A, in Lemma A.1. For the sake of simplicity of presentation, we choose $u^0_{\rm DG} = W(0)$ and $u^1_{\rm DG} = \dot{W}(0)$ here. By using an argument based on Banach's fixed point theorem, similar to the one presented in the previous section, and stimulated by the ideas in [15], we will show the existence and uniqueness of $u_{\rm DG}$. We shall also show that $u_{\rm DG}$ converges to the analytical solution u with optimal order as the spatial discretization parameter h converges to 0.

6.1. Definition of the fixed point map. We decompose

$$u - u_{\rm DG} = (u - W) - (u_{\rm DG} - W) \equiv \eta - \xi.$$

Then, with our choice of the numerical initial conditions $u^0_{\rm DG}$ and $u^1_{\rm DG}$, we have $\xi(0) = 0$ and $\dot{\xi}(0) = 0$. Hence,

$$\begin{aligned}(\ddot{\xi}, v) + B(u_{\rm DG}, v) - B(W, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\xi}]\!]\cdot[\![v]\!]\,{\rm d}s \\ = (\ddot{\eta}, v) - 2M_0(\eta, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\eta}]\!]\cdot[\![v]\!]\,{\rm d}s \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).\end{aligned}$$

Upon linearization of the term $B(u_{\rm DG}, v) - B(W, v)$, in terms of our earlier notation, we have that

$$\begin{aligned}(\ddot{\xi}, v) + \int_0^1 \tilde{b}(W + \tau(u_{\rm DG} - W); u_{\rm DG} - W, v)\,{\rm d}\tau + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\xi}]\!]\cdot[\![v]\!]\,{\rm d}s \\ = (\ddot{\eta}, v) - 2M_0(\eta, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\eta}]\!]\cdot[\![v]\!]\,{\rm d}s \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}).\end{aligned} \tag{6.3}$$

As in the case of the elliptic problem, we can simplify the notation considerably by defining the bilinear form $\tilde{B}(t, \varphi; \cdot, \cdot)$ by

$$\tilde{B}(t, \varphi; v, w) := \int_0^1 \tilde{b}(W(t) + \tau(\varphi - W(t)); v, w)\,{\rm d}t$$

and the linear functional $\rho(t; \cdot)$ by

$$\rho(t; v) := (\ddot{\eta}, v) - 2M_0(\eta, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\eta}]\!]\cdot[\![v]\!]\,{\rm d}s,$$

which allows us to rewrite (6.3) as

$$(\ddot{\xi}, v) + \tilde{B}(t, u_{\rm DG}(t); \xi, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\xi}]\!]\cdot[\![v]\!]\,{\rm d}s = \rho(t; v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{6.4}$$

We consider the set $\mathcal{J} \subset {\rm C}^1([0, T]; S^p(\Omega, \mathcal{T}_h, \mathbf{F})) \equiv Y$ defined by

$$\begin{aligned}\mathcal{J} :=& \big\{\psi \in Y \,:\, \|\psi - W\|_Y \\ :=& \max_{t\in[0,T]} \Big(\|\psi(t) - W(t)\|_{1,h} + \|\dot{\psi}(t) - \dot{W}(t)\|_{{\rm L}^2(\Omega)}\Big) \le C_*(u)h^r\big\},\end{aligned}$$

where $C_*(u)$ is a positive constant and $(d/2) + 1 < r \le \min(m, p)$. As in the elliptic case, by the first inverse inequality in (2.1), there exists $h_0 > 0$ such that, for all $h \in (0, h_0]$,

$$\psi \in \mathcal{J} \;\Rightarrow\; \nabla\psi(t) \in \mathcal{Z}_\delta \;\forall\; t \in [0, T]. \tag{6.5}$$

In addition, $\mathcal{J}$ is a closed, convex subset of Y.

Now, motivated by the form of (6.4) and the definition of ξ, similarly as in the case of the elliptic problem, we are led to the following definition of the fixed point map $\mathcal{N}$ on $\mathcal{J}$: If $\varphi \in \mathcal{J}$, the image $u_\varphi = \mathcal{N}(\varphi) \in \mathrm{C}^2([0,T]; S^p(\Omega, \mathcal{T}_h, \mathbf{F}))$ is defined as the solution to the following linear problem:

$$(\ddot{u}_\varphi - \ddot{W}, v) + \tilde{B}(t, \varphi(t); u_\varphi - W, v) + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma [\![\dot{u}_\varphi - \dot{W}]\!] \cdot [\![v]\!] \, \mathrm{d}s = \rho(t; v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}), \tag{6.6}$$

with $u_\varphi(0) = u^0_{\mathrm{DG}}$, $\dot{u}_\varphi(0) = u^1_{\mathrm{DG}}$. Clearly, this variational form can be rewritten as an explicit linear ordinary differential equation for u_φ, and hence $\mathcal{N}$ is well-defined. Our objective now is to show, via Banach's fixed point theorem, that the nonlinear mapping $\varphi \in \mathcal{J} \mapsto \mathcal{N}(\varphi)$ has a unique fixed point $u_{\mathrm{DG}} \in \mathcal{J}$.

6.2. Auxiliary results. In the analysis of the linear problem (6.6), it will be crucial to replace a term of the form

$$\tilde{B}(t, \varphi(t); \xi(t), \dot{\xi}(t))$$

by a total derivative. Since $\tilde{B}(t, \varphi(t); \cdot, \cdot)$ is not symmetric in its last two arguments, we split $\tilde{B}$ into a symmetric term and a remainder which can be controlled:

$$\tilde{B}(t, \varphi(t); v, w) = \tilde{B}^{(S)}(t, \varphi(t); v, w) + \tilde{B}^{(A)}(t, \varphi(t); v, w), \tag{6.7}$$

where

$$\tilde{B}^{(S)}(t, \varphi(t); v, w) := \int_0^1 \int_\Omega \sum_{i,\alpha,j,\beta=1}^{d} A^\tau_{i\alpha j\beta} \partial_{x_\alpha} w_i \partial_{x_\beta} v_j \, \mathrm{d}x \, \mathrm{d}\tau + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma [\![v]\!] \cdot [\![w]\!] \, \mathrm{d}s,$$

and

$$\tilde{B}^{(A)}(t, \varphi(t); v, w) := - \int_0^1 \sum_{i,\alpha,j,\beta=1}^{d} \left[\int_{\Gamma_{\mathrm{int}}} \left\langle A^\tau_{i\alpha j\beta} \nu_\alpha \partial_{x_\beta} v_j \right\rangle [\![w_i]\!] \, \mathrm{d}s + \int_{\Gamma_{\mathrm{D}}} A^\tau_{i\alpha j\beta} \nu_\alpha w_i \partial_{x_\beta} v_j \, \mathrm{d}s \right] \mathrm{d}\tau,$$

where $A^\tau_{i\alpha j\beta} := A_{i\alpha j\beta}(\nabla W(t) + \tau(\nabla \varphi(t) - \nabla W(t)))$. Note that $\tilde{B}^{(A)}(t, \varphi; \cdot, \cdot)$ is not skew-symmetric but asymmetric, i.e., simply, *not symmetric*.

Following the proof of Lemma 4.1 closely, we obtain for all $\varphi \in \mathcal{J}$ and for all $\alpha \geq \alpha_0$, where α_0 is as in Lemma 4.1,

$$\tilde{B}^{(S)}(t, \varphi(t); v, v) \geq \tfrac{1}{2} M_1 \|v\|^2_{1,h} - 2M_0 \|v\|^2_{\mathrm{L}^2(\Omega)} \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}), \tag{6.8}$$

and

$$\left| \tilde{B}^{(A)}(t, \varphi(t); v, w) \right| \lesssim \|\nabla v\|_{\mathrm{L}^2(\Omega)} \left(\int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma |[\![w]\!]|^2 \, \mathrm{d}s \right)^{1/2} \qquad \forall v, w \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{6.9}$$

In addition, we shall require an estimate on the expression

$$\tilde{B}^{(S)}_t(t, \varphi(t); v, w) := \int_0^1 \int_\Omega \sum_{i,\alpha,j,\beta=1}^{d} \left[\frac{\mathrm{d}}{\mathrm{d}t} A^\tau_{i\alpha j\beta} \right] \partial_{x_\alpha} w_i \partial_{x_\beta} v_j \, \mathrm{d}x \, \mathrm{d}\tau, \qquad v, w \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}),$$

where $A^\tau_{i\alpha j\beta}$ is as defined above. Upon setting

$$K'_\delta := \operatorname*{ess.sup}_{x\in\Omega,\tau\in[0,1]} \left(\sum_{i,\alpha,j,\beta=1}^{d} \left| \frac{\mathrm{d}}{\mathrm{d}t} A^\tau_{i\alpha j\beta} \right|^2 \right)^{1/2},$$

we deduce that

$$\left|\tilde{B}^{(S)}_t(t,\varphi(t);v,w)\right| \le K'_\delta \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\nabla w\|_{\mathrm{L}^2(\Omega)}.$$

To estimate K'_δ, consider

$$\begin{aligned} K'_\delta &= \operatorname*{ess.sup}_{x\in\Omega,\tau\in[0,1]} \left(\sum_{i,\alpha,j,\beta=1}^{d} \left|\nabla A_{i\alpha j\beta}(\nabla\psi(t,x))\right|^2 |\nabla\dot{W}(t) + \tau(\nabla\dot{\varphi}(t) - \nabla\dot{W}(t))|^2 \right)^{1/2} \\ &\le L_\delta \big(\|\nabla\dot{W}(t)\|_{\mathrm{L}^\infty(\Omega)} + \|\nabla\dot{\varphi}(t) - \nabla\dot{W}(t)\|_{\mathrm{L}^\infty(\Omega)} \big). \end{aligned}$$

As $\varphi \in \mathcal{J}$, the first of the inverse inequalities (2.1), the bound (A.2), and the definition of the set $\mathcal{J}$ yield

$$K'_\delta \lesssim \|\nabla\dot{W}(t)\|_{\mathrm{L}^\infty(\Omega)} + h^{-d/2} \|\nabla\dot{\varphi}(t) - \nabla\dot{W}(t)\|_{\mathrm{L}^2(\Omega)} \lesssim 1 + h^{r-d/2}.$$

Combining these estimates and recalling that, by hypothesis $r > (d/2) + 1$ and, a fortiori, $r > d/2$, we obtain for all $\varphi \in \mathcal{J}$, for all $t \in [0,T]$, and for all $h \in (0,1]$

$$\left|\tilde{B}^{(S)}_t(t,\varphi(t);v,w)\right| \lesssim \|\nabla v\|_{\mathrm{L}^2(\Omega)} \|\nabla w\|_{\mathrm{L}^2(\Omega)} \qquad \forall v,w \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}). \tag{6.10}$$

Finally, we shall require an estimate on the right-hand side $\rho(t;v)$ in (6.4). A straightforward computation gives

$$\begin{aligned} |\rho(t,v)| &\le \left(2\|\ddot{\eta}\|^2_{\mathrm{L}^2(\Omega)} + 8M_0^2 \|\eta\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma |[\![\dot{\eta}]\!]|^2 \,\mathrm{d}s \right)^{1/2} \\ &\qquad \left(\|v\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma |[\![v]\!]|^2 \,\mathrm{d}s \right)^{1/2} \\ &\lesssim h^r \left(\|v\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma |[\![v]\!]|^2 \,\mathrm{d}s \right)^{1/2}, \end{aligned} \tag{6.11}$$

where we used (A.2) and (A.7) to bound the different norms of η.

6.3. Convergence analysis. For the sake of notational simplicity, we define

$$\xi_\varphi = u_\varphi - W.$$

Testing (6.6) with $v = \dot{\xi}_\varphi$, and using the decomposition (6.7), we deduce that

$$(\ddot{\xi}_\varphi, \dot{\xi}_\varphi) + \tilde{B}^{(S)}(t,\varphi(t);\xi_\varphi,\dot{\xi}_\varphi) + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma |[\![\dot{\xi}_\varphi]\!]|^2 \,\mathrm{d}s = \rho(t;\dot{\xi}_\varphi) - \tilde{B}^{(A)}(t,\varphi(t);\xi_\varphi,\dot{\xi}_\varphi),$$

which can be rewritten as

$$\begin{aligned} &\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\Big[\|\dot{\xi}_\varphi\|^2_{\mathrm{L}^2(\Omega)} + \tilde{B}^{(S)}(t,\varphi(t);\xi_\varphi,\xi_\varphi) \Big] + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma |[\![\dot{\xi}_\varphi]\!]|^2 \,\mathrm{d}s \\ &\qquad = \rho(t;\dot{\xi}_\varphi) - \tilde{B}^{(A)}(t,\varphi(t);\xi_\varphi,\dot{\xi}_\varphi) - \frac{1}{2}\tilde{B}^{(S)}_t(t,\varphi;\xi_\varphi,\xi_\varphi). \end{aligned} \tag{6.12}$$

On noting that $\xi_\varphi(0) = 0$ and $\dot{\xi}_\varphi(0) = 0$, integrating the above identity in t, and multiplying by 2, we deduce from (6.8) that, for $\alpha \geq \alpha_0$ and $h \in (0, h_0]$,

$$\begin{aligned}
&\|\dot{\xi}_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)} + \tfrac{1}{2}M_1\|\xi_\varphi(t)\|^2_{1,h} - 2M_0\|\xi_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)} + 2\int_0^t \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma|[\![\dot{\xi}_\varphi(\tau)]\!]|^2 \,\mathrm{d}s\,\mathrm{d}\tau \\
&\leq \int_0^t \Big[2|\rho(\tau;\dot{\xi}_\varphi(\tau))| + 2\big|\tilde{B}^{(A)}(\tau,\varphi(\tau);\xi_\varphi(\tau),\dot{\xi}_\varphi(\tau))\big| + \big|\tilde{B}^{(S)}_t(\tau;\varphi(\tau);\xi_\varphi,\xi_\varphi)\big|\Big]\,\mathrm{d}\tau. \qquad (6.13)
\end{aligned}$$

Next we estimate the terms on the right-hand side, using (6.11), (6.9), and (6.10). Transferring the term $2M_0\|\xi_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)}$ to the right-hand side, we obtain

$$\begin{aligned}
&\|\dot{\xi}_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)} + \tfrac{1}{2}M_1\|\xi_\varphi(t)\|^2_{1,h} + 2\int_0^t \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma|[\![\dot{\xi}_\varphi(\tau)]\!]|^2 \,\mathrm{d}s\,\mathrm{d}\tau \\
&\quad\lesssim \|\xi_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)} + h^r \int_0^t \Big[\|\dot{\xi}_\varphi(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma|[\![\dot{\xi}_\varphi(\tau)]\!]|^2 \,\mathrm{d}s\Big]^{1/2} \mathrm{d}\tau \\
&\quad + \int_0^t \|\nabla\xi_\varphi(\tau)\|_{\mathrm{L}^2(\Omega)} \Big[\int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma|[\![\dot{\xi}_\varphi(\tau)]\!]|^2 \,\mathrm{d}s\Big]^{1/2} \mathrm{d}\tau + \int_0^t \|\nabla\xi_\varphi(\tau)\|^2_{\mathrm{L}^2(\Omega)} \,\mathrm{d}\tau. \qquad (6.14)
\end{aligned}$$

Using $\xi_\varphi(0) = 0$, the first term on the right-hand side can be estimated by

$$\|\xi_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)} = \Big\|\int_0^t \dot{\xi}_\varphi(\tau)\,\mathrm{d}\tau\Big\|^2_{\mathrm{L}^2(\Omega)} \leq T\int_0^t \|\dot{\xi}_\varphi(\tau)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}\tau.$$

Terms containing integrals over $[0,t]\times(\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}})$ in (6.14) can be absorbed into the third term on the left-hand side of (6.14) by apply the ε-inequality with sufficiently small ε (but independent of h). After normalization, we obtain

$$\begin{aligned}
&\|\dot{\xi}_\varphi(t)\|^2_{\mathrm{L}^2(\Omega)} + \|\xi_\varphi(t)\|^2_{1,h} + \int_0^t \int_{\Gamma_{\mathrm{int}}\cup\Gamma_{\mathrm{D}}} \sigma|[\![\dot{\xi}_\varphi(\tau)]\!]|^2 \,\mathrm{d}s\,\mathrm{d}\tau \\
&\qquad\lesssim h^{2r} + \int_0^t \Big[\|\dot{\xi}_\varphi(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \|\nabla\xi_\varphi(\tau)\|^2_{\mathrm{L}^2(\Omega)}\Big]\,\mathrm{d}\tau. \qquad (6.15)
\end{aligned}$$

Hence, an application of Gronwall's lemma gives

$$\max_{t\in[0,T]} \|\mathcal{N}(\varphi)(t) - W(t)\|_Y \lesssim h^r,$$

which allows us to deduce the existence of a constant $C_* = C_*(u)$, independent of h, such that, for $h \leq h_0$, $\mathcal{N}$ maps $\mathcal{J}$ into itself.

Remark 1. Since our strategy for proving that $\mathcal{N}$ maps $\mathcal{J}$ into itself was very similar to the one presented for the case of the quasi-linear elliptic problem considered earlier, we were more concise here than in the corresponding discussion for the elliptic problem. In particular, unlike our detailed analysis in the case of the elliptic problem where we made a deliberate effort to carefully track the constants in the bounds so as to be able to explicitly specify the value of the constant C_* featured in the definition of the set $\mathcal{J}$, here, for the sake of brevity, we refrained from doing so. As a matter of fact, the corresponding constant C_* can be found in an identical manner as in the case of the elliptic problem.

Next we prove that $\mathcal{N}$ is a contraction of $\mathcal{J}$ in the norm $\|\cdot\|_Y$. For this purpose, consider $u_\varphi = \mathcal{N}(\varphi) \in \mathcal{J}$ and $u_\psi = \mathcal{N}(\psi) \in \mathcal{J}$ defined analogously. Setting $\xi_\varphi = u_\varphi - W$ and $\xi_\psi = u_\psi - W$, we have that

$$(\ddot{\xi}_\varphi, v) + \tilde{B}(t, \varphi; \xi_\varphi, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\xi}_\varphi]\!] \cdot [\![v]\!] \,\mathrm{d}s = \rho(t; v), \quad \text{and}$$
$$(\ddot{\xi}_\psi, v) + \tilde{B}(t, \psi; \xi_\psi, v) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma[\![\dot{\xi}_\psi]\!] \cdot [\![v]\!] \,\mathrm{d}s = \rho(t; v) \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}),$$

subject to $\xi_\varphi(0) = \xi_\psi(0) = 0$ and $\dot{\xi}_\varphi(0) = \dot{\xi}_\psi(0) = 0$. By subtracting the second line from the first line, and testing with

$$v = \dot{\xi}_\varphi - \dot{\xi}_\psi = \dot{u}_\varphi - \dot{u}_\psi \equiv \dot{e},$$

where $e = u_\varphi - u_\psi$, we obtain

$$(\ddot{e}, \dot{e}) + \tilde{B}(t, \varphi; e, \dot{e}) + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma|[\![\dot{e}]\!]|^2 \,\mathrm{d}s = \tilde{B}(t, \psi; \xi_\psi, \dot{e}) - \tilde{B}(t, \varphi; \xi_\psi, \dot{e}).$$

By virtue of Lemma 4.3,

$$\left|\tilde{B}(t, \psi; \xi_\psi, \dot{e}) - \tilde{B}(t, \varphi; \xi_\psi, \dot{e})\right| \lesssim \|\nabla\varphi - \nabla\psi\|_{\mathrm{L}^\infty(\Omega)} \|\xi_\psi\|_{1,h} \|\dot{e}\|_{1,h}.$$

Thus, by using the same procedure as in the proof of the inclusion $\mathcal{N}(\mathcal{J}) \subset \mathcal{J}$, we obtain

$$\begin{aligned} &\|\dot{e}(t)\|^2_{\mathrm{L}^2(\Omega)} + \|e(t)\|^2_{1,h} + \int_0^t \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma|[\![\dot{e}(\tau)]\!]|^2 \,\mathrm{d}s \,\mathrm{d}\tau \\ &\quad\lesssim \int_0^t \left[\|\dot{e}(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \|\nabla e(\tau)\|^2_{\mathrm{L}^2(\Omega)}\right] \mathrm{d}\tau \\ &\quad\quad + \int_0^t \|\nabla\varphi(\tau) - \nabla\psi(\tau)\|_{\mathrm{L}^\infty(\Omega)} \|\xi_\psi(\tau)\|_{1,h} \|\dot{e}(\tau)\|_{1,h} \,\mathrm{d}\tau. \end{aligned} \tag{6.16}$$

As $u_\psi \in \mathcal{J}$, we have $\max_{t\in[0,T]} \|\xi_\psi(t)\|_{1,h} \le C_* h^r$, and, by the first inequality in (2.1), we also have that

$$\|\nabla\varphi(\tau) - \nabla\psi(\tau)\|_{\mathrm{L}^\infty(\Omega)} \lesssim h^{-d/2} \|\nabla\varphi(\tau) - \nabla\psi(\tau)\|_{\mathrm{L}^2(\Omega)}.$$

The only term on the right-hand side of (6.16) which cannot be directly controlled by any of the terms featured on the left-hand side of (6.16) is $\|\dot{e}(\tau)\|_{1,h}$. Employing the second inverse inequality in (2.1), we handle this term as follows:

$$\begin{aligned} \|\dot{e}(\tau)\|^2_{1,h} &= \|\nabla\dot{e}(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma|[\![\dot{e}(\tau)]\!]|^2 \,\mathrm{d}s \\ &\lesssim h^{-2} \|\dot{e}(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma|[\![\dot{e}(\tau)]\!]|^2 \,\mathrm{d}s. \end{aligned}$$

Inserting these bounds into (6.16), we obtain

$$\begin{aligned} &\|\dot{e}(t)\|^2_{\mathrm{L}^2(\Omega)} + \|e(t)\|^2_{1,h} + \int_0^t \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma|[\![\dot{e}(\tau)]\!]|^2 \,\mathrm{d}s \,\mathrm{d}\tau \\ &\quad\lesssim \int_0^t \left[\|\dot{e}(\tau)\|^2_{\mathrm{L}^2(\Omega)} + \|e(\tau)\|^2_{1,h}\right] \mathrm{d}\tau + h^{r-d/2-1} \int_0^t \|\nabla\varphi(\tau) - \nabla\psi(\tau)\|_{\mathrm{L}^2(\Omega)} \\ &\quad\quad \times \left[\|\dot{e}(\tau)\|^2_{\mathrm{L}^2(\Omega)} + h \int_{\Gamma_{\rm int}\cup\Gamma_{\rm D}} \sigma|[\![\dot{e}(\tau)]\!]|^2 \,\mathrm{d}s\right]^{1/2} \mathrm{d}\tau. \end{aligned}$$

Thus, by applying to the two last terms on the right-hand side of the ε-inequality $ab \le \frac{\varepsilon}{2}a^2 + \frac{1}{2\varepsilon}b^2$, with $\varepsilon > 0$ sufficiently small, we deduce from Gronwall's lemma that

$$\begin{aligned}\|\dot{u}_\varphi(t) - \dot{u}_\psi(t)\|^2_{\mathrm{L}^2(\Omega)} + \|u_\varphi(t) - u_\psi(t)\|^2_{1,h} &\lesssim h^{2(r-d/2-1)} \int_0^t \|\nabla\varphi(\tau) - \nabla\psi(\tau)\|^2_{\mathrm{L}^2(\Omega)}\,\mathrm{d}\tau \\ &\lesssim h^{2(r-d/2-1)}\|\varphi - \psi\|^2_Y,\end{aligned}$$

and thereby

$$\|\mathcal{N}(\varphi) - \mathcal{N}(\psi)\|_Y \lesssim h^{r-d/2-1}\|\varphi - \psi\|_Y \qquad \forall \varphi, \psi \in \mathcal{J},$$

which, in turn, implies that, for h sufficiently small, $\mathcal{N}$ is a contraction of $\mathcal{J}$ into itself in the norm $\|\cdot\|_Y$. Therefore, by Banach's fixed point theorem, for h sufficiently small, $\mathcal{N}$ has a unique fixed point, $u_{\mathrm{DG}} \in \mathcal{J}$, the semidiscrete discontinuous Galerkin finite element approximation to u defined by (6.2). In other words, for h sufficiently small,

$$\max_{t\in[0,T]} \Big(\|\dot{u}_{\mathrm{DG}}(t) - \dot{W}(t)\|_{\mathrm{L}^2(\Omega)} + \|u_{\mathrm{DG}}(t) - W(t)\|_{1,h}\Big) \le C_*(u)h^r, \qquad (d/2)+1 < r \le \min(m,p).$$

Combining the last bound with (A.1) and (A.7) we then deduce, for h sufficiently small, that

$$\max_{t\in[0,T]} \big(\|\dot{u}(t) - \dot{u}_{\mathrm{DG}}(t)\|_{\mathrm{L}^2(\Omega)} + \|u(t) - u_{\mathrm{DG}}(t)\|_{1,h}\big) \lesssim h^r, \qquad (d/2)+1 < r \le \min(m,p),$$

which is the desired optimal convergence estimate.

7. Extensions to other methods. It is straightforward to extend our error analysis to different discontinuous finite element methods. Note, for example, that in the elliptic case only Lemmas 4.1–4.3 are method-dependent. Once they are established, the remaining analysis is independent of the particular form of discretization used. We shall demonstrate this through the example of the discontinuous Galerkin finite element method (DGFEM) of Eyck and Lew [10], which is a particularly attractive candidate for variational problems since it is defined via a discrete energy principle.

The idea is to use the lifting operator introduced in [2] to find a gradient representation for the jumps across element interfaces to define a *discontinuous Galerkin (DG) gradient operator*. More precisely, for $v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$, let

$$\nabla_{\mathrm{DG}} v = \nabla v + \mathbf{R}(v),$$

where $\mathbf{R} : S^p(\Omega, \mathcal{T}_h, \mathbf{F}) \to \mathrm{C}_{\mathrm{pw}}(\overline{\Omega})^{d\times d}$ is defined by

$$\int_\Omega \mathbf{R}(v) : F\,\mathrm{d}x = -\int_{\Gamma_{\mathrm{int}}} [\![v]\!] \cdot \langle F\,\nu_{\mathrm{int}}\rangle\,\mathrm{d}s \qquad \forall F \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})^d.$$

We shall also use $\nabla^{i\alpha}_{\mathrm{DG}}$ to denote the (i,α) component of ∇_{DG}. It is straightforward to show that $\mathbf{R}$ is a bounded operator; more precisely,

$$\|\mathbf{R}(v)\|_{\mathrm{L}^2(\Omega)} \le C_L \left(\int_{\Gamma_{\mathrm{int}}} \sigma |[\![v]\!]|^2\,\mathrm{d}x\right)^{1/2} \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}), \tag{7.1}$$

where C_L is independent of h. Using the definition of the DG gradient, we define the discrete functional $J_h : S^p(\Omega, \mathcal{T}_h, \mathbf{F}) \to \mathbb{R}$ by

$$J_h(v) = \int_\Omega \left[W(\nabla_{\mathrm{DG}} v) - f \cdot v\right] \,\mathrm{d}x - \int_{\Gamma_{\mathrm{N}}} g_{\mathrm{N}} \cdot v \,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}} \sigma |[\![v]\!]|^2 \,\mathrm{d}s + \int_{\Gamma_{\mathrm{D}}} \sigma |v - g_{\mathrm{D}}|^2 \,\mathrm{d}s,$$

as an approximation to the functional J defined in (1.1). The resulting DGFEM for (1.2) is simply the Euler–Lagrange equation $\delta J_h(u_{\mathrm{DG}}) = 0$, where δJ_h, the first variation of J_h, is given by

$$\delta J_h(\varphi; v) = \int_\Omega \sum_{i,\alpha=1}^{d} S_{i\alpha}(\nabla_{\mathrm{DG}}\varphi) \nabla_{\mathrm{DG}}^{i,\alpha} v \,\mathrm{d}x + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma [\![\varphi]\!] \cdot [\![v]\!] \,\mathrm{d}s - \ell(v),$$

where $\ell(v)$ is defined as in (2.3). Since $\mathbf{R}(u) = 0$ if u is continuous on $\overline{\Omega}$, the method is consistent. Similarly, the second variation of J_h is defined by

$$\delta^2 J_h(\varphi; v, w) = \int_\Omega \sum_{i,\alpha,j,\beta=1}^{d} A_{i\alpha j\beta}(\nabla_{\mathrm{DG}}\varphi) \nabla_{\mathrm{DG}}^{i,\alpha} v \; \nabla_{\mathrm{DG}}^{j,\beta} w \,\mathrm{d}x + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma [\![v]\!] \cdot [\![w]\!] \,\mathrm{d}s.$$

Suppose that $u \in \mathrm{C}^1(\overline{\Omega})$ satisfies (3.2). While Lemma 3.2 cannot be applied directly, it is nevertheless straightforward to modify its proof to obtain for $h \le 1$, $\alpha \ge \alpha_0 = \alpha_0(C_r, K_\delta, M_1, M_0)$, and for all $\varphi \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ such that $\|\nabla_{\mathrm{DG}}\varphi - \nabla u\|_{\mathrm{L}^\infty(\Omega)} \le \delta \le M_1/(4L_\delta)$

$$\delta^2 J_h(\varphi; v, v) \ge \tfrac{1}{2} M_1 \|v\|_{1,h}^2 - 2M_0 \|v\|_{\mathrm{L}^2(\Omega)}^2 \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{7.2}$$

The boundedness and Lipschitz continuity of $\varphi \mapsto \delta^2 J_h(\varphi; \cdot, \cdot)$ over the set of all φ such that $\nabla_{\mathrm{DG}}\varphi \in \mathcal{M}_\delta$ can be obtained precisely as in Lemma 4.2 and 4.3. Using (7.1) we can again deduce that for $h \le h_0$

$$\varphi \in \mathcal{J} \;\Rightarrow\; \|\nabla_{\mathrm{DG}}\varphi - \nabla u\|_{\mathrm{L}^\infty(\Omega)} \le \delta,$$

and thus, the convergence analysis of section 5 can be repeated verbatim to obtain the existence of a solution u_{DG} to $\delta J_h(u_{\mathrm{DG}}; v) = 0$ for all $v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$, satisfying the optimal-order error estimate (5.10).

The analysis in the hyperbolic case can be generalized just as easily. The DGFEM based on the energy principle outlined above reads: For $t \in (0, T]$ find $u_{\mathrm{DG}}(t, \cdot) \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ such that

$$(\ddot{u}_{\mathrm{DG}}, v) + \delta J_h(u_{\mathrm{DG}}; v) + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma [\![\dot{u}_{\mathrm{DG}}]\!] \cdot [\![v]\!] \,\mathrm{d}s = \int_{\Gamma_{\mathrm{D}}} \sigma \dot{g}_{\mathrm{D}} \cdot v \,\mathrm{d}s \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{7.3}$$

Upon defining

$$\tilde{B}(t, \varphi(t); v, w) = \int_0^1 \delta^2 J_h(W(t) + \tau(\varphi(t) - W(t)); v, w) \,\mathrm{d}\tau,$$

the analysis proceeds almost exactly as in section 6. The only difference now is that, since $\tilde{B}(t, \varphi(t); \cdot, \cdot)$ is symmetric, we do not have to split it into a symmetric and an asymmetric part. Instead of (6.12), we will obtain

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\left[\|\dot{\xi}_\varphi\|_{\mathrm{L}^2}^2 + \tilde{B}(t, \varphi(t); \xi_\varphi, \xi_\varphi)\right] + \int_{\Gamma_{\mathrm{int}} \cup \Gamma_{\mathrm{D}}} \sigma |[\![\dot{\xi}_\varphi]\!]|^2 \,\mathrm{d}s = \rho(t; \dot{\xi}_\varphi) - \frac{1}{2}\tilde{B}_t(t, \varphi(t); \xi_\varphi, \xi_\varphi).$$

From the Lipschitz continuity of $\varphi \mapsto \delta^2 J_h(\varphi;\cdot,\cdot)$ on $\mathcal{J}$, we immediately obtain the bound on $\tilde{B}_t$ equivalent to (6.10), and we can thus proceed as in section 6.3 to prove the existence of a solution to (7.3) and an optimal error bound, identical to the one we had previously established.

8. Conclusions. We derived optimal-order convergence estimates in the broken H^1 norm for discontinuous Galerkin finite element approximations to second-order quasi-linear elliptic and hyperbolic systems of partial differential equations, using piecewise polynomials of degree $p > d/2$ in the elliptic case and of degree $p > d/2+1$ in the (spatially semidiscrete) hyperbolic case, where d is the spatial dimension of the problem. In the physically relevant cases of $d = 2$ and $d = 3$, these correspond to assuming that $p \geq 2$ and $p \geq 3$, respectively. These technical restrictions were also present in the work of Makridakis [15], whose techniques we have employed here. They occur, since we have used the inverse estimate (2.1) in order to obtain L^∞ bounds for elements of the set $\mathcal{J}$ defined, respectively, in sections 5 and 6.1, which in turn are required to obtain the uniform Gårding inequality of Lemma 3.2. However, we have reason to believe that the methods considered remain optimally convergent in the energy norm in these excluded cases as well; certainly, this is true for the nonlinear elliptic problem in the special case when the nonlinearity $\eta \mapsto S(\eta)$ is globally Lipschitz continuous and uniformly monotone (see [12]). The same statement would also follow immediately if one could prove directly, without involving the first inverse inequality in (2.1), that ∇u_{DG} is sufficiently close to ∇u in the L^∞-norm.

The main contribution of the paper is that these optimal-order, $\mathcal{O}(h^p)$, convergence rates have been proved without assuming that the nonlinear coefficient $S(\nabla u)$ appearing in the principal part of the operator is globally Lipschitz continuous or uniformly monotone (cf. (1.10)); instead, we assumed only local Lipschitz continuity of S and the Gårding inequality (3.2).

The main body of the paper was devoted to an analysis of the incomplete interior penalty method [9, 18]. However, we have demonstrated in section 7, where we showed how to extend all results to the variational DGFEM of Eyck and Lew [10], that the framework which we had developed should apply to virtually any discontinuous Galerkin discretization of the quasi-linear elliptic and hyperbolic equations considered. The crucial step is a proof of the coercivity estimate (5.3), using (a variation of) the broken Gårding inequality, stated in Lemma 3.2.

We note that all of our results can be straightforwardly extended to quasi-linear elliptic and hyperbolic partial differential equations where $S(\nabla u)$ is replaced by $S(u, \nabla u)$ under the same hypotheses; the presence of the lower-order nonlinearity causes no additional technical difficulties.

As our key objective here was to understand the analysis of discontinuous Galerkin approximations of locally Lipschitz spatial nonlinearities in quasi-linear elliptic and hyperbolic systems, we did not discuss fully discrete discontinuous Galerkin finite element approximations of quasi-linear hyperbolic problems. The convergence analysis of fully discrete schemes can be carried out using very similar theoretical tools to those presented here. We refer to [15], for example, for the corresponding analysis in the case of spatially H^1_0-conforming finite element methods which may serve as a starting point for further analytical considerations in that direction.

Appendix A. Bounds on the nonlinear projection error. The purpose of this section is to derive the required bounds on the error between a function u and its nonlinear elliptic projection W.

LEMMA A.1. *Let $u \in \mathrm{C}^2([0,T];\mathrm{H}^{m+1}(\Omega)^d)$, $m > d/2+1$, satisfy (3.2) with constants $M_1 > 0$ and $M_0 \geq 0$ which are independent of t. Suppose also that the family $\{\mathcal{T}_h\}_{h>0}$ satisfies* **H1**–**H4** *of section 2. Then there exists $h_0 > 0$ such that for $h \leq h_0$ there exists a solution $W(t) \in S^p(\Omega, \mathcal{T}_h, \mathbf{F})$ to the nonlinear equation*

$$B(W(t);v) + 2M_0(W(t),v) = B(u(t);v) + 2M_0(u(t),v) \qquad \forall v \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}).$$

Furthermore, $t \mapsto W(t)$ is twice differentiable in $[0,T]$ and satisfies

$$\|u(t) - W(t)\|_{1,h} \leq C_p h^r, \tag{A.1}$$

$$\|\dot{u}(t) - \dot{W}(t)\|_{1,h} \leq C_p'' h^r, \quad \text{and} \tag{A.2}$$

$$\|\ddot{u}(t) - \ddot{W}(t)\|_{1,h} \leq C_p'' h^r, \tag{A.3}$$

where C_p, C_p', and C_p'' are constants independent of h and t.

We skip the proof of existence of $W(t)$ and of the bound (A.1) which can be established by identical arguments to those in section 5 (see [17] for details). The proofs of (A.2) and (A.3) are given in the following two sections.

A.1. Bounds on $\dot{u} - \dot{W}$. Having established the existence of the nonlinear projection $W(t)$ of $u(t)$ for $t \in [0,T]$, we next prove the differentiability of the mapping $t \mapsto W(t)$. Suppose that $U \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ and $t \in [0,T]$. The mapping $V \mapsto B(U,V) - B(u(t),V) + 2M_0(U-u(t),V)$ is a bounded linear functional on $S^p(\Omega,\mathcal{T}_h,\mathbf{F})$; hence, by the Riesz representation theorem, there exists a unique (Riesz representer) $\mathcal{B}(t,U) \in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$ such that

$$(\mathcal{B}(t,U),V) = B(U,V) - B(u(t),V) + 2M_0(U-u(t),V) \qquad \forall V \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}).$$

This defines the (nonlinear) mapping

$$\mathcal{B} : (t,U) \in [0,T] \times S^p(\Omega,\mathcal{T}_h,\mathbf{F}) \mapsto \mathcal{B}(t,U) \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}).$$

It follows from the linearization process in section 4 and from Lemma 4.1 that the derivative of $(t,U) \mapsto \mathcal{B}(t,U)$ with respect to U, evaluated at $U = W(t)$, exists and is invertible for any $t \in [0,T]$. Note, furthermore, that $\mathcal{B}(t,W(t)) = 0$. Since $t \mapsto u(t)$ is differentiable, it follows that $(t,U) \mapsto \mathcal{B}(t,U)$ is differentiable in a neighborhood of $(t_0,W(t_0))$ for any $t_0 \in (0,T)$. We then deduce from the implicit function theorem that $t \mapsto W(t)$ is differentiable in $(0,T)$.

Set

$$u(t) - W(t) = (u(t) - \Pi_h u(t)) - (W(t) - \Pi_h u(t)) \equiv \eta - \xi.$$

We begin by noting that, according to the definition of $W(t)$,

$$\tilde{B}(t,W(t);\xi,v) + 2M_0(\xi,v) = \tilde{B}(t,u(t);\eta,v) + 2M_0(\eta,v) \qquad \forall v \in S^p(\Omega,\mathcal{T}_h,\mathbf{F}),$$

where

$$\tilde{B}(t,\varphi;v,w) = \int_0^t \tilde{b}(\Pi_h u(t) + \tau(\varphi - \Pi_h u(t));v,w)\,\mathrm{d}\tau.$$

After differentiation with respect to t, we obtain

$$\begin{aligned}\tilde{B}(t,W(t);\dot{\xi}(t),v) + 2M_0(\dot{\xi}(t),v) = \tilde{B}(t,u(t);\dot{\eta}(t),v) - \tilde{B}_t(t,W(t);\xi(t),v)\\ + \tilde{B}_t(t,u(t);\eta(t),v),\end{aligned} \tag{A.4}$$

where

$$\tilde{B}_t(t,\varphi(t);v,w)=\frac{\mathrm{d}}{\mathrm{d}t}\tilde{B}(t,\varphi(t);v,w)=\int_0^1\int_\Omega\sum_{i,\alpha,j,\beta=1}^{d}\Big[\frac{\mathrm{d}}{\mathrm{d}t}A^\tau_{i\alpha j\beta}\Big]\partial_{x_\alpha}w_i\partial_{x_\beta}v_j\ \mathrm{d}x\ \mathrm{d}\tau$$

for $v,w\in S^p(\Omega,\mathcal{T}_h,\mathbf{F})$, and $A^\tau_{i\alpha j\beta}$ is as before.

Arguing as in the proof of the bound (6.10), we obtain

(A.5)
$$\big|\tilde{B}_t(t,\varphi(t);v,w)\big|\lesssim\big(\|\nabla\Pi_h\dot{u}(t)\|_{\mathrm{L}^\infty(\Omega)}+\|\nabla\dot{\varphi}(t)-\nabla\Pi_h\dot{u}(t)\|_{\mathrm{L}^\infty(\Omega)}\big)\|v\|_{1,h}\|w\|_{1,h}$$

for all $\varphi\in\mathcal{Z}_\delta$. We note that

$$\|\nabla\Pi_h\dot{u}(t)\|_{\mathrm{L}^\infty(\Omega)}\le\|\nabla\dot{u}(t)\|_{\mathrm{L}^\infty(\Omega)}+C_5h^{r-d/2}\|\dot{u}\|_{\mathrm{H}^{r+1}(\Omega)}\lesssim 1,$$

where, in the last inequality, we made use of hypothesis **H3** whereby $r>(d/2)+1$, and, a fortiori, $r>d/2$.

In order to bound the last two terms in (A.4) we shall need to consider two specific choices of φ in (A.5): $\varphi=W$ and $\varphi=u$. For the case of $\varphi=W$ in (A.5), we shall use the following bound, which results on applying the first inverse inequality in (2.1):

$$\begin{aligned}\|\nabla\dot{W}(t)-\nabla\Pi_h\dot{u}(t)\|_{\mathrm{L}^\infty(\Omega)}&\lesssim h^{-d/2}\|\nabla\dot{W}(t)-\nabla\Pi_h\dot{u}(t)\|_{\mathrm{L}^2(\Omega)}\\&\lesssim h^{-d/2}\|\dot{W}(t)-\Pi_h\dot{u}(t)\|_{1,h}\approx h^{-d/2}\|\dot{\xi}\|_{1,h}.\end{aligned}$$

On the other hand, for the case of $\varphi=u$, we shall use the bound

$$\|\nabla\dot{u}(t)-\nabla\Pi_h\dot{u}(t)\|_{\mathrm{L}^\infty(\Omega)}\lesssim C_5h^{r-d/2}\|\dot{u}(t)\|_{\mathrm{H}^1(\Omega)}.$$

Thus, we obtain

$$\begin{aligned}&\big|\tilde{B}_t(t,W(t);\xi,v)\big|\lesssim\Big(1+h^{-d/2}\|\dot{\xi}\|_{1,h}\Big)\|\xi\|_{1,h}\|v\|_{1,h}\quad\text{and}\\&\big|\tilde{B}_t(t,u(t);\eta,v)\big|\lesssim\|\eta\|_{1,h}\|v\|_{1,h}.\end{aligned}$$

Upon testing (A.4) with $v=\dot{\xi}(t)$ and using (5.4) on the first term on its right-hand side, we obtain

$$\begin{aligned}\|\dot{\xi}(t)\|^2_{1,h}&\lesssim\|\dot{\eta}\|_{1,h}\|\dot{\xi}\|_{1,h}+\|\eta\|_{1,h}\|\dot{\xi}\|_{1,h}+\|\xi\|_{1,h}\|\dot{\xi}\|_{1,h}+h^{-d/2}\|\xi\|_{1,h}\|\dot{\xi}\|^2_{1,h}\\&\lesssim h^r\|\dot{\xi}\|_{1,h}+h^{r-d/2}\|\dot{\xi}\|^2_{1,h},\end{aligned}$$

where we also used the approximation properties of Π_h and estimate (A.1). Since $r>d/2$, there exists $h_2\in(0,\min(h_0,h_1)]$ such that for $h\in(0,h_2]$ the coefficient of $\|\dot{\xi}\|^2_{1,h}$ on the right-hand side is less than or equal to $\frac{1}{2}$. We can therefore bring this term to the left-hand side and divide by $\frac{1}{2}\|\dot{\xi}\|_{1,h}$ to finally obtain

$$\|\dot{\xi}\|_{1,h}=\|\dot{W}(t)-\Pi_h\dot{u}(t)\|_{1,h}\lesssim h^r$$

for $(d/2)+1<r\le\min(m,p)$, from which (A.2) follows immediately on invoking the approximation properties of Π_h.

A.2. Bounds on $\ddot{\eta} = \ddot{u} - \ddot{W}$. By proceeding in an identical manner as in the previous section we find that the mapping $t \mapsto \dot{W}(t)$ is differentiable on $(0,T)$, and we get, for $(d/2)+1 < r \leq \min(m,p)$, that

$$\|\ddot{W}(t) - \Pi_h \ddot{u}(t)\|_{1,h} \lesssim h^r \left(\|u(t)\|_{\mathrm{H}^{r+1}(\Omega)} + \|\dot{u}(t)\|_{\mathrm{H}^{r+1}(\Omega)} + \|\ddot{u}(t)\|_{\mathrm{H}^{r+1}(\Omega)} \right), \ h \in (0, h_2].$$

Invoking, once again, the approximation properties of Π_h, we deduce from the triangle inequality that, for $(d/2)+1 < r \leq \min(m,p)$, (A.3) holds.

Technically, the only additional step in this argument in comparison with that in the previous section is to establish a bound, similar to (A.5), on the term

$$\tilde{B}_{tt}(t, \varphi(t); v, w) = \frac{\mathrm{d}^2}{\mathrm{d}t^2} \tilde{B}(t, \varphi(t); v, w)$$

for $\varphi \in \mathcal{Z}_\delta$. Here we require a uniform bound on the fourth derivative of W, i.e., on the second derivatives

$$\frac{\partial^2}{\partial \eta_{\gamma k} \partial \eta_{\iota m}} A_{i\alpha j\beta}(\eta)$$

for $\eta \in \mathcal{M}_\delta$ and can otherwise argue similarly as in the proof of (6.10); hence our assumption $W \in \mathrm{C}^4(\mathbb{R}^{d\times d}; \mathbb{R})$ on the regularity of the stored energy function W was adopted in the introductory section of the paper.

A.3. $\mathbf{L^2}$-bounds. Since, by hypothesis **H4**, the family $\{\mathcal{T}_h\}_{h>0}$ is uniformly simplicially reducible (cf. also section 3), the broken Friedrichs inequality (cf. [3]) implies the existence of a positive constant C, independent of h, such that

$$\|v\|^2_{\mathrm{L}^2(\Omega)} \leq C \|v\|^2_{1,h} \qquad \forall v \in S^p(\Omega, \mathcal{T}_h, \mathbf{F}). \tag{A.6}$$

Here the constant C depends only on certain shape-regularity properties of the family $\{\mathcal{T}_h\}_{h>0}$, the penalty parameter α, and the Friedrichs constant for $\mathrm{H}^1_{\mathrm{D},0}(\Omega)$.

In fact, (A.6) can also be obtained from Lemma 3.1, in which case the corresponding constant C would depend on the constant C_r, the penalty parameter α, and the Friedrichs constant for $\mathrm{H}^1_{\mathrm{D},0}(\Omega)$.

Either way, on applying (A.6) to (A.1)–(A.3), we obtain

$$\|W(t) - u(t)\|_{\mathrm{L}^2(\Omega)} + \|\dot{W}(t) - u(t)\|_{\mathrm{L}^2(\Omega)} + \|\ddot{W}(t) - \ddot{u}(t)\|_{\mathrm{L}^2(\Omega)} \lesssim h^r. \tag{A.7}$$

While this bound is not optimal (the optimal rate would be $r+1$ rather than r), it is entirely adequate for the purposes of deriving an optimal bound on $u - u_{\mathrm{DG}}$ in the *energy norm* $\|\cdot\|_Y$.

REFERENCES

[1] S. S. Antman, *Nonlinear Problems of Elasticity*, Appl. Math. Sci. 107, 2nd ed., Springer, New York, 2005.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2001), pp. 1749–1779.

[3] S. C. Brenner, *Poincaré-Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

[4] S. C. Brenner, *Korn's inequalities for piecewise H^1 vector fields*, Math. Comp., 73 (2004), pp. 1067–1087.

[5] V. C. Chen and W. von Wahl, *Das Rand-Anfangswertproblem für quasilineare Wellengleichungen in Sobolev-räumen niedriger Ordnung*, J. Reine Angew. Math., 337 (1982), pp. 77–112.

[6] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, in Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978.

[7] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 3–50.

[8] C. M. Dafermos and W. J. Hrusa, *Energy methods for quasilinear hyperbolic initial-boundary value problems. Applications to elastodynamics*, Arch. Ration. Mech. Anal., 87 (1985), pp. 267–292.

[9] C. Dawson, S. Sun, and M. F. Wheeler, *Compatible algorithms for coupled flow and transport*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 2565–2580.

[10] T. Eyck and A. Lew, *Discontinuous Galerkin methods for nonlinear elasticity*, Int. J. Numer. Meth. Engrg., 67 (2006), pp. 1204–1243.

[11] M. E. Gurtin, *An Introduction to Continuum Mechanics*, Math. Sci. Eng. 158, Academic Press, New York, 1981.

[12] P. Houston, J. Robson, and E. Süli, *Discontinuous Galerkin finite element approximation of quasilinear elliptic boundary value problems. I. The scalar case*, IMA J. Numer. Anal., 25 (2005), pp. 726–749.

[13] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[14] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[15] C. G. Makridakis, *Finite element approximations of nonlinear elastic waves*, Math. Comp., 61 (1993), pp. 569–594.

[16] C. B. Morrey, Jr., *Multiple integrals in the calculus of variations*, Grundlehren Math. Wi. 130, Springer, New York, 1966.

[17] C. Ortner and E. Süli, *Discontinuous Galerkin Finite Element Approximation of Nonlinear Second-Order Elliptic and Hyperbolic Systems*, Technical report NA-06/05, Oxford University Computing Laboratory, London, 2006.

[18] S. Sun, *Discontinuous Galerkin Methods for Reactive Transport in Porous Media*, Ph.D. thesis, The University of Texas at Austin, 2003.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1398–1424

VALIDATED CONTINUATION FOR EQUILIBRIA OF PDEs*

SARAH DAY†, JEAN-PHILIPPE LESSARD‡, AND KONSTANTIN MISCHAIKOW§‡

Abstract. One of the most efficient methods for determining the equilibria of a continuous parameterized family of differential equations is to use predictor-corrector continuation techniques. In the case of partial differential equations this procedure must be applied to some finite-dimensional approximation, which of course raises the question of the validity of the output. We introduce a new technique that combines the information obtained from the predictor-corrector steps with ideas from rigorous computations and verifies that the numerically produced equilibrium for the finite-dimensional system can be used to explicitly define a set which contains a unique equilibrium for the infinite-dimensional partial differential equation. Using the Cahn–Hilliard and Swift–Hohenberg equations as models we demonstrate that the cost of this new validated continuation is less than twice the cost of the standard continuation method alone.

Key words. continuation, PDE, Swift–Hohenberg equations, validation

AMS subject classifications. 65G20, 65N30

DOI. 10.1137/050645968

1. Introduction.

The first step in understanding the dynamics of a nonlinear system of differential equations

$$u_t = f(u, \nu) \tag{1.1}$$

on a Hilbert space is to identify the set of equilibria $\mathcal{E} := \{(u, \nu) \mid f(u, \nu) = 0\}$. For many applications this can only be done using numerical methods. In particular, continuation provides an efficient technique for determining elements on branches of $\mathcal{E}$. Recall that this method involves a predictor and corrector step: given, within a prescribed tolerance, an equilibrium u_0 at parameter value ν_0, the predictor step produces an approximate equilibrium $\tilde{u}_1$ at nearby parameter value ν_1, and the corrector step, often based on a Newton-like operator, takes $\tilde{u}_1$ as its input and produces, once again within the prescribed tolerance, an equilibrium u_1 at ν_1.

With any numerical method there is the question of validity of the output as compared with the cost of computation. The goal of this paper is to argue that for a large and important class of partial differential equations (PDEs) the cost of validating the existence and uniqueness of equilibria is small when compared to the cost of identifying potential equilibria by means of a continuation method. Our interest in this question was motivated by the increasing development of computer-assisted proofs in the dynamics of infinite-dimensional systems (see [3], [10] and the references therein). As mathematicians we are willing to argue forcefully for the importance of rigorous verification and thus marginalize the cost. However, in reality for many

*Received by the editors November 23, 2005; accepted for publication (in revised form) January 16, 2007; published electronically July 18, 2007.
http://www.siam.org/journals/sinum/45-4/64596.html

†Department of Mathematics, The College of William and Mary, P.O. Box 8795, Williamsburg, VA 23187-8795 (sday@math.wm.edu). This author was partially supported by NSF DMS 9983660.

‡School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332-0160 (lessard@math.gatech.edu). The second author was partially supported by FQRNT, NSF DMS 0511115, and grants from D.O.E. and DARPA.

§Department of Mathematics, Rutgers University, Hill Center-Busch Campus, 110 Frelinghusen Rd., Piscataway, NJ 08854-8019 (mischaik@math.rutgers.edu). This author was partially supported by NSF DMS 0511115 and grants from D.O.E. and DARPA.

applications, researchers are often interested in investigating a variety of model PDEs at a multitude of parameter values to gain scientific insight rather than an answer to a particular question. This places a premium on minimizing computational cost, often leading to acceptance of the validity of numerical results simply based upon the reproducibility of the result at different levels of refinement. As we shall argue, the results of this paper suggest that this dichotomy need not exist, and we provide examples wherein it is demonstrated that by judicious use of the computations involved in the continuation method it is cheaper to validate the results than to reperform the continuation computation. We refer to the method we propose as *validated continuation*. As is made clear towards the end of the introduction, validated continuation is slightly weaker and computationally cheaper than rigorous continuation.

To the best of our knowledge this is the first attempt to integrate the techniques of rigorous computations with a continuation method; thus we focus on a clear presentation of the ideas as opposed to presenting the results in the most general possible setting. We make use of spectral methods, as they provide us with considerable control on truncation errors. To be more precise, assume that (1.1) takes the form

$$u_t = L(u,\nu) + \sum_{p=0}^{d} c_p(\nu) u^p, \tag{1.2}$$

where $L(\cdot,\nu)$ is a linear operator at parameter value ν and d is the degree of the polynomial nonlinearity. Typically, $c_1(\nu) = 0$ since linear terms are grouped under $L(\cdot,\nu)$. Expanding (1.2) using an orthogonal basis chosen appropriately in terms of the eigenfunctions of the linear operator $L(\cdot,\nu)$, the particular domain, and the boundary conditions results in a countable system of differential equations on the coefficients of the expanded solution.

To simplify the exposition, let us assume the expansion takes the form

$$\dot{u}_k = f_k(u,\nu) := \mu_k u_k + \sum_{p=0}^{d} \sum_{\sum n_i = k} (c_p)_{n_0} u_{n_1} \cdots u_{n_p}, \quad k = 0,1,2,\ldots, \tag{1.3}$$

where $\mu_k = \mu_k(\nu)$ are the parameter-dependent eigenvalues of $L(\cdot,\nu)$, and $\{u_n\}$ and $\{(c_p)_n\}$ are the coefficients of the corresponding expansions of the functions u and $c_p(\nu)$, respectively, with $u_n = u_{-n}$ and $(c_p)_n = (c_p)_{-n}$ for all n. In order to simplify the notation, for a fixed parameter ν, we use $f(u)$ to denote $f(u,\nu)$. The continuation method is applied to the m-dimensional system of ordinary differenial equations (ODEs) of the form

$$\dot{u}_k = \mu_k u_k + \sum_{p=0}^{d} \sum_{\substack{\sum n_i = k \\ |n_i| < m}} (c_p)_{n_0} u_{n_1} \cdots u_{n_p}, \quad k = 0,1,\ldots,m-1, \tag{1.4}$$

obtained by performing a Galerkin projection on (1.3). It is this truncation that introduces the most substantial concern for the validity of the results of the continuation method. In section 3 we present estimates that provide us with bounds on the errors. We obtain these bounds under the assumption of power decay rates in the coefficients $\{u_n\}$. Of course, such decay rates are directly related to the spatial smoothness of the equilibria, which in turn is governed, at least in part, by the linear operator $L(\cdot,\nu)$.

The theoretical justification for our proof of existence and uniqueness of equilibria is based on a componentwise version of the Banach fixed point theorem (see

Theorem 2.1), which itself represents a minor modification of a result of Yamamoto [9, Theorem 2.1]. A similar formulation can also be found in [4]. Recall that to apply the Banach fixed point theorem one must have a contraction mapping $T : X \to X$. With this in mind, we can state that it is appropriate to view our approach as a method by which the Newton-like iteration of the corrector step in the continuation process is used to construct a set X and the above estimates are used to verify that an appropriate generalization of the Newton-like operator is in fact a contraction. More precisely, let $\bar{u}$ be a numerical zero obtained from (1.4). In the orthogonal basis used to obtain (1.3) consider the set $X = \bar{u} + W(r)$ of $\bar{u}$, where $W(r)$ is of the form

$$W(r) = \prod_{k=0}^{m-1} [-r, r] \times \prod_{k=m}^{\infty} \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s} \right].$$

Observe that s indicates the decay rate of the coefficients and r is referred to as the *validation radius*. Our strategy, described in detail in section 3, is to produce a set of *radii polynomials*, $\{P_k(r)\}_{k=0,1,\ldots}$, whose coefficients are given explicitly in terms of the constants A_s, s, and (1.3). Theorem 3.4 guarantees that if there exists a validation radius $r > 0$ such that $P_k(r) < 0$ for all k, then there exists a unique equilibrium solution to (1.2) in the set $X = \bar{u} + W(r)$ built around the numerical equilibrium $\bar{u}$ produced by the continuation procedure. It is important to note that the conditions of Theorem 3.4 can be checked with a finite number of calculations.

As is indicated above the focus of this paper is on the computational efficacy of validated continuation and hence the following organization of the material. Section 2 contains the statement and proof of the aforementioned componentwise version of the Banach fixed point theorem, Theorem 2.1, without any indication of how this result can be used in practice. Section 3 provides the opposite extreme, an explicit set of formulas and steps and the assertion that their successful implementation leads, via Theorem 3.4, to the existence of a unique equilibrium in a specified set. The justification of this assertion and the relationship between Theorems 2.1 and 3.4 is presented in section 6. However, presenting the formulae in this fashion has two advantages. First, they contain all the necessary information should the reader wish to independently code and test the techniques suggested in this paper. Second, it allows for the presentation in section 4 of the comparison of the computational costs between traditional and validated continuation.

It should be emphasized that how one should best compare the costs between the two methods of continuation is not completely clear. In the standard approach m, the dimension of the system on which continuation is performed, is fixed. Thus traditionally a particular Galerkin projection dimension is chosen and continuation is performed. The results are checked by choosing a higher-dimensional projection, reperforming the continuation, and then deciding if the two calculations agree within a certain level of numerical tolerance. In validated continuation, m becomes a variable. In particular, if validation fails, then one has the option of choosing a higher-dimensional Galerkin projection. Equally important, failure of validation may be an indication that a higher-dimensional projection is necessary. In summary, validated continuation provides an internal check of consistency on the dimension of truncation from the infinite- to finite-dimensional problem, a feature which is not present in the traditional application of continuation methods.

With this in mind we have chosen to compare the computational costs as follows. First we restrict our attention to cubic nonlinearities. As is made clear by the formulae of section 3, in this case the cost of evaluating the nonlinearities and performing

Newton's method are both of order m^3. Thus, we can obtain a rough bound on the ratio of the cost of traditional versus validated continuation by counting the number of m^3 operations which need to be performed. These calculations suggest that for fixed m the cost of validated continuation is less than twice the cost of traditional continuation, that is, *it appears that it is cheaper to perform validated continuation than to perform traditional continuation and then check it against continuation performed on a higher-dimensional projection.* In section 5 this estimate is tested against actual computations for the Swift–Hohenberg equation and the Cahn–Hilliard equation. To ensure that these comparisons are fair, we employ standard floating point computations in both cases.

This last point raises an important distinction: validated continuation versus rigorous continuation. Using floating point calculations at all steps of the validated continuation does not allow one to control for roundoff errors, and hence one cannot rigorously conclude the existence of an equilibrium. Because the current computer technology treats floating point and interval arithmetic differently we chose not to make and present timed comparisons between the two for this paper. However, if specific steps in the validation argument are performed using interval arithmetic, then one obtains rigorous results on the existence of equilibria. Results of this type are presented in section 5 for a branch of equilibria of the Swift–Hohenberg equation.

We see the results of this paper as a first step in the direction of combining continuation methods with rigorous computations. With this in mind we conclude the paper in section 7 with a discussion of open questions and ongoing work. In particular, we return to the issue of the necessity of interval arithmetic computations.

2. Computational proofs for equilibria. Assume that following the expansion of a PDE into an appropriate orthogonal basis, we have a system of the form (1.3). Our goal is to prove that there is a unique equilibrium for (1.3) which lies in a small set containing a computed numerical equilibrium. Suppose $\bar{u}_F$ is a numerical equilibrium computed using an m-dimensional continuation procedure (as described in section 3) and $\bar{u} := (\bar{u}_F, 0, \dots)$ is the corresponding point in the infinite-dimensional space. We will consider a set of the form $\bar{u} + W$, where $W = \Pi_k \tilde{w}_k$,

$$\tilde{w}_k = \begin{cases} [-r, r], & 0 \le k < m, \\ \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s}\right], & k \ge m, \end{cases} \tag{2.1}$$

for some constants $r, A_s > 0$ and $s \ge 2$.

A particularly nice norm to use for this set (similar to the one used by Yamamoto in [9]) is the normalized sup norm

$$\|u\|_W := \sup_k \left\{ \frac{|u_k|}{|\tilde{w}_k|} \right\},$$

where $|\tilde{w}_k| := \max\{|x| \mid x \in \tilde{w}_k\}$. In this norm, $W = B(0,1)$ is the unit ball around 0, and $\bar{u} + W = B(\bar{u}, 1)$ is the unit ball around $\bar{u}$.

We will now reformulate our problem of studying equilibria for (1.3) by establishing an equivalent fixed point problem on $\bar{u} + W$. Suppose J is an invertible operator. Then u is an equilibrium solution of (1.3) if and only if u is a fixed point of

$$T(u) = u - Jf(u), \tag{2.2}$$

where f is given by (1.3). In practice, T is constructed to be a contraction (Newton-like) operator with $J \approx (Df(\bar{u}))^{-1}$ so that we may use Banach's fixed point theorem. We now frame this fixed point theorem in a more computational setting.

In the process of showing that T is a contraction, we first consider the following Lipschitz condition on $\bar{u} + W$:

$$\|T(x) - T(y)\|_W \leq K\|x - y\|_W \quad \text{for } x, y \in \bar{u} + W. \tag{2.3}$$

The question now becomes whether we can compute a contraction constant $K < 1$ satisfying (2.3). We begin by computing Lipschitz constants, K_n, for the component functions T_n on $\bar{u} + W$ satisfying

$$|T_n(x) - T_n(y)| \leq K_n\|x - y\|_W \quad \text{for } x, y \in \bar{u} + W. \tag{2.4}$$

If T is C^1, we may take K_n to be a bound on the derivative of T_n over $\bar{u} + W$. More explicitly,

$$\begin{aligned} K_n &\geq \sup |DT_n(\bar{u} + W) \cdot W| \\ &:= \sup_{b,c \in W} |DT_n(\bar{u} + b) \cdot c| . \end{aligned}$$

A constant K_n computed in this manner satisfies (2.4) by the following argument. For $x, y \in \bar{u} + W$, let $g_n(s) := T_n[sx + (1 - s)y]$. Applying the mean value theorem to g_n, we get the existence of $s_n \in [0, 1]$ such that $g_n(1) - g_n(0) = g'(s_n)$. Since the set $\bar{u} + W$ is convex, we get the existence of $z_n := s_n x + (1 - s_n)y \in \bar{u} + W$ such that

$$\begin{aligned} |T_n(x) - T_n(y)| &= |DT_n(z_n)(x - y)| \\ &= \left| DT_n(z_n) \frac{x - y}{\|x - y\|_W} \right| \|x - y\|_W. \end{aligned}$$

By construction of $\|\cdot\|_W$, $\frac{x-y}{\|x-y\|_W} \in W$. Now if $K := \sup_n \frac{K_n}{|\tilde{w}_n|} < \infty$, then, as the following argument shows, it satisfies (2.3):

$$\begin{aligned} \|T(x) - T(y)\|_W &= \sup_n \frac{|T_n(x) - T_n(y)|}{|\tilde{w}_n|} \\ &= \sup_n \frac{\left| DT_n(z_n) \frac{x-y}{\|x-y\|_W} \right| \|x - y\|_W}{|\tilde{w}_n|} \\ &\leq \sup_n \frac{K_n}{|\tilde{w}_n|} \|x - y\|_W \\ &= K\|x - y\|_W. \end{aligned}$$

THEOREM 2.1 (existence and uniqueness). *If for all n there exist bounds $Y_n \geq |T_n(\bar{u}) - \bar{u}_n|$ and K_n satisfying (2.4) such that*

$$Y_n + K_n - |\tilde{w}_n| < 0 \tag{2.5}$$

and

$$K := \sup_n \frac{K_n}{|\tilde{w}_n|} < 1, \tag{2.6}$$

then there exists a unique fixed point of T in $\bar{u} + W$.

Proof. The first inequality ensures that $T(\bar{u} + W) \subset \bar{u} + W$. This is true if and only if for every $u \in \bar{u} + W$, $\|T(u) - \bar{u}\|_W \leq 1$ or, equivalently, $\frac{|T_n(u) - \bar{u}_n|}{|\tilde{w}_n|} < 1$ for all n.

Let $u \in \bar{u} + W$. Then $\|u - \bar{u}\|_W \leq 1$ and for each n,

$$\begin{aligned} |T_n(u) - \bar{u}_n| &= |T_n(u) - T_n(\bar{u}) + T_n(\bar{u}) - \bar{u}_n| \\ &\leq |T_n(u) - T_n(\bar{u})| + |T_n(\bar{u}) - \bar{u}_n| \\ &\leq K_n \|u - \bar{u}\|_W + Y_n \\ &\leq Y_n + K_n \\ &< |\tilde{w}_n| \end{aligned}$$

by assumption (2.5). Therefore, $T(\bar{u}+W) \subset \bar{u}+W$. The second inequality guarantees that T is also a contraction. Thus, the result follows from Banach's fixed point theorem. ∎

Let us make the comment here that sufficient regularity of the equilibrium solutions will effectively reduce the infinite set of conditions listed in Theorem 2.1 to a finite list. In essence, the strong decay in the higher modes may be used to verify (2.5) simultaneously for all $n > N$ for some N. (In our case N is determined by the dimension used for continuation and the degree of the nonlinearity.) Furthermore, regularity of the equilibria may also be used to show that $K_n|\tilde{w}_n|^{-1}$ becomes a decreasing sequence. Therefore, (2.6) follows automatically from (2.5).

Perhaps an even more important point to make for our intended algorithmic approach in this paper is that $Y_n + K_n - |\tilde{w}_n|$ will be given as a polynomial in the validation radius r, the width of the set W in the low modes. Therefore, validating the existence of a unique equilibrium near $\bar{u}$ will amount to showing that it is possible to simultaneously solve a (finite) list of polynomial inequalities in r.

3. Validated continuation. The ideas outlined in section 2 for proving the existence of unique equilibria fit naturally with traditional continuation techniques for following branches of numerical equilibria. In particular, an approximation of a projection of the Newton operator given in (2.2) onto the appropriate m-dimensional subspace is an intrinsic element of the continuation algorithm. In this section, we discuss exploiting this relationship to automatically produce a validation of the existence of unique equilibria at each step of the continuation procedure.

Recall that following the expansion of the system in the appropriate basis, we have

$$\dot{u} = f(u, \nu), \tag{3.1}$$

where for $k = 0, 1, 2, \ldots$, $\mu_k = \mu_k(\nu)$, $(c_p)_n = (c_p(\nu))_n$, and

$$\dot{u}_k = f_k(u) = \mu_k u_k + \sum_{p=0}^{d} \sum_{\sum n_i = k} (c_p)_{n_0} u_{n_1} \cdots u_{n_p}. \tag{3.2}$$

A first step for implementing a continuation algorithm for studying a PDE is to perform a Galerkin projection. Let m be a fixed projection dimension and consider the following truncated version of our original expansion of the PDE given in (3.2). For $u_F := (u_0, \ldots, u_{m-1}) \in \mathbb{R}^m$, define $f^{(m)} : \mathbb{R}^m \to \mathbb{R}^m$ by $f^{(m)}(u_F) = (f_0^{(m)}(u_F), \ldots, f_{m-1}^{(m)}(u_F))$, where for $k = 0, \ldots, m-1$,

$$f_k^{(m)}(u_F) = \mu_k u_k + \sum_{p=0}^{d} \sum_{\substack{\sum n_i = k \\ |n_i| < m}} (c_p)_{n_0} u_{n_1} \cdots u_{n_p}.$$

The corresponding Galerkin projection of the original system (3.1) is then

$$\dot{u}_F = f^{(m)}(u_F, \nu). \tag{3.3}$$

This is the m-dimensional system to be studied numerically. Intuitively, we expect that if m is sufficiently large, (3.3) will capture the essential dynamics for the original system (3.1). In particular, given an equilibrium $\bar{u}_F$ for (3.3) we expect that there is a small set around $\bar{u} := (\bar{u}_F, 0, \dots)$ which contains a unique equilibrium solution for (3.1). Our approach is to study this relationship via the tools outlined in section 2.

3.1. Continuation for ODEs and Newton-like operator. A traditional continuation procedure involves iteration of predictor and corrector steps to trace out branches of equilibria. Under the assumption that at some parameter $\nu = \nu_0$ we have an equilibrium solution for (3.3), we want to continue the equilibrium as we vary ν.

(1) *Euler predictor:* Given an approximate equilibrium x_0 at ν_0, the *predictor* at $\nu_1 = \nu_0 + \Delta\nu$ is $x_1^{(0)} = x_0 + \dot{x}_0 \Delta\nu$, where

$$\dot{x}_0 = -f_x^{(m)}(x_0, \nu_0)^{-1} f_\nu^{(m)}(x_0, \nu_0). \tag{3.4}$$

(2) *Quasi-Newton corrector:* We now use the following quasi-Newton iterative scheme to improve our approximation at ν_1:

$$x_1^{(n+1)} = x_1^{(n)} - f_x^{(m)}({x_1}^{(0)}, \nu_1)^{-1} f^{(m)}(x_1^{(n)}, \nu_1). \tag{3.5}$$

If k is the total number of iterations of (3.5), then $\bar{u}_F := x_1^{(k)}$ and $f^{(m)}(\bar{u}_F, \nu_1) \approx 0$.

As before, define the corresponding point $\bar{u} = (\bar{u}_F, 0, \dots)$ in the infinite-dimensional space. We now use the information required for the next predictor step, the numerical inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$, to construct a Newton-like operator near $\bar{u}$ at the parameter value ν_1. Let $J_{F\times F}$ be the numerical inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$ and define the Newton-like operator T by

$$T(u) = u - Jf(u), \tag{3.6}$$

where

$$J := \begin{bmatrix} J_{F\times F} & & 0 & \\ & \mu_m^{-1} & & \\ 0 & & \mu_{m+1}^{-1} & \\ & & & \ddots \end{bmatrix}$$

is the block diagonal matrix which we expect to be close to $(Df(\bar{u}, \nu_1))^{-1}$. Note that T, J, and f all depend on the parameter ν. As in section 2, we will attempt to show that T is a contraction on a set of the form $\bar{u} + W$, where W has the form (2.1). We now emphasize the dependence of this set $W = W(r)$ on the validation radius r since this approach relies on finding an appropriate $r > 0$ to satisfy a set of conditions. The constants A_s and s may be determined by regularity arguments or otherwise set prior to the computations. As seen in the definition of $W(r)$, these constants determine the size of the region in which we are attempting to show the unique existence of an equilibrium solution.

3.2. Radii polynomials. We now present the formulae for *radii polynomials*. In order to focus on the applicability of validated continuation, the justification that these polynomials do, in fact, encode the required bounds Y_n and K_n in (2.5) for the Newton-like operator constructed in (3.6) is delayed to section 6.

Since the formulae for the polynomials are rather ungainly, let us begin by explicitly stating the information that is used to construct the coefficients.

- d is the degree of the nonlinearity of (1.2).
- m is the number of modes used in the Galerkin projection.
- $M \geq m$ is a computational parameter that allows for the use of explicit values for coefficients of $M - m$ additional modes to decrease truncation error bounds.
- $m_+ \geq m$ is a computational parameter that allows for the use of an additional structure in the model to get tighter truncation error bounds.
- $\bar{u}_F \in \mathbb{R}^m$ is the numerical zero produced by the predictor-corrector step.
- $J_{F\times F}$ is the numerical inverse obtained from the predictor-corrector step.
- $(c_p)_n$, $|n| < m$, are the coefficients from the expansion (1.3).
- μ_k, $k \geq 0$, are the eigenvalues for the linear operator L as expressed in (1.3) and

$$\bar{\mu} := \inf_{n \geq m_+} \{|\mu_n|\}.$$

  Note that if $|\mu_n|$ is monotonically increasing for $n \geq m_+$, then $\bar{\mu} = |\mu_{m_+}|$.
- s and A_s are positive constants that are related to the regularity of the equilibria.

Observe that given this information we can evaluate the vector

$$f_F(\bar{u}) := \begin{bmatrix} f_0(\bar{u}) \\ \vdots \\ f_{m-1}(\bar{u}) \end{bmatrix},$$

where

$$f_n(\bar{u}) = \mu_n \bar{u}_n + \sum_{p=0}^{d} \sum_{\substack{n_0+\cdots+n_p=n \\ |n_1|,\ldots,|n_p|<m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_p} \ .$$

We can also set

$$Y_k \geq \begin{cases} |J_{F\times F} f_F(\bar{u})|_k & \text{if } 0 \leq k < m, \\ \frac{|\sum_{p=2}^{d} (c_p \bar{u}^p)_k|}{|\mu_k|} & \text{if } k \geq m, \end{cases} \tag{3.7}$$

where

$$(c_p \bar{u}^p)_k = \sum_{\sum n_i = k} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_p}.$$

The following constants are all related to asymptotic bounds on the expansions of the numerical equilibrium $\bar{u}$ and the set $\bar{u} + W$. As such they are related to the

regularity of the equilibrium and the coefficients of (1.2). Define

$$\begin{aligned}
\alpha &:= \frac{2}{s-1} + 2 + 3.5 \cdot 2^s, \\
C_p &:= \max_k \{|(c_p)_0|, |(c_p)_k||k|^s\}, \\
\bar{A} &:= \max_{1 \le k < m} \{|\bar{u}_0|, |\bar{u}_k||k|^s\}, \\
A = A(r) &:= \max\{A_s, r(m-1)^s\}, \\
C(\bar{A}, A) &:= \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l\binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A(r)^l, \\
C_+(\bar{A}, A) &:= \begin{cases} \sum_{l=1}^{d} \sum_{p=2}^{d} l\binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l & \text{if } Y_k, R_k = 0 \text{ for all } k \ge m_+, \\ \sum_{p=0}^{d} \alpha^p C_p \bar{A}^p + \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l\binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l & \text{otherwise}, \end{cases} \\
V_F^{(0)} &:= |J_{F\times F}|\, R_F, \qquad V_F^{(1)} := \left| I_{F\times F} - J_{F\times F} \cdot Df^{(m)}(\bar{u}_F) \right| \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},
\end{aligned}$$

where $|\cdot|$ denotes the entrywise absolute value and for $k \in \{0, \dots, m-1\}$,

$$R_k := \sum_{\substack{\bar{n}=-\infty \\ |k-\bar{n}| \ge m}}^{\infty} \left| \sum_{p=1}^{d} p \sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \dots \bar{u}_{n_{p-1}} \right| \frac{A_s}{|k-\bar{n}|^s} \ .$$

Note that if all c_p have finite expansions, then $V_F^{(0)}$ requires only a finite computation. Observe also that the above implies that $\bar{u}_k \in \frac{\bar{A}}{k^s}[-1,1]$ and $\tilde{w}_k \subset \frac{A}{k^s}[-1,1]$ for all k.

The validation procedure also requires bounds on the errors due to truncating modes $k \ge m$. These bounds come in the following form:

$$\epsilon_n := \sum_{l=1}^{d} \sum_{p=l}^{d} l\binom{p}{l} \epsilon_n(p,l,M), \tag{3.8}$$

where

$$\begin{aligned}
&\epsilon_n(p,l,M) \\
&:= \min\left\{ \frac{p\alpha^{p-1} C_p \bar{A}^{p-l} A^l}{(M-1)^{s-1}(s-1)} \left[\frac{1}{(M-n)^s} + \frac{1}{(M+n)^s} \right], \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{n^s} \right\}
\end{aligned} \tag{3.9}$$

and

$$\begin{aligned}
&C_n(p,j,l,M) \\
&:= \sum_{|\bar{n}| < (p-l)(m-1)+M} \left| \sum_{\substack{\sum n_i = \bar{n} \\ |n_0| < M \\ |n_1|, \dots, |n_{p-l}| < m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right| \left(\sum_{\substack{\sum n_i + \bar{n} = n \\ m \le |n_1|, \dots, |n_j| < M}} \frac{A_s^j}{|n_1|^s \cdots |n_j|^s} \right).
\end{aligned} \tag{3.10}$$

For notational purposes, we also define m-vectors containing these bounds for modes $n = 0, \ldots, m-1$ as follows:

$$\epsilon_F := \begin{bmatrix} \epsilon_0 \\ \vdots \\ \epsilon_{m-1} \end{bmatrix} \quad \text{and} \quad C_F(p,j,l,M) := \begin{bmatrix} C_0(p,j,l,M) \\ \vdots \\ C_{m-1}(p,j,l,M) \end{bmatrix}.$$

Note that these bounds are computable in that they require only a finite number of computations. In addition, increasing the computational parameter M has the effect of increasing the computational work in order to decrease the bounds.

We now use bounds (3.9) and (3.10) to define radii polynomials, $P_n(r)$. These polynomials are designed to encode the bounds required by Theorem 2.1. More specifically, as is demonstrated in section 6, the polynomials are constructed so that $P_n(r) < 0$ implies that $Y_n + K_n - |\tilde{w}_n| < 0$ on the set

$$\bar{u} + W(r) = \bar{u} + \left(\prod_{k=0}^{m-1} [-r, r] \times \prod_{k=m}^{\infty} \left[-\frac{A_s}{k^s}, \frac{A_s}{k^s} \right] \right). \tag{3.11}$$

DEFINITION 3.1. *To simplify notation, the* finite radii polynomials, $P_0, \ldots, P_{m-1}$, *are given as an* m*-vector* $P_F(r) = (P_0(r), \ldots, P_{m-1}(r))^t$. *Define*

$$P_F(r) := \sum_{n=0}^{d} C_F(n) r^n, \tag{3.12}$$

where the coefficients are

$$C_F(n) := \begin{cases} C_F^Y + C_F^K(0), & n = 0, \\ C_F^K(1) - 1, & n = 1, \\ C_F^K(n), & n = 2, \ldots, d. \end{cases}$$

The right-hand terms are defined as follows. The individual terms of the vectors $C_F^K(i)$ *are chosen to satisfy*

$$\begin{aligned} C_k^K(i) \geq & \left(\sum_{l=\max\{2,i\}}^{d} \sum_{p=l}^{d} l \binom{p}{l} \binom{l}{i} |J_{F\times F}| C_F(p, l-i, l, M) \right. \\ & \left. + \begin{cases} |J_{F\times F}| \epsilon_F + V_F^{(0)}, & i = 0, \\ V_F^{(1)}, & i = 1, \\ 0, & \textit{otherwise} \end{cases} \right)_k \end{aligned} \tag{3.13}$$

and similarly

$$C_F^Y = Y_F, \tag{3.14}$$

where $|\cdot|$ *and the bounds are computed componentwise.*

Observe, again, that determining these bounds requires only a finite number of computations.

DEFINITION 3.2. *For $k \geq m$, the* tail radii polynomial *is*

$$P_k(r) = \begin{cases} \frac{|\sum_{p=0}^{d}(c_p \bar{u}^p)_k|}{|\mu_k|} + \frac{C(\bar{A},A(r))}{|\mu_k|k^s} - \frac{A_s}{k^s}, & m \leq k < m_+, \\ \frac{C_+(\bar{A},A)}{|\mu_k|k^s} - \frac{A_s}{k^s}, & k \geq m_+, \end{cases}$$

where, again,

$$(c_p \bar{u}^p)_k = \sum_{\sum n_i = k} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_p}.$$

DEFINITION 3.3. *Consider the radii polynomials consisting of the finite radii polynomials P_k, $k = 0, \ldots, m-1$, and the tail radii polynomials, P_k, $k \geq m$. A positive real number r is a* validation radius *if $P_k(r) < 0$ for all $k \geq 0$.*

The proof of the following theorem is presented in section 6.

THEOREM 3.4. *If there exists a validation radius $r > 0$ and the eigenvalues μ_k satisfy $|\mu_k| \to \infty$, then there exists a unique equilibrium solution of (3.1) in $\bar{u}+W(r)$.*

We now present a procedure for computing a validation radius that satisfies the hypotheses of Theorem 3.4. In particular, this procedure describes a natural order for defining the decay constants A_s, s, and A. The constants A_s and s reflect regularity properties of the equation and should be chosen either from numerical simulations or analysis. In this approach, we choose to treat $A = A(r)$ as a constant. The rationale for this choice is that from a computational perspective, we would like to find $r > 0$ solving simple constructions of the finite radii inequalities $P_0(r) < 0, \ldots, P_{m-1}(r) < 0$ without having to simultaneously control the more complicated effects from A on the coefficients of these polynomials as well as on the tail polynomials P_k, $k \geq m$. A practical way to achieve this goal is to set $A = A_s$ at the beginning of the procedure and then check in the end that a solution $r > 0$ to $P_0(r) < 0, \ldots, P_{m-1}(r) < 0$ also satisfies $r(m-1)^s \leq A_s$.

Here, for the sake of simplicity, we set $M = m$. If the truncation error bounds prove too large for the computations, then M should be increased as described in Remark 6.3 in section 6. Finally, we add a condition which reduces the check of the tail polynomials $P_k(r) < 0$, $k > m$, to a finite number of computations. The following procedure outlines this approach.

PROCEDURE 3.5. *Suppose that the eigenvalues μ_k are such that $|\mu_k| \to \infty$. Suppose further that we may choose $m, m_+, \bar{m} \in \mathbb{N}$, $\bar{m} \geq m_+ \geq m$, and $\bar{\mu} > 0$ such that*

1. *m is the Galerkin projection dimension used for numerical continuation;*
2. *m_+ is the parameter used in the computation of $C_+(\bar{A}, A)$; and*
3. *$\bar{m}$ measures where the tail terms are bounded from below by $\bar{\mu}$ as follows: for all $k \geq \bar{m}$, $|\mu_k| \geq |\bar{\mu}|$.*

Set $M = m$.

Remark: m should be chosen to give the expected nonzero modes along the bifurcation branch under study, and $\bar{m} = m_+ = (2d+1)(m-1)+1$ if $(c_p)_n = 0$ for all $n \neq 0$ and the eigenvalues μ_k are monotonically increasing in magnitude after $k = (2d+1)(m-1)$.

Fix the decay constants

$$s \geq 2 \quad \text{and} \quad A_s > 0. \tag{3.15}$$

Remark: In practice, A_s and s should be determined by regularity properties of the equation.

Set $A := A_s$. *Using the finite radii polynomials given in Definition* 3.1, *for* $k = 0, \ldots, m-1$, *numerically compute* $I_k := \{r > 0 \mid P_k(r) < 0\}$ *and*

$$\mathcal{I} := \bigcap_{k=0}^{m-1} I_k \ . \tag{3.16}$$

Check that $\mathcal{I} \neq \emptyset$.

Remark: If $\mathcal{I} = \emptyset$, *begin the procedure again either by choosing* m *larger or by choosing* s *larger and/or* A_s *smaller in* (3.15).

Check that there exists $\bar{r} \in \mathcal{I}$ *such that*

$$\bar{r} \leq \frac{A_s}{(m-1)^s} \ . \tag{3.17}$$

Remark: If such an $\bar{r}$ *exists, then* $A = A_s = \max\{A_s, \bar{r}(m-1)^s\}$. *This in turn implies that componentwise* $P_F(\bar{r}) < 0$. *If* $\bar{r}$ *does not exist, then begin the procedure again either by choosing* m *larger or by choosing* s *larger and/or* A_s *smaller in* (3.15).

Check the inequalities

$$P_m(\bar{r}) < 0, \ \ldots, \ P_{\bar{m}-1}(\bar{r}) < 0 \quad \text{and} \quad \frac{C(\bar{A}, A)}{|\bar{\mu}|} - A_s < 0.$$

Remark: If any of these inequalities fails, begin the procedure again either by choosing m *larger or by choosing* s *larger and/or* A_s *smaller in* (3.15).

Observe that if Procedure 3.5 is successful, the hypotheses of Theorem 3.4 are satisfied with validation radius $\bar{r}$.

4. Computational cost. We now provide a rough comparison of the cost of continuation with the cost of validated continuation for PDEs of the form

$$u_t = L(u, \nu) - u^3 \ . \tag{4.1}$$

Since the degree of the polynomial nonlinearity in (4.1) is cubic and we use a Newton-like operator in the continuation procedure, the most expensive terms of the computation involve m^3 operations, where m is the number of modes used in the Galerkin projection

$$f_k^{(m)}(u_F, \nu) = \mu_k(\nu) u_k - \sum_{\substack{n_1+n_2+n_3=k \\ |n_i|<m}} u_{n_1} u_{n_2} u_{n_3}, \qquad k = 0, \ldots, m-1. \tag{4.2}$$

With this in mind we count the number of m^3 operations for both approaches to obtain an estimate for the asymptotic costs and conclude with statistics obtained from calculations for the Swift–Hohenberg and Cahn–Hilliard equations.

4.1. Cost of continuation. We decompose the analysis of the cost of continuation into four steps, assuming that we begin with an approximate zero x_0 at ν_0.

Step 1. In order to get the Euler predictor (3.4), we need to evaluate the vector $-f_x^{(m)}(x_0, \nu_0)^{-1} f_\nu^{(m)}(x_0, \nu_0)$. This requires computing the m by m matrix $f_x^{(m)}(x_0^{(0)}, \nu_0)$, where for $0 \leq i, j < m$,

$$\left[f_x^{(m)}(x_0^{(0)}, \nu_0)\right]_{i+1,j+1} = \delta_{i,j}\mu_i - 3\Bigg(\sum_{\substack{n_1+n_2+j=i \\ |n_i|<m}} [x_0^{(0)}]_{|n_1|} [x_0^{(0)}]_{|n_2|} + \sum_{\substack{n_1+n_2-j=i \\ |n_i|<m}} [x_0^{(0)}]_{|n_1|} [x_0^{(0)}]_{|n_2|} \Bigg).$$

This involves the evaluation of $2m^2$ sums demanding $2m-1$ multiplications and $2m-2$ additions each. Therefore, determining $f_x^{(m)}(x_0^{(0)}, \nu_0)$ requires $8m^3$ operations. Next, we compute the LU decomposition of $f_x^{(m)}(x_0^{(0)}, \nu_0)$ in order to compute the action of its inverse on $f_\nu^{(m)}(x_0, \nu_0)$. This involves $\frac{2}{3}m^3$ operations. In our case, $f_\nu^{(m)}(x_0, \nu_0) = x_0$, requiring no additional cost. The predictor is then

$$\begin{cases} x_1^{(0)} = x_0 - \Delta\nu f_x^{(m)}(x_0, \nu_0)^{-1} x_0, \\ \nu_1 \quad = \nu_0 + \Delta\nu. \end{cases}$$

Step 2. We now start the corrector. To construct the quasi-Newton operator (3.5), we need the action of the inverse of $f_x^{(m)}$ at the predictor $(x_1^{(0)}, \nu_1)$. As seen before, it costs $8m^3$ to evaluate $f_x^{(m)}(x_1^{(0)}, \nu_1)$ and $\frac{2}{3}m^3$ to compute its inverse using LU decomposition. Note that we need to compute the LU decomposition only at the first step.

Step 3. At the jth iteration of (3.5), we need to evaluate $f^{(m)}(x_1^{(j-1)}, \nu_1)$. Its ith component is

$$[f^{(m)}(x_1^{(j-1)}, \nu_1)]_i = \mu_i(\nu_1)[x_1^{(j-1)}]_i - \sum_{\substack{n_1+n_2+n_3=i \\ |n_i|<m}} [x_1^{(j-1)}]_{|n_1|}[x_1^{(j-1)}]_{|n_2|}[x_1^{(j-1)}]_{|n_3|},$$

which requires at least $3m^2$ operations to evaluate. Since $f^{(m)}$ has m components, we get a total of $3m^3$. If k is the total number of iterations of the corrector, then this step requires $3km^3$ operations.

Step 4. The corrector ends when $||f^{(m)}(x_1^{(k)}, \nu_1)|| < \texttt{tolerance}$. Let $\bar{a}_F := x_1^{(k)}$. Evaluating the function at $(\bar{u}_F, \nu_1)$ is another $3m^3$. Now, note that we have to compute the action of the inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$ to get the predictor for the next step. Recall $J_{F\times F}$ is the numerical inverse of $f_x^{(m)}(\bar{u}_F, \nu_1)$ computed as before using an LU decomposition. Explicitly computing all the coefficients in $f_x^{(m)}(\bar{u}_F, \nu_1)$ requires an extra $2m^3$ operations. We do not count the m^3 involved to get the next predictor, since that is part of the next predictor-corrector step.

Combining the costs of the four above-mentioned steps suggests that the cost of one application of the predictor-corrector algorithm is on the order of $(20+3k)m^3$, where k is the number of iterations in the quasi-Newton corrector.

4.2. Cost of validation. We now show that the extra cost of performing validation for a cubic function ($d=3$) with constant function coefficients is of the order of $6m^3$ operations, where m is the projection dimension used for continuation. The additional cost comes primarily from computing the coefficients of the radii polynomials. In the following, we construct $m_+ = d(m-1)+1 = 3m-2$ polynomials $P_0, \ldots, P_{3m-3}$ using Procedure 3.5 and calculate the associated computational cost. Both to simplify the presentation and because this is what is used to perform the computations presented in section 5, we set $\bar{m} = m_+ = d(m-1)+1$, with $|\mu_k| \geq |\mu_{\bar{m}}|$ for all $k \geq \bar{m}$, and $M = m$. As described in Procedure 3.5, $A = A_s$ and we consider fixed $s > 2$ and $A_s > 0$.

The only nonlinear term of (4.1) is a monomial of degree 3. Thus, if $p \neq 3$, then $C_k(p, j, l, M) = 0$. In addition, we have set $M = m$. Hence, if $j \neq 0$, then $C_k(p, j, l, M) = 0$ (see Remark 6.3). Therefore, the only nonzero terms of this form

are

$$C_k(3,0,l,m) = \left| \sum_{\substack{n_1+n_2+n_3=k \\ |n_1|,|n_2|,|n_3|<m}} \bar{u}_{n_1} \cdots \bar{u}_{n_{3-l}} \right|. \tag{4.3}$$

Hence, by (3.13) we set

$$C_k^K(0) \geq (|J_{F\times F}|\epsilon_F)_k + V_k^{(0)} \tag{4.4}$$

for $0 \leq k < m$, and $|\cdot|$ denotes the componentwise absolute value. Note that it is possible to get an analytic upper bound on $V_k^{(0)}$ using Lemma 6.2, in which case computing $V_k^{(0)}$ doesn't require any m^3 operations. Hence, all necessary computations for $C_F^K(0)$ are of order less than m^3. Using (3.13),

$$C_k^K(1) \geq V_k^{(1)}$$

for $0 \leq k < m$ and evaluating $V_F^{(1)}$ does not require any m^3 operations.

Finally, combining (3.13) and (4.3),

$$C_F^K(2) \geq 6|J_{F\times F}|C_F(3,0,2,m),$$

where $C_n(3,0,2,m) = |\bar{u}_n|$, and

$$C_F^K(3) \geq 3|J_{F\times F}|C_F(3,0,3,m),$$

where $C_n(3,0,3,m) = 1$.

The last coefficient to compute to get all the finite radii polynomials (3.14) is

$$C_F^Y \geq |J_{F\times F} f_F(\bar{u})|,$$

where again $|\cdot|$ denotes the componentwise absolute value. This comes with no extra m^3 cost since $f_F(\bar{u}) = f^{(m)}(\bar{u}_F, \nu_1)$ was computed in Step 4 of the predictor-corrector algorithm.

The next step in Procedure 3.5 is checking for the existence of a validation radius $r > 0$. This requires finding the numerical zeros of each of the cubic polynomials $P_0, \ldots, P_{m-1}$, constructing $I_0, \ldots, I_{m-1}$, where I_k are closed intervals such that $I_k \subsetneq \{r > 0 | P_k(r) < 0\}$, and finally checking for a nonempty intersection $\mathcal{I} = \cap_{k=0}^{m-1} I_k$. All of these steps are of order less than m^3.

Assuming there exists a positive $\bar{r} \in \mathcal{I}$ such that $\bar{r}(m-1)^s \leq A_s$, we construct and evaluate the tail radii polynomials $P_m, \ldots, P_{3m-1}$ at $\bar{r}$. We compute Y_k using (3.7), which requires $6m^3$ operations since we need to evaluate $f_k(\bar{u})$ for $k = m, \ldots, 3m-3$.

Using Definition 3.2 and the assumption that $A = A_s$, we compute

$$C(\bar{A}, A) = \sum_{l=1}^{3} l \binom{3}{l} \alpha^3 \bar{A}^{3-l} A^l = 3\alpha^3 A_s (\bar{A} + A_s)^2.$$

This latter step and the remaining computations for Procedure 3.5 are all of order less than m^3.

In summary, the m^3 cost of computing the coefficients of the radii polynomials is $6m^3$. Thus the additional cost of validation is on the order of $6m^3$ operations.

4.3. Relative cost. Combining the results of sections 4.1 and 4.2 suggests that asymptotically the ratio of the cost of validated continuation to the cost of traditional continuation is

$$\frac{26+3k}{20+3k},$$

where k is the number of iterations performed in the corrector step. We tested this hypothesis again on two fourth order PDEs with cubic nonlinearities, Swift–Hohenberg and Cahn–Hilliard. The results are discussed in greater detail in section 5. For the moment we are interested only in the relative times of computation.

We performed validated continuation for 46 predictor-corrector steps involving a total of 90 quasi-Newton iterations for the cubic Swift–Hohenberg equation. We repeated the computations without validation. The ratio of elapsed time for validated continuation to the time used for continuation alone was ≈ 1.156. Given that we had an average of 90/46 iterations per predictor-corrector step, this is close to the rough estimate of $\frac{26+3\cdot 90/46}{20+3\cdot 90/46} \approx 1.232$ given by the above arguments.

Similarly, we performed validated continuation for 15 predictor-corrector steps involving a total of 37 quasi-Newton iterations for Cahn–Hilliard. Again, we repeated the computations without validation. The ratio of elapsed time for validated continuation to the time used for continuation alone was ≈ 1.173. Given that we had an average of 37/15 iterations per predictor-corrector step, the asymptotic ratio is $\frac{26+3\cdot 37/15}{20+3\cdot 37/15} \approx 1.219$.

The results of these computations are summarized in Table 4.1.

TABLE 4.1
Comparison of the asymptotic ratios.

PDE	m	$\frac{\#\text{ iterations}}{\#\text{ steps}}$	Experimental ratio	Estimated ratio $\frac{26+3k}{20+2k}$
Swift–Hohenberg	27	1.96	1.156	1.232
Cahn–Hilliard	60	1.65	1.173	1.219

5. Sample results. To demonstrate the practical applicability of validated continuation we turn to two model problems, Cahn–Hilliard and Swift–Hohenberg. In both cases we follow a branch of equilibria and validate at each parameter value of the continuation. In the case of Swift–Hohenberg we also use interval arithmetic to evaluate the radii polynomials, thus allowing us to rigorously verify the existence and uniqueness of the equilibria.

5.1. Cahn–Hilliard. The Cahn–Hilliard equation was introduced in [1] as a model for the process of phase separation of a binary alloy at a fixed temperature. On a one-dimensional domain it takes the form

$$\begin{aligned} u_t &= -\left(\frac{1}{\nu}u_{xx} + u - u^3\right)_{xx}, \quad x \in [0,1], \\ u_x &= u_{xxx} = 0 \quad \text{at } x = 0, 1. \end{aligned} \tag{5.1}$$

The assumption of an equal concentration of both alloys is formulated as

$$\int_0^1 u(x,\cdot)\,dx = 0.$$

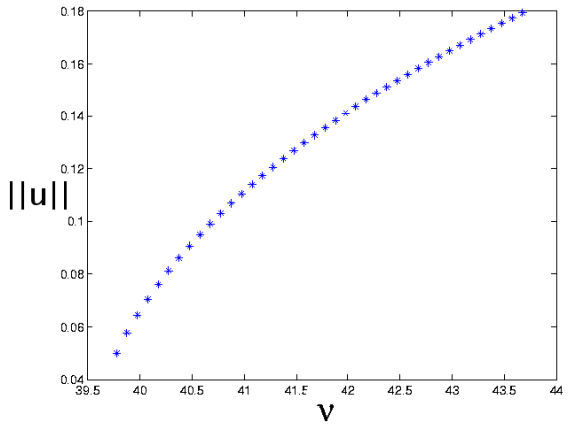


FIG. 5.1. *Validated continuation in ν for the Cahn–Hilliard equation on* $[0,1]$.

Note that when looking for the equilibrium solutions of (5.1), it is sufficient to work with the Allen–Cahn equation

$$\begin{aligned} &\frac{1}{\nu}u_{xx} + u - u^3 = 0, \\ &u_x = 0 \quad \text{at } x = 0, 1. \end{aligned} \tag{5.2}$$

Rewriting (5.2) in the form of (1.2), the linear operator is $L(\cdot,\nu) = \frac{1}{\nu}\frac{\partial^2}{\partial x^2} + 1$ and the polynomial nonlinearity is of degree $d = 3$ with coefficient functions

$$(c_p)_n = \begin{cases} -1, & p = 3 \text{ and } n = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Applying Procedure 3.5 with $M = m = 60$, $s = 3$, and $A_s = 0.01$ results in the branch of equilibria indicated in Figure 5.1, where each point represents the center of the infinite-dimensional validation set of the form $\bar{u} + W(\bar{r})$, containing a unique equilibrium of (5.1). These are the points used to obtain the cost estimates presented in Table 4.1. To avoid drowning the reader in large lists of numbers, we only provide the detailed numerical output at one parameter value.

VALIDATED RESULT 5.1. *Let* $\nu = 43.57415358799057$. *Then*

$$\bar{r} = 4.846104201261526 \times 10^{-8}$$

is a validation radius for the numerical zero $\bar{u}_F$ *given in Table* 5.1. *Thus, there exists a unique equilibrium for* (5.1) *in the validation set*

$$(\bar{u}_F, 0) + \prod_{k=0}^{59}[-\bar{r}, \bar{r}] \times \prod_{k=60}^{\infty}\left[-\frac{0.01}{k^3}, \frac{0.01}{k^3}\right].$$

5.2. Swift–Hohenberg. The Swift–Hohenberg equation

$$\begin{aligned} &u_t = f(u,\nu) = \left\{\nu - \left(1 + \frac{\partial^2}{\partial x^2}\right)^2\right\} u - u^3, \qquad u(\cdot,t) \in L^2\left(0, \frac{2\pi}{L_0}\right), \\ &u(x,t) = u\left(x + \frac{2\pi}{L_0}, t\right), \qquad u(-x,t) = u(x,t), \qquad \nu > 0, \end{aligned} \tag{5.3}$$

TABLE 5.1
The numerical zero $\bar{u}_F$ obtained by continuation for the Cahn–Hilliard equation at $\nu = 43.57415358799057$. Note that all even coefficients are 0.

k	$\bar{u}_k$
1	$1.773844149032812 \times 10^{-1}$
3	$-7.601617928785714 \times 10^{-4}$
5	$3.271672072176762 \times 10^{-6}$
7	$-1.408100160017936 \times 10^{-8}$
9	$6.060344382471457 \times 10^{-11}$
11	$-2.608320515803233 \times 10^{-13}$
13	$1.122598345048980 \times 10^{-15}$
15	$-4.831561184682242 \times 10^{-18}$
17	$2.079457485469691 \times 10^{-20}$
19	$-8.949770271275235 \times 10^{-23}$
21	$3.851880360024139 \times 10^{-25}$
23	$-1.657801422354123 \times 10^{-27}$
25	$7.134947464114615 \times 10^{-30}$
27	$-3.070770234245256 \times 10^{-32}$
29	$1.321605495419571 \times 10^{-34}$
31	$-5.687926883858248 \times 10^{-37}$
33	$2.447955395983479 \times 10^{-39}$
35	$-1.053537452697732 \times 10^{-41}$
37	$4.534120813401209 \times 10^{-44}$
39	$-1.951337823193323 \times 10^{-46}$
41	$8.397842606319005 \times 10^{-49}$
43	$-3.614086242431264 \times 10^{-51}$
45	$1.555336697148314 \times 10^{-53}$
47	$-6.693373497802139 \times 10^{-56}$
49	$2.880447985844179 \times 10^{-58}$
51	$-1.239563989182517 \times 10^{-60}$
53	$5.334225825486573 \times 10^{-63}$
55	$-2.295445428599939 \times 10^{-65}$
57	$9.877687199770852 \times 10^{-68}$
59	$-4.250458946966345 \times 10^{-70}$
≥ 60	0

was originally introduced to describe the onset of Rayleigh–Bénard heat convection [8], where L_0 is a fundamental wave number for the system size $2\pi/L_0$. The parameter ν corresponds to the Rayleigh number, and its increase is associated with the appearance of multiple solutions that exhibit complicated patterns. For the computations presented here we fixed $L_0 = 0.65$.

Rewriting (5.3) in the form of (1.2), the linear operator is $L(\cdot, \nu) = \nu - (1 + \frac{\partial^2}{\partial x^2})^2$ and the polynomial nonlinearity is of degree $d = 3$ with coefficient functions

$$(c_p)_n = \begin{cases} -1, & p = 3 \text{ and } n = 0, \\ 0, & \text{otherwise.} \end{cases}$$

Applying Procedure 3.5 with $M = m = 27$, $s = 4$, and $A_s = 0.002$ results in the branch of equilibria indicated in Figure 5.2, where each point represents the center of the infinite-dimensional validation set of the form $\bar{u} + W(\bar{r})$, containing a unique equilibrium of (5.3). Again, these are the points used to obtain the cost estimates presented in Table 4.1.

As in the case of the Cahn–Hilliard equation, we only include the output at one point on the branch of Figure 5.2.

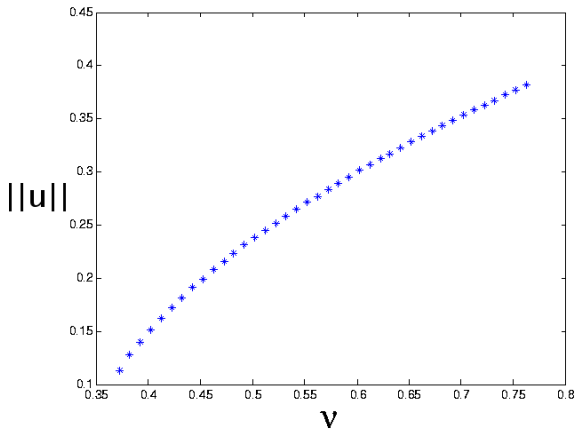


FIG. 5.2. *Validated continuation in ν for the Swift–Hohenberg equation at $L_0 = 0.65$.*

TABLE 5.2

The numerical zero $\bar{u}_F$ obtained by continuation for the Swift–Hohenberg equation at $\nu = 0.6674701641462312$ and $L_0 = 0.65$. All even coefficients are 0.

k	$\bar{u}_k$
1	$-3.359998711939212 \times 10^{-1}$
3	$4.824376413178060 \times 10^{-3}$
5	$-1.761066797314072 \times 10^{-5}$
7	$7.535865329757206 \times 10^{-8}$
9	$-2.790895103063484 \times 10^{-10}$
11	$9.411109491227775 \times 10^{-13}$
13	$-3.113936321690645 \times 10^{-15}$
15	$1.007016979585499 \times 10^{-17}$
17	$-3.200410295859874 \times 10^{-20}$
19	$1.003878817132397 \times 10^{-22}$
21	$-3.114244522738206 \times 10^{-25}$
23	$9.573156964813860 \times 10^{-28}$
25	$-2.920394630491221 \times 10^{-30}$
≥ 26	0

VALIDATED RESULT 5.2. *Let $\nu = 0.6674701641462312$. Then*

$$\bar{r} = 1.998167170445973 \ \times \ 10^{-9}$$

is a validation radius for the numerical zero $\bar{u}_F$ whose coefficient values are indicated in Table 5.2. *Thus, there exists a unique equilibrium solution for* (5.3) *in the validation set*

$$(\bar{u}_F, 0) + \prod_{k=0}^{26} [-\bar{r}, \bar{r}] \times \prod_{k=27}^{\infty} \left[-\frac{0.002}{k^4}, \frac{0.002}{k^4} \right].$$

Observe that in all the above-mentioned calculations, floating point roundoff errors have not been controlled; thus at this point one cannot claim that the validation results presented above are rigorous. However, with additional computational effort a computer-assisted proof can be obtained. To be more precise, our technique relies on the existence of a validation radius $\bar{r}$ making all radii polynomials strictly negative.

Hence, rigorous validation follows if the inequalities are satisfied when one includes bounds to control the possibility of floating point errors. The first step in checking these inequalities on this level is to obtain floating point outer bounds for the coefficients of the polynomials. This can be done by defining each entry of

$$\bar{u}_F,\ f^{(m)}(\bar{u}_F,\nu),\ J_{F\times F},\ f_x^{(m)}(\bar{u}_F,\nu),\ \mu_k(\nu),\ A_s,\ \text{and } s$$

to be an interval and then computing (3.13), (3.14), and the quantities in Definition 3.2 using interval arithmetic. The resulting radii polynomials, which we denote by $\tilde{P}_k$, have interval coefficients. Let $\bar{r}$ be the smallest representable number such that using interval arithmetic, the corresponding finite radii polynomials may be shown to be strictly contained in $(-\infty,0)$. Assume such an $\bar{r}$ exists. If, again using interval arithmetic, $\bar{r}(m-1)-A_s\subset(-\infty,0)$ and the intervals obtained from evaluating tail radii polynomials at $\bar{r}$ are strictly contained in $(-\infty,0)$, i.e., $\tilde{P}_k(\bar{r})\subset(-\infty,0)$ for all $k\geq m$, then the hypotheses of Theorem 3.4 are satisfied and we obtain a proof.

The above-mentioned computations were performed using the interval arithmetic package in MATLAB. Thus, we can state the following theorem.

THEOREM 5.3. *Each point in Figure* 5.2 *represents the center of an infinite-dimensional set of the form*

$$\bar{u}_F+\prod_{k=0}^{26}[-\bar{r},\bar{r}]\times\prod_{k=27}^{\infty}\left[-\frac{0.002}{k^4},\frac{0.002}{k^4}\right]$$

containing a unique equilibrium to (5.3).

The actual values for the various numerical zeros and validation radii are of limited interest and thus not presented. Of greater interest is understanding how large the errors induced by the floating point computations are as opposed to the magnitudes of the floating point computations of $P_k(\bar{r})$, $k\geq 0$, where $\bar{r}$ is the validation radius.

Let us restrict our attention to the equilibrium described by Validated Result 5.2. Following Procedure 3.5 at this parameter value, beginning using radii polynomials with interval coefficients, and performing the computations with interval arithmetic leads to an interval of potential validation radii

$$\mathcal{I}=[3.373873850437414\times10^{-9},9.003755731999980\times10^{-4}].$$

Hence, we choose $\bar{r}=3.373873850437415\times10^{-9}$. There are 53 inclusions that need to be satisfied, those arising from the $2m-2=52$ tail radii polynomials with interval coefficients and the one associated with inequality (3.17). The fact that the inclusions are satisfied leads to the conclusion of Theorem 5.3 at this parameter value. Again, rather than listing all 53 inclusions, let us focus on the two extremes, the interval closest to 0,

$$\tilde{P}_{27}(\bar{r})=-3.191484496597115\times10^{-11}\pm7.037497555236307\times10^{-24},$$

and the interval the farthest from 0,

$$-1.973098298147102\times10^{-3}\pm8.673617379884037\times10^{-19},$$

corresponding to inequality (3.17). Observe that in both cases, the width of the interval induced by the floating point errors is more than ten orders of magnitude smaller than the value of the center. Furthermore, this behavior is typical for all the validation computations that were performed. This suggests that it is reasonably safe to assume that a validated equilibrium is a true equilibrium.

6. Justification of radii polynomials. In this section, we describe the construction of the radii polynomials that were defined in section 3.2 and encode the bounds required for Theorem 2.1. We begin by computing the required bounds Y_n and K_n in (2.5) for the Newton-like operator constructed in (3.6).

Using a Taylor expansion of the Newton-like operator $T(u) = u - Jf(u)$ around the numerical equilibrium $\bar{u} = (\bar{u}_F, 0, 0, \dots)$ leads to

$$\begin{aligned}
&DT(\bar{u}+w')w = [I - J\cdot Df(\bar{u}+w')]\,w \\
&= \left(I - J\left(Df(\bar{u}) + D^2f(\bar{u})(w') + \cdots + \frac{D^l f(\bar{u})}{(l-1)!}(w')^{l-1} + \cdots + \frac{D^d f(\bar{u})}{(d-1)!}(w')^{d-1}\right)\right) w \\
&= [I - J\cdot Df(\bar{u})]w - J\left(\sum_{l=2}^{d}\frac{D^l f(\bar{u})}{(l-1)!}(w')^{l-1}\right) w \\
&= [I - J\cdot Df(\bar{u})]\,w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d}\frac{p!c_p\bar{u}^{p-l}(w')^{l-1}}{(l-1)!(p-l)!}\right) w \\
&= [I - J\cdot Df(\bar{u})]\,w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l}c_p\bar{u}^{p-l}(w')^{l-1}\right) w\ .
\end{aligned}$$

In the rest of the section, we will make use of the discrete convolution of bi-infinite vectors, i.e., considering two bi-infinite vectors $(a_j)_{j\in\mathbb{Z}}$, $(b_j)_{j\in\mathbb{Z}}$, we define their convolution by

$$(a*b)_k = \sum_{n=-\infty}^{\infty} a_n b_{k-n} = \sum_{\substack{k_1+k_2=k\\ k_i\in\mathbb{Z}}} a_{k_1}b_{k_2}, \quad k\in\mathbb{Z}\ .$$

Expanding into Fourier modes, we can write the nonlinear part in terms of convolution:

$$\begin{aligned}
DT(\bar{u}+w')w &= [I - J\cdot Df(\bar{u})]\,w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l}c_p\bar{u}^{p-l}(w')^{l-1}\right) * w \\
&= [I - J\cdot Df(\bar{u})]\,w - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l}(c_p\bar{u}^{p-l}) * (w')^{l-1} * w\right)\ . \qquad (6.1)
\end{aligned}$$

Thus,

$$(c_p\bar{u}^{p-l})*((w')^{l-1})*w = \left[\sum_{\bar{n}}\left(\sum_{\sum n_i=\bar{n}}(c_p)_{n_0}\bar{u}_{n_1}\cdots\bar{u}_{n_{p-l}}\right)\left(\sum_{\sum n_i=n-\bar{n}} w'_{n_1}\cdots w'_{n_{l-1}}w_{n_l}\right)\right]_n .$$

Here, $[\cdot]_n$ denotes the bi-infinite vector indexed by $n\in\mathbb{Z}$ and $(\cdot)_k$ denotes the entry at index k.

We use this expansion to compute the bounds

$$\begin{aligned}
K_k &\geq \max|(DT(\bar{u}+W)W)_k| \\
&\geq \max\left|[I - J\cdot Df(\bar{u})]\tilde{w} - J\left(\sum_{l=2}^{d}\sum_{p=l}^{d} l\binom{p}{l}(c_p\bar{u}^{p-l}) * \tilde{w}^l\right)\right|,
\end{aligned}$$

where, as in section 2, W has the form (2.1).

The block diagonal structure of J allows us to decompose (6.1) into a finite, m-dimensional piece and the infinite-dimensional tail terms. For the following, we adopt the notation $[\cdot]_F$ to denote the m-vector whose nth entry is computed at index value $n-1$ for $1 \le n \le m$, the subscript $\tilde{F}$ to denote the bi-infinite vector in which the kth entries for $|k| \ge m$ are set equal to 0, and the subscript $\tilde{I}$ to denote the bi-infinite vector in which the kth entries for $|k| < m$ are set equal to 0. We begin with the following decomposition of the finite part of the linear term:

$$\begin{aligned}
\{[I - J \cdot Df(\bar{u})]w\}_F &= w_F - [J \cdot Df(\bar{u})w]_F \\
&= w_F - J_{F\times F}\,[Df(\bar{u})w]_F \\
&= w_F - J_{F\times F} \cdot Df_F(\bar{u})w \\
&= w_F - J_{F\times F} \cdot \left[Df^{(m)}(\bar{u}_F)w_F + R_F(\bar{u}, w)\right] \\
&= \left[I_{F\times F} - J_{F\times F} \cdot Df^{(m)}(\bar{u}_F)\right] w_F - J_{F\times F} \cdot R_F(\bar{u}, w)\ , \qquad (6.2)
\end{aligned}$$

where for $k \in \{0, \dots, m-1\}$,

$$\begin{aligned}
R_k(\bar{u}, w) &:= \sum_{i=m}^{\infty} \frac{\partial f_k}{\partial u_i}(\bar{u}) w_i \\
&= \sum_{\substack{\bar{n}=-\infty \\ |k-\bar{n}|\ge m}}^{\infty} \left| \sum_{p=1}^{d} p \sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \dots \bar{u}_{n_{p-1}} \right| \frac{A_s}{|k-\bar{n}|^s}\ . \qquad (6.3)
\end{aligned}$$

It follows that

$$\begin{aligned}
[DT(\bar{u}+W)W]_F \subseteq &\left[I_{F\times F} - J_{F\times F} \cdot Df^{(m)}(\bar{u}_F)\right] \tilde{w}_F - J_{F\times F} \cdot R_F(\bar{u}, w) \\
&- \left(J_{F\times F} \sum_{l=2}^{d} \sum_{p=l}^{d} l \binom{p}{l} [(c_p \bar{u}^{p-l}) * \tilde{w}^l]_F \right). \qquad (6.4)
\end{aligned}$$

For $k \ge m$,

$$(DT(\bar{u}+W)W)_k \subseteq -J(k,k) \sum_{l=1}^{d} \sum_{p=l}^{d} l \binom{p}{l} ((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k. \qquad (6.5)$$

We now focus on finding bounds on the terms given in (6.4) and (6.5). First consider

$$((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k = \sum_{\bar{n}} \left(\sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right) \left(\sum_{\sum n_i + \bar{n} = k} \tilde{w}_{n_1} \cdots \tilde{w}_{n_l} \right), \qquad (6.6)$$

where p is the degree of the original monomial term of f and $l \in \{1, \dots, p\}$ is the order of the derivative being taken. One upper bound for (6.6) is given in the following lemma, which uses asymptotic bounds first listed in section 3.2.

LEMMA 6.1. *Let* $\alpha = \frac{2}{s-1} + 2 + 3.5 \cdot 2^s$, $\bar{u}_k \in \frac{\bar{A}}{k^s}[-1,1]$, $(c_p)_k \in \frac{C_p}{k^s}[-1,1]$, *and* $\tilde{w}_k \subset \frac{A}{k^s}[-1,1]$ *for all* k. *Then*

$$((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k \subseteq \begin{cases} \frac{\alpha^p C_p \bar{A}^{p-l} A^l}{|k|^s}[-1,1], & k \ne 0, \\ \alpha^p C_p \bar{A}^{p-l} A^l [-1,1], & k = 0. \end{cases}$$

Proof. Note that

$$\sum_{\bar{n}} \left(\sum_{\sum n_i = \bar{n}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right) \left(\sum_{\sum n_i + \bar{n} = k} \tilde{w}_{n_1} \cdots \tilde{w}_{n_l} \right)$$
$$\subseteq \sum_{\sum n_i = k} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \tilde{w}_{n_{p-l+1}} \cdots \tilde{w}_{n_p}$$
$$\subseteq \sum_{\sum n_i = k} \frac{C_p}{|n_0|^s} \frac{\bar{A}}{|n_1|^s} \cdots \frac{\bar{A}}{|n_{p-l}|^s} \frac{A}{|n_{p-l+1}|^s} \cdots \frac{A}{|n_p|^s} [-1, 1].$$

The remainder of the proof is a modification of [2, Lemma 5.8]. ∎

In most cases, especially when l is small relative to p, this bound will be too large to use for the low modes. In particular, $\bar{u}$ may be far from zero, resulting in a large constant $\bar{A}$. By taking k sufficiently large, the contraction given by $J(k,k) \approx {\mu_k}^{-1}$ will overcome the large bound. A more practical approach for obtaining bounds for the low modes is given by the following lemma. For flexibility in balancing numerical computations (requiring a finite number of operations) with analysis (to obtain truncation bounds), we choose $M \geq m$ to be the dimension used to split these sums.

LEMMA 6.2. *For $M \geq m$,*

$$((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k \subseteq \left(\sum_{j=0}^{l} \binom{l}{j} C_k(p, j, l, M) r^{l-j} + \epsilon_k(p, l, M) \right) [-1, 1].$$

Proof. This lemma is a modification of [2, Lemma 5.10] combined with Lemma 6.1. In [2, Lemma 5.10], the bound is split into finite sums and the tail term, bounded by

$$\frac{p \alpha^{p-1} C_p \bar{A}^{p-l} A^l}{(M-1)^{s-1}(s-1)} \left[\frac{1}{(M-k)^s} + \frac{1}{(M+k)^s} \right] [-1, 1].$$

We obtain a polynomial in r by rewriting the finite sums as follows:

$$\sum_{\bar{n}} \left(\sum_{\substack{\sum n_i = \bar{n} \\ |n_i| < M}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right) \left(\sum_{\substack{\sum n_i + \bar{n} = k \\ |n_i| < M}} \tilde{w}_{n_1} \cdots \tilde{w}_{n_l} \right)$$
$$= \sum_{\bar{n}} \left(\sum_{\substack{\sum n_i = \bar{n} \\ |n_1|, \ldots, |n_{p-l}| < m \\ |n_0| < M}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right) \left(\sum_{\substack{\sum n_i + \bar{n} = k \\ |n_i| < M}} \tilde{w}_{n_1} \cdots \tilde{w}_{n_l} \right)$$
$$= \sum_{\bar{n}} \left(\sum_{\substack{\sum n_i = \bar{n} \\ |n_1|, \ldots, |n_{p-l}| < m \\ |n_0| < M}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right) \left(\sum_{j=0}^{l} \binom{l}{j} \sum_{\substack{\sum n_i + \bar{n} = k \\ m \leq |n_1|, \ldots, |n_j| < M \\ |n_{j+1}|, \ldots, |n_l| < m}} \tilde{w}_{n_1} \cdots \tilde{w}_{n_l} \right)$$
$$= \sum_{\bar{n}} \left(\sum_{\substack{\sum n_i = \bar{n} \\ |n_1|, \ldots, |n_{p-l}| < m \\ |n_0| < M}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right)$$

$$\times \left(\sum_{j=0}^{l} \binom{l}{j} r^{l-j} [-1,1] \sum_{\substack{\sum n_i + \bar{n} = k \\ m \le |n_1|, \ldots, |n_j| < M \\ |n_{j+1}|, \ldots, |n_l| < m}} \frac{A_s^j}{|n_1|^s \cdots |n_j|^s} \right)$$

$$= \sum_{j=0}^{l} \binom{l}{j} r^{l-j} \sum_{|\bar{n}| < (p-l)(m-1)+M} [-1,1] \left| \sum_{\substack{\sum n_i = \bar{n} \\ |n_1|, \ldots, |n_{p-l}| < m \\ |n_0| < M}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right|$$

$$\times \left(\sum_{\substack{\sum n_i + \bar{n} = k \\ m \le |n_1|, \ldots, |n_j| < M \\ |n_{j+1}|, \ldots, |n_l| < m}} \frac{A_s^j}{|n_1|^s \cdots |n_j|^s} \right). \qquad \square$$

REMARK 6.3. *Note that in Lemma* 6.2*, $C_k(p,j,l,M)$ captures the contribution to the $(l-j)$th polynomial coefficient from the lth derivative of the pth monomial term of f in the Taylor expansion. If $M = m$, then $C_k(p,j,l,M) = 0$ for all $j > 0$ and*

$$C_k(p,0,l,m) = \left| \sum_{\substack{n_0 + \cdots + n_{p-l} = k \\ |n_0|, \ldots, |n_{p-l}| < m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_{p-l}} \right| .$$

For $M > m$ there is also a (small) contribution to the coefficients of higher degrees of r in the polynomials, while simultaneously decreasing the ϵ_k term. This offers a method for using additional computations to decrease the bound ϵ_k if this bound proves to be too large for the validation procedure.

For notational purposes, set ϵ_F, $C_F(p,j,l,M)$, $V_F^{(0)}$, and $V_F^{(1)}$, to be the m-vectors as defined in section 3.2. For $0 \le k < m$, we substitute the bounds from Lemma 6.2 into (6.4),

$$\begin{aligned}
&(DT(\bar{u}+W)W)_k \subseteq r V_k^{(1)}[-1,1] + V_k^{(0)}[-1,1] \\
&\quad + \left(-J_{F\times F} \sum_{l=2}^{d} \sum_{p=l}^{d} l \binom{p}{l} \left(\sum_{j=0}^{l} \binom{l}{j} (C_F(p,j,l,M) r^{l-j} + \epsilon_F(p,l,M)) \right) [-1,1] \right)_k \\
&= \left(|J_{F\times F}| \epsilon_F \right)_k [-1,1] + r V_k^{(1)}[-1,1] + V_k^{(0)}[-1,1] \\
&\quad + \left(\sum_{l=2}^{d} \sum_{p=l}^{d} \sum_{j=0}^{l} r^{l-j} l \binom{p}{l} \binom{l}{j} |J_{F\times F}| C_F(p,j,l,M) \right)_k [-1,1] \\
&= \left(|J_{F\times F}| \epsilon_F + V_F^{(0)} \right)_k [-1,1] + r V_k^{(1)}[-1,1] \\
&\quad + \left(\sum_{i=0}^{d} r^i \sum_{l=\max\{2,i\}}^{d} \sum_{p=l}^{d} l \binom{p}{l} \binom{l}{i} |J_{F\times F}| C_F(p,l-i,l,M) \right)_k [-1,1],
\end{aligned}$$

where $|\cdot|$ denotes entrywise absolute value. For $0 \le k < m$, set K_k to be

$$K_k := \sum_{i=0}^{d} C_k^K(i) r^i \ge |(DT(\bar{u}+W)W)_k|,$$

where $C_k^K(i)$ satisfies (3.13).

Recall that our goal is to find a polynomial bound for $Y_k + K_k - |\tilde{w}_k|$ for Theorem 2.1. This requires also computing the bounds for Y_k satisfying the following equation:

$$\begin{aligned} Y_k &\geq |(T(\bar{u}) - \bar{u})_k| \\ &= |[-Jf(\bar{u})]_k| \\ &= \left| \left(-J \left[\mu_n \bar{u}_n + \sum_{p=0}^{d} \sum_{\substack{n_0+\cdots+n_p=n \\ |n_1|,\ldots,|n_p|<m}} (c_p)_{n_0} \bar{u}_{n_1} \cdots \bar{u}_{n_p} \right]_n \right)_k \right|. \end{aligned} \tag{6.7}$$

Therefore, for $k < m$, set $Y_k = C_k^Y$, where C_F^Y is given by (3.14). Note that these terms involve the Galerkin projection of f at $\bar{u}$ onto the first m modes and, therefore, are expected to be small.

For $0 \leq k < m$, we now combine our bounds for Y_k with the bounds for K_k to compute the coefficients of the polynomials $P_k(r)$ giving the bounds $Y_k + K_k - |\tilde{w}_k|$. This leads us to the definition of the finite radii polynomials presented in Definition 3.1.

In modes $k \geq m$, we use Lemma 6.1 and (6.5) to obtain

$$\begin{aligned} (DT(\bar{u}+W)W)_k &\subseteq -J(k,k) \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l \binom{p}{l} ((c_p \bar{u}^{p-l}) * \tilde{w}^l)_k \\ &\subseteq \frac{1}{|\mu_k| k^s} \sum_{l=1}^{d} \sum_{p=\max\{2,l\}}^{d} l \binom{p}{l} \alpha^p C_p \bar{A}^{p-l} A^l [-1,1]. \end{aligned} \tag{6.8}$$

Therefore, set K_k, $k \geq m$, such that

$$K_k \geq \frac{C(\bar{A}, A)}{|\mu_k| k^s}. \tag{6.9}$$

Recall (6.7). For $k \geq m$, choose Y_k (compare with (3.7)) such that

$$\begin{aligned} Y_k &\geq |(T(\bar{u}) - \bar{u})_k| \\ &= |-J(k,k)(f_k(\bar{u}))| \\ &= \frac{|\sum_{p=2}^{d} (c_p \bar{u}^p)_k|}{|\mu_k|}. \end{aligned} \tag{6.10}$$

Using Lemma 6.1,

$$\frac{|\sum_{p=2}^{d} (c_p \bar{u}^p)_k|}{|\mu_k|} \subseteq \sum_{p=2}^{d} \frac{\alpha C_p \bar{A}^p}{|\mu_k||k|^s} [-1,1]. \tag{6.11}$$

These bounds are overestimates and should only be used for large k. In fact, if the coefficient functions c_p have finite Fourier expansions (as in the examples we consider in section 5), then $Y_k = 0$ for k sufficiently large.

We may now define the polynomial bounds for $Y_k + K_k - |\tilde{w}_k|$ in the tail modes. Suppose the bounds Y_k are numerically or analytically computed for $m \leq k < m_+$.

Then for $k \geq m$, the tail radii polynomial (see Definition 3.2) satisfies

$$P_k(r) = Y_k + K_k(r) - \frac{A_s}{k^s}$$
$$= \begin{cases} \frac{|\sum_{p=2}^{d}(c_p \bar{u}^p)_k|}{|\mu_k|} + \frac{C(\bar{A},A)}{|\mu_k|k^s} - \frac{A_s}{k^s}, & m \leq k < m_+, \\ \frac{C_+(\bar{A},A)}{|\mu_k|k^s} - \frac{A_s}{k^s}, & k \geq m_+, \end{cases}$$

Checking that $P_k < 0$ for $k \geq m$ reduces to checking the inequalities $P_m < 0, \ldots, P_{m_+-1} < 0$ and, by rearranging terms,

$$C_+(\bar{A}, A) < |\mu_k| A_s. \tag{6.12}$$

Therefore, the assumption that $|\mu_k|$ is growing in k ensures that (6.12) may be verified for all $k \geq m$ with only a finite number of checks. More explicitly, computing a lower bound on $|\mu_k|$, $k \geq m_+$, would allow us to verify all inequalities of type (6.12), $k \geq m_+$, in one step. Indeed, since $\frac{C_+(\bar{A},A)}{|\bar{\mu}|} - A_s < 0$, and $f_k(\bar{u}) = 0$ and $|\mu_k| \geq |\bar{\mu}|$ for all $k \geq \bar{m} \geq m_+$,

$$\begin{aligned} P_k(\bar{r}) &= Y_k + K_k - \frac{A_s}{k^s} \\ &= \frac{C_+(\bar{A},A)}{|\mu_k|k^s} - \frac{A_s}{k^s} \\ &\leq \frac{C_+(\bar{A},A)}{|\bar{\mu}|k^s} - \frac{A_s}{k^s} \\ &< 0. \end{aligned}$$

We have now constructed the radii polynomials to give the bounds required for Theorem 2.1.

Recall that $r > 0$ is a *validation radius* if $P_k(r) < 0$ for all radii polynomials P_k as defined in Definitions 3.1 and 3.2. We may now prove Theorem 3.4 from section 3.2.

Proof of Theorem 3.4. The radii polynomials have been constructed so that $P_k(r) < 0$ for all k ensures that the first condition of Theorem 2.1 is satisfied. Since the first condition is satisfied, we also have that $\frac{K_k}{|\tilde{w}_k|} < 1$ for all k. Finally, since $|\mu_k| \to \infty$,

$$\frac{K_k}{|\tilde{w}_k|} = \frac{\frac{C_+(\bar{A},A)}{|\mu_k|k^s}}{\frac{A_s}{k^s}} = \frac{C_+(\bar{A},A)}{A_s|\mu_k|} \to 0.$$

Therefore, $K := \sup\left\{\frac{K_k}{|\tilde{w}_k|}\right\} < 1$ and the second and final hypothesis in Theorem 2.1 is also satisfied.

7. Concluding remarks. As is indicated in the introduction, the purpose of this paper is to communicate the essential ideas of our proposed validation method. As such we have presented it in a somewhat limited setting. Thus, we conclude with a range of comments, beginning with obvious generalizations describing ongoing work and ending with some open questions.

The particular choice of the abstract expression for the expansion of the PDE (1.3) was chosen because it was appropriate for the application to Cahn–Hilliard (5.1) and Swift–Hohenberg (5.3). Hopefully it is clear that a different choice of boundary conditions or symmetries does not affect the essential estimates. It is expected,

but remains to be checked, that the form of the estimates can be lifted to parabolic PDEs on rectangular domains (see [6], where similar estimates were used to study the equilibria of the Cahn–Hilliard equation on the unit square) and to systems of such PDEs. We also believe that generalizing this technique to pseudoarclength continuation should be fairly straightforward. Furthermore, treating the parameter ν as an interval allows us to prove the existence and uniqueness of a branch of solutions over the interval $\tilde{\nu}$. By adapting the predictor step length, this approach may be used to prove existence and uniqueness along continuous, finite branches of equilibria.

While there are numerous directions in which our validation technique can be expanded or improved, we focus on the following four.

- Observe that if (1.2) has a polynomial nonlinearity of order d, then straightforward evaluation of the nonlinear term in (1.4) involves on the order of m^d operations. In a forthcoming work [5], this computational cost is reduced by the use of the fast Fourier transform.
- For the computations presented in this paper, we fixed $M = m$. This was done for the sake of simplicity of presentation. Clearly, the success of validation strongly depends on upper bounds presented in Lemma 6.2. In general, for fixed m, choosing $M > m$ increases the computational cost but provides a smaller bound for the truncation error ϵ_k. Improved bounds should facilitate validated continuation with a smaller projection dimension m, which decreases the computational cost. The exact tradeoff is currently being explored.
- The computational strategy adopted for this work is to fix A_s and s throughout the continuation procedure. In particular, in the Swift–Hohenberg example we obtained 40 successful predictor-corrector steps with $A_s = 0.002$ and $s = 4$ held constant over a parameter range of length 0.4. We were able to do this because we chose a projection dimension $m = 27$, which is unnecessarily large. For example, with $m = 11$, $A_s = 0.002$, and $s = 4.52$ we were able to perform a validated continuation over a parameter range of length 0.01. In this case, we obtained $s = 4.52$ by fixing A_s and seeking a successful s by trial and error. This suggests that it is worthwhile to develop a method for choosing A_s and s adaptively during the validated continuation procedure.
- As pointed out in section 5, the floating point errors are many orders of magnitude smaller than the magnitude of the radii polynomials evaluated at the validation radius. This suggests that it might be possible to compute a priori bounds on the floating point errors from which one could conclude that the validation computations are in fact rigorous computations. The techniques in [7] might prove useful for this purpose.

Acknowledgments. The authors would like to thank L. Dieci for numerous helpful conversations concerning continuation methods and M. Gameiro, R. Beardmore, and the referees for helpful comments on the layout of the paper.

REFERENCES

[1] J. W. Cahn and J. E. Hilliard, *Free energy of a nonuniform system* I. *Interfacial free energy*, J. Chem. Phys., 28 (1958), pp. 258–267.

[2] S. Day, *A Rigorous Numerical Method in Infinite Dimensions*, Ph.D. thesis, Georgia Institute of Technology, Atlanta, 2003.

[3] S. Day, Y. Hiraoka, K. Mischaikow, and T. Ogawa, *Rigorous numerics for global dynamics: A study of the Swift–Hohenberg equation*, SIAM J. Appl. Dyn. Syst., 4 (2005), pp. 1–31.

[4] Z. Galias and P. Zgliczyński, *An interval method for finding fixed points and periodic orbits of infinite dimensional discrete dynamical systems*, Internat. J. Chaos Birfur., to appear.
[5] M. Gameiro, J.-P. Lessard, and K. Mischaikow, *Rigorous Continuation over Long Parameter Ranges for Equilibria of PDEs*, preprint.
[6] S. Maier-Paape, U. Miller, K. Mischaikow, and T. Wanner, *Structure of the Cahn-Hilliard equation on a square*, Internat. J. Chaos Birfur., to appear.
[7] M. Mrozek, *Rigorous error analysis of numerical algorithms via symbolic computations*, J. Symbolic Comput., 22 (1996), pp. 435–458.
[8] J. B. Swift and P. C. Hohenberg, *Hydrodynamic fluctuations at the convective instability*, Phys. Rev. A (3), 15 (1977), p. 319.
[9] N. Yamamoto, *A numerical verification method for solutions of boundary value problems with local uniqueness by Banach's fixed-point theorem*, SIAM J. Numer. Anal., 35 (1998), pp. 2004–2013.
[10] P. Zgliczyński and K. Mischaikow, *Rigorous numerics for partial differential equations: The Kuramoto-Sivashinsky equation*, Found. Comput. Math., 1 (2001), pp. 255–288.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1425–1452

CONVERGENCE OF A NUMERICAL METHOD FOR SOLVING DISCONTINUOUS FOKKER–PLANCK EQUATIONS*

HONGYUN WANG†

Abstract. In studies of molecular motors, the stochastic motion is modeled using the Langevin equation. If we consider an ensemble of motors, the probability density is governed by the corresponding Fokker–Planck equation. Average quantities, such as average velocity, effective diffusion coefficient, and randomness parameter, can be calculated from the probability density. A numerical method was previously developed to solve Fokker–Planck equations [H. Wang, C. Peskin, and T. Elston, *J. Theoret. Biol.*, 221 (2003), pp. 491–511]. It preserves detailed balance, which ensures that if the system is forced to an equilibrium, the numerical solution will be the same as the Boltzmann distribution. Here we study the convergence of this numerical method when the potential has a finite number of discontinuities at half numerical grid points. We prove that this numerical method is stable and is consistent with the differential equation. Based on the consistency analysis, we propose a modified version of this numerical method to eliminate the first order error term caused by the discontinuity. We also show that in the presence of discontinuities, detailed balance is a necessary condition for converging to the correct solution. This explains why the central difference method converges to a wrong solution.

Key words. Fokker–Planck equation, detailed balance, consistency, stability, convergence

AMS subject classification. 65M

DOI. 10.1137/050639442

1. Introduction.

Molecular motors operate in an environment dominated by thermal fluctuations [1]. In general, a molecular motor has many internal and external degrees of freedom. Of these degrees of freedom, there is one associated with the main function of the motor, its unidirectional motion. For example, the γ shaft of an F_1 ATPase rotates with respect to the hexamer formed by three pairs of α and β subunits [2, 3, 4], and a kinesin dimer walks along a microtubule [5, 6]. In studies of molecular motors, it is natural to follow the motor along the dimension of its unidirectional motion [7, 8, 9]. The effects of the other degrees of freedom are modeled in the mean field potential affecting the unidirectional motion.

To introduce the governing equations for molecular motors, we start with the one-dimensional motion of a small particle in a fluid environment subject to a potential, $V(x)$, where x is the coordinate along the dimension of motion. The particle is subject to the viscous drag force, the force derived from the potential, and the Brownian force. Both the drag force and the Brownian force are the results of bombardments by surrounding fluid molecules. The drag force is the mean, and the Brownian force is the fluctuation part of the random force caused by bombardments. The particle is governed by Newton's second law:

$$m\frac{dv}{dt} = -\zeta v - V'(x) + \sqrt{2k_BT\zeta}\,\frac{dW(t)}{dt}, \tag{1.1}$$

where m is the mass and v the velocity of the particle. In (1.1), $W(t)$ is the Weiner process. The drag force on the particle, ζv, is proportional to the velocity, and ζ is

*Received by the editors September 2, 2005; accepted for publication (in revised form) January 16, 2007; published electronically July 18, 2007. This work was partially supported by National Science Foundation grant DMS-0317937.

http://www.siam.org/journals/sinum/45-4/63944.html

†Department of Applied Mathematics and Statistics, Mail Stop SOE2, University of California Santa Cruz, Santa Cruz, CA 95064 (hongwang@ams.ucsc.edu).

called the drag coefficient. The magnitude of the Brownian force is related to the drag coefficient and is given by $\sqrt{2k_BT\zeta}$, which is a result of the fluctuation-dissipation theorem [11, 13]. Here k_B is the Boltzmann constant and T the absolute temperature.

For a bead of radius a, the drag coefficient and the mass, respectively, are [1]

$$\zeta = 6\pi\eta a, \qquad m = \frac{4}{3}\pi\rho a^3, \tag{1.2}$$

where ρ is the density and η the viscosity of the surrounding fluid. Equation (1.1) can be written as

$$\frac{dv}{dt} = -\frac{\zeta}{m}\left(v - \left[-\frac{1}{\zeta}V'(x) + \sqrt{2D}\frac{dW(t)}{dt}\right]\right), \tag{1.3}$$

where $D = \frac{k_BT}{\zeta}$ is the diffusion constant [1]. It is important to notice that the ratio $\frac{\zeta}{m}$ is inversely proportional to the square of the radius of particle

$$\frac{\zeta}{m} = \frac{9\eta}{2\rho a^2} = O\left(\frac{1}{a^2}\right). \tag{1.4}$$

Thus, for a small particle, $\frac{\zeta}{m}$ is very large. In this case, (1.3) is well approximated by

$$v = \left[-\frac{1}{\zeta}V'(x) + \sqrt{2D}\frac{dW(t)}{dt}\right]. \tag{1.5}$$

The reduction from (1.3) to (1.5) in the limit of large $\frac{\zeta}{m}$ is called the Kramers–Smoluchowski approximation. Writing (1.5) as a stochastic differential equation for x, we have

$$\frac{dx}{dt} = -\frac{1}{\zeta}V'(x) + \sqrt{2D}\frac{dW(t)}{dt}. \tag{1.6}$$

This is the Langevin equation without the inertia term, governing the stochastic motion of a small particle subject to potential $V(x)$ [12].

In molecular motors, the mechanical motion is coupled to the chemical reaction. The general mathematical framework used in modeling molecular motors is a system of coupled Langevin equations. Each Langevin equation in the coupled system has the form of (1.6) with a periodic potential $V_S(x)$, where S represents the current chemical state of the motor system [7, 9, 4]:

$$\frac{dx}{dt} = -\frac{1}{\zeta}V_S'(x) + \sqrt{2D}\frac{dW(t)}{dt}. \tag{1.7}$$

Here $1 \le S \le N$, and N is the number of possible chemical states of the motor system. The period of these potentials is usually determined by the step size of the motor. For example, a kinesin dimer walks on a microtubule with 8-nm steps [6]. The chemical reaction of the motor system (the stochastic jumping of the motor system among the chemical states) is governed by a discrete Markov process (a jump process).

The motor behavior (such as the average velocity) can be calculated by following the stochastic evolution (mechanical motion and chemical reaction) of the motor in Monte Carlo simulations. However, results obtained with Monte Carlo simulations have statistical errors and converge very slowly. If we calculate the ensemble average

by following a large number of motors, then the statistical error is inversely proportional to the square root of the number of motors in the ensemble. Furthermore, when the potential $\psi(x)$ is not smooth, there are numerical difficulties in Monte Carlo simulations. Fortunately, average quantities can be calculated more efficiently by following the probability density of the motor.

Let us consider an ensemble of motors, each evolving in time independently and stochastically according to Langevin equation (1.7). Let $\rho_S(x,t)$ be the probability density that the motor is at position x and in chemical state S at time t. The time evolution of $\rho_S(x,t)$ is governed by the Fokker–Planck equation corresponding to Langevin equation (1.7) [12]:

$$\frac{\partial \rho_S}{\partial t} = D\frac{\partial}{\partial x}\left[\frac{V_S'(x)}{k_BT}\rho_S + \frac{\partial \rho_S}{\partial x}\right] + \sum_{j=1}^{N} k_{j\to S}(x)\rho_j, \qquad S = 1, 2, \ldots, N, \tag{1.8}$$

where, for $j \neq S$, $k_{j\to S}(x)$ is the chemical transition rate from state j to state S. $k_{S\to S}(x)$ is the total rate of jumping out of state S and is given by

$$k_{S\to S}(x) \equiv -\sum_{j\neq S} k_{S\to j}(x). \tag{1.9}$$

A simple way to model molecular motors is to average $V_S'(x)$ over all chemical states weighted by the steady state probability density functions of these states [10]. Let $\psi'(x)$ be the weighted average of $V_S'(x)$ over all chemical states:

$$\psi'(x) \equiv \frac{1}{\sum_{S=1}^{N}\rho_S(x)}\sum_{S=1}^{N}\rho_S(x)V_S'(x). \tag{1.10}$$

$\psi(x)$ can be viewed as the motor's mean field free energy landscape. The mechanical motion of the motor can be modeled using Langevin equation (1.6) with potential $\psi(x)$. Let L denote the period of $V_S(x)$. We immediately see that $\psi'(x)$ is also periodic with period L. However, $\psi(x)$ may not be periodic. As a matter of fact, for a molecular motor undergoing a unidirectional motion powered by a chemical reaction, $\psi(x)$ must not be periodic. If $\psi(x)$ is periodic, then there is no energy available to drive the motor forward because there is no free energy change going from one period to the next. For the motor to go forward, there must be a free energy drop going from one period to the next. Since $\psi'(x)$ is periodic, the energy landscape $\psi(x)$ is a tilted periodic potential:

$$\psi(x+L) = \psi(x) - \Delta\psi, \tag{1.11}$$

where $\Delta\psi > 0$ is the energy made available from the chemical reaction to drive the motor forward in one period. An example of tilted periodic potential is shown in Figure 1.1. This is also the potential we will use in numerical simulations in section 7. If the energy landscape $\psi(x)$ is simply a constant slope downhill, then the energy $\Delta\psi$ is utilized uniformly in one period to generate a constant motor force. If the slope of $\psi(x)$ is not a constant, then the motor force varies with the motor position within one period. As shown in Figure 1.1, the energy landscape $\psi(x)$ may not be monotonic. In that case, the motor depends on the Brownian fluctuations from the surrounding fluid to get over the free energy barrier. Of course, the energy source for driving the motor forward eventually comes from the free energy drop $\Delta\psi$, which rectifies

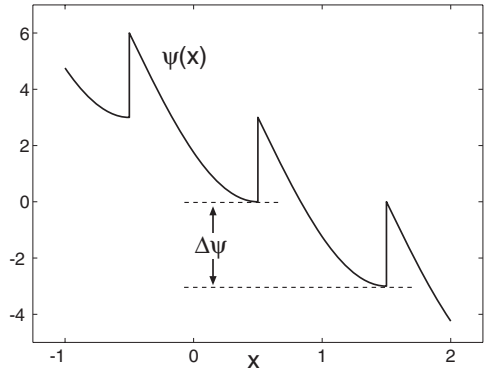


FIG. 1.1. *Graph of the tilted periodic potential used in numerical simulations in section 7.*

the forward fluctuations. This mechanism of driving the motor forward by rectifying thermal fluctuations is called ratchet [22].

If we use the mean field potential (1.10), then the mechanical motion of the motor is governed by the Langevin equation:

$$\frac{dx}{dt} = -\frac{1}{\zeta}\psi'(x) + \sqrt{2D}\frac{dW(t)}{dt}. \tag{1.12}$$

Equation (1.12) has been used in studies of motors [14, 15]. In addition to its mathematical simplicity, another advantage of using (1.12) is that the energy landscape $\psi(x)$ can be extracted from single molecule experimental data [10]. The Fokker–Planck equation corresponding to Langevin equation (1.12) is [12]

$$\frac{\partial \rho}{\partial t} = D\frac{\partial}{\partial x}\left[\frac{\psi'(x)}{k_B T}\rho + \frac{\partial \rho}{\partial x}\right]. \tag{1.13}$$

In [16], a robust numerical method (hereafter referred to as Method 1) was designed for solving Fokker–Planck equations (1.8) and (1.13). When the potential is smooth, the proof of convergence of Method 1 is straightforward [16]. But that does not provide an accurate theoretical explanation for the robust performance of Method 1. The strength of Method 1 is that it works fairly well even if the potential is discontinuous. In this paper, we are going to prove the convergence of Method 1 for the model equation (1.13) when the potential is piecewise smooth and has a finite number of discontinuities at half numerical grid points. First, we nondimensionalize (1.13). The dimensionless independent variables and functions are defined as

$$\tilde{x} = x\frac{1}{L}, \qquad \tilde{t} = t\frac{D}{L^2},$$

$$\tilde{\psi}(\tilde{x}) = \psi(x)\frac{1}{k_B T}, \qquad \tilde{\rho}(\tilde{x}) = \rho(x)L, \qquad \Delta\tilde{\psi} = \Delta\psi\frac{1}{k_B T}.$$

Since we are going to work with the dimensionless variables and functions, let us drop $\sim$ from the notations. The dimensionless version of (1.13) is

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x}\left[\psi'(x)\rho + \frac{\partial \rho}{\partial x}\right], \tag{1.14}$$

where $\psi(x)$ satisfies

$$\psi(x+1) = \psi(x) - \Delta\psi, \qquad \Delta\psi > 0.$$

Equation (1.14) can be viewed as a special case of (1.13) with $L = 1$, $D = 1$, and $k_B T = 1$.

The rest of the paper is organized as follows. In section 2, we discuss the two conditions that the exact solution of a discontinuous Fokker–Planck equation must satisfy at a discontinuity. We derive the two conditions for the exact solution at a discontinuity by rewriting the Fokker–Planck equation as a heat equation with discontinuous heat conductivity and discontinuous specific heat capacity. The first condition is the continuity of the "heat flux," which corresponds to the conservation of heat. The second condition is the continuity of the "temperature," which follows from the regularizing properties of the heat equation. In section 3, we describe the construction and properties of Method 1 developed in [16]. The most important property of Method 1 is that it preserves detailed balance. In section 4, we prove that Method 1 is stable with respect to a norm that is equivalent to the 2-norm. We start by showing that the steady state solution of the half discrete method is unique and all positive. This allows us to define a weighted 2-norm using the steady state solution as the weight function. We proceed to show that the steady state solution is bounded away from 0 and from infinity, independent of the numerical grid size. It follows that the weighted 2-norm is equivalent to the standard 2-norm. We then show that with respect to the weighted 2-norm, Method 1 is unconditionally stable. In section 5, we analyze the consistency of Method 1 when the potential is discontinuous. We show that away from the discontinuity, the local truncation error of Method 1 on the exact solution is $O(k(k^2+h^2))$. At the discontinuity, the local truncation error on the exact solution is $O(1)$. However, if we perturb the exact solution by a term of the order $O(h)$, then the local truncation error on the perturbed solution is $O(k(k^2+h))$. Once we have both the stability and consistency, the Lax equivalence theorem [17] implies that Method 1 converges to the correct solution of the differential equation. In section 6, we propose a modified version of Method 1 to eliminate the first order error term caused by the discontinuity (hereafter referred to as Method 2). As a result, the modified method (Method 2) is second order accurate even in the presence of discontinuities. In section 7, we carry out numerical simulations using a discontinuous potential to compare the performance of the central difference method, Method 1, and Method 2. The central difference method converges to a wrong solution. Both Method 1 and Method 2 converge to the correct solution. We also show that in the presence of discontinuities, detailed balance is a necessary condition for converging to the correct solution. This explains the defect of the central difference method.

2. Exact solution of a discontinuous Fokker–Planck equation and conditions at the discontinuity. In this section, we discuss the two conditions that the exact solution of a discontinuous Fokker–Planck equation must satisfy at a discontinuity. We derive the two conditions for the exact solution at a discontinuity by rewriting the Fokker–Planck equation as a heat equation with discontinuous heat conductivity and discontinuous specific heat capacity. The first condition is the continuity of the "heat flux." The second condition is the continuity of the "temperature."

We study the case where potential $\psi(x)$ is piecewise smooth and has a finite number of discontinuities in one period. Without loss of generality, we assume that there is only one discontinuity at x_d in $[0, 1]$. More specifically, we assume that $\psi(x)$ is two smooth functions connected by the discontinuity. That is, $\psi(x)$ is smooth in

$[0, x_d]$ if $\psi(x_d)$ is redefined as $\psi(x_d) = \lim_{x \to x_d^-} \psi(x)$, and $\psi(x)$ is smooth in $[x_d, 1]$ if $\psi(x_d)$ is redefined as $\psi(x_d) = \lim_{x \to x_d^+} \psi(x)$. For simplicity, in this paper a smooth function means it is infinitely differentiable, and so we can use as many terms of its Taylor expansion as we want.

When $\psi(x)$ is discontinuous at x_d, $\psi'(x)$ is not a regular function. If the system is brought to an equilibrium, the equilibrium solution is given by the Boltzmann distribution:

$$\rho(x) = \frac{1}{Z} e^{-\psi(x)}, \qquad Z = \int_0^1 e^{-\psi(x)} dx,$$

which is discontinuous at x_d. Thus, we should expect $\rho(x,t)$ to be discontinuous as a function of x at x_d. In modeling molecular motors, a discontinuous potential is simply a mathematical abstraction. In reality, the discontinuity represents a very narrow transition region in which the potential is smooth but changes dramatically. When the potential is smooth, we can rewrite Fokker–Planck equation (1.14) as

$$\frac{\partial \rho}{\partial t} = \frac{\partial}{\partial x} \left(e^{-\psi(x)} \frac{\partial e^{\psi(x)} \rho}{\partial x} \right). \tag{2.1}$$

Let us introduce $u(x,t) \equiv e^{\psi(x)} \rho(x,t)$. The equation above can be written in the form

$$\frac{\partial e^{-\psi(x)} u}{\partial t} = \frac{\partial}{\partial x} \left(e^{-\psi(x)} \frac{\partial u}{\partial x} \right). \tag{2.2}$$

Equation (2.2) has the form of a heat equation. In (2.2), $u(x,t)$ can be viewed as the "temperature," $e^{-\psi(x)}$ on the right-hand side as the heat conductivity, $e^{-\psi(x)}$ on the left-hand side as the specific heat capacity, and $e^{-\psi(x)} u(x,t) = \rho(x,t)$ as the heat. Equation (2.2) is equivalent to Fokker–Planck equation (1.14) when the potential is smooth. So it is natural for us to use the exact solution of (2.2) to define the exact solution of Fokker–Planck equation (1.14) when the potential is discontinuous. The biggest advantage of using (2.2) is that we can avoid the nonconservative product $\psi'(x)\rho(x,t)$ in Fokker–Planck equation (1.14). When both $\psi(x)$ and $\rho(x,t)$ are discontinuous, it is highly nontrivial to interpret the nonconservative product $\psi'(x)\rho(x,t)$ in Fokker–Planck equation (1.14) (for example, in [23], nonconservative product of the form $\frac{dw}{dx} g(w)$ is defined as a Borel measure).

In (2.2), when $\psi(x)$ is discontinuous, both the heat conductivity and the specific heat capacity are discontinuous. Away from the discontinuity, the exact solution of (2.2) satisfies differential equation (2.2) in the classical sense. At the discontinuity, the exact solution of (2.2) is constrained by two conditions. The first condition is the continuity of the "heat flux," which corresponds to the conservation of heat. The second condition is the continuity of the "temperature," which follows from the regularizing properties of the heat equation. The continuity of the "temperature" also reflects the physical nature of the heat conduction process: temperature gradient is relaxed by the heat flow that is induced by the temperature gradient. In particular, any isolated discontinuity in temperature will be removed immediately by heat conduction.

Now we write the two conditions in terms of $\rho(x,t)$ and $\psi(x)$. The first condition (continuity of "heat flux") is

$$\left(\psi'(x)\rho + \frac{\partial \rho}{\partial x} \right) \Big|_{x = x_d^-} = \left(\psi'(x)\rho + \frac{\partial \rho}{\partial x} \right) \Big|_{x = x_d^+}. \tag{2.3}$$

For Fokker–Planck equation (1.14), condition (2.3) means that the probability flux into the discontinuity is the same as the probability flux out of the discontinuity (i.e., the conservation of probability at the discontinuity), which corresponds to the well-known Rankine–Hugoniot condition for weak solutions of hyperbolic equations [24]. The second condition (continuity of "temperature") is

$$\left(e^{\psi(x)}\rho(x,t)\right)\Big|_{x=x_d^-} = \left(e^{\psi(x)}\rho(x,t)\right)\Big|_{x=x_d^+}. \tag{2.4}$$

Equations (2.3) and (2.4) are the two conditions that the exact solution of a discontinuous Fokker–Planck equation must satisfy at the discontinuity. If a numerical method is based on conservation of probability, then the numerical solution will automatically satisfy condition (2.3). As we will see in section 7, condition (2.4) is related to detailed balance. If a numerical method does not preserve detailed balance, then the numerical solution may converge to a wrong solution that does not satisfy condition (2.4).

3. The numerical method. In this section, we summarize Method 1 proposed in [16]. In the spatial discretization of (1.14), we divide the period $[0,1]$ into M subintervals of size $h = 1/M$. Each subinterval is represented by its center (a site), and the numerical grid is formed as

$$h = \frac{1}{M}, \qquad x_j = \frac{h}{2} + jh, \qquad x_{j-1/2} = \frac{h}{2} + \left(j - \frac{1}{2}\right) h.$$

Since the underlying stochastic evolution (1.12) is a continuous Markov process, we discretize it as a jump process (discrete Markov process). The idea of using a jump process on discrete sites to approximate a continuous Markov process was originated in [18] and in an unpublished result by C. Peskin. As shown in Figure 3.1, in the spatial discretization, subinterval j is $\left[x_{j-1/2}, x_{j+1/2}\right]$ and its center is x_j. The motor system can reside only on a set of discrete sites $\{x_j\}$. In a single jump, it can jump only to one of the two adjacent sites. Let $h \cdot p_j(t)$ be the probability that the motor system is at site x_j at time t in the jump process. $p_j(t)$ can be viewed as

$$p_j(t) \approx \frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} \rho(x,t)dx \approx \rho(x_j,t). \tag{3.1}$$

Let $F_{j+1/2}$ be the rate of jumping from x_j to x_{j+1} (forward jump) and $B_{j+1/2}$ the rate of jumping from x_{j+1} to x_j (backward jump). The numerical probability flux through $x_{j+1/2}$ is

$$J_{j+1/2} = h\left(F_{j+1/2}\, p_j - B_{j+1/2}\, p_{j+1}\right). \tag{3.2}$$

The time evolution of $p_j(t)$ is governed by the conservation of probability:

$$\begin{aligned} \frac{dp_j}{dt} &= \frac{1}{h}\left(J_{j-1/2} - J_{j+1/2}\right) \\ &= \left(F_{j-1/2}\, p_{j-1} - B_{j-1/2}\, p_j\right) - \left(F_{j+1/2}\, p_j - B_{j+1/2}\, p_{j+1}\right). \end{aligned} \tag{3.3}$$

Before we describe how the jump rates $F_{j+1/2}$ and $B_{j+1/2}$ are calculated in Method 1, we would like to point out that (3.3) is a very general framework. It can even accommodate the central difference method, which can be cast into the form of (3.3) with

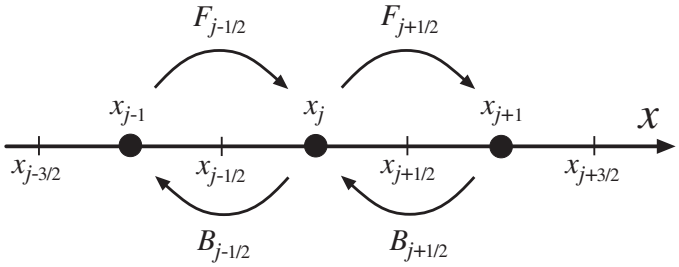


FIG. 3.1. *Spatial discretization of* (1.14). *The motor system is restricted to a set of discrete sites* $\{x_j\}$ *and can jump only to adjacent sites.*

jump rates

$$F_{j+1/2}^{(CD)} = \frac{1}{h^2}\left(1 - \frac{\delta\psi_{j+1/2}}{2}\right),$$
$$B_{j+1/2}^{(CD)} = \frac{1}{h^2}\left(1 + \frac{\delta\psi_{j+1/2}}{2}\right), \tag{3.4}$$

where

$$\delta\psi_{j+1/2} = \psi(x_{j+1}) - \psi(x_j). \tag{3.5}$$

Notice, however, that the jump rates (3.4) associated with the central difference method may be negative. Even in the limit of $h \to 0$, the jump rates (3.4) may be negative when the potential $\psi(x)$ is discontinuous. This explains why the central difference method converges to a wrong solution (as we will see in section 7).

In Method 1 [16], the jump rates are calculated based on local approximate solutions. In calculating $F_{j+1/2}$ and $B_{j+1/2}$, we make two assumptions:

1. In $[x_{j-1/2}, x_{j+3/2}]$, potential $\psi(x)$ is linear with slope $\delta\psi_{j+1/2}/h$. This assumption is to make the method simple and easy to implement. Under this assumption, the potential in $[x_{j-1/2}, x_{j+3/2}]$ is given by

$$\psi(x) = C + \frac{\delta\psi_{j+1/2}}{h} \cdot x, \tag{3.6}$$

   where $\delta\psi_{j+1/2}$ is defined in (3.5).

2. Let $\rho_{j+1/2}(x)$ be the steady state solution of (1.14) in $[x_{j-1/2}, x_{j+3/2}]$ with linear potential (3.6) and subject to the condition

$$\frac{1}{h}\int_{x_{j-1/2}}^{x_{j+1/2}} \rho_{j+1/2}(x)dx = p_j,$$
$$\frac{1}{h}\int_{x_{j+1/2}}^{x_{j+3/2}} \rho_{j+1/2}(x)dx = p_{j+1}. \tag{3.7}$$

   In the jump process, the probability flux through $x_{j+1/2}$ is given by that of $\rho_{j+1/2}(x)$. This assumption is a key component of Method 1 [16]. Instead of using the Taylor expansion, the numerical approximation is based on local approximate solutions. The consequence of this approach is that detailed balance is preserved, and Method 1 works well even if the potential is discontinuous.

The probability flux of $\rho_{j+1/2}(x)$ is derived in [16] and is given by

$$J = \frac{1}{h} \cdot \frac{\delta\psi_{j+1/2}}{e^{\delta\psi_{j+1/2}} - 1} \left(p_j - e^{\delta\psi_{j+1/2}}\, p_{j+1} \right). \tag{3.8}$$

Comparing the theoretical flux (3.8) with the numerical flux (3.2), we immediately obtain

$$\begin{aligned} F_{j+1/2} &= \frac{1}{h^2} \frac{\delta\psi_{j+1/2}}{e^{\delta\psi_{j+1/2}} - 1}, \\ B_{j+1/2} &= \frac{1}{h^2} \frac{\delta\psi_{j+1/2}\, e^{\delta\psi_{j+1/2}}}{e^{\delta\psi_{j+1/2}} - 1}, \end{aligned} \tag{3.9}$$

where $\delta\psi_{j+1/2} = \psi(x_{j+1}) - \psi(x_j)$, as defined in (3.5). It is important to notice that the jump rates (3.9) are always positive. It is straightforward to verify that $F_{j+1/2}$ and $B_{j+1/2}$ are both positive when $\delta\psi_{j+1/2} \neq 0$. When $\psi_{j+1/2} = 0$, we have

$$\begin{aligned} F_{j+1/2} &= \lim_{\delta\psi \to 0} \frac{1}{h^2} \frac{\delta\psi}{e^{\delta\psi} - 1} = \frac{1}{h^2}, \\ B_{j+1/2} &= \lim_{\delta\psi \to 0} \frac{1}{h^2} \frac{\delta\psi\, e^{\delta\psi}}{e^{\delta\psi} - 1} = \frac{1}{h^2}. \end{aligned}$$

The property that the jump rates given in (3.9) are always positive will be used in the stability analysis below. The jump rates given in (3.9) also satisfy detailed balance [16]:

$$\frac{F_{j+1/2}}{B_{j+1/2}} = e^{\psi(x_j) - \psi(x_{j+1})}. \tag{3.10}$$

In the time dimension, (3.3) is discretized using a Crank–Nicolson-type method [19]. Let p_j^n be the numerical approximation for $p_j(nk)$, where k is the time step. The fully discrete method is

$$\begin{aligned} p_j^{n+1} = p_j^n + k \Bigg\{ & \left(F_{j-1/2} \frac{p_{j-1}^n + p_{j-1}^{n+1}}{2} - B_{j-1/2} \frac{p_j^n + p_j^{n+1}}{2} \right) \\ & - \left(F_{j+1/2} \frac{p_j^n + p_j^{n+1}}{2} - B_{j+1/2} \frac{p_{j+1}^n + p_{j+1}^{n+1}}{2} \right) \Bigg\}. \end{aligned} \tag{3.11}$$

In the calculation of average velocity and/or effective diffusion coefficient, (1.14) is solved with the periodic boundary condition [16]. In the analysis of subsequent sections, we always assume the periodic boundary condition:

$$p_j^n = p_{j+M}^n, \qquad \psi(x+1) = \psi(x) - \Delta\psi.$$

We will prove the stability and consistency of (3.11) when potential $\psi(x)$ is piecewise smooth and has a finite number of discontinuities at half numerical grid points.

4. Stability of the numerical method. In this section, we prove the stability of Method 1 [16]. We first show that the steady state solution of the half discrete method (3.3) is unique and is all positive. Furthermore, we show that the maximum of the steady state solution is bounded by the minimum multiplied by a constant that

is determined by the underlying physical problem but is independent of the numerical grid size. Thus, we can define a weighted 2-norm using the steady state solution as the weighting function, and the weighted 2-norm so defined is equivalent to the 2-norm. Then we prove that the fully discrete method (3.11) is unconditionally stable with respect to the weighted 2-norm.

For the convenience of mathematical discussion, we introduce vector notations for numerical solutions in one period:

$$\begin{aligned}\vec{p}^{\,n} &\equiv (p_1^n,\ p_2^n,\ \ldots,\ p_M^n)\,,\\ \vec{q} &\equiv (q_1,\ q_2,\ \ldots,\ q_M)\,,\\ \vec{r} &\equiv (r_1,\ r_2,\ \ldots,\ r_M)\,.\end{aligned}$$

Here the superscript n denotes the time level. Remember all solutions are periodic. Let $\vec{q}$ be the steady state solution of (3.3). $\vec{q}$ satisfies the equation

$$(F_{j-1/2}\,q_{j-1} - B_{j-1/2}\,q_j) - (F_{j+1/2}\,q_j - B_{j+1/2}\,q_{j+1}) = 0, \qquad j = 1, 2, \ldots, M, \tag{4.1}$$

and satisfies the constraint

$$h\sum_{j=1}^{M} q_j = 1. \tag{4.2}$$

Condition (4.2) corresponds to $\int_0^1 \rho(x)dx = 1$. We now show that $\vec{q}$ is unique, is all positive, and is bounded away from 0 and from infinity, independent of the numerical grid size.

THEOREM 4.1. *Suppose $\vec{q}$ satisfies* (4.1). *Then $\vec{q}$ is either all zeros or all positive or all negative.*

Proof. Suppose $\vec{q}$ is not all zeros. Otherwise, there is no need to continue. Without loss of generality, we assume that there is an index j_0 such that $q_{j_0} > 0$. Otherwise, we simply consider $-\vec{q}$, which also satisfies (4.1). Starting at j_0, we first search to the left for a nonpositive element. If we cannot find a nonpositive element over a distance of M grid points, then all elements are positive because the solution is periodic.

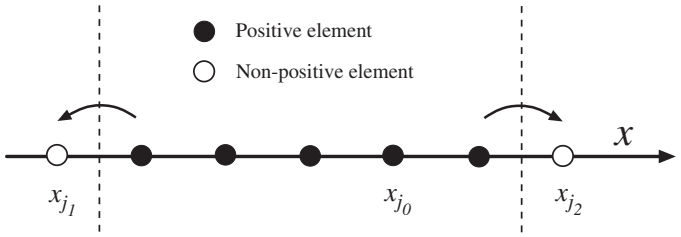


FIG. 4.1. *Starting from a positive element, $q_{j_0} > 0$, we either find that all elements are positive or find a positive region bounded by nonpositive elements. The latter leads to a contradiction.*

Suppose q_{j_1} is the first nonpositive element found in the search to the left and q_{j_2} is the first nonpositive element found in the search to the right. As shown in Figure 4.1, we have a positive region bounded by nonpositive elements:

$$q_{j_1} \le 0, \qquad q_{j_2} \le 0, \qquad \text{and } q_j > 0 \qquad \text{for } j_1 < j < j_2.$$

We are going to show that this leads to a contradiction. Summing (4.1) from $j = j_1+1$ to $j = j_2 - 1$, we obtain

$$\left(B_{j_1+1/2}\, q_{j_1+1} - F_{j_1+1/2}\, q_{j_1}\right) + \left(F_{j_2-1/2}\, q_{j_2-1} - B_{j_2-1/2}\, q_{j_2}\right) = 0. \tag{4.3}$$

The two terms on the left-hand side of (4.3) are the net probability fluxes to the outside of the region. Recall that the jump rates in Method 1 are always positive. Because $q_{j_1+1} > 0$ and $q_{j_1} \le 0$, the first term is positive. Similarly, the second term is also positive. Thus, the total net probability flux to the outside of the region is positive. This contradicts that $\vec{q}$ is a steady state. Mathematically, the contradiction arises in (4.3), where the two positive terms sum to zero. Therefore, if one element is positive, then all elements must be positive. ∎

Remark. The proof presented here can be extended to Fokker–Planck equations of higher dimensions, in which the positive region is bounded by a combination of nonpositive elements and a periodic boundary. The total net probability flux to the outside of the region is positive, which contradicts the steady state assumption.

THEOREM 4.2. *Suppose $\vec{q}$ satisfies* (4.1) *and* (4.2). *Then $\vec{q}$ is unique and all positive.*

Proof. Suppose both $\vec{q}$ and $\vec{p}$ satisfy (4.1) and (4.2). Let $\vec{r} = \vec{q} - \vec{p}$. Then $\vec{r}$ satisfies (4.1). Applying Theorem 4.1 to $\vec{r}$ yields that $\vec{r}$ is either all zeros or all positive or all negative. Since $\vec{r}$ also satisfies $h\sum_{j=1}^{M} r_j = 0$, $\vec{r}$ must be all zeros. Consequently, $\vec{q}$ is unique. Applying Theorem 4.1 to $\vec{q}$ and using (4.2), we obtain that $\vec{q}$ is all positive. ∎

Remark. This theorem shows that the solution of (4.1) with condition (4.2) is unique and all positive. In other words, Method 1 yields a unique steady state solution and preserves the positivity of probability. As pointed above, this result can be extended to Fokker–Planck equations of higher dimensions.

Now we consider the 2-norm and the weighted 2-norm defined as

$$\|\vec{p}\|_2 \equiv \left(h\sum_{j=1}^{M} (p_j)^2\right)^{\frac{1}{2}},$$

$$\|\vec{p}\|_\psi \equiv \left(h\sum_{j=1}^{M} \frac{1}{q_j}\,(p_j)^2\right)^{\frac{1}{2}}, \qquad h = \frac{1}{M}, \tag{4.4}$$

where $\vec{q} = (q_1,\ q_2,\ \ldots,\ q_M)$ is the solution of (4.1) with condition (4.2). In the analysis below, $\vec{q} = (q_1,\ q_2,\ \ldots,\ q_M)$ is reserved to denote this steady state solution. The 2-norm is denoted by $\|\bullet\|_2$. The weighted 2-norm is denoted by $\|\bullet\|_\psi$ because the weighting function $\vec{q}$ depends on potential $\psi(x)$.

THEOREM 4.3. *Suppose the free energy drop satisfies $\Delta\psi = \psi(x) - \psi(x+1) > 0$. Then the steady state probability flux of* (4.1), *$\tilde{J} = h\left(F_{j+1/2}\, q_j - B_{j+1/2}\, q_{j+1}\right)$, must be positive.*

Proof. We use proof by contradiction. Suppose $\tilde{J} \le 0$. Recall that Method 1 satisfies detailed balance (3.10). It follows that

$$q_{j+1} \ \ge\ \frac{F_{j+1/2}}{B_{j+1/2}}\, q_j = e^{\psi(x_j)-\psi(x_{j+1})}\, q_j.$$

Applying the inequality for $j,\ j+1,\ \ldots$, we obtain

$$q_{j+M} \ \ge\ e^{\psi(x_j)-\psi(x_{j+M})}\, q_j = e^{\Delta\psi}\, q_j > q_j,$$

which contradicts the periodic condition $q_{j+M} = q_j$. In the above, we have used the condition $\Delta\psi > 0$. Therefore, when $\Delta\psi$ is positive, $\tilde{J}$ must be positive. ∎

Remark. The key component in the proof of this theorem is detailed balance (3.10). A method preserving detailed balance has the advantage that the direction of chemical reaction and mechanical motion is preserved in numerical results. That is, the method will not produce a numerical result in which the motor system goes upward along the free energy landscape. This property is very important in studies of molecular motors.

THEOREM 4.4. *Suppose $\Delta\psi > 0$, and $\vec{q}$ satisfies* (4.1) *and* (4.2). *Then we have*

$$\max_j q_j \le e^{C_\psi}, \qquad \min_j q_j \ge e^{-C_\psi}, \tag{4.5}$$

where C_ψ depends on potential $\psi(x)$ but is independent of the numerical grid size.

Proof. Equation (4.1) implies that

$$\left(F_{j+1/2}\, q_j - B_{j+1/2}\, q_{j+1}\right) = \frac{\tilde{J}}{h}, \quad \text{independent of } j, \tag{4.6}$$

where $\tilde{J}$ is the steady state probability flux. Applying Theorem 4.3, we have $\tilde{J} > 0$, and (4.6) becomes

$$\left(F_{j+1/2}\, q_j - B_{j+1/2}\, q_{j+1}\right) \ge 0.$$

Recall that Method 1 satisfies detailed balance (3.10). It follows that

$$q_{j+1} \;\le\; \frac{F_{j+1/2}}{B_{j+1/2}}\, q_j = e^{\psi(x_j)-\psi(x_{j+1})}\, q_j. \tag{4.7}$$

Let $q_l = \min_j q_j$. Applying inequality (4.7) for $j = l$, $j = l+1$, ..., we get

$$q_j \;\le\; e^{\psi(x_l)-\psi(x_j)}\, q_l \quad \text{for } j \ge l. \tag{4.8}$$

Let $C_\psi = \max_x \max_{x\le y\le x+1} (\psi(x) - \psi(y))$. C_ψ is a constant independent of the numerical grid size. Taking the maximum of both sides of (4.8) over $l \le j \le l + M$, and noticing that $\vec{q}$ is periodic, we obtain

$$\max_j q_j \;\le\; e^{C_\psi}\, q_l = e^{C_\psi} \min_j q_j. \tag{4.9}$$

Since $\vec{q}$ also satisfies (4.2), we have

$$\min_j q_j \le h \sum_{j=1}^{M} q_j = 1, \qquad \max_j q_j \ge h \sum_{j=1}^{M} q_j = 1. \tag{4.10}$$

Equations (4.9) and (4.10) lead immediately to (4.5). ∎

Remark. This theorem shows that the weighted 2-norm is equivalent to the 2-norm

$$e^{-C_\psi} \|\vec{p}\|_2 \;\le\; \|\vec{p}\|_\psi \;\le\; e^{C_\psi} \|\vec{p}\|_2. \tag{4.11}$$

The extension of this theorem to Fokker–Planck equations of higher dimensions is still an open problem (we believe the theorem is valid for Fokker–Planck equations of higher dimensions, but the proof is still an open problem). We are going to use the weighted 2-norm to study the stability of the fully discrete method (3.11).

THEOREM 4.5. *Let $\vec{p}^{\,n} = (p_1^n, p_2^n, \ldots, p_M^n)$ denote the solution of the fully discrete method* (3.11) *at time level n. Then we have*

$$\|\vec{p}^{\,n+1}\|_\psi \le \|\vec{p}^{\,n}\|_\psi. \tag{4.12}$$

Proof. First, we write (3.11) as

$$p_j^{n+1} - p_j^n = \frac{k}{h}\left(J_{j-1/2}^{n+1/2} - J_{j+1/2}^{n+1/2}\right), \tag{4.13}$$

where

$$J_{j+1/2}^{n+1/2} = h\left(F_{j+1/2}\, p_j^{n+1/2} - B_{j+1/2}\, p_{j+1}^{n+1/2}\right), \tag{4.14}$$

$$p_j^{n+1/2} = \frac{p_j^{n+1} + p_j^n}{2}.$$

Multiplying both sides of (4.13) by $2\,h\,p_j^{n+1/2}$, dividing by q_j, and summing over j yields

$$h\sum_{j=1}^{M}\frac{1}{q_j}\,(p_j^{n+1})^2 - h\sum_{j=1}^{M}\frac{1}{q_j}\,(p_j^n)^2 = 2k\sum_{j=1}^{M}\frac{1}{q_j}p_j^{n+1/2}\left(J_{j-1/2}^{n+1/2} - J_{j+1/2}^{n+1/2}\right).$$

Applying summation by parts and using the periodic condition, we get

$$\|\vec{p}^{\,n+1}\|_\psi^2 - \|\vec{p}^{\,n}\|_\psi^2 = -2k\sum_{j=1}^{M}\left(\frac{1}{q_j}p_j^{n+1/2} - \frac{1}{q_{j+1}}p_{j+1}^{n+1/2}\right)J_{j+1/2}^{n+1/2}. \tag{4.15}$$

Let $r_j = \frac{p_j^{n+1/2}}{q_j}$. Here, for simplicity and without causing confusion, we have dropped the superscript $(n+1/2)$ from r_j. We write the probability flux $J_{j+1/2}^{n+1/2}$ as

$$\begin{aligned} J_{j+1/2}^{n+1/2} &= h\left(F_{j+1/2}\, p_j^{n+1/2} - B_{j+1/2}\, p_{j+1}^{n+1/2}\right) \\ &= h\left(F_{j+1/2}\, q_j\, r_j - B_{j+1/2}\, q_{j+1}\, r_{j+1}\right) \\ &= h\left(\frac{F_{j+1/2}\, q_j + B_{j+1/2}\, q_{j+1}}{2}\right)(r_j - r_{j+1}) + \tilde{J}\cdot\frac{r_j + r_{j+1}}{2}. \end{aligned}$$

Here $\tilde{J} = h\left(F_{j+1/2}\, q_j - B_{j+1/2}\, q_{j+1}\right)$ is the steady state probability flux, which is a constant independent of j. Substituting $J_{j+1/2}^{n+1/2}$ into (4.15), we have

$$\begin{aligned} \|\vec{p}^{\,n+1}\|_\psi^2 - \|\vec{p}^{\,n}\|_\psi^2 = &-kh\sum_{j=1}^{M}\left(F_{j+1/2}\, q_j + B_{j+1/2}\, q_{j+1}\right)(r_j - r_{j+1})^2 \\ &- k\sum_{j=1}^{M}\tilde{J}\left((r_j)^2 - (r_{j+1})^2\right). \end{aligned} \tag{4.16}$$

Recall that in Method 1, the steady state solution $\vec{q}$ is all positive and jump rates are all positive. Consequently, the first term on the right-hand side of (4.16) is nonpositive.

Applying summation by parts and using the periodic condition, we obtain that the second term on the right-hand side of (4.16) is zero. Thus, we conclude that

$$\|\vec{p}^{\,n+1}\|_\psi^2 - \|\vec{p}^{\,n}\|_\psi^2 \le 0, \tag{4.17}$$

which immediately leads to (4.12). □

Remark 1. This theorem shows that the fully discrete method (3.11) is unconditionally stable with respect to the weighted 2-norm $\| \bullet \|_\psi$ even if potential $\psi(x)$ is discontinuous.

Remark 2. This theorem can be extended to Fokker–Planck equations of higher dimensions.

5. Consistency and convergence of the numerical method. We study the consistency of Method 1 when potential $\psi(x)$ is piecewise smooth and has a finite number of discontinuities in one period at half numerical grid points. Without loss of generality, we assume that there is only one discontinuity at x_d in $[0,1]$, where $x_d = x_{l+1/2}$ is a half numerical grid point (that is, x_d is at the boundary between two numerical subintervals). Below we will show that away from the discontinuity, the local truncation error of Method 1 on the exact solution is $O\left(k(k^2+h^2)\right)$. At the discontinuity, the local truncation error on the exact solution is $O(1)$. However, if we perturb the exact solution by a term of the order $O(h)$, then the local truncation error on the perturbed solution (instead of the exact solution) is $O\left(k(k^2+h)\right)$.

5.1. Local truncation error away from the discontinuity. We rewrite the fully discrete method (3.11) in the flux form

$$p_j^{n+1} = p_j^n + \frac{k}{h}\left(\frac{J_{j-1/2}^n + J_{j-1/2}^{n+1}}{2} - \frac{J_{j+1/2}^n + J_{j+1/2}^{n+1}}{2}\right), \tag{5.1}$$

where the numerical probability flux is

$$J_{j+1/2}^n = h\left(F_{j+1/2}\, p_j^n - B_{j+1/2}\, p_{j+1}^n\right). \tag{5.2}$$

Let $\rho(x,t)$ be the exact solution of (1.14) subject to conditions (2.3) and (2.4). Let ρ_j^n denote the exact solution on the numerical grid:

$$\rho_j^n = \rho(x_j, t_n).$$

The local truncation error is defined as the residual term when the numerical method is applied to the exact solution ρ_j^n. Integrating the differential equation (1.14) over $[x_{j-1/2}, x_{j+1/2}] \times [t_n, t_{n+1}]$, we have

$$\begin{aligned}\int_{x_{j-1/2}}^{x_{j+1/2}} \rho(x,t_{n+1})dx - \int_{x_{j-1/2}}^{x_{j+1/2}} \rho(x,t_n)dx \\ = \int_{t_n}^{t_{n+1}} J(x_{j-1/2},t)dt - \int_{t_n}^{t_{n+1}} J(x_{j+1/2},t)dt,\end{aligned} \tag{5.3}$$

where $J(x,t) = -\left(\psi'\rho + \frac{\partial \rho}{\partial x}\right)$ is the exact probability flux in (1.14). Using the trapezoidal rule to approximate the integrals on the right-hand side yields

$$\begin{aligned}\int_{t_n}^{t_{n+1}} J(x_{j+1/2},t)dt = k\left(\frac{J(x_{j+1/2},t_n) + J(x_{j+1/2},t_{n+1})}{2}\right) \\ - \frac{k^3}{12}\frac{\partial^2}{\partial t^2} J(x_{j+1/2}, t_{n+1/2}) + O(k^5).\end{aligned}$$

Since the probability flux, $J(x,t)$, is continuous across the discontinuity, the time derivatives of $J(x,t)$ are also continuous across the discontinuity. Suppose $x_{j+1/2}$ is the location of discontinuity. Then we have

$$\frac{\partial^2}{\partial t^2} J(x^+_{j+1/2}, t_{n+1/2}) = \frac{\partial^2}{\partial t^2} J(x^-_{j+1/2}, t_{n+1/2}).$$

Using the assumption that everything is smooth on both sides of the discontinuity, we obtain

$$\begin{aligned}\frac{\partial^2}{\partial t^2} J(x^-_{j+1/2}, t_{n+1/2}) - \frac{\partial^2}{\partial t^2} J(x_{j-1/2}, t_{n+1/2}) \\ = h \frac{\partial^3}{\partial t^2 \partial x} J(\xi, t_{n+1/2}) = O(h),\end{aligned}$$

where $x_{j-1/2} < \xi < x_{j+1/2}$. Expanding the integrals on the left-hand side of (5.3), using the results we just derived for the integrals on the right-hand side of (5.3), and then dividing (5.3) by h, we arrive at

$$\begin{aligned}\rho_j^{n+1} - \rho_j^n = \frac{k}{h} \Bigg(& \frac{J(x_{j-1/2}, t_n) + J(x_{j-1/2}, t_{n+1})}{2} \\ & - \frac{J(x_{j+1/2}, t_n) + J(x_{j+1/2}, t_{n+1})}{2} \Bigg) + O\left(k(k^2 + h^2)\right).\end{aligned} \tag{5.4}$$

Let $J^n_{j+1/2}\{\rho\}$ denote the numerical probability flux on the exact solution $\rho(x,t)$:

$$J^n_{j+1/2}\{\rho\} = h\left(F_{j+1/2}\, \rho_j^n - B_{j+1/2}\, \rho_{j+1}^n\right). \tag{5.5}$$

We expand $F_{j+1/2}$, $B_{j+1/2}$, ρ_j^n, and ρ_j^{n+1} around $x = x_{j+1/2}$ to obtain the following expansion for $J^n_{j+1/2}\{\rho\}$ away from the discontinuity:

$$J^n_{j+1/2}\{\rho\} = J(x_{j+1/2}, t_n) - h^2\, u(x_{j+1/2}, t_n) + O\left(h^3\right), \tag{5.6}$$

where $u(x,t)$ is a function consisting of various derivatives of $\psi(x)$ and $\rho(x,t)$. The derivation of (5.6) is presented in Appendix A. Notice that $u(x,t)$ is smooth in the region where $\psi(x)$ and $\rho(x,t)$ are smooth. Away from the discontinuity, $u(x,t)$ satisfies

$$u(x_{j+1/2}, t_n) - u(x_{j-1/2}, t_n) = O(h).$$

Substituting (5.6) into (5.4), we obtain that, away from the discontinuity, ρ_j^n satisfies

$$\begin{aligned}\rho_j^{n+1} = \rho_j^n + \frac{k}{h} \left(\frac{J^n_{j-1/2}\{\rho\} + J^{n+1}_{j-1/2}\{\rho\}}{2} - \frac{J^n_{j+1/2}\{\rho\} + J^{n+1}_{j+1/2}\{\rho\}}{2} \right) \\ + O\left(k(k^2 + h^2)\right).\end{aligned} \tag{5.7}$$

That is, away from the discontinuity, the local truncation error on the exact solution $\rho(x,t)$ is $O\left(k(k^2 + h^2)\right)$.

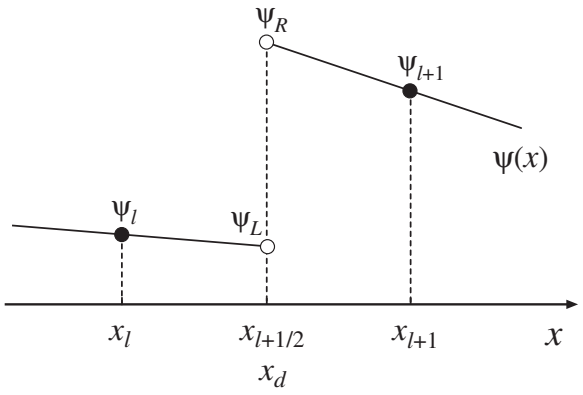


FIG. 5.1. *Schematic diagram of the discontinuity at* $x_d = x_{l+1/2}$.

5.2. Local truncation error on the perturbed solution. Now let us look at the numerical probability flux at the discontinuity. As shown in Figure 5.1, the discontinuity is at $x_d = x_{l+1/2}$. We introduce several shorthand notations:

$$\rho_L^n = \rho(x,t_n)|_{x=x_d^-}\,, \qquad \rho_R^n = \rho(x,t_n)|_{x=x_d^+}\,,$$

$$\psi_L = \psi(x)|_{x=x_d^-}\,, \qquad \psi_R = \psi(x)|_{x=x_d^+}\,.$$

The numerical probability flux on the exact solution $\rho(x,t)$ is

$$\begin{aligned} J_{l+1/2}^n\{\rho\} &= h\left(F_{l+1/2}\rho_l^n - B_{l+1/2}\rho_{l+1}^n\right) \\ &= \frac{1}{h}\,\frac{\psi_{l+1}-\psi_l}{e^{\psi_{l+1}}-e^{\psi_l}}\left(e^{\psi_l}\rho_l^n - e^{\psi_{l+1}}\rho_{l+1}^n\right) \\ &= \frac{1}{h}\,\frac{\psi_{l+1}-\psi_l}{e^{\psi_{l+1}}-e^{\psi_l}}\left(e^{\psi_l}\rho_l^n - e^{\psi_L}\rho_L^n + e^{\psi_R}\rho_R^n - e^{\psi_{l+1}}\rho_{l+1}^n\right). \end{aligned} \tag{5.8}$$

Here we have used condition (2.4): $e^{\psi_L}\rho_L^n = e^{\psi_R}\rho_R^n$. Expanding $e^{\psi_l}\rho_l^n$ around $x = x_d^-$ and $e^{\psi_{l+1}}\rho_{l+1}^n$ around $x = x_d^+$, we get

$$\begin{aligned} &e^{\psi_l}\rho_l^n - e^{\psi_L}\rho_L^n \\ &\quad= -\frac{h}{2}e^{\psi_L}\left(\psi'(x)\rho + \frac{\partial\rho}{\partial x}\right)\bigg|_{x=x_d^-} + h^2\,v_L(t_n) + h^3\,w_L(t_n) + O(h^4) \\ &\quad= \frac{h}{2}e^{\psi_L}J(x_{l+1/2},t_n) + h^2\,v_L(t_n) + h^3\,w_L(t_n) + O(h^4), \end{aligned}$$

$$\begin{aligned} &e^{\psi_R}\rho_R^n - e^{\psi_{l+1}}\rho_{l+1}^n \\ &\quad= \frac{h}{2}e^{\psi_R}J(x_{l+1/2},t_n) + h^2\,v_R(t_n) + h^3\,w_R(t_n) + O(h^4), \end{aligned}$$

where $v_L(t)$, $v_R(t)$, $w_L(t)$, and $w_R(t)$ are smooth functions of t. Substituting these two expansions into (5.8), we have

$$J_{l+1/2}^n\{\rho\} = J(x_{l+1/2},t_n) + v_0(t_n) + h\,v_1(t_n) + h^2\,v_2(t_n) + O(h^3), \tag{5.9}$$

where

$$v_0(t) = \left[\frac{\psi_R - \psi_L}{e^{\psi_R} - e^{\psi_L}}\left(\frac{e^{\psi_L} + e^{\psi_R}}{2}\right) - 1\right] J(x_{l+1/2}, t_n). \tag{5.10}$$

Again, $v_0(t)$, $v_1(t)$, and $v_2(t)$ are smooth functions of t. Combining (5.9), which gives numerical flux at the discontinuity, and (5.6), which is valid away from the discontinuity, we obtain

$$\begin{aligned} J^n_{j-1/2}\{\rho\} - J^n_{j+1/2}\{\rho\} &= \big(J(x_{j-1/2}, t_n) - J(x_{j+1/2}, t_n)\big) + O(h^3) \\ &+ h^2\left(u(x_d^+, t_n) - u(x_d^-, t_n)\right) a_j + h\left(v_0(t_n) + h\, v_1(t_n) + h^2 \tilde{v}_2(t_n)\right) b_j, \end{aligned} \tag{5.11}$$

where

$$a_j = \begin{cases} \frac{1}{2}, & j = l, \\ \frac{1}{2}, & j = l+1, \\ -h\frac{M}{M-2} & \text{otherwise}, \end{cases} \tag{5.12}$$

$$b_j = \begin{cases} -\frac{1}{h}, & j = l, \\ +\frac{1}{h}, & j = l+1, \\ 0 & \text{otherwise}, \end{cases} \tag{5.13}$$

and

$$\tilde{v}_2(t) = v_2(t) + \frac{u(x_d^+, t) + u(x_d^-, t)}{2}. \tag{5.14}$$

In (5.11), the first term on the right-hand side is the desired result; the second term is of the order $O(h^3)$; the third term is of the order $O(h^2)$ at the discontinuity and of the order $O(h^3)$ elsewhere; the fourth term is of the order $O(1)$, which is the term we need to deal with.

Terms in (5.11) contribute to the local truncation error. In general, the global error is one order lower than the local truncation error. However, the $O(1)$ term in (5.11) does not necessarily imply that the error in the numerical solution is $O(1/h)$ or is $O(1)$. As we will see below, the $O(1)$ term in (5.11) actually leads to an $O(h)$ term in the global error. The key is that the $O(1)$ term in (5.11) can be eliminated by perturbing the exact solution by an $O(h)$ term. So if we use the perturbed exact solution to calculate the local truncation error, then the $O(1)$ term in (5.11) disappears. In other words, the numerical method is consistent with the perturbed exact solution and the perturbed exact solution converges to the exact solution. The successful elimination of the $O(1)$ term in (5.11) by perturbing the exact solution by an $O(h)$ term depends on the special structure of vector $\{b_j\}$ given in (5.13). For that purpose, we have the theorem below.

THEOREM 5.1. *Suppose $\vec{r}$ is periodic, satisfies the equation*

$$h\left(F_{j-1/2}\, r_{j-1} - B_{j-1/2}\, r_j\right) - h\left(F_{j+1/2}\, r_j - B_{j+1/2}\, r_{j+1}\right) = b_j, \tag{5.15}$$

and satisfies the condition

$$\sum_{j=1}^{M} r_j = 0, \tag{5.16}$$

where b_j is defined in (5.13). *Then there exists a constant C_b, independent of the numerical grid size, such that*

$$\max_j |r_j| \le C_b.$$

Proof. The proof of Theorem 5.1 is presented in Appendix B. ∎

Now we use the result of Theorem 5.1 to eliminate the $O(1)$ error term in (5.11), and we consider the perturbed solution given below:

$$\tilde{\rho}_j^n \equiv \rho_j^n - h\left(v_0(t_n) + h\, v_1(t_n) + h^2\, \tilde{v}_2(t_n)\right) r_j = \rho_j^n + O(h), \tag{5.17}$$

where r_j is the solution of (5.15) and (5.16) in Theorem 5.1. It is straightforward to verify that the perturbed solution $\tilde{\rho}_j^n$ satisfies

$$J_{j-1/2}^n\{\tilde{\rho}\} - J_{j+1/2}^n\{\tilde{\rho}\} = \left(J(x_{j-1/2}, t_n) - J(x_{j+1/2}, t_n)\right) + O(h^2), \tag{5.18}$$

$$\tilde{\rho}_j^{n+1} - \tilde{\rho}_j^n = \rho_j^{n+1} - \rho_j^n + O(kh). \tag{5.19}$$

Substituting (5.18) and (5.19) into (5.4), we obtain

$$\begin{aligned}\tilde{\rho}_j^{n+1} = \tilde{\rho}_j^n + \frac{k}{h}\left(\frac{J_{j-1/2}^n\{\tilde{\rho}\} + J_{j-1/2}^{n+1}\{\tilde{\rho}\}}{2} - \frac{J_{j+1/2}^n\{\tilde{\rho}\} + J_{j+1/2}^{n+1}\{\tilde{\rho}\}}{2}\right)\\ + O\left(k(k^2+h)\right).\end{aligned} \tag{5.20}$$

That is, the local truncation error on the perturbed solution $\tilde{\rho}_j^n$ is $O\left(k(k^2+h)\right)$.

5.3. Convergence of the numerical method. Once we have both the stability and the consistency, the convergence follows, in principle, from the Lax equivalence theorem [17]. More specifically, we write the numerical method (3.11) in the vector-operator form

$$\vec{p}^{\,n+1} = L\,\vec{p}^{\,n}, \tag{5.21}$$

where L is the linear operator representing the numerical method. Stability (4.12) implies

$$\|L\|_\psi \le 1, \tag{5.22}$$

where $\|L\|_\psi$ is the induced operator norm defined by

$$\|L\|_\psi \equiv \max_{\|\vec{p}\|_\psi = 1} \|L\vec{p}\|_\psi.$$

Consistency (5.20) on the perturbed solution implies

$$\vec{\tilde{\rho}}^{\,n+1} = L\,\vec{\tilde{\rho}}^{\,n} + O\left(k(k^2+h)\right), \tag{5.23}$$

where $\vec{\tilde{\rho}}^{\,n}$ is the perturbed solution given in (5.17). Subtracting (5.23) from (5.21) yields

$$\vec{p}^{\,n+1} - \vec{\tilde{\rho}}^{\,n+1} = L\left[\vec{p}^{\,n} - \vec{\tilde{\rho}}^{\,n}\right] + O\left(k(k^2+h)\right). \tag{5.24}$$

Taking $\| \bullet \|_\psi$ norm of the both sides, using the stability, and summing over n, we have

$$\|\vec{p}^{\,n} - \vec{\tilde{\rho}}^{\,n}\|_\psi \le T \cdot O(k^2 + h), \qquad nk \le T. \tag{5.25}$$

Using the fact that $\vec{\tilde{\rho}}^{\,n} - \vec{\rho}^{\,n} = O(h)$, we obtain

$$\|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_\psi \le T \cdot O(k^2 + h). \tag{5.26}$$

Using Theorem 4.4, we see that (5.26) implies the convergence in the 2-norm:

$$\|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_2 \le e^{C_\psi} \|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_\psi \le e^{C_\psi}\, T \cdot O(k^2 + h). \tag{5.27}$$

If $k \le O(h)$, then (5.26) also implies the pointwise convergence

$$\|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_\infty \le \frac{1}{\sqrt{h}} \|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_2 \le e^{C_\psi}\, T \cdot O(\sqrt{h}). \tag{5.28}$$

6. The modified numerical method. In the consistency analysis of the previous section, we see the connection between the leading term in the local truncation error and the jump rates. Based on the lessons we learned in the consistency analysis, we will design a new set of jump rates to eliminate the first order error term caused by the discontinuity. The modified numerical method (hereafter referred to as Method 2) is as simple as Method 1. We will show that Method 2 is second order accurate even if the potential is discontinuous.

In the local truncation error in (5.20), the leading term $O(kh)$ comes from the term $v_0(t_n)$ in (5.9). At the discontinuity, if we use Method 1 defined in (3.9), then we have

$$v_0(t_n) = \left[\frac{\psi_R - \psi_L}{e^{\psi_R} - e^{\psi_L}} \left(\frac{e^{\psi_L} + e^{\psi_R}}{2} \right) - 1 \right] J(x_{l+1/2}, t_n) \neq 0.$$

This suggests a way of improving the performance of Method 1. We need to design the jump rates such that $v_0(t_n) = 0$ at the discontinuity. For that purpose, we propose Method 2:

$$\begin{aligned} F_{j+1/2} &= \frac{1}{h^2} \frac{2e^{\psi_j}}{e^{\psi_j} + e^{\psi_{j+1}}}, \\ B_{j+1/2} &= \frac{1}{h^2} \frac{2e^{\psi_{j+1}}}{e^{\psi_j} + e^{\psi_{j+1}}}. \end{aligned} \tag{6.1}$$

If we use Method 2 defined in (6.1), then we have

$$v_0(t_n) = \left[\frac{2}{e^{\psi_L} + e^{\psi_R}} \left(\frac{e^{\psi_L} + e^{\psi_R}}{2} \right) - 1 \right] J(x_{l+1/2}, t_n) = 0.$$

It can be shown that expansion (5.6) is still valid for Method 2. Consequently, expansion (5.11) is valid for Method 2 where $v_0(t_n) = 0$. Recall that for Method 1, the $O(1)$ term in (5.11) leads to an $O(h)$ error term in the numerical solution. For Method 2, $v_0(t_n) = 0$ kills the $O(1)$ term in (5.11). The third term in (5.11) is $O(h^2)$, and the remaining part of the fourth term in (5.11) is $O(h)$. Because of the special structures of vector $\{b_j\}$ in (5.13) and vector $\{a_j\}$ in (5.12), both the third term and the remaining part of the fourth term in (5.11) can be eliminated by perturbing the

exact solution by an $O(h^2)$ term. The elimination of these two terms by a perturbation of $O(h^2)$ makes Method 2 a second order method. For eliminating the third term in (5.11), we have the theorem below.

THEOREM 6.1. *Suppose $\vec{\sigma}$ is periodic, satisfies the equation*

$$h\left(F_{j-1/2}\,\sigma_{j-1} - B_{j-1/2}\,\sigma_j\right) - h\left(F_{j+1/2}\,\sigma_j - B_{j+1/2}\,\sigma_{j+1}\right) = a_j, \tag{6.2}$$

and satisfies the condition

$$\sum_{j=1}^{M} \sigma_j = 0, \tag{6.3}$$

where a_j is defined in (5.12). *Then there exists a constant C_a, independent of the numerical grid size, such that*

$$\max_j |\sigma_j| \le C_a.$$

Proof. The proof of Theorem 6.1 is similar to that of Theorem 5.1 and is skipped. ∎

Now we use the results of Theorems 5.1 and 6.1 to eliminate the third and fourth terms in (5.11). For Method 2, we consider the perturbed solution:

$$\begin{aligned} \hat{\rho}_j^n &\equiv \rho_j^n - h^2\left(v_1(t_n) + h\,\tilde{v}_2(t_n)\right) r_j - h^2\left(u(x_d^+, t_n) - u(x_d^-, t_n)\right)\sigma_j \\ &= \rho_j^n + O(h^2), \end{aligned} \tag{6.4}$$

where r_j is the solution of (5.15) and (5.16), and σ_j is the solution of (6.2) and (6.3). Theorems 5.1 and 6.1 guarantee that both r_j and σ_j are bounded by a constant, independent of the numerical grid size. The perturbed solution $\hat{\rho}_j^n$ satisfies

$$J_{j-1/2}^n\{\hat{\rho}\} - J_{j+1/2}^n\{\hat{\rho}\} = \left(J(x_{j-1/2}, t_n) - J(x_{j+1/2}, t_n)\right) + O(h^3), \tag{6.5}$$

$$\hat{\rho}_j^{n+1} - \hat{\rho}_j^n = \rho_j^{n+1} - \rho_j^n + O(kh^2). \tag{6.6}$$

Thus, the local truncation error on the perturbed solution $\hat{\rho}_j^n$ is $O\left(k(k^2+h^2)\right)$. Repeating the derivation from (5.21) to (5.28), we obtain

$$\|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_2 \le e^{C_\psi}\, T \cdot O(k^2 + h^2), \tag{6.7}$$

$$\|\vec{p}^{\,n} - \vec{\rho}^{\,n}\|_\infty \le e^{C_\psi}\, T \cdot O(h^{\frac{3}{2}}). \tag{6.8}$$

The error bound for the ∞-norm (6.8) is derived assuming the worst case scenario. As we will see in the numerical example below, the ∞-norm of the error is usually of the same order as the 2-norm of the error.

Method 2 defined in (6.1) can be viewed as constructed by using the standard finite difference on (2.1) with a special approximation for the heat conductivity:

$$e^{-\psi(x_{j+1/2})} \approx \frac{2}{e^{\psi(x_j)} + e^{\psi(x_{j+1})}}.$$

This special approximation is essential for Method 2 to achieve second order accuracy at discontinuities. Method 1 developed in [16] can be viewed as constructed by using the standard finite difference on (2.1) with approximation

$$e^{-\psi(x_{j+1/2})} \approx \frac{\psi(x_{j+1}) - \psi(x_j)}{e^{\psi(x_{j+1})} - e^{\psi(x_j)}}.$$

The numerical method previously developed by Elston and Doering in [18] can be viewed as constructed by using the standard finite difference on (2.1) with approximation

$$e^{-\psi(x_{j+1/2})} \approx e^{-\left(\frac{\psi(x_j)+\psi(x_{j+1})}{2}\right)}.$$

7. Numerical results and discussions. Now we compare the performance of three numerical methods on a model problem with discontinuous potential. For the model problem, we select the potential

$$\psi(x) = \begin{cases} 6 - 6\sin\left(\frac{\pi}{2}(x+0.5)\right), & 0 < x < 0.5, \\ 3 - 6\sin\left(\frac{\pi}{2}(x-0.5)\right), & 0.5 < x < 1. \end{cases} \tag{7.1}$$

The graph of this discontinuous potential is shown in Figure 1.1. It has a discontinuity of amplitude 3 at $x = 0.5$. We use the initial condition

$$\rho(x, 0) = 1 + \cos(2\pi x). \tag{7.2}$$

We run simulations to $t = 1$ with a wide range of values for spatial step h, and we use time step $k = h$. We compare the performance of the central difference method (3.4), Method 1 (3.9), and Method 2 (6.1). We define the error as the difference between the numerical solution obtained with a finite value of h and the converged target (the converged target is not necessarily the correct solution of the differential equation). The behavior of the error so defined tells us whether or not a method converges. However, it does not tell us whether or not the converged target is the correct solution. We estimate the error as follows. Suppose $\vec{p}^{\,n}(h)$ is the numerical solution obtained with spatial step h and time step $k = h$. The error of $\vec{p}^{\,n}(h)$ is estimated as

$$\text{error}(h) \approx C_p \left\| \vec{p}^{\,n}(h) - \vec{p}^{\,n}\left(\frac{h}{2}\right) \right\|, \tag{7.3}$$

where C_p is a constant depending on the order of the method. For methods of first order or higher, C_p is between 1 and 2. Here we simply use $C_p = 1$ (in the worst case scenario, we underestimate the error by a factor of 2). The order of a method is estimated as

$$\text{order}(h) \approx \log_2 \left(\frac{\text{error}(h)}{\text{error}\left(\frac{h}{2}\right)} \right). \tag{7.4}$$

Figure 7.1 shows the estimated errors and estimated orders for the three methods. We follow the behavior of both the 2-norm and the ∞-norm of the estimated error. Here we are solving a nondimensionalized Fokker–Planck equation. Both the time step and the spatial step are dimensionless. So $k = h$ is just a convenient choice. Strictly speaking, the error shown in Figure 7.1 is the total error in time and space estimated by comparing the numerical solution obtained using (h, k) to that of $(\frac{h}{2}, \frac{k}{2})$. However, we find that the difference in numerical solution between (h, k) and $(h, \frac{k}{2})$ is much smaller than that between (h, k) and $(\frac{h}{2}, \frac{k}{2})$ (results not shown), which indicates that the error shown in Figure 7.1 is mainly due to the spatial discretization.

For Method 1 (the second row in Figure 7.1), the estimated error decreases as $k = h$ is reduced. From the convergence analysis in the previous sections, we know that

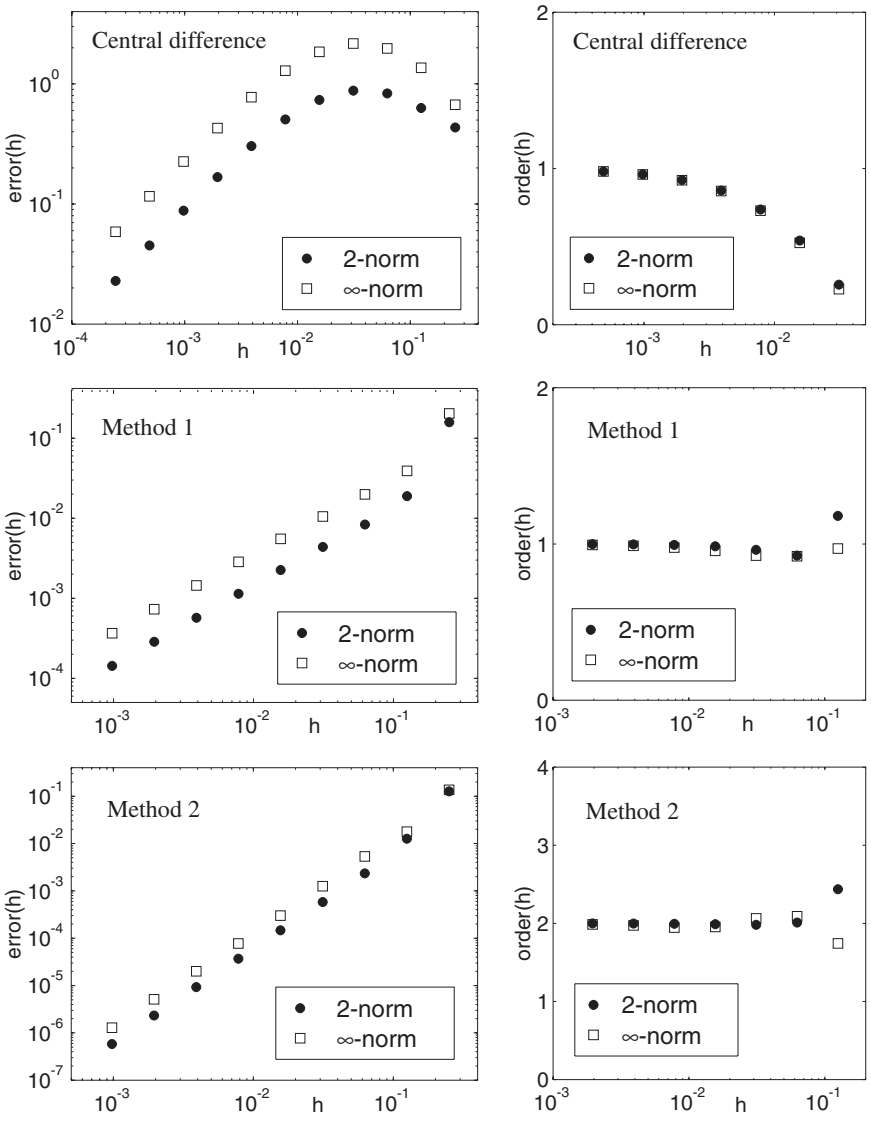


FIG. 7.1. *Estimated errors (first column) and estimated orders (second column) for the central difference method (first row), Method* 1 *(second row), and Method* 2 *(third row).*

the converged target must be the correct solution of the Fokker–Planck equation. In the presence of a discontinuity, the estimated order of accuracy of Method 1 developed in [16] is 1 for both the 2-norm and the ∞-norm. This result is consistent with the error bounds (5.27) and (5.28) derived in the previous sections. Notice that the 2-norm and the ∞-norm of the estimated error are of the same order. This tells us that although the first order error is caused by the discontinuity, it is spread to the whole region by diffusion.

For Method 2 (the third row in Figure 7.1), the estimated error decreases more rapidly as $k = h$ is reduced. The convergence analysis in the previous sections guarantees that Method 2 converges to the correct solution of the Fokker–Planck equation.

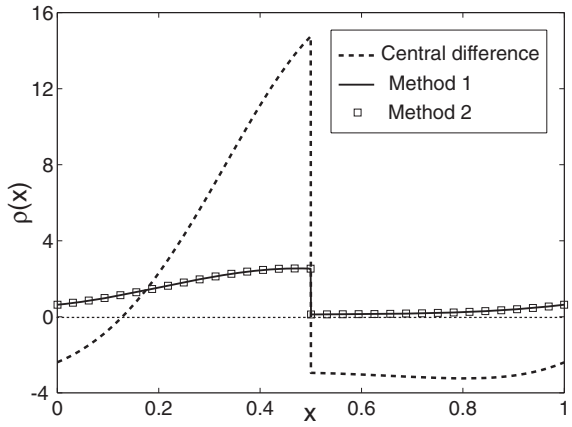


FIG. 7.2. *Numerical probability densities at $t = 1$ obtained with spatial and time steps $h = k = \frac{1}{4096}$ using, respectively, the central difference method, Method* 1*, and Method* 2*.*

Even in the presence of a discontinuity, the estimated order of accuracy for Method 2 is 2 for both the 2-norm and the ∞-norm. This result is consistent with the error bounds (6.7) and (6.8) derived in the previous section.

For the central difference method (the first row in Figure 7.1), it appears that the estimated error converges to zero as $k = h$ goes to zero. The convergence to the target is very slow. Even with $k = h = \frac{1}{4096}$, both the 2-norm and the ∞-norm of the error are still above 10^{-2}. But this is not the only defect of the central difference method. The fatal defect of the central difference method is that it converges to a wrong solution that does not satisfies condition (2.4) at the discontinuity. Figure 7.2 shows the numerical probability densities at $t = 1$ for the three methods. Notice that the probability density obtained using the central difference method is *negative* for $x > 0.5$. This is definitely wrong because the probability can never be negative. This defect of the central difference method is caused by the fact that the jump rates (3.4) may be negative at the discontinuity. Suppose the discontinuity is at $x_{l+1/2}$. The numerical probability flux at $x_{l+1/2}$ is

$$\begin{aligned} J_{l+1/2} &= h\left(F_{l+1/2}\, p_l - B_{l+1/2}\, p_{l+1}\right) \\ &= h\, B_{l+1/2}\, p_l \left[\frac{F_{l+1/2}}{B_{l+1/2}} - \frac{p_{l+1}}{p_l}\right]. \end{aligned} \tag{7.5}$$

A necessary condition for converging to the correct solution is

$$\begin{aligned} &\lim_{h\to 0} p_l = \rho(x_d^-, t), \\ &\lim_{h\to 0} p_{l+1} = \rho(x_d^+, t), \\ &\lim_{h\to 0} J_{l+1/2} = \text{finite}. \end{aligned}$$

Applying condition (2.4) yields

$$\lim_{h\to 0} \frac{p_{l+1}}{p_l} = \frac{\rho(x_d^+, t)}{\rho(x_d^-, t)} = e^{\psi(x_d^-) - \psi(x_d^+)}.$$

Multiplying (7.5) by h, using $h^2 B_{l+1/2} = O(1)$, and taking the limit as $h \to 0$, we obtain

$$\lim_{h\to 0} \frac{F_{l+1/2}}{B_{l+1/2}} = e^{\psi(x_d^-)-\psi(x_d^+)},$$

which reduces to detailed balance (3.10). Therefore, in the presence of discontinuities, detailed balance is a necessary condition for converging to the correct solution of the differential equation. Both Method 1 developed in [16] and Method 2 proposed in this paper satisfy detailed balance. The central difference method does not. For the model problem (7.1), we have

$$e^{\psi(x_d^-)-\psi(x_d^+)} = e^{-3},$$
$$\lim_{h\to 0} \frac{F^{(CD)}_{l+1/2}}{B^{(CD)}_{l+1/2}} = \frac{1 - \frac{\psi(x_d^+)-\psi(x_d^-)}{2}}{1 + \frac{\psi(x_d^+)-\psi(x_d^-)}{2}} = \frac{-1}{5} < 0.$$

The negative value of $\frac{F_{l+1/2}}{B_{l+1/2}}$ will force the ratio $\frac{p_{l+1}}{p_l}$ to be negative, which leads to negative probability in the numerical solution of the central difference method.

In conclusion, we have proved that Method 1 developed in [16] is stable and is consistent with the Fokker–Planck equation. Method 1 converges to the correct solution of the Fokker–Planck equation, and the 2-norm of the error behaves like $O(k^2 + h)$ when the potential is discontinuous. Numerical results indicate that the ∞-norm of the error is of the same order. Based on the consistency analysis, we proposed a modified version of Method 1 to eliminate the first order error caused by the discontinuity. The modified numerical method (Method 2) is guaranteed to converge to the correct solution, and the 2-norm of the error behaves like $O(k^2 + h^2)$ even in the presence of discontinuities. Again, numerical results indicate that the ∞-norm of the error is of the same order.

In stochastic ratchets with discontinuous force [20], also known as sharp stochastic ratchets [21], the potential is continuous, but the derivative of the potential is discontinuous. Originally in [20] and subsequently in [21], the transport in stochastic ratchets was studied where a particle is driven by a continuous piecewise linear potential, the Brownian noise (white noise), and an additional colored noise. They derived analytic expressions for the steady state particle current for various asymptotic limits. Now we look at the convergence of numerical methods in the case of sharp stochastic ratchets. When the potential is continuous and piecewise smooth, both Method 1 and Method 2 (the modified numerical method) converge, and the error behaves like $O(k^2 + h^2)$. This can be seen by going back to section 5. In the local truncation error in (5.20), the leading term $O(kh)$ comes from the term $v_0(t_n)$ in (5.9). $v_0(t_n)$ is nonzero only at discontinuities. For Method 1, $v_0(t_n)$ is given in (5.10). At a discontinuity on the derivative of a continuous function, we have $\psi_L = \psi_R$ and consequently $v_0(t_n) = 0$. Thus, Method 1 is second order when the potential is continuous. Method 2 (the modified numerical method) is already second order even when the potential is discontinuous.

All of the conclusions above for Method 1 are also true for the numerical method previously developed by Elston and Doering [18]. More specifically, the numerical stability proved in section 4 depends on two main features of the numerical method, (i) all jump rates being positive and (ii) detailed balance being preserved, which are satisfied in the method of [18]. The numerical consistence away from the discontinuities is obtained by doing Taylor expansion. The key feature we utilized in section 5

to derive the numerical consistence at the discontinuities is again detailed balance. Therefore, all the analysis in sections 4 and 5 for Method 1 can be extended to the method in [18].

In the numerical simulations above, we also examined the behavior of the central difference method. We found that in the presence of discontinuities, the central difference method converges to a wrong solution that does not satisfy condition (2.4). We showed that in the presence of discontinuities, detailed balance is a necessary condition for converging to the correct solution. Both Method 1 and Method 2 satisfy detailed balance. The central difference method does not, which explains the fatal defect of the central difference method.

Appendix A. In this appendix, we derive (5.6). We start by expanding function $\frac{x}{e^x-1}$ around $x=0$:

$$\frac{x}{e^x-1} = 1 - \frac{1}{2}x + \frac{1}{12}x^2 + 0 \times x^3 + O\left(x^4\right). \tag{A.1}$$

Using (A.1) to expand jump rates $F_{j+1/2}$ and $B_{j+1/2}$ in terms of $\delta\psi_{j+1/2}$, we get

$$\begin{aligned} F_{j+1/2} &= \frac{1}{h^2}\frac{\delta\psi_{j+1/2}}{e^{\delta\psi_{j+1/2}}-1} \\ &= \frac{1}{h^2}\left[1 - \frac{1}{2}\delta\psi_{j+1/2} + \frac{1}{12}(\delta\psi_{j+1/2})^2 + O\left((\delta\psi_{j+1/2})^4\right)\right], \end{aligned} \tag{A.2}$$

$$\begin{aligned} B_{j+1/2} &= \frac{1}{h^2}\frac{\delta\psi_{j+1/2}\, e^{\delta\psi_{j+1/2}}}{e^{\delta\psi_{j+1/2}}-1} \\ &= \frac{1}{h^2}\frac{(-\delta\psi_{j+1/2})}{e^{-\delta\psi_{j+1/2}}-1} \\ &= \frac{1}{h^2}\left[1 + \frac{1}{2}\delta\psi_{j+1/2} + \frac{1}{12}(\delta\psi_{j+1/2})^2 + O\left((\delta\psi_{j+1/2})^4\right)\right]. \end{aligned} \tag{A.3}$$

Substituting $F_{j+1/2}$, $B_{j+1/2}$, ρ_j^n, and ρ_{j+1}^n into (5.2) yields

$$\begin{aligned} J_{j+1/2}^n &= h\left(F_{j+1/2}\,\rho_j^n - B_{j+1/2}\,\rho_{j+1}^n\right) \\ &= \frac{1}{h}\left[\left(\rho_j^n - \rho_{j+1}^n\right) - \frac{1}{2}\delta\psi_{j+1/2}\left(\rho_j^n + \rho_{j+1}^n\right)\right. \\ &\qquad \left. + \frac{1}{12}(\delta\psi_{j+1/2})^2\left(\rho_j^n - \rho_{j+1}^n\right) + O\left((\delta\psi_{j+1/2})^4\right)\right]. \end{aligned} \tag{A.4}$$

Expanding $\delta\psi_{j+1/2}$, ρ_j^n, and ρ_{j+1}^n around $x = x_{j+1/2}$, we have

$$\delta\psi_{j+1/2} = h\,\psi'_{j+1/2} + \frac{h^3}{24}\psi'''_{j+1/2} + O\left(h^5\right),$$

$$\rho_j^n - \rho_{j+1}^n = -h\left(\frac{\partial\rho}{\partial x}\right)_{j+1/2}^n - \frac{h^3}{24}\left(\frac{\partial^3\rho}{\partial x^3}\right)_{j+1/2}^n + O\left(h^5\right),$$

$$\rho_j^n + \rho_{j+1}^n = 2\left(\rho(x,t)\right)_{j+1/2}^n + \frac{h^2}{4}\left(\frac{\partial^2\rho}{\partial x^2}\right)_{j+1/2}^n + O\left(h^4\right).$$

Here we used the shorthand notation $(g(x,t))_j^n = g(x_j, t_n)$. Substituting these expansions into (A.4), we obtain the expansion for the probability flux

$$J_{j+1/2}^n = -\left(\psi'\rho + \frac{\partial \rho}{\partial x}\right)_{j+1/2}^n - h^2\left(\frac{1}{24}\frac{\partial^3 \rho}{\partial x^3} + \frac{1}{8}\psi'\frac{\partial^2 \rho}{\partial x^2} + \frac{1}{12}(\psi')^2\frac{\partial \rho}{\partial x} + \frac{1}{24}\psi'''\rho\right)_{j+1/2}^n + O\left(h^3\right),$$

which corresponds to (5.6).

Appendix B. In this appendix, we prove Theorem 5.1. We first rewrite the numerical probability flux as

$$\begin{aligned} J_{j+1/2} &= h\left(F_{j+1/2}\, r_j - B_{j+1/2}\, r_{j+1}\right) \\ &= \frac{1}{h}\,\frac{\psi_{j+1}-\psi_j}{e^{\psi_{j+1}}-e^{\psi_j}} e^{\psi_j}\, r_j - \frac{1}{h}\,\frac{\psi_{j+1}-\psi_j}{e^{\psi_{j+1}}-e^{\psi_j}} e^{\psi_{j+1}}\, r_{j+1} \\ &= \frac{1}{h}\,\frac{\psi_{j+1}-\psi_j}{e^{\psi_{j+1}}-e^{\psi_j}}\left(e^{\psi_j}\, r_j - e^{\psi_{j+1}}\, r_{j+1}\right). \end{aligned} \tag{B.1}$$

Suppose r_j is the solution of (5.15) and (5.16). Let $\tilde{r}_j = e^{\psi_j}\, r_j$. $\tilde{r}_j$ satisfies the equation

$$\frac{1}{h}\,\frac{\psi_j-\psi_{j-1}}{e^{\psi_j}-e^{\psi_{j-1}}}\left(\tilde{r}_{j-1}-\tilde{r}_j\right) - \frac{1}{h}\,\frac{\psi_{j+1}-\psi_j}{e^{\psi_{j+1}}-e^{\psi_j}}\left(\tilde{r}_j-\tilde{r}_{j+1}\right) = b_j \tag{B.2}$$

and the condition

$$\sum_{j=1}^{M} e^{-\psi_j}\,\tilde{r}_j = 0. \tag{B.3}$$

We construct $\tilde{r}_j$ starting at $j = l+1$ with

$$\tilde{r}_{l+1} = -c_1 \quad \text{and} \quad \frac{1}{h}\,\frac{\psi_{l+2}-\psi_{l+1}}{e^{\psi_{l+2}}-e^{\psi_{l+1}}}\left(\tilde{r}_{l+1}-\tilde{r}_{l+2}\right) = -c_2,$$

where c_1 and c_2 are two coefficients to be determined. Because b_j, as defined in (5.13), satisfies $b_j = 0$ for $j = l+2, \ldots, M+l-1$, we immediately obtain that

$$\frac{1}{h}\,\frac{\psi_{j+1}-\psi_j}{e^{\psi_{j+1}}-e^{\psi_j}}\left(\tilde{r}_j-\tilde{r}_{j+1}\right) = -c_2 \quad \text{for } j = l+2, \ldots, M+l-1.$$

This allows us to write $\tilde{r}_{j+1}$ in terms of $\tilde{r}_j$:

$$\tilde{r}_{j+1} = \tilde{r}_j + c_2\left(h\frac{e^{\psi_{j+1}}-e^{\psi_j}}{\psi_{j+1}-\psi_j}\right).$$

Summing over j, we get

$$\tilde{r}_i = -c_1 + c_2\sum_{j=l+1}^{i-1}\left(h\frac{e^{\psi_{j+1}}-e^{\psi_j}}{\psi_{j+1}-\psi_j}\right) \quad \text{for } i = l+2, \ldots, M+l. \tag{B.4}$$

For $\tilde{r}_j$ to solve (B.2), it needs to satisfy

$$\frac{1}{h}\,\frac{\psi_{l+1}-\psi_l}{e^{\psi_{l+1}}-e^{\psi_l}}\left(\tilde{r}_l-\tilde{r}_{l+1}\right) = \frac{1}{h} - c_2. \tag{B.5}$$

Using the fact that $\tilde{r}_j$ is periodic and substituting (B.4) into (B.5) yields an equation for c_2:

$$c_2 \frac{\psi_{l+1} - \psi_l}{e^{\psi_{l+1}} - e^{\psi_l}} \sum_{j=l+1}^{M+l-1} \left(h \frac{e^{\psi_{j+1}} - e^{\psi_j}}{\psi_{j+1} - \psi_j} \right) = 1 - c_2 h. \tag{B.6}$$

It follows that

$$c_2 = \left[h + \frac{\psi_{l+1} - \psi_l}{e^{\psi_{l+1}} - e^{\psi_l}} \sum_{j=l+1}^{M+l-1} \left(h \frac{e^{\psi_{j+1}} - e^{\psi_j}}{\psi_{j+1} - \psi_j} \right) \right]^{-1}. \tag{B.7}$$

The sum in (B.7) is approximately an integral

$$\sum_{j=l+1}^{M+l-1} \left(h \frac{e^{\psi_{j+1}} - e^{\psi_j}}{\psi_{j+1} - \psi_j} \right) = \int_0^1 e^{\psi(x)} dx + O(h).$$

Substituting this result into (B.7), we have

$$c_2 = \frac{e^{\psi_R} - e^{\psi_L}}{\psi_R - \psi_L} \left[\int_0^1 e^{\psi(x)} dx \right]^{-1} + O(h). \tag{B.8}$$

Thus, for h small enough, c_2 is positive and bounded.

c_1 is determined by condition (B.3). Notice that $\tilde{r}_j$, as given in (B.4), is monotonically increasing for $j = l+1, \ldots, M+l$. If $c_1 = 0$, then we have $\tilde{r}_{l+1} = 0$ and $\tilde{r}_j > 0$ for $j = l+2, \ldots, M+l$. Consequently, we have $\sum_{j=1}^{M} e^{-\psi_j} \tilde{r}_j > 0$. Now we select c_1 to make $\tilde{r}_{M+l} = 0$:

$$\hat{c}_1 = c_2 \sum_{j=l+1}^{M+l-1} \left(h \frac{e^{\psi_{j+1}} - e^{\psi_j}}{\psi_{j+1} - \psi_j} \right) = \frac{e^{\psi_R} - e^{\psi_L}}{\psi_R - \psi_L} + O(h).$$

In this case, we have $\tilde{r}_{M+l} = 0$ and $\tilde{r}_j < 0$ for $j = l+1, \ldots, M+l-1$. Consequently, we have $\sum_{j=1}^{M} e^{-\psi_j} \tilde{r}_j < 0$. The value of c_1 that satisfies condition (B.3) is between 0 and $\hat{c}_1$. Thus, for h small enough, c_1 is positive and bounded:

$$0 < c_1 < \frac{e^{\psi_R} - e^{\psi_L}}{\psi_R - \psi_L} + O(h). \tag{B.9}$$

Substituting (B.8) and (B.9) into (B.4), we conclude that

$$\max_j |\tilde{r}_j| \leq \frac{e^{\psi_R} - e^{\psi_L}}{\psi_R - \psi_L} + O(h), \tag{B.10}$$

which leads directly to the conclusion of Theorem 5.1.

Acknowledgment. The author would like to thank anonymous referees for their constructive comments and suggestions in improving the manuscript.

REFERENCES

[1] H. C. Berg, *Random Walks in Biology*, Princeton University Press, Princeton, NJ, 1993.

[2] J. Abrahams, A. Leslie, R. Lutter, and J. Walker, *Structure at* 2.8*A resolution of F*1-*ATPase from bovine heart mitochondria*, Nature, 370 (1994), pp. 621–628.
[3] H. Noji, R. Yasuda, M. Yoshida, and K. Kinosita, *Direct observation of the rotation of F*1-*ATPase*, Nature, 386 (1997), pp. 299–302.
[4] H. Wang and G. Oster, *Energy transduction in the F*1 *motor of ATP synthase*, Nature, 396 (1998), pp. 279–282.
[5] C. Coppin, D. Pierce, L. Hsu, and R. Vale, *The load dependence of kinesin's mechanical cycle*, Proc. Nat. Acad. Sci. U.S.A., 94 (1997), pp. 8539–8544.
[6] K. Visscher, M. Schnitzer, and S. Block, *Single kinesin molecules studied with a molecular force clamp*, Nature, 400 (1999), pp. 184–189.
[7] J. Prost, J. Chauwin, L. Peliti, and A. Ajdari, *Asymmetric pumping of particles*, Phys. Rev. Lett., 72 (1994), pp. 2652–2655.
[8] R. Astumian, *Thermodynamics and kinetics of a Brownian motor*, Science, 276 (1997), pp. 917–922.
[9] T. Elston, H. Wang, and G. Oster, *Energy transduction in ATP synthase*, Nature, 391 (1998), pp. 510–514.
[10] H. Wang, *Mathematical theory of molecular motors and a new approach for uncovering motor mechanism*, IEE Proc. Nanobiotechnol., 150 (2003), pp. 127–133.
[11] F. Reif, *Fundamentals of Statistical and Thermal Physics*, McGraw–Hill, New York, 1985.
[12] H. Risken, *The Fokker-Planck Equation*, 2nd ed., Springer, Berlin, 1989.
[13] A. Einstein, *Investigation on the Theory of the Brownian Motion*, Dover, New York, 1956.
[14] C. Peskin, G. Odell, and G. Oster, *Cellular motions and thermal fluctuations: The Brownian ratchet*, Biophys. J., 65 (1993), pp. 316–324.
[15] T. C. Elston and C. S. Peskin, *The role of protein flexibility in molecular motor function: Coupled diffusion in a tilted periodic potential*, SIAM J. Appl. Math., 60 (2000), pp. 842–867.
[16] H. Wang, C. Peskin, and T. Elston, *A robust numerical algorithm for studying biomolecular transport processes*, J. Theoret. Biol., 221 (2003), pp. 491–511.
[17] R. D. Richtmyer and K. W. Morton, *Difference Methods for Initial Value Problems*, Wiley-Interscience, New York, 1967.
[18] T. Elston and C. Doering, *Numerical and analytical studies of nonequilibrium fluctuation induced transport processes*, J. Statist. Phys., 83 (1996), pp. 359–383.
[19] M. L. Juncosa and D. Young, *On the Crank-Nicolson procedure for solving parabolic partial differential equations*, Proc. Cambridge Philos. Soc., 53 (1957), pp. 448–461.
[20] C. R. Doering, L. A. Dontcheva, and M. M. Klosek, *Constructive role of noise: Fast fluctuation asymptotics of transport in stochastic ratchets*, Chaos, 8 (1998), pp. 643–649.
[21] M. M. Kłosek and R. W. Cox, *Steady-state currents in sharp stochastic ratchets*, Phys. Rev. E (3), 60 (1999), pp. 3727–3735.
[22] R. Feynman, R. Leighton, and M. Sands, *The Feynman Lectures on Physics*, Addison–Wesley, Reading, MA, 1963.
[23] G. Dal Maso, Ph. Le Floch, and F. Murat, *Definition and weak stability of non-conservative products*, J. Math. Pures Appl. (9), 74 (1995), pp. 483–548.
[24] E. Godlewski and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Appl. Math. Sci. 118, Springer, New York, 1996.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1453–1471

ON CONVERGENCE OF A DOMAIN DECOMPOSITION METHOD FOR A SCALAR CONSERVATION LAW*

HUAZHONG TANG† AND GERALD WARNECKE‡

Abstract. In this paper, we prove convergence of a domain decomposition method for one-dimensional scalar conservation laws by dealing carefully with nonconservative terms at the interface of subdomains. The method consists of an explicit scheme in some subdomains and an implicit scheme in other subdomains with a numerical flux being the same as the one used in the explicit scheme. Although such a multidomain algorithm is not strictly conservative, the conservation error $CE(0, N\Delta t)$ is equal to $\mathcal{O}(\Delta t)$ regardless of the smoothness of the solution. Finally, two test examples are given to validate convergence and the computational efficiency of the present method.

Key words. hyperbolic conservation laws, domain decomposition method, explicit scheme, implicit scheme, convergence

AMS subject classifications. 65M06, 35L65, 65M99, 76M12

DOI. 10.1137/040607423

1. Introduction.

This paper is devoted to the study of convergence of a domain decomposition method (DDM) or multidomain algorithm for scalar conservation laws

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \tag{1.1}$$

with initial data

$$u(x, 0) = u_0(x), \tag{1.2}$$

where $x \in \mathbb{R}$, $t \geq 0$, and we assume that u_0 has compact support. We are considering admissible weak solutions to this initial value problem that satisfy entropy conditions such as the Kružkov entropy inequalities [20].

Hyperbolic conservation laws are of great practical importance since they arise in fluid flows, for example, reactive flows, groundwater flows, non-Newtonian flows, traffic flows, and two-phase flows in oil reservoirs. They also govern a variety of physical phenomena that appear in aeronautics, astrophysics, meteorology, semiconductors, financial modeling, front propagation, and other areas.

During the past few decades there has been a considerable amount of activity related to the construction of finite difference schemes for equations of type (1.1) and applications; see, for instance, [10, 13, 17, 19, 21, 22, 28, 29, 35] and the references therein. Explicit methods have been proved to be very efficient in capturing moving discontinuities or fronts, such as shock waves, detonation waves, and contact

*Received by the editors April 27, 2004; accepted for publication (in revised form) February 16, 2007; published electronically July 25, 2007. This research was partially supported by the National Basic Research Program under grant 2005CB321703, the National Natural Science Foundation of China (10431050, 10576001), the Alexander von Humboldt Foundation, and the Deutsche Forschungsgemeinschaft (DFG Wa 633/10-3).

http://www.siam.org/journals/sinum/45-4/60742.html

†LMAM, School of Mathematical Sciences, Peking University, Beijing 100871, People's Republic of China (hztang@math.pku.edu.cn).

‡Institut für Analysis und Numerik, Otto-von-Guericke Universität Magdeburg, 39106 Magdeburg, Germany (Gerald.Warnecke@mathematik.uni-magdeburg.de).

discontinuities. However, they need a small time step size satisfying a Courant–Friedrichs–Lewy (CFL) condition to guarantee stability. For an implicit scheme, the time step size is also often constrained by convergence. An implicit scheme usually requires solving a nonlinear equation by some iteration method, and thus it is very time-consuming. Even for the same time step size as used in the explicit case, unsteady solutions of the implicit schemes are less accurate; see [19]. However, the implicit scheme is very attractive in simulating steady state solutions. The spatial step sizes and the "signal" speeds are the two main elements that limit a choice of the time step size. Hence, when solving numerically some initial boundary value problems (IBVP) for nonlinear partial differential equations (PDEs), it may occur that in some spatial regions there is the need for a smaller time step than in other regions. Typical examples are numerical simulations of viscous fluid flows on nonuniform meshes and other computations of solutions to PDEs on an adaptive mesh [2, 25]. Due to the above reason, the large time step schemes [23, 36, 37] and the local time step schemes [2, 7, 25, 32, 33] become attractive. The large time step schemes satisfy the CFL condition by automatically increasing the stencil with the size of the time step. They can correctly give the location of shocks with virtually no smearing, but they seem to be inconvenient in practical applications, especially in treating boundary conditions.

The local time step schemes are restricted only by a local stability condition rather than the traditional global stability condition dominated by the smallest cells. The schemes studied in [2, 25] are conservative, but they suffer a loss of consistency near a time grid interface in terms of truncation errors. The schemes proposed in [32] are consistent but slightly nonconservative. Recently, Berger, Helzel, and LeVeque [3] presented an h-box method approximating hyperbolic conservation laws on irregular grids that had a time step restriction based on a reference grid cell length that could be orders of magnitude larger than the smallest grid cell arising in the discretization.

The discrete conservation of a numerical algorithm for (1.1) is important in order to keep the correct location of the discontinuities. Hou and LeFloch [16] showed that if a nonconservative scheme for (1.1) converges, it converges to a solution of $\partial_t u + \partial_x f(u) = \mu$, where μ is a Borel measure source term that is expected to be zero in the region where the solution u is smooth and concentrated where u is not smooth. Tang and Zhou [30] analyzed the conservation error of a numerical solution caused by a nonconservative interface matching for grid interfaces and showed that the error had an upper bound when the solution itself was bounded. Even so, nonconservative schemes are also valuable in some practical applications and have been implemented successfully, for example, in computations of compressible multifluids [1] and fluid flows on an overlapping grid [26].

The DDM has been widely used in solving elliptic or parabolic equations and parallel computation of some large-scale problems. There also exist some works on applications in solving hyperbolic conservation laws. For example, Quarteroni [27] used DDMs for systems of conservation laws in connection with a spectral approximation. Grott, Chernigovskij, and Glatzel [14] applied an implicit scheme with vastly different time scales to the numerical simulation of stellar instabilities by using a DDM. However, due to a possible loss of conservation, there seems to be no theoretical analysis of a DDM for hyperbolic conservation laws.

This paper attempts to conduct a study of convergence of a DDM for hyperbolic conservation laws. The DDM we consider consists of an explicit scheme in some subdomains and an implicit scheme in other subdomains. Due to the local implicit character, such a multidomain algorithm can be used to reduce the CFL restriction

on the time step size and improve efficiency of simulating numerically some problems with multitime scale phenomena. The DDM is not strictly conservative, but we can prove that under a local stability condition for the explicit schemes, the approximate solutions constructed by the DDM will converge to the unique entropy solution to the scalar conservation laws. The main technical point in our convergence proof is to treat nonconservative terms suitably.

This paper is organized as follows. In section 2, we introduce the DDM. We consider only one space dimension. Section 3 is devoted to a study of nonlinear stability of the algorithm introduced in section 2, including the maximum principle, total variation (TV) stability, and L^1-continuity in time. In section 4 we give a proof of a cell entropy inequality. In section 5 we focus on a study of convergence of the approximate solutions constructed by the DDM. Section 6 presents numerical experiments to validate the theoretical results derived in sections 3–5.

2. A domain decomposition method. We define the domains $\Omega^1 = \{x | x \le a \text{ or } x \ge b\}$ and $\Omega^2 = \{x | a \le x \le b\}$, where a and b are two given constants, $-\infty < a < b < \infty$. Each of them is equipped with an individual mesh with a varying space step size $\Delta x_{j+\frac{1}{2}} = x_{j+1} - x_j$, where x_j denotes the coordinate of the jth grid point. We also discretize in time. Define $\Delta t_n = t_{n+1} - t_n > 0$, $n \ge 0$. As a measure of refinement we introduce $\delta = \max_{j\in\mathbb{Z},n\ge 0}\{\Delta x_{j+\frac{1}{2}}, \Delta t_n\}$. Moreover, throughout this paper, we will also use the following notation: $x_{j+\frac{1}{2}} = \frac{1}{2}(x_{j+1} + x_j)$, $\Omega^1_\delta = \{j | j < j_1 \text{ or } j \ge j_2, j \in \mathbb{Z}\}$, and $\Omega^2_\delta = \{j | j_1 \le j < j_2, j \in \mathbb{Z}\}$, where $j_1 = j_1(\delta), j_2 = j_2(\delta)$, $x_{j_1} = a$, and $x_{j_2} = b$. We also assume that $\Delta x_{j+\frac{1}{2}}/\Delta x_{j-\frac{1}{2}} = 1 + \mathcal{O}(\delta^\alpha)$ for all $j \in \mathbb{Z}$, or $\Delta t_{n-1}/\Delta t_n = 1 + \mathcal{O}(\delta^\alpha)$ for all $n \in \mathbb{N}$, where α is a positive number.

The initial condition in (1.2) is projected onto the space of piecewise constant functions as

$$u^0_{j+\frac{1}{2}} = \frac{1}{\Delta x_{j+\frac{1}{2}}} \int_{x_j}^{x_{j+1}} u_0(x)\, dx, \quad j \in \mathbb{Z}. \tag{2.1}$$

Our DDM is described as follows. A three-point explicit conservative scheme,

$$u^{n+1}_{j+\frac{1}{2}} = u^n_{j+\frac{1}{2}} - \lambda^n_{j+\frac{1}{2}} \left(h(u^n_{j+\frac{1}{2}}, u^n_{j+\frac{3}{2}}) - h(u^n_{j-\frac{1}{2}}, u^n_{j+\frac{1}{2}}) \right), \tag{2.2}$$

is used in the subdomain Ω^1, while a three-point implicit conservative scheme,

$$u^{n+1}_{j+\frac{1}{2}} = u^n_{j+\frac{1}{2}} - \lambda^n_{j+\frac{1}{2}} \left(h(u^{n+1}_{j+\frac{1}{2}}, u^{n+1}_{j+\frac{3}{2}}) - h(u^{n+1}_{j-\frac{1}{2}}, u^{n+1}_{j+\frac{1}{2}}) \right), \tag{2.3}$$

is adopted in the subdomain Ω^2, where $\lambda^n_{j+\frac{1}{2}} = \Delta t_n/\Delta x_{j+\frac{1}{2}}$. The numerical flux $h(u, v)$ is assumed to be $C^1(\mathbb{R}^2)$ and satisfy $|h_1(u, v)|, |h_2(u, v)| \le L < \infty$ for all $u, v \in \mathbb{R}$, where $h_1(\cdot, \cdot)$ and $h_2(\cdot, \cdot)$ denote the partial derivatives of $h(\cdot, \cdot)$ with respect to its first and second arguments, respectively. Moreover, we also assume that the numerical flux $h(u, v)$ is monotone; that is to say, $h(u, v)$ is nondecreasing in the first variable and nonincreasing in the second variable. The numerical flux $h(u, v)$ satisfies the consistency condition

$$h(u, u) = f(u). \tag{2.4}$$

Obviously, conservation of the total mass cannot be preserved exactly; for example, if we assume that u approaches the same constant when $j \to \pm\infty$, then we

have

$$\sum_{j\in\mathbb{R}} u^{n+1}_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}} \equiv \sum_{j<j_1} u^{n+1}_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}} + \sum_{j\geq j_2} u^{n+1}_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}} + \sum_{j_1\leq j<j_2} u^{n+1}_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}}$$
$$= \sum_{j\in\mathbb{R}} u^{n}_{j+\frac{1}{2}}\Delta x_{j+\frac{1}{2}} + CE(t_n, t_{n+1}), \tag{2.5}$$

where the additional term $CE(t_n, t_{n+1})$ is defined by

$$CE(t_n,t_{n+1}) := \Delta t_n\Big(h(u^{n+1}_{j_1-\frac{1}{2}}, u^{n+1}_{j_1+\frac{1}{2}}) - h(u^{n}_{j_1-\frac{1}{2}}, u^{n}_{j_1+\frac{1}{2}}) + h(u^{n}_{j_2-\frac{1}{2}}, u^{n}_{j_2+\frac{1}{2}})$$
$$- h(u^{n+1}_{j_2-\frac{1}{2}}, u^{n+1}_{j_2+\frac{1}{2}})\Big) \leq L\Delta t_n\Big(|u^{n}_{j_1-\frac{1}{2}} - u^{n+1}_{j_1-\frac{1}{2}}| + |u^{n}_{j_1+\frac{1}{2}} - u^{n+1}_{j_1+\frac{1}{2}}|$$
$$+ |u^{n}_{j_2-\frac{1}{2}} - u^{n+1}_{j_2-\frac{1}{2}}| + |u^{n}_{j_2+\frac{1}{2}} - u^{n+1}_{j_2+\frac{1}{2}}|\Big),$$

which is generated by the DDM algorithm at the interfaces of two subdomains. It can be considered as a measure of conservation error. It is obvious that the term $CE(t_n, t_{n+1})$ will be $\mathcal{O}(\Delta t_n^2)$ in the smooth regions of the solution, but it is only $\mathcal{O}(\Delta t_n)$ if a discontinuity is interacting with the corresponding interface $\{x = x_{j_1}, t_n \leq t \leq t_{n+1}\}$ or $\{x = x_{j_2}, t_n \leq t \leq t_{n+1}\}$. If we assume that the time step size is a constant, i.e., $\Delta t_n = \Delta t$, then the conservation error $CE(0, N\Delta t)$ is equal to

$$CE(0,N\Delta t) = \sum_{n=0}^{N-1} CE\big((n-1)\Delta t, n\Delta t\big) = \Delta t\Big(h(u^{N}_{j_1-\frac{1}{2}}, u^{N}_{j_1+\frac{1}{2}})$$
$$- h(u^{0}_{j_1-\frac{1}{2}}, u^{0}_{j_1+\frac{1}{2}})\Big) + \Delta t\left(h(u^{0}_{j_2-\frac{1}{2}}, u^{0}_{j_2+\frac{1}{2}}) - h(u^{N}_{j_2-\frac{1}{2}}, u^{N}_{j_2+\frac{1}{2}})\right)$$
$$\leq L\Delta t_n\Big(|u^{N}_{j_1-\frac{1}{2}} - u^{0}_{j_1-\frac{1}{2}}| + |u^{N}_{j_1+\frac{1}{2}} - u^{0}_{j_1+\frac{1}{2}}|$$
$$+ |u^{N}_{j_2-\frac{1}{2}} - u^{0}_{j_2-\frac{1}{2}}| + |u^{N}_{j_2+\frac{1}{2}} - u^{0}_{j_2+\frac{1}{2}}|\Big). \tag{2.6}$$

From (2.6), we may conclude that the conservative error $CE(0, N\Delta t)$ is generally $\mathcal{O}(\Delta t)$ if the solution is smooth, or a discontinuity with a finite strength is interacting with the interface $x = x_{j_1}$ or x_{j_2}. It means that the above DDM is not a bad approximation to (1.1). It is worth noting that the conservation error $CE(0, N\Delta t)$ of a general nonconservative scheme is usually equal to $\mathcal{O}(\Delta t)$ in the smooth regions of the solution and to $\mathcal{O}(1)$ when the solution is discontinuous. In practical applications, we can generally avoid the interaction of discontinuities with the interface of subdomains by decomposing the domains appropriately at the different times. In test cases where the shock position is known, for instance, one can avoid decomposing there. In applications where this is not the case an approximate estimate of shock positions, e.g., by a numerical shock indicator, would have to be used. Nevertheless, it seems to be difficult to improve the total nonconservation error $CE(0, N\Delta t)$ in theory because

$$CE(0,N\Delta t) = \sum_{n=0}^{N-1} CE\big((n-1)\Delta t, n\Delta t\big) = N\mathcal{O}(\Delta t^2) = \mathcal{O}(\Delta t).$$

Remark 2.1. The two numerical fluxes adopted in (2.2) and (2.3) have been assumed to be the same. This will be crucial to get convergence of the approximate

solutions constructed by the present algorithm. However, we may use a θ-scheme to replace the above fully implicit scheme (2.3), for example,

$$u_{j+\frac{1}{2}}^{n+1} = u_{j+\frac{1}{2}}^{n} + \lambda_{j+\frac{1}{2}}^{n}\Delta_+\left(\theta h(u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}) + (1-\theta)h(u_{j-\frac{1}{2}}^{n}, u_{j+\frac{1}{2}}^{n})\right), \tag{2.7}$$

where θ is a given constant, $0 \le \theta \le 1$, and Δ_+ denotes the forward difference operator in space. When $\theta = 0$, we have a trivial case: the explicit scheme (2.2) is exploited in the whole computational domain. For this trivial case, the time step size is constrained by the global CFL condition; convergence of the nonoverlapping DDM can be obtained by following the proof of convergence of a single domain method.

The motivation of using different schemes within different subdomains is to relax the CFL restriction and improve the efficiency of simulating numerically some problems with multitime scale phenomena. Moreover, the results derived in the following are also analogously valid if the number of subdomains is any finite number m. Further, they hold for any multipoint monotone scheme when the fluxes are interpreted accordingly.

3. The main properties. In this section we shall analyze mainly the nonlinear stability of the DDM (2.2) and (2.3) introduced in the last section. Before doing this, we first show existence of a solution of (2.3) with given boundary conditions at x_{j_1} and x_{j_2}. A similar question has been discussed by several authors; see, e.g., [9, 11, 24]. By using the inverse positivity property and Brouwer's fixed point theorem, Fuhrmann [11] gave a proof of existence and uniqueness of solutions of certain systems of algebraic equations with off-diagonal nonlinearity, which arise, e.g., from stable finite volume discretizations of viscous conservation laws; see [12]. Lucier [24], as well as Evje and Karlsen [9], used the theory of accretive operators introduced in [4, 18] and the Crandall–Liggett theorem [5] to prove the existence and uniqueness of solutions of their system of algebraic equations, too.

LEMMA 3.1. *Assume that the initial data* $\{u_{j+\frac{1}{2}}^{0} | j \in \mathbb{Z}\}$ *are given by* (2.1)*, that* $\{u_{j+\frac{1}{2}}^{n+1} | j \in \Omega_\delta^1\}$ *is the explicit solution of* (2.2) *at* $t = t_{n+1}$*, and that* $\{u_{j+\frac{1}{2}}^{n+1} | j \in \Omega_\delta^2\}$ *is defined by the implicit equation* (2.3)*. Then the solution* $\{u_{j+\frac{1}{2}}^{n+1} | j \in \Omega_\delta^2\}$ *exists uniquely; that is, the boundary value problem of* (2.3) *with boundary conditions at* x_{j_1} *and* x_{j_2} *computed by the explicit scheme* (2.2) *admits a unique solution. Moreover, if we assume that the initial data* $\{u_{j+\frac{1}{2}}^{0} | j \in \mathbb{Z}\}$ *are bounded, a bound for* $u_{j+\frac{1}{2}}^{n+1}$ *for all* $j \in \Omega_\delta^2$ *exists for fixed* δ *and independent of* n*.*

Proof. We consider the boundary value problem of the implicit scheme (2.3), where the boundary values are obtained from the computed values of the explicit scheme (2.2) at x_{j_1} and x_{j_2}. It can be written in a matrix form as follows:

$$\begin{cases} \boldsymbol{u}^{n+1} + \Delta t_n A(\boldsymbol{u}^{n+1}) = \boldsymbol{u}^n, \\ u_{j_1-\frac{1}{2}}^{n+1} \text{ and } u_{j_2+\frac{1}{2}}^{n+1} \text{ are given by the explicit scheme (2.2)}, \end{cases} \tag{3.1}$$

where $\boldsymbol{u} = (u_{j_1+\frac{1}{2}}, \dots, u_{j_2-\frac{1}{2}})^T$, $A(\boldsymbol{u}) = \left(\frac{1}{\Delta x_{j_1+\frac{1}{2}}} a_{j_1+\frac{1}{2}}(\boldsymbol{u}), \dots, \frac{1}{\Delta x_{j_2-\frac{1}{2}}} a_{j_2-\frac{1}{2}}(\boldsymbol{u})\right)^T$, and

$$a_{j+\frac{1}{2}}(\boldsymbol{u}) = h(u_{j+\frac{1}{2}}, u_{j+\frac{3}{2}}) - h(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}}), \quad j \in \Omega_\delta^2.$$

It remains to show the existence of a solution of the scheme (3.1) following [9] which is based on [24]. The existence needed here is on a finite dimensional space. If the

domain is semi-infinite or infinite by switching the two domains, we have to consider a suitable version of ℓ^1 as our solution space. It suffices to show that the operator $A : \ell^1 \longrightarrow \ell^1$ is accretive; i.e., the operator A satisfies

$$\mathcal{I} := \sum_{j\in\Omega_\delta^2} \operatorname{sign}(w_{j+\frac{1}{2}})\big(a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v})\big) \geq 0$$

for all $\boldsymbol{u}$ and $\boldsymbol{v}$ under consideration, e.g., satisfying $u_{j_1-\frac{1}{2}} = v_{j_1-\frac{1}{2}}$ and $u_{j_2+\frac{1}{2}} = v_{j_2+\frac{1}{2}}$, where $w_{j+\frac{1}{2}} = u_{j+\frac{1}{2}} - v_{j+\frac{1}{2}}$. In fact, because for each $j \in \Omega_\delta^2$ and a positive constant $c \geq 2L$, the inequality

$$\begin{aligned}\Big|cw_{j+\frac{1}{2}} - \big(a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v})\big)\Big| &\equiv \Big|c|w_{j+\frac{1}{2}}| - \operatorname{sign}(w_{j+\frac{1}{2}})\big(a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v})\big)\Big| \\ &\geq c|w_{j+\frac{1}{2}}| - \operatorname{sign}(w_{j+\frac{1}{2}})\big(a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v})\big)\end{aligned}$$

holds, we have

$$\mathcal{I} \geq -\sum_{j\in\Omega_\delta^2} \Big|cw_{j+\frac{1}{2}} - \big(a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v})\big)\Big| + c\sum_{j\in\Omega_\delta^2} |w_{j+\frac{1}{2}}|.$$

Define

$$a_j = \begin{cases} \frac{h(u_{j-\frac{1}{2}},u_{j+\frac{1}{2}}) - h(v_{j-\frac{1}{2}},u_{j+\frac{1}{2}})}{u_{j-\frac{1}{2}} - v_{j-\frac{1}{2}}} & \text{if } u_{j-\frac{1}{2}} \neq v_{j-\frac{1}{2}}, \\ h_1(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}}) & \text{otherwise,} \end{cases}$$

$$b_j = \begin{cases} \frac{h(v_{j-\frac{1}{2}},u_{j+\frac{1}{2}}) - h(v_{j-\frac{1}{2}},v_{j+\frac{1}{2}})}{u_{j+\frac{1}{2}} - v_{j+\frac{1}{2}}} & \text{if } u_{j+\frac{1}{2}} \neq v_{j+\frac{1}{2}}, \\ h_2(v_{j-\frac{1}{2}}, v_{j+\frac{1}{2}}) & \text{otherwise.} \end{cases}$$

Obviously, we have $a_j \geq 0$ and $b_j \leq 0$, because the numerical flux $h(\cdot,\cdot)$ is monotone. Moreover, we can rewrite $a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v})$ as follows:

$$\begin{aligned}a_{j+\frac{1}{2}}(\boldsymbol{u}) - a_{j+\frac{1}{2}}(\boldsymbol{v}) &= h(u_{j+\frac{1}{2}}, u_{j+\frac{3}{2}}) - h(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}}) - h(v_{j+\frac{1}{2}}, v_{j+\frac{3}{2}}) + h(v_{j-\frac{1}{2}}, v_{j+\frac{1}{2}}) \\ &= (a_{j+1} - b_j)w_{j+\frac{1}{2}} + b_{j+1}w_{j+\frac{3}{2}} - a_j w_{j-\frac{1}{2}}.\end{aligned}$$

Because of $c \geq (a_{j+1} - b_j) \geq 0$ and $w_{j_1-\frac{1}{2}} = w_{j_2+\frac{1}{2}} = 0$, we have

$$\begin{aligned}\mathcal{I} &\geq c\sum_{j\in\Omega_\delta^2} |w_{j+\frac{1}{2}}| - \sum_{j\in\Omega_\delta^2} (c - a_{j+1} + b_j)\,|w_{j+\frac{1}{2}}| + \sum_{j\in\Omega_\delta^2} b_{j+1}|w_{j+\frac{3}{2}}| - \sum_{j\in\Omega_\delta^2} a_j|w_{j-\frac{1}{2}}| \\ &= a_{j_1+1}|w_{j_1+\frac{1}{2}}| - b_{j_2-1}|w_{j_2-\frac{1}{2}}| \geq 0.\end{aligned}$$

Therefore, the operator A is accretive.

On the other hand, because the numerical flux $h(u,v) \in C^1(\mathbb{R}^2)$, the operator A is not only accretive but also m-accretive; i.e., for all positive numbers λ, $I + \lambda A$ is a surjection, where I denotes the identity operator. Therefore, in view of the well-known results of Lucier [24] as well as Crandall and Liggett [5], we can conclude the existence of a unique solution of (3.1).

Since for fixed δ there is only a finite number of real terms in (3.1), $u_{j+\frac{1}{2}}^{n+1}$, $j_1 \leq j < j_2$, can be computed if $u_{j+\frac{1}{2}}^{n} \in \mathbb{R}$. Moreover, if we assume that the initial data

$\{u^0_{j+\frac{1}{2}} | j \in \mathbb{Z}\}$ are bounded, a bound for $u^{n+1}_{j+\frac{1}{2}}$ for all $j \in \Omega^2_\delta$ exists for fixed δ and independent of n, which will be proved later. □

Mimicking the proof of Lemma 3.1, the following result is easily proved.

LEMMA 3.2. *Assuming that* $\{u^{n+1}_{j_1-\frac{1}{2}}, u^{n+1}_{j_2+\frac{1}{2}}, u^n_{j+\frac{1}{2}}, j = j_1, \ldots, j_2 - 1\}$ *and* $\{v^{n+1}_{j_1-\frac{1}{2}}, v^{n+1}_{j_2+\frac{1}{2}}, v^n_{j+\frac{1}{2}}, j = j_1, \ldots, j_2 - 1\}$ *are two arbitrarily given initial data and boundary conditions, then the two corresponding solutions of* (3.1), $\boldsymbol{u}^{n+1}$ *and* $\boldsymbol{v}^{n+1}$, *satisfy*

$$\begin{aligned}\sum_{j=j_1}^{j_2-1} |u^{n+1}_{j+\frac{1}{2}} - v^{n+1}_{j+\frac{1}{2}}| \Delta x_{j+\frac{1}{2}} &\le \sum_{j=j_1}^{j_2-1} |u^n_{j+\frac{1}{2}} - v^n_{j+\frac{1}{2}}| \Delta x_{j+\frac{1}{2}} \\ &\quad + \Delta t_n L \left(|u^{n+1}_{j_1-\frac{1}{2}} - v^{n+1}_{j_1-\frac{1}{2}}| + |u^{n+1}_{j_2+\frac{1}{2}} - v^{n+1}_{j_2+\frac{1}{2}}| \right). \end{aligned} \tag{3.2}$$

Moreover, if $u^{n+1}_{j_1-\frac{1}{2}}$, $u^{n+1}_{j_2+\frac{1}{2}}$, *and* $u^n_{j+\frac{1}{2}}$, $j = j_1, \ldots, j_2 - 1$, *are bounded, so is* $u^{n+1}_{j+\frac{1}{2}}$ *for all* $j \in \Omega^2_\delta$.

Remark 3.1. If we assume that the time step sizes satisfy a suitable restriction, then we may also present a proof of Lemma 3.1 by using the uniform monotonicity theorem of Dekker and Verwer; see page 147 of the book [8].

3.1. Maximum principle. In this subsection we show the existence of a maximum principle satisfied by the solutions of the DDM (2.2) and (2.3).

LEMMA 3.3. *If the initial data* $\{u^0_{j+\frac{1}{2}}\}_{j\in\mathbb{Z}}$ *satisfy*

$$m \le u^0_{j+\frac{1}{2}} \le M \quad \forall j \in \mathbb{Z}, \tag{3.3}$$

then under the local CFL condition

$$\lambda^n_{j+\frac{1}{2}} \max_{u,v,w,z\in\mathcal{A}} \{|h_1(u,v)| + |h_2(w,z)|\} \le 1 \tag{3.4}$$

for all $j \in \Omega^1_\delta$, *where* $\mathcal{A} = \{w \in L^\infty(\mathbb{R}) \big| ||w||_{L^\infty(\mathbb{R})} \le ||u^0||_{L^\infty(\mathbb{R})}\}$, *we have*

$$m \le u^{n+1}_{j+\frac{1}{2}} \le M \quad \forall j \in \mathbb{Z},\ n \ge 0. \tag{3.5}$$

Proof. Assume that $m \le u^n_{j+\frac{1}{2}} \le M$ for all $j \in \mathbb{Z}$. First, we consider the solutions in the domain Ω^1_δ. We obtain the estimate

$$\begin{aligned} u^{n+1}_{j+\frac{1}{2}} - m &= u^n_{j+\frac{1}{2}} - m - \lambda^n_{j+\frac{1}{2}} \left(h(u^n_{j+\frac{1}{2}}, u^n_{j+\frac{3}{2}}) - h(m, u^n_{j+\frac{3}{2}}) + h(m,m) - h(m, u^n_{j+\frac{1}{2}}) \right) \\ &\quad - \lambda^n_{j+\frac{1}{2}} \left(h(m, u^n_{j+\frac{3}{2}}) - h(m,m) + h(m, u^n_{j+\frac{1}{2}}) - h(u^n_{j-\frac{1}{2}}, u^n_{j+\frac{1}{2}}) \right) \ge 0 \end{aligned}$$

for all $j \in \Omega^1_\delta$ by applying the CFL condition (3.4) to the right-hand side (RHS) of the first line and by using the monotonicity property of the flux h in the second line. Similarly, we can also get that $u^{n+1}_{j+\frac{1}{2}} \le M$ for all $j \in \Omega^1_\delta$ under the hypotheses of the lemma.

Next, we show $m \le u^{n+1}_{j+\frac{1}{2}} \le M$ for all $j \in \Omega^2_\delta$. To do this, we introduce a small positive constant, $0 < \beta \ll 1$, and rewrite (2.3) in the following form:

$$(1+\beta) u^{n+1}_{j+\frac{1}{2}} = \beta u^n_{j+\frac{1}{2}} + \widetilde{u}^{n+1}_{j+\frac{1}{2}}, \tag{3.6}$$

where

$$\widetilde{u}_{j+\frac{1}{2}}^{n+1} = u_{j+\frac{1}{2}}^{n+1} - \beta\lambda_{j+\frac{1}{2}}^{n}\left(h(u_{j+\frac{1}{2}}^{n+1}, u_{j+\frac{3}{2}}^{n+1}) - h(u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1})\right). \tag{3.7}$$

Define $\widetilde{M} = \max_{j\in\Omega_\delta^2}\{u_{j+\frac{1}{2}}^{n+1}\}$ and $\widetilde{m} = \min_{j\in\Omega_\delta^2}\{u_{j+\frac{1}{2}}^{n+1}\}$. They are all finite, due to Lemma 3.2. Following the analysis for $j \in \Omega_\delta^1$, we may prove

$$\min\{\widetilde{m}, m\} \le \widetilde{u}_{j+\frac{1}{2}}^{n+1} \le \max\{\widetilde{M}, M\} \quad \forall j \in \Omega_\delta^2$$

under the CFL-type condition

$$\beta\lambda_{j+\frac{1}{2}}^{n} \max_{u,v,w,z\in\widetilde{\mathcal{A}}}\{|h_1(u,v)| + |h_2(w,z)|\} \le 1 \tag{3.8}$$

for all $j \in \Omega_\delta^2$, where $\widetilde{\mathcal{A}} := [\min\{\widetilde{m}, m\}, \max\{\widetilde{M}, M\}]$. Note that (3.8) is slightly more general than (3.4), but since β may be arbitrarily small there is essentially no further restriction to the time step. From (3.6), we get

$$\beta m + \min\{\widetilde{m}, m\} \le (1+\beta)u_{j+\frac{1}{2}}^{n+1} \le \beta M + \max\{\widetilde{M}, M\}, \quad j \in \Omega_\delta^2. \tag{3.9}$$

If $\widetilde{M} > M$, then we have from (3.9)

$$(1+\beta)\widetilde{M} \le \beta M + \widetilde{M}, \quad \widetilde{M} = \max_{j\in\Omega_\delta^2}\{u_{j+\frac{1}{2}}^{n+1}\}.$$

This leads to a contradiction. Therefore, $u_{j+\frac{1}{2}}^{n+1} \le M, j \in \mathbb{Z}$. Similarly, we can also get $u_{j+\frac{1}{2}}^{n+1} \ge m$ for all $j \in \mathbb{Z}$. The proof is completed. □

LEMMA 3.4. *Under the local CFL condition*

$$\lambda_j^n \max_{u,v,w,z\in\mathcal{A}}\{|h_1(u,v)| + |h_2(w,z)|\} \le 1 \tag{3.10}$$

for all $j \in \Omega_\delta^1$, *where* $\lambda_j^n = \max\{\lambda_{j+\frac{1}{2}}^n, \lambda_{j-\frac{1}{2}}^n\}$, *the DDM* (2.2) *and* (2.3) *is total variation diminishing; i.e.,*

$$TV(u^{n+1}) \le TV(u^n) \equiv \sum_{j\in\mathbb{Z}} |\Delta_+ u_{j-\frac{1}{2}}^n|, \quad n \ge 0, \tag{3.11}$$

where Δ_+ *denotes the forward difference operator in space.*

Proof. First, we rewrite (2.2) and (2.3) in an incremental form as follows:

$$u_{j+\frac{1}{2}}^{n+1} = u_{j+\frac{1}{2}}^n + C_{j+1}^n\Delta_+ u_{j+\frac{1}{2}}^n - D_j^n\Delta_+ u_{j-\frac{1}{2}}^n, \ j \in \Omega_\delta^1, \tag{3.12}$$

$$u_{j+\frac{1}{2}}^{n+1} = u_{j+\frac{1}{2}}^n + C_{j+1}^{n+1}\Delta_+ u_{j+\frac{1}{2}}^{n+1} - D_j^{n+1}\Delta_+ u_{j-\frac{1}{2}}^{n+1}, \ j \in \Omega_\delta^2, \tag{3.13}$$

where the incremental coefficients are defined by

$$C_{j+1}^\nu = \lambda_{j+\frac{1}{2}}^n \begin{cases} \frac{f(u_{j+\frac{1}{2}}^\nu) - h(u_{j+\frac{1}{2}}^\nu, u_{j+\frac{3}{2}}^\nu)}{\Delta_+ u_{j+\frac{1}{2}}^\nu}, & \Delta_+ u_{j+\frac{1}{2}}^\nu \ne 0, \\ h_2(u_{j+\frac{1}{2}}^\nu, u_{j+1/2}^\nu) & \text{otherwise}, \end{cases}$$

$$D_j^\nu = \lambda_{j+\frac{1}{2}}^n \begin{cases} \frac{f(u_{j+\frac{1}{2}}^\nu) - h(u_{j-\frac{1}{2}}^\nu, u_{j+\frac{1}{2}}^\nu)}{\Delta_+ u_{j-\frac{1}{2}}^\nu}, & \Delta_+ u_{j-\frac{1}{2}}^\nu \ne 0, \\ h_1(u_{j-\frac{1}{2}}^\nu, u_{j+\frac{1}{2}}^\nu) & \text{otherwise}, \end{cases}$$

where $\nu = n$ or $n+1$. Subtracting (3.12) (or (3.13)) at $j-\frac{1}{2}$ from (3.12) (or (3.13)) at $j+\frac{1}{2}$ for $j < j_1$ and $j > j_2$ (or $j_2 > j > j_1$) gives

$$\Delta_+ u^{n+1}_{j-\frac{1}{2}} = (1 - C^n_j - D^n_j)\Delta_+ u^n_{j-\frac{1}{2}} + C^n_{j+1}\Delta_+ u^n_{j+\frac{1}{2}} + D^n_{j-1}\Delta_+ u^n_{j-\frac{3}{2}}, \tag{3.14}$$

$$(1 + C^{n+1}_j + D^{n+1}_j)\Delta_+ u^{n+1}_{j-\frac{1}{2}} = \Delta_+ u^n_{j-\frac{1}{2}} + C^{n+1}_{j+1}\Delta_+ u^{n+1}_{j+\frac{1}{2}} + D^{n+1}_{j-1}\Delta_+ u^{n+1}_{j-\frac{3}{2}}. \tag{3.15}$$

We also subtract (3.12) at $j_1 - \frac{1}{2}$ from (3.13) at $j_1 + \frac{1}{2}$, and have

$$(1 + D^{n+1}_{j_1})\Delta_+ u^{n+1}_{j_1-\frac{1}{2}} = (1 - C^n_{j_1})\Delta_+ u^n_{j_1-\frac{1}{2}} + C^{n+1}_{j_1+1}\Delta_+ u^{n+1}_{j_1+\frac{1}{2}} + D^n_{j_1-1}\Delta_+ u^n_{j_1-\frac{3}{2}}. \tag{3.16}$$

Similarly, near the interface $x = x_{j_2}$, we have

$$(1 + C^{n+1}_{j_2})\Delta_+ u^{n+1}_{j_2-\frac{1}{2}} = (1 - D^n_{j_2})\Delta_+ u^n_{j_2-\frac{1}{2}} + C^n_{j_2+1}\Delta_+ u^n_{j_2+\frac{1}{2}} + D^{n+1}_{j_2-1}\Delta_+ u^{n+1}_{j_2-\frac{3}{2}}. \tag{3.17}$$

Under (3.10) and the monotonicity property of the numerical flux $h(u,v)$, the coefficients C^ν_j and D^ν_j in (3.14)–(3.17) are all nonnegative; i.e.,

$$C^\nu_j \geq 0, \quad D^\nu_j \geq 0 \quad \text{for all } j \in \mathbb{Z} \text{ and } \nu = n \text{ or } n+1,$$
$$C^n_j + D^n_j \leq \lambda^n_j \max\{|h_1| + |h_2|\} \leq 1.$$

Thus, taking the absolute value of (3.14)–(3.17) and using the triangle inequality, we get

$$\begin{aligned} |\Delta_+ u^{n+1}_{j-\frac{1}{2}}| \leq{}& \big(1 - (C^n_j + D^n_j)\big)|\Delta_+ u^n_{j-\frac{1}{2}}| + C^n_{j+1}|\Delta_+ u^n_{j+\frac{1}{2}}| \\ &+ D^n_{j-1}|\Delta_+ u^n_{j-\frac{3}{2}}|, \ j < j_1 \text{ or } j > j_2, \end{aligned} \tag{3.18}$$

$$\begin{aligned} (1 + C^{n+1}_j + D^{n+1}_j)|\Delta_+ u^{n+1}_{j-\frac{1}{2}}| \leq{}& |\Delta_+ u^n_{j-\frac{1}{2}}| + C^{n+1}_{j+1}|\Delta_+ u^{n+1}_{j+\frac{1}{2}}| \\ &+ D^{n+1}_{j-1}|\Delta_+ u^{n+1}_{j-\frac{3}{2}}|, \ j_1 < j < j_2, \end{aligned} \tag{3.19}$$

$$\begin{aligned} (1 + D^{n+1}_{j_1})|\Delta_+ u^{n+1}_{j_1-\frac{1}{2}}| \leq{}& (1 - C^n_{j_1})|\Delta_+ u^n_{j_1-\frac{1}{2}}| \\ &+ C^{n+1}_{j_1+1}|\Delta_+ u^{n+1}_{j_1+\frac{1}{2}}| + D^n_{j_1-1}|\Delta_+ u^n_{j_1-\frac{3}{2}}|, \end{aligned} \tag{3.20}$$

$$\begin{aligned} (1 + C^{n+1}_{j_2})|\Delta_+ u^{n+1}_{j_2-\frac{1}{2}}| \leq{}& (1 - D^n_{j_2})|\Delta_+ u^n_{j_2-\frac{1}{2}}| + C^n_{j_2+1}|\Delta_+ u^n_{j_2+\frac{1}{2}}| \\ &+ D^{n+1}_{j_2-1}|\Delta_+ u^{n+1}_{j_2-\frac{3}{2}}|. \end{aligned} \tag{3.21}$$

Summing (3.18) from $j = -\infty$ to $j_1 - 1$ and from $j = j_2 + 1$ to ∞, (3.19) from $j_1 + 1$ to $j_2 - 1$, (3.20), and (3.21), we get by shifting indices

$$TV(u^{n+1}) \leq TV(u^n).$$

This completes the proof. ∎

Remark 3.2. Let us take the following numerical flux [32]:

$$h(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}}) = \frac{\Delta x_{j+\frac{1}{2}} f(u_{j-\frac{1}{2}}) + \Delta x_{j-\frac{1}{2}} f(u_{j+\frac{1}{2}})}{\Delta x_{j+\frac{1}{2}} + \Delta x_{j-\frac{1}{2}}} - Q_j \frac{u_{j+\frac{1}{2}} - u_{j-\frac{1}{2}}}{\Delta x_{j+\frac{1}{2}} + \Delta x_{j-\frac{1}{2}}}, \tag{3.22}$$

where the numerical viscosity coefficient $Q_j = Q(u_{j+\frac{1}{2}}, u_{j-\frac{1}{2}}; \Delta x_{j+\frac{1}{2}}, \Delta x_{j-\frac{1}{2}})$ is chosen such that $h(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}})$ is monotone. As an example, we define Q_j by

$$Q_j = \max_u \{|f'(u)|\} \max\{\Delta x_{j+\frac{1}{2}}, \Delta x_{j-\frac{1}{2}}\}. \tag{3.23}$$

For this special numerical flux $h(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}})$ given by (3.22) and (3.23), we have

$$C_j + D_j = \frac{\Delta t_n}{\Delta x_{j-\frac{1}{2}} \Delta x_{j+\frac{1}{2}}} Q_j = \frac{\Delta t_n}{\min\{\Delta x_{j-\frac{1}{2}}, \Delta x_{j+\frac{1}{2}}\}} \max_u\{|f'(u)|\}.$$

Thus, the TV-stability condition now becomes

$$\frac{\Delta t_n}{\min\{\Delta x_{j-\frac{1}{2}}, \Delta x_{j+\frac{1}{2}}\}} \max_u\{|f'(u)|\} \le 1 \quad \forall j \in \Omega_\delta^1.$$

3.2. L^1-continuity in time.

LEMMA 3.5. *Let $\mu > 1$ and assume $C_j^{n+1} + D_j^{n+1} \le \mu$ for all $j \in \Omega_2^\delta$, $n \in \mathbb{Z}$. Under the hypothesis of Lemma* 3.4*, we have*

$$||u^m - u^n||_{L^1(\mathbb{R})} \equiv \sum_{j\in\mathbb{Z}} |u_{j+\frac{1}{2}}^m - u_{j+\frac{1}{2}}^n| \Delta x_{j+\frac{1}{2}} \le \frac{1+\mu}{\lambda}(t_m - t_n) TV(u_0), \tag{3.24}$$

where $\frac{1}{\lambda} = \max_{j,n}\{\frac{1}{\lambda_{j+\frac{1}{2}}^n}\}$.

Note that due to the larger time step in a part of the domain, i.e., Ω_δ^2, the term $C_j^{n+1} + D_j^{n+1}$ is not necessarily bounded by 1. For this purpose we introduce the constant $\mu \ge 1$ as in the formulation of the lemma. This constant is assumed to be chosen independently of spatial and time step sizes.

Proof. From (3.12)–(3.13), we have

$$(u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n)\Delta x_{j+\frac{1}{2}} = \frac{\Delta t_n}{\lambda_{j+\frac{1}{2}}^n}(C_{j+1}^n \Delta_+ u_{j+\frac{1}{2}}^n - D_j^n \Delta_+ u_{j-\frac{1}{2}}^n),\ j \in \Omega_\delta^1, \tag{3.25}$$

$$(u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n)\Delta x_{j+\frac{1}{2}} = \frac{\Delta t_n}{\lambda_{j+\frac{1}{2}}^n}(C_{j+1}^{n+1} \Delta_+ u_{j+\frac{1}{2}}^{n+1} - D_j^{n+1} \Delta_+ u_{j-\frac{1}{2}}^{n+1}),\ j \in \Omega_\delta^2. \tag{3.26}$$

Taking the absolute value of these equations and using the triangle inequality, we get

$$|u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n| \Delta x_{j+\frac{1}{2}} \le \frac{\Delta t_n}{\lambda_{j+\frac{1}{2}}^n}(C_{j+1}^n |\Delta_+ u_{j+\frac{1}{2}}^n| + D_j^n |\Delta_+ u_{j-\frac{1}{2}}^n|),\ j \in \Omega_\delta^1, \tag{3.27}$$

$$|u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n| \Delta x_{j+\frac{1}{2}} \le \frac{\Delta t_n}{\lambda_{j+\frac{1}{2}}^n}(C_{j+1}^{n+1} |\Delta_+ u_{j+\frac{1}{2}}^{n+1}| + D_j^{n+1} |\Delta_+ u_{j-\frac{1}{2}}^{n+1}|),\ j \in \Omega_\delta^2. \tag{3.28}$$

Summing (3.27) from $j = -\infty$ to $j_1 - 1$ and from j_2 to ∞, and (3.28) from j_1 to $j_2 - 1$, we get by shifting indices

$$\sum_{j=-\infty}^{\infty} |u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n| \Delta x_{j+\frac{1}{2}} \le \frac{\Delta t_n}{\lambda}\left(TV(u^n) + \mu TV(u^{n+1})\right) \le \frac{(1+\mu)\Delta t_n}{\lambda} TV(u^n). \tag{3.29}$$

Here the conclusion of Lemma 3.4 has been used. From this inequality, we may complete the proof. ∎

4. A cell entropy inequality. This section concerns the study of a cell entropy inequality satisfied by the solutions of the DDM (2.2) and (2.3). In our analysis we will make use of the Kruzkov entropy pairs (U,F): $U(u;k)=|u-k|$, $F(u;k)=\operatorname{sign}(u-k)\big(f(u)-f(k)\big)$ for any $k\in\mathbb{R}$, and the notation "$\vee$" and "$\wedge$" defined by $a\vee b=\max\{a,b\}$, and $a\wedge b=\min\{a,b\}$.

LEMMA 4.1. *The solutions of the scheme* (2.2) *satisfy the cell entropy inequality*

$$U(u^{n+1}_{j+\frac12};k)-U(u^n_{j+\frac12};k)+\lambda^n_{j+\frac12}\Delta_+H(u^n_{j-\frac12},u^n_{j+\frac12};k)\le 0,\quad j\in\Omega^1_\delta, \tag{4.1}$$

under the CFL condition (3.4)*, while the solutions of the scheme* (2.3) *satisfy*

$$U(u^{n+1}_{j+\frac12};k)-U(u^n_{j+\frac12};k)+\lambda^n_{j+\frac12}\Delta_+H(u^{n+1}_{j-\frac12},u^{n+1}_{j+\frac12};k)\le 0,\quad j\in\Omega^2_\delta, \tag{4.2}$$

where the numerical entropy flux is defined by $H(u^\nu_{j-\frac12},u^\nu_{j+\frac12};k)=h(u^\nu_{j-\frac12}\vee k,u^\nu_{j+\frac12}\vee k)-h(u^\nu_{j-\frac12}\wedge k,u^\nu_{j+\frac12}\wedge k)$*, which satisfies the consistency condition*

$$H(u,u;k)=F(u;k).$$

Proof. Mimicking the proof of Crandall and Majda given in [6], we can deduce the cell entropy inequality (4.1). Rewrite the difference equation (2.2) as

$$u^{n+1}_{j+\frac12}=G(u^n_{j-\frac12},u^n_{j+\frac12},u^n_{j+\frac32}).$$

Because $G(u,v,w)$ satisfies $k=G(k,k,k)$ and is monotonously increasing with respect to its arguments, under the CFL condition (3.4), we find that

$$\begin{aligned}
u^{n+1}_{j+\frac12}\vee k&=G(u^n_{j-\frac12},u^n_{j+\frac12},u^n_{j+\frac32})\vee G(k,k,k)\le G(k\vee u^n_{j-\frac12},k\vee u^n_{j+\frac12},k\vee u^n_{j+\frac32}),\\
-u^{n+1}_{j+\frac12}\wedge k&=-G(u^n_{j-\frac12},u^n_{j+\frac12},u^n_{j+\frac32})\wedge G(k,k,k)\le -G(k\wedge u^n_{j-\frac12},k\wedge u^n_{j+\frac12},k\wedge u^n_{j+\frac32}).
\end{aligned}$$

Adding these two inequalities gives (4.1), i.e.,

$$\begin{aligned}
|u^{n+1}_{j+\frac12}-k|&\le G(k\vee u^n_{j-\frac12},k\vee u^n_{j+\frac12},k\vee u^n_{j+\frac32})-G(k\wedge u^n_{j-\frac12},k\wedge u^n_{j+\frac12},k\wedge u^n_{j+\frac32})\\
&=|u^n_{j+\frac12}-k|-\beta\lambda^n_{j+\frac12}\Delta_+H(u^n_{j-\frac12},u^n_{j+\frac12};k).
\end{aligned}$$

If using the CFL-type condition (3.8), we can get the inequality (4.2) from the difference equations (3.7) and (3.6) in a similar way. In the following we begin to prove the inequality (4.2) by a case-by-case procedure, without using the condition (3.8).

Case 1. $u^{n+1}_{j+\frac12},u^{n+1}_{j-\frac12},u^{n+1}_{j+\frac32}\ge k$. From the scheme (2.3), we have

$$|u^{n+1}_{j+\frac12}-k|-(u^n_{j+\frac12}-k)+\lambda^n_{j+\frac12}\Delta_+\Big(h(u^{n+1}_{j-\frac12},u^{n+1}_{j+\frac12})-h(k,k)\Big)=0$$

and get (4.2); i.e.,

$$|u^{n+1}_{j+\frac12}-k|-|u^n_{j+\frac12}-k|+\lambda^n_{j+\frac12}\Delta_+\Big(h(u^{n+1}_{j-\frac12},u^{n+1}_{j+\frac12})-h(k,k)\Big)\le 0.$$

Case 2. $u^{n+1}_{j+\frac12},u^{n+1}_{j-\frac12},u^{n+1}_{j+\frac32}\le k$. From the scheme (2.3), we have

$$|k-u^{n+1}_{j+\frac12}|-(k-u^n_{j+\frac12})+\lambda^n_{j+\frac12}\Delta_+\Big(h(k,k)-h(u^{n+1}_{j-\frac12},u^{n+1}_{j+\frac12})\Big)=0$$

and get (4.2); i.e.,

$$|u^{n+1}_{j+\frac{1}{2}}-k|-|u^{n}_{j+\frac{1}{2}}-k|+\lambda^n_{j+\frac{1}{2}}\Delta_+\Big(h(k,k)-h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})\Big)\le 0.$$

Case 3. $u^{n+1}_{j+\frac{1}{2}},u^{n+1}_{j+\frac{3}{2}}\ge k\ge u^{n+1}_{j-\frac{1}{2}}$. Substituting (2.3) into the left-hand side (LHS) of (4.2), we have

$$\begin{aligned}\text{LHS}(4.2)&=(u^n_{j+\frac{1}{2}}-k)-|u^n_{j+\frac{1}{2}}-k|-\lambda^n_{j+\frac{1}{2}}\Delta_+h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})\\&\quad+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j+\frac{1}{2}},u^{n+1}_{j+\frac{3}{2}})-h(k,k)-h(k,u^{n+1}_{j+\frac{1}{2}})+h(u^{n+1}_{j-\frac{1}{2}},k)\Big)\\&=(u^n_{j+\frac{1}{2}}-k)-|u^n_{j+\frac{1}{2}}-k|+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})-h(k,u^{n+1}_{j+\frac{1}{2}})\\&\quad+h(u^{n+1}_{j-\frac{1}{2}},k)-h(k,k)\Big)\le 0,\end{aligned}$$

where we have used the monotonicity property of the numerical flux $h(u,v)$.

Case 4. $u^{n+1}_{j+\frac{1}{2}},u^{n+1}_{j-\frac{1}{2}}\ge k\ge u^{n+1}_{j+\frac{3}{2}}$. This case is similar to Case 3. We have

$$\begin{aligned}\text{LHS}(4.2)&=(u^n_{j+\frac{1}{2}}-k)-|u^n_{j+\frac{1}{2}}-k|-\lambda^n_{j+\frac{1}{2}}\Delta_+h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})\\&\quad+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j+\frac{1}{2}},k)-h(k,u^{n+1}_{j+\frac{3}{2}})-h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})+h(k,k)\Big)\\&=(u^n_{j+\frac{1}{2}}-k)-|u^n_{j+\frac{1}{2}}-k|+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j+\frac{1}{2}},k)-h(u^{n+1}_{j+\frac{1}{2}},u^{n+1}_{j+\frac{3}{2}})\\&\quad+h(k,k)-h(k,u^{n+1}_{j+\frac{3}{2}})\Big)\le 0.\end{aligned}$$

Case 5. $u^{n+1}_{j+\frac{3}{2}},u^{n+1}_{j-\frac{1}{2}}\ge k\ge u^{n+1}_{j+\frac{1}{2}}$. Similarly, we have

$$\begin{aligned}\text{LHS}(4.2)&=(k-u^n_{j+\frac{1}{2}})-|u^n_{j+\frac{1}{2}}-k|+\lambda^n_{j+\frac{1}{2}}\Delta_+h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})\\&\quad+\lambda^n_{j+\frac{1}{2}}\Big(h(k,u^{n+1}_{j+\frac{3}{2}})-h(u^{n+1}_{j+\frac{1}{2}},k)-h(u^{n+1}_{j-\frac{1}{2}},k)+h(k,u^{n+1}_{j+\frac{1}{2}})\Big)\\&=(k-u^n_{j+\frac{1}{2}})-|u^n_{j+\frac{1}{2}}-k|+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j+\frac{1}{2}},u^{n+1}_{j+\frac{3}{2}})-h(u^{n+1}_{j+\frac{1}{2}},k)\\&\quad+h(k,u^{n+1}_{j+\frac{1}{2}})-h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})+h(k,u^{n+1}_{j+\frac{3}{2}})-h(k,k)\\&\quad+h(k,k)-h(u^{n+1}_{j-\frac{1}{2}},k)\Big)\le 0.\end{aligned}$$

Case 6. $u^{n+1}_{j+\frac{1}{2}}\ge k\ge u^{n+1}_{j+\frac{3}{2}},u^{n+1}_{j-\frac{1}{2}}$. This case is similar to Case 5. We have

$$\begin{aligned}\text{LHS}(4.2)&=(u^n_{j+\frac{1}{2}}-k)-|u^n_{j+\frac{1}{2}}-k|-\lambda^n_{j+\frac{1}{2}}\Delta_+h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})\\&\quad+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j+\frac{1}{2}},k)-h(k,u^{n+1}_{j+\frac{3}{2}})-h(k,u^{n+1}_{j+\frac{1}{2}})+h(u^{n+1}_{j-\frac{1}{2}},k)\Big)\\&=(u^n_{j+\frac{1}{2}}-k)-|u^n_{j+\frac{1}{2}}-k|+\lambda^n_{j+\frac{1}{2}}\Big(h(u^{n+1}_{j+\frac{1}{2}},k)-h(u^{n+1}_{j+\frac{1}{2}},u^{n+1}_{j+\frac{3}{2}})\\&\quad+h(u^{n+1}_{j-\frac{1}{2}},u^{n+1}_{j+\frac{1}{2}})-h(k,u^{n+1}_{j+\frac{1}{2}})+h(u^{n+1}_{j-\frac{1}{2}},k)-h(k,k)\\&\quad+h(k,k)-h(k,u^{n+1}_{j+\frac{3}{2}})\Big)\le 0.\end{aligned}$$

Case 7. $u_{j-\frac{1}{2}}^{n+1} \geq k \geq u_{j+\frac{3}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}$. We have

$$\begin{aligned}\text{LHS}(4.2) &= (k-u_{j+\frac{1}{2}}^n) - |u_{j+\frac{1}{2}}^n - k| + \lambda_{j+\frac{1}{2}}^n \Delta_+ h(u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}) \\ &\quad + \lambda_{j+\frac{1}{2}}^n \Big(h(k,k) - h(u_{j+\frac{1}{2}}^{n+1}, u_{j+\frac{3}{2}}^{n+1}) - h(u_{j-\frac{1}{2}}^{n+1}, k) + h(k, u_{j+\frac{1}{2}}^{n+1}) \Big) \\ &= (k-u_{j+\frac{1}{2}}^n) - |u_{j+\frac{1}{2}}^n - k| + \lambda_{j+\frac{1}{2}}^n \Big(h(k, u_{j+\frac{1}{2}}^{n+1}) - h(u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}) \\ &\quad + h(k,k) - h(u_{j-\frac{1}{2}}^{n+1}, k) \Big) \leq 0.\end{aligned}$$

Case 8. $u_{j+\frac{3}{2}}^{n+1} \geq k \geq u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}$. It is similar to Case 7. We have

$$\begin{aligned}\text{LHS}(4.2) &= (k-u_{j+\frac{1}{2}}^n) - |u_{j+\frac{1}{2}}^n - k| + \lambda_{j+\frac{1}{2}}^n \Delta_+ h(u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}) \\ &\quad + \lambda_{j+\frac{1}{2}}^n \Big(h(k, u_{j+\frac{3}{2}}^{n+1}) - h(u_{j+\frac{1}{2}}^{n+1}, k) - h(k,k) + h(u_{j-\frac{1}{2}}^{n+1}, u_{j+\frac{1}{2}}^{n+1}) \Big) \\ &= (k-u_{j+\frac{1}{2}}^n) - |u_{j+\frac{1}{2}}^n - k| + \lambda_{j+\frac{1}{2}}^n \Big(h(u_{j+\frac{1}{2}}^{n+1}, u_{j+\frac{3}{2}}^{n+1}) - h(u_{j+\frac{1}{2}}^{n+1}, k) \\ &\quad h(k, u_{j+\frac{3}{2}}^{n+1}) - h(k,k) \Big) \leq 0.\end{aligned}$$

The proof is completed. ∎

5. Convergence of the algorithm. We focus in this section on the study of the convergence of the DDM (2.2) and (2.3) introduced in section 2. Treating the nonconservative terms carefully is the key.

Let $L^1_{loc}(\Omega)$ denote the space of all functions u that are integrable on compact subsets of Ω, and define the step function u_δ as follows:

$$u_\delta(x,t) = u_{j+\frac{1}{2}}^n \quad \text{for } (x,t) \in (x_j, x_{j+1}) \times [t_n, t_{n+1}) \tag{5.1}$$

if $j \in \Omega_\delta^1 \cup \Omega_\delta^2$.

Using Lemmas 3.3–3.5, we can prove the following theorem.

THEOREM 5.1. *If $u_0(x) \in L^\infty(\mathbb{R}) \cap L^1(\mathbb{R}) \cap BV(\mathbb{R})$, then the approximate family of solutions u_δ constructed by the DDM algorithm converges as δ tends to zero. The limit is a function u in $L^1_{loc}(\mathbb{R} \times \mathbb{R}^+)$ which is a weak solution of* (1.1) *with initial data* (1.2).

Proof. From Lemmas 3.3–3.5, the sequence $\{u_\delta\}$ is bounded in $L^\infty(\mathbb{R}) \cap L^1(\mathbb{R}) \cap BV(\mathbb{R})$. Then by Helly's compactness theorem, we can extract a subsequence still labeled $\{u_\delta\}$ which converges towards a function u in $L^1_{loc}(\mathbb{R}\times\mathbb{R}^+)$. In the following we shall prove that the function u is a weak solution of (1.1) and (1.2). We mimic the proof of the Lax–Wendroff theorem. For the sake of clarity throughout the proof, we still use $u_{j+\frac{1}{2}}^\nu$ to denote $u_\delta(x_{j+\frac{1}{2}}, t_\nu)$, and h_j^ν to denote $h\big(u_\delta(x_{j-\frac{1}{2}}, t_\nu), u_\delta(x_{j+\frac{1}{2}}, t_\nu)\big)$, where $\nu = n$ or $n+1$. Multiplying (2.2) and (2.3) by a smooth test function $\phi(x_{j+\frac{1}{2}}, t_n) \in C_0^\infty(\mathbb{R}^2)$ and summing it with respect to n and j gives

$$\begin{aligned}&\sum_{j=-\infty}^{\infty} \left\{ \sum_{n=0}^{\infty} \big(u_{j+\frac{1}{2}}^{n+1} - u_{j+\frac{1}{2}}^n\big) \phi(x_{j+\frac{1}{2}}, t_n) \right\} \Delta x_{j+\frac{1}{2}} + \sum_{n=0}^{\infty} \left\{ \sum_{j=-\infty}^{j_1-1} (h_{j+1}^n - h_j^n) \phi(x_{j+\frac{1}{2}}, t_n) \right. \\ &\left. + \sum_{j=j_1}^{j_2-1} (h_{j+1}^{n+1} - h_j^{n+1}) \phi(x_{j+\frac{1}{2}}, t_n) + \sum_{j=j_2}^{\infty} (h_{j+1}^n - h_j^n) \phi(x_{j+\frac{1}{2}}, t_n) \right\} \Delta t_n = 0.\end{aligned} \tag{5.2}$$

Now, we use I_0 to denote the first term at the LHS of (5.2), and I_i, $i = 1, 2, 3$, to denote the ith term in braces in the second term at the LHS of (5.2). If we use "summation by parts," then the term I_0 becomes

$$I_0 = \sum_{j=-\infty}^{\infty} \left\{ -u^0_{j+\frac{1}{2}} \phi(x_{j+\frac{1}{2}}, t_0) - \sum_{n=1}^{\infty} u^n_{j+\frac{1}{2}} \Delta^t_+ \phi(x_{j+\frac{1}{2}}, t_{n-1}) \right\} \Delta x_{j+\frac{1}{2}}, \tag{5.3}$$

where Δ^t_+ denotes the forward difference operator in time. Similarly, we also have

$$I_1 = h^n_{j_1} \phi(x_{j_1-\frac{1}{2}}, t_n) - \sum_{j=-\infty}^{j_1-1} h^n_j \Delta_+ \phi(x_{j-\frac{1}{2}}, t_n), \tag{5.4}$$

$$I_2 = \left\{ h^{n+1}_{j_2} \phi(x_{j_2-\frac{1}{2}}, t_n) - h^{n+1}_{j_1} \phi(x_{j_1-\frac{1}{2}}, t_n) \right\} - \sum_{j=j_1}^{j_2-1} h^{n+1}_j \Delta_+ \phi(x_{j-\frac{1}{2}}, t_n), \tag{5.5}$$

where the term $h^{n+1}_{j_1} \phi(x_{j_1-\frac{1}{2}}, t_n)$ has been added and subtracted above, and

$$I_3 = - h^n_{j_2} \phi(x_{j_2-\frac{1}{2}}, t_n) - \sum_{j=j_2}^{\infty} h^n_j \Delta_+ \phi(x_{j-\frac{1}{2}}, t_n). \tag{5.6}$$

Here we have similarly added and subtracted the term $h^n_{j_2} \phi(x_{j_2-\frac{1}{2}}, t_n)$. Hence, (5.2) becomes, after using summation by parts as above,

$$\begin{aligned} 0 = & \sum_{j=-\infty}^{\infty} \left\{ -u^0_{j+\frac{1}{2}} \phi(x_{j+\frac{1}{2}}, t_0) - \sum_{n=1}^{\infty} u^n_{j+\frac{1}{2}} \Delta^t_+ \phi(x_{j+\frac{1}{2}}, t_{n-1}) \right\} \Delta x_{j+\frac{1}{2}} \\ & - \sum_{n=0}^{\infty} \left\{ \sum_{j=-\infty}^{j_1-1} h^n_j \Delta_+ \phi(x_{j-\frac{1}{2}}, t_n) + \sum_{j=j_1}^{j_2-1} h^{n+1}_j \Delta_+ \phi(x_{j-\frac{1}{2}}, t_n) + \sum_{j=j_2}^{\infty} h^n_j \Delta_+ \phi(x_{j-\frac{1}{2}}, t_n) \right\} \\ & + \sum_{n=0}^{\infty} \left\{ \phi(x_{j_2-\frac{1}{2}}, t_n) \Delta t_n \Delta^t_+ h^n_{j_2} - \phi(x_{j_1-\frac{1}{2}}, t_n) \Delta t_n \Delta^t_+ h^n_{j_1} \right\}. \end{aligned} \tag{5.7}$$

Again using summation by parts, we can rewrite the third term at the RHS of (5.7) as follows:

$$\begin{aligned} I_4 := & \phi(x_{j_1-\frac{1}{2}}, t_0) \Delta t_0 - \phi(x_{j_2-\frac{1}{2}}, t_0) \Delta t_0 - \sum_{n=1}^{\infty} \left\{ h^n_{j_2} \Delta^t_+ (\phi(x_{j_2-\frac{1}{2}}, t_{n-1}) \Delta t_{n-1}) \right. \\ & \left. - h^n_{j_1} \Delta^t_+ (\phi(x_{j_1-\frac{1}{2}}, t_{n-1}) \Delta t_{n-1}) \right\}. \end{aligned} \tag{5.8}$$

Under the assumption that $\Delta t_n = \Delta t_{n-1} + \mathcal{O}(\delta^{1+\alpha})$, we have

$$\begin{aligned} & \sum_{n=1}^{\infty} h^n_{\hat{j}} \Delta^t_+ \left(\phi(x_{\hat{j}-\frac{1}{2}}, t_{n-1}) \Delta t_{n-1} \right) \\ & \quad = \sum_{n=1}^{\infty} h^n_{\hat{j}} \left\{ \phi(x_{\hat{j}-\frac{1}{2}}, t_n) \left(\Delta t_{n-1} + \mathcal{O}(\delta^{1+\alpha}) \right) - \phi(x_{\hat{j}-\frac{1}{2}}, t_{n-1}) \Delta t_{n-1} \right\} \\ & \quad = \sum_{n=1}^{\infty} h^n_{\hat{j}} \left\{ \Delta t^2_{n-1} \phi_t(x_{\hat{j}-\frac{1}{2}}, \eta_{n-1}) + \phi(x_{\hat{j}-\frac{1}{2}}, t) \mathcal{O}(\delta^{1+\alpha}) \right\} \longrightarrow 0 \end{aligned} \tag{5.9}$$

as $\delta \longrightarrow 0$, where $\hat{j} = j_2$ or j_1, and $\eta_{n-1} \in [t_{n-1}, t_n]$. Therefore, it follows that the term $\lim_{\delta \to 0} I_4 = 0$ uniformly.

Taking into account the above results, we prove that (5.7) gives

$$-\int_{\mathbb{R}} u_0(x)\phi(x,0)\,dx = \int_{\mathbb{R}\times\mathbb{R}^+} \left(u\frac{\partial \phi}{\partial t} + f(u)\frac{\partial \phi}{\partial x}\right)\,dxdt, \tag{5.10}$$

as δ tends to zero. The proof is completed. □

THEOREM 5.2. *Let u be a weak solution to the initial value problem* (1.1) *and* (1.2). *Further, let u_δ be a sequence of approximate solutions constructed by* (2.2) *and* (2.3). *Suppose that u is the limit in $L^1_{loc}(\mathbb{R}\times\mathbb{R}^+)$ of the sequence of the approximate solutions u_δ for $\delta \to 0$. Then u is the unique entropy solution to* (1.1) *and* (1.2) *satisfying the entropy condition*

$$-\int_{\mathbb{R}\times\mathbb{R}^+} \left(|u-k|\frac{\partial \phi}{\partial t} + sign(u-k)(f(u)-f(k))\frac{\partial \phi}{\partial x}\right)\,dxdt \le 0 \tag{5.11}$$

for all nonnegative test function $\phi \in C_0^\infty(\mathbb{R}\times\mathbb{R}^+)$ and all real numbers k.

The proof of this theorem is analogous to that of the previous procedure in the proof of the last theorem and will be omitted here.

6. Numerical experiments. In this section we present two numerical experiments to validate the previous results of the domain decomposition method given in section 2.

Example 1. We use our algorithm (2.2) and (2.3) to solve the following initial value problem of the inviscid Burgers equation:

$$\frac{\partial u}{\partial t} + \frac{\partial\left(\frac{1}{2}u^2\right)}{\partial x} = 0, \tag{6.1}$$

$$u(x,0) = \begin{cases} 1 & \text{if } -1 \le x \le 0, \\ 0.1 & \text{otherwise.} \end{cases} \tag{6.2}$$

The numerical flux is taken as $h(u_{j-\frac{1}{2}}, u_{j+\frac{1}{2}}) = \frac{1}{2}(u_{j-\frac{1}{2}})^2$, because $u(x,t)$ is nonnegative for all $x \in \mathbb{R}$ and $t \ge 0$. The computational domain $[-2,2]$ is divided into two subdomains: $\Omega^1 = \{x|x < -1\} \cup \{x|x > 1\}$ and $\Omega^2 = \{x| -1 \le x \le 1\}$. Each subdomain is again partitioned into small cells with length $\Delta x = 0.01$. It means that the global domain is divided into 400 cells. In Figure 1, we show the numerical solutions at $t = 1$ obtained by two single domain algorithms: the fully explicit upwind scheme and the fully implicit upwind scheme, respectively. The time step size $\Delta t = 9.5 \times 10^{-3}$ has been used. The solid line is the exact solution obtained by the method of characteristics and the jump condition.

Figure 2 gives the numerical solutions at $t = 1$ calculated by our multidomain algorithm (2.2) and (2.3) with two different time step sizes $\Delta t = 9.5 \times 10^{-3}$ and $\Delta t = 9.5 \times 10^{-2}$, respectively. The results show that the correct location of discontinuities has been obtained; it is in accordance with the exact solution. The numerical solutions at $t = 1$ and 2 shown in Figure 3 are calculated by our multidomain algorithm (2.2) and (2.3) with the time step sizes $\Delta t = 9.5 \times 10^{-3}$ as well as $\Omega^1 = \{x|x < -1\} \cup \{x|x > 0\}$ and $\Omega^2 = \{x| -1 \le x \le 0\}$. The discontinuities have also been resolved correctly. In this case, the shock wave is passing through the interface of Ω^1 and Ω^2.

However, from Figures 1 and 2, we observe that the solutions calculated by using the fully implicit upwind scheme are more dissipative than those of the explicit upwind

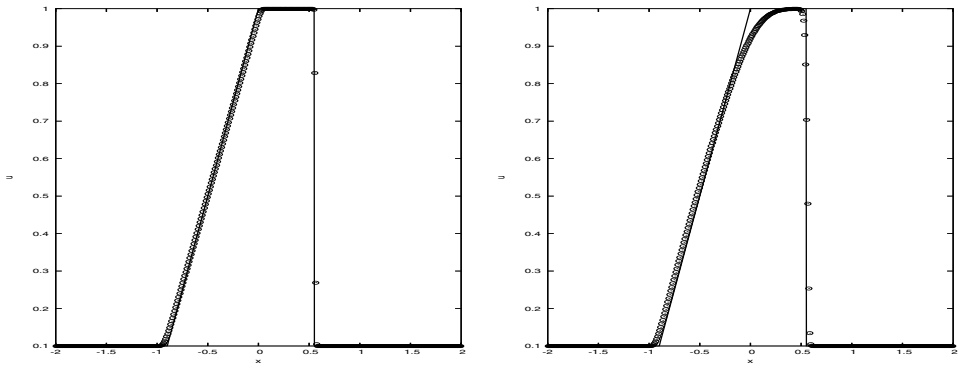


FIG. 1. *Example* 1: *Comparison of the single-domain solutions ("circle") with the exact solutions. Left: explicit upwind scheme; right: implicit upwind scheme.*

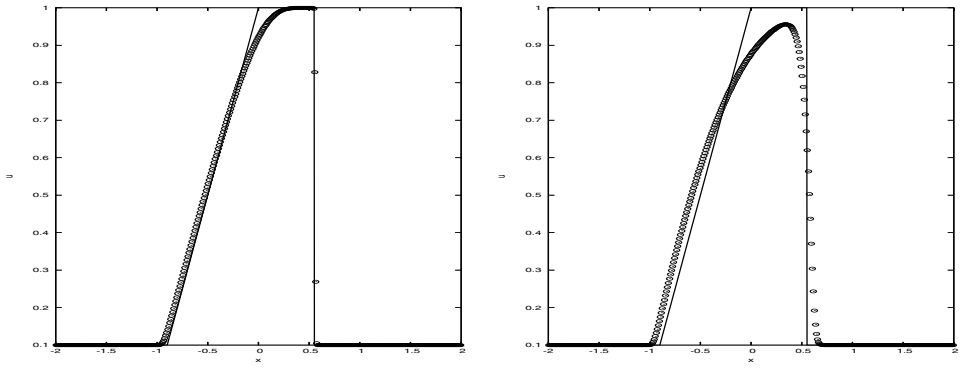


FIG. 2. *Example* 1: *Comparison of multidomain solutions ("circle") with the exact solutions. Left:* $\Delta t = 9.5 \times 10^{-3}$*; right:* $\Delta t = 9.5 \times 10^{-2}$.

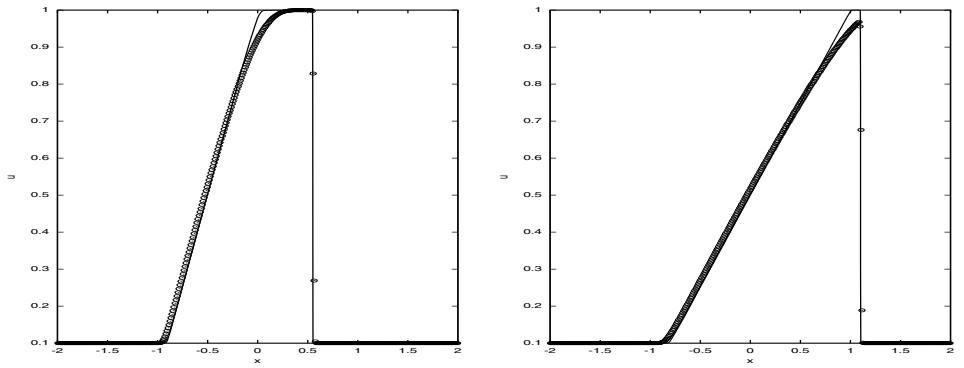


FIG. 3. *Example* 1: *Comparison of multidomain solutions ("circle") with the exact solutions. Left:* $t = 1$*; right:* $t = 2$.

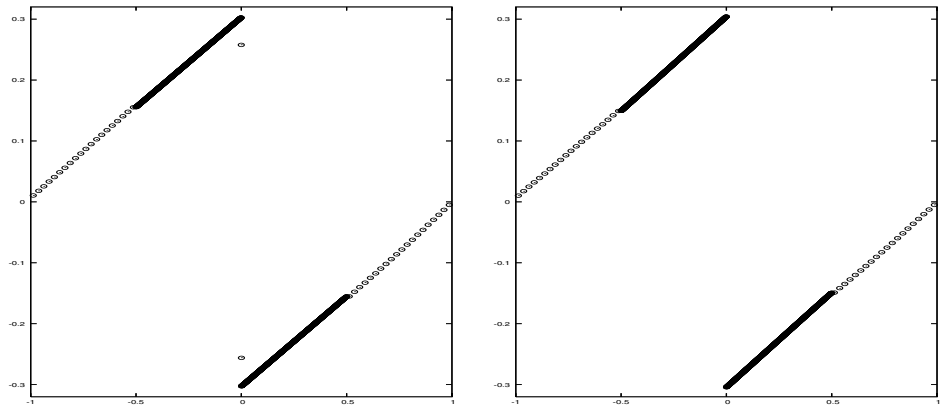


FIG. 4. *Example* 2: *the computed solutions. Left: the fully explicit upwind scheme; right: the DDM.*

scheme. If the time step size is taken larger, resolution of numerical solutions will be decreased; see, e.g., Figure 2. Therefore, it will be meaningful to develop high resolution DDMs for hyperbolic conservation laws.

Even though the numerical solutions obtained by using the DDM (2.2) and (2.3) are less accurate in the above example, this DDM is still very attractive in relaxing the restriction of the time step size and improving the computational efficiency of the DDM algorithm when we solve the steady state problem. In the following, we give a simple example to demonstrate this.

Example 2. This example is to solve the following IBVP of the Burgers equation subject to the periodic data [19]:

$$u(x,0) = \sin(\pi(x+1)), \quad x \in [-1,1). \tag{6.3}$$

The numerical flux $h(u,v)$ is taken as the traditional upwind flux, but an entropy fix should be used to avoid the sonic point glitch [31]. The computational domain $[-1,1]$ is first divided into two subdomains: $\Omega^1 = \{x|x < -0.5\} \cup \{x|x > 0.5\}$ and $\Omega^2 = \{x| -0.5 \le x \le 0.5\}$. The domain Ω^2 is chosen in the region where $f'(u) = u$ is largest and leads to the most severe time step restriction. On the other hand, there the shock appears, and we expect to need a fine spatial resolution for accuracy. The subdomain Ω^1 is then partitioned into 20 large cells, and Ω^2 is divided into 2000 small cells. We now give a comparison of our partitioning method on the two domains with an explicit computation using an explicit method.

The computed solutions at $t = 3$ are shown in Figure 4. We see that the steady state solution of the DDM is slightly more accurate than the explicit upwind scheme. Here we used a fine mesh as in Ω^2 on the whole domain for the explicit scheme with the usual CFL condition. The CFL number used was 0.9. It spent the CPU time of 1.06s. When we use the DDM, the time step size is determined only by the explicit part, and a CPU time of 0.06s is needed.

7. Concluding remarks. The use of DDMs for conservation laws is an emerging field. The potential payoff will lie in a reduction of computing time possibly in conjunction with parallelization and the use of different solvers in different regions of the computational domain. The second author is pursuing the latter approach for

advection-diffusion equations and reaction-diffusion systems in parallel work, e.g., in [15].

A sound analytical foundation of the DDMs is needed. For conservation laws it is currently very difficult to obtain numerical analysis for systems of equations, even impossible for real multidimensional systems and for methods that are higher than first order. The entropy consistency in conjunction with the need to use limiters is the key problem for higher order methods. For conservation laws the most important issue is conservativity. This and other essential properties need careful study in the scalar case, as was done in this paper. The method described in this paper is already an improvement over previous approaches. An important further step is to find a practical solution for the multidimensional systems appearing in practical applications. This development must be guided by the type of analysis we presented in our paper, especially concerning the issue of conservativity.

The analysis and the computational results in this paper suggest that the approach of mixing implicit and explicit methods should be considered further from a computational point of view. Next to considering variants of the computations in this paper, methods of at least second order are needed for practical purposes, even if their analytical basis is still incomplete.

Acknowledgment. The authors would like to thank the referees for many helpful suggestions during the revision of the paper.

REFERENCES

[1] R. Abgrall and S. Karni, *Computations of compressible multifluids*, J. Comput. Phys., 169 (2001), pp. 594–623.

[2] M. J. Berger, *Stability of interfaces with mesh refinement*, Math. Comp., 45 (1985), pp. 301–318.

[3] M. J. Berger, C. Helzel, and R. J. LeVeque, *h-box methods for the approximation of hyperbolic conservation laws on irregular grids*, SIAM J. Numer. Anal., 41 (2003), pp. 893–918.

[4] F. E. Browder, *Nonlinear mappings of nonexpansive and accretive type in Banach spaces*, Bull. Amer. Math. Soc., 73 (1967), pp. 875–882.

[5] M. G. Crandall and T. M. Liggett, *Generation of semi-groups of nonlinear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.

[6] M. G. Crandall and A. Majda, *Monotone difference approximations for scalar conservation laws*, Math. Comp., 34 (1980), pp. 1–21.

[7] C. Dawson and R. Kirby, *High resolution schemes for conservation laws with locally varying time steps*, SIAM J. Sci. Comput., 22 (2001), pp. 2256–2281.

[8] K. Dekker and J. G. Verwer, *Stability of Runge-Kutta Methods for Stiff Nonlinear Differential Equations*, North-Holland, Amsterdam, 1984.

[9] S. Evje and K. H. Karlsen, *Degenerate convection-diffusion equations and implicit monotone difference schemes*, in Hyperbolic Problems: Theory, Numerics, Applications, M. Fey and R. Jeltsch, eds., Birkhäuser, Basel, Switzerland, 1999, pp. 285–294.

[10] H. Freistühler and G. Warnecke, *Hyperbolic Problems: Theory, Numerics, Applications*, Birkhäuser, Basel, Switzerland, 2001.

[11] J. Fuhrmann, *Existence and uniqueness of solutions of certain systems of algebraic equations with off-diagonal nonlinearity*, Appl. Numer. Math., 37 (2001), pp. 359–370.

[12] J. Fuhrmann and H. Langmach, *Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws*, Appl. Numer. Math., 37 (2001), pp. 201–230.

[13] E. Godlewski and P. A. Raviart, *Numerical Approximation of Hyperbolic Systems of Conservation Laws*, Springer, New York, 1996.

[14] M. Grott, S. Chernigovskij, and W. Glatzel, *Simulation of stellar instabilities with vastly different timescales using domain decomposition*, Mon. Not. R. Astron. Soc., 344 (2003), pp. 1119–1130.

[15] W. Heineken and G. Warnecke, *Partitioning methods for reaction-diffusion problems*, Appl. Numer. Math., 56 (2006), pp. 981–1000.
[16] T. Y. Hou and P. G. LeFloch, *Why nonconservative schemes converge to wrong solutions: Error analysis*, Math. Comp., 62 (1994), pp. 497–530.
[17] S. Jin and Z. P. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions,* Comm. Pure Appl. Math., 48 (1995), pp. 235–276.
[18] T. Kato, *Nonlinear semigroups and evolution equations*, J. Math. Soc. Japan, 19 (1967), pp. 508–520.
[19] D. Kröner, *Numerical Schemes for Conservation Laws*, Wiley, Chichester, UK, Teubner, Stuttgart, 1997.
[20] S. N. Kružkov, *First order quasi-linear equations with several space variables*, USSR Math. Sb., 10 (1970), pp. 217–243.
[21] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.
[22] R. J. LeVeque, *Finite Volume Methods for Hyperbolic Problems*, Cambridge University Press, Cambridge, UK, 2002.
[23] R. J. LeVeque, *Large time step shock-capturing techniques for scalar conservation laws*, SIAM J. Numer. Anal., 19 (1982), pp. 1091–1109.
[24] B. J. Lucier, *On non-local monotone difference schemes for scalar conservation laws*, Math. Comp., 47 (1986), pp. 19–36.
[25] S. Osher and R. Sanders, *Numerical approximations to nonlinear conservation laws with locally varying time and space grids*, Math. Comp., 41 (1983), pp. 321–336.
[26] E. Pärt-Enander and B. Sjögreen, *Conservative and non-conservative interpolation between overlapping grids for finite volume solutions of hyperbolic problems*, Comput. & Fluids, 23 (1994), pp. 551–574.
[27] A. Quarteroni, *Domain decomposition methods for systems of conservation laws: Spectral collocation approximations*, SIAM J. Sci. Statist. Comput., 11 (1990), pp. 1029–1052.
[28] R. Sanders, *On convergence of monotone finite difference schemes with variable spatial difference*, Math. Comp., 40 (1983), pp. 91–106.
[29] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Lecture Notes in Math. 1697, A. Quarteroni, ed., Springer, Berlin, 1998, pp. 325–343.
[30] H. S. Tang and T. Zhou, *On nonconservative algorithms for grid interfaces*, SIAM J. Numer. Anal., 37 (1999), pp. 173–193.
[31] H. Z. Tang, *On the sonic point glitch*, J. Comput. Phys., 202 (2005), pp. 507–532.
[32] H. Z. Tang and G. Warnecke, *A class of high resolution schemes for hyperbolic conservation laws and convection-diffusion equations with varying time and space grids*, J. Comput. Math., 24 (2006), pp. 121–140.
[33] H. Tang and G. Warnecke, *A class of high resolution difference schemes for nonlinear Hamilton–Jacobi equations with varying time and space grids*, SIAM J. Sci. Comput., 26 (2005), pp. 1415–1431.
[34] T. Tang and Z.-H. Teng, *Viscosity methods for piecewise smooth solutions to scalar conservation laws*, Math. Comp., 66 (1997), pp. 495–526.
[35] E. F. Toro, *Riemann Solvers and Numerical Methods for Fluid Dynamics*, 2nd ed., Springer, Berlin, 1999.
[36] J. Wang and G. Warnecke, *On entropy consistency of large time step schemes* I. *The Godunov and Glimm schemes*, SIAM J. Numer. Anal., 30 (1993), pp. 1229–1251.
[37] J. Wang and G. Warnecke, *On entropy consistency of large time step schemes* II. *Approximate Riemann solvers*, SIAM J. Numer. Anal., 30 (1993), pp. 1252–1267.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1472–1494

WHY FINITE ELEMENT DISCRETIZATIONS CAN BE FACTORED BY TRIANGULAR HIERARCHICAL MATRICES*

MARIO BEBENDORF†

Abstract. Although the asymptotic complexity of direct methods for the solution of large sparse finite element systems arising from second-order elliptic partial differential operators is far from being optimal, these methods are often preferred over modern iterative methods. This is mainly due to their robustness. In this article it is shown that an approximate LU decomposition exists which can be computed in the algebra of hierarchical matrices with almost linear complexity and with the same robustness as the classical LU decomposition. Low-precision approximations may be used for preconditioning iterative solvers. As a byproduct we prove that Schur complements of stiffness matrices can be approximated with almost linear complexity.

Key words. approximate LU decomposition, fast direct solution, preconditioning, hierarchical matrices

AMS subject classifications. 35C20, 65F05, 65F50, 65N30

DOI. 10.1137/060669747

1. Introduction.

The finite element discretization of Dirichlet boundary value problems

$$\begin{aligned} Du &= f \quad \text{in } \Omega, \\ u &= g \quad \text{on } \partial\Omega \end{aligned}$$

with general second-order elliptic partial differential operators

$$Du = -\operatorname{div}\left[C\nabla u + c'u\right] + c'' \cdot \nabla u + c_0 u \tag{1}$$

and possibly rough coefficients $c_{ij}, c_i', c_j'', c_0 \in L^\infty(\Omega)$, $i,j = 1,\dots,d$, on a bounded Lipschitz domain $\Omega \subset \mathbb{R}^d$ leads to large sparse linear systems

$$Ax = b, \quad A \in \mathbb{R}^{n\times n}. \tag{2}$$

In this article we assume that for the discretization of (1) quasi-uniform finite elements are used.

The numerical solution of linear systems (2) is usually done iteratively by Krylov subspace methods. The sparse coefficient matrix A enters these solvers only through the matrix-vector product. Iterative methods are particularly efficient if an approximate solution of relatively low accuracy is sought. However, the number of iterations may be large depending on the distribution of the eigenvalues of A. For instance, the spectral condition number of finite element discretizations of the second-order operator D from (1) grows like $n^{2/d}$ for large n but also depends significantly on the coefficients of D. Hence, an appropriate preconditioner has to be employed in order to improve the convergence properties. This usually requires tailoring the preconditioner to the respective application.

*Received by the editors September 13, 2006; accepted for publication (in revised form) February 20, 2007; published electronically July 25, 2007. This work was supported by the DFG priority program SPP 1146 "Modellierung inkrementeller Umformverfahren."
http://www.siam.org/journals/sinum/45-4/66974.html
†Fakultät für Mathematik und Informatik, Universität Leipzig, Johannisgasse 26, D-04103 Leipzig, Germany (bebendorf@math.uni-leipzig.de).

The lack of robustness of iterative methods is the reason for the fact that direct solvers are still in use even for large-scale problems. The latter are based on factorizations of the coefficient matrix A into easily invertible matrices. The disadvantage of direct methods is that these factors suffer from so-called *fill-in*; i.e., compared with the sparsity of A considerably more entries of the factors will be nonzero. This usually happens to all entries within the bandwidth of the original matrix, which for Galerkin matrices of operators (1) scales like $n^{1-1/d}$ even if the bandwidth has been reduced, for instance, by the *reverse Cuthill–McKee* (RCM) algorithm [9]. Hence, the fill-in will lead to a computational complexity of order $n^{3-2/d}$. Instead of reducing the bandwidth, the aim of the *minimum degree algorithm* [26, 15] and *nested dissection* [13] is to reduce fill-in. When $d = 2$, the fill-in for nested dissection is of the order $n \log n$, and the complexity is $n^{3/2}$; see [14]. For $d > 2$ the complexity scales like $n^{3-3/d}$; cf. [23]. Constants, however, are attractively small, making direct methods the methods of choice if n is not too large or if we are to solve problems in two spatial dimensions. Here, recent multifrontal solvers (see [1] and the references therein) can be used.

In the last decade many fast algorithms for the direct solution of boundary value problems have been proposed. If Ω is an interval, then the Green function G of ordinary differential operators has the property that

$$G(x,y) = \begin{cases} u_1(x)v_1(y), & x \geq y, \\ u_2(x)v_2(y), & x \leq y, \end{cases}$$

with appropriately chosen functions u_1, v_1, u_2, and v_2. This fact is, for instance, exploited by the algorithm in [30]. The algebraic analogue of the above property are *semiseparable matrices*, which can be factored with linear complexity; cf. [11, 12, 8, 32]. The presence of such structures, which allow an exact factorization, is, however, restricted to one-dimensional problems. For higher dimensions, fast algorithms require approximation; see, for instance, [21] in which the inverse is generated in compressed form.

The aim of this article is to present a new approach that merges the advantages of direct and iterative methods. This will be achieved by generalizing the classical LU decomposition to an approximate LU decomposition, which on one side inherits the robustness of the classical LU decomposition, while on the other side has logarithmic-linear complexity independently of the spatial dimension. The efficiency of this approximate LU decomposition will, however, depend on the accuracy ε although the dependence is only logarithmic. In the limiting case $\varepsilon = 0$, the proposed hierarchical LU decomposition is nothing but a partitioned LU decomposition of a matrix thereby inheriting its $n^{3-2/d}$ asymptotic complexity. Hence, the proposed method aims at applications which admit an approximation of low accuracy. Since the discretization of (1) introduces an approximation error which cannot be removed no matter how accurately the linear system (2) is solved, it is sufficient to compute an LU decomposition with a precision which is of the same order of magnitude as the discretization error. Another application is approximate inverse preconditioning. The results of [4] on approximate preconditioners also hold for approximate inverses obtained by approximate LU decomposition.

Since fill-in will also occur during an approximate LU decomposition, we make use of the structure of hierarchical matrices, by which appropriate dense matrices can be treated with almost linear complexity. Consequently, the bandwidths of the factors L and U will not be an issue. In recent years fast methods for the treatment of large

dense matrices $M \in \mathbb{R}^{n\times n}$ have considerably spread. After the introduction of the *fast multipole method* [25] and the *panel-clustering method* [19], numerous methods have been developed based on low-rank approximations

$$M_{ts} \approx VW^T$$

of appropriate subblocks M_{ts} in the rows and columns $t, s \subset \{1, \dots, n\}$ of M, where $V \in \mathbb{R}^{t\times k}$, $W \in \mathbb{R}^{s\times k}$, and k is small compared with $|t|$ and $|s|$. While the fast multipole method was aiming at an efficient approximate evaluation of matrix-vector products, by the structure of *hierarchical matrices* ($\mathcal{H}$-matrices) (see [17, 18]), one efficiently approximates the matrix entries. Basically, $\mathcal{H}$-matrices are matrices that are of low rank on each block of a certain partition resulting from a recursive subdivision of the set of matrix indices. In addition to the efficient matrix-vector multiplication (also with the transposed matrix) this structure provides approximate operations such as matrix addition, matrix-matrix multiplication, and matrix inversion of fully populated matrices with almost linear complexity. Furthermore, $\mathcal{H}$-matrices can be stored in an almost linear amount of units of memory. In the case of matrices arising from the discretization of integral equations, $\mathcal{H}$-matrices can be efficiently constructed from few of the original matrix entries; see [5, 6]. The existence of $\mathcal{H}$-matrix approximants of almost linear complexity is not self-evident. Recently [7, 2] it was shown that the inverse of finite element discretizations of operators of type (1) can be approximated by $\mathcal{H}$-matrices with a blockwise rank that depends logarithmically on both the number of unknowns n and the accuracy ε. Interestingly, this approximation is very robust with respect to nonsmooth coefficients. The aim of this article is to extend the existence theory to the factors of the LU decomposition. Compared with the $\mathcal{H}$-inverse, the presented $\mathcal{H}$-LU decomposition can be computed in significantly less time while keeping the same robustness with respect to the coefficients of D.

The hierarchical LU decomposition differs conceptually from the so-called *incomplete LU factorization* (ILU); see [27]. The ILU overcomes the problem of fill-in by setting to zero entries in the factors L and U outside of the sparsity pattern of A. Although the ILU can be equipped with a thresholding parameter, it will always result in more or less sparse factors, while the hierarchical LU decomposition is a data-sparse representation of fully (up to the bandwidth) populated matrix approximants.

The structure of this article is as follows: In section 2 a brief review of the structure of $\mathcal{H}$-matrices will be given. The existence of $\mathcal{H}$-matrix approximants, which was proved in [7, 2], will be used to show existence of $\mathcal{H}$-matrix approximants to the factors L and U in section 3. In contrast to the inverse of a finite element Galerkin matrix A, its LU decomposition has no analytic equivalent. It is thus surprising that the matrix partition which has proved useful for elliptic problems can also be used for the approximation of the factors L and U. For the proof of this main result of the article, we first show that each Schur complement in A can be approximated by $\mathcal{H}$-matrices. It will be seen that this knowledge will be sufficient to show that the factors L and U have $\mathcal{H}$-matrix approximants. The complexity estimates show the same dependence on the coefficients of operator (1) as the estimates for the $\mathcal{H}$-inverse. Hence, the asymptotic complexity and the robustness of the hierarchical inverse are inherited by the $\mathcal{H}$-LU decomposition. The ideas of nested dissection can also be used to improve the efficiency of $\mathcal{H}$-matrices; cf. [24]. As for the classical LU decomposition, the $\mathcal{H}$-LU decomposition preserves zero blocks introduced by this special kind of ordering. We remark that the approximation results of this article are obviously valid for the induced matrix partition although we will confine ourselves to standard partitions.

In section 4 the existence result for the factors L and U is used to lay theoretical ground to an algorithm for the approximate factorization. This algorithm uses the $\mathcal{H}$-matrix arithmetic and is related to the partitioned LU decomposition procedure. As a consequence of the partitioned approach, only a limited version of pivoting is possible. Once the matrix partition has been generated from the mesh information, an approximate LU decomposition can be obtained from any Galerkin matrix in a purely algebraic way. Finally, in section 5 numerical results for elliptic partial differential operators with nonsmooth coefficients will confirm our analysis. It will be seen that the proposed approximate LU decomposition can be computed, stored, and used during forward/backward substitution with almost linear complexity. The comparison with MUMPS (see [1]) shows that the exact and the approximate LU decompositions scale similarly for two-dimensional problems. The reduced asymptotic complexity of the $\mathcal{H}$-LU decomposition will be observable in three spatial dimensions.

2. Hierarchical matrices. In this section we will briefly review the structure of $\mathcal{H}$-matrices originally introduced by Hackbusch [17] and Hackbusch and Khoromskij [18]. We will describe the two principles on which the efficiency of $\mathcal{H}$-matrices is based. These are the hierarchical partitioning of the matrix into blocks and the blockwise restriction to low-rank matrices. These principles were also used in the mosaic-skeleton method [31].

In this article we will consider matrices $A \in \mathbb{R}^{n\times n}$ with entries

$$a_{ij} = a(\varphi_j, \varphi_i), \quad i,j = 1,\dots,n, \tag{3}$$

where a is a bilinear form and φ_i are basis functions with supports $X_i := \operatorname{supp} \varphi_i$, $i \in I := \{1,\dots,n\}$. Here, it is crucial that the basis functions φ_i are locally supported. Matrices of type (3) arise, for instance, from the Galerkin method, which is frequently used to discretize operators of type (1). If a arises from the variational formulation of differential operators, then A is a sparse matrix. A will, however, be fully populated in general if a incorporates a nonlocal operator.

In order to be able to approximate each block $t \times s$, $t, s \subset I$, of A by a matrix of low rank, i.e.,

$$A_{ts} \approx VW^T, \quad V \in \mathbb{R}^{t\times k}, W \in \mathbb{R}^{s\times k},$$

where k is small compared with $|t|$ and $|s|$, $t\times s$ has to satisfy a certain condition which is caused by the operator D hidden in a. In the field of elliptic partial differential operators the corresponding Green function $G(x,y)$ has an algebraic singularity for $x = y$. Hence, the following condition on $t \times s$ has proved useful:

$$\min\{\operatorname{diam} X_t, \operatorname{diam} X_s\} < \eta \operatorname{dist}(X_t, X_s), \tag{4}$$

where $\eta > 0$ is a given real number which typically is chosen from the interval $[0.5, 1.5]$. Blocks $t \times s$ satisfying (4) will be called *admissible*. The support X_t of a cluster t is the union of the supports of the basis functions corresponding to the indices contained in t:

$$X_t := \bigcup_{i\in t} X_i.$$

The *far-field* $\mathcal{F}_\eta(t)$ of $t \subset I$ is defined as

$$\mathcal{F}_\eta(t) := \{i \in I : \eta \operatorname{dist}(X_i, X_t) > \operatorname{diam} X_t\},$$

and by $\mathcal{N}_\eta(t) := I \setminus \mathcal{F}_\eta(t)$ we denote the *near-field* of t. As usual we set

$$\operatorname{diam} X = \sup_{x,y\in X} |x-y| \quad \text{and} \quad \operatorname{dist}(X,Y) = \inf_{x\in X,\, y\in Y} |x-y|$$

for two bounded sets $X, Y \subset \mathbb{R}^d$. Hence, (4) is equivalent to the condition $s \subset \mathcal{F}_\eta(t)$ or $t \subset \mathcal{F}_\eta(s)$. Note that (4) implies that the partition we are looking for has to be refined towards the diagonal of A, since the diagonal entries arise from the interaction of the same basis functions; i.e., $\operatorname{dist}(X_t, X_s) = 0$ for all blocks $t \times s$ containing the diagonal.

The construction of a partition P of the matrix indices $I \times I$ consisting of admissible blocks or blocks which are small enough is usually based on cluster trees. A tree T_I satisfying the following conditions is called a *cluster tree* for I:

(i) I is the root of T_I;

(ii) if $t \in T_I$ is not a leaf, then t has sons $t_1, t_2 \in T_I$, so that $t = t_1 \dot\cup t_2$.[1]

The set of sons of $t \in T_I$ is denoted by $\mathcal{S}(t)$, while $\mathcal{L}(T_I)$ stands for the set of leaves of the tree T_I.

A cluster tree is usually generated by recursive subdivision of I. For practical purposes the recursion should be stopped if a certain cardinality $n_{\min}$ of the clusters is reached, rather than subdividing the clusters until only one index is left. The depth of T_I, i.e., the maximum distance of a vertex to the root of the tree increased by one, will be denoted by p, which for quasi-uniform grids can be guaranteed to be of the order $\log n$.

Note that by moving the indices of the first son t_1 to the beginning of t we can always obtain contiguous clusters; i.e., for $t \in T_I$ there are $t_{\min}, t_{\max} \in \mathbb{N}$ such that

$$t = \{i \in I : t_{\min} \le i \le t_{\max}\}.$$

This rearrangement induces a reordering of the index set I.

Remark 2.1. Since for each subdivision we have only two possibilities for arranging the indices, i.e., $t = [t_1, t_2]$ or $t = [t_2, t_1]$, the above construction leaves room for only $2^p n_{\min}!$ permutations of I (the size of the leaves in T_I is assumed to be exactly $n_{\min}$). Hence, building the cluster tree determines the numbering of the indices in I up to $\mathcal{O}(n)$ permutations.

Using a cluster tree, which contains a hierarchy of partitions of I, a block cluster tree $T_{I\times I}$ is constructed by recursively subdividing each block $t \times s$ into four disjoint subblocks $t_1 \times s_1$, $t_1 \times s_2$, $t_2 \times s_1$ and $t_2 \times s_2$, $t_1, t_2 \in \mathcal{S}(t)$, $s_1, s_2 \in \mathcal{S}(s)$, starting from the set of matrix indices $I \times I$. This recursion stops in blocks which satisfy (4) or are small enough. The resulting set of leaves $P := \mathcal{L}(T_{I\times I})$ is a partition with the desired properties. For a detailed description of the partitioning of a matrix into admissible subblocks the reader is referred to [4]. An important property of P is that a constant $c_{\mathrm{sp}} > 0$ exists such that for each set of indices $t \subset I$ there are at most c_{sp} blocks $t \times s \in P$ and at most c_{sp} blocks $s \times t \in P$ with some set of indices $s \subset I$; cf. [16].

On P we define the set of $\mathcal{H}$-matrices (see Figure 1) with blockwise rank k by

$$\mathcal{H}(T_{I\times I}, k) := \{M \in \mathbb{R}^{I\times I} : \operatorname{rank} M_b \le k \text{ for all } b \in \mathcal{L}(T_{I\times I})\}.$$

Note that $\mathcal{H}(T_{I\times I}, k)$ is not a linear space since the sum of two rank-k matrices exceeds rank k in general.

[1] In the case of nested dissection a subdivision into three sons is required.

Remark 2.2. For a block $B \in \mathbb{R}^{t\times s}$ the low-rank representation $B = VW^T$, $V \in \mathbb{R}^{t\times k}$, $W \in \mathbb{R}^{s\times k}$, is advantageous compared with the entrywise representation only if $k(|t| + |s|) \leq |t|\,|s|$. For the sake of simplicity in this article we will, however, assume that each block has the low-rank representation. Employing the entrywise representation for appropriate blocks will accelerate the algorithms.

Exploiting the hierarchical structure of $M \in \mathcal{H}(T_{I\times I}, k)$, it can be shown that the storage requirement for $M \in \mathcal{H}(T_{I\times I}, k)$ is of the order $kn \log n$. Multiplying M by a vector can be done with $\mathcal{O}(kn \log n)$ arithmetical operations. Two $\mathcal{H}$-matrices from $\mathcal{H}(T_{I\times I}, k)$ can be added with complexity $\mathcal{O}(k^2 n \log n)$ provided that the sum can be approximated with the desired accuracy by a matrix having blockwise rank at most k. The complexity of computing a rounded product of two $\mathcal{H}$-matrices is $\mathcal{O}(k^2 n(\log n)^2)$; see [17, 18, 16]. It was already mentioned in [17] that LU decompositions can be computed in the algebra of $\mathcal{H}$-matrices with complexity $k^2 n(\log n)^2$ assuming that arising blocks are approximated by rank-k matrices. In order to be able to deduce almost linear complexity from these complexity estimates it is crucial to know how the blockwise rank k depends on the accuracy ε of the approximation or, conversely, which accuracy is associated with a given k. In the extreme case $k \sim n$ the $\mathcal{H}$-matrix operations would be as expensive as the usual ones. Analyzing the relation between k and ε is the main purpose of this article. It will turn out that k depends logarithmically on both ε and n.

2.1. Bandwidth and $\mathcal{H}$-matrices. Although $\mathcal{H}$-matrices are primarily aiming at dense matrices, the stiffness matrix A of the differential operator D from (1) is in $\mathcal{H}(T_{I\times I}, n_{\min})$ and can be stored in this format with complexity $\mathcal{O}(n)$. This can be seen by the following arguments. If $b \in P$ is admissible, then the supports of the basis functions are pairwise disjoint. Hence, the matrix entries in this block vanish. In the remaining case, b does not satisfy (4). Then the size of one of the clusters is less than or equal to $n_{\min}$. In either case, the rank of A_b does not exceed $n_{\min}$. The last observation is of particular importance since it will allow us to compute an LU decomposition using approximate arithmetical operations on the set of $\mathcal{H}$-matrices; see section 4.

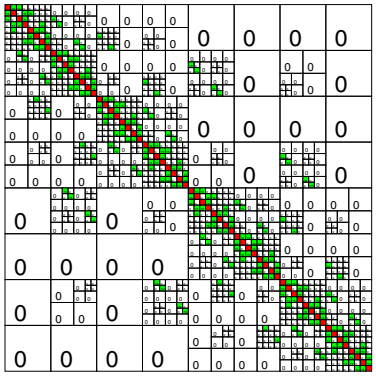


FIG. 1. *A sparse $\mathcal{H}$-matrix with its rank distribution.*

The efficiency of the usual LU decomposition is determined by the bandwidth of A unless special reordering techniques such as nested dissection are employed. The reason for this is that although A is sparse, the factors L and U will in general be fully

populated up to the bandwidth. Since $\mathcal{H}$-matrices are able to handle dense matrices with almost linear complexity, the bandwidth of A is not an issue when using this structure. Due to the reordering of indices required when building the cluster tree, we even obtain a bandwidth which is of order n. This will result in an enormous fill-in and is unavoidable as can be seen by the following example.

A matrix entry a_{ij} in the Galerkin matrix A will in general be nonzero if the supports of the associated basis functions φ_i and φ_j have a nonempty intersection. For simplicity we investigate the situation which occurs for a regular triangulation of the unit square in $\mathbb{R}^2$. Assume that after two subdivision steps this square has been subdivided into four smaller squares of the same size each containing $n/4$ supports. During the subdivision, the indices are reordered so that the kth square contains the indices $(k-1)n/4+1$ to $kn/4$, $k=1,\dots,4$. Hence, the first and the last square contain indices which differ by at least $n/2$. These squares intersect in the center of the original square. Therefore, at this point the supports of two basis functions φ_i and φ_j with $|i-j|\geq n/2$ intersect. This situation persists when the subdivision is continued since the indices are rearranged only within each subsquare.

2.2. Where can $\mathcal{H}$-matrices be applied? One of the first applications of the structure of $\mathcal{H}$-matrices was the acceleration of both the building process and the matrix-vector multiplication of discrete integral operators with smooth kernels having an algebraic singularity at $x=y$. This kind of integral operator arises, for instance, from the boundary element method. For such operators the adaptive cross approximation (ACA) algorithm [5, 6] can be used to generate the low-rank approximants from few of the original matrix entries.

In addition to discretizations of integral operators with smooth kernel functions, in [7, 2] it was shown that inverses of discrete elliptic differential operators with measurable coefficients can be approximated on partitions satisfying (4). Since the analysis of this article will be based on approximations of the inverse, we state the main result of [2]. Let the operator D from (1) be uniformly elliptic; i.e., for the coefficient $C(x)\in\mathbb{R}^{d\times d}$ of D it holds that C is symmetric with $c_{ij}\in L^\infty(\Omega)$ and

$$0<\lambda\leq\lambda(x)\leq\Lambda$$

for all eigenvalues $\lambda(x)$ of $C(x)$ and almost all $x\in\Omega$. Furthermore, let $e_h(u):=\|u-P_hu\|_{L^2(\Omega)}$ be the finite element error, where $P_h: H^1_0(\Omega)\to V_h$ is the Ritz projector mapping $u\in H^1_0(\Omega)$ to its finite element solution u_h; i.e., the solution of $a(u_h,v_h)=\ell(v_h)$ for all $v_h\in V_h$ with a linear form ℓ. We assume that the finite element method converges in the following sense:

$$e_h(u)\leq\varepsilon_h\|f\|_{L^2(\Omega)}\quad\text{for all } u=D^{-1}f,\ f\in L^2(\Omega), \tag{5}$$

where $\varepsilon_h\to 0$ as $h\to 0$. Note that due to a possible lack of regularity of D one cannot guarantee a specific rate of convergence.

THEOREM 2.3. *Let p be the depth of the cluster tree T_I defined in the beginning of section 2. Then there is $C_{\mathcal{H}}\in\mathcal{H}(T_{I\times I},k)$ with $k:=p^2\log^{d+1}(p/\varepsilon_h)$ such that*

$$\|A^{-1}-C_{\mathcal{H}}\|_2<c\,\varepsilon_h\|A^{-1}\|_2,$$

where $c=c(D,\Omega,\eta)>0$ depends on the size of coefficients of D, the diameter of Ω, and the cluster parameter η. If $\varepsilon_h=\mathcal{O}(h^\beta)$ for some $\beta>0$, then $k=\mathcal{O}(\log^{d+3}n)$ holds.

Remark 2.4. Since the proof of Theorem 2.3 is based on the finite element error estimate (5), we were only able to show existence of approximants with an accuracy which is of the order of the finite element error ε_h. This is not a restriction since a higher accuracy in the approximation of the inverse would be superposed by the finite element error in the solution anyhow. However, numerical experiments show that the above result is true for any accuracy. Therefore, in this article we assume that for any $\varepsilon > 0$ there is $C_{\mathcal{H}} \in \mathcal{H}(T_{I\times I}, k)$ with $k := |\log \varepsilon|^{d+1}(\log n)^2$ such that

$$\|A^{-1} - C_{\mathcal{H}}\|_2 < c\,\varepsilon \|A^{-1}\|_2,$$

where $c > 0$ depends on the size of the coefficients of D, the diameter of Ω, and η.

2.3. Schur complements. Among other applications, the efficient treatment of Schur complements is of particular importance for domain decomposition methods (see, for instance, [29]). In this section it will be shown that Schur complements of subblocks of A can be approximated by $\mathcal{H}$-matrices. This result will lay ground to our main aim, the approximation of the factors L and U arising from the LU decomposition of A.

Assume that the Galerkin stiffness matrix $A \in \mathbb{R}^{n\times n}$ is partitioned in the following way:

$$A = \begin{bmatrix} A_{rr} & A_{rr'} \\ A_{r'r} & A_{r'r'} \end{bmatrix}, \tag{6}$$

where $r \subset I$ and $r' := I \setminus r$. We will show that the Schur complement

$$S := A_{r'r'} - A_{r'r} A_{rr}^{-1} A_{rr'}$$

of A_{rr} in A can be approximated by an $\mathcal{H}$-matrix with blockwise rank k, where k depends only logarithmically on both the approximation accuracy ε and n. For this purpose it is crucial to notice that A_{rr} in (6) is nothing but the Galerkin matrix of D if we replace Ω by the subdomain X_r. Hence, Theorem 2.3 guarantees that an $\mathcal{H}$-matrix approximant for A_{rr}^{-1} exists. Additionally, we assume that there is a constant $c > 0$ such that

$$\|A_{rr}^{-1}\|_2 \le c\|A^{-1}\|_2. \tag{7}$$

The previous estimate holds, for instance, if D is self-adjoint and its associated bilinear form a is V_h-coercive in the sense that there is $\alpha > 0$ satisfying

$$\alpha = \inf_{v_h \in V_h} \frac{a(v_h, v_h)}{\|v_h\|_{H^1}^2}.$$

Then

$$\|A_{rr}^{-1}\|_2 = \sup_{x\in\mathbb{R}^I} \frac{\|Rx\|_2^2}{a(JRx, JRx)} \le \frac{1}{c_1 h^d} \sup_{x\in\mathbb{R}^I} \frac{\|JRx\|_{H^1}^2}{a(JRx, JRx)} \le \frac{1}{\alpha c_1 h^d} \tag{8a}$$

$$= \frac{1}{c_1 h^d} \sup_{u\in H^1(\Omega)} \frac{\|Jx\|_{H^1}^2}{a(Jx, Jx)} \le c' \frac{c_2}{c_1} \sup_{x\in\mathbb{R}^I} \frac{\|x\|_2^2}{a(Jx, Jx)} = c' \frac{c_2}{c_1} \|A^{-1}\|_2, \tag{8b}$$

where R denotes the restriction of $\mathbb{R}^I$ to $\mathbb{R}^r$ and $Jx := \sum_{i\in I} x_i \varphi_i$ with the finite element basis functions φ_i. In (8) we have used that for quasi-uniform discretizations it holds that

$$c_1 \|x\|_2^2 \le h^{-d} \|Jx\|_{L^2}^2 \le c_2 \|x\|_2^2 \quad \text{for all } x \in \mathbb{R}^I$$

and that $\|\cdot\|_{H^1} \le c'\|\cdot\|_{L^2}$ on $H_0^1(\Omega)$.

Let $t \in T_I$ be a cluster. By

$$N(t) = \{i \in I : \operatorname{dist}(X_i, X_t) = 0\} \tag{9}$$

we denote a neighborhood of t. Furthermore, we define the ratios

$$q := \frac{\max_{i\in I} \operatorname{diam} X_i}{\min_{t\in T_I} \operatorname{diam} X_t} \quad \text{and} \quad \bar q := \max_{t\in T_I} \frac{\operatorname{diam} X_{N(t)}}{\operatorname{diam} X_t} \le 1 + 2q.$$

Since the minimal cluster size $n_{\min}$ is usually chosen larger than 20, one can expect that realistic values for $\bar q$ are close to 1. The size of q depends on the uniformity of $\{X_i\}_{i\in I}$. We need the following basic lemma, which states that the neighborhood of t is in the far-field of the neighborhood of s if t is in the far-field of s.

LEMMA 2.5. *Let $0 < \eta < (q+\bar q)^{-1}$. If $t \subset \mathcal{F}_\eta(s)$, then*

$$N(t) \subset \mathcal{F}_{\tilde\eta}(N(s)), \quad \textit{where} \quad \tilde\eta = \frac{\bar q\eta}{1-(q+\bar q)\eta}.$$

Proof. Since $\max_{i\in I} \operatorname{diam} X_i \le q \operatorname{diam} X_s$, we obtain for $x \in X_{N(t)}$ and $y \in X_{N(s)}$ that

$$\begin{aligned} |x-y| &\ge \operatorname{dist}(X_t, X_s) - \max_{i\in I} \operatorname{diam} X_i - \operatorname{diam} X_{N(s)} \\ &> \left(\frac{1}{\eta} - q\right) \operatorname{diam} X_s - \operatorname{diam} X_{N(s)} \\ &\ge \left[\frac{1}{\bar q}\left(\frac{1}{\eta} - q\right) - 1\right] \operatorname{diam} X_{N(s)}, \end{aligned}$$

which proves the assertion. ∎

Approximation results for $\mathcal{H}$-matrices are usually derived for each block of the partition. By the following estimates it is possible to relate the blockwise norms of an $\mathcal{H}$-matrix to its global norm. If we are interested in the Frobenius norm, estimates for each block E_b, $b \in P$, immediately lead to an estimate for $E \in \mathbb{R}^{I\times I}$ due to

$$\|E\|_F^2 = \sum_{b\in P} \|E_b\|_F^2.$$

For the spectral norm the situation is a bit more difficult. We can, however, use the following lemma together with the structure of P.

LEMMA 2.6. *We consider the $\nu \times \nu$ block matrix*

$$E = \begin{bmatrix} E_{11} & \dots & E_{1\nu} \\ \vdots & & \vdots \\ E_{\nu 1} & \dots & E_{\nu\nu} \end{bmatrix} \tag{10}$$

with $E_{ij} \in \mathbb{R}^{m_i\times n_j}$, $i,j = 1,\dots,\nu$. Then it holds that

$$\max_{i,j=1,\dots,\nu} \|E_{ij}\|_2 \le \|E\|_2 \le \left(\max_{i=1,\dots,\nu} \sum_{j=1}^{\nu} \|E_{ij}\|_2\right)^{1/2} \left(\max_{j=1,\dots,\nu} \sum_{i=1}^{\nu} \|E_{ij}\|_2\right)^{1/2}. \tag{11}$$

Proof. Let $u = [u_1, \dots, u_\nu]^T \in \mathbb{R}^n$, where $n = \sum_{j=1}^{\nu} n_j$. Observe that

$$\|Eu\|_2^2 = \sum_{i=1}^{\nu} \left\| \sum_{j=1}^{\nu} E_{ij} u_j \right\|_2^2 \le \sum_{i=1}^{\nu} \left(\sum_{j=1}^{\nu} \|E_{ij}\|_2 \|u_j\|_2 \right)^2 = \|\hat{E}\hat{u}\|_2^2,$$

where $\hat{E} \in \mathbb{R}^{\nu\times\nu}$ has the entries $\hat{E}_{ij} = \|E_{ij}\|_2$ and $\hat{u} \in \mathbb{R}^\nu$ is the vector with components $\hat{u}_j = \|u_j\|_2$, $j = 1, \dots, \nu$. It is well known that $\|\hat{E}\|_2^2 \le \|\hat{E}\|_1 \|\hat{E}\|_\infty$. Hence,

$$\|\hat{E}\hat{u}\|_2^2 \le \|\hat{E}\|_1 \|\hat{E}\|_\infty \|\hat{u}\|_2^2 = \|\hat{E}\|_1 \|\hat{E}\|_\infty \|u\|_2^2$$

gives the first part of the assertion. The lower bound follows from the fact that the spectral norm of any subblock of E is bounded by the spectral norm of E. ∎

An important consequence of (11) is that for matrices (10) vanishing in all but μ blocks in each row and each column it follows that

$$\max_{i,j=1,\dots,\nu} \|E_{ij}\|_2 \le \|E\|_2 \le \mu \max_{i,j=1,\dots,\nu} \|E_{ij}\|_2.$$

The previous estimate was also proved in [16] with a different technique. This equivalence of the global and the blockwise spectral norm is useful in translating a blockwise error to a global one. When relative error estimates are to be derived, we will additionally need to estimate how a blockwise norm relation is carried over to the whole matrix.

LEMMA 2.7. *Let P be the leaves of a block cluster tree $T_{I\times I}$. Then for $E, F \in \mathcal{H}(T_{I\times I}, k)$ it holds that*

(i) $\max_{b\in P} \|E_b\|_2 \le \|E\|_2 \le c_{\mathrm{sp}} p \max_{b\in P} \|E_b\|_2$;

(ii) $\|E\|_2 \le c_{\mathrm{sp}} p \|F\|_2$ *provided* $\max_{b\in P} \|E_b\|_2 \le \max_{b\in P} \|F_b\|_2$.

Proof. Let E_ℓ denote the part of E which corresponds to the blocks of P from the ℓth level $T_{I\times I}^{(\ell)}$ of $T_{I\times I}$; i.e.,

$$(E_\ell)_b = \begin{cases} E_b, & b \in T_{I\times I}^{(\ell)} \cap P, \\ 0, & \text{else.} \end{cases}$$

Since E_ℓ has tensor structure with at most c_{sp} blocks per row or block column, Lemma 2.6 gives $\|E_\ell\|_2 \le c_{\mathrm{sp}} \max_{b\in T_{I\times I}^{(\ell)}\cap P} \|E_b\|_2$ such that

$$\|E\|_2 \le \sum_{\ell=0}^{p-1} \|E_\ell\|_2 \le c_{\mathrm{sp}} \sum_{\ell=0}^{p-1} \max_{b\in T_{I\times I}^{(\ell)}\cap P} \|E_b\| \le c_{\mathrm{sp}} p \max_{b\in P} \|E_b\|_2.$$

The estimate

$$\max_{b\in P} \|E_b\|_2 \le \max_{b\in P} \|F_b\|_2 \le \|F\|_2$$

gives the second part of the assertion. ∎

Using the last three lemmas, we can now prove that the Schur complement S of finite element Galerkin matrices A can be approximated with almost linear complexity.

THEOREM 2.8. *Let the finite element Galerkin matrix $A \in \mathbb{R}^{n\times n}$ be partitioned as in* (6). *Then for the Schur complement*

$$S = A_{r'r'} - A_{r'r} A_{rr}^{-1} A_{rr'}$$

of A_{rr}, $r \neq \emptyset$, in A and all $\varepsilon > 0$ there is $S_{\mathcal{H}} \in \mathcal{H}(T_{r'\times r'}, k_S)$, where $k_S \sim |\log\varepsilon|^{d+1}(\log|r|)^2$, such that

$$\|S - S_{\mathcal{H}}\|_2 < \kappa p\varepsilon\|A\|_2, \tag{12}$$

where $\kappa := \|A\|_2\|A^{-1}\|_2$ denotes the spectral condition number of A.

Proof. We have to show that for each admissible subblock $t \times s \in P$ of $r' \times r'$ and any prescribed accuracy $\varepsilon > 0$ we can find a low-rank matrix which approximates S_{ts} with accuracy ε. Since $t \times s$ is admissible, $(A_{r'r'})_{ts} = 0$ holds. Hence,

$$S_{ts} = -A_{tr}A_{rr}^{-1}A_{rs} = -\sum_{i,j\in r} A_{ti}(A_{rr}^{-1})_{ij}A_{js}.$$

If $i \notin N(t)$, where $N(t)$ is defined in (9), then $A_{ti} = 0$. If, on the other hand, $j \notin N(s)$, then $A_{js} = 0$. With the notation $N'(t) := N(t) \cap r$, we have

$$S_{ts} = -\sum_{i\in N'(t),\, j\in N'(s)} A_{ti}(A_{rr}^{-1})_{ij}A_{js}.$$

Since $t \times s$ is admissible, $t \subset \mathcal{F}_\eta(s)$ or $s \subset \mathcal{F}_\eta(t)$ holds. According to Lemma 2.5, it follows that $N(t) \subset \mathcal{F}_{\tilde\eta}(N(s))$ or $N(s) \subset \mathcal{F}_{\tilde\eta}(N(t))$ is valid. Following Theorem 2.3 (with η replaced by $\tilde\eta$), there are $X \in \mathbb{R}^{N'(t)\times k}$ and $Y \in \mathbb{R}^{N'(s)\times k}$ with $k \sim |\log\varepsilon|^{d+1}(\log|r|)^2$ such that

$$\|(A_{rr}^{-1})_{N'(t)N'(s)} - XY^T\|_2 < \varepsilon\|A_{rr}^{-1}\|_2.$$

Let X and Y be extended to $\hat X \in \mathbb{R}^{r\times k}$ and $\hat Y \in \mathbb{R}^{r\times k}$ by adding zero rows. Observe that

$$\begin{aligned} A_{tr}\hat X\hat Y^T A_{rs} &= \sum_{i\in N'(t),\, j\in N'(s)} \sum_{\ell=1}^{k} A_{ti}X_{i\ell}Y_{j\ell}A_{js} \\ &= \sum_{\ell=1}^{k} \left(\sum_{i\in N'(t)} A_{ti}X_{i\ell}\right)\left(\sum_{j\in N'(s)} Y_{j\ell}A_{js}\right) =: VW^T, \end{aligned}$$

where $V \in \mathbb{R}^{t\times k}$, $W \in \mathbb{R}^{s\times k}$ have the entries

$$V_{t\ell} := \sum_{i\in N'(t)} A_{ti}X_{i\ell} \quad\text{and}\quad W_{s\ell} := \sum_{j\in N'(s)} Y_{j\ell}A_{js}, \qquad \ell = 1,\dots,k.$$

Define $B \in \mathbb{R}^{r\times r}$ with entries

$$b_{ij} = \begin{cases} (A_{rr}^{-1})_{ij} & \text{if } i \in N'(t) \text{ and } j \in N'(s), \\ 0 & \text{else;} \end{cases}$$

then using (7) it follows that

$$\begin{aligned} \|S_{ts} - VW^T\|_2 &= \|A_{tr}(B - \hat X\hat Y^T)A_{rs}\|_2 \le \|A_{tr}\|_2\|(A_{rr}^{-1})_{N'(t)N'(s)} - XY^T\|_2\|A_{rs}\|_2 \\ &\le \varepsilon\|A_{tr}\|_2\|A_{rr}^{-1}\|\|A_{rs}\|_2 < c\kappa\varepsilon\|A\|_2. \end{aligned}$$

The assertion follows from Lemma 2.7. ∎

Since the spectral condition number κ grows polynomially with n and since the accuracy ε enters the complexity estimate only through the logarithm, we can get rid of the factor κ in (12) if the rank k_S is increased by adding a logarithmic factor.

3. Hierarchical *LU* decomposition. Assume that all minors of A are nonzero. Then A can be factored as

$$A = LU,$$

where L is a unit lower triangular and U is an upper triangular matrix. In this section it will be shown that the factors L and U can be approximated by $\mathcal{H}$-matrices $L_{\mathcal{H}}$ and $U_{\mathcal{H}}$ if any Schur complement in A has this property. Note that the following proof consists of algebraic arguments only. Hence, the LU decompositions can be accelerated also for problems that do not stem from finite element applications as long as the Schur complements are known to have an approximant in the set of $\mathcal{H}$-matrices.

When computing pointwise LU decompositions, usually pivoting is performed in order to avoid zero or almost zero pivots. For block versions of the LU algorithm the possibilities of pivoting are limited if the blocking is given. In our case we can choose from only two possible pivots, block $t_1 \times t_1$ or block $t_2 \times t_2$, if the LU decomposition of a block $t \times t$, $t = t_1 \cup t_2$, is to be computed. Hence, the accuracy analysis cannot rely on the advantages of pivoting.

In order to show that L and U can be approximated by $\mathcal{H}$-matrices it seems natural to define the approximants

$$\tilde{L} = \begin{bmatrix} \tilde{L}_{11} & \\ \tilde{L}_{21} & \tilde{L}_{22} \end{bmatrix} \quad \text{and} \quad \tilde{U} = \begin{bmatrix} \tilde{U}_{11} & \tilde{U}_{12} \\ & \tilde{U}_{22} \end{bmatrix}$$

recursively as

$$\tilde{L}_{11}\tilde{U}_{11} = A_{11} + E_{11}, \tag{13a}$$

$$\tilde{L}_{11}\tilde{U}_{12} = A_{12} + E_{12}, \tag{13b}$$

$$\tilde{L}_{21}\tilde{U}_{11} = A_{21} + E_{21}, \tag{13c}$$

$$\tilde{L}_{22}\tilde{U}_{22} = A_{22} - \tilde{L}_{21}\tilde{U}_{12} + E_{22}, \tag{13d}$$

replacing appropriate subblocks of the arising Schur complements with low-rank matrices thereby introducing the error terms E_{12} and E_{21}. The errors E_{11} and E_{22} could then be estimated by the error analysis of the block LU decomposition; see [10]. The problem with this approach is that the arising Schur complements (see (13d)) are not the original complements but complements that contain the perturbations from all previous approximation steps. Since it cannot be guaranteed that these perturbed complements can be approximated by $\mathcal{H}$-matrices and since their distance to the exact complements leads to unattractive estimates, we have to go a different way for the proof. In the following subsection we first find a recursive relation between the Schur complement of a block b and the complements of its subblocks.

3.1. A hierarchy of Schur complements. Let $A \in \mathbb{R}^{n\times n}$ and $t, s \subset I$. With the notations $\hat{t} = \{i \in I : i \le \max t\}$ and $\hat{s} = \{j \in I : j \le \max s\}$ the Schur complement for the block $t \times s$ in $A_{\hat{t}\hat{s}}$ is defined as

$$S(t,s) = A_{ts} - A_{tr}A_{rr}^{-1}A_{rs}, \tag{14}$$

where $r = \{i \in I : i < \min t \cup s\}$; see Figure 2. Note that in the case $r = \emptyset$ this definition is meant to result in $S(t,s) = A_{ts}$. For $t = s$ the expression $S(t,s)$ is the usual Schur complement of A_{rr} in $A_{\hat{t}\hat{t}}$. Note that if t_1, t_2 are the sons of t, then

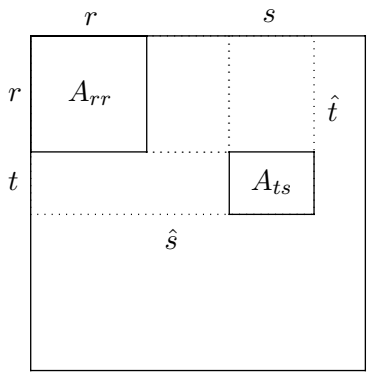


FIG. 2. *Schur complement of a block* $t \times s$.

$S(t_2, t_2)$ does not coincide with the subblock in the rows t_2 and columns t_2 of $S(t, t)$ in general. The following lemma will show the right relation between the complement of a block and the complements of its subblocks. For the ease of notation we first consider the case of blocks on the diagonal.

LEMMA 3.1. *Let* $t \in T_I$ *and let* t_1, t_2 *be its sons. Then*

$$S(t,t) = \begin{bmatrix} S(t_1,t_1) & S(t_1,t_2) \\ S(t_2,t_1) & S(t_2,t_2) + S(t_2,t_1)S(t_1,t_1)^{-1}S(t_1,t_2) \end{bmatrix}.$$

Proof. Let $S(t, t)$ be decomposed in the following way:

$$S(t,t) = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

For the blocks (t_1, t_1), (t_1, t_2), and (t_2, t_1) the definition of r from (14) results in $r = \{i \in I : i < \min t\}$. Hence, we obtain

$$S(t_1,t_2) = A_{t_1t_2} - A_{t_1r}A_{rr}^{-1}A_{rt_2} = S_{12}.$$

Similarly, one sees that $S(t_1, t_1) = S_{11}$ and $S(t_2, t_1) = S_{21}$. It remains to show that

$$S_{22} = S(t_2,t_2) + S_{21}S_{11}^{-1}S_{12}.$$

Let $\bar{r} = \{i \in I : i < \min t\}$. Then from the definition of $S(t, t)$ it follows that

$$S(t_2,t_2) = A_{t_2t_2} - \begin{bmatrix} A_{t_2\bar{r}} & A_{t_2t_1} \end{bmatrix} \begin{bmatrix} A_{\bar{r}\bar{r}} & A_{\bar{r}t_1} \\ A_{t_1\bar{r}} & A_{t_1t_1} \end{bmatrix}^{-1} \begin{bmatrix} A_{\bar{r}t_2} \\ A_{t_1t_2} \end{bmatrix}.$$

Since

$$\begin{bmatrix} A_{\bar{r}\bar{r}} & A_{\bar{r}t_1} \\ A_{t_1\bar{r}} & A_{t_1t_1} \end{bmatrix}^{-1} = \begin{bmatrix} A_{\bar{r}\bar{r}}^{-1} & -A_{\bar{r}\bar{r}}^{-1}A_{\bar{r}t_1}S_{11}^{-1} \\ 0 & S_{11}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{t_1\bar{r}}A_{\bar{r}\bar{r}}^{-1} & I \end{bmatrix},$$

we have

$$\begin{aligned} S(t_2,t_2) &= A_{t_2t_2} - \begin{bmatrix} A_{t_2\bar{r}} & A_{t_2t_1} \end{bmatrix} \begin{bmatrix} A_{\bar{r}\bar{r}}^{-1} & -A_{\bar{r}\bar{r}}^{-1}A_{\bar{r}t_1}S_{11}^{-1} \\ 0 & S_{11}^{-1} \end{bmatrix} \begin{bmatrix} I & 0 \\ -A_{t_1\bar{r}}A_{\bar{r}\bar{r}}^{-1} & I \end{bmatrix} \begin{bmatrix} A_{\bar{r}t_2} \\ A_{t_1t_2} \end{bmatrix} \\ &= A_{t_2t_2} - \begin{bmatrix} A_{t_2\bar{r}} & A_{t_2t_1} \end{bmatrix} \begin{bmatrix} A_{\bar{r}\bar{r}}^{-1} & -A_{\bar{r}\bar{r}}^{-1}A_{\bar{r}t_1}S_{11}^{-1} \\ 0 & S_{11}^{-1} \end{bmatrix} \begin{bmatrix} A_{\bar{r}t_2} \\ S_{12} \end{bmatrix} \\ &= A_{t_2t_2} - \begin{bmatrix} A_{t_2\bar{r}}A_{\bar{r}\bar{r}}^{-1} & S_{21}S_{11}^{-1} \end{bmatrix} \begin{bmatrix} A_{\bar{r}t_2} \\ S_{12} \end{bmatrix} = S_{22} - S_{21}S_{11}^{-1}S_{12}, \end{aligned}$$

which proves the assertion. ∎

Since each block $t \times s$ in the upper triangular part, i.e., $\max t \le \min s$, can be embedded into the block $r \times r$, $r = \{i \in I : \min t \le i \le \max s\}$, Lemma 3.1 gives

$$S(t,s) = \begin{bmatrix} S(t_1,s) \\ S(t_2,s_1) + S(t_2,t_1)S(t_1,t_1)^{-1}S(t_1,s) \end{bmatrix}. \tag{15}$$

Similarly, for each block $t \times s$ in the lower triangular part, i.e., $\max s \le \min t$, it holds that

$$S(t,s) = \begin{bmatrix} S(t,s_1) & S(t,s_2) + S(t_1,s_1)S(s_1,s_1)^{-1}S(s_1,s_2) \end{bmatrix}.$$

3.2. Constructing the factors $L_{\mathcal{H}}$ and $U_{\mathcal{H}}$. We assume that the corresponding Schur complement $S(t,s)$ for each admissible block $t \times s \in \mathcal{L}(T_{I\times I})$ can be approximated by a matrix of low rank with arbitrary accuracy; i.e, for all $\varepsilon > 0$ there is $\tilde{S}(t,s) \in \mathbb{R}^{t\times s}$ of rank $k_S \sim (\log n)^{\alpha} |\log \varepsilon|^{\beta}$ with some $\alpha, \beta > 0$ such that

$$\|S(t,s) - \tilde{S}(t,s)\|_2 \le \varepsilon \|A\|_2. \tag{16}$$

According to Theorem 2.8, this assumption is fulfilled, for instance, in the case of finite element stiffness matrices.

In order to define the factors $L(t)$ and $U(t)$ of $S(t,t) = L(t)U(t)$, $t \in T_I \setminus \mathcal{L}(T_I)$, we set

$$L(t) := \begin{bmatrix} L(t_1) & 0 \\ S(t_2,t_1)U(t_1)^{-1} & L(t_2) \end{bmatrix} \quad \text{and} \quad U(t) := \begin{bmatrix} U(t_1) & L(t_1)^{-1}S(t_1,t_2) \\ 0 & U(t_2) \end{bmatrix}, \tag{17}$$

where

$$L(t_1)U(t_1) = S(t_1,t_1), \qquad L(t_2)U(t_2) = S(t_2,t_2),$$

and t_1, t_2 are the sons of t. If $t \in \mathcal{L}(T_I)$, then $L(t)$ and $U(t)$ are defined by the pointwise LU decomposition. Note that since

$$L(t)U(t) = \begin{bmatrix} L(t_1)U(t_1) & S(t_1,t_2) \\ S(t_2,t_1) & L(t_2)U(t_2) + S(t_2,t_1)S(t_1,t_1)^{-1}S(t_1,t_2) \end{bmatrix},$$

we obtain $L(t)U(t) = S(t,t)$ due to Lemma 3.1. The following lemma shows that the off-diagonal blocks in (17) can be approximated by hierarchical matrices. For its proof we will make use of

$$\|S(t,t)^{-1}\|_2 \le \|A_{\hat{t}\hat{t}}^{-1}\|_2 \le c\|A^{-1}\|_2, \tag{18}$$

where $\hat{t} := \{i \in I : i \le \max t\}$. Estimate (18) follows from (7) and the fact that $S(t,t)^{-1}$ is the $t \times t$ subblock of $A_{\hat{t}\hat{t}}^{-1}$.

LEMMA 3.2. *Let X, Y solve $L(t)X = S(t,s)$ and $YU(t) = S(s,t)$, where $\max t \le \min s$. Then X and Y can be approximated by $\tilde{X} \in \mathcal{H}(T_{t\times s}, k_S)$ and $\tilde{Y} \in \mathcal{H}(T_{s\times t}, k_S)$ such that*

$$\|X - \tilde{X}\|_2 \le c\kappa p\varepsilon \|U(t)\|_2 \quad \text{and} \quad \|Y - \tilde{Y}\|_2 \le c\kappa p\varepsilon \|L(t)\|_2.$$

The bound k_S on the blockwise rank was defined in (16).

Proof. We first prove by induction that X can be approximated by a matrix $\tilde{X} \in \mathcal{H}(T_{t\times s}, k_S)$ such that on each admissible subblock $t' \times s'$ of $t \times s$ it holds that

$$\|X_{t's'} - \tilde{X}_{t's'}\|_2 \le c\kappa\varepsilon\|U(t')\|_2. \tag{19}$$

If $t \times s$ is an admissible leaf in T_I, then we have assumed (see (16)) that $S(t,s)$ can be approximated by a matrix $\tilde{S}(t,s) \in \mathbb{R}^{t\times s}$ of rank at most k_S. Hence, the rank of $\tilde{X} := L(t)^{-1}\tilde{S}(t,s)$ cannot exceed k_S, and we have that

$$\begin{aligned}\|X - \tilde{X}\|_2 &= \|L(t)^{-1}[S(t,s) - \tilde{S}(t,s)]\|_2 = \|U(t)S(t,t)^{-1}[S(t,s) - \tilde{S}(t,s)]\|_2\\ &\le \varepsilon\|S(t,t)^{-1}\|_2\|A\|_2\|U(t)\|_2 \le c\kappa\varepsilon\|U(t)\|_2\end{aligned}$$

due to (18). If $t \times s$ is not a leaf, then t has sons t_1, t_2 and s has sons s_1, s_2. Define $X_{11} \in \mathbb{R}^{t_1\times s_1}$, $X_{12} \in \mathbb{R}^{t_1\times s_2}$, $X_{21} \in \mathbb{R}^{t_2\times s_1}$, and $X_{22} \in \mathbb{R}^{t_2\times s_2}$ by

$$L(t_1)X_{11} = S(t_1, s_1), \quad L(t_1)X_{12} = S(t_1, s_2)$$

and

$$L(t_2)X_{21} = S(t_2, s_1), \quad L(t_2)X_{22} = S(t_2, s_2),$$

respectively. By induction we know that X_{11}, X_{12}, X_{21}, and X_{22} can be approximated by $\mathcal{H}$-matrices $\tilde{X}_{11}$, $\tilde{X}_{12}$, $\tilde{X}_{21}$, and $\tilde{X}_{22}$ to the subtrees of $T_{I\times I}$ with roots $t_1 \times s_1$, $t_1 \times s_2$, $t_2 \times s_1$, and $t_2 \times s_2$, respectively. Hence,

$$X = \begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}$$

satisfies

$$\begin{aligned}L(t)X &= \begin{bmatrix} L(t_1) & 0 \\ S(t_2,t_1)U(t_1)^{-1} & L(t_2) \end{bmatrix}\begin{bmatrix} X_{11} & X_{12} \\ X_{21} & X_{22} \end{bmatrix}\\ &= \begin{bmatrix} S(t_1,s) \\ S(t_2,s) + S(t_2,t_1)S(t_1,t_1)^{-1}S(t_1,s) \end{bmatrix} = S(t,s)\end{aligned}$$

due to the definition (17) of $L(t)$ and (15) and can be approximated by

$$\tilde{X} := \begin{bmatrix} \tilde{X}_{11} & \tilde{X}_{12} \\ \tilde{X}_{21} & \tilde{X}_{22} \end{bmatrix} \in \mathcal{H}(T_{t\times s}, k_S)$$

satisfying (19). The assertion follows from Lemma 2.7, because

$$\|X - \tilde{X}\|_2 \le c_{\mathrm{sp}} p \max_{t'\times s' \in T_{t\times s}} \|X_{t's'} - \tilde{X}_{t's'}\|_2 \le cc_{\mathrm{sp}}\kappa p\varepsilon \max_{t'\in T_t} \|U(t')\|_2 \le cc_{\mathrm{sp}}\kappa p\varepsilon\|U(t)\|_2.$$

The proof for Y can be done analogously. ∎

The following theorem is the main result of this article. Although the proved error estimates have a Wilkinson style, these theorems are not meant as estimates on the backward stability of an algorithm for the approximation by $\mathcal{H}$-matrices. However, they show that $\mathcal{H}$-matrix approximants exist and that their complexity depends logarithmically on the accuracy $\varepsilon \to 0$. A similar estimate obviously holds for the Cholesky decomposition.

THEOREM 3.3. *Assume that* (16) *holds and let L and U be the unique unit lower and upper triangular factors of A. Then there are unit lower and upper triangular matrices $L_{\mathcal{H}}, U_{\mathcal{H}} \in \mathcal{H}(T_{I\times I}, k_S)$ such that*

$$\|A - L_{\mathcal{H}}U_{\mathcal{H}}\|_2 \le c\kappa p\varepsilon\|L\|_2\|U\|_2 + \mathcal{O}(\varepsilon^2).$$

Proof. Since $A = S(I,I) = L(I)U(I)$, it follows from the uniqueness of the LU decomposition that $L = L(I)$ and $U = U(I)$. According to Lemma 3.2, there are $\mathcal{H}$-matrices $L_{\mathcal{H}}, U_{\mathcal{H}} \in \mathcal{H}(T_{I\times I}, k_S)$ satisfying

$$\|L - L_{\mathcal{H}}\|_2 \le c\kappa p\varepsilon\|L\|_2 \quad \text{and} \quad \|U - U_{\mathcal{H}}\|_2 \le c\kappa p\varepsilon\|U\|_2.$$

As a consequence we have

$$\begin{aligned}\|A - L_{\mathcal{H}}U_{\mathcal{H}}\|_2 &\le \|(L - L_{\mathcal{H}})U\|_2 + \|L(U - U_{\mathcal{H}})\|_2 + \|(L - L_{\mathcal{H}})(U - U_{\mathcal{H}})\|_2\\ &\le \|L - L_{\mathcal{H}}\|_2\|U\|_2 + \|L\|_2\|U - U_{\mathcal{H}}\|_2 + \|L - L_{\mathcal{H}}\|_2\|U - U_{\mathcal{H}}\|_2\\ &\le [2c\kappa p\varepsilon + c^2\kappa^2p^2\varepsilon^2]\|L\|_2\|U\|_2,\end{aligned}$$

which proves the assertion. ∎

Since the complexity depends logarithmically on the accuracy ε, the blockwise rank of the factors $L_{\mathcal{H}}$ and $U_{\mathcal{H}}$ compared with the rank of the inverse bears an additional logarithmic factor. However, from the following numerical experiments it will be seen that the complexity of the $\mathcal{H}$-LU decomposition is much smaller than the complexity of the $\mathcal{H}$-inverse in practice.

4. Computing the hierarchical *LU* decomposition. In the last section we have seen that the factors L and U from an LU decomposition of A can be approximated by $\mathcal{H}$-matrices $L_{\mathcal{H}}$ and $U_{\mathcal{H}}$ whenever the Schur complements in A possess this property. Although the construction used for the proof could in principle be used to compute $L_{\mathcal{H}}$ and $U_{\mathcal{H}}$, for an improved efficiency, we prefer to use another method which is based on the partitioned LU decomposition.

On the set $\mathcal{H}(T_{I\times I}, k)$ of hierarchical matrices approximate versions of the usual matrix operations such as addition, matrix-matrix multiplication, and inversion can be defined; cf. [17, 18, 16]. The rounding precision these operations are performed with will be denoted by $\varepsilon_{\mathcal{H}}$. The hierarchical LU decomposition can then be computed using these operations during the block LU decomposition instead of the usual ones. The LU decomposition of $\mathcal{H}$-matrices in a format that does not account for general $\mathcal{H}$-matrices has already been used in [22]. The first algorithm for the factorization of general $\mathcal{H}$-matrices was published in [3]; the ideas of nested dissection were first applied to $\mathcal{H}$-matrices in [24].

In order to define the $\mathcal{H}$-LU decomposition we exploit the hierarchical block structure of a block A_{tt}, $t \in T_I \setminus \mathcal{L}(T_I)$:

$$A_{tt} = \begin{bmatrix} A_{t_1t_1} & A_{t_1t_2} \\ A_{t_2t_1} & A_{t_2t_2}\end{bmatrix} = \begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2}\end{bmatrix}\begin{bmatrix} U_{t_1t_1} & U_{t_1t_2} \\ & U_{t_2t_2}\end{bmatrix},$$

where $t_1, t_2 \in T_I$ denote the sons of t in T_I. Hence, the LU decomposition of a block A_{tt} is reduced to the following four problems on the sons of $t \times t$:

(i) Compute $L_{t_1t_1}$ and $U_{t_1t_1}$ from the LU decomposition $L_{t_1t_1}U_{t_1t_1} = A_{t_1t_1}$.
(ii) Compute $U_{t_1t_2}$ from $L_{t_1t_1}U_{t_1t_2} = A_{t_1t_2}$.
(iii) Compute $L_{t_2t_1}$ from $L_{t_2t_1}U_{t_1t_1} = A_{t_2t_1}$.

(iv) Compute $L_{t_2t_2}$ and $U_{t_2t_2}$ from the LU decomposition $L_{t_2t_2}U_{t_2t_2} = A_{t_2t_2} - L_{t_2t_1}U_{t_1t_2}$.

If a block $t \times t \in \mathcal{L}(T_{I\times I})$ is a leaf, the usual pivoted LU decomposition is employed. For (i) and (iv) two LU decompositions of half the size have to be computed. In order to solve (ii), i.e., solve a problem of the structure $L_{tt}B_{ts} = A_{ts}$ for B_{ts}, where L_{tt} is a lower triangular matrix and $t \times s \in T_{I\times I}$, we use a recursive block forward substitution: If the block $t \times s$ is not a leaf in $T_{I\times I}$, from the decompositions of the blocks A_{ts}, B_{ts}, and L_{tt} into their subblocks (t_1, t_2 and s_1, s_2 are again the sons of t and s, respectively)

$$\begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2} \end{bmatrix} \begin{bmatrix} B_{t_1s_1} & B_{t_1s_2} \\ B_{t_2s_1} & B_{t_2s_2} \end{bmatrix} = \begin{bmatrix} A_{t_1s_1} & A_{t_1s_2} \\ A_{t_2s_1} & A_{t_2s_2} \end{bmatrix},$$

one observes that B_{ts} can be found from the equations

$$\begin{aligned} L_{t_1t_1}B_{t_1s_1} &= A_{t_1s_1}, \\ L_{t_1t_1}B_{t_1s_2} &= A_{t_1s_2}, \\ L_{t_2t_2}B_{t_2s_1} &= A_{t_2s_1} - L_{t_2t_1}B_{t_1s_1}, \\ L_{t_2t_2}B_{t_2s_2} &= A_{t_2s_2} - L_{t_2t_1}B_{t_1s_2}, \end{aligned}$$

which are again of type (ii). If, on the other hand, $t \times s$ is a leaf, the usual forward substitution is applied. Similarly, one can solve (iii) by recursive block backward substitution.

The complexity of the above recursions is mainly determined by the complexity of the hierarchical matrix-matrix multiplication, which can be estimated as $\mathcal{O}(k^2n(\log n)^2)$ for two matrices from $\mathcal{H}(T_{I\times I}, k)$; cf. [16]. Each operation is carried out with precision $\varepsilon_{\mathcal{H}}$. A result [10] on the stability analysis of the block LU decomposition states that the product LU is backward stable in the following sense:

$$\|A - LU\|_2 < c(n)\varepsilon_{\mathcal{H}}(\|A\|_2 + \|L\|_2\|U\|_2).$$

Provided that $\|L\|_2\|U\|_2 \approx \|A\|_2$, the relative accuracy of LU will hence be of order $\varepsilon_{\mathcal{H}}$. Employing the $\mathcal{H}$-matrix arithmetic, it is therefore possible to generate an approximate LU decomposition of an $\mathcal{H}$-matrix $A \in \mathcal{H}(T_{I\times I}, k)$ to any prescribed accuracy with almost linear complexity.

Remark 4.1. Although the intermediate results of the presented algorithm (i)–(iv) are guaranteed to be $\mathcal{H}$-matrices, the blockwise rank k is not known. Note that our theory cannot be applied to this construction of the LU decomposition since the computed Schur complements in (iv) are approximate ones. Nevertheless, it will be seen from the numerical experiments that k still depends logarithmically on both the accuracy ε and the number of unknowns n.

In the case of positive definite matrices A it is possible to define an $\mathcal{H}$-version of the Cholesky decomposition of a block A_{tt}, $t \in T_I \setminus \mathcal{L}(T_I)$:

$$A_{tt} = \begin{bmatrix} A_{t_1t_1} & A_{t_1t_2} \\ A_{t_1t_2}^T & A_{t_2t_2} \end{bmatrix} = \begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2} \end{bmatrix} \begin{bmatrix} L_{t_1t_1} & \\ L_{t_2t_1} & L_{t_2t_2} \end{bmatrix}^T.$$

This factorization is recursively computed by

$$\begin{aligned} L_{t_1t_1}L_{t_1t_1}^T &= A_{t_1t_1}, \\ L_{t_1t_1}L_{t_2t_1}^T &= A_{t_1t_2}, \\ L_{t_2t_2}L_{t_2t_2}^T &= A_{t_2t_2} - L_{t_2t_1}L_{t_2t_1}^T, \end{aligned}$$

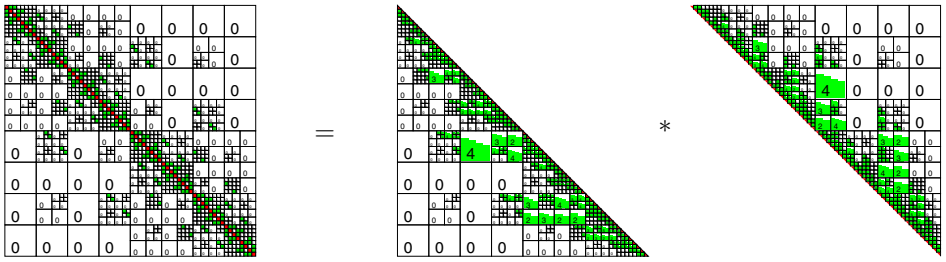
using the usual Cholesky decomposition on the leaves of $T_{I\times I}$. The second equation $L_{t_1t_1}L_{t_2t_1}^T = A_{t_1t_2}$ is solved for $L_{t_2t_1}$ in a way similar to how $U_{t_1t_2}$ was previously obtained in the LU decomposition.

Once A has been decomposed, the solution of $Ax = b$ can be found by forward/backward substitution: $L_{\mathcal{H}}y = b$ and $U_{\mathcal{H}}x = y$. Since $L_{\mathcal{H}}$ and $U_{\mathcal{H}}$ are $\mathcal{H}$-matrices, y_t, $t \in T_I \setminus \mathcal{L}(T_I)$, can be computed recursively by solving the following systems for y_{t_1} and y_{t_2}:

$$L_{t_1t_1}y_{t_1} = b_{t_1} \quad \text{and} \quad L_{t_2t_2}y_{t_2} = b_{t_2} - L_{t_2t_1}y_{t_1}.$$

If $t \in \mathcal{L}(T_I)$ is a leaf, a usual triangular solver is used. The backward substitution can be done analogously. The complexity of this forward/backward substitution is determined by the complexity of the hierarchical matrix-vector multiplication, which is $\mathcal{O}(kn\log n)$ if a matrix from $\mathcal{H}(T_{I\times I}, k)$ is multiplied by a vector.

5. Numerical results. In this section we make use of the above algorithms for the computation of approximate LU decompositions of finite element stiffness matrices in two and three spatial dimensions. The emphasis in these tests is laid on robustness with respect to varying coefficients of the underlying operator.



All computations were carried out on an Athlon64 PC (2 GHz) with 4 GB of core memory. For compiling the $\mathcal{H}$-matrix library,[2] the Intel compiler version 9.0 was used.

5.1. Two-dimensional diffusion. As a first set of tests we consider the Dirichlet boundary value problem

$$\begin{aligned} -\operatorname{div}\alpha(x)\nabla u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $\Omega := (0,1)^2$ is the unit square in $\mathbb{R}^2$ and $\alpha(x)$ is a random number from the interval $[0,a]$ for each $x = (x_1, x_2) \in \Omega$ satisfying $x_1 > x_2$. In the remaining part of Ω the coefficient α is set to 1. The amplitude a will be used to demonstrate that the presented method is not sensitive to nonsmooth coefficients.

The main aim of these two-dimensional tests is to show that the computational complexity of the presented hierarchical LU decomposition is almost linear, thereby confirming our estimates. In the following tables we compare for different problem sizes n and for different amplitudes a the computational effort to decompose the stiffness matrix of the problem from above. Since the discrete operator is symmetric positive definite, we actually compute the Cholesky decomposition LL^T. Table 1

[2] A C++ implementation of the $\mathcal{H}$-matrix structure can be obtained from the following web site: http://www.mathematik.uni-leipzig.de/~bebendorf/AHMED.html.

TABLE 1
$\mathcal{H}$-LU for two-dimensional diffusion with $a = 1$.

	$\varepsilon_{\mathcal{H}} = 1_{10}{-}2$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}4$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}6$			
n	T_{LL^T}	MB	k	$\mathcal{E}_A$	T_{LL^T}	MB	k	$\mathcal{E}_A$	T_{LL^T}	MB	k	$\mathcal{E}_A$
39 061	1.4	32	14	$1.2_{10}{-}3$	1.9	39	15	$2.9_{10}{-}5$	2.5	46	17	$2.5_{10}{-}7$
78 961	3.4	68	10	$1.5_{10}{-}3$	4.9	85	12	$3.0_{10}{-}5$	6.5	100	13	$1.4_{10}{-}7$
159 201	8.0	147	14	$1.9_{10}{-}3$	11.5	185	16	$3.2_{10}{-}5$	15.2	221	17	$3.1_{10}{-}7$
318 096	19.3	307	11	$1.5_{10}{-}3$	28.9	399	12	$2.9_{10}{-}5$	38.9	480	14	$2.3_{10}{-}7$
638 401	45.0	657	14	$1.9_{10}{-}3$	66.5	857	16	$3.2_{10}{-}5$	92.3	1046	17	$4.2_{10}{-}7$
1 276 900	109.3	1358	11	$1.1_{10}{-}3$	167.8	1827	12	$2.9_{10}{-}5$	241.9	2237	14	$2.3_{10}{-}7$
2 556 801	258.5	2858	14	$1.9_{10}{-}3$	387.8	3902	16	$3.2_{10}{-}5$	566.6	4857	17	$3.6_{10}{-}7$

shows the time T_{LL^T} for computing the hierarchical Cholesky decomposition in seconds, its memory consumption (MB), the maximum rank k among the blocks, and the backward error

$$\mathcal{E}_A := \frac{\|A - L_{\mathcal{H}} L_{\mathcal{H}}^T\|_2}{\|A\|_2}.$$

The minimal block size (see section 2) was chosen to be $n_{\min} = 50$. Apparently, the computational effort grows almost linearly with n, while the backward error seems to be independent of n. Increasing the rounding precision $\varepsilon_{\mathcal{H}}$ directly results in a smaller backward error. Table 2 shows the same values for $a = 10^9$ with only slight changes of the results. The dependence of the computational effort and of the backward error on the amplitude a is surprisingly weak, demonstrating the robustness of this approximate LU decomposition.

TABLE 2
$\mathcal{H}$-LU for two-dimensional diffusion with $a = 10^9$.

	$\varepsilon_{\mathcal{H}} = 1_{10}{-}2$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}4$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}6$			
n	T_{LL^T}	MB	k	$\mathcal{E}_A$	T_{LL^T}	MB	k	$\mathcal{E}_A$	T_{LL^T}	MB	k	$\mathcal{E}_A$
39 061	1.4	32	14	$1.2_{10}{-}3$	1.9	39	15	$2.9_{10}{-}5$	2.4	45	17	$2.8_{10}{-}7$
78 961	3.3	68	10	$1.5_{10}{-}3$	4.8	85	12	$3.0_{10}{-}5$	6.5	100	13	$1.4_{10}{-}7$
159 201	8.1	147	14	$1.9_{10}{-}3$	11.3	185	16	$3.2_{10}{-}5$	15.3	221	17	$3.1_{10}{-}7$
318 096	19.2	307	11	$1.5_{10}{-}3$	28.7	399	12	$2.9_{10}{-}5$	38.8	480	14	$2.5_{10}{-}7$
638 401	45.1	657	14	$1.9_{10}{-}3$	66.6	857	16	$3.2_{10}{-}5$	92.2	1046	17	$4.2_{10}{-}7$
1 276 900	109.4	1358	11	$1.1_{10}{-}3$	167.6	1827	12	$2.9_{10}{-}5$	241.5	2237	14	$2.3_{10}{-}7$
2 556 801	258.3	2858	14	$1.9_{10}{-}3$	387.5	3878	16	$3.2_{10}{-}5$	567.0	4857	17	$3.8_{10}{-}7$

5.2. Convection-diffusion problems. In the next tests operators of type

$$D = -\Delta + c \cdot \nabla$$

are considered. The convection coefficient c is randomly chosen, i.e., $c(x) \in [-a, a]^2$ for each $x \in \Omega = (0, 1)^2$. Since the standard finite element method becomes unstable for large a (i.e., the stiffness matrix may become singular), we have restricted ourselves to the cases $a = 10, 100$. In the convection dominant case the *streamline diffusion finite element method* (cf. [20]) has to be used. Note that the limiting case $a \to \infty$ is not covered by the present theory but will be treated in a forthcoming article.

Tables 3 and 4 show the same quantities as Tables 1 and 2. Note that in this case symmetry of the stiffness matrix could not be exploited.

Due to memory limitations the example with $n = 1\,276\,900$ and $\varepsilon_{\mathcal{H}} = 1_{10}{-}6$ failed to compute. Although the computational effort has increased compared with

TABLE 3
$\mathcal{H}$-LU for convection-diffusion problems with $a = 10$.

	$\varepsilon_{\mathcal{H}} = 1_{10}{-}2$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}4$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}6$			
n	T_{LU}	MB	k	$\mathcal{E}_A$	T_{LU}	MB	k	$\mathcal{E}_A$	T_{LU}	MB	k	$\mathcal{E}_A$
39 061	4.5	65	14	$1.1_{10}{-}3$	5.8	79	15	$2.9_{10}{-}5$	7.1	92	17	$2.6_{10}{-}7$
78 961	10.5	136	10	$1.5_{10}{-}3$	14.3	171	12	$3.0_{10}{-}5$	17.8	201	13	$1.4_{10}{-}7$
159 201	25.7	269	14	$1.7_{10}{-}3$	33.8	374	16	$3.2_{10}{-}5$	43.8	447	17	$3.9_{10}{-}7$
318 096	61.2	618	10	$1.5_{10}{-}3$	83.0	804	12	$2.9_{10}{-}5$	112.8	969	13	$2.3_{10}{-}7$
638 401	152.3	1319	14	$1.7_{10}{-}3$	219.4	1726	16	$3.2_{10}{-}5$	318.5	2110	17	$4.1_{10}{-}7$
1 276 900	356.9	2728	10	$2.5_{10}{-}2$	551.6	3671	12	$3.0_{10}{-}5$	–	–	–	–

TABLE 4
$\mathcal{H}$-LU for convection-diffusion problems with $a = 100$.

	$\varepsilon_{\mathcal{H}} = 1_{10}{-}2$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}4$				$\varepsilon_{\mathcal{H}} = 1_{10}{-}6$			
n	T_{LU}	MB	k	$\mathcal{E}_A$	T_{LU}	MB	k	$\mathcal{E}_A$	T_{LU}	MB	k	$\mathcal{E}_A$
39 061	4.5	65	14	$1.3_{10}{-}3$	5.8	79	16	$3.2_{10}{-}5$	7.1	92	17	$3.5_{10}{-}7$
78 961	10.4	136	10	$1.6_{10}{-}3$	14.0	171	12	$3.5_{10}{-}5$	18.0	201	13	$2.6_{10}{-}7$
159 201	25.6	296	14	$1.7_{10}{-}3$	33.6	374	16	$3.2_{10}{-}5$	43.0	447	17	$3.6_{10}{-}7$
318 096	59.3	618	10	$1.5_{10}{-}3$	84.0	803	12	$3.8_{10}{-}5$	112.3	969	13	$2.6_{10}{-}7$
638 401	152.3	1318	14	$1.8_{10}{-}3$	218.8	1726	16	$3.2_{10}{-}5$	315.4	2109	17	$3.9_{10}{-}7$
1 276 900	357.4	2726	10	$1.6_{10}{-}3$	550.2	3671	12	$3.4_{10}{-}5$	–	–	–	–

the diffusion problem, it still scales almost linearly. A dependence on the coefficient c can hardly be observed.

Table 5 shows the time T_{LU} and the amount of memory needed to compute the exact LU decomposition using MUMPS version 4.6.3; cf. [1]. Both methods seem to scale in a similar way with n while MUMPS is 5 times faster than the $\mathcal{H}$-LU factorization.

TABLE 5
Results of MUMPS for convection-diffusion problems.

n	T_{LU}	MB
39 061	0.9	31
78 961	2.2	69
159 201	5.1	151
318 096	11.5	329
638 401	29.2	712
1 276 900	72.1	1564

5.3. Three-dimensional diffusion. As we have mentioned in section 2, the structure of $\mathcal{H}$-matrices can be equally applied to any quasi-uniform finite element discretization of Ω given by just the grid information. In order to demonstrate that the $\mathcal{H}$-LU decomposition is also efficient for three-dimensional problems, we test the proposed method on three tetrahedral discretizations ($n = 25\,011$, $n = 217\,225$, and $n = 1\,848\,951$ of unknowns) of the volume shown in Figure 3. The meshes were generated using NETGEN [28].

We consider the Dirichlet boundary value problem

$$\begin{aligned} -\operatorname{div} C(x)\nabla u &= f \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

where $C(x) \in \mathbb{R}^{3\times 3}$ is a symmetric positive definite matrix for all $x \in \Omega$. The

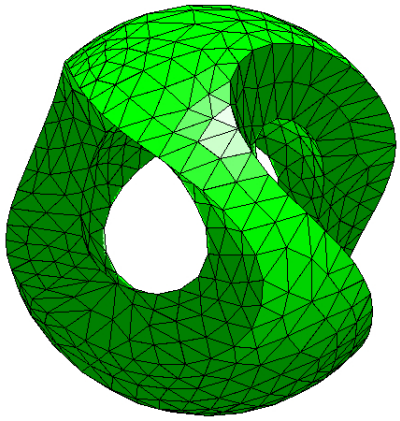


FIG. 3. *The computational domain.*

coefficients c_{ij}, $i,j = 1,2,3$, are set to one in the left half-space and to a random number from the interval $[0,1]$ in each point of the right half-space.

In Table 6 we compare the numerical effort for generating a hierarchical Cholesky decomposition. The minimal cluster size was chosen to be $n_{\min} = 50$. In the second column the time that was needed to compute the approximate Cholesky decomposition is shown for different rounding precisions $\varepsilon_{\mathcal{H}}$. The memory consumption can be found in the third column. Columns four and five contain the maximum rank among the blocks and the backward error. Compared with the two-dimensional problems from section 5.1, the CPU time for decomposing the matrix has increased. However, a behavior similar to that in the two-dimensional tests can be observed: The computational complexity scales almost linearly, and the backward error does not seem to depend on n. Again, the proposed method is able to adapt itself to the varying coefficients.

TABLE 6
Decompositions for three-dimensional diffusion.

	$n = 25\,011$				$n = 217\,225$				$n = 1\,848\,951$			
$\varepsilon_{\mathcal{H}}$	T_{LL^T}	MB	k	$\mathcal{E}_A$	T_{LL^T}	MB	k	$\mathcal{E}_A$	T_{LL^T}	MB	k	$\mathcal{E}_A$
$1_{10}{-}1$	1.7s	14	35	$3.0_{10}{-}2$	39.4s	173	35	$3.3_{10}{-}2$	869.9s	2029	187	$4.7_{10}{-}2$
$1_{10}{-}2$	2.6s	18	41	$3.3_{10}{-}3$	65.7s	261	47	$5.3_{10}{-}3$	1662.1s	3323	206	$5.9_{10}{-}3$
$1_{10}{-}3$	3.4s	21	46	$3.3_{10}{-}4$	97.2s	344	58	$5.7_{10}{-}4$	–	–	–	–
$1_{10}{-}4$	4.2s	25	47	$3.4_{10}{-}5$	137.4s	434	82	$4.5_{10}{-}5$	–	–	–	–
MUMPS	1.8s	26			85.6s	547			–	–		

The comparison with MUMPS shows that the $\mathcal{H}$-LU decomposition is as fast as this direct solver for the smallest problem. The situation changes, however, for larger problems. The $\mathcal{H}$-LU factorization becomes more efficient. MUMPS was not able to compute the problem with $n = 1\,848\,951$ unknowns on our system. Here, the higher asymptotic complexity of direct solvers is revealed. The proposed approximate LU decomposition still scales linearly and hence becomes more efficient. This is especially true for low-precision approximations which are sufficient if the LU decomposition is used for preconditioning.

Table 7 contains the time T_S required to solve a linear system by forward/backward

TABLE 7
Time for solving and solution error.

	$n = 25\,011$		$n = 217\,225$		$n = 1\,848\,951$	
$\varepsilon_{\mathcal{H}}$	T_S	$\mathcal{E}_x$	T_S	$\mathcal{E}_x$	T_S	$\mathcal{E}_x$
$1_{10}-1$	0.01s	$1.1_{10}-1$	0.20s	$4.0_{10}-1$	2.79s	$6.7_{10}-1$
$1_{10}-2$	0.02s	$1.5_{10}-2$	0.37s	$9.8_{10}-2$	3.95s	$3.2_{10}-1$
$1_{10}-3$	0.02s	$9.2_{10}-4$	0.49s	$7.7_{10}-3$	–	–
$1_{10}-4$	0.02s	$5.7_{10}-5$	0.51s	$4.9_{10}-4$	–	–
MUMPS	0.04s	$3.4_{10}-15$	0.61s	$3.6_{10}-14$	–	

substitution. The third column contains the error

$$\mathcal{E}_x := \frac{\|\tilde{x} - x\|_2}{\|x\|_2}$$

of the solution vector $\tilde{x}$ compared with the exact solution x.

6. Conclusion. We have laid theoretical ground for the approximation of the factors L and U of the LU decomposition of discrete elliptic operators by $\mathcal{H}$-matrices. Furthermore, an algorithm for the computation of an approximate LU decomposition with almost linear complexity was presented. The comparison with MUMPS shows that an improved asymptotic complexity of the approximate LU decomposition can be observed for problems posed in three spatial dimensions. The hierarchical LU decomposition is robust with respect to varying coefficients of the differential operator and is especially efficient for low-precision applications. Hence, it is well suited, for instance, for computing black-box preconditioners at least for the class of second-order elliptic operators.

REFERENCES

[1] P. R. AMESTOY, I. S. DUFF, J.-Y. L'EXCELLENT, AND J. KOSTER, *A fully asynchronous multifrontal solver using distributed dynamic scheduling*, SIAM J. Matrix Anal. Appl., 23 (2001), pp. 15–41.

[2] M. BEBENDORF, *Efficient inversion of Galerkin matrices of general second order elliptic differential operators*, Math. Comp., 74 (2005), pp. 1179–1199.

[3] M. BEBENDORF, *Hierarchical LU decomposition based preconditioners for BEM*, Computing, 74 (2005), pp. 225–247.

[4] M. BEBENDORF, *Approximate inverse preconditioning of finite element discretizations of elliptic operators with nonsmooth coefficients*, SIAM J. Matrix Anal. Appl., 27 (2006), pp. 909–929.

[5] M. BEBENDORF, *Approximation of boundary element matrices*, Numer. Math., 86 (2000), pp. 565–589.

[6] M. BEBENDORF AND S. RJASANOW, *Adaptive low-rank approximation of collocation matrices*, Computing, 70 (2003), pp. 1–24.

[7] M. BEBENDORF AND W. HACKBUSCH, *Existence of $\mathcal{H}$-matrix approximants to the inverse FE-matrix of elliptic operators with L^∞-coefficients*, Numer. Math., 95 (2003), pp. 1–28.

[8] S. CHANDRASEKARAN AND M. GU, *Fast and stable algorithms for banded plus semiseparable systems of linear equations*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 373–384.

[9] E. CUTHILL AND J. MCKEE, *Reducing the bandwidth of sparse symmetric matrices*, in Proceedings of the 24th National Conference of the ACM, 1969, pp. 157–172.

[10] J. W. DEMMEL, N. J. HIGHAM, AND R. SCHREIBER, *Stability of block LU factorization*, Numer. Linear Algebra Appl., 2 (1995), pp. 173–190.

[11] Y. EIDELMAN AND I. GOHBERG, *Inversion formulas and linear complexity algorithm for diagonal plus semiseparable matrices*, Comput. Math. Appl., 33 (1997), pp. 69–79.

[12] Y. EIDELMAN AND I. GOHBERG, *A look-ahead block Schur algorithm for diagonal plus semiseparable matrices*, Comput. Math. Appl., 35 (1998), pp. 25–34.

[13] A. GEORGE, *Nested dissection of a regular finite element mesh*, SIAM J. Numer. Anal., 10 (1973), pp. 345–363.

[14] A. George and J. Liu, *Computer Solution of Large Sparse Positive Definite Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1981.
[15] A. George and J. W. H. Liu, *The evolution of the minimum degree ordering algorithm*, SIAM Rev., 31 (1989), pp. 1–19.
[16] L. Grasedyck and W. Hackbusch, *Construction and arithmetics of $\mathcal{H}$-matrices*, Computing, 70 (2003), pp. 295–334.
[17] W. Hackbusch, *A sparse matrix arithmetic based on $\mathcal{H}$-matrices.* I. *Introduction to $\mathcal{H}$-matrices*, Computing, 62 (1999), pp. 89–108.
[18] W. Hackbusch and B. N. Khoromskij, *A sparse $\mathcal{H}$-matrix arithmetic.* II. *Application to multi-dimensional problems*, Computing, 64 (2000), pp. 21–47.
[19] W. Hackbusch and Z. P. Nowak, *On the fast matrix multiplication in the boundary element method by panel clustering*, Numer. Math., 54 (1989), pp. 463–491.
[20] T. J. R. Hughes and A. Brooks, *A multidimensional upwind scheme with no cross-wind diffusion*, in Finite Element Methods for Convection Dominated Flows, AMD 34, T. J .R. Hughes, ed., ASME, New York, 1979, pp. 19–35.
[21] P. Jones, J. Ma, and V. Rokhlin, *A fast direct algorithm for the solution of the Laplace equation on regions with fractal boundaries*, J. Comput. Phys., 113 (1994), pp. 35–51.
[22] M. Lintner, *Lösung der 2D Wellengleichung mittels hierarchischer Matrizen*, Doctoral thesis, Technische Universität München, Munich, Germany, 2002.
[23] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, *Geometric separators for finite-element meshes*, SIAM J. Sci. Comput., 19 (1998), pp. 364–386.
[24] S. Rjasanow, I. Ibraghimov, and K. Straube, *Hierarchical Cholesky decomposition of sparse matrices arising from curl-curl-equations*, J. Numer. Math., to appear.
[25] V. Rokhlin, *Rapid solution of integral equations of classical potential theory*, J. Comput. Phys., 60 (1985), pp. 187–207.
[26] D. J. Rose, *A graph-theoretic study of the numerical solution of sparse positive definite systems of linear equations*, in Graph Theory and Computing, Academic Press, New York, 1972, pp. 183–217.
[27] Y. Saad, *Iterative Methods for Sparse Linear Systems*, PWS Publishing, Boston, 1996.
[28] J. Schöberl, *NETGEN – An advancing front 2D/3D-mesh generator based on abstract rules*, Comput. Visual. Sci., 1 (1997), pp. 41–52.
[29] B. F. Smith, P. E. Bjørstad, and W. D. Gropp, *Domain Decomposition: Parallel Multilevel Methods for Elliptic Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1996.
[30] P. Starr and V. Rokhlin, *On the numerical solution of two-point boundary value problems.* II, Comm. Pure Appl. Math., 47 (1994), pp. 1117–1159.
[31] E. Tyrtyshnikov, *Mosaic-skeleton approximations*, Calcolo, 33 (1996), pp. 47–57.
[32] E. Van Camp, N. Mastronardi, and M. Van Barel, *Two fast algorithms for solving diagonal-plus-semiseparable linear systems*, J. Comput. Appl. Math., 164/165 (2004), pp. 731–747.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1495–1516

MESH INDEPENDENT SUPERLINEAR PCG RATES VIA COMPACT-EQUIVALENT OPERATORS*

OWE AXELSSON† AND JÁNOS KARÁTSON‡

Abstract. The subject of the paper is the mesh independent convergence of the preconditioned conjugate gradient (PCG) method for nonsymmetric elliptic problems. The approach of equivalent operators is involved, in which one uses the discretization of another suitable elliptic operator as preconditioning matrix. By introducing the notion of compact-equivalent operators, it is proved that for a wide class of elliptic problems the superlinear convergence of the obtained PCG method is mesh independent under finite element discretizations; that is, the rate of superlinear convergence is given in the form of a sequence which is mesh independent and is determined only by the elliptic operators.

Key words. nonsymmetric elliptic problems, conjugate gradient method, preconditioning, equivalent operators, superlinear convergence, mesh independence

AMS subject classifications. 65J10, 65F10, 65N15

DOI. 10.1137/06066391X

1. Introduction.

The conjugate gradient (CG) method is a widespread way of solving large linear algebraic systems, such as those arising from discretized elliptic problems, in particular when combined with a suitable preconditioning. For nonsymmetric systems several CG algorithms exist [2, 4], including the common CGN method based on normal equations. Since its first presentation in [21] the convergence of the CG method has been well established, as summarized in [4]. The convergence theory of the CG method often involves linear operators in Hilbert space; see both classical and recent results [14, 15, 20, 27, 31, 32] and the authors' papers [6, 7, 8, 22, 24]. A basic reason to use Hilbert space theory is to derive mesh independence of the convergence estimates, by which it can be shown that the preconditioned CG (PCG) method can be competitive with multigrid methods [14].

The theory of equivalent operators in Hilbert space has proved to provide an efficient clear framework for the convergence study of the PCG method for elliptic problems [14, 18, 26]. Thereby one uses the discretization of a suitable linear elliptic operator as preconditioning matrix; see also [10, 12, 32]. As a main result, mesh independence of linear convergence rates is rigorously characterized in [14, 26]. We note that in [18], for proper boundary conditions, when the preconditioned operator is a compact perturbation of the identity, then convergence is expected to be faster than any linear rate.

Our goal is to complete the above results on the preconditioned CGN (PCGN) method by showing that for a class of elliptic problems, the superlinear convergence of the iteration is mesh independent under finite element method (FEM) discretizations. This means that a bound on the rate of superlinear convergence is given in the form of a sequence which is mesh independent and is determined only by the

*Received by the editors June 28, 2006; accepted for publication (in revised form) February 21, 2007; published electronically August 1, 2007.
http://www.siam.org/journals/sinum/45-4/66391.html

†Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden, and Institute of Geonics AS CR, 708 00 Ostrava, Czech Republic (owea@it.uu.se).

‡Department of Applied Analysis, ELTE University, H-1117 Budapest, Hungary (karatson@cs.elte.hu). This author was supported by NKTH Öveges Program and Hungarian Research grant OTKA T 043765.

elliptic operators. To describe the suitable class of problems, we introduce the notion of compact-equivalent operators, which expresses that preconditioning one with the other yields a compact perturbation of the identity. This notion and the convergence result give a refinement of the case of equivalent operators: roughly speaking, if the two operators (the original and preconditioner) are equivalent, then the corresponding PCG method provides mesh independent linear convergence, whereas if the two operators are compact-equivalent, then the PCG method provides mesh independent superlinear convergence.

Our present results are extensions of the earlier ones [8, 24], where such mesh independence was proved for the generalized conjugate gradient-least squares (GCG-LS) method for elliptic Dirichlet problems, but with severe restrictions: except for some special cases, both the original and preconditioning operators had to contain constant coefficients. Now we show that two elliptic operators, with homogeneous Dirichlet conditions on the same portion of the boundary, are compact-equivalent if and only if their principal parts coincide up to a constant factor. Within this class, the proof of the mesh independence result then contains no restrictions except standard smoothness and coercivity assumptions on the operators.

Our characterization of compact-equivalence provides, in fact, a limitation on the scope of the mesh independent superlinear convergence property. Since the principal parts of compact-equivalent operators must coincide (up to a constant), preconditioning methods like replacing rough diffusion coefficients by simpler, e.g., constant, ones are not covered by this setting except, of course, the case when the variable coefficient problem can be easily rewritten by suitable scaling to a constant coefficient problem, as for scalar coefficients. In fact, one cannot expect superlinear convergence for such non–compact-equivalent operators since, as shown in [15], convergence of the CG method may be only linear if an operator is not a compact perturbation of a constant times the identity.

The paper is organized as follows: the required background is given in section 2, compact-equivalent operators are introduced and characterized in section 3, and the mesh independence result is proved in section 4. Some closing remarks are found in section 5.

2. Background.

2.1. Conjugate gradient algorithms.

Let us consider a linear system

$$Bu = f \tag{1}$$

with a given nonsingular matrix $B \in \mathbf{R}^{n\times n}$, $f \in \mathbf{R}^n$ and solution u. Let $\langle .,. \rangle$ be a given inner product on $\mathbf{R}^n$ and, denoting by B^* the adjoint of B w.r.t. this inner product, assume that $B + B^* > 0$, i.e., is positive definite.

If B is self-adjoint, then the standard CG method reads as follows [4, 31]: let $u_0 \in \mathbf{R}^n$ be arbitrary, $d_0 := Bu_0 - f$; for given u_k and d_k, with $\hat{r}_k := Bu_k - f$, we let

$$u_{k+1} = u_k - \alpha_k d_k, \text{ where } \alpha_k = \frac{\langle \hat{r}_k, d_k \rangle}{\langle Bd_k, d_k \rangle}; \quad d_{k+1} = \hat{r}_{k+1} + \beta_k d_k, \text{ where } \beta_k = \frac{\|\hat{r}_{k+1}\|^2}{\|\hat{r}_k\|^2}. \tag{2}$$

Then, using the error vector $e_k = u_k - u$ and its energy norm $\|e_k\|_B = \langle Be_k, e_k \rangle^{1/2}$, respectively, and with the decomposition $B = I + C$ (where I is the identity matrix),

the following celebrated estimate holds [4, 31]:

$$\left(\frac{\|e_k\|_B}{\|e_0\|_B}\right)^{1/k} \le \frac{2}{k}\,\|B^{-1}\|\sum_{j=1}^{k}|\lambda_j(C)| \qquad (k=1,2,\dots,n), \tag{3}$$

which shows superlinear convergence if the eigenvalues $|\lambda_1(C)| \ge |\lambda_2(C)| \ge \cdots$ approach zero.

Since this result is basic for the whole paper, and for completeness, we present a derivation of (3) following [4]. The optimality of the CG method implies

$$\frac{\|e_k\|_B}{\|e_0\|_B} \le \min_{P_k\in\pi_k^1}\max_{\lambda\in\sigma(B)\}}|P_k(\lambda)|,$$

where π_k^1 denotes the set of polynomials P_k of degree k with $P_k(0)=1$. Let $\lambda_j := \lambda_j(B)$ and $\mu_j := \lambda_j(C)\,(=\lambda_j-1)$. Then the polynomials $P_k(\lambda) := \prod_{j=1}^{k}(1-\frac{\lambda}{\lambda_j})$ satisfy $P_k(\lambda_i)=0$ $(i=1,\dots,k)$ and

$$\max_{\lambda\in\sigma(B)\}}|P_k(\lambda)| = \max_{i\ge k+1}\prod_{j=1}^{k}\left|1-\frac{\lambda_i}{\lambda_j}\right| = \max_{i\ge k+1}\prod_{j=1}^{k}\frac{|\mu_j-\mu_i|}{|1+\mu_j|} \le \max_{i\ge k+1}\prod_{j=1}^{k}\frac{2|\mu_j|}{|1+\mu_j|}\,;$$

hence, using the arithmetic-geometric inequality,

$$\max_{\lambda\in\sigma(B)\}}|P_k(\lambda)|^{1/k} \le \frac{2}{k}\sum_{j=1}^{k}\frac{|\mu_j|}{|1+\mu_j|} \le \frac{2}{k}\left(\sup\frac{1}{|\lambda_j|}\right)\sum_{j=1}^{k}|\mu_j|,$$

which yields (3).

For nonsymmetric B, several CG algorithms exist (see, e.g., [2, 4, 13]). The GCG-LS method [3, 4] is defined directly for (1) and produces an estimate similar to that of (3) if B is normal. Mesh independent bounds in [8, 24] for (3) for some elliptic problems have been given using the GCG-LS method. Alternatively, one can consider the normal equation and apply a symmetric CG algorithm, which we will do in this paper. For clearness, let us hereby consider a nonsymmetric linear system

$$Au = b \tag{4}$$

with a given nonsingular matrix $A\in\mathbf{R}^{n\times n}$ and vector $b\in\mathbf{R}^n$. Let us apply the iteration (2) for the equation $A^*Au = A^*b$, i.e., with $B=A^*A$ and $f=A^*b$. Then, with notations $s_k=\hat r_k$ and $r_k = A^{-*}\hat r_k$, we obtain the following algorithmic form, often called the CGN method: let $u_0\in\mathbf{R}^n$ be arbitrary, $r_0 := Au_0-b$, $s_0 := d_0 := A^*r_0$; for given d_k, u_k, r_k, and s_k, we let

$$\begin{cases} z_k = Ad_k, \\ \alpha_k = \dfrac{\langle r_k, z_k\rangle}{\|z_k\|^2}\,, \quad u_{k+1} = u_k - \alpha_k d_k\,, \quad r_{k+1} = r_k - \alpha_k z_k\,; \\ s_{k+1} = A^*r_{k+1}, \\ \beta_k = \dfrac{\|s_{k+1}\|^2}{\|s_k\|^2}\,, \quad d_{k+1} = s_{k+1} + \beta_k d_k. \end{cases} \tag{5}$$

Let us consider the decomposition

$$A = I + K.$$

Then, using the relations $B = I + (K^* + K + K^*K)$, $\|e_k\|_B = \|Ae_k\| = \|r_k\|$, and $\|B^{-1}\| \le \nu^{-1}$, where $\nu := \min_{x\in\mathbf{R}^n} \frac{\|Ax\|^2}{\|x\|^2}$, estimate (3) can be reformulated as

$$\left(\frac{\|r_k\|}{\|r_0\|}\right)^{1/k} \le \frac{2}{k\nu} \sum_{i=1}^{k} \Big(\big|\lambda_i(K^*+K)\big| + \lambda_i(K^*K) \Big) \qquad (k = 1, 2, \dots, n). \tag{6}$$

The goal of this paper is to derive a mesh independent bound for (6) when (4) comes from a preconditioned discretized elliptic PDE using suitable equivalent operators.

2.2. Singular values of compact operators. Let H be a real Hilbert space. We shall consider compact operators, i.e., operators C such that the image (Cv_i) of any bounded sequence (v_i) contains a convergent subsequence.

DEFINITION 2.1. (i) We call $\lambda_i(F)$ $(i = 1, 2, \dots)$ the *ordered eigenvalues* of a compact self-adjoint linear operator F in H if each of them is repeated as many times as its multiplicity and $|\lambda_1(F)| \ge |\lambda_2(F)| \ge \cdots$.

(ii) The *singular values* of a compact operator C in H are

$$s_i(C) := \lambda_i(C^*C)^{1/2} \qquad (i = 1, 2, \dots),$$

where $\lambda_i(C^*C)$ are the ordered eigenvalues of C^*C. In particular, if C is self-adjoint, then $s_i(C) = |\lambda_i(C)|$.

Some useful properties of compact operators are listed below.

PROPOSITION 2.2. *Let C be a compact operator in H. Then the following properties hold.*

(a) *For any $k \in \mathbf{N}^+$ and any orthonormal vectors $u_1, \dots, u_k \in H$,*

$$\sum_{i=1}^{k} \big|\langle Cu_i, u_i\rangle\big| \le \sum_{i=1}^{k} s_i(C)\,.$$

(b) *If B is a bounded linear operator in H, then*

$$s_i(BC) \le \|B\|\, s_i(C) \qquad (i = 1, 2, \dots).$$

(c) *(Variational characterization of the eigenvalues.) If C is also self-adjoint, then*

$$\big|\lambda_i(C)\big| = \min_{H_{i-1}\subset H} \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{\big|\langle Cu, u\rangle\big|}{\|u\|^2}\,,$$

where H_{i-1} stands for an arbitrary $(i-1)$-dimensional subspace.

(d) *If a sequence $(u_i) \subset H$ satisfies $\langle u_i, u_j\rangle = \langle Cu_i, u_j\rangle = 0$ $(i \neq j)$, then*

$$\inf_i\, |\langle Cu_i, u_i\rangle| / \|u_i\|^2 = 0.$$

Proof. Statements (a) and (b) are the consequences of [16, Chap. VI, Corollary 3.3 and Proposition 1.3, resp.]; for statement (c), see [17, Theorem III.9.1]. To prove (d), assume the contrary that the infimum equals $\delta > 0$. We may assume that $\langle Cu_i, u_i\rangle$ has constant sign (otherwise we can consider a subsequence that has constant sign). Then the orthonormal sequence $v_i := u_i/\|u_i\|$ satisfies for all $i \neq j$

$$\begin{aligned} 2\delta &\le |\langle Cv_i, v_i\rangle + \langle Cv_j, v_j\rangle| = |\langle C(v_i - v_j), v_i - v_j\rangle| \\ &\le \|C(v_i - v_j)\|\,\|v_i - v_j\| = \sqrt{2}\|C(v_i - v_j)\|; \end{aligned}$$

hence the image (Cv_i) of the bounded sequence (v_i) contains no convergent subsequence (i.e., C is not compact). ∎

3. Compact-equivalent operators in Hilbert space. In this section we introduce and characterize compact-equivalent operators. Roughly speaking, the compact-equivalence of the unbounded operators N and L expresses that $N^{-1}L$ is a compact perturbation of a constant times the identity. To avoid difficulties with domains and ranges, our definition will use a weak form of the operators in a suitable energy space H_S. In particular, no regularity is required in the case of elliptic operators.

The fact that a compact perturbation of a constant times identity is a bounded operator implies that compact-equivalent operators are equivalent in the sense of [14]. Hence, when we characterize compact-equivalent elliptic operators (under standard smoothness and coercivity assumptions), we can a priori assume that they have homogeneous Dirichlet conditions on the same portion of the boundary [26]. Within this class, compact-equivalence will hold if and only if the principal parts of the operators coincide up to some constant.

3.1. Basic definitions. In what follows, let H be a real Hilbert space. Let S be a (generally unbounded) linear symmetric operator in H which is coercive; i.e., there exists $p > 0$ such that $\langle Su, u\rangle \ge p\|u\|^2$ $(u \in D(S))$. Then the energy space H_S is the completion of $D(S)$ under the inner product $\langle u, v\rangle_S = \langle Su, v\rangle$, and the coercivity implies $H_S \subset H$. The corresponding S-norm is denoted by $\|u\|_S$, and the space of bounded linear operators on H_S by $B(H_S)$.

DEFINITION 3.1. *Let S be a linear symmetric coercive operator in H. We say that a linear operator L in H is S-bounded and S-coercive, and write $L \in BC_S(H)$ if the following properties hold:*

(i) *$D(L) \subset H_S$ and $D(L)$ is dense in H_S in the S-norm;*

(ii) *there exists $M > 0$ such that*

$$|\langle Lu, v\rangle| \le M\|u\|_S\|v\|_S \qquad (u, v \in D(L));$$

(iii) *there exists $m > 0$ such that*

$$\langle Lu, u\rangle \ge m\|u\|_S^2 \qquad (u \in D(L)).$$

DEFINITION 3.2. *For any $L \in BC_S(H)$, let $L_S \in B(H_S)$ be defined by*

$$\langle L_S u, v\rangle_S = \langle Lu, v\rangle \qquad (u, v \in D(L)).$$

Remark 1.

(a) The above definition makes sense since L_S is the bounded linear operator on H_S that represents the unique extension to H_S of the densely defined S-bounded bilinear form $u, v \mapsto \langle Lu, v\rangle$.

(b) L_S is coercive on H_S.
(c) If in particular $R(L) \subset R(S)$ (where $R(.)$ denotes the range), then $L_S\big|_{D(L)} = S^{-1}L$.

Remark 2. Definition 3.2 uses the idea of weak form of operators from [26]. Namely, if H_S is a subspace of $H^1(\Omega)$ consisting of functions vanishing on a fixed portion of the boundary, then L_S coincides with the weak operator L_w using (2.15) in [26].

Now let us consider an operator equation

$$Lu = g, \tag{7}$$

where $L \in BC_S(H)$ and $g \in H$.

DEFINITION 3.3. *We call $u \in H_S$ the weak solution of equation* (7) *if*

$$\langle L_S u, v\rangle_S = \langle g, v\rangle \qquad (v \in H_S). \tag{8}$$

Remark 3.
(a) For all $g \in H$ the weak solution of (7) exists and is unique. This follows in the usual way from the Lax–Milgram theorem, since $v \mapsto \langle g, v\rangle$ is a bounded linear functional on H_S by the coercivity of S.
(b) If $u \in D(L)$, then u satisfies (7) (i.e., it is a strong solution) if and only if u is a weak solution.

3.2. Compact-equivalent operators. We can introduce the notion of compact-equivalence within the previously described setting as follows.

DEFINITION 3.4. *Let L and N be S-bounded and S-coercive operators in H. We call L and N compact-equivalent in H_S if*

$$L_S = \mu N_S + Q_S \tag{9}$$

for some constant $\mu > 0$ and compact operator $Q_S \in B(H_S)$.

Remark 4. (i) It follows in a straightforward way that the property compact-equivalence is an equivalence relation.

(ii) In the special case $R(L) \subset R(N)$, compact-equivalence of L and N means that $N^{-1}L$ is a compact perturbation of a constant times the identity in the space H_S. Indeed, it is easy to see that here $N^{-1}L = N_S^{-1}L_S\big|_{D(L)}$, and by definition the latter is the perturbation of μI with the operator $N_S^{-1}Q_S\big|_{D(L)}$, which is compact since N_S^{-1} is bounded. (In the general case the "weakly preconditioned" form $N_S^{-1}L_S$ is also a compact perturbation.)

Now we characterize compact-equivalence for elliptic operators. Let $H = L^2(\Omega)$ and let us define the operators

$$N_1 u \equiv -\mathrm{div}\,(A_1 \nabla u) + \mathbf{b}_1 \cdot \nabla u + c_1 u \qquad \text{for } u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_{A_1}} + \alpha_1 u_{|\Gamma_N} = 0,$$
$$N_2 u \equiv -\mathrm{div}\,(A_2 \nabla u) + \mathbf{b}_2 \cdot \nabla u + c_2 u \qquad \text{for } u_{|\Gamma_D} = 0,\ \tfrac{\partial u}{\partial \nu_{A_2}} + \alpha_2 u_{|\Gamma_N} = 0,$$

where $\frac{\partial u}{\partial \nu_{A_i}} = A_i\,\nu \cdot \nabla u$ denotes the weighted normal derivative. (The formal domain of N_i to be used in Definition 3.2 consists of those $u \in H^2(\Omega)$ that satisfy the above boundary conditions; however, this is used nowhere else.) The following properties hold, where $i = 1, 2$.

ASSUMPTIONS 3.2.

(i) $\Omega \subset \mathbf{R}^d$ is a bounded piecewise C^1 domain; Γ_D, Γ_N are disjoint open measurable subparts of $\partial\Omega$ such that $\partial\Omega = \overline{\Gamma}_D \cup \overline{\Gamma}_N$.

(ii) $A_i \in C^1(\overline{\Omega}, \mathbf{R}^{d\times d})$ and for all $x \in \overline{\Omega}$ the matrix $A_i(x)$ is symmetric; $\mathbf{b}_i \in C^1(\overline{\Omega})^d$, $c_i \in L^\infty(\Omega)$, $\alpha_i \in L^\infty(\Gamma_N)$.

(iii) We have the coercivity properties $\min_{\lambda\in\sigma(A_i(x))} \lambda \geq p > 0$ with p independent of x, $\hat{c}_i := c_i - \frac{1}{2}\operatorname{div}\mathbf{b}_i \geq 0$ in Ω and $\hat{\alpha}_i := \alpha_i + \frac{1}{2}(\mathbf{b}_i \cdot \nu) \geq 0$ on Γ_N.

(iv) Either $\Gamma_D \neq \emptyset$, or $\hat{c}_i$ or $\hat{\alpha}_i$ has a positive lower bound.

For the study of such operators we define the space

$$H^1_D(\Omega) := \{u \in H^1(\Omega) : u_{|\Gamma_D} = 0\} \quad \text{with } \langle u, v\rangle_S := \int_\Omega (G\,\nabla u\cdot\nabla v + huv) + \int_{\Gamma_N} \beta uv\, d\sigma, \tag{10}$$

where G has the same properties as A_i above in (ii)–(iii), and $h \in L^\infty(\Omega)$, $h \geq 0$, if $\Gamma_D \neq \emptyset$ and $h \geq \delta_0 > 0$ if $\Gamma_D = \emptyset$, and further, $\beta \in L^\infty(\Gamma_N)$ and $\beta \geq 0$. Then $H^1_D(\Omega)$ is the energy space H_S of the operator $Su := -\operatorname{div}(G\nabla u) + hu$ on $D(S) := \{u \in H^2(\Omega) : u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_{G+\beta u}}{}_{|\Gamma_N} = 0\}$. It is easy to check the properties in Definition 3.1 from the above assumptions, which means that $N_1, N_2 \in BC_S(L^2(\Omega))$.

PROPOSITION 3.5. *The elliptic operators N_1 and N_2 are compact-equivalent in $H^1_D(\Omega)$ if and only if their principal parts coincide up to some constant $\mu > 0$, i.e., $A_1 = \mu A_2$.*

Proof. We have for all $u, v \in H^1_D(\Omega)$

$$\langle (N_i)_S u, v\rangle_S = \int_\Omega \Big(A_i\ \nabla u \cdot \nabla v + (\mathbf{b}_i \cdot \nabla u)v + c_i uv\Big)\, dx + \int_{\Gamma_N} \alpha_i uv\, d\sigma\,.$$

Hence

$$(N_1)_S - \mu(N_2)_S = J_S + Q_S,$$

where, using notations $\mathbf{b} := \mathbf{b}_1 - \mu\mathbf{b}_2$, $c := c_1 - \mu c_2$, and $\alpha := \alpha_1 - \mu\alpha_2$, we have

$$\begin{aligned} \langle J_S u, v\rangle_S &= \int_\Omega (A_1 - \mu A_2)\ \nabla u \cdot \nabla v\, dx\,, \\ \langle Q_S u, v\rangle_S &= \int_\Omega \Big((\mathbf{b}\cdot\nabla u)v + cuv\Big)\, dx + \int_{\Gamma_N} \alpha uv\, d\sigma\,. \end{aligned} \tag{11}$$

Here Q_S is compact, which is known [18] when N_1 and N_2 have the same boundary conditions. Otherwise we use the equality

$$\begin{aligned} \int_\Omega (\mathbf{b}\cdot\nabla u)v\, dx &= -\int_\Omega u(\mathbf{b}\cdot\nabla v)\, dx - \int_\Omega (\operatorname{div}\mathbf{b})uv\, dx \\ &\quad + \int_{\Gamma_N} (\mathbf{b}\cdot\nu)\, uv\, d\sigma \quad (u, v \in H^1_D(\Omega)) \end{aligned} \tag{12}$$

whence, using notations $\tilde{c} := c - \operatorname{div}\mathbf{b}$ and $\tilde{\alpha} := \alpha + \mathbf{b}\cdot\nu$,

$$\|Q_S u\|_S = \sup_{\substack{v\in H^1_D(\Omega)\\ \|v\|_S=1}} |\langle Q_S u, v\rangle_S| = \sup_{\substack{v\in H^1_D(\Omega)\\ \|v\|_S=1}} \left| -\int_\Omega u(\mathbf{b}\cdot\nabla v)\, dx + \int_\Omega \tilde{c}uv\, dx + \int_{\Gamma_N} \tilde{\alpha}\, uv\, d\sigma \right|.$$

Using the embedding estimates

$$\|v\|_{L^2(\Omega)} \le C_\Omega \|v\|_S\,, \qquad \|v\|_{L^2(\Gamma_N)} \le C_{\Gamma_N}\|v\|_S \qquad (v \in H^1_D(\Omega)) \tag{13}$$

(where $C_\Omega, C_{\Gamma_N} > 0$) and $\|\nabla v\|_{L^2(\Omega)} \le p^{-1/2}\|v\|_S$, and letting $K_1 := p^{-1/2}\|\mathbf{b}\|_{L^\infty(\Omega)} + C_\Omega\|\tilde{c}\|_{L^\infty(\Omega)}$, $K_2 := C_{\Gamma_N}\|\tilde{\alpha}\|_{L^\infty(\Gamma_N)}$, we obtain

$$\|Q_S u\|_S \le K_1\|u\|_{L^2(\Omega)} + K_2\|u\|_{L^2(\Gamma_N)} \tag{14}$$

whence Q_S is compact.

It remains to prove that if $A_1 \ne \mu A_2$, then J_S is not compact. Using Proposition 2.2(d), it suffices to find a sequence $(u_i) \subset H^1_0(\Omega) \subset H^1_D(\Omega)$ satisfying

$$\langle u_i, u_j\rangle_S = \langle J_S u_i, u_j\rangle_S = 0 \qquad (i \ne j), \tag{15}$$

$$\inf_i\, |\langle J_S u_i, u_i\rangle_S|/\|u_i\|_S^2 = \delta > 0. \tag{16}$$

Let $A := A_1 - \mu A_2$. Since A is not identically zero, there is $x_0 \in \Omega$ such that $A_0 := A(x_0) \ne 0$. Here A_0 is symmetric; hence there is $u_0 \in H^1_0(\Omega)$ such that $\int_\Omega A_0\ \nabla u_0 \cdot \nabla u_0 \ne 0$. Let

$$\varepsilon := \left|\int_\Omega A_0\ \nabla u_0 \cdot \nabla u_0\right| \Big/ \left(\int_\Omega |\nabla u_0|^2\right), \qquad \Omega_{\varepsilon/2} := \{x \in \Omega :\ \|A(x) - A_0\| < \varepsilon/2\},$$

which is an open set since A is continuous. Fix $z' \in \Omega$, and for any $z \in \Omega$ and $R > 0$ let $\Omega_{z,R} := \{x \in \mathbf{R}^d :\ z' + R(x - z) \in \Omega\}$. Let $z_i \in \Omega$, $R_i > 0$ $(i \in \mathbf{N}^+)$ such that $\Omega_i := \Omega_{z_i,R_i} \subset \Omega_{\varepsilon/2}$ and $\overline{\Omega}_i$ are pairwise disjoint sets. We define $u_i \in H^1_0(\Omega)$ by $u_i(x) := u_0\big(z' + R_i(x - z_i)\big)$ for $x \in \Omega_i$ and $u_i(x) := 0$ for $x \in \Omega \setminus \Omega_i$. Since $\operatorname{supp} u_i = \overline{\Omega}_i$ are disjoint, (15) is satisfied. Further, using the fact $\Omega_i \subset \Omega_{\varepsilon/2}$ and a linear transformation $\Omega_i \to \Omega$ in the integral, we obtain

$$\begin{aligned}\frac{|\langle J_S u_i, u_i\rangle_S|}{\int_{\Omega_i}|\nabla u_i|^2} &= \frac{\left|\int_{\Omega_i} A\ \nabla u_i \cdot \nabla u_i\right|}{\int_{\Omega_i}|\nabla u_i|^2} \ge \frac{\left|\int_{\Omega_i} A_0\ \nabla u_i \cdot \nabla u_i\right|}{\int_{\Omega_i}|\nabla u_i|^2} - \frac{\varepsilon}{2}\\ &= \frac{\left|\int_\Omega A_0\ \nabla u_0 \cdot \nabla u_0\right|}{\int_\Omega |\nabla u_0|^2} - \frac{\varepsilon}{2} = \frac{\varepsilon}{2}.\end{aligned}$$

Since for $u \in H^1_0(\Omega)$ we have $\|u\|_S^2 \le C \cdot \int_\Omega |\nabla u|^2$, the above estimate yields (16) with $\delta = \frac{\varepsilon}{2C} > 0$. ∎

4. Compact-equivalent preconditioning and mesh independent superlinear convergence rates. We prove the mesh independent convergence results for the PCG method in four stages. First we consider symmetric preconditioning operators, which are more straightforward to handle. Then, by suitable modifications of the proof, we turn to arbitrary preconditioning operators (in the studied coercive framework) where the general result is obtained. In both the symmetric and nonsymmetric cases we first consider an abstract Hilbert space level and then derive the corresponding estimates for elliptic problems.

For simplicity we will consider compact-equivalence with $\mu = 1$ in (9), which is clearly no restriction, since if a preconditioner N_S satisfies $L_S = \mu N_S + Q_S$, then we can consider the preconditioner μN_S instead.

4.1. The abstract operator equation and its discretization. Let us consider the operator equation

$$Lu = g, \tag{17}$$

where $L \in BC_S(H)$ and $g \in H$, and let $u \in H_S$ be the weak solution as in Definition 3.3. Equation (17) will be solved numerically using a Galerkin discretization: let

$$V_h = span\{\varphi_1, \ldots, \varphi_n\} \subset H_S\,,$$

where φ_i are linearly independent, be a given finite-dimensional subspace and let

$$\mathbf{L}_h := \Big\{ \langle L_S \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n}.$$

Finding the discrete solution $u_h \in V_h$ requires solving the $n \times n$ system

$$\mathbf{L}_h\, \mathbf{c} = \mathbf{b} \tag{18}$$

with $\mathbf{b} = \{\langle g, \varphi_j \rangle\}_{j=1}^{n}$. Since $L \in BC_S(H)$, the symmetric part of $\mathbf{L}_h$ is positive definite; hence system (18) has a unique solution. Moreover, if a sequence of such subspaces V_h satisfies $\inf_{v \in V_h} \|u - v\|_S \to 0$ for all $u \in H_S$, then the coercivity of L_S implies in the standard way [9] that u_h converges to the exact weak solution in the H_S-norm.

4.2. Symmetric preconditioning in Hilbert space. We introduce the stiffness matrix of S,

$$\mathbf{S}_h = \Big\{ \langle \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n}, \tag{19}$$

as preconditioner for system (18), and wish to solve

$$\mathbf{S}_h^{-1} \mathbf{L}_h\, \mathbf{c} = \tilde{\mathbf{b}} \tag{20}$$

(with $\tilde{\mathbf{b}} = \mathbf{S}_h^{-1}\mathbf{b}$) using the CG method . Let us endow $\mathbf{R}^n$ with the $\mathbf{S}_h$-inner product $\langle \mathbf{c}, \mathbf{d} \rangle_{\mathbf{S}_h} := \mathbf{S}_h\, \mathbf{c} \cdot \mathbf{d}$. Then the $\mathbf{S}_h$-adjoint of $\mathbf{S}_h^{-1}\mathbf{L}_h$ is $\mathbf{S}_h^{-1}\mathbf{L}_h^T$; hence we apply the CG algorithm (5) with $A = \mathbf{S}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T$.

Let us now assume that L and S are compact-equivalent with $\mu = 1$. In this special case (9) holds with $N_S = I$:

$$L_S = I + Q_S\,. \tag{21}$$

Hence, letting

$$\mathbf{Q}_h = \Big\{ \langle Q_S \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n},$$

system (20) takes the form

$$(\mathbf{I}_h + \mathbf{S}_h^{-1}\mathbf{Q}_h)\, \mathbf{c} = \tilde{\mathbf{b}}, \tag{22}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix. Using (6), the CG algorithm (5) thus provides the estimate

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \le \frac{2}{k\nu_h} \sum_{i=1}^{k} \Big(\big|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T + \mathbf{S}_h^{-1}\mathbf{Q}_h)\big| + \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\mathbf{S}_h^{-1}\mathbf{Q}_h) \Big) \tag{23}$$

$(k = 1, 2, \dots, n)$, where

$$\nu_h = \min_{\substack{\mathbf{c}\in\mathbf{R}^n \\ \mathbf{c}\neq\mathbf{0}}} \frac{\|\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c}\|_{\mathbf{S}_h}^2}{\|\mathbf{c}\|_{\mathbf{S}_h}^2}\,. \tag{24}$$

Our goal is to give a bound on (23) that is independent of the subspace V_h.

PROPOSITION 4.1. *Let L be S-bounded and S-coercive. Let $\mathbf{S}_h$, $\mathbf{Q}_h$ be defined as above and let $s_i(Q_S)$ and $\lambda_i(Q_S^* + Q_S)$ $(i = 1, 2, \dots)$ denote the corresponding singular values, respectively, ordered eigenvalues where Q_S, defined in* (21)*, is compact on H_S. Then the following relations hold:*

(a) $$\sum_{i=1}^{k} \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{S}_h^{-1}\mathbf{Q}_h) \le \sum_{i=1}^{k} s_i(Q_S)^2 \qquad (k = 1, \dots, n),$$

(b) $$\sum_{i=1}^{k} \big|\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T + \mathbf{S}_h^{-1}\mathbf{Q}_h)\big| \le \sum_{i=1}^{k} \big|\lambda_i(Q_S^* + Q_S)\big| \qquad (k = 1, \dots, n),$$

(c) $$\nu_h \ge m^2 \quad \textit{for } \nu_h \textit{ in } (24), \quad \textit{where} \quad m := \inf_{\substack{u\in D(L) \\ u\neq 0}} \frac{\langle Lu, u\rangle}{\|u\|_S^2}\,.$$

Proof. (a) Let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{S}_h^{-1}\mathbf{Q}_h)$ $(i = 1, \dots, n)$ and let $\mathbf{c}^i = (c_1^i, \dots, c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors such that

$$\mathbf{S}_h\,\mathbf{c}^i \cdot \mathbf{c}^l = \delta_{il} \qquad (i, l = 1, \dots, n), \tag{25}$$

where $\cdot$ denotes the ordinary inner product on $\mathbf{R}^n$. Then

$$\mathbf{S}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{Q}_h\,\mathbf{c}^i = \lambda_i \qquad (i = 1, \dots, n). \tag{26}$$

Let $\mathbf{d}^i := \mathbf{S}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i$ for all i; that is,

$$\mathbf{S}_h\,\mathbf{d}^i = \mathbf{Q}_h\,\mathbf{c}^i, \tag{27}$$

which turns (26) into

$$\mathbf{S}_h\mathbf{d}^i \cdot \mathbf{d}^i = \lambda_i\,. \tag{28}$$

Now let $u_i = \sum_{j=1}^n c_j^i\varphi_j \in V_h$ and $z_i = \sum_{j=1}^n d_j^i\varphi_j \in V_h$ $(i = 1, \dots, n)$. Then (28) yields

$$\|z_i\|_S^2 = \lambda_i\,. \tag{29}$$

Further, for all $v = \sum_{j=1}^n p_j\varphi_j \in V_h$, with notation $\mathbf{p} = (p_1, \dots, p_n) \in \mathbf{R}^n$, (27) yields $\mathbf{S}_h\,\mathbf{d}^i \cdot \mathbf{p} = \mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{p}$, which implies

$$\langle z_i, v\rangle_S = \langle Q_S u_i, v\rangle_S \qquad (v \in V_h);$$

i.e., z_i is the orthogonal projection of $Q_S u_i \in H_S$ into V_h. Therefore $\|z_i\|_S \le \|Q_S u_i\|_S$, and (29) provides

$$\sum_{i=1}^{k} \lambda_i \le \sum_{i=1}^{k} \|Q_S u_i\|_S^2 = \sum_{i=1}^{k} \langle Q_S^* Q_S u_i, u_i\rangle_S. \tag{30}$$

Here $\langle u_i, u_l\rangle_S = \mathbf{S}_h\,\mathbf{c}^i \cdot \mathbf{c}^l$ for all $i,l = 1,\dots,n$; hence by (25) the vectors u_i are orthonormal in H_S. Therefore Proposition 2.2(a) for the operator $C = Q_S^* Q_S$ in the space H_S yields the desired estimate.

(b) The proof is similar to that of (a). Now let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T + \mathbf{S}_h^{-1}\mathbf{Q}_h)$ and let $\mathbf{c}^i = (c_1^i,\dots,c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors with property (25). Then

$$(\mathbf{Q}_h^T + \mathbf{Q}_h)\,\mathbf{c}^i = \lambda_i\,\mathbf{S}_h\,\mathbf{c}^i \qquad (i = 1,\dots,n),$$

and (25) yields

$$\lambda_i \;=\; (\mathbf{Q}_h^T + \mathbf{Q}_h)\,\mathbf{c}^i \cdot \mathbf{c}^i = 2\,\mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{c}^i.$$

For $u_i = \sum_{j=1}^n c_j^i \varphi_j \in V_h$ we thus obtain

$$\sum_{i=1}^k \big|\lambda_i\big| \;= 2\sum_{i=1}^k \big|\langle Q_S u_i, u_i\rangle_S\big| = \sum_{i=1}^k \big|\langle (Q_S^* + Q_S)u_i, u_i\rangle_S\big|, \tag{31}$$

and Proposition 2.2(a) for the operator $C = Q_S^* + Q_S$ in the space H_S yields the desired estimate.

(c) We have

$$\min_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \frac{\|\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c}\|_{\mathbf{S}_h}}{\|\mathbf{c}\|_{\mathbf{S}_h}} = \min_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \frac{\|\mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c}\|_{\mathbf{S}_h}\|\mathbf{c}\|_{\mathbf{S}_h}}{\|\mathbf{c}\|^2_{\mathbf{S}_h}} \geq \min_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \frac{\langle \mathbf{S}_h^{-1}\mathbf{L}_h\mathbf{c},\mathbf{c}\rangle_{\mathbf{S}_h}}{\|\mathbf{c}\|^2_{\mathbf{S}_h}}$$

$$= \min_{\substack{\mathbf{c}\in\mathbf{R}^n\\ \mathbf{c}\neq\mathbf{0}}} \frac{\mathbf{L}_h\,\mathbf{c}\cdot\mathbf{c}}{\mathbf{S}_h\,\mathbf{c}\cdot\mathbf{c}} = \min_{\substack{u\in V_h\\ u\neq 0}} \frac{\langle L_S u, u\rangle_S}{\|u\|_S^2} \geq \inf_{\substack{u\in H_S\\ u\neq 0}} \frac{\langle L_S u, u\rangle_S}{\|u\|_S^2}$$

$$= \inf_{\substack{u\in D(L)\\ u\neq 0}} \frac{\langle L_S u, u\rangle_S}{\|u\|_S^2} = \inf_{\substack{u\in D(L)\\ u\neq 0}} \frac{\langle Lu, u\rangle}{\|u\|_S^2} = m,$$

where the density of $D(L)$ in H_S has been used. □

In virtue of (23) and Proposition 4.1, we have proved the following theorem.

THEOREM 4.2. *Let L be S-bounded and S-coercive, and let L and S be compact-equivalent with $\mu = 1$. Let the compact operator Q_S be as in* (21). *Then for any subspace $V_h = span\{\varphi_1,\dots,\varphi_n\} \subset H_S$, the CG algorithm* (5) *with $\mathbf{S}_h$-inner product, applied for the $n \times n$ preconditioned system* (20), *yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \leq\; \varepsilon_k \qquad (k = 1,2,\dots,n), \tag{32}$$

$$\textit{where}\quad \varepsilon_k = \frac{2}{km^2}\sum_{i=1}^k \Big(\big|\lambda_i(Q_S^* + Q_S)\big| + s_i(Q_S)^2\Big) \;\to 0 \qquad (\textit{as } k\to\infty) \tag{33}$$

and $(\varepsilon_k)_{k\in\mathbf{N}^+}$ is a sequence independent of n and V_h.

4.3. Symmetric preconditioning for discretized elliptic problems.

4.3.1. General elliptic equations. Let us consider an elliptic problem

$$\begin{cases} Lu \equiv -\mathrm{div}\,(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu = g, \\ u_{|\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \end{cases} \tag{34}$$

where L satisfies Assumptions 3.2 and $g \in L^2(\Omega)$. We define $H^1_D(\Omega) = \{u \in H^1(\Omega) : u_{|\Gamma_D} = 0\}$; then Assumptions 3.2 ensure that problem (34) has a unique weak solution $u \in H^1_D(\Omega)$. Now let $V_h = span\{\varphi_1, \dots, \varphi_n\} \subset H^1_D(\Omega)$ be a given FEM subspace. We seek the FEM solution $u_h \in V_h$, which requires solving the $n \times n$ system

$$\mathbf{L}_h\,\mathbf{c} = \mathbf{b}, \tag{35}$$

where

$$\big(\mathbf{L}_h\big)_{i,j} = \int_\Omega \Big(A\,\nabla\varphi_i \cdot \nabla\varphi_j + (\mathbf{b}\cdot\nabla\varphi_i)\varphi_j + c\varphi_i\varphi_j \Big) \; + \int_{\Gamma_N} \alpha\varphi_i\varphi_j \, d\sigma$$

and $\mathbf{b}_j = \int_\Omega g\varphi_j$ $(j = 1, \dots, n)$. Following subsection 4.2, we define a preconditioner for system (35) as the discretization of a suitable symmetric elliptic operator. Let

$$Su := -\mathrm{div}\,(A\,\nabla u) + hu \qquad \text{for } \; u \in H^2(\Omega) : \; u_{|\Gamma_D} = 0, \; \tfrac{\partial u}{\partial \nu_A} + \beta u_{|\Gamma_N} = 0, \tag{36}$$

where $h \in L^\infty(\Omega)$ and $h \geq 0$ if $\Gamma_D \neq \emptyset$ and $h \geq \delta_0 > 0$ if $\Gamma_D = \emptyset$, and, further, $\beta \in L^\infty(\Gamma_N)$ and $\beta \geq 0$. The corresponding inner product on $H^1_D(\Omega)$ is

$$\langle u, v\rangle_S := \int_\Omega (A\;\nabla u \cdot \nabla v + huv) + \int_{\Gamma_N} \beta uv \, d\sigma\,. \tag{37}$$

We introduce the matrix

$$\mathbf{S}_h = \Big\{ \langle \varphi_i, \varphi_j\rangle_S \Big\}_{i,j=1}^n \tag{38}$$

as preconditioner for system (35), and then solve system (20) using the CG algorithm (5) with the $\mathbf{S}_h$-inner product and with $A = \mathbf{S}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T$.

THEOREM 4.3. *Let $V_h \subset H^1_D(\Omega)$ be an arbitrary FEM subspace and consider the FEM discretization* (35) *of problem* (34)*, using the stiffness matrix $\mathbf{S}_h$ as preconditioner. Then the superlinear convergence of the preconditioned CG method is mesh independent in the sense of Theorem* 4.2*; i.e., we have*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \leq \; \varepsilon_k \qquad (k = 1, 2, \dots, n) \tag{39}$$

for the mesh independent sequence $\varepsilon_k \to 0$ from (33).

Proof. The coercivity and boundedness assumptions on the coefficients of L and S imply in a standard way that L is S-bounded and S-coercive. Proposition 3.5 yields that L and S are compact-equivalent in $H^1_D(\Omega)$ if the latter is endowed with the inner product (37). Therefore Theorem 4.2 is valid with the compact operator Q_S defined via

$$\langle Q_S u, v\rangle_S = \int_\Omega \Big((\mathbf{b}\cdot\nabla u)v + (c-h)uv \Big) + \int_{\Gamma_N} (\alpha-\beta)uv \, d\sigma \qquad (u, v \in H^1_D(\Omega)), \tag{40}$$

which satisfies (21). □

We note that the above result is an extension of [8], where the mesh independence property has been proved for Dirichlet boundary conditions when either S is the symmetric part of L, or both L and S have constant coefficients.

Remark 5. Finding the correction terms in algorithm (5) with the present choice $A = \mathbf{S}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T$ is equivalent to the auxiliary problems

$$\begin{aligned} &\text{find } z_k \in V_h: \qquad \langle z_k, v\rangle_S = \langle L_S d_k, v\rangle_S \qquad (v \in V_h),\\ &\text{find } s_{k+1} \in V_h: \qquad \langle s_{k+1}, v\rangle_S = \langle L_S^* r_{k+1}, v\rangle_S \qquad (v \in V_h); \end{aligned}$$

i.e., z_k and s_{k+1} are the FEM solutions in V_h of the symmetric elliptic problems of the form $Sz_k = Ld_k$ and $Ss_{k+1} = L^* r_{k+1}$ with the boundary conditions of (36).

PROPOSITION 4.4. *Under the conditions of Theorem* 4.3*, the sequence* ε_k *in* (39) *satisfies*

$$\varepsilon_k \le \frac{4s}{k} \sum_{i=1}^{k} \frac{1}{\mu_i}, \tag{41}$$

where μ_i $(i \in \mathbf{N}^+)$ *are the solutions of the eigenvalue problem*

$$Su = \mu u, \quad u_{|\Gamma_D} = 0, \quad r\left(\frac{\partial u}{\partial \nu_A} + \beta u\right)_{|\Gamma_N} = \mu u \tag{42}$$

and $s, r > 0$ *are constants defined below. When the asymptotics* $\mu_i = O(i^{2/d})$ *holds (in particular, for Dirichlet boundary conditions),*

$$\varepsilon_k \le O\left(\frac{\log k}{k}\right) \quad \text{if } d = 2 \quad \text{and} \quad \varepsilon_k \le O\left(\frac{1}{k^{2/d}}\right) \quad \text{if } d \ge 3. \tag{43}$$

Proof. From (40) and (12) for $v = u$, letting $d = c - h$ and $\gamma = \alpha - \beta$, we obtain

$$\begin{aligned} \langle Q_S u, u\rangle_S &= \int_\Omega \left(d - \frac{1}{2}(\operatorname{div}\mathbf{b})\right) u^2 \; + \int_{\Gamma_N} \left(\gamma + \frac{1}{2}(\mathbf{b}\cdot\nu)\right) u^2 \, d\sigma \\ &\le C_1 \|u\|^2_{L^2(\Omega)} + C_2 \|u\|^2_{L^2(\Gamma_N)}. \end{aligned}$$

We have $\big|\langle (Q_S^* + Q_S)u, u\rangle_S\big| = 2\big|\langle Q_S u, u\rangle_S\big|$; hence the variational characterization of the eigenvalues yields

$$\big|\lambda_i(Q_S^* + Q_S)\big| = \min_{H_{i-1}\subset H_S} \max_{\substack{u \perp H_{i-1} \\ u \neq 0}} \frac{\big|\langle (Q_S^* + Q_S)u, u\rangle_S\big|}{\|u\|_S^2}$$

$$\le 2 \min_{H_{i-1}\subset H_S} \max_{\substack{u \perp H_{i-1} \\ u \neq 0}} \frac{C_1\|u\|^2_{L^2(\Omega)} + C_2\|u\|^2_{L^2(\Gamma_N)}}{\|u\|_S^2},$$

where H_{i-1} stands for an arbitrary $(i-1)$-dimensional subspace. On the other hand, here Q_S falls into the type (11), and hence (14) implies

$$\|Q_S u\|_S^2 \le 2K_1^2 \|u\|^2_{L^2(\Omega)} + 2K_2^2 \|u\|^2_{L^2(\Gamma_N)}.$$

Since $s_i(Q_S)^2 = \lambda_i(Q_S^* Q_S)$ and $\langle Q_S^* Q_S u, u\rangle_S = \|Q_S u\|_S^2$, we obtain as above that

$$s_i(Q_S)^2 = \min_{H_{i-1}\subset H_S} \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{\langle Q_S^* Q_S u, u\rangle_S}{\|u\|_S^2}$$

$$\leq \min_{H_{i-1}\subset H_S} \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{2K_1^2\|u\|_{L^2(\Omega)}^2 + 2K_2^2\|u\|_{L^2(\Gamma_N)}^2}{\|u\|_S^2}\,.$$

Altogether, letting $s := \frac{C_1+K_1^2}{m^2}$, $r := \frac{C_1+K_1^2}{C_2+K_2^2}$, formula (33) implies

$$\varepsilon_k \leq \frac{4s}{k}\sum_{i=1}^{k}\hat{\mu}_i, \qquad \text{where } \hat{\mu}_i = \min_{H_{i-1}\subset H_S} \max_{\substack{u\perp H_{i-1}\\ u\neq 0}} \frac{\|u\|_{L^2(\Omega)}^2 + \frac{1}{r}\|u\|_{L^2(\Gamma_N)}^2}{\|u\|_S^2},$$

in which the fraction equals $1/\mu$ for (42); hence the equality $\hat{\mu}_i = \frac{1}{\mu_i}$ follows from the variational characterization of the eigenvalues.

Estimate (43) follows from the asymptotics $\mu_i = O(i^{2/d})$ by an elementary calculation. For Dirichlet boundary conditions, this asymptotic behavior can be found in [11]. ∎

Remark 6. To the authors' knowledge the asymptotic behavior $\mu_i = O(i^{2/d})$ is not known for general (other than Dirichlet) boundary conditions. However, for the simple special case $-\Delta u = \mu u$, $\frac{\partial u}{\partial \nu}\big|_{\partial\Omega} = \mu u$, where Ω is a disc in $\mathbf{R}^2$, one can easily verify via the sign properties of the Bessel functions that μ_i are asymptotic to the Dirichlet eigenvalues and hence also satisfy $\mu_i = O(i^{2/d})$. This suggests a wider validity of this asymptotic rate.

Remark 7. It is of interest to compare the estimates (43), obtained in the context of the CGN method, to those valid for the GCG-LS method. In [8] we have proved $\varepsilon_k \leq O\big(k^{-1/2}\big)$ in two dimensions on the unit square for the GCG-LS method under the same preconditioning (for Dirichlet boundary conditions, and using explicit formulae for the eigenvalues). Using the same technique, one can similarly derive $\varepsilon_k \leq O\big(k^{-1/d}\big)$ in d dimensions (on the unit cube). That is, comparing with (43), we see that the decay rate of ε_k for the CGN method is almost or exactly (in two or more dimensions, respectively) the square of the decay rate for the GCG method, which compensates for the extra work of solving two auxiliary problems in the preconditioned CGN iteration.

4.3.2. An example: Convection-diffusion equations with Helmholtz preconditioners. As a special case of the preceding subsection, let us consider the case of a convection-diffusion operator L in (34) and a preconditioning operator S with constant coefficients. Namely, if $A \equiv I$ in (34), then we have the problem

$$\text{(44)} \qquad \begin{cases} Lu \equiv -\Delta u + \mathbf{b}(x)\cdot\nabla u + c(x)u = g(x), \\ u_{|\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu} + \alpha(x)u_{|\Gamma_N} = 0, \end{cases}$$

where for clearness, the dependence of the coefficients on x has now been indicated unlike before. Let us define the preconditioning operator

$$\text{(45)} \qquad Su := -\Delta u + hu \qquad \text{for } u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu} + \beta u_{|\Gamma_N} = 0,$$

where $h, \beta \in \mathbf{R}$ are constants such that $h \geq 0$ if $\Gamma_D \neq \emptyset$ and $h > 0$ if $\Gamma_D = \emptyset$, and further, $\beta \geq 0$. (For constant $\mathbf{b}$ and Dirichlet boundary conditions, the analysis of linear convergence in [25] proposes $h = O(|\mathbf{b}|^2)$ as an efficient choice.)

Then the auxiliary problems with this preconditioning are discrete Helmholtz problems with constant coefficients. For such problems various fast solvers are available (like fast Fourier transform, cyclic reduction, or multigrid; see, e.g., [19, 28, 30]), which, together with the mesh independence result of Theorem 4.3, turns $\mathbf{S}_h$ into an efficient preconditioner. We point out that this is an extension of [8], where the mesh independence property has been proved for Dirichlet boundary conditions under the strong restriction that the operator L itself has constant coefficients.

4.3.3. Elliptic systems. Analogously to subsection 4.3.1, we can consider elliptic systems

$$\left.\begin{array}{l} L_i u \equiv -\mathrm{div}\,(A_i \nabla u_i) + \mathbf{b}_i \cdot \nabla u_i + \sum_{j=1}^{l} V_{ij} u_j = g_i, \\ u_{i\,|\Gamma_D} = 0, \qquad \frac{\partial u_i}{\partial \nu_A} + \alpha_i u_{i\,|\Gamma_N} = 0 \end{array}\right\} \qquad (i = 1, \ldots, l), \tag{46}$$

where Ω, A_i, and α_i are as in Assumptions 3.2, $\mathbf{b}_i \in C^1(\overline{\Omega})^N$, $g_i \in L^2(\Omega)$, $V_{ij} \in L^\infty(\Omega)$. We assume that $\mathbf{b}_i$ and the matrix $V = \left\{V_{ij}\right\}_{i,j=1}^{l}$ satisfy the coercivity property

$$\lambda_{min}(V + V^T) - \max_i \mathrm{div}\,\mathbf{b}_i \geq 0$$

pointwise on Ω, where λ_{min} denotes the smallest eigenvalue; then system (46) has a unique weak solution $u \in H_D^1(\Omega)^l$. Such systems arise, e.g., from suitable time discretization and Newton linearization of transport systems, which often consist of a huge number of equations [33]. Now we choose an FEM subspace $V_h \subset H_D^1(\Omega)^l$ and look for the solution of the corresponding algebraic system $\mathbf{L}_h\,\mathbf{c} = \mathbf{b}$. We define the preconditioning operator $S = (S_1, \ldots, S_l)$ as the l-tuple of independent operators

$$S_i u_i := -\mathrm{div}\,(A_i \nabla u) + h_i u \qquad \text{for } u_i \in H^2(\Omega) :\ u_{i\,|\Gamma_D} = 0,\ \frac{\partial u_i}{\partial \nu_A} + \beta_i u_{i\,|\Gamma_N} = 0 \tag{47}$$

$(i = 1, \ldots, l)$ with the conditions of (36), and let $\mathbf{S}_h$ be the stiffness matrix of S in $H_D^1(\Omega)^l$.

Then, similarly to subsection 4.3.1, one can verify that the superlinear convergence of the preconditioned CG method is mesh independent in the sense of Theorem 4.2; i.e., (32)–(33) hold.

This result is an extension of [24] where the above preconditioning has been introduced and its efficient parallelizability has been demonstrated; on the other hand, the mesh independence property was proved there for Dirichlet boundary conditions under strong restrictions on the matrix V (antisymmetric, or normal when the operator L itself has constant coefficients).

4.4. Nonsymmetric preconditioning in Hilbert space. Now let N be a general (possibly nonsymmetric) S-bounded and S-coercive operator which is compact-equivalent to L with $\mu = 1$; i.e., (9) becomes

$$L_S = N_S + Q_S\,. \tag{48}$$

We introduce the stiffness matrix of N_S,

$$\mathbf{N}_h = \Big\{ \langle N_S \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n},$$

as preconditioner for system (18), and wish to solve

$$\mathbf{N}_h^{-1} \mathbf{L}_h \, \mathbf{c} = \hat{\mathbf{b}} \tag{49}$$

(with $\hat{\mathbf{b}} = \mathbf{N}_h^{-1}\mathbf{b}$) using the CG method . Since N is nonsymmetric, in order to define an inner product on $\mathbf{R}^n$ we preserve the stiffness matrix of S on V_h; i.e., using (19) we endow $\mathbf{R}^n$ with the $\mathbf{S}_h$-inner product $\langle \mathbf{c}, \mathbf{d} \rangle_{\mathbf{S}_h} := \mathbf{S}_h \, \mathbf{c} \cdot \mathbf{d}$ as earlier. Then the $\mathbf{S}_h$-adjoint of $\mathbf{N}_h^{-1}\mathbf{L}_h$ is $\mathbf{S}_h^{-1}\mathbf{L}_h^T \, \mathbf{N}_h^{-T}\mathbf{S}_h$; hence we apply the CG algorithm (5) with $A = \mathbf{N}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T \mathbf{N}_h^{-T}\mathbf{S}_h$.

Letting

$$\mathbf{Q}_h = \Big\{ \langle Q_S \varphi_i, \varphi_j \rangle_S \Big\}_{i,j=1}^{n},$$

system (20) takes the form

$$(\mathbf{I}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h) \, \mathbf{c} = \hat{\mathbf{b}}, \tag{50}$$

where $\mathbf{I}_h$ is the $n \times n$ identity matrix. Using (6), the CG algorithm (5) thus provides

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}} \right)^{1/k} \le \frac{2}{k\nu_h} \sum_{i=1}^{k} \Big(\lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T \, \mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h) + \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T \, \mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h) \Big) \tag{51}$$

$(k = 1, 2, \ldots, n)$, where

$$\nu_h = \min_{\mathbf{c} \in \mathbf{R}^n} \frac{\|\mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}\|_{\mathbf{S}_h}^2}{\|\mathbf{c}\|_{\mathbf{S}_h}^2} \, . \tag{52}$$

Again, our goal is to give a bound on (51) that is independent of V_h.

PROPOSITION 4.5. *Let L and N be S-bounded and S-coercive operators, in particular,*

$$m := \inf_{\substack{u \in D(L) \\ u \neq 0}} \frac{\langle Lu, u \rangle}{\|u\|_S^2} > 0, \qquad \hat{m} := \inf_{\substack{u \in D(N) \\ u \neq 0}} \frac{\langle Nu, u \rangle}{\|u\|_S^2} > 0,$$
$$\hat{M} := \sup_{\substack{u \in D(N) \\ u \neq 0}} \frac{|\langle Nu, v \rangle|}{\|u\|_S \|v\|_S} > 0,$$

and let Q_S be a compact operator on H_S. Let $\mathbf{S}_h$, $\mathbf{N}_h$, and $\mathbf{Q}_h$ be defined as above, and let $s_i(Q_S)$ $(i = 1, 2, \ldots)$ denote the singular values of Q_S. Then the following relations hold:

(a) $$\sum_{i=1}^{k} \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T \, \mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h) \le \frac{1}{\hat{m}^2} \sum_{i=1}^{k} s_i(Q_S)^2 \qquad (k = 1, \ldots, n),$$

(b) $$\sum_{i=1}^{k} \big| \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T \, \mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h) \big| \le \frac{2}{\hat{m}} \sum_{i=1}^{k} s_i(Q_S) \qquad (k = 1, \ldots, n),$$

(c) $$\nu_h \geq \frac{m^2}{\hat{M}^2}\,.$$

Proof. (a) We proceed in a manner similar to that of Proposition 4.1. Let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h)$ $(i=1,\ldots,n)$ and let $\mathbf{c}^i = (c_1^i,\ldots,c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors with property (25). Then

$$\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i = \lambda_i \qquad (i=1,\ldots,n). \tag{53}$$

Let $\mathbf{d}^i := \mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i$ for all i; that is,

$$\mathbf{N}_h\,\mathbf{d}^i = \mathbf{Q}_h\,\mathbf{c}^i\,. \tag{54}$$

For this $\mathbf{d}^i$ and λ_i, similarly to Proposition 4.1, we have (28) and, letting $u_i = \sum_{j=1}^n c_j^i\varphi_j \in V_h$ and $z_i = \sum_{j=1}^n d_j^i\varphi_j \in V_h$, we obtain (29). Further, for all $v = \sum_{j=1}^n p_j\varphi_j \in V_h$, with notation $\mathbf{p} = (p_1,\ldots,p_n) \in \mathbf{R}^n$, (54) yields $\mathbf{N}_h\,\mathbf{d}^i \cdot \mathbf{p} = \mathbf{Q}_h\,\mathbf{c}^i \cdot \mathbf{p}$, which means

$$\langle N_S z_i, v\rangle_S = \langle Q_S u_i, v\rangle_S \qquad (v \in V_h).$$

From this we have

$$\|z_i\|_S^2 \leq \frac{1}{\hat{m}}\,\langle N_S z_i, z_i\rangle_S = \frac{1}{\hat{m}}\,\langle Q_S u_i, z_i\rangle_S \leq \frac{1}{\hat{m}}\,\|Q_S u_i\|_S\|z_i\|_S;$$

hence $\|z_i\|_S \leq \frac{1}{\hat{m}}\,\|Q_S u_i\|_S$. Then from (29)

$$\sum_{i=1}^k \lambda_i \leq \frac{1}{\hat{m}^2}\sum_{i=1}^k \|Q_S u_i\|_S^2 = \frac{1}{\hat{m}^2}\sum_{i=1}^k \langle Q_S^* Q_S u_i, u_i\rangle_S, \tag{55}$$

whence the desired estimate follows in the same way as from (30) in Proposition 4.1.

(b) Now let $\lambda_i := \lambda_i(\mathbf{S}_h^{-1}\mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h + \mathbf{N}_h^{-1}\mathbf{Q}_h)$ and let $\mathbf{c}^i = (c_1^i,\ldots,c_n^i) \in \mathbf{R}^n$ be corresponding eigenvectors with property (25). Then

$$\lambda_i = \lambda_i\,\mathbf{S}_h\,\mathbf{c}^i\cdot\mathbf{c}^i = \mathbf{Q}_h^T\,\mathbf{N}_h^{-T}\mathbf{S}_h\,\mathbf{c}^i\cdot\mathbf{c}^i + \mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i\cdot\mathbf{c}^i = 2\,\mathbf{S}_h\mathbf{N}_h^{-1}\mathbf{Q}_h\,\mathbf{c}^i\cdot\mathbf{c}^i = 2\,\mathbf{Q}_h\,\mathbf{c}^i\cdot\mathbf{e}^i,$$

where $\mathbf{e}^i := \mathbf{N}_h^{-T}\mathbf{S}_h\,\mathbf{c}^i$ for all i. Here for all $v = \sum_{j=1}^n p_j\varphi_j \in V_h$, with notation $\mathbf{p} = (p_1,\ldots,p_n) \in \mathbf{R}^n$, we obtain $\mathbf{e}^i\cdot\mathbf{N}_h\,\mathbf{p} = \mathbf{S}_h\,\mathbf{c}^i\cdot\mathbf{p}$, which means $\langle w_i, N_S v\rangle_S = \langle u_i, v\rangle_S$ for all $v \in V_h$, where $w_i = \sum_{j=1}^n e_j^i\varphi_j$ and $u_i = \sum_{j=1}^n c_j^i\varphi_j$, or

$$\langle N_S^* w_i, v\rangle_S = \langle u_i, v\rangle_S \qquad (v \in V_h). \tag{56}$$

Denote by P the orthogonal projection of H_S onto V_h. Then (56) yields $u_i = PN_S^* w_i$. Here the linear mapping $(PN_S^*)_{|V_h} : V_h \to V_h$ is one-to-one, since for all $v \in V_h$

$$\langle PN_S^* v, v\rangle_S = \langle N_S^* v, v\rangle_S = \langle N_S v, v\rangle_S \geq \hat{m}\|v\|_S^2\,. \tag{57}$$

Therefore

$$\mathbf{Q}_h\,\mathbf{c}^i\cdot\mathbf{e}^i = \langle Q_S u_i, w_i\rangle_S = \langle Q_S u_i, (PN_S^*)^{-1}_{|V_h} u_i\rangle_S = \langle u_i,\, Q_S^*(PN_S^*)^{-1}_{|V_h} u_i\rangle_S.$$

Here the operator $(PN_S^*)^{-1}_{|V_h}$ has a norm-preserving extension $\hat{N}$ from V_h onto H_S (namely, with $\hat{N}\big|_{(V_h)\perp} := 0$), and from (57) we have $\|\hat{N}\| \le \frac{1}{\hat{m}}$. Altogether, we obtain

$$\sum_{i=1}^{k} |\lambda_i| = 2\sum_{i=1}^{k} \big|\langle Q_S^*(PN_S^*)^{-1}_{|V_h} u_i,\, u_i\rangle_S\big| \;=\; 2\sum_{i=1}^{k} \big|\langle Q_S^*\hat{N}u_i,\, u_i\rangle_S\big| \;\le\; 2\sum_{i=1}^{k} s_i\big(Q_S^*\hat{N}\big)$$
$$\le \frac{2}{\hat{m}}\sum_{i=1}^{k} s_i\big(Q_S^*\big) = \frac{2}{\hat{m}}\sum_{i=1}^{k} s_i\big(Q_S\big)$$

(where, in the inequalities, statements (a) and (b) of Proposition 2.2 have been used, respectively).

(c) Let $\mathbf{c} \in \mathbf{R}^n$ be arbitrary, $\mathbf{d} := \mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}$. Let $u = \sum_{j=1}^n c_j\varphi_j \in V_h$ and $z = \sum_{j=1}^n d_j\varphi_j \in V_h$. Then $m\|u\|_S^2 \le \langle L_S u, u\rangle_S = \mathbf{L}_h\,\mathbf{c}\cdot\mathbf{c} = \mathbf{N}_h\,\mathbf{d}\cdot\mathbf{c} = \langle N_S z, u\rangle_S \le \|N_S z\|_S\|u\|_S$; hence

$$m\|u\|_S \le \|N_S z\|_S$$

and

$$\frac{\|\mathbf{N}_h^{-1}\mathbf{L}_h\mathbf{c}\|^2_{\mathbf{S}_h}}{\|\mathbf{c}\|^2_{\mathbf{S}_h}} = \frac{\mathbf{S}_h\,\mathbf{d}\cdot\mathbf{d}}{\mathbf{S}_h\,\mathbf{c}\cdot\mathbf{c}} = \frac{\|z\|_S^2}{\|u\|_S^2} \ge m^2\,\frac{\|z\|_S^2}{\|N_S z\|_S^2} \ge \frac{m^2}{\hat{M}^2}. \qquad \square$$

By virtue of (51) and Proposition 4.5, we have proved the following theorem.

THEOREM 4.6. *Let L and N be S-bounded and S-coercive operators that are compact-equivalent in H_S with $\mu = 1$. Let the compact operator Q_S be as in* (48)*. Then for any subspace $V_h = span\{\varphi_1, \dots, \varphi_n\} \subset H_S$, the CG algorithm* (5) *with $\mathbf{S}_h$-inner product, applied for the $n \times n$ preconditioned system* (49)*, yields*

$$\left(\frac{\|r_k\|_{\mathbf{S}_h}}{\|r_0\|_{\mathbf{S}_h}}\right)^{1/k} \le\; \varepsilon_k \qquad (k = 1, 2, \dots, n), \tag{58}$$

$$\textit{where}\quad \varepsilon_k = \frac{2\hat{M}^2}{km^2}\sum_{i=1}^{k}\left(\frac{2}{\hat{m}}\,s_i(Q_S) + \frac{1}{\hat{m}^2}\,s_i(Q_S)^2\right) \to 0 \qquad (\textit{as } k\to\infty) \tag{59}$$

and $(\varepsilon_k)_{k\in\mathbf{N}^+}$ is a sequence independent of n and V_h.

Remark 8. When one preconditions L with N, a useful choice for the operator S is the symmetric part of N: i.e., if $D(N) = D(N^*)$, then $S = (N + N^*)/2$, and if $D(N) \ne D(N^*)$, then S is an operator that generates the inner product satisfying $\langle u, v\rangle_S := \frac{1}{2}(\langle Nu, v\rangle + \langle u, Nv\rangle)$ for $u, v \in D(N)$; see [23]. Then in Proposition 4.5 we have $\langle Nu, u\rangle = \|u\|_S^2$ $(u \in D(N))$, and hence $\hat{m} = 1$.

4.5. Nonsymmetric preconditioning for discretized elliptic problems. This section contains our most general result for elliptic operators: in the studied coercive framework, preconditioning with an arbitrary operator N that is compact-equivalent with L provides mesh independent superlinear convergence. Besides its theoretical aspect, the importance of this property will be shown below by some practical examples as well. Let us first consider the elliptic problem (34)

$$\begin{cases} Lu \equiv\; -\mathrm{div}\,(A\,\nabla u) + \mathbf{b}\cdot\nabla u + cu = g, \\ u_{|\Gamma_D} = 0, \qquad \frac{\partial u}{\partial \nu_A} + \alpha u_{|\Gamma_N} = 0, \end{cases} \tag{60}$$

and let us define the nonsymmetric preconditioning operator

(61)
$$Nu := -\mathrm{div}\,(A\,\nabla u) + \mathbf{w}\cdot\nabla u + zu \quad \text{for } u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu_A} + \eta u_{|\Gamma_N} = 0$$

for some properly chosen functions $\mathbf{w}, z, \eta$, where L and N satisfy Assumptions 3.2 in the obvious sense, and further, $g \in L^2(\Omega)$. Accordingly, the preconditioner for the discretized problem (35) is the nonsymmetric stiffness matrix

$$\left(\mathbf{N}_h\right)_{i,j} = \int_\Omega \Big(A\,\nabla\varphi_i\cdot\nabla\varphi_j + (\mathbf{w}\cdot\nabla\varphi_i)\,\varphi_j + z\varphi_i\varphi_j \Big) \ + \int_{\Gamma_N} \eta\varphi_i\varphi_j\,d\sigma\,.$$

We use the same energy space as in the symmetric case, i.e., $H_S = H^1_D(\Omega)$ with inner product (37). We then solve the preconditioned system using the CG algorithm (5) with the $\mathbf{S}_h$-inner product and with $A = \mathbf{N}_h^{-1}\mathbf{L}_h$ and $A^* = \mathbf{S}_h^{-1}\mathbf{L}_h^T\mathbf{N}_h^{-T}\mathbf{S}_h$.

THEOREM 4.7. *Let $V_h \subset H^1_D(\Omega)$ be an arbitrary FEM subspace and consider the FEM discretization* (35) *of problem* (34)*, using the stiffness matrix $\mathbf{N}_h$ as preconditioner. Then the superlinear convergence of the preconditioned CG method is mesh independent in the sense of Theorem* 4.6*; i.e.,* (58)–(59) *hold.*

Proof. The proof is similar to that of Theorem 4.3, but now Theorem 4.6 is applied in $H^1_D(\Omega)$. ∎

Examples. Let us consider problem (44); i.e., when in (60) we have

$$Lu = -\Delta u + \mathbf{b}(x)\cdot\nabla u + c(x)u,$$

where for clarity, the dependence of the coefficients on x has now been indicated. For convection-dominated problems (i.e., when $|\mathbf{b}|$ is large), the inclusion of nonsymmetric terms in N may turn it into a much better approximation of L than a symmetric preconditioner like (45). Although the preconditioner N thus becomes nonsymmetric as is L itself, the solution of the auxiliary problems can still remain considerably simpler than the original one. We illustrate this with two examples.

1. One can propose a preconditioning operator with constant coefficients:

(62)
$$Nu = -\Delta u + \mathbf{w}\cdot\nabla u + zu \qquad \text{for } u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0,$$

where $\mathbf{w} \in \mathbf{R}^d$, $z, \eta \in \mathbf{R}$ are constants such that $z \geq 0$ if $\Gamma_D \neq \emptyset$ and $z > 0$ if $\Gamma_D = \emptyset$, and further, $\eta \geq 0$. Owing to the fact that N has constant coefficients, one can rely on efficient solution methods for the auxiliary problems. Here one can use either multigrid or multilevel methods, or (if Ω is rectangular or the boundary conditions allow the problem to be easily embedded into a rectangular domain) fast direct solvers for separable equations are available; see, e.g., [29].

2. The preconditioning operator (62) can be further simplified if one convection coefficient is dominating. Assume that, say, $b_1(x)$ has considerably larger values than $b_j(x)$ $(j \geq 2)$. Then one can include only one nonsymmetric coefficient, i.e., propose the preconditioning operator

(63)
$$Nu = -\Delta u + w_1\frac{\partial u}{\partial x_1} + zu \qquad \text{for } u \in H^2(\Omega):\ u_{|\Gamma_D} = 0,\ \frac{\partial u}{\partial \nu} + \eta u_{|\Gamma_N} = 0,$$

where $w_1, z, \eta \in \mathbf{R}$ are constants with the same properties as those required for (62). In this case (above all, if $b_1(x)$ are large), the presence of the term $w_1 \frac{\partial u}{\partial x_1}$ itself may turn N into a much better approximation of L. Nevertheless, since this term is one-dimensional, the solution of the auxiliary problems remains considerably simpler than the original one, e.g., via local one-dimensional Green's functions [5]. (The above operator N has been proposed in [8], where the mesh independence result of the PCG method has been proved for Dirichlet boundary conditions under the strong restriction that the operator L itself has constant coefficients.)

Analogously to the symmetric case in subsection 4.3.3, the above results can be extended to systems in a straightforward way. Namely, let us consider system (46) and introduce the preconditioning operator N as an l-tuple of decoupled operators N_i, where each N_i is of the type (61). Then the superlinear convergence of the preconditioned CG method is mesh independent in the sense of Theorem 4.6; i.e., (58)–(59) hold. Since N_i are decoupled, the resulting algorithm is parallelizable. This turns it into an efficient method if, for instance, each N_i is like (62), or the problem itself is in one dimension, which may occur, e.g., after using some method of splitting in meteorological models with several components; see [33].

5. Some closing remarks.

5.1. Conclusions and notes on numerical realization. The main results of this paper can be summarized as follows. If two elliptic operators are compact-equivalent (which requires that their principal parts coincide up to a constant factor and they have homogeneous Dirichlet conditions on the same portion of the boundary), then the PCGN method provides mesh independent superlinear convergence; i.e., a bound on the rate of superlinear convergence is given in the form of a sequence which is mesh independent and is determined only by the elliptic operators. The analogous result holds for suitable elliptic systems where, as an additional advantage, the preconditioning operator can be chosen to be decoupled. Various further examples have been shown on the efficient choice of compact-equivalent preconditioners.

For the GCG-LS method we have obtained similar earlier results in [8, 24], but with severe restrictions: except for some special cases, both the original and preconditioning operators had to contain constant coefficients, and further, only Dirichlet boundary conditions have been considered. On the other hand, numerical experiments in [24] suggest that the restrictions are probably mostly technical, since a similar superlinear behavior has been observed for test problems with or without these conditions. Remark 7 suggests that the PCGN and GCG-LS methods require the same order of operations for prescribed accuracy; hence there is no a priori preference for one over the other. In any case, a favorable property for the PCGN iteration is the generality of the underlying theory, clarified in the present paper.

The PCGN algorithm has been applied in the setting of subsection 4.3.3 as an inner iterative solver for Newton's method for nonlinear nonsymmetric elliptic systems in [1]. As in the above-mentioned experiments in [24], an efficient performance of the compact-equivalent preconditioning has been observed. Further numerical experiments are beyond the scope and length of this paper.

When realizing the equivalent operator preconditioning for a problem with a second order operator with variable coefficients, one can use an inner-outer iteration method, i.e., precondition in the outer iterations with the given second order operator and use inner iterations to solve this equation. For the superlinear rate to remain, the inner iteration errors must not be of an order greater than that for the first order part of the operator. For optimal complexity of the overall computations to hold, one

should then solve the arising inner systems with an optimal order of computational complexity, i.e., proportional to the degrees of freedom used in the discretization of the differential equation.

5.2. On singular perturbation problems. For singular perturbation problems such as

$$L_\varepsilon u \equiv -\varepsilon\Delta u + \mathbf{b}\cdot\nabla u + cu = f$$

(plus boundary conditions), where $\varepsilon > 0$ but $\varepsilon \ll \|\mathbf{b}\|$, one cannot neglect the first order term when forming an efficient preconditioner. Such problems are characterized by thin boundary and/or interior layers, and the diffusion term plays a noticeable role only in the layer. This property is not exploited in preconditioners like (62). A possible approach for handling such problems therefore is to use the following defect-correction method:

$$L_{\delta(x)}(u_{k+1} - u_k) = f - L_\varepsilon u \qquad (k \in \mathbf{N}^+),$$

where u_0 is given and in practice only one or two steps need to be performed. Here

$$L_{\delta(x)}u := -\delta(x)\Delta u + \mathbf{b}\cdot\nabla u + cu,$$

where $\delta(x) = 0$ outside the layers and increases continuously along each characteristic line (defined by the velocity vector $\mathbf{b}$) from zero to ε in the layers. The widths of the layers are typically chosen as $\varepsilon\log(1/\varepsilon)$. To solve the correction equation by iteration, one can form a preconditioner S by using the operator $\mathbf{b}\cdot\nabla u + hu$ outside the layers and $-\delta(x)\Delta u + \mathbf{b}\cdot\nabla u + hu$ in the layers for some properly chosen function $h \geq 0$. The analysis of the problem will not be considered further in the present paper.

REFERENCES

[1] I. Antal and J. Karátson, *A mesh independent superlinear algorithm for some nonlinear nonsymmetric elliptic systems*, Comput. Math. Appl., to appear.

[2] S. F. Ashby, T. A. Manteuffel, and P. E. Saylor, *A taxonomy for conjugate gradient methods*, SIAM J. Numer. Anal., 27 (1990), pp. 1542–1568.

[3] O. Axelsson, *A generalized conjugate gradient, least square method*, Numer. Math., 51 (1987), pp. 209–227.

[4] O. Axelsson, *Iterative Solution Methods*, Cambridge University Press, Cambridge, UK, 1994.

[5] O. Axelsson and S. V. Gololobov, *A combined method of local Green's functions and central difference method for singularly perturbed convection-diffusion problems*, J. Comput. Appl. Math., 161 (2003), pp. 245–257.

[6] O. Axelsson and J. Karátson, *On the rate of convergence of the conjugate gradient method for linear operators in Hilbert space*, Numer. Funct. Anal., 23 (2002), pp. 285–302.

[7] O. Axelsson and J. Karátson, *Symmetric part preconditioning for the conjugate gradient method in Hilbert space*, Numer. Funct. Anal. Optim., 24 (2003), pp. 455–474.

[8] O. Axelsson and J. Karátson, *Superlinearly convergent CG methods via equivalent preconditioning for nonsymmetric elliptic operators*, Numer. Math., 99 (2004), pp. 197–223.

[9] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North–Holland, Amsterdam, 1978.

[10] P. Concus and G. H. Golub, *A generalized conjugate gradient method for nonsymmetric systems of linear equations*, in Computing Methods in Applied Sciences and Engineering, Lecture Notes in Econom. and Math. Systems 134, R. Glowinski and J.-L. Lions, eds., Springer, Berlin, 1976, pp. 56–65.

[11] R. Courant and D. Hilbert, *Methods of Mathematical Physics*, Vol. II, Wiley Classics Library, John Wiley & Sons, New York, 1989.

[12] H. C. Elman and M. H. Schultz, *Preconditioning by fast direct methods for nonself-adjoint nonseparable elliptic equations*, SIAM J. Numer. Anal., 23 (1986), pp. 44–57.
[13] V. Faber and T. Manteuffel, *Necessary and sufficient conditions for the existence of a conjugate gradient method*, SIAM J. Numer. Anal., 21 (1984), pp. 352–362.
[14] V. Faber, T. Manteuffel, and S. V. Parter, *On the theory of equivalent operators and application to the numerical solution of uniformly elliptic partial differential equations*, Adv. in Appl. Math., 11 (1990), pp. 109–163.
[15] Z. Fortuna, *Some convergence properties of the conjugate gradient method in Hilbert space*, SIAM J. Numer. Anal., 16 (1979), pp. 380–384.
[16] I. Gohberg, S. Goldberg, and M. A. Kaashoek, *Classes of Linear Operators*, Vol. I, Oper. Theory Adv. Appl. 49, Birkhäuser Verlag, Basel, Switzerland, 1990.
[17] I. Gohberg and S., Goldberg, *Basic Operator Theory*, Birkhäuser, Boston, MA, 1981.
[18] C. I. Goldstein, T. A. Manteuffel, and S. V. Parter, *Preconditioning and boundary conditions without H_2 estimates: L_2 condition numbers and the distribution of the singular values*, SIAM J. Numer. Anal., 30 (1993), pp. 343–376.
[19] W. Hackbusch, *Multigrid Methods and Applications*, Springer Ser. Comput. Math. 4, Springer, Berlin, 1985.
[20] R. M. Hayes, *Iterative methods of solving linear problems in Hilbert space*, Nat. Bur. Standards Appl. Math. Ser, 39 (1954), pp. 71–104.
[21] M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. Res. Nat. Bur. Standards, Sect. B, 49 (1952), pp. 409–436.
[22] J. Karátson, *Mesh independent superlinear convergence estimates of the conjugate gradient method for some equivalent self-adjoint operators*, Appl. Math., 50 (2005), pp. 277–290.
[23] J. Karátson, *Superlinear PCG Algorithms: Symmetric Part Preconditioning and Boundary Conditions*, Preprint 2006-10, Department of Applied Analysis, ELTE University, Budapest, Hungary; available online from http://www.cs.elte.hu/applanal/eng/preprint_eng.html.
[24] J. Karátson and T. Kurics, *Superlinearly convergent PCG algorithms for some nonsymmetric elliptic systems*, J. Comput. Appl. Math., to appear.
[25] T. Manteuffel and J. Otto, *Optimal equivalent preconditioners*, SIAM J. Numer. Anal., 30 (1993), pp. 790–812.
[26] T. A. Manteuffel and S. V. Parter, *Preconditioning and boundary conditions*, SIAM J. Numer. Anal., 27 (1990), pp. 656–694.
[27] O. Nevanlinna, *Convergence of Iterations for Linear Equations*, Birkhäuser Verlag, Basel, Switzerland, 1993.
[28] T. Rossi and J. Toivanen, *A parallel fast direct solver for block tridiagonal systems with separable matrices of arbitrary dimension*, SIAM J. Sci. Comput., 20 (1999), pp. 1778–1796.
[29] P. N. Swarztrauber, *A direct method for the discrete solution of separable elliptic equations*, SIAM J. Numer. Anal., 11 (1974), pp. 1136–1150.
[30] P. N. Swarztrauber, *The methods of cyclic reduction, Fourier analysis and the FACR algorithm for the discrete solution of Poisson's equation on a rectangle*, SIAM Rev., 19 (1977), pp. 490–501.
[31] R. Winther, *Some superlinear convergence results for the conjugate gradient method*, SIAM J. Numer. Anal., 17 (1980), pp. 14–17.
[32] O. Widlund, *A Lanczos method for a class of nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 15 (1978), pp. 801–812.
[33] Z. Zlatev, *Computer Treatment of Large Air Pollution Models*, Kluwer Academic Publishers, Dordrecht, Boston, London, 1995.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1517–1543

A FINITE ELEMENT METHOD FOR 3D EXTERIOR OSEEN FLOWS: ERROR ESTIMATES*

PAUL DEURING†

Abstract. Stationary Oseen flows in a 3D exterior domain are discretized by applying a P1-P1 stabilized finite element method to the Oseen system in a truncated exterior domain, with an implicit pointwise artificial boundary condition on the truncating boundary. Error estimates are proved for this discretization. The paper extends an approach which was introduced by Guirguis and Gunzburger [*RAIRO Modél. Math. Anal. Numér.*, 21 (1987), pp. 445–464] for exterior Stokes flows. The stabilized P1-P1 method in question was introduced by Rebollo [*Numer. Math.*, 79 (1998), pp. 283–319].

Key words. Oseen flows, exterior domains, stabilized finite element methods, error estimates

AMS subject classifications. 35Q30, 65N30, 76D05

DOI. 10.1137/06065492X

1. Introduction.

We consider the following Oseen system in an exterior domain Ω, with a Dirichlet boundary condition on $\partial\Omega$ and with zero velocity at infinity:

$$-\Delta u + \tau \cdot \partial_1 u + \nabla \pi = f, \quad \operatorname{div} u = 0 \quad \text{in } \Omega, \tag{1.1}$$
$$u \,|\, \partial\Omega = (-1,0,0), \quad |u(x)| \to 0 \quad \text{for } |x| \to \infty,$$

where $\Omega = \mathbb{R}^3 \backslash \overline{\mathcal{P}}$ for some open bounded polyhedron $\mathcal{P} \subset \mathbb{R}^3$ with Lipschitz boundary. Problem (1.1) arises as a linearization of a mathematical model for the steady motion without rotation of a rigid body in a viscous incompressible fluid. We refer to [6] for more details; here we mention only that $\tau \in (0,\infty)$ is the Reynolds number of the fluid, and the vector $(-1,0,0)$ appearing in the boundary condition on $\partial\Omega$ corresponds to the normalized steady velocity of the rigid body. Under suitable assumptions on f, problem (1.1) admits a unique solution (u,π) with $\nabla u \in L^2(\Omega)^9$ and $\pi \in L^2(\Omega)$. In the following, this solution will be called "exterior flow." We will study a finite element method for computing this exterior flow in a region Ω_S around $\partial\Omega$, for some fixed $S > 0$ with $\mathbb{R}^3 \backslash \Omega \subset B_S$. (For any $r > 0$, we denote by B_r the open ball with center at the origin and with radius r, and we set $\Omega_r := \Omega \cap B_r$.) As computational domains, we consider polyhedrons $P_{h,R}$, which are larger than Ω_S and have parameters $h \in (0,S)$ and $R \in (S,\infty)$ that may be interpreted as follows: The number R indicates that $P_{h,R}$ approximates the truncated exterior domain Ω_R in a suitable sense. As for the parameter h, it refers to the way we decompose $P_{h,R}$ into tetrahedrons. These tetrahedrons are assumed to have a diameter of order h when situated near $\partial\Omega$, and of order $h \cdot R$ when located near ∂B_R. The transition of the mesh size from values of order h near $\partial\Omega$ to about $h \cdot R$ near ∂B_R is performed in such a way that the aspect ratio of our tetrahedrons remains bounded as a function of h and R. We refer to assumptions (A1)–(A8) in section 2 for more details.

On each polyhedron $P_{h,R}$, we define a pair $(V_{h,R}, M_{h,R})$ of P1 finite element spaces. Then, in (2.11)–(2.13) below we introduce a variational problem, which is

*Received by the editors March 23, 2006; accepted for publication (in revised form) January 9, 2007; published electronically August 8, 2007.
http://www.siam.org/journals/sinum/45-4/65492.html
†Laboratoire de Mathématiques Pures et Appliquées, Université du Littoral, B.P. 699, F-62228 Calais cédex, France (Paul.Deuring@lmpa.univ-littoral.fr).

based on these spaces and is a variant of the P1-P1 stabilized finite element method proposed by Rebollo [17] for the Navier–Stokes system in a fixed bounded domain. We show that a solution $(u_{h,R}, \pi_{h,R})$ of (2.11)–(2.13) approximates (u, π) in the sense that an error estimate is valid with essential features that may be stated as follows:

$$\|\nabla(u - u_{h,R})\|_2 \ + \ \|(\pi - \pi_{h,R}) \,|\, \Omega_S\|_2 \ \le \ C \cdot (h^t + h \cdot \ln(R/S) + R^{-1}) \tag{1.2}$$

for $h \in (0, h_0)$, $R \in (R_0, \infty)$, where $h_0 \in (0, S/2)$, $R_0 \in (S, \infty)$ are constants depending on Ω, S and on certain mesh parameters. Inequality (1.2) is valid under the assumption that the quantity $h \cdot R$ is bounded by some arbitrary but fixed constant. The exponent $t \in (0, 1]$ in (1.2) describes the regularity of u and π near the Lipschitz boundary $\partial\Omega$. We may take $t = 1$ if u is H^2 and if π is H^1 in a neighborhood of $\partial\Omega$, but this case will not arise in general because Ω is a nonconvex Lipschitz domain. The constant C in (1.2) depends on Ω, S, t, τ, on an upper bound for $h \cdot R$, on some mesh parameters, on certain quantities related to the exterior flow (u, π) (see (2.1)–(2.5)), and on a bilinear form used to stabilize P1-P1 finite elements (see (2.6)). A more detailed version of (1.2) may be found in Theorem 2.1 below.

Our discrete variational problem (2.11)–(2.13) may be considered as a discretization of a boundary value problem in Ω_R consisting of the Oseen system

$$-\Delta u_R + \tau \cdot \partial_1 u_R + \nabla \pi_R = f \,|\, \Omega_R, \quad \operatorname{div} u_R = 0 \quad \text{in } \Omega_R, \tag{1.3}$$

with the same boundary condition on $\partial\Omega$ as in (1.1):

$$u_R \,|\, \partial\Omega = (-1, 0, 0), \tag{1.4}$$

and with the ensuing pointwise "artificial" boundary condition on the sphere ∂B_R:

$$\sum_{j=1}^{3} \Big(\partial_j v_k(x) - \delta_{jk} \cdot \varrho(x) \Big) \cdot (x_j / R) + \Big(R^{-1} + (\tau/2) \cdot (1 - x_1/R) \Big) \cdot v_k(x) = 0 \tag{1.5}$$

for $x \in \partial B_R$, $1 \le k \le 3$. This boundary value problem can be written as a variational problem which admits a unique solution (u_R, π_R) in $H^1(\Omega_R)^3 \times L^2(\Omega_R)$. As was shown in [7, Theorem 7.1 with $\widetilde{\tau} = 0$ (linear case)], this solution satisfies the relation

$$\|\nabla(u_R - u)\|_2 \ = \ O(R^{-1}) \quad \text{for } R \to \infty. \tag{1.6}$$

This result motivated the choice of the discrete equations (2.11)–(2.13), and it explains the term R^{-1} on the right-hand side of (1.2): this term corresponds to the truncation error exhibited in [7]. It should be remarked, however, that relation (1.6) will not be used in the work at hand. Instead the solution $(u_{h,R}, \pi_{h,R})$ of our finite element variational problem (2.11)–(2.13) will be compared directly with the exterior flow (u, π), and no continuous intermediate problem on Ω_R or $P_{h,R}$ will be considered. In this way, we will be better able to exploit the asymptotics of u and π. A crucial feature of this asymptotics is the existence of a "wake region," which means that in the downstream direction, the velocity u converges more slowly than elsewhere to its boundary value at infinity (see [9, p. 374]). This wake region will be taken into consideration in two ways. First, an asymmetric surface integral on the outer boundary of $P_{h,R}$ enters into the definition of the bilinear form a appearing in our

discrete variational problem (2.11)–(2.13). Second, our error estimates involve bounds of u and ∇u with respect to pointwise norms with inhomogeneous weights. We remark that our error bounds additionally depend on homogeneously weighted L^2-norms of the second derivatives of u and of the gradient of π, with the weights being defined in an implicit way: we take the standard L^2-norm over annular domains $B_R \backslash B_{R/2}$ for $R \in (2 \cdot S, \infty)$ and then multiply by powers of R. A complete list of the quantities related to u and π and entering into our error bounds may be found in (2.1)–(2.4).

Of course, the finite element approximation we consider gives rise not only to a truncation error but also to a discretization error. Concerning that latter component of the total error, we indicate that the velocity part u of the exterior flow is assumed to be an H^{1+t}-function near $\partial\Omega$, and an H^2-function far from $\partial\Omega$. Analogous conditions are imposed on the pressure (H^t near $\partial\Omega$, H^1 far from Ω). As a consequence, if the error is measured by the L^2-norm of the gradient of the velocity, we expect our P1-P1 finite element method to give rise to a discretization error of order h^t related to the approximation of u and π near $\partial\Omega$, and to an error of order h which may be ascribed to the discretization of velocity and pressure far from $\partial\Omega$. This explains the terms h^t and $h \cdot \ln(R/S)$ on the right-hand side of (1.2) and indicates that these terms describe the discretization error in an optimal way, except for the factor $\ln(R/S)$. However, we cannot see how to remove this factor.

It might be asked why we have chosen Rebollo's finite element method for our discretization of (1.3)–(1.5). There are essentially two reasons for this choice. First, Rebollo [17] uses P1-P1 elements and circumvents the LBB condition by introducing a stabilization term ("pressure stabilization") which does not involve any parameter. Thus, implementing Rebollo's method is relatively simple. Second, due to [17], optimal error estimates are available when this method is applied to the stationary Navier–Stokes system in a fixed bounded domain. The theory in [17] even covers a streamlined upwind Petrov–Galerkin (SUPG)-type stabilization related to the convective term ("velocity stabilization"). Thus Rebollo's article presents a coherent theory for several aspects of finite element discretizations of stationary Navier–Stokes flows. Such a theory seemed to be a good starting point for attempting to discretize problem (1.3)–(1.5). As it turns out, the principal results from [17] may be generalized to our situation. This generalization takes the form of estimate (1.2), which we establish here for our approximate Oseen flows, but which remains valid when the stationary Navier–Stokes system with Oseen term and with small Reynolds number is considered, and when Rebollo's version of SUPG is included in the discrete problem. These two points—nonlinearity and SUPG—will be the subject of a separate paper. However, the main difficulties of our theory are discussed in the work at hand because they already arise with the discrete Oseen problem (2.11)–(2.13) without velocity stabilization. These difficulties are essentially due to four features: the Oseen term $\tau \cdot \partial_1 u$, the graded mesh, a nonvanishing mean value of the pressure, and the parameter R, which must be controlled in all our estimates, in addition to the quantity h. Since none of these features appear in [17], it is not astonishing that the proofs in that reference do not carry over to our situation. In fact, the present article represents a considerable extension of the theory in [17].

The argument we present here might also work for mixed finite element methods without stabilization, under one condition: the mixed method in question must satisfy the LBB condition on the type of grid considered here, with a constant independent of *h and R*. This condition is fulfilled by P1-P1 finite elements augmented by bubble functions (the "mini element"), as was shown in [3] and is stated in Theorem 3.1 below.

However, it seems to be an open question whether an analogue of Theorem 3.1 is valid for other kinds of LBB-stable finite elements.

Our approach to the discretization of exterior domains was inspired by Guirguis and Gunzburger [14], who approximated exterior Stokes flows by solutions of finite element variational problems in truncated exterior domains; also see [15, sections 16.3 and 16.4]. Guirguis and Gunzburger were led to these finite element problems by discretizing the Stokes system in a truncated exterior domain with a suitable pointwise artificial boundary condition on the truncating surface. This boundary condition determines the quality of the approximation. The error bounds considered in [14] depend on the given exterior Stokes flow via weighted L^2-norms involving second order or higher derivatives of the velocity, as well as the gradient or higher derivatives of the pressure [14, Theorem 5.2].

We further remark that Goldstein [11] introduced the type of graded meshes we consider in the work at hand. Due to these meshes, the complexity of our finite element method (2.11)–(2.13) on $P_{h,R}$ is proportional to $h^{-3} \cdot \ln(R/S)$, and thus exhibits only a logarithmic dependence on R; see [11], [12] in this respect.

Further articles related to the computation of exterior Stokes, Oseen, or Navier–Stokes flows, but not closely linked to the work at hand, are listed in [6]. To our knowledge, there are no previous articles presenting error estimates for discretizations of 3D exterior Oseen or Navier–Stokes flows.

2. Notation. Statement of our finite element variational problem. Main results. For $U \subset \mathbb{R}^3$, put $U^c := \mathbb{R}^3 \backslash U$. If $r \in (0,\infty)$, $x \in \mathbb{R}^3$, put $B_r(x) := \{y \in \mathbb{R}^3 \ : \ |y-x| < r\}$, $B_r := B_r(0)$. If U is (Lebesgue) measurable, we denote the Lebesgue measure of U by $|U|$. Let $V \subset \mathbb{R}^3$ be open. For functions $v : V \mapsto \mathbb{R}$, $w, \widetilde{w} : V \mapsto \mathbb{R}^3$ with appropriate regularity, the notation $\partial_l v$ for $1 \le l \le 3$, $\partial^a v$ for $a \in \mathbb{N}_0^3$, ∇v, Δv, $\operatorname{div} w$, $(\widetilde{w} \cdot \nabla) w$ stands for partial derivatives, with obvious meanings. If $p \in [1,\infty)$, the standard L^p-norm of functions on V is denoted by $\| \ \|_{p,V}$. Let $m \in \mathbb{N}$. We write $H^m(V)$ for the usual Sobolev space of order m and exponent 2. The usual norm of that space is denoted by $\| \ \|_{m,2,V}$. The subspace $H_0^1(V)$ of $H^1(V)$ is defined in the standard way. We write $H^{-1}(V)^3$ for the dual space of $H_0^1(V)^3$. If $s \in (0,2)$ with $s \ne 1$, the symbol $H^s(V)$ stands for the usual fractional-order Sobolev space, with its intrinsic norm denoted by $\| \ \|_{s,2,V}$ (see [1, 7.48]). We will also use the seminorm $| \ |_{m,2,V}$ defined by

$$|v|_{m,2,V} := \left(\sum_{a \in \mathbb{N}_0^3,\, |a|=m} \|\partial^a v\|_{2,V}^2 \right)^{1/2} \qquad \text{for } \ v \in W^{m,p}(V),$$

where $|a| := a_1 + a_2 + a_3$ for $a \in \mathbb{N}_0^3$. We write $H^m_{loc}(V)$ for the space of all functions $v : V \mapsto \mathbb{R}$ such that $v \,|\, K \in H^m(K)$ for any $K \subset \mathbb{R}^3$ open with $\overline{K} \subset V$. If $(\mathcal{H}, \| \ \|)$ is a normed space consisting of functions from V into $\mathbb{R}$, we use the symbol $\| \ \|$ also for the norm $(\sum_{i=1}^3 \|v_i\|^2)^{1/2}$ of fields $v : V \mapsto \mathbb{R}^3$ with $v_i \in \mathcal{H}$ $(1 \le i \le 3)$. Moreover, if $W \subset \mathbb{R}^3$ with $V \subset W$, and if $w : W \mapsto \mathbb{R}$ is a function with $w \,|\, V \in \mathcal{H}$, we write $\|w\|$ instead of $\|w \,|\, V\|$. This convention will not give rise to ambiguities because the notation for our norms is chosen in such a way that the domain of reference is always indicated.

For $R \in (0,\infty)$, put $\Omega_R := B_R \cap \Omega$. We fix some number $S \in (0,\infty)$ with $\Omega^c \subset B_S$. As we already remarked in section 1, the domain Ω_S should be considered as the region where we want to compute the flow (u,π). Denote $U_0 := \Omega_S$, $U_j := B_{2^j \cdot S} \backslash B_{2^{j-1} \cdot S}$ for $j \in \mathbb{N}$. It will be convenient to use the notation $U_{-1} := \emptyset$.

In the rest of this section, we will always assume that the parameter h belongs to $(0,\, S/2)$, and R to (S, ∞), except when indicated otherwise.

For such h and R, we choose an open polyhedron $P_{h,R} \subset \mathbb{R}^3$ and closed tetrahedrons $K_1^{(h,R)}, \ldots, K_{k(h,R)}^{(h,R)} \subset P_{h,R}$, with $k(h,R) \in \mathbb{N}$, such that the following assumptions (A1)–(A8) are valid. (In order to simplify notation, we will always write k instead of $k(h,R)$, and K_l instead of $K_l^{(h,R)}$, for $1 \le l \le k$.):

(A1) $\overline{P_{h,R}} = \bigcup_{l=1}^{k} K_l$.

(A2) $K_l \subset \Omega_R$; $K_l \cap K_m$ is either empty or a common vertex or a common side or a common face of K_l and K_m, for $l, m \in \{1, \ldots, k\}$ with $l \neq m$.

(A3) There is $\sigma_1 > 0$, independent of h and R, such that

$$\sigma_1 \cdot 2^j \cdot h \le \sup\big\{\, r \in (0,\infty) \,:\, B_r(x) \subset K_l \text{ for some } x \in K_l \,\big\},$$

for $l \in \{1, \ldots, k\}$, $j \in \mathbb{N}_0$ with $K_l \cap U_j \neq \emptyset$. Moreover, $diam\, K_l \le 2^j \cdot h$ for such l and j.

(A4) $B_S \cap P_{h,R} = \Omega_S$; if $l \in \{1, \ldots, k\}$ and if $x \in \partial P_{h,R} \backslash \partial\Omega$ is a vertex of K_l, then $x \in \partial B_R$.

(A5) The domain $P_{h,R} \cup \Omega^c$ is convex.

(A6) There is $\sigma_2 \in (0,\infty)$, independent of h and R, such that for $l \in \{1, \ldots, k\}$, the macroelement $(K_l)_\Delta := \cup\big\{\, K_m \,:\, m \in \{1, \ldots, k\} \text{ with } K_m \cap K_l \neq \emptyset \,\big\}$ is star-shaped with respect to the ball $B_{\sigma_2 \cdot diam\, K_l}(x)$, for some $x \in (K_l)_\Delta$. (See [2, (4.2.2)] for the notion "star-shaped with respect to a ball.")

(A7) For any $l \in \{1, \ldots, k\}$, at least one vertex of K_l is located in the (open) set $P_{h,R}$.

(A8) There is $\varphi_1 \in (0, \pi/2)$, independent of h and R, such that the relation

$$x + \big\{\, z \in \mathbb{R}^3 \backslash \{0\} \,:\, |x|^{-1} \cdot |z|^{-1} \cdot (x \cdot z) \ge \cos\varphi_1 \,\big\} \subset \mathbb{R}^3 \backslash (P'_{h,R} \cup \Omega^c)$$

holds for any $x \in \partial P'_{h,R} \backslash \partial\Omega$, where $P'_{h,R}$ denotes the interior of the union of the tetrahedrons K_l with $K_l \subset \Omega_{2\cdot S}$.

Assumptions (A1) and (A2) specify that $P_{h,R}$ is a subset of Ω_R, and $\mathcal{T}_{h,R}$ is a triangulation of $P_{h,R}$ without hanging nodes. By assumption (A3), we specify Goldstein's mesh-grading process and require the aspect ratio of our mesh cells to stay away from zero. Condition (A6) is necessary for constructing interpolation operators of Clément type; see [2, section 4.8], where (A6) is used implicitly. As concerns (A5), (A7), and (A8), these assumptions are needed so that Theorem 3.1 is valid. Moreover, the convexity of $P_{h,R} \cup \Omega^c$ required in (A5) and our assumptions on Ω imply that $P_{h,R}$ is Lipschitz bounded (see [13, Corollary 1.2.2.3]). This means in particular that the outward unit normal to $P_{h,R}$ is well defined [16, pp. 88–89]. We will denote it by n. It will be convenient to use the abbreviation $\partial_{h,R} := \partial P_{h,R} \backslash \partial\Omega$. Thus $\partial_{h,R}$ is the "outer part" of the boundary of $P_{h,R}$, that is, the surface which cuts off the exterior domain Ω. We put

$$W_{h,R} := \big\{\, v \in H^1(P_{h,R})^3 \,:\, v \,|\, \partial\Omega = 0 \,\big\},$$

$$\|v\|^{(h,R)} := \big(\, \|\nabla v\|^2_{2,P_{h,R}} + R^{-1} \cdot \|v\|^2_{2,\partial_{h,R}} \,\big)^{1/2} \quad \text{for } v \in W_{h,R}.$$

It follows from [6, Theorem 3.4] that the mapping $\|\ \|^{(h,R)}$ is a norm on $W_{h,R}$ which is equivalent to the norm $\|\ \|_{1,2,P_{h,R}}$, but this equivalence involves constants depending on R.

Turning to problem (1.1), we fix $\tau \in (0,\infty)$ and a function $f : \Omega \mapsto \mathbb{R}^3$ such that $f \,|\, \Omega_S \in L^2(\Omega_S)^3$ and $\sup\{\, |f(x)| \cdot |x|^\sigma \ : \ x \in B_S^c \,\} < \infty$ for some $\sigma \in (4,\infty)$. By [9, Theorems VII.2.1, VII.1.1; Lemma VII.1.1], [7, Theorem 4.13], there are unique functions $u \in H^2_{loc}(\Omega)^3,\ \pi \in H^1_{loc}(\Omega) \cap L^2(\Omega)$ such that $\nabla u \in L^2(\Omega)^9$ and such that the pair (u,π) verifies (1.1). This pair (u,π) is the exterior flow we will consider in the following. Since $\partial\Omega$ is only Lipschitz bounded, we cannot expect H^2-regularity of u or H^1-regularity of π near $\partial\Omega$. Instead we have only

$$\mathcal{A}_1 := \|u\|_{1+t,2,\Omega_{4\cdot S}} \ + \ \|\pi\|_{t,2,\Omega_{4\cdot S}} \ < \ \infty \quad \text{for some } \ t \in (0,1], \tag{2.1}$$

with $t < 1$ in general. This parameter t will be kept fixed throughout. Of course, the radius $4 \cdot S$ may be replaced by any number $R \in (0,\infty)$ with $\Omega^c \subset B_R$, but the choice $R = 4 \cdot S$ will be convenient in the following. To our knowledge, no direct proof is available for relation (2.1), which, however, follows from regularity results for the Stokes system in Lipschitz domains (see [8]). By [7, Theorem 5.6], we further have

(2.2)

$$\begin{aligned}\mathcal{A}_2 := {} & \sup\{\, |u(x)| \cdot |x| \cdot (1 + \tau \cdot (|x| - x_1)) \ : \ x \in B_S^c \,\} \\ & + \ \sup\{\, |\pi(x)| \cdot |x|^2 \ : \ x \in B_S^c \,\} \\ & + \ \sup\{\, |\nabla u(x)| \cdot \big[\, |x|^{-2} + \tau^{1/2} \cdot |x|^{-3/2} \cdot (1 + \tau \cdot (|x| - x_1))^{-3/2} \,\big]^{-1} \ : \ x \in B_S^c \,\} \\ & < \infty.\end{aligned}$$

Moreover, by [9, Theorem VII.1.1], [5, Theorem 5.2; Theorem 7.1 with $\widetilde{\tau} = 0$], the following relations are valid:

(2.3)

$$\begin{aligned}\mathcal{A}_3 := {} & \sup\{\, \|\, \partial_l \partial_m u\|_{2,\, B_R \setminus B_{\delta\cdot R}} \cdot \big(\, (1-\delta)^{1/2} \cdot R^{-1} + R^{-3/2} \,\big)^{-1} \ : \\ & \qquad\qquad R \in [2\cdot S,\, \infty),\ \delta \in [1/2,\, 1],\ 1 \le l, m \le 3 \,\} \\ & + \ \sup\{\, \|\, \nabla\pi\|_{2,\, B_R \setminus B_{R/2}} \cdot R \ + \ \|\partial_1 u\|_{2,\, B_R \setminus B_{R/2}} \cdot R \ : \ R \in [2\cdot S,\, \infty) \,\} \\ & < \infty,\end{aligned}$$

(2.4)

$$\begin{aligned}\mathcal{A}_4 := {} & \sup\{\, \big(\, \|\nabla u\|_{2,\, \partial_{h,R}} \ + \ \tau \cdot \|\, (1 - n_1) \cdot u\|_{2,\, \partial_{h,R}} \\ & \qquad + \|\pi\|_{2,\, \partial_{h,R}} \,\big)\, / (h + R^{-1}) \ + \ \|u\|_{2,\, \partial_{h,R}} \ : \ R \in [4\cdot S,\, \infty),\ h \in (0, S_0) \,\} \\ & < \infty,\end{aligned}$$

with a constant $S_0 \in (0,\, S/4)$. Put

$$\mathcal{A} := \max\{\, \mathcal{A}_1,\, \ldots,\, \mathcal{A}_4 \,\}. \tag{2.5}$$

This quantity $\mathcal{A}$ characterizes the regularity and the asymptotic behavior near infinity of the exterior flow (u,π) and contains all the information on (u,π) that we will need in the following. Next we introduce the finite element variational problem with solution

assumed to approximate our exterior flow. To this end, we define the following finite element spaces:

$$V_{h,R} := \big\{\, v \in C^0(\overline{P_{h,R}})^3 \;:\; v \,|\, K_l \in P_1(K_l)^3 \;\text{ for }\; 1 \le l \le k \,\big\},$$

$$Y_{h,R} := \big\{\, v \in V_{h,R} \;:\; v|\partial\Omega = 0 \,\big\},$$

$$M_{h,R} := \big\{\, \varrho \in C^0(\overline{P_{h,R}}) \;:\; \varrho \,|\, K_l \in P_1(K_l) \;\text{ for }\quad 1 \le l \le k \,\big\},$$

where $P_1(K_l)$ denotes the space of all polynomials over K_l of degree at most 1 ($l \in \{1, \ldots, k\}$). We write $\mathcal{B}_{h,R}$ for the space of all functions $v : \overline{P_{h,R}} \mapsto \mathbb{R}$ such that for $l \in \{1, \ldots, k\}$, we have $v \,|\, K_l = \alpha_l \cdot b_{K_l}$ for some $\alpha_l \in \mathbb{R}$, where b_{K_l} is the standard bubble function on K_l (that is, the polynomial of order 4 on K_l vanishing on ∂K_l as defined in [17, p. 287]). For $v, w \in H^1(P_{h,R})^3,\ q \in L^2(P_{h,R})$, we set

$$a(v,w) := a_{h,R,\tau}(v,w) := \int_{P_{h,R}} \left(\sum_{k=1}^{3} \partial_k v \cdot \partial_k w + \tau \cdot \partial_1 v \cdot w \right) dx$$

$$+ \int_{\partial_{h,R}} \Big(R^{-1} + \ (\tau/2) \cdot (1 - n_1) \Big) \cdot (v \cdot w)\, do_x,$$

$$c(v,q) := c_{h,R}(v,q) := -\int_{P_{h,R}} \operatorname{div} v \cdot q \, dx.$$

Next, following [17], we want to introduce a stabilization term which allows us to circumvent the LBB condition. To this end, we consider a bilinear symmetric form $A := A_{h,R} : H_0^1(P_{h,R})^3 \times H_0^1(P_{h,R})^3 \mapsto \mathbb{R}$ with

$$A(V,V) \ge \alpha \cdot \|\nabla V\|^2_{2,P_{h,R}}, \quad |\, A(V,W)\,| \le \alpha^{-1} \cdot \|\nabla V\|_{2,P_{h,R}} \cdot \|\nabla W\|_{2,P_{h,R}} \tag{2.6}$$

for $V, W \in \mathcal{B}^3_{h,R}$, where the constant $\alpha \in (0,\infty)$ is to be independent of h and R. Let $H_0^1(P_{h,R})^3$ be equipped with the gradient norm, and denote the corresponding norm of $H^{-1}(P_{h,R})^3$ by $\|\ \|_{-1,2,P_{h,R}}$. Then by (2.6) and the Lax–Milgram theorem, there is a unique operator $\mathcal{R} := \mathcal{R}_A : H^{-1}(P_{h,R})^3 \mapsto \mathcal{B}^3_{h,R}$ such that

$$A\big(\,\mathcal{R}(F),\, W\big) = F(W) \quad \text{for }\ F \in H^{-1}(P_{h,R})^3,\ W \in \mathcal{B}^3_{h,R}, \tag{2.7}$$

and this operator verifies the following inequality:

$$\alpha \cdot \|\nabla \mathcal{R}(F)\|^2_{2,P_{h,R}} \le A\big(\,\mathcal{R}(F), \mathcal{R}(F)\,\big) = F\big(\,\mathcal{R}(F)\,\big) \le \|F\|_{-1,2,P_{h,R}} \cdot \|\nabla \mathcal{R}(F)\|_{2,P_{h,R}}.$$

Hence

$$\|\nabla \mathcal{R}(F)\|_{2,P_{h,R}} \le \alpha^{-1} \cdot \|F\|_{-1,2,P_{h,R}} \quad \text{for }\ F \in H^{-1}(P_{h,R})^3; \tag{2.8}$$

compare [17, p. 288]. In particular, $\mathcal{R}$ is continuous. It is easy to see that $\mathcal{R} \,|\, \mathcal{B}^3_{h,R}$ is one-to-one if a function $W \in \mathcal{B}^3_{h,R}$ is identified with the element F_W of $H^{-1}(P_{h,R})^3$ given by $F_W(v) := \int_{P_{h,R}} W \cdot v \, dx$ for $v \in H_0^1(P_{h,R})^3$. Thus we may think of $\mathcal{R}$ as a kind of interpolation operator from $H^{-1}(P_{h,R})^3$ into $\mathcal{B}^3_{h,R}$.

For $q \in L^2(P_{h,R})$, we may define a functional $\nabla q \in H^{-1}(P_{h,R})^3$ by setting

$$\nabla q(w) := -\int_{P_{h,R}} q \cdot \operatorname{div} w \, dx \quad \text{for }\ w \in H_0^1(P_{h,R})^3.$$

Then

$$\|\nabla q\|_{-1,2,P_{h,R}} \le 3 \cdot \|q\|_{2,P_{h,R}} \quad \text{for } q \in L^2(P_{h,R})^3. \tag{2.9}$$

When we write $\mathcal{R}(\nabla q)$ in the following, for some q in $L^2(P_{h,R})^3$ (in particular for $q \in M_{h,R}$), then the term ∇q is to be understood as a functional in the preceding sense. Observe that by (2.7),

$$A\big(\mathcal{R}(\nabla q), W\big) = -\int_{P_{h,R}} q \cdot \operatorname{div} W \, dx \quad \text{for } q \in M_{h,R},\ W \in \mathcal{B}^3_{h,R}. \tag{2.10}$$

Now we may formulate our finite element problem. It reads as follows: For $F \in Y'_{h,R}$, find $u_{h,R} = u_{h,R,\tau,A,F} \in V_{h,R}$, $\pi_{h,R} = \pi_{h,R,\tau,A,F} \in M_{h,R}$ such that

$$a(u_{h,R}, w) + c(w, \pi_{h,R}) = F(w) \quad \text{for } w \in Y_{h,R}, \tag{2.11}$$

$$c(u_{h,R}, q) = A\big(\mathcal{R}(\nabla \pi_{h,R}), \mathcal{R}(\nabla q)\big) \quad \text{for } q \in M_{h,R}, \tag{2.12}$$

$$u_{h,R} \,|\, \partial\Omega = (-1, 0, 0). \tag{2.13}$$

In [6], it was shown that a solution $(u_{h,R}, \pi_{h,R})$ to (2.11)–(2.13) exists and is unique. The term $A\big(\mathcal{R}(\nabla \pi_{h,R}), \mathcal{R}(\nabla q)\big)$ serves to circumvent the LBB condition ("pressure stabilization") and may be considered as a generalization of the static condensation operator used to eliminate the bubble functions of the mini element. In fact, if we make the most obvious choice for A, that is, $A(v, w) := \int_\Omega \nabla v \cdot \nabla w \, dx$ $(v, w \in H^1_0(P_{h,R})^3)$, then the term $A\big(\mathcal{R}(\nabla \pi_{h,R}), \mathcal{R}(\nabla q)\big)$ corresponds to just that operator (see [17, p. 304]). But instead of restricting ourselves to this special case, we chose the more abstract framework based on the bilinear symmetric form A with (2.6) because such a framework covers at least one more interesting example, namely, the Brezzi–Pitkäranta pressure stabilization (see [17, p. 304]) and may have other applications as well.

Our error estimates may now be stated in the form of the following theorem. Recall that the quantities σ_1, σ_2, φ_1, α, and $\mathcal{A}$ were introduced in (A3), (A6), (A8), (2.6), and (2.5), respectively.

THEOREM 2.1. *Let $h \in (0, h_0)$, $R \in (R_0, \infty)$, where $h_0 \in (0, S/8]$ and $R_0 \in [8 \cdot S, \infty)$ are constants depending only on Ω, S, σ_1, σ_2, and φ_1. Further suppose that $F(w) = \int_{P_{h,R}} f \cdot w \, dx$ for $w \in Y_{h,R}$, and that $(u_{h,R}, \pi_{h,R}) \in V_{h,R} \times M_{h,R}$ is a solution of (2.11)–(2.13). Then*

$$\begin{aligned} &\Big((\|u - u_{h,R}\|^{(h,R)})^2 + \tau \cdot \|u - u_{h,R}\|^2_{2,\partial P_{h,R}} \\ &\qquad + \| \nabla \mathcal{R}(\nabla(\pi - \pi_{h,R})) \|^2_{2,P_{h,R}} + \|\pi - \pi_{h,R}\|^2_{2,\Omega_{2\cdot S}} \Big)^{1/2} \\ &\qquad \le C_1(\tau, h \cdot R) \cdot \max\{1, \tau^{-1/2}\} \cdot \mathcal{A} \cdot (h^t + h \cdot \ln(R/S) + R^{-1}). \end{aligned} \tag{2.14}$$

The constant $C_1(\tau, h \cdot R)$ depends on Ω, S, σ_1, σ_2, φ_1, and α and is an increasing function of τ and $h \cdot R$.

A remark is perhaps in order with respect to our estimate of the pressure error $\pi - \pi_{h,R}$ in (2.14). In that inequality, the restriction of this error to $\Omega_{2\cdot S}$ is evaluated in the L^2-norm, and an interpolation of this error in $\mathcal{B}^3_{h,R}$ is estimated with respect to the H^1-norm. It remains an open problem whether the L^2-norm of $\pi - \pi_{h,R}$ in

$P_{h,R}$—instead of $\Omega_{2\cdot S}$—decays in the same way for $h+R^{-1}\to 0$, as does the left-hand side of (2.14), or whether it decays at all.

We remark that the notation $\partial_{h,R}$, $\mathcal{A}$, $\mathcal{R}$, $u_{h,R}$, $\pi_{h,R}$, $\Gamma_{h,R}$ will be used frequently in the following.

3. Auxiliary results. We will apply the following version of the LBB condition for the mini element.

THEOREM 3.1. *There are constants* $C_0\in(0,\infty)$, $\gamma_0\in(0,\,S/4]$, $R_0\in[8\cdot S,\,\infty)$ *such that for* $h\in(0,\,\gamma_0)$, $R\in(R_0,\infty)$, $\varrho\in M_{h,R}$,

$$\|\varrho\|_{2,P_{h,R}}\le C_0\cdot\sup\Big\{\int_{P_{h,R}}\varrho\cdot\operatorname{div}(w+W)\,dx/\|w+W\|^{(h,R)}\ :\ w\in Y_{h,R},\ W\in B^3_{h,R},\ w+W\neq 0\Big\}.$$

The constants C_0, γ_0, R_0 *depend on the parameters* σ_1, σ_2, *and* φ_1 (*see* (A3), (A6), (A8), *respectively*) *and on* Ω *and* S.

Note that the constant C_0 does not depend on either h or R. Theorem 3.1 is a slightly simplified form of [3, Theorem 4.1]. (The spherical boundary ∂B_R in [3] is replaced here by the polyhedral boundary $\partial_{h,R}$.) By [4, eqs. (2.3), (2.7); Corollary 2.2], there is a constant $S_1\in(0,\,S/8]$ such that

$$\int_{\partial_{h,R}}do\ \le\ 2\cdot\int_{\partial B_R}do\quad\text{for }\ h\in(0,S_1)\ \text{ and for }\ R\in(8\cdot S,\,\infty).\tag{3.1}$$

We put $h_0:=\min\{\gamma_0,S_0,S_1\}$, with S_0 from (2.4) and γ_0 from Theorem 3.1 above. This means in particular that $h_0<S/8$. In the following, we write $\mathcal{C}$ for constants which depend only on Ω, S, and the parameters $\sigma_1,\sigma_2,\varphi_1$, and α from (A3), (A6), (A8), and (2.6), respectively. By $\mathcal{C}(\tau)$ we denote constants which depend on $\Omega,S,\sigma_1,\sigma_2,\varphi_1,\alpha$ and which are increasing functions of τ. We write $\mathcal{C}(\tau,h\cdot R)$ for constants which, in addition, are increasing functions of $h\cdot R$. In the following, the quantities h and R are always assumed to belong to $(0,h_0)$ and (R_0,∞), respectively. We set

$$J:=J_R:=\min\big\{\,j\in\mathbb{N}\ :\ 2^j\cdot S\ge R\,\big\},\tag{3.2}$$

$$A_{h,R}:=\cup\big\{\,K_l\ :\ l\in\{1,\dots,k\}\text{ with }\ K_l\cap\partial B_R\neq\emptyset\,\big\}.$$

In the next lemma, we introduce functions from $M_{h,R}$ with support in a single layer of mesh cells near $\partial_{h,R}$; compare [6, Lemma 3.1] and the remarks on its proof given in [6].

LEMMA 3.1. *Let* $m\in M_{h,R}$. *Denote by* $\overline{m}$ *the uniquely determined element from* $M_{h,R}$ *such that for any* $l\in\{1,\dots,k\}$ *and for any vertex* x *of* K_l, *the relation* $\overline{m}(x)=m(x)$ *holds if* $x\in\partial B_R$, *and* $\overline{m}(x)=0$ *otherwise. Then*

$$m\,|\,\partial_{h,R}=\overline{m}\,|\,\partial_{h,R},\quad \operatorname{supp}(\overline{m})\subset A_{h,R},\quad \|\overline{m}\|_2\le\mathcal{C}\cdot\|m\|_{2,A_{h,R}}.$$

We further note the following consequences of (A3) and (A4). These consequences should be obvious, except perhaps relation (3.5), which may be proved in the same

way as [3, Lemma 4.2]:

$$\text{(3.3)}\quad diam\, K_l \le 2^J \cdot h \le 2 \cdot h \cdot R/S \ \text{ for } 1 \le l \le k; \quad A_{h,R} \subset \overline{P_{h,R}} \backslash B_{R\cdot(1-2\cdot h/S)};$$

$$\text{(3.4)}\quad (K_l)_\Delta \subset \overline{P_{h,R}} \backslash B_{R\cdot(1-4\cdot h/S)} \ \text{ for } \ l \in \{1, \dots, k\} \ \text{ with } \ K_l \cap \partial B_R \ne \emptyset;$$

$$\text{(3.5)}\quad \Omega_{R\cdot(1-h^2/S^2)^{1/2}} \subset P_{h,R} \subset \Omega_R, \quad \text{so } \ B_R \backslash (P_{h,R} \cup \Omega^c) \subset B_R \backslash B_{R\cdot(1-h^2/S^2)^{1/2}};$$

$$\text{(3.6)}\quad \sum_{l=1}^{k} \int_{(K_l)_\Delta} v\, dx \ \le \ \mathcal{C} \cdot \int_{P_{h,R}} v\, dx \quad \text{for any } \ v \in L^1(P_{h,R}) \text{ with } v \ge 0;$$

$$\text{(3.7)}\quad \sum_{l=1,\ K_l \cap U_j \ne \emptyset}^{k} \int_{(K_l)_\Delta} v\, dx \ \le \ \mathcal{C} \cdot \int_{(U_{j-1} \cup U_j \cup U_{j+1}) \cap P_{h,R}} v\, dx$$

$$\text{for } \ v \ \text{ as in } \ (3.6) \ \text{ and for } \ 0 \le j \le J;$$

$$\text{(3.8)}\quad \sum_{l=1,\ K_l \cap \partial B_R \ne \emptyset}^{k} \int_{(K_l)_\Delta} v\, dx \ \le \ \mathcal{C} \cdot \int_{P_{h,R} \backslash B_{R\cdot(1-4\cdot h/S)}} v\, dx \quad \text{for } \ v \ \text{ as in } \ (3.6);$$

$$\text{(3.9)}\quad \int_{P_{h,R}} v\, dx \ = \ \sum_{j=0}^{J} \int_{U_j \cap P_{h,R}} v\, dx \quad \text{for } \ v \in L^1(P_{h,R}).$$

Next we introduce two Clément-type interpolation operators. By [2, section 4.8], (A3), and (A6), there are linear operators $\Pi_{h,R} : H^1(P_{h,R})^3 \mapsto V_{h,R}$, $\widetilde{\Pi}_{h,R} : L^2(P_{h,R}) \mapsto M_{h,R}$ with $\Pi_{h,R}(w) \,|\, \partial\Omega = 0$ for $w \in W_{h,R}$; $\Pi_{h,R}(v) = v$ for $v \in V_{h,R}$; $\widetilde{\Pi}_{h,R}(\varrho) = \varrho$ for $\varrho \in M_{h,R}$,

$$\text{(3.10)}\quad |\Pi_{h,R}(w) - w|_{r,2,K_l} \ \le \ \widetilde{C} \cdot (diam\, K_l)^{\nu - r} \cdot |w|_{\nu,2,(K_l)_\Delta}$$

for $r \in \{0,1\}$, $\nu \in \{1,\, 2\}$, $w \in H^1(P_{h,R})^3$, $1 \le l \le k$ with $w \,|\, (K_l)_\Delta \in H^\nu\big((K_l)_\Delta\big)^3$; and

$$\text{(3.11)}\quad |\widetilde{\Pi}_{h,R}(\varrho) - \varrho|_{r,2,K_l} \ \le \ \widetilde{C} \cdot (diam\, K_l)^{1-r} \cdot |\varrho|_{1,2,(K_l)_\Delta}$$

for $r \in \{0,1\}$, $\varrho \in L^2(P_{h,R})$, $l \in \{1, \dots, k\}$ with $\varrho \,|\, (K_l)_\Delta \in H^1\big((K_l)_\Delta\big)$, where the constant $\widetilde{C}$ depends only on σ_1 and σ_2. For simplicity, we have written K_l and $(K_l)_\Delta$ instead of the interior of K_l and $(K_l)_\Delta$, respectively. Note that since $\Pi_{h,R}(v) \,|\, \partial\Omega = 0$ for $v \in W_{h,R}$, $\Pi_{h,R}(v) = v$ for $v \in V_{h,R}$, and because $u \,|\, \partial\Omega = (-1,0,0)$ on $\partial\Omega$, we have $\Pi_{h,R}(u) \,|\, \partial\Omega = (-1,0,0)$. Let us draw some conclusions from these relations.

Corollary 3.1. *The ensuing estimates are valid:*

$$\text{(3.12)}\quad \|\, \Pi_{h,R}(w) - w\|_{2,\partial_{h,R}} \ \le \ \mathcal{C} \cdot (h \cdot R)^{\nu - 1/2} \cdot |w|_{\nu,2,\, P_{h,R} \backslash B_{R\cdot(1-4\cdot h/S)}}$$

for $\nu \in \{1,\, 2\}$, $w \in W_{h,R}$ *with* $w \,|\, P_{h,R} \backslash B_{R\cdot(1-4\cdot h/S)} \in H^\nu(P_{h,R} \backslash B_{R\cdot(1-4\cdot h/S)})^3$;

$$\text{(3.13)}\quad |\, \Pi_{h,R}(w) - w\,|_{r,2,P_{h,R}} \ \le \ \mathcal{C} \cdot (h \cdot R)^{1-r} \cdot \|\nabla w\|_{2,P_{h,R}}$$

$$\text{for } \ w \in H^1(P_{h,R})^3, \ \ r \in \{0,\, 1\},$$

$$\text{(3.14)}\quad \|\Pi_{h,R}(w) - w\|_{6,K_l} \ \le \ \mathcal{C} \cdot (diam\, K_l) \cdot |w|_{2,2,(K_l)_\Delta}$$

$$\text{for } \ l \in \{1, \dots, k\}, \ w \in H^1(P_{h,R})^3 \ \text{ with } \ w \,|\, (K_l)_\Delta \in H^2((K_l)_\Delta)^3.$$

Proof. For the first inequality, we refer to (3.8) and to the proof of [4, Theorem 3.1]. The second is a consequence of (3.10), (3.6), and (3.3). The third may be proved by transforming its left-hand side into an integral over a standard tetrahedron, then using the Sobolev imbedding of H^1 into L^6 on this tetrahedron, returning to the domain of integration K_l, and finally applying (3.10). Also see [10, (I.A.14)]. ∎

COROLLARY 3.2. *For $r \in \{0,1\}$, $l \in \{1, \ldots, k\}$ with $K_l \cap \Omega_{2\cdot S} \neq \emptyset$, we have*

$$|\Pi_{h,R}(u) - u|_{r,2,K_l} \;\le\; \mathcal{C} \cdot h^{1+t-r} \cdot \|u\|_{1+t,\,2,\,(K_l)_\Delta},$$

$$\|\widetilde{\Pi}_{h,R}(\pi) - \pi\|_{2,K_l} \;\le\; \mathcal{C} \cdot h^{t} \cdot \|\pi\|_{t,2,\,(K_l)_\Delta}.$$

Proof. The corollary follows by interpolation from (3.10), (3.11), and (A3). ∎

COROLLARY 3.3. *Put $w := \Pi(u)_{h,R} - u$. Then, with $\mathcal{A}$ from (2.5),*

$$h^{-1-t} \cdot \|w\|_{2,\Omega_{2\cdot S}} + h^{-1/2-t} \cdot \|w\|_{3,\Omega_{2\cdot S}} + h^{-t} \cdot \|\nabla w\|_{2,\Omega_{2\cdot S}} \;\le\; \mathcal{C} \cdot \mathcal{A}; \tag{3.15}$$

$$(h^2 \cdot 2^j)^{-1} \cdot \|w\|_{2,\,U_j \cap P_{h,R}} \;+\; (h^{3/2} \cdot 2^{j/2})^{-1} \cdot \|w\|_{3,\,U_j \cap P_{h,R}}$$
$$+\; h^{-1} \cdot \big(\|w\|_{6,\,U_j \cap P_{h,R}} + \|\nabla w\|_{2,\,U_j \cap P_{h,R}}\big) \;\le\; \mathcal{C} \cdot \mathcal{A} \qquad \textit{for } j \in \{2, \ldots, J\}; \tag{3.16}$$

$$\|w\|_{6,P_{h,R}} + \|\nabla w\|_{2,P_{h,R}} + \|\widetilde{\Pi}_{h,R}(\pi) - \pi\|_{2,P_{h,R}} \;\le\; \mathcal{C} \cdot \mathcal{A} \cdot \big(\, h^t + h \cdot \ln(R/S)\,\big); \tag{3.17}$$

$$\|w\|_{3,P_{h,R}} \le \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot (h^{3/2} \cdot R^{1/2} + h^{t+1/2});\; \|w\|_{2,P_{h,R}} \le \mathcal{C} \cdot \mathcal{A} \cdot (h^2 \cdot R + h^{t+1}); \tag{3.18}$$

$$\|w\|_{2,\partial_{h,R}} \;\le\; \mathcal{C} \cdot \mathcal{A} \cdot h^{3/2}. \tag{3.19}$$

Proof. Denote $\mathcal{I}_0 := \big\{\, l \in \{1, \ldots, k\} \;:\; K_l \cap \Omega_{2\cdot S} \neq \emptyset \,\big\}$. Then

$$\|w\|_{2,\,\Omega_{2\cdot S}} \le \left(\sum_{l \in \mathcal{I}_0} \|w\|^2_{2,K_l}\right)^{1/2} \;\le\; \mathcal{C} \cdot h^{1+t} \cdot \left(\sum_{l \in \mathcal{I}_0} \|u\|^2_{1+t,\,2,\,(K_l)_\Delta}\right)^{1/2} \tag{3.20}$$
$$\le \mathcal{C} \cdot h^{1+t} \cdot \|u\|_{1+t,\,2,\,\Omega_{4\cdot S}} \;\le\; \mathcal{C} \cdot \mathcal{A} \cdot h^{1+t},$$

where we used Corollary 3.2, (3.7), (2.1), and (2.5). The term $\|\nabla w\|_{2,\,\Omega_{2\cdot S}}$ may be evaluated in the same way. Noting that $\|w\|_{6,\,\Omega_{2\cdot S}} \le \mathcal{C} \cdot \|w\|_{1,2,\,\Omega_{2\cdot S}}$ according to a Sobolev inequality, we obtain an estimate of $\|w\|_{3,\,\Omega_{2\cdot S}}$ by interpolation between an L^6- and an L^2-estimate. A similar computation as in (3.20) yields, in view of (3.10), (A3), and (3.7):

$$\|w\|_{2,\,U_j \cap P_{h,R}} \;\le\; \mathcal{C} \cdot 2^{2\cdot j} \cdot h^2 \cdot |u|_{2,\,2,\,U_{j-1} \cup U_j \cup U_{j+1}} \quad \text{for } j \in \{2, \ldots, J\}.$$

It follows by (2.3) and (2.5) that $\|w\|_{2,\,U_j \cap P_{h,R}} \;\le\; \mathcal{C} \cdot \mathcal{A} \cdot h^2 \cdot 2^j$. This explains the estimate of $(h^2 \cdot 2^j)^{-1} \cdot \|w\|_{2,\,U_j \cap P_{h,R}}$ in (3.16). The other terms on the left-hand side of that inequality may be dealt with in a similar way. (Use (3.14) for the L^6-estimate, and use interpolation for the estimate of the L^3-norm.) The estimates of w and ∇w in (3.17) and (3.18) follow from (3.16), (3.15), (3.9), and the inequality $J \le \mathcal{C} \cdot \ln(R/S)$ (see (3.2)). Estimate (3.12) yields $\|w\|_{2,\partial_{h,R}} \;\le\; \mathcal{C} \cdot (h \cdot R)^{3/2} \cdot |u|_{2,\,2,\,B_R \setminus B_{R\cdot(1-4\cdot h/S)}}$. This observation, (2.3) with $\delta = 1 - 4 \cdot h/S$, and (2.5) yield inequality (3.19). The term

$\|\widetilde{\Pi}_{h,R}(\pi)-\pi\|_{2,P_{h,R}}$ may be evaluated by using Corollary 3.2, (3.11), an argument as in (3.20), as well as (3.7), (2.1), (2.3), and (2.5). ∎

LEMMA 3.2. *Let* $W \in \mathcal{B}^3_{h,R}$. *Then*

$$\|\nabla W\|_{2,K_l} \;\le\; \mathcal{C}\cdot(\mathit{diam}\, K_l)^{-5/2}\cdot\Big|\int_{K_l} W\; dx\Big| \quad \textit{for } l\in\{1,\dots,k\}; \tag{3.21}$$

$$\|W\|_{2,\,U_j\cap P_{h,R}} \;\le\; \mathcal{C}\cdot 2^j\cdot h\cdot\|\nabla W\|_{2,\,(U_{j-1}\cap U_j\cap U_{j+1})\cap P_{h,R}} \quad \textit{for } 1\le j\le J; \tag{3.22}$$

$$\|W\|_{2,P_{h,R}} \;\le\; \mathcal{C}\cdot h\cdot R\cdot\|\nabla W\|_{2,P_{h,R}}. \tag{3.23}$$

Proof. The lemma may be proved via transformations to a reference tetrahedron; compare [17, Lemma 4.1b]. ∎

4. Estimate of the velocity error by the pressure error. In this section, the velocity error $u-u_{h,R}$ and an interpolation in $\mathcal{B}^3_{h,R}$ of the gradient of the pressure error $\pi-\pi_{h,R}$ are estimated by a quantity depending on $\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)$ (see Theorem 4.1). Since the latter term differs from the pressure error only by the interpolation error $\pi-\widetilde{\Pi}_{h,R}(\pi)$, our result may thus be considered as an estimate of the velocity error by the pressure error.

In the following lemma, the velocity error and the interpolation via $\mathcal{R}$ of $\nabla(\pi-\pi_{h,R})$ in $\mathcal{B}^3_{h,R}$ are estimated by a sum of seven terms $|\mathcal{N}_1|$ to $|\mathcal{N}_7|$; these terms will be evaluated in the proof of Theorem 4.1. The main difficulty consists in dealing with the term $\mathcal{N}_4$, because we were not able to prove that $\|\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}}$ decreases for $h+1/R\to 0$ (section 5). Therefore we had to estimate $|\mathcal{N}_4|$ in such a way that we obtained an upper bound in which the critical quantity $\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)$ is multiplied by negative powers of R and/or positive powers of h.

LEMMA 4.1. *Put*

$$\mathcal{N}_1 := \big(\,\|u-\Pi_{h,R}(u)\|^{(h,R)}\,\big)^2 \;+\; \tau\cdot\|u-\Pi_{h,R}(u)\|^2_{2,\partial_{h,R}},$$

$$\mathcal{N}_2 := a\big(\,\Pi_{h,R}(u)-u,\,\Pi_{h,R}(u)-u_{h,R}\,\big), \quad \mathcal{N}_3 := c\big(\,\Pi_{h,R}(u)-u_{h,R},\,\widetilde{\Pi}_{h,R}(\pi)-\pi\,\big),$$

$$\mathcal{N}_4 := c\big(\,\Pi_{h,R}(u)-u,\,\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)\,\big),$$

$$\mathcal{N}_5 := \big|\,\|\nabla\mathcal{R}\big(\,\nabla(\pi-\pi_{h,R})\,\big)\|^2_{2,P_{h,R}} - \|\nabla\mathcal{R}\big(\,\nabla(\widetilde{\Pi}_{h,R}(\pi)-\pi_{h,R})\,\big)\|^2_{2,P_{h,R}}\,\big|,$$

$$\mathcal{N}_6 := A\Big(\,\mathcal{R}\big(\,\nabla\widetilde{\Pi}_{h,R}(\pi)\,\big),\,\mathcal{R}\big(\,\nabla(\widetilde{\Pi}_{h,R}(\pi)-\pi_{h,R})\,\big)\,\Big),$$

$$\mathcal{N}_7 := \int_{\partial_{h,R}}\sum_{k=1}^{3}\left(\sum_{j=1}^{3}(\partial_j u_k-\delta_{jk}\cdot\pi)\cdot n_j \;+\; \big(\,R^{-1}+(\tau/2)\cdot(1-n_1)\,\big)\cdot u_k\right)$$
$$\cdot(\Pi_{h,R}(u)-u_{h,R})_k\; do.$$

Then

$$(\|u-u_{h,R}\|^{(h,R)})^2+\tau\cdot\|u-u_{h,R}\|^2_{2,\partial P_{h,R}}+\|\nabla\mathcal{R}\big(\nabla(\pi-\pi_{h,R})\big)\|^2_{2,P_{h,R}}\le\mathcal{C}\cdot\sum_{i=1}^{7}|\mathcal{N}_i|.$$

Proof. Denote

$$z := \Pi_{h,R}(u) - u_{h,R}, \quad I := \big(\|u - u_{h,R}\|^{(h,R)} \big)^2 + \tau \cdot \|u - u_{h,R}\|^2_{2,\partial_{h,R}}.$$

Then

$$I \le 2 \cdot \big(\mathcal{N}_1 + (\|z\|^{(h,R)})^2 + \tau \cdot \|z\|^2_{2,\partial_{h,R}} \big) \ \le \ \mathcal{C} \cdot \big(\mathcal{N}_1 + \mathcal{N}_2 + a(u - u_{h,R},\, z) \big)$$

$$= \mathcal{C} \cdot \left[\mathcal{N}_1 + \mathcal{N}_2 + \mathcal{N}_7 + \int_{P_{h,R}} (-\Delta u + \tau \cdot \partial_1 u + \nabla \pi) \cdot z \; dx - c(z, \pi) - a(u_{h,R}, z) \right].$$

Lemma 4.1 follows from this estimate, (2.6) , (1.1), and (2.11)–(2.13). ∎

Next we establish three auxiliary results which will be needed in order to evaluate $\mathcal{N}_4$ and $\mathcal{N}_6$.

LEMMA 4.2. *Define a function* $W(u) := W_{h,R}(u) \in \mathcal{B}^3_{h,R}$ *by setting*

$$W(u)_i(x) := \left(\int_{K_l} b_{K_l} \, dy \right)^{-1} \cdot \int_{K_l} (\Pi_{h,R}(u) - u)_i \, dy \cdot b_{K_l}(x)$$

for $x \in K_l$, $l \in \{1, \ldots, k\}$, $i \in \{1, 2, 3\}$, *where* b_{K_l} *is as introduced in section* 2. *Then we have for* $m \in M_{h,R}$, *with* $\overline{m}$ *defined in Lemma* 3.1,

$$|c(\Pi_{h,R}(u) - u, \, \overline{m})| \ + \ |c(\, W(u), \, \overline{m}\,)| \ \le \ \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot h \cdot R^{-1/2} \cdot \|\overline{m}\|_{2,P_{h,R}}.$$

Proof. Abbreviate $I(h,R) := \big\{ \, l \in \{1, \ldots, k\} \, : \, K_l \cap \partial B_R \neq \emptyset \, \big\}$. Applying (3.10), (3.3), (3.8), (2.3) with $\delta = 1 - 4 \cdot h/S$, and (2.5), we get

$$|c(\Pi_{h,R}(u) - u, \, \overline{m})| \ \le \ \mathcal{C} \cdot \left(\sum_{l \in I(h,R)} \|\nabla(\Pi_{h,R}(u) - u)\|^2_{2,K_l} \right)^{1/2} \cdot \|\overline{m}\|_{2,P_{h,R}}$$

$$\le \mathcal{C} \cdot h \cdot R \cdot |u|_{2,2,\, P_{h,R} \setminus B_{R \cdot (1 - 4 \cdot h/S)}} \cdot \|\overline{m}\|_{2,P_{h,R}} \ \le \ \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot h \cdot R^{-1/2} \cdot \|\overline{m}\|_{2,P_{h,R}}.$$

We further observe that $|c\big(\, W(u), \, \overline{m}\,\big)| \le \mathcal{C} \cdot \|\nabla W(u)\|_{2,A_{h,R}} \cdot \|\overline{m}\|_{2,P_{h,R}}$, with $A_{h,R}$ as defined in (3.2). Moreover, by (3.21), the definition of $W(u)$, (3.10), (3.3), (3.8), (2.3) with $\delta = 1 - 4 \cdot h/S$, and (2.5), we find

$$\|\nabla W(u)\|^2_{2,A_{h,R}} = \sum_{l \in I(h,R)} \|\nabla W(u)\|^2_{2,K_l} \ \le \ \mathcal{C} \cdot \sum_{l \in I(h,R)} (diam\, K_l)^{-5} \cdot \left| \int_{K_l} W(u) \; dx \right|^2$$

$$= \mathcal{C} \cdot \sum_{l \in I(h,R)} (diam\, K_l)^{-5} \cdot \left| \int_{K_l} (\Pi_{h,R}(u) - u) \; dx \right|^2$$

$$\le \mathcal{C} \cdot \sum_{l \in I(h,R)} (diam\, K_l)^{-2} \cdot \|\Pi_{h,R}(u) - u\|^2_{2,K_l}$$

$$\le \mathcal{C} \cdot h^2 \cdot R^2 \cdot |u|^2_{2,\, 2,\, P_{h,R} \setminus B_{R \cdot (1 - 4 \cdot h/S)}} \ \le \ \mathcal{C}(h \cdot R) \cdot \mathcal{A}^2 \cdot h^2 \cdot R^{-1}.$$

The lemma follows from the preceding inequalities. ∎

LEMMA 4.3. *Define the function $W(u)$ as in Lemma 4.2. Then*

$$|c\big(W(u),\,\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)\big)|$$
$$\leq\ \mathcal{C}\cdot\mathcal{A}\cdot\big(h^t+h\cdot\ln(R/S)\big)\cdot\|\nabla\mathcal{R}\big(\nabla(\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi))\big)\|_{2,P_{h,R}}.$$

Proof. With the abbreviation $\mathcal{T}:=\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)$, we find by (2.10) and (2.6) that

(4.1)
$$\big|\,c\big(W(u),\,\mathcal{T}\big)\big|=|A\big(\mathcal{R}(\nabla\mathcal{T}),\,W(u)\big)|\leq\mathcal{C}\cdot\|\nabla\mathcal{R}(\nabla\mathcal{T})\|_{2,P_{h,R}}\cdot\|\nabla W(u)\|_{2,P_{h,R}}.$$

By the same reasoning as in the proof of Lemma 4.2, we get

$$\|\nabla W(u)\|_{2,P_{h,R}}\ \leq\ \mathcal{C}\cdot\left(\sum_{l=1}^{k}(diam\,K_l)^{-2}\cdot\|\Pi_{h,R}(u)-u\|^2_{2,K_l}\right)^{1/2}. \tag{4.2}$$

Abbreviate $Z:=\big\{\,l\in\{1,\,\dots,\,k\}\ :\ K_l\subset B_{2\cdot S}\,\big\}$. We have $diam\,K_l\geq\mathcal{C}\cdot h$ for $l\in Z$ by (A3), so with (3.15) we get

$$\sum_{l\in Z}(diam\,K_l)^{-2}\cdot\|\Pi_{h,R}(u)-u\|^2_{2,K_l}\ \leq\ \mathcal{C}\cdot h^{-2}\cdot\|\Pi_{h,R}(u)-u\|^2_{2,\Omega_{2\cdot S}}\ \leq\ \mathcal{C}\cdot\mathcal{A}^2\cdot h^{2\cdot t}.$$

On the other hand, using (3.10), (A3), and (3.7), we find

$$\sum_{l=1,\,l\notin Z}^{k}(diam\,K_l)^{-2}\cdot\|\Pi_{h,R}(u)-u\|^2_{2,K_l}\ \leq\ \sum_{l=1,\,l\notin Z}^{k}(diam\,K_l)^{2}\cdot|u|^2_{2,\,2,\,(K_l)_\Delta}$$
$$\leq\ \sum_{j=2}^{J}(2^j\cdot h)^2\cdot\sum_{l=1,\,K_l\cap U_j\neq\emptyset}^{k}\big|u\big|^2_{2,\,2,\,(K_l)_\Delta}\ \leq\ \sum_{j=2}^{J}(2^j\cdot h)^2\cdot\big|u\big|^2_{2,\,2,\,U_{j-1}\cup U_j\cup U_{j+1}}.$$

But for $j\in\{2,\,\dots,\,J\}$, we have $\big|u\big|_{2,\,2,\,U_{j-1}\cup U_j\cup U_{j+1}}\ \leq\ \mathcal{C}\cdot\mathcal{A}\cdot2^{-j}$ by (2.3) and (2.5). Combining the preceding estimates beginning with (4.2), we obtain

$$\|\nabla W(u)\|_{2,P_{h,R}}\ \leq\ \mathcal{C}\cdot\mathcal{A}\cdot(h^t+J\cdot h)\ \leq\ \mathcal{C}\cdot\mathcal{A}\cdot\big(h^t+h\cdot\ln(R/S)\big).$$

Lemma 4.3 now follows with (4.1). ∎

LEMMA 4.4. *Let $W\in\mathcal{B}^3_{h,R}$. Then*

$$\Big|\int_{P_{h,R}}\pi\cdot\operatorname{div}W\,dx\Big|\ \leq\ \mathcal{C}\cdot\mathcal{A}\cdot\big(h^t+h\cdot\ln(R/S)\big)\cdot\|\nabla W\|_{2,P_{h,R}}.$$

Proof. Let B denote the interior of the set $\cup\big\{\,K_l:1\leq l\leq k,\ K_l\cap\overline{\Omega_S}\neq\emptyset\,\big\}$. Note that $\overline{B}\subset\overline{U_0}\cup\overline{U_1}=\overline{\Omega_{2\cdot S}}$ by (A3). Define operators $T_1:L^2(B)\mapsto\mathbb{R}$, $T_2:H^1(B)\mapsto\mathbb{R}$ by $T_i(z):=\int_B z\cdot\operatorname{div}W\,dx$, with $z\in L^2(B)$ in the case $i=1$, and $z\in H^1(B)$ if $i=2$. Observe that

$$T_1\,|\,H^1(B)=T_2,\qquad|T_1(z)|\ \leq\ \mathcal{C}\cdot\|z\|_{2,B}\cdot\|\nabla W\|_{2,P_{h,R}}\quad\text{for }\ z\in L^2(B),$$

$$|T_2(z)|\ =\ \Big|\int_B\nabla z\cdot W\,dx\Big|\ \leq\ \mathcal{C}\cdot\|\nabla z\|_{2,B}\cdot\|W\|_{2,B}\ \leq\ \mathcal{C}\cdot\|\nabla z\|_{2,B}\cdot\|\nabla W\|_{2,P_{h,R}}\cdot h$$

for $z \in H^1(B)$, where we refer to (3.22) for the last inequality. It follows by interpolation (see [2, Theorem 14.2.3, Proposition 14.1.5], for example) that

(4.3)
$$\Big| \int_B \pi \cdot \operatorname{div} W \, dx \Big| \le \mathcal{C} \cdot \|\pi\|_{t,2,B} \cdot \|\nabla W\|_{2,P_{h,R}} \cdot h^t \le \mathcal{C} \cdot \mathcal{A} \cdot \|\nabla W\|_{2,P_{h,R}} \cdot h^t,$$

where we used the relation $\|\pi\|_{t,2,B} \le \|\pi\|_{t,2,\Omega_{2\cdot S}} \le \mathcal{A}$ (see (2.1), (2.5)) in the last inequality. On the other hand, since $P_{h,R}\backslash B \subset P_{h,R}\backslash\Omega_S = P_{h,R}\backslash U_0$, with a partial integration and by (2.3), (2.5) we get

$$\Big| \int_{P_{h,R}\backslash B} \pi \cdot \operatorname{div} W \, dx \Big| \le \sum_{j=1}^{J} \int_{U_j \cap P_{h,R}} |\nabla \pi| \cdot |W| \, dx \le \mathcal{C} \cdot \mathcal{A} \cdot \sum_{j=1}^{J} 2^{-j} \cdot \|W\|_{2,\,U_j \cap P_{h,R}}. \tag{4.4}$$

Combining (4.3), (4.4), and (3.22) yields the lemma. ∎

Now we are in a position to prove the main result of this section.

THEOREM 4.1. *Let $\mathcal{D}_{h,R}$ be an abbreviation of the term*

$$\Big((\|u - u_{h,R}\|^{(h,R)})^2 + \tau \cdot \|u - u_{h,R}\|^2_{2,\partial P_{h,R}} + \| \nabla \mathcal{R}(\nabla(\pi - \pi_{h,R})) \|^2_{2,P_{h,R}} \Big)^{1/2}.$$

Then

$$\begin{aligned} \mathcal{D}^2_{h,R} \le\ & \mathcal{C}(\tau, h \cdot R) \cdot \max\{1, \tau^{-1/2}\} \cdot \mathcal{A} \cdot (h^t + h \cdot \ln(R/S) + R^{-1}) \\ & \cdot \big(\mathcal{D}_{h,R} + \mathcal{A} \cdot (h^t + h \cdot \ln(R/S)) \big) \\ & + \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot h \cdot R^{-1} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_2 \\ & + \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot h \cdot R^{-1/2} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi) - m_{h,R}\|_2, \end{aligned}$$

with $m_{h,R} := |P_{h,R}|^{-1} \cdot \int_{P_{h,R}} (\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)) \, dx$ (mean value of $\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)$). Moreover,

$$\|\nabla \mathcal{R}\big(\nabla \big(\widetilde{\Pi}_{h,R}(\pi) - \pi_{h,R} \big) \big)\|_{2,P_{h,R}} \le \mathcal{C} \cdot \big(\mathcal{D}_{h,R} + \mathcal{A} \cdot (h^t + h \cdot \ln(R/S)) \big). \tag{4.5}$$

Proof. Abbreviate $\mathcal{H} := h^t + h \cdot \ln(R/S)$. Let us estimate the terms $\mathcal{N}_1$ to $\mathcal{N}_7$ introduced in Lemma 4.1. We begin by observing that as a direct consequence of (3.17) and (3.19),

$$|\mathcal{N}_1| \le \mathcal{C}(\tau) \cdot (\mathcal{A} \cdot \mathcal{H})^2.$$

In view of finding an upper bound for $|\mathcal{N}_2|$, we remark that $\|\Pi_{h,R}(u) - u\|^{(h,R)} \le \mathcal{C} \cdot \mathcal{A} \cdot \mathcal{H}$ by (3.17) and (3.19). Hence

$$\|\Pi_{h,R}(u) - u_{h,R}\|^{(h,R)} \le \|u - u_{h,R}\|^{(h,R)} + \mathcal{C} \cdot \mathcal{A} \cdot \mathcal{H} \le \mathcal{D}_{h,R} + \mathcal{C} \cdot \mathcal{A} \cdot \mathcal{H}. \tag{4.6}$$

On the other hand, after a partial integration of the Oseen term, we get

$$\begin{aligned}|\mathcal{N}_2| \;\le\; \mathcal{C}\cdot\big(&\|\nabla\big(\,\Pi_{h,R}(u)-u\,\big)\|_{2,P_{h,R}}\cdot\;\|\nabla\big(\,\Pi_{h,R}(u)-u_{h,R}\,\big)\|_{2,P_{h,R}}\\ &+(R^{-1}+\tau)\cdot\|\Pi_{h,R}(u)-u\|_{2,\partial_{h,R}}\cdot\|\Pi_{h,R}(u)-u_{h,R}\|_{2,\partial_{h,R}}\\ &+\tau\cdot\|\Pi_{h,R}(u)-u\|_{2,P_{h,R}}\cdot\|\nabla\big(\,\Pi_{h,R}(u)-u_{h,R}\,\big)\|_{2,P_{h,R}}\big).\end{aligned}$$

This estimate, (3.17)–(3.19), and (4.6) imply

$$|\mathcal{N}_2| \;\le\; \mathcal{C}(\tau,h\cdot R)\cdot\mathcal{A}\cdot\mathcal{H}\cdot(\mathcal{D}_{h,R}+\mathcal{A}\cdot\mathcal{H}).$$

As an immediate consequence of (3.17) and (4.6), we get

$$|\mathcal{N}_3| \;\le\; \mathcal{C}\cdot\mathcal{A}\cdot\mathcal{H}\cdot(\mathcal{D}_{h,R}+\mathcal{A}\cdot\mathcal{H}).$$

In order to deal with $\mathcal{N}_5$ and $\mathcal{N}_6$, denote

$$\begin{aligned}&\mathcal{R}^{(1)}:=\mathcal{R}\big(\,\nabla(\widetilde{\Pi}_{h,R}(\pi)-\pi_{h,R})\,\big),\;\;\mathcal{R}^{(2)}:=\mathcal{R}\big(\,\nabla(\pi-\pi_{h,R})\,\big),\\ &\mathcal{R}^{(3)}:=\mathcal{R}\big(\,\nabla(\widetilde{\Pi}_{h,R}(\pi)-\pi)\,\big),\end{aligned}$$

and observe that $\|\nabla\mathcal{R}^{(2)}\|_{2,P_{h,R}}\le\mathcal{D}_{h,R}$. We find

$$\begin{aligned}|\mathcal{N}_5|&=\big(\,\|\nabla\mathcal{R}^{(1)}\|_{2,P_{h,R}}+\|\nabla\mathcal{R}^{(2)}\|_{2,P_{h,R}}\,\big)\cdot\|\nabla\mathcal{R}^{(3)}\|_{2,P_{h,R}}\\ &\le\big(\,2\cdot\|\nabla\mathcal{R}^{(2)}\|_{2,P_{h,R}}+\|\nabla\mathcal{R}^{(3)}\|_{2,P_{h,R}}\,\big)\cdot\|\nabla\mathcal{R}^{(3)}\|_{2,P_{h,R}}.\end{aligned}$$

On the other hand, by (2.8), (2.9), and (3.17),

$$\|\nabla\mathcal{R}^{(3)}\|_{2,P_{h,R}}\le\mathcal{C}\cdot\|\,\widetilde{\Pi}_{h,R}(\pi)-\pi\|_{2,P_{h,R}}\le\mathcal{C}\cdot\mathcal{A}\cdot\mathcal{H}.$$

Thus we get

$$|\mathcal{N}_5|\le\mathcal{C}\cdot(\mathcal{D}_{h,R}+\mathcal{A}\cdot\mathcal{H})\cdot\mathcal{A}\cdot\mathcal{H};\quad\|\nabla\mathcal{R}^{(1)}\|_{2,P_{h,R}}\le\mathcal{C}\cdot(\mathcal{D}_{h,R}+\mathcal{A}\cdot\mathcal{H}).$$

In particular we have shown (4.5). By (2.10), the relation

$$|\mathcal{N}_6|=\Big|\int_{P_{h,R}}\widetilde{\Pi}_{h,R}(\pi)\cdot\operatorname{div}\mathcal{R}^{(1)}\,dx\Big|\;\le\;|\mathcal{I}_1|+|\mathcal{I}_2| \tag{4.7}$$

holds, with

$$\mathcal{I}_1:=\int_{P_{h,R}}\big(\,\widetilde{\Pi}_{h,R}(\pi)-\pi\,\big)\cdot\operatorname{div}\mathcal{R}^{(1)}\,dx,\quad\mathcal{I}_2:=\int_{P_{h,R}}\pi\cdot\operatorname{div}\mathcal{R}^{(1)}\,dx.$$

Due to (3.17), we find

$$|\mathcal{I}_1|\;\le\;\mathcal{C}\cdot\|\widetilde{\Pi}_{h,R}(\pi)-\pi\|_{2,P_{h,R}}\cdot\|\nabla\mathcal{R}^{(1)}\|_{2,P_{h,R}}\;\le\;\mathcal{C}\cdot\mathcal{A}\cdot\mathcal{H}\cdot\|\nabla\mathcal{R}^{(1)}\|_{2,P_{h,R}}.$$

By combining this inequality with the estimate of $\mathcal{I}_2$ in Lemma 4.4 and then applying (4.5), we obtain

$$|\mathcal{N}_6|\;\le\;\mathcal{C}\cdot\mathcal{A}\cdot\mathcal{H}\cdot\|\nabla\mathcal{R}^{(1)}\|_{2,P_{h,R}}\;\le\;\mathcal{C}\cdot\mathcal{A}\cdot\mathcal{H}\cdot(\mathcal{A}\cdot\mathcal{H}+\mathcal{D}_{h,R}).$$

Now we turn to the term $\mathcal{N}_4$. For $q \in M_{h,R}$, put $m(q) := |P_{h,R}|^{-1} \cdot \int_{P_{h,R}} q\, dx$ (mean value of q). This means that $m_{h,R} = m(\,\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\,)$, where $m_{h,R}$ is as introduced in Theorem 4.1. Since $R \geq R_0 \geq 8 \cdot S$ (see Theorem 3.1) and $\Omega^c \subset B_S$, and because of (3.5), we have $|P_{h,R}| \geq \mathcal{C} \cdot R^3$. Hence with Hölder's inequality,

$$|m_{h,R}| \leq |P_{h,R}|^{-1/2} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}} \leq \mathcal{C} \cdot R^{-3/2} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}}. \tag{4.8}$$

Now by a partial integration we obtain

$$\begin{aligned} |c\big(\,\Pi_{h,R}(u) - u,\, m_{h,R}\,\big)| &= \Big| \int_{\partial_{h,R}} \big(\,(\Pi_{h,R}(u) - u) \cdot n\,\big) \cdot m_{h,R}\; do \Big| \\ &\leq\; \mathcal{C} \cdot |m_{h,R}| \cdot \|\Pi_{h,R}(u) - u\|_{2,\partial_{h,R}} \cdot \left(\int_{\partial_{h,R}} do \right)^{1/2} \\ &\leq\; \mathcal{C} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}} \cdot \|\Pi_{h,R}(u) - u\|_{2,\partial_{h,R}} \cdot R^{-1/2}. \end{aligned} \tag{4.9}$$

The last inequality follows from (4.8) and (3.1). Inequalities (4.9) and (3.19) imply

$$\begin{aligned} |c(\Pi_{h,R}(u) - u,\, m_{h,R})| &\leq \mathcal{C} \cdot \mathcal{A} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}} \cdot h^{3/2} \cdot R^{-1/2} \\ &\leq \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot \|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}} \cdot h/R. \end{aligned} \tag{4.10}$$

Next define $W(u)$ as in Lemma 4.2, and abbreviate $p := \pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi) - m_{h,R}$. Recall that $\mathcal{R}^{(1)}$ is an abbreviation for $\mathcal{R}(\nabla(\widetilde{\Pi}_{h,R}(\pi) - \pi_{h,R}))$. Then

$$\begin{aligned} |\mathcal{N}_4| \leq\; & |c\big(\,\Pi_{h,R}(u) - u,\, p - \overline{p}\,\big)| + |c\big(\,\Pi_{h,R}(u) - u,\, \overline{p}\,\big)| \\ & + |c\big(\,\Pi_{h,R}(u) - u,\, m_{h,R}\,\big)|, \end{aligned} \tag{4.11}$$

with $\overline{p}$ defined as in Lemma 3.1. Using the main trick from the proof of the stability of the mini element, and recalling that $(p - \overline{p})\,|\,\partial_{h,R} = 0$ and $(\Pi_{h,R}(u) - u)\,|\,\partial\Omega = 0$, by a partial integration and by the definition of the function $W(u) \in \mathcal{B}^3_{h,R}$, we get

$$|c\big(\,\Pi_{h,R}(u) - u,\, p - \overline{p}\,\big)| \;=\; \left| \sum_{l=1}^{k} \nabla(p - \overline{p}) \cdot \int_{K_l} W(u)\; dx \right| \;=\; |c\big(\,W(u),\, p - \overline{p}\,\big)|.$$

It follows that

$$\begin{aligned} |c\big(\,\Pi_{h,R}(u) - u,\, p - \overline{p}\,\big)| &\leq |c\big(\,W(u),\, p\,\big)| + |c\big(\,W(u),\, \overline{p}\,\big)| \\ &= |c\big(\,W(u),\, \pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\,\big)| + |c\big(\,W(u),\, \overline{p}\,\big)|, \end{aligned}$$

where the last equation holds since $W(u) \in \mathcal{B}^3_{h,R}$. Now apply Lemmas 4.3 and 4.2 to obtain

$$|c\big(\,\Pi_{h,R}(u) - u,\, p - \overline{p}\,\big)| \;\leq\; \mathcal{C}(h \cdot R) \cdot \mathcal{A} \cdot \big(\,\mathcal{H} \cdot \|\nabla \mathcal{R}^{(1)}\|_{2,P_{h,R}} \;+\; \|\overline{p}\|_{2,P_{h,R}} \cdot h \cdot R^{-1/2}\,\big). \tag{4.12}$$

But the terms $|c\big(\Pi_{h,R}(u)-u,\,\overline{p}\,\big)|$ and $|c\big(\Pi_{h,R}(u)-u,\,m_{h,R}\,\big)|$ were already estimated in Lemma 4.2 and (4.10), respectively. We further take into account that $\|\overline{p}\|_{2,P_{h,R}} \le \|p\|_{2,P_{h,R}}$ by Lemma 3.1. Combining these estimates with (4.5), (4.11), and (4.12), we get

$$|\mathcal{N}_4| \le \mathcal{C}(h\cdot R)\cdot\mathcal{A}\cdot\Big(\mathcal{H}\cdot(\mathcal{A}\cdot\mathcal{H}+\mathcal{D}_{h,r}) + h\cdot R^{-1}\cdot\|\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}} + h\cdot R^{-1/2}\cdot\|\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)-m_{h,R}\|_{2,P_{h,R}}\Big).$$

We further observe that by (2.4) and (2.5),

$$|\mathcal{N}_7| \;\le\; \mathcal{C}(h\cdot R)\cdot\mathcal{A}\cdot R^{-1}\cdot\|\Pi_{h,R}(u)-u_{h,R})\|_{2,\partial_{h,R}}.$$

On the other hand, by referring to (3.19) we find

$$\|\Pi_{h,R}(u)-u_{h,R}\|_{2,\partial_{h,R}} \le \|\Pi_{h,R}(u)-u\|_{2,\partial_{h,R}} \;+\; \|u-u_{h,R}\|_{2,\partial_{h,R}} \le \mathcal{C}\cdot\mathcal{A}\cdot h^{3/2}+\tau^{-1/2}\cdot\mathcal{D}_{h,R}.$$

Thus we have

$$|\mathcal{N}_7| \le \mathcal{C}(h\cdot R)\cdot\max\{1,\tau^{-1/2}\}\cdot\mathcal{A}\cdot R^{-1}\cdot(\mathcal{D}_{h,R}+\mathcal{A}\cdot\mathcal{H}).$$

The preceding estimates of $|\mathcal{N}_1|$ to $|\mathcal{N}_7|$ yield the upper bound of $\mathcal{D}_{h,R}^2$ given in Theorem 4.1. ∎

5. A bound for the pressure error. Theorem 4.1 leaves the problem of how to deal with the terms $\|\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}}$ and $\|\pi_{h,R}-\widetilde{\Pi}_{h,R}(\pi)-m_{h,R}\|_{2,P_{h,R}}$. This is the subject of the present section, where these terms will be majorized by bounds containing a factor of R or $h^{-1/2}$ (Corollary 5.1), which seems to pollute our error estimates. However, in view of the factor $h\cdot R^{-1}$ or $h\cdot R^{-1/2}$ with which these critical terms are multiplied in Theorem 4.1, such bounds are just sufficient to obtain the desired decay rate $h^t+h\cdot\ln(R/S)+R^{-1}$ of the velocity error. Concerning the proof of our estimate of the critical terms, we will reduce this estimate to an evaluation of $c(w,\,\widetilde{\Pi}_{h,R}(\pi)-\pi_{h,R})$, for $w\in Y_{h,R}$. In the ensuing lemma, we introduce a decomposition of $c(w,\,\widetilde{\Pi}_{h,R}(\pi)-\pi_{h,R})$ which will be the starting point of this argument.

LEMMA 5.1. *For $w\in W_{h,R}$, set*

$$\mathcal{M}_1(w) := -a(u-u_{h,R},\,w),\quad \mathcal{M}_2(w) := c(w,\,\widetilde{\Pi}_{h,R}(\pi)-\pi),$$

$$\mathcal{M}_3(w) := \int_{\partial_{h,R}}\sum_{k=1}^{3}\left(\sum_{j=1}^{3}(D_ju_k-\delta_{jk}\cdot\pi)\cdot n_j+(R^{-1}+(\tau/2)\cdot(1-n_1))\cdot u_k\right)\cdot w_k\;do.$$

Then

$$c(w,\,\widetilde{\Pi}_{h,R}(\pi)-\pi_{h,R}) \;=\; \sum_{i=1}^{3}\mathcal{M}_i(w)\quad \textit{for}\;\; w\in W_{h,R}.$$

Proof. The lemma follows from (1.1) and (2.11)–(2.13). ∎

LEMMA 5.2. *Let $q \in M_{h,R}$, and put $\psi := G(q - m(q))$, with the operator G introduced in Theorem A.2, and with $m(q) := |P_{h,R}|^{-1} \cdot \int_{P_{h,R}} q\, dx$. Set $Z := \Pi_{h,R}(\psi)$, $\overline{Z} := (\overline{Z_i})_{1 \le i \le 3}$, with $\overline{Z_i}$ defined as in Lemma* 3.1. *Then*

$$\|\nabla Z\|_{2,P_{h,R}} + \|Z\|_{2,\Omega_S} \;\le\; \mathcal{C} \cdot \|q - m(q)\|_{2,P_{h,R}},$$

$$\|\overline{Z}\|_{2,\partial_{h,R}} \;=\; \|Z\|_{2,\partial_{h,R}} \;\le\; \mathcal{C} \cdot R^{1/2} \cdot \|q - m(q)\|_{2,P_{h,R}},$$

$$\|\nabla \overline{Z}\|_{2,P_{h,R}} \;\le\; \mathcal{C} \cdot h^{-1/2} \cdot \|q - m(q)\|_{2,P_{h,R}}.$$

Proof. The first inequality in the lemma follows from (3.13) and Theorem A.2 (estimate of ∇Z), and from (3.10), (3.7), and Theorem A.2 (estimate of $Z \,|\, \Omega_{2 \cdot S}$). Concerning the second, we observe that $\|\overline{Z}\|_{2,\partial_{h,R}} = \|Z\|_{2,\partial_{h,R}}$ by Lemma 3.1, and

$$\|Z\|_{2,\partial_{h,R}} \le \|Z - \psi\|_{2,\partial_{h,R}} + \|\psi\|_{2,\partial_{h,R}} \;\le\; \mathcal{C} \cdot (h \cdot R)^{1/2} \cdot \|\nabla \psi\|_{2,P_{h,R}} + R^{1/2} \cdot \|\psi\|^{(h,R)}$$

by (3.12). Now the second inequality stated in the lemma follows with Theorem A.2. Let us finally consider the term $\|\nabla \overline{Z}\|_{2,P_{h,R}}$. The usual technique (transformation to a reference tetrahedron) yields the inverse estimate $\|\nabla \overline{Z}\|_{2,K_l} \le \mathcal{C} \cdot (diam\, K_l)^{-1} \cdot \|\overline{Z}\|_{2,K_l}$ for $1 \le l \le k$. Since $K_l \cap \partial B_R \ne \emptyset$ for any $l \in \{1, \ldots, k\}$ with $\overline{Z} \,|\, K_l \ne 0$, by referring to (A3), (3.2), and Lemma 3.1, we thus obtain

$$\|\nabla \overline{Z}\|_{2,P_{h,R}} \le \mathcal{C} \cdot (h \cdot R)^{-1} \cdot \|\overline{Z}\|_{2,A_{h,R}} \le \mathcal{C} \cdot (h \cdot R)^{-1} \cdot \|Z\|_{2,A_{h,R}}.$$

But the relations in (3.3) and (A.3) yield

$$\|Z\|_{2,A_{h,R}} \;\le\; \mathcal{C} \cdot \big(h \cdot R \cdot \|\nabla Z\|_{2,P_{h,R}} \;+\; (h \cdot R)^{1/2} \cdot \|Z\|_{2,\partial_{h,R}} \big).$$

The last inequality in the lemma follows from the two preceding estimates and from the first and second estimates in the lemma. ∎

LEMMA 5.3. *In the situation of Lemma* 5.2, *we have*

$$\|q - m(q)\|^2_{2,P_{h,R}} \;\le\; |c(Z - \overline{Z},\, q)| \;+\; \mathcal{C} \cdot h^{-1/2} \cdot \|\nabla \mathcal{R}(\nabla q)\|_{2,P_{h,R}} \cdot \|q - m(q)\|_{2,P_{h,R}}.$$

Proof. We define a function $W \in \mathcal{B}^3_{h,R}$ by

$$W_i(x) := \left(\int_{K_l} b_{K_l}\, dy \right)^{-1} \cdot \int_{K_l} \big(\psi \;-\; (Z - \overline{Z}) \big)_i \, dy \cdot b_{K_l}(x)$$

for $x \in K_l$, $1 \le l \le k$, $1 \le i \le 3$. By Theorem A.2, the function ψ introduced in the preceding lemma verifies the equations

$$\operatorname{div} \psi = q - m(q), \quad \psi \,|\, \partial\Omega = 0, \quad \psi \cdot n = 0 \;\text{ on } \partial_{h,R}.$$

Therefore after a partial integration of the integral of the product $m(q) \cdot \operatorname{div} \psi$, we obtain

$$\|q - m(q)\|^2_{2,P_{h,R}} \tag{5.1}$$

$$= \int_{P_{h,R}} q \cdot \operatorname{div} \psi \, dx \;=\; -c(Z - \overline{Z},\, q) + \int_{P_{h,R}} q \cdot \operatorname{div} \big(\psi - (Z - \overline{Z}) \big)\, dx.$$

On the other hand, recalling the properties of ψ and the equation $(Z-\overline{Z})\,|\,\partial_{h,R}=0$ (Lemma 3.1), and applying the trick from the proof of LBB stability of the mini element in a similar way as in the proof of (4.12), we obtain

$$\int_{P_{h,R}} q\cdot \operatorname{div}\big(\,\psi\;-\;(Z-\overline{Z})\,\big)\;dx\;=\;\int_{P_{h,R}} q\cdot \operatorname{div} W\;dx\;=\;-A\big(\,\mathcal{R}(\nabla q),\,W\big).$$

The last equation is valid due to (2.10). With (2.6) it follows that

$$\Big|\int_{P_{h,R}} q\cdot \operatorname{div}\big(\,\psi\;-\;(Z-\overline{Z})\big)\;dx\Big|\;\le\;\mathcal{C}\cdot\|\nabla\mathcal{R}(\nabla q)\|_{2,P_{h,R}}\cdot\|\nabla W\|_{2,P_{h,R}}. \tag{5.2}$$

But with (3.21) and the definition of W,

$$\|\nabla W\|_{2,P_{h,R}}\;\le\;\mathcal{C}\cdot\left(\sum_{l=1}^{k}(diam\,K_l)^{-2}\cdot\|\psi\;-\;(Z-\overline{Z})\|_{2,K_l}^2\right)^{1/2}; \tag{5.3}$$

compare the proof of Lemma 4.2 and (4.2). Put $I(h,R):=\{l\in\{1,\dots,k\}\;:\;K_l\cap\partial B_R\neq\emptyset\}$. Then, referring to Lemma 3.1, (A3), and notation from (3.2), we obtain

$$\begin{aligned}\sum_{l=1}^{k}(diam\,K_l)^{-2}\cdot\|\overline{Z}\|_{2,K_l}^2\;&\le\;\mathcal{C}\cdot\sum_{l\in I(h,R)}(diam\,K_l)^{-2}\cdot\|Z\|_{2,K_l}^2\\
&\le\;\mathcal{C}\cdot\left((h\cdot R)^{-2}\cdot\|\psi\|_{2,A_{h,R}}^2\;+\sum_{l\in I(h,R)}(diam\,K_l)^{-2}\cdot\|\psi-Z\|_{2,K_l}^2\right)\\
&\le\;\mathcal{C}\cdot\Big((h\cdot R)^{-2}\cdot\|\psi\|_{2,A_{h,R}}^2+\|\nabla\psi\|_{2,P_{h,R}}^2\Big)\\
&\le\;\mathcal{C}\cdot\Big((h\cdot R)^{-1}\cdot\|\psi\|_{2,A_{h,R}}+\|q-m(q)\|_{2,P_{h,R}}\Big)^2,\end{aligned} \tag{5.4}$$

where the last two inequalities are a consequence of (3.10), (3.8), and Theorem A.2. Moreover, by (3.3) and (A.3) we find

$$\begin{aligned}\|\psi\|_{2,A_{h,R}}&\le\|\psi\|_{2,\,P_{h,R}\setminus B_{R\cdot(1-2\cdot h/S)}}\\
&\le\mathcal{C}\cdot\big(\,h\cdot R\cdot\|\nabla\psi\|_{2,P_{h,R}}\;+\;(h\cdot R)^{1/2}\cdot\|\psi\|_{2,\partial_{h,R}}\,\big).\end{aligned}$$

Hence $(h\cdot R)^{-1}\cdot\|\psi\|_{2,A_{h,R}}\;\le\;\mathcal{C}\cdot h^{-1/2}\cdot\|q-m(q)\|_{2,P_{h,R}}$ by Theorem A.2. From this estimate, (5.3), (3.10), and (3.6), we may conclude that

$$\|\nabla W\|_{2,P_{h,R}}\le\mathcal{C}\cdot h^{-1/2}\cdot\|q-m(q)\|_{2,P_{h,R}}.$$

This inequality, (5.1), and (5.2) imply the lemma. □

THEOREM 5.1. *In the situation of Lemma* 5.2, *and with the abbreviation* $\mathcal{D}_{h,R}$ *defined in Theorem* 4.1, *we have*

$$\begin{aligned}&|a(u-u_{h,R},\,Z)|\\
&\le\;\mathcal{C}(\tau,h\cdot R)\cdot\big(\,\mathcal{D}_{h,R}+\mathcal{A}\cdot\big(\,h^t+h\cdot\ln(R/S)\,\big)\,\big)\cdot\|q-m(q)\|_{2,P_{h,R}}\cdot R^{1/2};\end{aligned} \tag{5.5}$$

$$|a(u-u_{h,R},\,\overline{Z})|\;\le\;\mathcal{C}(\tau,h\cdot R)\cdot\mathcal{D}_{h,R}\cdot\|q-m(q)\|_{2,P_{h,R}}\cdot h^{-1/2}; \tag{5.6}$$

$$|a(u-u_{h,R},\,w)|\;\le\;\mathcal{C}(\tau)\cdot\mathcal{D}_{h,R}\cdot\|w\|^{(h,R)}\cdot R\quad\textit{for }\;w\in W_{h,R}. \tag{5.7}$$

Proof. For brevity, we put $e_{h,R} := u - u_{h,R}$. We have

$$|a(e_{h,R},\, Z)| \;\le\; \sum_{i=1}^{3} |I_i| \quad \text{with} \quad I_1 := \int_{P_{h,R}} \nabla e_{h,R} \cdot \nabla Z \; dx, \tag{5.8}$$

$$I_2 := \tau \cdot \int_{P_{h,R}} \partial_1 e_{h,R} \cdot Z \; dx, \quad I_3 := \int_{\partial_{h,R}} \big(R^{-1} + (\tau/2) \cdot (1 - n_1) \big) \cdot \big(e_{h,R} \cdot Z \big) \; do.$$

The main difficulty consists in estimating the term I_2. In order to obtain such an estimate, we begin by observing that the function ψ from Lemma 5.2 fulfills the relation $\psi \,|\, P_{h,R} \backslash B_S = \nabla v \,|\, P_{h,R} \backslash B_S$ for some function $v \in H^2(P_{h,R} \cup \Omega^c)$; see Theorem A.2. This means that $\partial_1 \psi(x) = \nabla \psi_1(x)$ for $x \in P_{h,R} \backslash \overline{B_S}$. As a consequence, after some partial integrations, we obtain the decomposition $I_2 = \tau \cdot \sum_{i=1}^{7} J_i$, where

$$J_1 := \int_{\Omega_S} \partial_1 e_{h,R} \cdot Z \; dx, \quad J_2 := \int_{\partial B_S} e_{h,R}(x) \cdot S^{-1} \cdot (-x_1 \cdot \psi(x) + \psi_1(x) \cdot x) \; do_x,$$

$$J_3 := \int_{P_{h,R} \backslash B_S} \partial_1 e_{h,R} \cdot (Z - \psi) \; dx, \quad J_4 := \int_{\partial_{h,R}} e_{h,R} \cdot (n_1 \cdot \psi - \psi_1 \cdot n) \; do,$$

$$J_5 := -\int_{\Omega_S} \operatorname{div} e_{h,R} \cdot \psi_1 \; dx, \quad J_6 := \int_{P_{h,R}} \operatorname{div} e_{h,R} \cdot (\psi - Z)_1 \; dx,$$

$$J_7 := \int_{P_{h,R}} \operatorname{div} u_{h,R} \cdot Z_1 \; dx.$$

By (3.13) and Theorem A.2,

$$\begin{aligned} |J_3| + |J_6| &\le \mathcal{C} \cdot \|\nabla e_{h,R}\|_{2,P_{h,R}} \cdot \|Z - \psi\|_{2,P_{h,R}} \\ &\le \mathcal{C}(h \cdot R) \cdot \|e_{h,R}\|^{(h,R)} \cdot \|q - m(q)\|_{2,P_{h,R}}. \end{aligned} \tag{5.9}$$

A standard trace theorem and Theorems A.1 and A.2 yield

$$|J_2| \;\le\; \mathcal{C} \cdot \|e_{h,R}\|_{1,2,\Omega_S} \cdot \|\psi\|_{1,2,\Omega_S} \;\le\; \mathcal{C} \cdot \|e_{h,R}\|^{(h,R)} \cdot \|q - m(q)\|_{2,P_{h,R}}. \tag{5.10}$$

By Theorem A.2 and Lemma 5.2,

$$\begin{aligned} |J_1| + |J_5| &\le \mathcal{C} \cdot \|\nabla e_{h,R}\|_{2,P_{h,R}} \cdot (\|Z\|_{2,\Omega_S} + \|\psi\|_{2,\Omega_S}) \\ &\le \mathcal{C} \cdot \|e_{h,R}\|^{(h,R)} \cdot \|q - m(q)\|_{2,P_{h,R}}. \end{aligned} \tag{5.11}$$

By Theorem A.2, we further get

$$|J_4| \;\le\; \mathcal{C} \cdot \|e_{h,R}\|_{2,\partial_{h,R}} \cdot \|\psi\|_{2,\partial_{h,R}} \;\le\; \mathcal{C} \cdot R^{1/2} \cdot \tau^{-1/2} \cdot \mathcal{D}_{h,R} \cdot \|q - m(q)\|_{2,P_{h,R}}. \tag{5.12}$$

This leaves us to deal with the term J_7. To this end, we observe that $|J_7|$ is bounded by $\mathcal{C} \cdot \|\nabla Z\|_{2,P_{h,R}} \cdot \|\mathcal{R}(\nabla \pi_{h,R})\|_{2,P_{h,R}}$, as follows from (2.12) and (2.10). We may conclude with Lemma 5.2 and (3.23) that

$$|J_7| \;\le\; \mathcal{C}(h \cdot R) \cdot \|\nabla \mathcal{R}(\nabla \pi_{h,R})\|_{2,P_{h,R}} \cdot \|q - m(q)\|_{2,P_{h,R}}.$$

This estimate, the inequality $\|e_{h,R}\|^{(h,R)} \le \mathcal{D}_{h,R}$, and (5.9)–(5.12) imply

$$|I_2| \le \mathcal{C}(\tau, h\cdot R)\cdot R^{1/2}\cdot (\mathcal{D}_{h,R} + \|\nabla\mathcal{R}(\nabla\pi_{h,R})\|_{2,P_{h,R}})\cdot \|q-m(q)\|_{2,P_{h,R}}. \tag{5.13}$$

As for I_3, we have

$$|I_3| \;\le\; \mathcal{C}(\tau)\cdot (\|e_{h,R}\|^{(h,R)} + \tau^{1/2}\cdot \|e_{h,R}\|_{2,\partial_{h,R}})\cdot \|Z\|_{2,\partial_{h,R}}.$$

Hence

$$\begin{aligned} |I_3| &\le \mathcal{C}(\tau)\cdot \mathcal{D}_{h,R}\cdot \|Z\|_{2,\partial_{h,R}} \\ &\le \mathcal{C}(\tau, h\cdot R)\cdot R^{1/2}\cdot \mathcal{D}_{h,R}\cdot \|q-m(q)\|_{2,P_{h,R}}, \end{aligned} \tag{5.14}$$

where we used Lemma 5.2. Observe in addition that by (2.6) and (2.10),

$$\|\nabla\mathcal{R}(\nabla\pi)\|^2_{2,P_{h,R}} \le \alpha^{-1}\cdot A\big(\mathcal{R}(\nabla\pi),\, \mathcal{R}(\nabla\pi)\big) = -\alpha^{-1}\cdot \int_{P_{h,R}} \pi\cdot \operatorname{div}\mathcal{R}(\nabla\pi)\,dx.$$

Hence with Lemma 4.4 and the definition of $\mathcal{D}_{h,R}$,

$$\|\nabla\mathcal{R}(\nabla\pi_{h,R})\|_{2,P_{h,R}} \;\le\; \mathcal{D}_{h,R} + \mathcal{C}\cdot\mathcal{A}\cdot\big(h^t + h\cdot\ln(R/S)\big). \tag{5.15}$$

We finally note that $|I_1|$ is bounded by $\mathcal{C}\cdot\mathcal{D}_{h,R}\cdot\|q-m(q)\|_{2,P_{h,R}}$, where we applied Lemma 5.2 again. This inequality and estimates (5.13)–(5.15) imply (5.5). Turning to the proof of (5.6), we have to estimate the terms I_1, I_2, and I_3 with Z replaced by $\overline{Z}$. Denoting these terms by $\widetilde{I}_i$, for $i\in\{1,\,2,\,3\}$, we use Lemma 3.1, (3.3), (A.3), and Lemma 5.2 to obtain

$$\begin{aligned} |\widetilde{I}_2| &\le \mathcal{C}(\tau)\cdot\|\nabla e_{h,R}\|_{2,P_{h,R}}\cdot\|Z\|_{2,A_{h,R}} \\ &\le \mathcal{C}(\tau, h\cdot R)\cdot R^{1/2}\cdot\|\nabla e_{h,R}\|_{2,P_{h,R}}\cdot\|q-m(q)\|_{2,P_{h,R}}. \end{aligned}$$

Taking account of this inequality, and estimating the terms $\widetilde{I}_1$ and $\widetilde{I}_3$ in a straightforward way (Hölder's inequality, Lemma 5.2; note that $R^{-1/2}\cdot\|e_{h,R}\|_{2,\partial_{h,R}} \le \mathcal{D}_{h,R}$ and $\tau^{1/2}\cdot\|e_{h,R}\|_{2,\partial_{h,R}} \le \mathcal{D}_{h,R}$), we arrive at (5.6). Inequality (5.7) may be established by combining Hölder's inequality, the preceding estimates of $\|e_{h,R}\|_{2,\partial_{h,R}}$, and the inequalities

$$\|w\|_{2,P_{h,R}} \le \mathcal{C}\cdot R\cdot\|w\|_{6,P_{h,R}} \le \mathcal{C}\cdot R\cdot\|w\|^{(h,R)} \;\text{ for }\; w\in W_{h,R}$$

(Theorem A.1). □

Now we are in a position to estimate the terms $\|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi) - m_{h,R}\|_{2,P_{h,R}}$ and $\|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}}$.

Corollary 5.1. *Let $\mathcal{D}_{h,R}$ and $m_{h,R}$ be defined as in Theorem* 4.1. *Then*

$$\begin{aligned} &\|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi) - m_{h,R}\|_{2,P_{h,R}} \\ &\quad\le\; \mathcal{C}(\tau, h\cdot R)\cdot\big(\mathcal{D}_{h,R} + \mathcal{A}\cdot(h^t + h\cdot\ln(R/S) + R^{-1})\big)\cdot h^{-1/2}; \end{aligned} \tag{5.16}$$

$$\begin{aligned} &\|\pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)\|_{2,P_{h,R}} \\ &\quad\le\; \mathcal{C}(\tau, h\cdot R)\cdot\big(\mathcal{D}_{h,R} + \mathcal{A}\cdot(h^t + h\cdot\ln(R/S) + R^{-1})\big)\cdot R. \end{aligned} \tag{5.17}$$

Proof. We use the notation introduced in Lemma 5.2, but suppose that $q = \pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)$. Hence $m(q) = m_{h,R}$. Then, by (4.5) and Lemma 5.3,

$$\|q - m(q)\|^2_{2,P_{h,R}} \le |c(Z - \overline{Z},\, q)| + \mathcal{C}\cdot h^{-1/2}\cdot \Big(\mathcal{D}_{h,R} + \mathcal{A}\cdot\big(h^t + h\cdot\ln(R/S)\big)\Big)\cdot \|q - m(q)\|_{2,P_{h,R}}. \tag{5.18}$$

But by Lemma 5.1, $|c(Z - \overline{Z},\, q)| \le \sum_{i=1}^{3}\big(|\mathcal{M}_i(Z)| + |\mathcal{M}_i(\overline{Z})|\big)$, with the terms $\mathcal{M}_i(Z)$ and $\mathcal{M}_i(\overline{Z})$ ($i \in \{1,\, 2,\, 3\}$) defined as in that lemma. As a consequence of (2.4), (2.5), and Lemma 5.2, we get

$$\begin{aligned}|\mathcal{M}_3(Z)| + |\mathcal{M}_3(\overline{Z})| &\le \mathcal{C}\cdot\mathcal{A}\cdot(h + R^{-1})\cdot R^{1/2}\cdot\|q - m(q)\|_{2,P_{h,R}}\\ &\le \mathcal{C}(h\cdot R)\cdot\mathcal{A}\cdot(h + R^{-1})\cdot\|q - m(q)\|_{2,P_{h,R}}\cdot h^{-1/2}.\end{aligned} \tag{5.19}$$

By (3.17) and Lemma 5.2, we have

$$|\mathcal{M}_2(Z)| + |\mathcal{M}_2(\overline{Z})| \le \mathcal{C}\cdot\mathcal{A}\cdot\big(h^t + h\cdot\ln(R/S)\big)\cdot\|q - m(q)\|_{2,P_{h,R}}\cdot h^{-1/2}. \tag{5.20}$$

The terms $\mathcal{M}_1(Z)$ and $\mathcal{M}_1(\overline{Z})$ were estimated in Theorem 5.1. By combining this observation, inequalities (5.19) and (5.20), the splitting of $|c(Z - \overline{Z},\, q)|$ provided by Lemma 5.1 and indicated above, and estimate (5.18), we obtain (5.16).

The starting point of our proof of (5.17) is the estimate in Theorem 3.1, applied with $\varrho = q$. Due to this inequality, and in view of Lemma 5.1, the problem reduces to estimating the terms $|\mathcal{M}_i(w)|$ for $w \in W_{h,R},\ i \in \{1,\, 2,\, 3\}$, and the term $|c(W, q)|$ for $W \in \mathcal{B}^3_{h,R}$. First take $w \in W_{h,R}$. The relations in (2.4) and (2.5) and the estimate $\|w\|_{2,\partial_{h,R}} \le R^{1/2}\cdot\|w\|^{(h,R)}$ yield

$$|\mathcal{M}_3(w)| \le \mathcal{C}\cdot\mathcal{A}\cdot(h + R^{-1})\cdot\|w\|^{(h,R)}\cdot R^{1/2}.$$

Moreover, inequality (3.17) implies

$$|\mathcal{M}_2(w)| \le \mathcal{C}\cdot\|\nabla w\|_2\cdot\mathcal{A}\cdot\big(h^t + h\cdot\ln(R/S)\big).$$

The term $\mathcal{M}_1(w)$ was estimated in (5.7). Next take $W \in \mathcal{B}^3_{h,R}$. With (2.10), (2.6), and (4.5), we get

$$\begin{aligned}|c(W, q)| = |A\big(\mathcal{R}(\nabla q),\, W\big)| &\le \mathcal{C}\cdot\|\nabla\mathcal{R}(\nabla q)\|_{2,P_{h,R}}\cdot\|\nabla W\|_{2,P_{h,R}}\\ &\le \mathcal{C}\cdot\Big(\mathcal{D}_{h,R} + \mathcal{A}\cdot\big(h^t + h\cdot\ln(R/S)\big)\Big)\cdot\|\nabla W\|_{2,P_{h,R}}.\end{aligned}$$

We note in addition that $\|w\|^{(h,R)} + \|\nabla W\|_{2,P_{h,R}} \le \mathcal{C}\cdot\|w + W\|^{(h,R)}$ for $w \in W_{h,R},\ W \in \mathcal{B}^3_{h,R}$ (see [17, p. 291]). These results taken together yield (5.17). ∎

6. Proof of Theorem 2.1. Let us again use the notation $\mathcal{D}_{h,R}$ introduced in Theorem 4.1. Put $\mathcal{H} := h^t + h\cdot\ln(R/S)$. Theorem 4.1 and Corollary 5.1 imply

$$\mathcal{D}^2_{h,R} \le \mathcal{C}(\tau, h\cdot R)\cdot\max\{1,\, \tau^{-1/2}\}\cdot\mathcal{A}\cdot(\mathcal{H} + R^{-1})\cdot\big(\mathcal{D}_{h,R} + \mathcal{A}\cdot(\mathcal{H} + R^{-1})\big).$$

Hence

$$\mathcal{D}_{h,R} \le \mathcal{C}(\tau, h\cdot R)\cdot\max\{1,\, \tau^{-1/2}\}\cdot\mathcal{A}\cdot(\mathcal{H} + R^{-1}). \tag{6.1}$$

Thus, in order to complete the proof of (2.14), we still have to evaluate $\|\pi-\pi_{h,R}\|_{2,\Omega_{2\cdot S}}$. Denote $q := \pi_{h,R} - \widetilde{\Pi}_{h,R}(\pi)$. By [6, Theorem 3.5], there is a function $w \in H^1(P_{h,R})^3$ with $w \,|\, \partial\Omega = 0,\ \operatorname{div} w(x) = q(x)$ for $x \in \Omega_{2\cdot S},\ \operatorname{div} w(x) = 0$ for $x \in P_{h,R}\backslash B_{2\cdot S}$, and such that $\Pi_{h,R}(w)$ verifies the estimates stated in [6, Lemma 3.4], with g replaced by $q \,|\, \Omega_{2\cdot S}$. For brevity, put $v := \Pi_{h,R}(w)$. By [6, Lemma 4.3], we have

$$\|q\|^2_{2,\Omega_{2\cdot S}} \ \le\ |c(v,q)| \ +\ \mathcal{C}\cdot\big(\,\|\nabla\mathcal{R}(\nabla q)\|_{2,P_{h,R}} \ +\ \|q\|_{2,P_{h,R}}\cdot h^{1/2}\cdot R^{-3/2}\,\big)^2. \tag{6.2}$$

On the other hand, we know by (5.17) that

$$\|q\|_{2,P_{h,R}}\cdot h^{1/2}\cdot R^{-3/2} \le\ \mathcal{C}(\tau, h\cdot R)\cdot\big(\,\mathcal{D}_{h,R} + \mathcal{A}\cdot(\mathcal{H}+R^{-1})\,\big). \tag{6.3}$$

Recall the terms $\mathcal{M}_i(v)$ $(1 \le i \le 3)$ defined in Lemma 5.1. The reference [6, Lemma 3.4] and the definitions in (2.4), (2.5) imply

$$|\mathcal{M}_3(v)| \ \le\ \mathcal{C}\cdot\mathcal{A}\cdot(h+R^{-1})\cdot\|v\|_{2,\partial_{h,R}} \ \le\ \mathcal{C}\cdot\mathcal{A}\cdot(h+R^{-1})\cdot\|q\|_{2,\Omega_{2\cdot S}}. \tag{6.4}$$

Using (3.17) and [6, Lemma 3.4], we get

$$|\mathcal{M}_2(v)| \ \le\ \mathcal{C}\cdot\|\nabla v\|_{2,P_{h,R}}\cdot\mathcal{A}\cdot\mathcal{H} \ \le\ \mathcal{C}\cdot\mathcal{A}\cdot\mathcal{H}\cdot\|q\|_{2,\Omega_{2\cdot S}}. \tag{6.5}$$

Again referring to [6, Lemma 3.4], we obtain

$$\begin{aligned}|\mathcal{M}_1(v)| \ &\le\ \mathcal{C}(\tau)\cdot\big(\,\|u-u_{h,R}\|^{(h,R)} + \tau^{1/2}\cdot\|u-u_{h,R}\|_{2,\partial_{h,R}}\,\big)\\ &\quad\cdot\big(\,\|\nabla v\|_{2,P_{h,R}} + \|v\|_{2,\partial_{h,R}} + \|v\|_{2,P_{h,R}}\,\big) \ \le\ \mathcal{C}(\tau)\cdot\mathcal{D}_{h,R}\cdot\|q\|_{2,\Omega_{2\cdot S}}.\end{aligned} \tag{6.6}$$

As a consequence of (6.4)–(6.6) and Lemma 5.1, we find

$$|c(v,q)| \ \le\ \mathcal{C}(\tau)\cdot\big(\,\mathcal{D}_{h,R} + \mathcal{A}\cdot(\mathcal{H}+R^{-1})\,\big)\cdot\|q\|_{2,\Omega_{2\cdot S}}.$$

Now we combine the preceding inequality with (6.2), (6.3), and (4.5) to obtain, by a simple shoestring argument,

$$\|q\|_{2,\Omega_{2\cdot S}} \ \le\ \mathcal{C}(\tau, h\cdot R)\cdot\big(\,\mathcal{D}_{h,R} + \mathcal{A}\cdot(\mathcal{H}+R^{-1})\,\big). \tag{6.7}$$

From (6.7) and (3.17) we may conclude that

$$\|\pi-\pi_{h,R}\|_{2,\Omega_{2\cdot S}} \ \le\ \mathcal{C}(\tau, h\cdot R)\cdot(\mathcal{D}_{h,R} + \mathcal{A}\cdot(\mathcal{H}+R^{-1})\,\big),$$

so (2.14) follows with (6.1).

Appendix. In this appendix, we prove some results from analysis we applied above. We begin by considering two inequalities already used in [6].

THEOREM A.1. *The following inequalities are valid:*

$$\|v\|_{2,\Omega_{2\cdot S}} + \|v\|_{6,P_{h,R}} \le \mathcal{C}\cdot\|v\|^{(h,R)} \quad \textit{for}\ \ v \in W_{h,R},$$

$$\|v\|_{2,\Omega_{2\cdot S}} \le \mathcal{C}\cdot\big(\,\|v\|_{2,P_{h,R}\cup\Omega^c} + R^{-1/2}\cdot\|v\|_{2,\partial_{h,R}}\,\big) \quad \textit{for}\ \ v \in H^1(P_{h,R}\cup\Omega^c)^3.$$

Proof. For the estimate of $\|v\|_{2,\Omega_{2\cdot S}}$ in both cases $v \in W_{h,R}$ and $v \in H^1(P_{h,R}\cup\Omega^c)^3$, we refer to [4, Theorem 3.4] and [6, Theorem 3.3]. Concerning the term $\|v\|_{6,P_{h,R}}$, it may be evaluated by applying [6, Theorem 3.4]. ∎

In the first three inequalities of the ensuing lemma, functions with domain $P_{h,R}$ are estimated near or on $\partial_{h,R}$. The lemma indicates how the constants in these inequalities depend on h and R.

LEMMA A.1. *Denote* $\widetilde{R} := R \cdot (1 - h^2/S^2)^{1/2}$. *Then, for* $v \in H^1(P_{h,R})^3$,

$$\|v\|_{2,\,P_{h,R}\setminus B_{\widetilde{R}}} \;\le\; \mathcal{C} \cdot \big(h^2 \cdot R \cdot \|\nabla v\|_{2,P_{h,R}} + h \cdot R^{1/2} \cdot \|v\|_{2,\,\partial B_{\widetilde{R}}} \big), \tag{A.1}$$

$$\|v\|_{2,\partial_{h,R}} \;\le\; \mathcal{C} \cdot \big(h \cdot \widetilde{R}^{1/2} \cdot \|\nabla v\|_{2,P_{h,R}} + \|v\|_{2,\,\partial B_{\widetilde{R}}} \big), \tag{A.2}$$

$$\|v\|_{2,\,P_{h,R}\setminus B_{R\cdot(1-4\cdot h/S)}} \;\le\; \mathcal{C} \cdot \big(h \cdot R \cdot \|\nabla v\|_{2,P_{h,R}} + (h \cdot R)^{1/2} \cdot \|v\|_{2,\partial_{h,R}} \big). \tag{A.3}$$

(*Note that* $\Omega_{\widetilde{R}} \subset P_{h,R}$ *by* (3.5).) *Moreover,*

$$\|v\|_{2,\,\widetilde{R}^{-1}\cdot(P_{h,R}\setminus B_{\widetilde{R}})} \;\le\; \mathcal{C} \cdot (\|\nabla v\|_{2,\,\widetilde{R}^{-1}\cdot P_{h,R}} + \|v\|_{2,\,\partial B_1}) \tag{A.4}$$

$$\text{for } v \in H^1(\widetilde{R}^{-1} \cdot P_{h,R}),$$

$$\|v\|_{2,\,\partial B_{\widetilde{R}}} \;\le\; \mathcal{C} \cdot \big(\widetilde{R}^{1/2} \cdot \|\nabla v\|_{2,B_{\widetilde{R}}} + \widetilde{R}^{-1/2} \cdot \|v\|_{2,B_{\widetilde{R}}} \big) \quad \text{for } v \in H^1(B_{\widetilde{R}})^3. \tag{A.5}$$

Proof. Inequalities (A.1)–(A.3) may be proved in the same way as [5, Lemma 7.1] or [4, Theorems 3.2 and 3.3]. We leave the details to the reader. The relation (A.4) may be obtained from (A.1) by a scaling argument, whereas estimate (A.5) may be reduced by a scaling argument to a standard trace estimate for functions from $H^1(B_1)^3$. □

Finally, we consider the divergence equation on $P_{h,R}$, which we solve by using the L^2-theory for the Laplace operator with Neumann boundary conditions on convex domains.

THEOREM A.2. *Let* $g \in L^2(P_{h,R})$ *with* $\int_{P_{h,R}} g\, dx = 0$. *Then there is a function* $G(g) \in H^1(P_{h,R})^3$ *with*

$$\operatorname{div} G(g) = g, \quad G(g)\,|\,\partial\Omega = 0, \quad G(g)\cdot n = 0 \;\text{ on }\; \partial_{h,R},$$

$$G(g)\,|\,P_{h,R}\setminus B_S = \nabla v(g)\,|\,P_{h,R}\setminus B_S \quad \text{for some function } v(g) \in H^2(P_{h,R}\cup\Omega^c),$$

$$\|G(g)\|^{(h,R)} + \|G(g)\|_{2,\Omega_{2\cdot S}} \;\le\; \mathcal{C}\cdot\|g\|_{2,P_{h,R}}.$$

Proof. Denote $\widetilde{R} := R \cdot (1 - h^2/S^2)^{1/2}$, $U := \widetilde{R}^{-1} \cdot \big(P_{h,R} \cup \Omega^c)$, and let $n^{(U)}$ denote the outward unit normal to U. Note that $B_1 \subset U$ (see (3.5)). Put

$$\mathcal{P} := \Big\{ v \in H^2(U) \;:\; \partial u/\partial n^{(U)} = 0 \;\text{ on }\; \partial U, \;\; \int_{B_1} v\, dx = 0 \Big\}.$$

Since U is convex (see (A5)), by [13, Theorem 3.1.2.3, Lemma 3.2.1.1], we have

$$\|v\|_{2,2,U} \;\le\; \mathcal{C} \cdot (\|\Delta v\|_{2,U} + \|v\|_{2,U}) \quad \text{for } v \in H^2(U) \text{ with } \partial v/\partial n^{(U)} = 0. \tag{A.6}$$

This inequality is valid with the same constant for any convex domain in $\mathbb{R}^3$. In particular, it is valid with the same constant for all domains $U = \widetilde{R}^{-1} \cdot (P_{h,R} \cup \Omega^c)$ with $h \in (0, h_0)$, $R \in (R_0, \infty)$. Let us show that

$$\|v\|_{2,2,U} \;\le\; \mathcal{C} \cdot \|\Delta v\|_{2,U} \quad \text{for any } \; v \in \mathcal{P}. \tag{A.7}$$

By first using (A.4) and then estimating $\|v\|_{2,\partial B_1}$ by a standard trace theorem, we get $\|v\|_{2,\,U\setminus B_1} \le \mathcal{C}\cdot(\|\nabla v\|_{2,U}+\|v\|_{2,B_1})$ for $v\in H^1(U)$. By Poincaré's inequality for functions with mean value zero on B_1 (see [9, Theorem II.4.3], for example), we may deduce from the previous estimate that $\|v\|_{2,U}\le \mathcal{C}\cdot\|\nabla v\|_{2,U}$ for $v\in\mathcal{P}$. By partial integration and due to the relation $\partial v/\partial n^{(U)}=0$ for $v\in\mathcal{P}$, the inequality

$$\|\nabla v\|_{2,U}^2 \le \mathcal{C}\cdot\|\Delta v\|_{2,U}\cdot\|v\|_{2,U} \le (\mathcal{C}/\epsilon)\cdot\|\Delta v\|_{2,U}^2+\epsilon\cdot\|v\|_{2,U}^2$$

holds for $\epsilon\in(0,\infty)$, $v\in\mathcal{P}$. By choosing ϵ sufficiently small, we may deduce (A.7) from the two preceding estimates and from (A.6).

Now we return to the function g given in the theorem. Let $\widetilde{g}$ be the zero extension of g to $P_{h,R}\cup\Omega^c$. Recall that $P_{h,R}\cup\Omega^c$ is convex (see (A5)) and $\int_{P_{h,R}} g\,dx=0$. Thus, according to [10, Theorem I.1.9], [13, Theorem 3.2.1.3], there is a function $v(g)\in H^2\big(P_{h,R}\cup\Omega^c\big)$ with $\Delta v(g)=\widetilde{g}$, $\partial v(g)/\partial n=0$ on $\partial_{h,R}$. Without loss of generality, we may suppose that $\int_{B_{\widetilde{R}}} v(g)\,dx=0$. Put $\widetilde{v}(x):=v(g)(\widetilde{R}\cdot x)$ for $x\in U$. Then $\widetilde{v}\in\mathcal{P}$, and hence inequality (A.7) holds with v replaced by $\widetilde{v}$. It follows by a scaling argument that

$$|v(g)|_{2,2,\,P_{h,R}\cup\Omega^c} + R^{-1}\cdot\|\nabla v(g)\|_{2,\,P_{h,R}\cup\Omega^c} \le \mathcal{C}\cdot\|g\|_{2,P_{h,R}}. \tag{A.8}$$

On the other hand, using (A.2) and (A.5) with v replaced by $\nabla v(g)$, and using (A.8) as well, we see that

$$R^{-1/2}\cdot\|\nabla v(g)\|_{2,\partial_{h,R}} \le \mathcal{C}\cdot\|g\|_{2,P_{h,R}}. \tag{A.9}$$

Since $\widetilde{g}\,|\,\Omega=0$, we get $\int_{\partial\Omega}\nabla v(g)\cdot n^{(\Omega)}\,do=0$, where we write $n^{(\Omega)}$ for the exterior unit normal to Ω. Thus, according to [9, Exercise III.3.4], there is a function $\mathcal{F}\in H^1(\Omega_S)^3$ with $\operatorname{div}\mathcal{F}=0$ and

$$\mathcal{F}\,|\,\partial\Omega=-\nabla v(g)\,|\,\partial\Omega,\quad \mathcal{F}\,|\,\partial B_S=0,\quad \|\mathcal{F}\|_{1,2,\Omega_S}\le\mathcal{C}\cdot\|\nabla v(g)\,|\,\partial\Omega\|_{1/2,\,2,\partial\Omega},$$

where $\|\ \|_{1/2,\,2,\partial\Omega}$ denotes an arbitrary but fixed norm of the usual fractional order Sobolev space $H^{1/2}(\partial\Omega)^3$ on the Lipschitz boundary $\partial\Omega$. The last inequality and a standard trace theorem imply

$$\|\mathcal{F}\|_{1,2,\Omega_S}\le\mathcal{C}\cdot\|\nabla v(g)\|_{1,2,\Omega_S}. \tag{A.10}$$

Let $\widetilde{\mathcal{F}}$ denote the zero extension of $\mathcal{F}$ to $P_{h,R}$, and put $G(g):=\widetilde{\mathcal{F}}+\nabla v(g)\,|\,P_{h,R}$. Then the functions $G(g)$ and $v(g)$ verify the properties stated in Theorem A.2. In particular, the inequality at the end of that theorem follows from (A.8)–(A.10) and from the first inequality in Theorem A.1 with v replaced by $\nabla v(g)$. □

REFERENCES

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

[2] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Springer, New York, 2002.

[3] P. Deuring, *A stable mixed finite element method on truncated exterior domains*, RAIRO Modél. Math. Anal. Numér., 32 (1998), pp. 283–305.

[4] P. Deuring, *Approximating exterior flows by flows on truncated exterior domains: Piecewise polygonal artificial boundaries*, in Proceedings of the 4th European Conference on Elliptic and Parabolic Problems (Rolduc and Gaeta, 2001), J. Bemelmans et al., eds., World Scientific, River Edge, NJ, 2002, pp. 364–376.

[5] P. Deuring, *Exterior stationary Navier-Stokes flows in* 3*D with non-zero velocity at infinity: Asymptotic behaviour of the second derivatives of the velocity*, Comm. Partial Differential Equations, 30 (2005), pp. 987–1020.
[6] P. Deuring, *Stability of a finite element method for* 3*D exterior stationary Navier–Stokes flows*, Appl. Math., 52 (2007), pp. 59–94.
[7] P. Deuring and S. Kračmar, *Exterior stationary Navier-Stokes flows in* 3*D with non-zero velocity at infinity: Approximation by flows in bounded domains*, Math. Nachr., 269/270 (2004), pp. 86–115.
[8] E. B. Fabes, C. E. Kenig, and G. C. Verchota, *The Dirichlet problem for the Stokes system on Lipschitz domains*, Duke Math. J., 57 (1988), pp. 769–792.
[9] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol.* I. *Linearized Steady Problems*, rev. ed., Springer, New York, 1998.
[10] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
[11] C. I. Goldstein, *The finite element method with nonuniform mesh sizes for unbounded domains*, Math. Comp., 36 (1981), pp. 387–404.
[12] C. I. Goldstein, *Multigrid methods for elliptic problems in unbounded domains*, SIAM J. Numer. Anal., 30 (1993), pp. 159–183.
[13] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[14] G. H. Guirguis and M. D. Gunzburger, *On the approximation of the exterior Stokes problem in three dimensions*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 445–464.
[15] M. D. Gunzburger, *Finite Element Methods for Viscous Incompressible Flows*, Academic Press, Boston, 1989.
[16] J. Nečas, *Les Méthodes Directes en Théorie des Équations Elliptiques*, Masson, Paris, 1967.
[17] T. C. Rebollo, *A term by term stabilization algorithm for finite element solution of incompressible flow problems*, Numer. Math., 79 (1998), pp. 283–319.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1544–1569

hp-VERSION DISCONTINUOUS GALERKIN FINITE ELEMENT METHOD FOR SEMILINEAR PARABOLIC PROBLEMS*

ANDRIS LASIS† AND ENDRE SÜLI‡

Abstract. We consider the *hp*-version discontinuous Galerkin finite element method (*hp*-DGFEM) with interior penalty for semilinear parabolic equations with locally Lipschitz continuous nonlinearity, subject to mixed nonhomogeneous Dirichlet–nonhomogeneous Neumann boundary conditions. Our main concern is the error analysis of the (spatially) semidiscrete *hp*-DGFEM on shape-regular spatial meshes. We derive error bounds under various hypotheses on the regularity of the solution, for both the symmetric and nonsymmetric versions of DGFEM.

Key words. *hp*-finite element methods, discontinuous Galerkin methods, semilinear parabolic PDEs

AMS subject classifications. 65N12, 65N15, 65N30

DOI. 10.1137/050642125

1. Introduction.

Let Ω be a bounded open polyhedral domain in $\mathbb{R}^d$, $d \geq 2$, with a Lipschitz-continuous boundary. We consider the semilinear parabolic partial differential equation (PDE)

$$u' - \Delta u = f(x,t,u) \quad \text{in} \quad \Omega \times (0,T], \tag{1.1}$$

where $u' := \partial u/\partial t$ and $T > 0$. It will be assumed throughout that f is a real-valued function defined on $\Omega \times (0,T] \times \mathbb{R}$ which satisfies the following assumption:

(**A**) $f(\cdot\,,\cdot\,,v) : (x,t) \in \Omega \times (0,T] \mapsto f(x,t,v) \in \mathbb{R}$ is measurable in $(x,t) \in \Omega \times (0,T]$ for all $v \in \mathbb{R}$, with $f(x,t,0) = 0$ for all $(x,t) \in \Omega \times (0,T]$, and the mapping $f(x,t,\cdot\,) : v \in \mathbb{R} \mapsto f(x,t,v) \in \mathbb{R}$ is locally Lipschitz continuous for a.e. $(x,t) \in \Omega \times (0,T]$, in the sense that there exist real numbers $G_f > 0$ and $\gamma \geq 0$ such that

$$|f(x,t,w) - f(x,t,v)| \leq G_f(1+|w|+|v|)^\gamma\,|w-v| \quad \begin{cases} \forall w,\, v \in \mathbb{R}, \\ \text{a.e. } (x,t) \in \Omega \times (0,T]. \end{cases} \tag{1.2}$$

We shall suppose that $0 \leq \gamma < \infty$ when $d = 2$, and $0 \leq \gamma \leq 2/(d-2)$ when $d \geq 3$. The trivial case of $\gamma = 0$ corresponds to assuming that the function f is globally Lipschitz continuous in its third argument.

Let $\partial\Omega$ denote the union of all $(d-1)$-dimensional open faces of the polyhedron Ω. Upon decomposing $\partial\Omega$ into two parts, $\Gamma_{\rm D}$ and $\Gamma_{\rm N}$, so that $\bar{\Gamma}_{\rm D} \cup \bar{\Gamma}_{\rm N} = \overline{\partial\Omega}$ and $\Gamma_{\rm D}$ has positive $(d-1)$-dimensional Hausdorff measure, and denoting by $\nu = (\nu_1, \ldots, \nu_d)^\top$

*Received by the editors October 7, 2005; accepted for publication (in revised form) January 16, 2007; published electronically August 8, 2007. Part of the work discussed in this paper was completed while the authors were visiting the Isaac Newton Institute for Mathematical Sciences in Cambridge, UK, during the six-month program *Computational Challenges in Partial Differential Equations* (January 20–July 4, 2003).

http://www.siam.org/journals/sinum/45-4/64212.html

†Merrill Lynch, 2 King Edward Street, London EC1A 1HQ, UK (Andris_Lasis@ml.com). This authors acknowledges the financial support received from Oxford University's Clarendon Fund and the OUCL Bursary Scheme.

‡Computing Laboratory, University of Oxford, Wolfson Building, Parks Road, Oxford OX1 3QD, UK (Endre.Suli@comlab.ox.ac.uk).

the unit outward normal vector to $\partial\Omega$, we impose Dirichlet and Neumann boundary conditions on Γ_{D} and Γ_{N}, respectively:

$$\begin{aligned} u &= g_{\mathrm{D}} \qquad \text{on} \qquad \Gamma_{\mathrm{D}} \times (0,T], \\ \nabla u \cdot \nu &= g_{\mathrm{N}} \qquad \text{on} \qquad \Gamma_{\mathrm{N}} \times (0,T], \end{aligned} \tag{1.3}$$

with $g_{\mathrm{D}} \in \mathrm{H}^{1/2}(\Gamma_{\mathrm{D}})$ and $g_{\mathrm{N}} \in \mathrm{L}^2(\Gamma_{\mathrm{N}})$. Given a function $u_0 \in \mathrm{L}^2(\Omega)$, we supplement (1.1) and (1.3) with the initial condition

$$u = u_0 \quad \text{on} \quad \Omega \times \{0\}\,. \tag{1.4}$$

As the solution to the problem (1.1)–(1.4) may exhibit blowup in finite time, we shall assume that, for the potential blowup time $T^\star \in (0,\infty]$, the time interval $[0,T]$ on which the problem is considered excludes the blowup time, i.e., $T < T^\star$.

Discontinuous Galerkin finite element methods (DGFEMs) were introduced in the early 1970s for the numerical solution of first-order hyperbolic problems (see [33, 28, 26, 25, 10, 11, 12, 15, 34, 35]). Simultaneously, but independently, they were proposed as nonstandard schemes for the numerical approximation of second-order elliptic equations [32, 16, 1]. In recent years there has been renewed interest in this class of methods due to their favorable properties, which include a high degree of locality, stability in the absence of streamline-diffusion stabilization for convection-dominated diffusion problems [22], and the flexibility of locally varying the polynomial degree in adaptive *hp*-version approximations, since no pointwise continuity requirements are imposed at the element interfaces [23]. Much attention has been devoted to the analysis of DG methods applied to scalar nonlinear hyperbolic equations and hyperbolic systems [21, 7, 8], as well as to several other types of nonlinear equations, including the Hamilton–Jacobi equations [24], the nonlinear Schrödinger equation [27], and various other nonlinear problems [9]. The analysis of the spatial discretization of nonlinear parabolic problems by the interior penalty DGFEM (see [1]) was pursued by Rivière and Wheeler in [36], where the nonlinearity was assumed to be *globally* Lipschitz continuous with respect to the unknown solution.

In this work we shall be concerned with the error analysis of the *hp*-version interior penalty discontinuous Galerkin finite element method (*hp*-DGFEM) on shape-regular meshes, for the initial-boundary value problem (1.1)–(1.4). In particular, we focus on the spatial semidiscretization of the problem; however, unlike [1] and [36], we shall suppose that the nonlinearity satisfies only the *local* Lipschitz condition (1.2). As we shall see, this relaxation of the hypothesis on f leads to technical difficulties which are not present in the case when $f(x,t,\,\cdot\,)$ is globally Lipschitz continuous.

The paper is structured as follows. In section 2 we state the broken weak formulation of the problem. The error analysis of the *hp*-DGFEM approximation is discussed in section 3. We begin by establishing the local Lipschitz continuity of the mapping $w \mapsto f(\,\cdot\,,t,w(\,\cdot\,)) : \mathrm{L}^{2(\gamma+1)}(\Omega) \to \mathrm{L}^2(\Omega)$; we then show the continuity and coercivity of the bilinear form $B(\,\cdot\,,\,\cdot\,)$ appearing in the broken weak formulation of the initial-boundary value problem under consideration. Finally, we define the broken elliptic projector induced by $B(\,\cdot\,,\,\cdot\,)$ and state its approximation properties in the L^2 and broken H^1 norms. Section 3.2 contains the error analysis of the nonsymmetric version of the interior penalty *hp*-DGFEM: we prove an a priori error bound in the broken $\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ norm that is h-optimal and p-suboptimal (by half an order of p). Full *hp*-optimality of the error bound in the broken $\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ norm can be easily restored by hypothesizing piecewise regularity of the solution in augmented Sobolev

spaces instead of classical Sobolev spaces, as was done in [17] in the elliptic case; for the sake of brevity we shall not pursue this line of study here since the necessary modifications are quite straightforward. Section 3.3 is concerned with the error analysis of the symmetric version of the interior penalty *hp*-DGFEM, where we derive a priori error bounds in the broken $\mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega))$ norm (which is stronger than the broken $\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ norm, in which the error bound for the nonsymmetric version of the method was derived) as well as in the $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$ norm. Unlike the case when the nonlinearity is globally Lipschitz continuous, corresponding to the particular choice of $\gamma = 0$ in (1.2), for $\gamma > 0$ a broken counterpart of the Sobolev–Poincaré inequality has to be used to complete the error analyses of the symmetric and nonsymmetric versions of *hp*-DGFEM. The variant of the Sobolev–Poincaré inequality considered here is inspired by the work of Brenner [5]. A further ingredient of our error analysis is an adaptation to DG methods of a continuity argument due to Thomée and Wahlbin (cf. pages 382–384 in [38]). Section 4 contains some final comments on our results in this work. We remark that the analysis of the fully discrete counterpart of the method considered here, with DGFEM time discretization, proceeds in much the same manner as our analysis of the spatially semidiscrete method and is therefore omitted (cf. [29] for details and exhaustive numerical experiments). We also note that the extension of our arguments to more general second-order semilinear parabolic equations, where the Laplace operator is replaced by a general linear second-order elliptic operator in divergence form, is also straightforward (e.g., by combining the analysis presented here with that in [22]). For an extension of the analysis in this paper to a class of quasi-linear parabolic problems we refer to [29].

2. Broken weak formulation. Throughout the paper, $\mathrm{W}^{s,q}(\Omega)$ will signify the usual Sobolev space on Ω, of differentiability-index s and integrability index q, equipped with the Sobolev norm $\|\cdot\|_{s,q,\Omega}$ and seminorm $|\cdot|_{s,q,\Omega}$. In the case when $q = 2$, we shall write $\mathrm{H}^s(\Omega) := \mathrm{W}^{s,2}(\Omega)$ and suppress the index q in the notation of the norm and seminorm, writing $\|\cdot\|_{s,\Omega}$ and $|\cdot|_{s,\Omega}$, respectively. For a Banach space X equipped with a norm $\|\cdot\|$, the space $\mathrm{L}^q(0,T;X)$ consists of all strongly measurable functions $\mathbf{u} : (0,T) \to X$ with the norm

$$\|\mathbf{u}\|_{\mathrm{L}^q(0,T;X)} := \left(\int_0^T \|\mathbf{u}(t)\|^q \,\mathrm{d}t\right)^{1/q} < \infty \quad \text{for} \quad 1 \le q < \infty,$$

and with

$$\|\mathbf{u}\|_{\mathrm{L}^\infty(0,T;X)} := \operatorname*{ess.\,sup}_{0\le t\le T} \|\mathbf{u}(t)\| < \infty \quad \text{for} \quad q = \infty.$$

The Sobolev space $\mathrm{W}^{1,q}(0,T;X)$ consists of all functions $\mathbf{u} \in \mathrm{L}^q(0,T;X)$ such that $\mathbf{u}'$ exists in the weak sense and belongs to $\mathrm{L}^q(0,T;X)$, with the associated norm

$$\|\mathbf{u}\|_{\mathrm{W}^{1,q}(0,T;X)} := \left(\int_0^T \{\|\mathbf{u}(t)\|^q + \|\mathbf{u}'(t)\|^q\} \,\mathrm{d}t\right)^{1/q} < \infty \quad \text{for} \quad 1 \le q < \infty,$$

and with

$$\|\mathbf{u}\|_{\mathrm{W}^{1,\infty}(0,T;X)} := \operatorname*{ess.\,sup}_{0<t<T} (\|\mathbf{u}(t)\| + \|\mathbf{u}'(t)\|).$$

In the context of the initial-boundary value problem under consideration, $\mathbf{u}(t) = u(\cdot\,,t)$; with a slight abuse of notation, we shall simply write u in place of $\mathbf{u}$. Also, for the sake of brevity, we shall write $\mathrm{H}^1(0,T;X) := \mathrm{W}^{1,2}(0,T;X)$.

Let $\mathcal{T}_h$ be a subdivision of Ω into disjoint open elements κ such that $\bar{\Omega} = \cup_{\kappa\in\mathcal{T}_h}\bar{\kappa}$, where $\mathcal{T}_h$ is regular or 1-irregular; i.e., each face of κ has at most one hanging node. We let $h_\kappa := \mathrm{diam}(\bar{\kappa})$ and $h := \max_{\kappa\in\mathcal{T}_h} h_\kappa$; it will be assumed throughout that $h \le 1$. We assume that the family of subdivisions $\{\mathcal{T}_h\}$ is shape regular (see pages 61, 113, and Remark 2.2 on page 114, in [4]), and require each $\kappa \in \mathcal{T}_h$ to be an affine image $F_\kappa(\hat{\kappa})$ of a fixed master element $\hat{\kappa}$ for all $\kappa \in \mathcal{T}_h$, where $\hat{\kappa}$ is the open unit simplex or the open unit hypercube in $\mathbb{R}^d$. For a nonnegative integer p, we denote by $\mathcal{P}_p(\hat{\kappa})$ the set of all polynomials of degree p or less on $\hat{\kappa}$; if $\hat{\kappa}$ is the open unit hypercube in $\mathbb{R}^d$ we also consider $\mathcal{Q}_p(\hat{\kappa})$, the set of all tensor-product polynomials on $\hat{\kappa}$ of degree p or less in each coordinate direction. To each $\kappa \in \mathcal{T}_h$ we assign a nonnegative integer p_κ (the local polynomial degree) and a nonnegative integer s_κ (the local Sobolev index), collect the p_κ, s_κ, and F_κ into vectors $\mathbf{p} = \{p_\kappa : \kappa \in \mathcal{T}_h\}$, $\mathbf{s} = \{s_\kappa : \kappa \in \mathcal{T}_h\}$, and $\mathbf{F} = \{F_\kappa : \kappa \in \mathcal{T}_h\}$, respectively, and consider the finite element space

$$S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}) := \left\{v \in \mathrm{L}^2(\Omega) : v|_\kappa \circ F_\kappa \in \mathcal{R}_{p_\kappa}(\hat{\kappa}),\, \kappa \in \mathcal{T}_h\right\}, \tag{2.1}$$

where $\mathcal{R}$ is either $\mathcal{P}$ or $\mathcal{Q}$ if $\hat{\kappa}$ is the open unit hypercube in $\mathbb{R}^d$, and $\mathcal{R}$ is $\mathcal{P}$ if $\hat{\kappa}$ is the open unit simplex in $\mathbb{R}^d$.

We shall assume that the polynomial degree vector $\mathbf{p}$, with $p_\kappa \ge 1$ for each $\kappa \in \mathcal{T}_h$, has *bounded local variation*; i.e., there exists a constant $\rho \ge 1$, independent of h, such that, for any pair of elements κ and κ' in $\mathcal{T}_h$ which share a $(d-1)$-dimensional face,

$$\rho^{-1} \le p_\kappa/p_{\kappa'} \le \rho. \tag{2.2}$$

For $q \in [1,\infty)$, we assign to the subdivision $\mathcal{T}_h$ the broken Sobolev space of composite order $\mathbf{s}$,

$$\mathrm{W}^{\mathbf{s},q}(\Omega,\mathcal{T}_h) := \{u \in \mathrm{L}^q(\Omega) : u|_\kappa \in \mathrm{W}^{s_\kappa,q}(\kappa)\ \forall\ \kappa \in \mathcal{T}_h\},$$

equipped with the broken Sobolev norm and seminorm, respectively,

$$\|u\|_{\mathbf{s},q,\mathcal{T}_h} := \left(\sum_{\kappa\in\mathcal{T}_h} \|u\|^q_{s_\kappa,q,\kappa}\right)^{1/q}, \quad |u|_{\mathbf{s},q,\mathcal{T}_h} := \left(\sum_{\kappa\in\mathcal{T}_h} |u|^q_{s_\kappa,q,\kappa}\right)^{1/q}.$$

When $s_\kappa = s$ for all $\kappa \in \mathcal{T}_h$, we write $\mathrm{W}^{s,q}(\Omega,\mathcal{T}_h)$, $\|u\|_{s,q,\mathcal{T}_h}$, $|u|_{s,q,\mathcal{T}_h}$, and for $q = 2$ we let $\mathrm{H}^{\mathbf{s}} := \mathrm{W}^{\mathbf{s},2}$ and omit the index q in the notation of the norm and seminorm.

Let $\mathcal{E}$ denote the set of all open $(d-1)$-dimensional faces of the subdivision $\mathcal{T}_h$, containing the smallest common $(d-1)$-dimensional interfaces e of neighboring elements. We denote by $\mathcal{E}_{\mathrm{int}}$ the set of all faces in $\mathcal{E}$ that are contained in Ω, and we let $\Gamma_{\mathrm{int}} := \{x \in \Omega : x \in e \text{ for some } e \in \mathcal{E}_{\mathrm{int}}\}$. Further, we denote by $\mathcal{E}_\partial$ the set of all $(d-1)$-dimensional boundary faces. Assuming that each $e \in \mathcal{E}_\partial$ is a subset of the interior of exactly one of Γ_{D} and Γ_{N}, we label the associated sets of faces by $\mathcal{E}_{\mathrm{D}}$ and $\mathcal{E}_{\mathrm{N}}$. Given that $e \in \mathcal{E}_{\mathrm{int}}$, there exist positive integers i, j such that $i > j$ and that κ_i and κ_j share the face e; we define the jump of $v \in \mathrm{W}^{\mathbf{s},q}(\Omega,\mathcal{T}_h)$, $s_\kappa > 1/q$, $\kappa \in \mathcal{T}_h$, across e and the mean value of v on e by

$$[\![v]\!]_e := v|_{\partial\kappa_i\cap e} - v|_{\partial\kappa_j\cap e} \quad \text{and} \quad \{\!\{v\}\!\}_e := \tfrac{1}{2}\left(v|_{\partial\kappa_i\cap e} + v|_{\partial\kappa_j\cap e}\right),$$

respectively, with $\partial\kappa$ denoting the union of all open faces of the element κ. With each face e we associate the unit normal vector ν pointing from the element κ_i to κ_j when $i > j$; when the face belongs to $\mathcal{E}_\partial$, we choose ν to be the unit outward normal vector.

With this notation, we introduce the bilinear form

(2.3)
$$B(w,v) := \sum_{\kappa\in\mathcal{T}_h}\int_\kappa \nabla w\cdot\nabla v\,\mathrm{d}x + \int_{\Gamma_\mathrm{D}}\{\theta(\nabla v\cdot\nu)w - (\nabla w\cdot\nu)v\}\,\mathrm{d}s + \int_{\Gamma_\mathrm{D}}\sigma wv\,\mathrm{d}s$$
$$+\int_{\Gamma_\mathrm{int}}\{\theta\{\!\{\nabla v\cdot\nu\}\!\}[\![w]\!] - \{\!\{\nabla w\cdot\nu\}\!\}[\![v]\!]\}\,\mathrm{d}s + \int_{\Gamma_\mathrm{int}}\sigma[\![w]\!][\![v]\!]\,\mathrm{d}s$$

and the linear functional

$$\ell(v) := \int_{\Gamma_\mathrm{N}} g_\mathrm{N} v\,\mathrm{d}s + \theta\int_{\Gamma_\mathrm{D}}(\nabla v\cdot\nu)g_\mathrm{D}\,\mathrm{d}s + \int_{\Gamma_\mathrm{D}}\sigma g_\mathrm{D} v\,\mathrm{d}s. \tag{2.4}$$

Here σ is called the *discontinuity-penalization parameter* and is defined by

$$\sigma|_e = \sigma_e \quad \text{for} \quad e\in\mathcal{E}_\mathrm{int}\cup\mathcal{E}_\partial,$$

where σ_e is a positive constant on the face e. The precise choice of σ_e will be discussed in section 3. The subscript e in these definitions will be suppressed when no confusion is likely to occur. The parameter θ takes its values in the interval $[-1,1]$. Since for $\theta\neq\pm1$ the analysis of the method is similar to that in the case of $\theta=\pm1$, for the sake of brevity we shall suppose throughout that $\theta\in\{-1,1\}$. The choice of $\theta=-1$ leads to a symmetric bilinear form $B(\,\cdot\,,\,\cdot\,)$; we call the associated method the *symmetric interior penalty*, or SIP, method. On the other hand, for $\theta=1$ the bilinear form $B(\,\cdot\,,\,\cdot\,)$ is nonsymmetric, but it is coercive for any $\sigma>0$; we call the corresponding method the *nonsymmetric interior penalty*, or NSIP, method. In order to distinguish between the two methods, we shall label the bilinear form (2.3) and the linear functional (2.4) with indices S and NS in the symmetric and nonsymmetric cases, corresponding to $\theta=-1$ and $\theta=1$, respectively.

Then, the broken weak formulation of the problem (1.1)–(1.4) reads as follows:

(2.5) Find $u\in\mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))\cap\mathrm{L}^2(0,T;\mathfrak{A})$ such that

$$\int_\Omega u'v\,\mathrm{d}x + B(u,v) - \int_\Omega f(x,t,u)v\,\mathrm{d}x = \ell(v)\quad\forall v\in\mathrm{H}^2(\Omega,\mathcal{T}_h),$$
$$u(0)=u_0,$$

where by $\mathfrak{A}$ we denote the function space

$$\mathfrak{A} = \left\{w\in\mathrm{H}^2(\Omega,\mathcal{T}_h) : w,\ \nabla w\cdot\nu \text{ are continuous across each } e\in\mathcal{E}_\mathrm{int}\right\}.$$

Note that if $u\in\mathrm{H}^1(0,T;\mathrm{L}^2(\Omega))\cap\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ is a weak solution of (1.1)–(1.4) and $u\in\mathrm{L}^2(0,T;\mathfrak{A})$, then u also solves (2.5); in what follows we shall always assume that such a u exists.

The *hp*-DGFEM approximation of problem (1.1)–(1.4) is as follows:

(2.6) Find $u_{\mathrm{DG}} \in \mathrm{H}^1(0,T;S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}))$ such that

$$\int_\Omega u'_{\mathrm{DG}} v \,\mathrm{d}x + B(u_{\mathrm{DG}},v) - \int_\Omega f(x,t,u_{\mathrm{DG}})v\,\mathrm{d}x = \ell(v) \quad \forall v \in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}),$$

$$u_{\mathrm{DG}}(0) = u_0^{\mathrm{DG}},$$

where u_0^{DG} denotes an approximation of the function u_0 from the finite element space $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$. The purpose of this paper is to quantify the size of the error between the solution u to (2.5) and its *hp*-DGFEM approximation u_{DG} in various norms.

Equation (2.6) can be interpreted as a system of ordinary differential equations (ODEs) for the coefficients in the expansion of $u_{\mathrm{DG}}(\,\cdot\,,t)$ in terms of basis functions of the finite-dimensional space $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$. Thus, (2.6) represents a nonautonomous system of ODEs with locally Lipschitz continuous right-hand side, given that $f(x,t,\,\cdot\,)$ is locally Lipschitz continuous, uniformly in $(x,t) \in \Omega \times (0,T]$, and the other terms are linear. By Carathéodory's theorem (see Theorems II.4.1 and II.4.5 in [39]) this, in turn, implies the existence of a unique local solution to (2.6) on a certain maximal subinterval $[0,t_{\star\star})$ of $[0,T]$. In fact, we shall show below that u_{DG} exists on the whole of $[0,T]$.

3. Error analysis. Before embarking on the analysis of (2.6), we state and prove some preliminary results.

3.1. Preliminary results. We begin by establishing the local Lipschitz-continuity of $w \mapsto f(\,\cdot\,,t,w(\,\cdot\,))$ as a mapping from $\mathrm{L}^{2(\gamma+1)}(\Omega)$ to $\mathrm{L}^2(\Omega)$.

LEMMA 3.1. *Suppose that f satisfies* (**A**). *Then, there exists a positive constant $C_f = C_f(\gamma, G_f, |\Omega|)$ such that*

$$\|f(\,\cdot\,,t,w) - f(\,\cdot\,,t,v)\|_{0,\Omega} \le C_f \|w-v\|_{0,2(\gamma+1),\Omega} \times \left(1 + \|w\|^\gamma_{0,2(\gamma+1),\Omega} + \|v\|^\gamma_{0,2(\gamma+1),\Omega}\right) \tag{3.1}$$

for all w, $v \in \mathrm{L}^{2(\gamma+1)}(\Omega)$ and a.e. $t \in (0,T]$.

Proof. Let $q = 2(\gamma+1)$. Note that if $d = 2$, $0 \le \gamma < \infty$, then $2 \le q < \infty$, and if $d \ge 3$, $0 \le \gamma \le 2/(d-2)$, then $2 \le q \le 2d/(d-2)$. Let us suppose that $\gamma > 0$ and therefore $q > 2$; for $\gamma = 0$, (3.1) trivially holds with $C_f = G_f/3$. Let w, $v \in \mathrm{L}^{2(\gamma+1)}(\Omega)$; from (1.2), by Hölder's inequality, for a.e. $t \in (0,T]$ we have

$$\begin{aligned}\|f(\,\cdot\,,t,w) - f(\,\cdot\,,t,v)\|^2_{0,\Omega} &\le G_f^2 \int_\Omega (w-v)^2(1+|w|+|v|)^{2\gamma}\,\mathrm{d}x \\ &\le G_f^2 \left(\int_\Omega |w-v|^{2\cdot q/2}\,\mathrm{d}x\right)^{2/q} \left(\int_\Omega (1+|w|+|v|)^{2\gamma\cdot(1-2/q)^{-1}}\,\mathrm{d}x\right)^{1-2/q}.\end{aligned}$$

As $1-2/q=(q-2)/q=2\gamma/q$ and $q>2$, we have

$$\begin{aligned}\|f(\,\cdot\,,t,w)-f(\,\cdot\,,t,v)\|_{0,\Omega}^2 &\le G_f^2\|w-v\|_{0,q,\Omega}^2\left(\int_\Omega(1+|w|+|v|)^q\,\mathrm{d}x\right)^{2\gamma/q}\\ &\le G_f^2\|w-v\|_{0,q,\Omega}^2\left(|\Omega|^{1/q}+\|w\|_{0,q,\Omega}+\|v\|_{0,q,\Omega}\right)^{2\gamma}\\ &\le C_f^2\|w-v\|_{0,q,\Omega}^2\left(1+\|w\|_{0,q,\Omega}^{\gamma}+\|v\|_{0,q,\Omega}^{\gamma}\right)^2,\end{aligned}$$

and hence (3.1) for a.e. $t\in(0,T]$ and all $w,v\in\mathrm{L}^q(\Omega)$, $q=2(\gamma+1)$. ∎

We equip $\mathrm{H}^1(\Omega,\mathcal{T}_h)$ with the norm $\|\cdot\|_{1,h}$ defined by

$$\|w\|_{1,h}:=\left(\sum_{\kappa\in\mathcal{T}_h}\|\nabla w\|_{0,\kappa}^2+\int_{\Gamma_{\mathrm{D}}}\sigma w^2\,\mathrm{d}s+\int_{\Gamma_{\mathrm{int}}}\sigma\,[\![w]\!]^2\,\mathrm{d}s\right)^{1/2},\tag{3.2}$$

where σ is the positive discontinuity-penalization parameter which was introduced after (2.4). In addition, we define the norm $|||\cdot|||_{1,h}$ by

$$|||w|||_{1,h}:=\left(\sum_{\kappa\in\mathcal{T}_h}\|\nabla w\|_{0,\kappa}^2+|\Gamma_{\mathrm{D}}|^{-1}\int_{\Gamma_{\mathrm{D}}}w^2\,\mathrm{d}s+\sum_{e\in\mathcal{E}_{\mathrm{int}}}h_e^{-1}\int_e[\![w]\!]^2\,\mathrm{d}s\right)^{1/2}.\tag{3.3}$$

The parameter σ will be chosen so that $\sigma|_e=\sigma_e$ on each face $e\in\mathcal{E}$ and $\sigma_e\ge C_\sigma/h_e$, where C_σ is a positive constant whose value will be fixed later on; here h_e denotes the diameter of the face e. With this choice of σ, by noting that $1\le h_e^{-1}\le\sigma_e/C_\sigma$ (since $h\le1$) and that $|\Gamma_{\mathrm{D}}|^{-1}\int_{\Gamma_{\mathrm{D}}}w^2\,\mathrm{d}s=|\Gamma_{\mathrm{D}}|^{-1}\sum_{e\in\mathcal{E}_{\mathrm{D}}}1\cdot\int_e w^2\,\mathrm{d}s$, we have that

$$|||w|||_{1,h}\le C\|w\|_{1,h}\qquad\forall w\in\mathrm{H}^1(\Omega,\mathcal{T}_h),\tag{3.4}$$

where C is a constant independent of h and w. The next lemma will play a key role.

LEMMA 3.2 (broken Sobolev–Poincaré inequality). *There exists a positive constant C, independent of h, such that for any $q\in[1,\infty)$ when $d=2$ and any $q\in[1,2d/(d-2)]$ when $d\ge3$,*

$$\|w\|_{0,q,\Omega}\le C|||w|||_{1,h}\qquad\forall w\in\mathrm{H}^1(\Omega,\mathcal{T}_h).\tag{3.5}$$

Proof. From [30, Theorem 3.7], using the notation therein, we define Ψ as in Example 3.6 of that paper, with $\psi\in\mathrm{L}^2(\partial\Omega)$, $\psi\equiv0$ on Γ_{N}, to obtain (3.5). ∎

LEMMA 3.3. *Suppose that f satisfies* (**A**). *Then, there exists a positive constant $C_f=C_f(\gamma,G_f,d,|\Omega|)$ such that*

$$|(f(\,\cdot\,,t,u)-f(\,\cdot\,,t,v),w)|\le C_f\|u-v\|_{0,\Omega}\left(1+|||u|||_{1,h}^{\gamma}+|||v|||_{1,h}^{\gamma}\right)\|w\|_{1,h}\tag{3.6}$$

for all u, v, $w\in\mathrm{H}^1(\Omega,\mathcal{T}_h)$ and a.e. $t\in(0,T]$.

Proof. Let $u,v,w\in\mathrm{H}^1(\Omega,\mathcal{T}_h)$. Let $p>1$ and $1/p+1/q=1$. By applying

Hölder's inequality, we obtain that, for a.e. $t \in (0,T]$,

$$\begin{aligned}|(f(\cdot,t,u)-f(\cdot,t,v),w)| &\le G_f \int_\Omega |u-v|\,(1+|u|+|v|)^\gamma\,|w|\,\mathrm{d}x\\ &\le G_f \left(\int_\Omega |u-v|^2\,\mathrm{d}x\right)^{1/2}\left(\int_\Omega (1+|u|+|v|)^{2\gamma}\,|w|^2\,\mathrm{d}x\right)^{1/2}\\ &\le G_f \left(\int_\Omega |u-v|^2\,\mathrm{d}x\right)^{1/2}\left(\int_\Omega (1+|u|+|v|)^{2\gamma p}\,\mathrm{d}x\right)^{1/2p}\left(\int_\Omega |w|^{2q}\,\mathrm{d}x\right)^{1/2q}.\end{aligned}$$

When $d \ge 3$ we take $p = d/2$, $q = d/(d-2)$; while if $d = 2$, we take any $p > 1$ and put $q = p/(p-1)$. The desired inequality then follows by using the broken Sobolev–Poincaré inequality (3.5), with q thus defined, and by applying (3.4). ∎

We recall the following approximation result for the space $S^{\mathbf{P}}(\Omega,\mathcal{T}_h,\mathbf{F})$.

LEMMA 3.4. *Suppose that $\kappa \in \mathcal{T}_h$ with $h_\kappa = \mathrm{diam}(\bar\kappa)$ and $u|_\kappa \in \mathrm{H}^{k_\kappa}(\kappa)$ for some $k_\kappa \ge 0$; then there exists a sequence of algebraic polynomials $z_{p_\kappa}^{h_\kappa}(u) \in \mathcal{R}_{p_\kappa}(\kappa)$, $p_\kappa \ge 1$, such that for any l, with $0 \le l \le s_\kappa$,*

$$\|u - z_{p_\kappa}^{h_\kappa}(u)\|_{l,\kappa} \le C\,\frac{h_\kappa^{s_\kappa - l}}{p_\kappa^{k_\kappa - l}}\,\|u\|_{k_\kappa,\kappa}, \tag{3.7}$$

and when $k_\kappa > 1/2$, then

$$\|u - z_{p_\kappa}^{h_\kappa}(u)\|_{0,e_\kappa} \le C\,\frac{h_\kappa^{s_\kappa - 1/2}}{p_\kappa^{k_\kappa - 1/2}}\,\|u\|_{k_\kappa,\kappa}; \tag{3.8}$$

further, if $k_\kappa > 3/2$, then

$$\|\nabla(u - z_{p_\kappa}^{h_\kappa}(u))\|_{0,e_\kappa} \le C\,\frac{h_\kappa^{s_\kappa - 3/2}}{p_\kappa^{k_\kappa - 3/2}}\,\|u\|_{k_\kappa,\kappa}, \tag{3.9}$$

where e_κ is any face $e_\kappa \subset \partial\kappa$, $s_\kappa = \min\{p_\kappa + 1, k_\kappa\}$, and C is a constant independent of u, h_κ, and p_κ but dependent on $k = \max_{\kappa\in\mathcal{T}_h} k_\kappa$.

Proof. For the proof of (3.7), see [3, Lemma 4.5] for $d = 2$ (the argument being analogous when $d > 2$). By using the multiplicative trace inequality

$$\|u\|_{0,\partial\kappa} \le C(d)\left(h_\kappa^{-1/2}\|u\|_{0,\kappa} + \|u\|_{0,\kappa}^{1/2}\|\nabla u\|_{0,\kappa}^{1/2}\right),$$

we obtain (3.8) and (3.9) from (3.7). ∎

Remark 3.5. If the reference element $\hat\kappa$ is the d-dimensional hypercube, instead of the Babuška–Suri projector $z_{p_\kappa}^{h_\kappa}$ we can use the Jackson-type quasi-interpolation operator $J_{p_\kappa}^{k_\kappa}$ (for its definition see [14, Chapter 7]; the error bounds are presented in [31, Theorem A.3]).

We require the following bound on the bilinear form $B(\cdot,\cdot)$ (see [22] for a proof), a key ingredient of which is the inverse inequality

$$\|\nabla w\|_{0,\partial\kappa\cap\Gamma_{\mathrm{D}}}^2 \le C_{\mathrm{inv}}\frac{p_\kappa^2}{h_\kappa}\|\nabla w\|_{0,\kappa}^2, \tag{3.10}$$

where the constant C_{inv} depends only on the shape-regularity constant of the family $\{\mathcal{T}_h\}$ (see Schwab [37, Theorem 4.76, inequality (4.6.4)]).

LEMMA 3.6. *There exists a positive constant C, independent of the discretization parameters, such that the following inequality holds for all $v \in \mathrm{H}^1(\Omega, \mathcal{T}_h)$ and all $w \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$:*

$$
\begin{aligned}
|B(v,w)| \le C\,\|w\|_{1,h} \int_{\Gamma_{\mathrm{D}}} \sigma\,|v|^2\,\mathrm{d}s + \int_{\Gamma_{\mathrm{int}}} \sigma\,[\![v]\!]^2\,\mathrm{d}s + \sum_{\kappa\in\mathcal{T}_h} \|\nabla v\|^2_{0,\kappa} \\
+ \sum_{\kappa\in\mathcal{T}_h} \left(\|\sqrt{\tau} v\|^2_{0,\partial\kappa\cap\Gamma_{\mathrm{D}}} + \|\sigma^{-1/2}\nabla v\|^2_{0,\partial\kappa\cap\Gamma_{\mathrm{D}}}\right) \\
+ \sum_{\kappa\in\mathcal{T}_h} \left(\|\sqrt{\tau}\,[\![v]\!]\,\|^2_{0,\partial\kappa\cap\Gamma_{\mathrm{int}}} + \|\sigma^{-1/2}\nabla v\|^2_{0,\partial\kappa\cap\Gamma_{\mathrm{int}}}\right)^{1/2},
\end{aligned}
\tag{3.11}
$$

where $\tau_e = \{\!\{p^2\}\!\}_e/h_e$ and h_e is the diameter of a face $e \subset \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{D}}$; when $e \in \mathcal{E}_{\mathrm{D}}$ the contribution from outside Ω in the definition of τ_e is set to 0.

Next, we shall investigate the coercivity of the bilinear form $B : S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) \times S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) \to \mathbb{R}$ defined by (2.3). In the nonsymmetric case (with $\theta = 1$) coercivity follows directly as, for any $w \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, we have $B_{\mathrm{NS}}(w,w) = \|w\|^2_{1,h}$.

Consider now the symmetric bilinear form (2.3) (with $\theta = -1$). We have, for any $w \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$,

$$
\begin{aligned}
B_{\mathrm{S}}(w,w) = \sum_{\kappa\in\mathcal{T}_h} \|\nabla w\|^2_{0,\kappa} + \int_{\Gamma_{\mathrm{D}}} \left(\sigma w^2 - 2w(\nabla w\cdot\nu)\right)\,\mathrm{d}s \\
+ \int_{\Gamma_{\mathrm{int}}} \left(\sigma\,[\![w]\!]^2 - 2\,[\![w]\!]\,\{\!\{\nabla w\cdot\nu\}\!\}\right)\,\mathrm{d}s.
\end{aligned}
$$

Clearly, the integrands in the last two terms need not be positive for $w \neq 0$ unless σ is chosen sufficiently large: the purpose of the analysis that now follows is to assess just how large σ needs to be to ensure coercivity of $B_{\mathrm{S}}(\,\cdot\,,\,\cdot\,)$ over $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F}) \times S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$.

For any positive number τ_e we have

$$
-2\int_{\Gamma_{\mathrm{D}}} w(\nabla w\cdot\nu)\,\mathrm{d}s \ge -\sum_{e\in\mathcal{E}_{\mathrm{D}}} \left(\int_e \tau_e w^2\,\mathrm{d}s + \int_e \tau_e^{-1}(\nabla w\cdot\nu)^2\,\mathrm{d}s\right).
$$

Omitting the summation sign but retaining the minus sign in front of it, we see that the second term on the right-hand side can be further bounded below by using the inverse inequality (3.10), the shape-regularity condition (to relate h_κ to h_e, where κ is the element whose face $e \in \mathcal{E}_{\mathrm{D}}$ is) and letting $p_e = p_\kappa$ for $e \subset \partial\kappa\cap\Gamma_{\mathrm{D}}$; by absorbing all constants into C_τ, we thus obtain

$$
-\int_e \tau_e^{-1}(\nabla w\cdot\nu)^2\,\mathrm{d}s \ge -\int_e \tau_e^{-1}\,|\nabla w|^2\,|\nu|^2\,\mathrm{d}s \ge -\tau_e^{-1} C_\tau \frac{p_e^2}{h_e}\,\|\nabla w\|^2_{0,\kappa},
$$

and hence

$$
-2\int_{\Gamma_{\mathrm{D}}} w(\nabla w\cdot\nu)\,\mathrm{d}s \ge -\sum_{e\in\mathcal{E}_{\mathrm{D}}} \left(\int_e \tau_e w^2\,\mathrm{d}s + \tau_e^{-1} C_\tau \frac{p_e^2}{h_e}\|\nabla w\|^2_{0,\kappa}\right).
$$

Similarly, for the term involving faces $e \in \mathcal{E}_{\rm int}$, we have, using the bounded local variation condition (to relate p_κ^2 to $\{\!\{p^2\}\!\}_e$),

$$-2\int_{\Gamma_{\rm int}} [\![w]\!]\,\{\!\{\nabla w\cdot\nu\}\!\}\,{\rm d}s \geq -\sum_{e\in\mathcal{E}_{\rm int}}\left(\int_e \tau_e\,[\![w]\!]^2\,{\rm d}s + \tau_e^{-1}C_\tau\frac{\{\!\{p^2\}\!\}_e}{2h_e}\left(\|\nabla w\|^2_{0,\kappa'}+\|\nabla w\|^2_{0,\kappa''}\right)\right);$$

here κ' and κ'' are the two elements that have e as their common face.

Thanks to our assumption that no face e of any element $\kappa\in\mathcal{T}_h$ contains more than one hanging node, it follows that no element κ can have more than $2d\cdot 2^{d-1}=2^d d$ faces if $\hat\kappa$ is the d-dimensional hypercube, or more than $(d+1)d$ faces if $\hat\kappa$ is the d-dimensional simplex. On writing $c_d=\max\left\{2^d d,(d+1)d\right\}=2^d d$, we then let $\tau_e:=c_dC_\tau\{\!\{p^2\}\!\}_e/h_e$ for $e\in\mathcal{E}_{\rm D}\cup\mathcal{E}_{\rm int}$ (with the convention that, for $e\in\mathcal{E}_{\rm D}$, $\{\!\{p^2\}\!\}_e=p_e^2/2=p_\kappa^2/2$ where κ is the element in $\mathcal{T}_h$ with face e); hence,

$$B_{\rm S}(w,w)\geq\frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\|\nabla w\|^2_{0,\kappa}+\int_{\Gamma_{\rm D}}(\sigma-\tau)w^2\,{\rm d}s+\int_{\Gamma_{\rm int}}(\sigma-\tau)\,[\![w]\!]^2\,{\rm d}s.$$

Choosing σ_e appropriately, i.e., letting $\sigma_e=C_\sigma\{\!\{p^2\}\!\}_e/h_e$ with the penalty constant $C_\sigma>0$ large enough (that is, with $C_\sigma>c_dC_\tau$), will ensure that $\sigma_e>\tau_e$ for all $e\in\mathcal{E}_{\rm int}\cup\mathcal{E}_{\rm D}$, and hence that the symmetric bilinear form $B_{\rm S}(\cdot,\cdot)$ is coercive. When $e\in\mathcal{E}_{\rm D}$ the contribution from outside Ω in the definition of σ_e is set to 0. Thus, we adopt the following hypothesis concerning the choice of the penalty constant C_σ:

(**B**) In the *nonsymmetric* case (when $\theta=1$) we take any $C_\sigma>0$. In the *symmetric* case (when $\theta=-1$) we take $C_\sigma>c_dC_\tau$.

We summarize our findings about the bilinear forms $B_{\rm NS}(\,\cdot\,,\,\cdot\,)$ and $B_{\rm S}(\,\cdot\,,\,\cdot\,)$ in the next lemma.

LEMMA 3.7. *The nonsymmetric bilinear form $B_{\rm NS}(\,\cdot\,,\,\cdot\,)$ is coercive in the norm $\|\,\cdot\,\|_{1,h}$ over the space $\mathrm{H}^1(\Omega,\mathcal{T}_h)\times\mathrm{H}^1(\Omega,\mathcal{T}_h)$; more precisely,*

$$B_{\rm NS}(w,w)=\|w\|^2_{1,h}\qquad\forall w\in\mathrm{H}^1(\Omega,\mathcal{T}_h).$$

With the constant C_σ chosen as in (**B**)*, the symmetric bilinear form $B_{\rm S}(\,\cdot\,,\,\cdot\,)$ induces a norm $\|\,\cdot\,\|_B$ on the finite element space $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$; moreover, there exists a positive constant c_0 such that*

$$B_{\rm S}(w,w)=\|w\|_B^2\geq c_0\|w\|^2_{1,h}\qquad\forall w\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}),$$

i.e., $B_{\rm S}(\,\cdot\,,\,\cdot\,)$ is coercive in $\|\,\cdot\,\|_{1,h}$ over the space $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})\times S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$.

Let us consider the projection operator Π from $\mathrm{H}^2(\Omega,\mathcal{T}_h)$ onto the finite element space $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ defined (for $u\in\mathrm{H}^2(\Omega,\mathcal{T}_h)$) by

$$B(u-\Pi u,v)=0\qquad\forall v\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}).\tag{3.12}$$

As $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ is a finite-dimensional linear space, the existence of a unique Πu in $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ for each $u\in\mathrm{H}^2(\Omega,\mathcal{T}_h)$ follows from the coercivity of the bilinear form $B(\,\cdot\,,\,\cdot\,)$ over $S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})\times S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$; Πu will be referred to as the *broken elliptic projection* of u, and $\Pi:\mathrm{H}^2(\Omega,\mathcal{T}_h)\to S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ as the *broken elliptic projector*.

Next, we establish bounds on the approximation error $u-\Pi u$, in the H^1 and L^2 norms, for the broken elliptic projector Π.

Suppose that $\Omega \subset \mathbb{R}^d$, $d \geq 2$, is a bounded open polyhedral domain with Lipschitz-continuous boundary. We shall say that Ω is H^2-*regular* if, for any $u \in \mathrm{H}^1_0(\Omega)$ with $\Delta u \in \mathrm{L}^2(\Omega)$, u belongs to $\mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$ and there exists a constant $c_* = c_*(\Omega, d) > 0$, independent of u, such that $\|u\|_{\mathrm{H}^2(\Omega)} \leq c_* \|\Delta u\|_{\mathrm{L}^2(\Omega)}$.

LEMMA 3.8. *Suppose that $u|_\kappa \in \mathrm{H}^{k_\kappa}(\kappa)$ for some Sobolev index $k_\kappa \geq 2$ and each $\kappa \in \mathcal{T}_h$. Let Πu be the projection of $u \in \mathrm{H}^2(\Omega, \mathcal{T}_h)$ onto $S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, defined by (3.12), with $p_\kappa \geq 1$ for each $\kappa \in \mathcal{T}_h$, and*

$$\sigma_e = C_\sigma \{\!\{p^2\}\!\}|_e / h_e,$$

*where C_σ is as in (**B**), and h_e is the diameter of a face $e \in \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{D}}$; when $e \in \mathcal{E}_{\mathrm{D}}$, the contribution from outside Ω in the definition of σ_e is set to 0. Then, the following error bound holds:*

$$\|u - \Pi u\|^2_{1,h} \leq C \sum_{\kappa \in \mathcal{T}_h} \frac{h_\kappa^{2s_\kappa - 2}}{p_\kappa^{2k_\kappa - 3}} \|u\|^2_{k_\kappa, \kappa}, \tag{3.13}$$

where $s_\kappa = \min\{p_\kappa + 1, k_\kappa\}$, and the constant C is independent of u, p_κ, and h_κ but depends on $k = \max_{\kappa \in \mathcal{T}_h} k_\kappa$, the parameter ρ in (2.2), and C_σ.

Moreover, if $\theta = -1$, Γ_N is empty (i.e., $\Gamma_{\mathrm{D}} = \partial\Omega$) and Ω is an H^2-regular polyhedral domain, we have

$$\|u - \Pi u\|^2_{0,\Omega} \leq C \left(\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa} \right) \sum_{\kappa \in \mathcal{T}_h} \frac{h_\kappa^{2s_\kappa - 2}}{p_\kappa^{2k_\kappa - 3}} \|u\|^2_{k_\kappa, \kappa}, \tag{3.14}$$

where $s_\kappa = \min\{p_\kappa + 1, k_\kappa\}$, and the constant C is independent of u, p_κ, and h_κ but depends on $k = \max_{\kappa \in \mathcal{T}_h} k_\kappa$, the parameter ρ in (2.2), the constant C_σ, and Ω.

Proof. By recalling the definition of the norm (3.2) and applying Lemma 3.7, we have that $B(w, w) \geq c_0 \|w\|^2_{1,h}$ for all $w \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$. On writing $u - \Pi u = (u - z^{h_\kappa}_{p_\kappa}(u)) + (z^{h_\kappa}_{p_\kappa}(u) - \Pi u) =: \eta + \xi$, with the projection operator $u \mapsto z^{h_\kappa}_{p_\kappa}(u)$ defined as in Lemma 3.4, and taking $v = \xi$ in the definition of the broken elliptic projector (3.12), we then deduce that

$$c_0 \|\xi\|^2_{1,h} \leq B(\xi, \xi) = B((u - \Pi u) - \eta, \xi) = -B(\eta, \xi) \leq |B(\eta, \xi)|.$$

By Lemma 3.6 with $v = \eta$, $w = \xi$, and the above inequality, noting that $\sigma_e = C_\sigma \{\!\{p^2\}\!\}|_e / h_e > \tau_e$, with h_e the diameter of a face $e \in \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{D}}$ and with the contribution from outside Ω in $\{\!\{p^2\}\!\}|_e$ set to 0 for $e \in \mathcal{E}_{\mathrm{D}}$, we have that

$$\|\xi\|^2_{1,h} \leq C \sum_{\kappa \in \mathcal{T}_h} \Big(\|\sqrt{\sigma}\eta\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{D}}} + \|\sqrt{\sigma}\,[\![\eta]\!]\,\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{int}}} + \|\nabla\eta\|^2_{0,\kappa} + \|\sigma^{-1/2}\nabla\eta\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{D}}} + \|\sigma^{-1/2}\nabla\eta\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{int}}} \Big).$$

Using the triangle inequality $\|u - \Pi u\|_{1,h} \leq \|\eta\|_{1,h} + \|\xi\|_{1,h}$ and recalling the definition of the norm $\|\cdot\|_{1,h}$, we then obtain the bound

$$\|u - \Pi u\|^2_{1,h} \leq C \sum_{\kappa \in \mathcal{T}_h} \Big(\|\sqrt{\sigma}\eta\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{D}}} + \|\sqrt{\sigma}\,[\![\eta]\!]\,\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{int}}} + \|\nabla\eta\|^2_{0,\kappa} + \|\sigma^{-1/2}\nabla\eta\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{D}}} + \|\sigma^{-1/2}\nabla\eta\|^2_{0,\partial\kappa \cap \Gamma_{\mathrm{int}}} \Big). \tag{3.15}$$

From Lemma 3.4, inequalities (3.7)–(3.9), we deduce that

$$\|\eta\|^2_{0,\partial\kappa} \le C\frac{h_\kappa^{2s_\kappa-1}}{p_\kappa^{2k_\kappa-1}}\|u\|^2_{k_\kappa,\kappa},\ \|\nabla\eta\|^2_{0,\partial\kappa} \le C\frac{h_\kappa^{2s_\kappa-3}}{p_\kappa^{2k_\kappa-3}}\|u\|^2_{k_\kappa,\kappa},\ \|\eta\|^2_{1,\kappa} \le C\frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-2}}\|u\|^2_{k_\kappa,\kappa}.$$

Applying these bounds to the right-hand side of (3.15), choosing σ_e as assumed in the statement of the lemma, and noting the bounded local variation condition (2.2) and the shape-regularity of $\mathcal{T}_h$ to relate h_e to h_κ, we obtain the bound

$$\|u-\Pi u\|^2_{1,h} \le C\sum_{\kappa\in\mathcal{T}_h}\left(\frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-2}} + \frac{p_\kappa^2}{h_\kappa}\frac{h_\kappa^{2s_\kappa-1}}{p_\kappa^{2k_\kappa-1}}\right)\|u\|^2_{k_\kappa,\kappa},$$

and hence (3.13).

To estimate $\|u-\Pi u\|_{0,\Omega}$ in the case of $\theta=-1$ and $\Gamma_{\rm D}=\partial\Omega$, when Ω is H^2-regular, we shall use the Aubin–Nitsche duality argument (see [6]). Let $(\,\cdot\,,\,\cdot\,)$ signify the L^2 inner product over Ω. Then,

$$\|u-\Pi u\|_{0,\Omega} = \sup_{\substack{g\in \mathrm{L}^2(\Omega)\\ g\neq 0}} \frac{(u-\Pi u, g)}{\|g\|_{0,\Omega}}. \tag{3.16}$$

For $g\in\mathrm{L}^2(\Omega)$ fixed, $g\neq 0$, let $w=w_g\in\mathrm{H}^1_0(\Omega)$ be the weak solution of the problem

$$\begin{aligned} -\Delta w &= g \qquad \text{in} \qquad \Omega,\\ w &= 0 \qquad \text{on} \qquad \partial\Omega. \end{aligned} \tag{3.17}$$

Since Ω is H^2-regular, we have that $w\in\mathrm{H}^2(\Omega)\cap\mathrm{H}^1_0(\Omega)$, and there exists a positive constant c_*, independent of g and w, such that

$$\|w\|_{2,\Omega} \le c_*\|g\|_{0,\Omega}. \tag{3.18}$$

Moreover, $w\in\mathrm{C}^1(\Omega)$ by [18, Corollary 8.36]. The SIP DGFEM approximation of the problem (3.17) is as follows:

$$\text{Find}\ \ w_{\rm DG}\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})\ \ \text{such that}\ \ B_{\rm S}(w_{\rm DG},v)=\ell_g(v)\qquad \forall v\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F}),$$

where $B_{\rm S}(w,v)$ is defined by (2.3) with $\theta=-1$, and $\ell_g(v)=(g,v)+\ell_{\rm S}(v)$, where $\ell_{\rm S}(v)$ is defined by (2.4) with $\theta=-1$ and $g_{\rm D}=0$ on $\Gamma_{\rm D}=\partial\Omega$ (also, as $\Gamma_{\rm N}=\emptyset$, the integral over $\Gamma_{\rm N}$ in (2.4) vanishes); clearly, then, $\ell_{\rm S}(v)=0$ for all v in $\mathrm{H}^2(\Omega,\mathcal{T}_h)$.

Using the fact that $w\in\mathrm{H}^2(\Omega)\cap\mathrm{H}^1_0(\Omega)\cap\mathrm{C}^1(\Omega)$, we deduce that $B_{\rm S}(w,v)=\ell_g(v)$ for all $v\in\mathrm{H}^2(\Omega,\mathcal{T}_h)$. Moreover, by the definition of the broken elliptic projector (3.12), $B_{\rm S}(u-\Pi u,v)=0$ for all $v\in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$, and hence

$$\begin{aligned}(u-\Pi u,g) &= (g,u-\Pi u)=\ell_g(u-\Pi u)=B_{\rm S}(w,u-\Pi u)\\ &= B_{\rm S}(u-\Pi u,w)=B_{\rm S}(u-\Pi u,w-z^{h_\kappa}_{p_\kappa}(w)).\end{aligned}$$

Further, on denoting $\eta_w:=w-z^{h_\kappa}_{p_\kappa}(w)$, by (3.11), and noting that $\sigma_e=C_\sigma\{\!\{p^2\}\!\}|_e/h_e>\tau_e$, with h_e the diameter of a face $e\in\mathcal{E}_{\rm int}\cup\mathcal{E}_{\rm D}$ and with the contribution from outside

Ω in $\{\!\{p^2\}\!\}|_e$ set to 0 for $e \in \mathcal{E}_{\rm D}$, we have that

(3.19)
$$(u-\Pi u, g) = B_{\rm S}(u-\Pi u, \eta_w) \le C\,\|u-\Pi u\|_{1,h} \left\{ \int_{\Gamma_{\rm D}} \sigma\,|\eta_w|^2\,{\rm d}s + \int_{\Gamma_{\rm int}} \sigma\,[\![\eta_w]\!]^2\,{\rm d}s \right.$$
$$\left. + \sum_{\kappa\in\mathcal{T}_h} \left(\|\nabla\eta_w\|^2_{0,\kappa} + \|\sigma^{-1/2}\nabla\eta_w\|^2_{0,\partial\kappa\cap\Gamma_{\rm D}} + \|\sigma^{-1/2}\nabla\eta_w\|^2_{0,\partial\kappa\cap\Gamma_{\rm int}} \right) \right\}^{1/2}.$$

Applying (3.13) and inequalities (3.7)–(3.9) from Lemma 3.4 to the right-hand side of (3.19), choosing σ_e as described in the statement of Lemma 3.8 above, and noting the bounded local variation condition (2.2) and the shape-regularity of $\mathcal{T}_h$ to relate h_e to h_κ, we obtain

$$(u-\Pi u, g) \le C \left(\sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\,\|u\|^2_{k_\kappa,\kappa} \times \sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa}\|w\|^2_{2,\kappa} \right)^{1/2}.$$

By observing that

$$\sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa}\|w\|^2_{2,\kappa} \le \left(\max_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa} \right) \sum_{\kappa\in\mathcal{T}_h} \|w\|^2_{2,\kappa} = \left(\max_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa} \right) \|w\|^2_{2,\Omega}$$

and noting (3.18), we obtain

$$(u-\Pi u, g) \le C \left(\left(\max_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa} \right) \sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\,\|u\|^2_{k_\kappa,\kappa} \right)^{1/2} \|g\|_{0,\Omega},$$

and therefore, for any $g \in {\rm L}^2(\Omega)$, $g \neq 0$,

$$\frac{(u-\Pi u, g)}{\|g\|_{0,\Omega}} \le C \left(\left(\max_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa} \right) \sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\,\|u\|^2_{k_\kappa,\kappa} \right)^{1/2}.$$

Recalling (3.16), taking the supremum over $g \in {\rm L}^2(\Omega)$, $g \neq 0$, and squaring the resulting expression yields (3.14). ∎

Remark 3.9. When $d \in \{2,3\}$, any convex bounded open polyhedral domain $\Omega \subset \mathbb{R}^d$ is ${\rm H}^2$-regular in the sense defined above (see Theorem 3.1.2.1 on page 139 of [19]). A similar result holds for mixed homogeneous Dirichlet–homogeneous Neumann boundary conditions, provided that the internal angles of the polyhedron are sufficiently small (cf. [13, Chapter 8]), with the definition of ${\rm H}^2$-regularity suitably adjusted. The error bound (3.14) still holds then in this, more general, case. For simplicity, though, we have confined ourselves to the case when $\Gamma_{\rm N} = \emptyset$.

3.2. Error analysis of the nonsymmetric version of DGFEM. Let the bilinear form B be as in (2.3). In this section we shall be concerned with the nonsymmetric version of DGFEM corresponding to $\theta = 1$ in (2.3), so we write $B_{\rm NS}(\,\cdot\,,\,\cdot\,)$ in place of $B(\,\cdot\,,\,\cdot\,)$. Our aim is to derive a bound on the ${\rm H}^1$ norm of the error $u - u_{\rm DG}$. Here $u_{\rm DG}$ is the NSIP DGFEM approximation of the analytical solution u. We decompose the error as $u - u_{\rm DG} = \eta + \xi$, where $\eta := u - \Pi u$ and $\xi := \Pi u - u_{\rm DG}$, with

Π denoting the broken elliptic projector defined in (3.12) with $\theta = 1$. We assume for simplicity that the initial value is chosen as $u_0^{\mathrm{DG}} = \Pi u_0$, and thus $\xi(0) = 0$.

As in (**B**), we shall suppose throughout this section that $\sigma_e = C_\sigma \{\!\{p^2\}\!\}_e / h_e$, where $C_\sigma > 0$ is an arbitrary positive constant, and h_e is the diameter of a face $e \in \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{D}}$; when $e \in \mathcal{E}_{\mathrm{D}}$, the contribution from outside Ω in the definition of σ_e is set to 0.

LEMMA 3.10. *Let f satisfy* (**A**) *and assume that $u \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega))$. Suppose further that*

(a) *$p_\kappa \geq 2$ and $u|_\kappa \in \mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))$ with $k_\kappa \geq 3\frac{1}{2}$ on each $\kappa \in \mathcal{T}_h$;*

(b) *the hp-mesh is quasi-uniform in the sense that there exists a positive constant C_0 such that*

$$\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2} \leq C_0 \min_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2}. \tag{3.20}$$

Then, there exists $h_0 \in (0,1]$ and a positive constant C independent of the discretization parameters, such that for all $h \in (0,h_0]$, $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$, and for all $t \in [0,T]$ we have

$$\int_0^t \|(u - u_{\mathrm{DG}})(s)\|_{1,h}^2 \,\mathrm{d}s \leq C \int_0^t \left\{ \|\eta(s)\|_{1,h}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} \mathrm{d}s. \tag{3.21}$$

Proof. Let $t_{\star\star} \in (0,T]$ be such that $u_{\mathrm{DG}}(t) \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$ exists for all $t \in [0, t_{\star\star})$. The existence of such a $t_{\star\star}$ is ensured by Carathéodory's theorem (see [39, Theorems II.4.1 and II.4.5]). Thus, either $t_{\star\star} = T$, or $t_{\star\star} < T$ and $\limsup_{t \to t_{\star\star}} \|u_{\mathrm{DG}}(t)\|_{1,h} = +\infty$. In fact, we shall show below that, for h sufficiently small, $t_{\star\star} = T$.

From the formulation of the *hp*-DGFEM (2.6), for all $v \in S^{\mathbf{p}}(\Omega, \mathcal{T}_h, \mathbf{F})$, we have

$$\int_\Omega u_{\mathrm{DG}}' v \,\mathrm{d}x + B_{\mathrm{NS}}(u_{\mathrm{DG}}, v) = \int_\Omega f(x,t,u_{\mathrm{DG}}) v \,\mathrm{d}x + \ell_{\mathrm{NS}}(v) \tag{3.22}$$

for all $t \in (0, t_{\star\star})$. On the other hand, the broken weak formulation (2.5) implies that

$$\begin{aligned} \int_\Omega (\Pi u') v \,\mathrm{d}x + B_{\mathrm{NS}}(\Pi u, v) &= \int_\Omega f(x,t,u) v \,\mathrm{d}x + \ell_{\mathrm{NS}}(v) \\ &\quad + \int_\Omega (\Pi u' - u') v \,\mathrm{d}x + B_{\mathrm{NS}}(\Pi u - u, v) \end{aligned} \tag{3.23}$$

for all $v \in \mathrm{H}^2(\Omega, \mathcal{T}_h)$ and all $t \in (0,T]$. Upon subtracting (3.22) from (3.23) and taking $v = \xi = \Pi u - u_{\mathrm{DG}}$, we obtain

$$\int_\Omega \xi' \xi \,\mathrm{d}x + B_{\mathrm{NS}}(\xi, \xi) = \int_\Omega \{f(x,t,u) - f(x,t,u_{\mathrm{DG}})\} \xi \,\mathrm{d}x - \int_\Omega \eta' \xi \,\mathrm{d}x - B_{\mathrm{NS}}(\eta, \xi)$$

for all $t \in (0, t_{\star\star})$. By virtue of (3.12) we have $B_{\mathrm{NS}}(\eta, \xi) = 0$. Hence, by noting that $\|\xi\|_{1,h}^2 = B_{\mathrm{NS}}(\xi, \xi)$, we deduce from the above identity that

$$\begin{aligned} \frac{1}{2} \frac{\mathrm{d}}{\mathrm{d}t} \|\xi\|_{0,\Omega}^2 + \|\xi\|_{1,h}^2 &\leq \left| \int_\Omega \{f(x,t,u) - f(x,t,\Pi u)\} \xi \,\mathrm{d}x \right| \\ &\quad + \left| \int_\Omega \{f(x,t,\Pi u) - f(x,t,u_{\mathrm{DG}})\} \xi \,\mathrm{d}x \right| + \left| \int_\Omega \eta' \xi \,\mathrm{d}x \right| \end{aligned} \tag{3.24}$$

for all $t \in (0, t_{\star\star})$. By the Cauchy–Schwarz inequality and Cauchy's inequality, with $\varepsilon_1 > 0$, we have

$$\left| \int_\Omega \eta' \xi \, \mathrm{d}x \right| \leq \|\eta'\|_{0,\Omega} \|\xi\|_{0,\Omega} \leq \frac{\varepsilon_1}{2} \|\eta'\|_{0,\Omega}^2 + \frac{1}{2\varepsilon_1} \|\xi\|_{0,\Omega}^2.$$

By the same argument, with ε_2, $\varepsilon_3 > 0$, we have

$$\left| \int_\Omega \{f(x,t,u) - f(x,t,\Pi u)\} \, \xi \, \mathrm{d}x \right| \leq \frac{\varepsilon_2}{2} \|f(\,\cdot\,,t,u) - f(\,\cdot\,,t,\Pi u)\|_{0,\Omega}^2 + \frac{1}{2\varepsilon_2} \|\xi\|_{0,\Omega}^2,$$

$$\left| \int_\Omega \{f(x,t,\Pi u) - f(x,t,u_{\mathrm{DG}})\} \, \xi \, \mathrm{d}x \right| \leq \frac{\varepsilon_3}{2} \|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2 + \frac{1}{2\varepsilon_3} \|\xi\|_{0,\Omega}^2.$$

Further, by Lemma 3.1, upon absorbing all constants into C and using the broken Sobolev–Poincaré inequality and (3.4), for a.e. $t \in (0,T]$ we have that

$$\begin{aligned}
&\|f(\,\cdot\,,t,u) - f(\,\cdot\,,t,\Pi u)\|_{0,\Omega}^2 \\
&\quad \leq C_f^2 \|\eta\|_{0,2(\gamma+1),\Omega}^2 \left(1 + \|u\|_{0,2(\gamma+1),\Omega}^{\gamma} + \|\Pi u\|_{0,2(\gamma+1),\Omega}^{\gamma}\right)^2 \\
&\quad \leq C \|\eta\|_{0,2(\gamma+1),\Omega}^2 \left(1 + \|u\|_{0,2(\gamma+1),\Omega}^{2\gamma} + \|\Pi u - u\|_{0,2(\gamma+1),\Omega}^{2\gamma}\right) \\
&\quad = C \|\eta\|_{0,2(\gamma+1),\Omega}^2 \left(1 + \|u\|_{0,2(\gamma+1),\Omega}^{2\gamma} + \|\eta\|_{0,2(\gamma+1),\Omega}^{2\gamma}\right) \\
&\quad \leq C \|\eta\|_{0,2(\gamma+1),\Omega} \left(1 + \|u\|_{0,2(\gamma+1),\Omega}^{2\gamma} + \|\eta\|_{1,h}^{2\gamma}\right) \\
&\quad \leq C \|\eta\|_{0,2(\gamma+1),\Omega}^2,
\end{aligned}$$

where the constant $C > 0$ depends only on the domain Ω, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$. For ease of writing, the dependence of the norms of u, η, and Πu on t in the last chain of inequalities has been suppressed.

Applying these bounds on the right-hand side of (3.24) with $\varepsilon_1 = \varepsilon_2 = 1$ (the value of ε_3 will be fixed below) and absorbing all constants into C_1 and $C_2 = C_2(\varepsilon_3)$, we obtain

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \|\xi(t)\|_{0,\Omega}^2 + 2\|\xi(t)\|_{1,h}^2 \leq C_1 (\|\eta(t)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(t)\|_{0,\Omega}^2) + C_2 \|\xi(t)\|_{0,\Omega}^2 \\
+ \varepsilon_3 \|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2
\end{aligned} \tag{3.25}$$

for a.e. $t \in (0, t_{\star\star})$.

To bound $\|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2$, we first note that, by a very similar argument to the one above, we have, for a.e. $t \in (0, t_{\star\star})$,

$$\|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2 \leq C \|\xi(t)\|_{0,2(\gamma+1),\Omega}^2 \left(1 + \|\xi(t)\|_{0,2(\gamma+1),\Omega}^{2\gamma}\right), \tag{3.26}$$

where the constant $C > 0$ depends only on the domain Ω, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$.

For $\mathcal{T}_h$ and the polynomial degree vector $\mathbf{p}$ fixed, let $t_\star = t_\star(\mathcal{T}_h, \mathbf{p}) \in (0, t_{\star\star}]$ be the largest time such that u_{DG} exists for all $t \in [0, t_\star]$ and $\|\xi(t)\|_{1,h}^2 \leq 1$ for all $t \in [0, t_\star]$; the existence of such a $t_\star$ follows from the definition of $t_{\star\star}$, together with the fact that $t \mapsto \|\xi(t)\|_{1,h}^2$ is continuous in the neighborhood of $t = 0$ and $\|\xi(0)\|_{1,h}^2 = 0$. Our aim is to show that $t_\star = T$ for all h, sufficiently small; thereby, we will have also shown that $t_{\star\star} = T$. We have that

$$\|\xi(t)\|_{0,2(\gamma+1),\Omega}^2 \leq \text{Const.}\|\xi(t)\|_{1,h}^2 \qquad \forall t \in [0, t_\star]$$

by the broken Sobolev–Poincaré inequality (see Lemma 3.2) and (3.4); here Const. is a constant that is independent of the discretization parameters and t. This and (3.26), together with the fact that $\|\xi(t)\|_{1,h}^2 \leq 1$ for all $t \in [0, t_\star]$, imply that, for a.e. $t \in (0, t_\star]$,

$$\|f(\,\cdot\,, t, \Pi u) - f(\,\cdot\,, t, u_{\mathrm{DG}})\|_{0,\Omega}^2 \leq \tilde{C}\|\xi(t)\|_{1,h}^2,$$

where the constant $\tilde{C} > 0$ depends only on the domain Ω, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0, t_\star)$.

On choosing $\varepsilon_3 \tilde{C} \leq 1$, after integration from 0 to $t \leq t_\star$ and noting that $\xi(0) = 0$, the inequality (3.25) yields that

$$\begin{aligned}\|\xi(t)\|_{0,\Omega}^2 + \int_0^t \|\xi(s)\|_{1,h}^2 \,\mathrm{d}s \leq & C_1 \int_0^t \left\{\|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2\right\} \mathrm{d}s \\ & + C_2 \int_0^t \|\xi(s)\|_{0,\Omega}^2 \,\mathrm{d}s \qquad \forall t \in [0, t_\star],\end{aligned} \tag{3.27}$$

with the constant $C_1 > 0$ depending only on the domain Ω, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0, T)$.

We can make the first integral on the right-hand side of (3.27) as small as we like (for example, by fixing the local polynomial degree p_κ on each element $\kappa \in \mathcal{T}_h$ and reducing $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$). In particular, let us take $h_0 \in (0, 1]$ so small that, for all $h \leq h_0$ and $t \in [0, T]$, the following inequality holds:

$$C_1 \int_0^t \left\{\|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2\right\} \mathrm{d}s < \frac{1}{1+T}\,\mathrm{e}^{-C_2 T} \times C_{\mathrm{inv}}^{-1} C_0^{-2} \left(\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2}\right)^2,$$

where C_{inv} is the constant from the inverse inequality

$$\|\xi(t)\|_{1,h}^2 \leq C_{\mathrm{inv}} \left(\max_{\kappa \in \mathcal{T}_h} \frac{p_\kappa^2}{h_\kappa}\right)^2 \|\xi(t)\|_{0,\Omega}^2 \qquad \forall t \in [0, t_\star]. \tag{3.28}$$

We note in passing that in order to be able to extract the factor $(\max_{\kappa \in \mathcal{T}_h}(h_\kappa/p_\kappa^2))^2$ above from $\|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2$ with *strict* inequality (by using (3.5), (3.4), and (3.13)), we require hypothesis (a) of the lemma; in particular, the desired strict inequality cannot be achieved when $p_\kappa = 1$, and hence we have our assumption that $p_\kappa \geq 2$ for all $\kappa \in \mathcal{T}_h$.

Thereby, (3.27) yields

$$\|\xi(t)\|_{0,\Omega}^2 + \int_0^t \|\xi(s)\|_{1,h}^2 \,\mathrm{d}s < \frac{\mathrm{e}^{-C_2 T}}{1+T} \times C_{\mathrm{inv}}^{-1} C_0^{-2} \left(\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2}\right)^2 + C_2 \int_0^t \|\xi(s)\|_{0,\Omega}^2 \,\mathrm{d}s$$

for all $t \in [0, t_\star]$. The Gronwall–Bellman inequality then implies that

$$\|\xi(t)\|_{0,\Omega}^2 < C_{\rm inv}^{-1} C_0^{-2} \left(\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2} \right)^2 \qquad \forall t \in [0, t_\star].$$

By the inverse inequality (3.28) we have that,

$$\|\xi(t)\|_{1,h}^2 < C_0^{-2} \left(\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2} \right)^2 \left(\max_{\kappa \in \mathcal{T}_h} \frac{p_\kappa^2}{h_\kappa} \right)^2 = C_0^{-2} \left(\max_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2} \right)^2 \left(\min_{\kappa \in \mathcal{T}_h} \frac{h_\kappa}{p_\kappa^2} \right)^{-2}$$

for all $t \in [0, t_\star]$, which, by the quasi-uniformity hypothesis (b) above, is ≤ 1.

Thus, for $h \leq h_0$, we have that $\|\xi(t)\|_{1,h}^2 < 1$ for all $t \in [0, t_\star]$. By continuity of the mapping $t \mapsto \|\xi(t)\|_{1,h}^2$ on $[0, t_\star]$ it follows that $t_\star = t_{\star\star}$, provided that $h \in (0, h_0]$ (otherwise $t_\star$ would *not* be the largest real number in $(0, t_{\star\star}]$ such that $\|\xi(t)\|_{1,h}^2 \leq 1$ for all $t \in [0, t_\star]$). Now, since $\|\xi(t)\|_{1,h}^2 < 1$ for all $t \in [0, t_{\star\star}]$, and hence $\limsup_{t \to t_{\star\star}} \|\xi(t)\|_{1,h} \leq 1$, it follows by the definition of ξ and the triangle inequality that

$$\limsup_{t \to t_{\star\star}} \|u_{\rm DG}(t)\|_{1,h} \leq 1 + \limsup_{t \to t_{\star\star}} \|\Pi u(t)\|_{1,h} \leq \text{Const.}$$

Therefore $t_{\star\star}$ cannot be strictly smaller than T (if it were, then we would have that $\limsup_{t \to t_{\star\star}} \|u_{\rm DG}\|_{1,h} = +\infty$). To summarize, we have shown that, for $h \leq h_0$, $u_{\rm DG}$ exists on the whole of the interval $[0, T]$ and $\|\xi(t)\|_{1,h} \leq 1$ for all $t \in [0, T]$.

From (3.27), by the Gronwall–Bellman inequality, we then obtain

$$\|\xi(t)\|_{0,\Omega}^2 + \int_0^t \|\xi(s)\|_{1,h}^2 \, {\rm d}s \leq C \int_0^t \left\{ \|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} {\rm d}s \quad \forall t \in [0, T],$$

and hence, in particular,

$$\int_0^t \|\xi(s)\|_{1,h}^2 \, {\rm d}s \leq C \int_0^t \left\{ \|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} {\rm d}s \qquad \forall t \in [0, T],$$

with the constant $C > 0$ depending only on the domain Ω, the quasi-uniformity constant C_0, the final time T, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0, T)$.

Employing the triangle inequality and applying the broken Sobolev–Poincaré inequality and (3.4), we deduce that

$$\int_0^t \|(u - u_{\rm DG})(s)\|_{1,h}^2 \, {\rm d}s \leq C \int_0^t \left\{ \|\eta(s)\|_{1,h}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} {\rm d}s \qquad \forall t \in [0, T]. \qquad \square$$

Lemma 3.10 yields the following error bound for the NSIP DGFEM (2.6).

THEOREM 3.11. *Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a bounded polyhedral domain with Lipschitz-continuous boundary, let $\{\mathcal{T}_h\}$ be a family of shape-regular and hp-quasi-uniform subdivisions of Ω (cf.* (b) *in Lemma* 3.10*), and suppose that $\mathbf{p}$ is a polynomial degree vector of bounded local variation. Let each face $e \in \mathcal{E}_{\rm int} \cup \mathcal{E}_{\rm D}$ be assigned a positive real number*

$$\sigma_e = \{\!\{p^2\}\!\}_e / h_e, \tag{3.29}$$

where h_e is the diameter of e, with the convention that for $e \in \mathcal{E}_{\rm D}$ the contributions from outside Ω in the definition of σ_e are set to 0*. Suppose that f satisfies* (**A**) *and*

$u \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega))$. *Then, if* $p_\kappa \geq 2$ *and* $u|_\kappa \in \mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))$ *with* $k_\kappa \geq 3\frac{1}{2}$ *on each* $\kappa \in \mathcal{T}_h$, *there exists* $h_0 \in (0,1]$ *such that for all* $h \in (0,h_0]$, $h = \max_{\kappa\in\mathcal{T}_h} h_\kappa$, *and all* $t \in [0,T]$, *the solution* $u_{\mathrm{DG}}(\,\cdot\,,t) \in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ *of the NSIP DGFEM* (2.6) *satisfies the following error bound:*

$$\|u-u_{\mathrm{DG}}\|^2_{\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega,\mathcal{T}_h))} \leq C\sum_{\kappa\in\mathcal{T}_h}\frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\|u\|^2_{\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))}, \tag{3.30}$$

with $1 \leq s_\kappa \leq \min\{p_\kappa+1,k_\kappa\}$, $p_\kappa \geq 2$ *on each* $\kappa \in \mathcal{T}_h$, *where* C *is a positive constant depending only on the domain* Ω, *the shape-regularity and quasi-uniformity constants of* $\mathcal{T}_h$, *the final time* T, *the exponent* γ *in the growth condition for the function* f, *the parameter* ρ *in* (2.2), $k = \max_{\kappa\in\mathcal{T}_h} k_\kappa$, *and the Lebesgue and Sobolev norms of* u *over the time interval* $(0,T)$.

Proof. As before, let us choose the projector Π to be the broken elliptic projector defined by (3.12), with $\theta = 1$. From Lemma 3.8 we have the bound

$$\|\eta(s)\|^2_{1,h} \leq C\sum_{\kappa\in\mathcal{T}_h}\frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\|u(s)\|^2_{k_\kappa,\kappa} \qquad \forall s\in[0,T],$$

with $1 \leq s_\kappa \leq \min\{p_\kappa+1,k_\kappa\}$, $p_\kappa \geq 1$, on each $\kappa \in \mathcal{T}_h$.

By differentiating (3.12) with respect to t we deduce that $B(u'-\Pi u',v)=0$ for all $v \in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$. Hence, by applying Lemma 3.8 to u', we obtain that

$$\|\eta'(s)\|^2_{1,h} \leq C\sum_{\kappa\in\mathcal{T}_h}\frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\|u'(s)\|^2_{k_\kappa,\kappa} \qquad \forall s\in[0,T],$$

and with $1 \leq s_\kappa \leq \min\{p_\kappa+1,k_\kappa\}$, $p_\kappa \geq 1$, on each $\kappa \in \mathcal{T}_h$. Therefore, by the broken Sobolev–Poincaré inequality (3.5) and (3.4), an identical bound holds for the norm $\|\eta'(s)\|_{0,\Omega}$, for all $s \in [0,T]$.

Applying these bounds in the right-hand side of (3.21) for $t \in [0,T]$, we obtain the desired bound, with $1 \leq s_\kappa \leq \min\{p_\kappa+1,k_\kappa\}$ and $p_\kappa \geq 2$ on each $\kappa \in \mathcal{T}_h$, where C is a positive constant depending only on the domain Ω, the shape-regularity and quasi-uniformity constants of $\mathcal{T}_h$, the final time T, the exponent γ in the growth condition for the function f, the parameter ρ in (2.2), $k = \max_{\kappa\in\mathcal{T}_h} k_\kappa$, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$. □

When f is *globally* Lipschitz continuous the hypotheses (a) and (b) stated in Lemma 3.10 are redundant, as (3.27) holds automatically for all $t \in [0,T]$, and it is not necessary to separately prove that $\|\xi(t)\|_{1,h} \leq 1$ for all $t \in [0,T]$ and all h sufficiently small. As we shall now see, (a) and (b) also are redundant in the case of the SIP DGFEM.

3.3. Error analysis of the symmetric version of the DGFEM. The symmetric version of the interior penalty DGFEM appeared in the literature much earlier than the nonsymmetric formulation; see Wheeler [16]. It was not widely accepted as an effective numerical method until very recently, due to the additional condition on the minimum size of the penalty parameter which is required to ensure the coercivity of the bilinear form of the method. The renewed interest in the symmetric formulation of the interior penalty DGFEM for second-order elliptic problems can be attributed to the optimality of its convergence rate in the L^2 norm as well as for linear functionals of the solution. Indeed, the nonsymmetric formulation of the interior penalty method

for second-order elliptic problems suffers from a lack of adjoint consistency (see [2]), and, in general, it results in suboptimal a priori error bounds in the L^2 norm and in linear functionals of the solution unless the polynomial degree $p_\kappa = p$, $\kappa \in \mathcal{T}_h$, where p is odd (see [20]). Thanks to its adjoint consistency, the symmetric version of the interior penalty DGFEM does not suffer from these drawbacks.

We state our first result about the accuracy of the symmetric version of the *hp*-DGFEM. We shall assume below that u_{DG} is the SIP DGFEM approximation to the analytical solution u and $u_0^{\mathrm{DG}} = \Pi u_0$, where Π is the broken elliptic projector defined by (3.12) with $\theta = -1$.

THEOREM 3.12. *Let $\Omega \subset \mathbb{R}^d$, $d \geq 2$, be a bounded polyhedral domain with a Lipschitz-continuous boundary. Suppose that $\{\mathcal{T}_h\}$ is a family of shape-regular subdivisions of Ω and $\mathbf{p}$ is a polynomial degree vector of bounded local variation. Let each face $e \in \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{D}}$ be assigned the positive real number*

$$\sigma_e = C_\sigma \{\!\{p^2\}\!\}_e / h_e, \tag{3.31}$$

where h_e is the diameter of e, with the convention that for $e \in \mathcal{E}_{\mathrm{D}}$ the contributions from outside Ω in the definition of σ_e are set to 0, and let C_σ be as in (**B**). *Suppose that the function f satisfies* (**A**). *Then, if $u|_\kappa \in \mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))$, $k_\kappa \geq 2$, $\kappa \in \mathcal{T}_h$, and $u \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega))$, there exists $h_0 \in (0,1]$ such that, for all $h \in (0,h_0]$, $h = \max_{\kappa\in\mathcal{T}_h} h_\kappa$, and $t \in [0,T]$, the solution $u_{\mathrm{DG}}(\cdot,t) \in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ of the SIP DGFEM* (2.6) *satisfies the following error bound:*

$$\operatorname*{ess.\,sup}_{0\leq t\leq T} \|u(t) - u_{\mathrm{DG}}(t)\|_{1,h}^2 \leq C \sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}} \|u\|^2_{\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))}, \tag{3.32}$$

with $1 \leq s_\kappa \leq \min\{p_\kappa+1, k_\kappa\}$, $p_\kappa \geq 1$, for $\kappa \in \mathcal{T}_h$, where C is a positive constant depending only on the domain Ω, the shape-regularity constant of $\mathcal{T}_h$, the final time T, the exponent γ in the growth condition for the function f, the parameter ρ in (2.2), *the Lebesgue and Sobolev norms of u over the interval $(0,T)$, and $k = \max_{\kappa\in\mathcal{T}_h} k_\kappa$.*

Proof. Let $t_{\star\star} \in (0,T]$ be such that $u_{\mathrm{DG}}(t) \in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ exists for all $t \in [0,t_{\star\star})$. Again, the existence of such a $t_{\star\star}$ is ensured by Carathéodory's theorem (see [39, Theorems II.4.1 and II.4.5]). Thus, either $t_{\star\star} = T$, or $t_{\star\star} < T$ and $\limsup_{t\to t_{\star\star}} \|u_{\mathrm{DG}}(t)\|_{1,h} = +\infty$. In fact, we shall show below that, for h sufficiently small, $t_{\star\star} = T$. Let us write $u - u_{\mathrm{DG}} = (u - \Pi u) + (\Pi u - u_{\mathrm{DG}}) =: \eta + \xi$. By the same argument as in the proof of Lemma 3.10, upon subtracting (3.22) from (3.23) and choosing $v = \xi'$, we obtain for a.e. $t \in (0,t_{\star\star})$ that

$$\|\xi'\|_{0,\Omega}^2 + B_{\mathrm{S}}(\xi,\xi') = \int_\Omega \{f(x,t,u) - f(x,t,u_{\mathrm{DG}})\}\,\xi'\,\mathrm{d}x - \int_\Omega \eta'\xi'\,\mathrm{d}x - B_{\mathrm{S}}(\eta,\xi').$$

By virtue of (3.12), $B_{\mathrm{S}}(\eta,\xi') = 0$. With the constant C_σ in (3.31) chosen large enough, the symmetric bilinear form $B_{\mathrm{S}}(\,\cdot\,,\,\cdot\,)$ is coercive and therefore defines an inner product on $\mathrm{H}^1(\Omega,\mathcal{T}_h)$, which induces the norm $\|\,\cdot\,\|_B$ on this space (cf. Lemma 3.7). Hence we deduce that $B_{\mathrm{S}}(\xi,\xi') = \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_B^2$. We thereby infer from the above equality that

$$\begin{aligned}\|\xi'\|_{0,\Omega}^2 + \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_B^2 \leq{}& \left|\int_\Omega \eta'\xi'\,\mathrm{d}x\right| + \left|\int_\Omega \{f(x,t,u) - f(x,t,\Pi u)\}\,\xi'\,\mathrm{d}x\right| \\ &+ \left|\int_\Omega \{f(x,t,\Pi u) - f(x,t,u_{\mathrm{DG}})\}\,\xi'\,\mathrm{d}x\right|\end{aligned} \tag{3.33}$$

for a.e. $t \in (0, t_{\star\star})$. By the Cauchy–Schwarz inequality and Cauchy's inequality, we have

$$\left| \int_\Omega \eta' \xi' \, \mathrm{d}x \right| \leq \|\eta'\|_{0,\Omega} \|\xi'\|_{0,\Omega} \leq \frac{\varepsilon_1}{2} \|\eta'\|_{0,\Omega}^2 + \frac{1}{2\varepsilon_1} \|\xi'\|_{0,\Omega}^2,$$

and, similarly,

$$\left| \int_\Omega \{f(x,t,u) - f(x,t,\Pi u)\} \xi' \, \mathrm{d}x \right| \leq \frac{\varepsilon_2}{2} \|f(\,\cdot\,,t,u) - f(\,\cdot\,,t,\Pi u)\|_{0,\Omega}^2 + \frac{1}{2\varepsilon_2} \|\xi'\|_{0,\Omega}^2,$$

$$\left| \int_\Omega \{f(x,t,\Pi u) - f(x,t,u_{\mathrm{DG}})\} \xi' \, \mathrm{d}x \right| \leq \frac{\varepsilon_3}{2} \|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2 + \frac{1}{2\varepsilon_3} \|\xi'\|_{0,\Omega}^2,$$

with $\varepsilon_1, \varepsilon_2, \varepsilon_3 > 0$. Also, by the same argument as in the proof of Lemma 3.10, we have, for a.e. $t \in [0,T]$, that

$$\|f(\,\cdot\,,t,u) - f(\,\cdot\,,t,\Pi u)\|_{0,\Omega}^2 \leq C \|\eta(t)\|_{0,2(\gamma+1),\Omega}^2,$$

where the constant $C > 0$ depends only on the domain Ω, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$. Choosing $\varepsilon_1, \varepsilon_2, \varepsilon_3$ such that $\varepsilon_1^{-1} + \varepsilon_2^{-1} + \varepsilon_3^{-1} \leq 2$, and inserting the above bounds into (3.33), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t} \|\xi\|_B^2 \leq C_1 \left(\|\eta\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'\|_{0,\Omega}^2 \right) + \tilde{C}_2 \|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2 \tag{3.34}$$

for all $t \in (0, t_{\star\star})$. To bound $\|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2$ we note that, by the same argument as in (3.26) above, for a.e. $t \in (0, t_{\star\star})$, we have

$$\|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2 \leq C \|\xi(t)\|_{0,2(\gamma+1),\Omega}^2 \left(1 + \|\xi(t)\|_{0,2(\gamma+1),\Omega}^{2\gamma} \right),$$

where the constant $C > 0$ depends only on the domain Ω, the exponent γ in the growth condition for the function f, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$.

For $\mathcal{T}_h$ and the polynomial degree vector $\mathbf{p}$ fixed, let $t_\star = t_\star(\mathcal{T}_h, \mathbf{p})$ denote the largest time $t \in (0, t_{\star\star}]$ such that $u_{\mathrm{DG}}(t)$ exists for all $t \in [0, t_\star]$ and $\|\xi(t)\|_{1,h} \leq 1$ for all $t \in [0, t_\star]$; the existence of such a $t_\star$ is guaranteed by the definition of $t_{\star\star}$, together with the fact that $t \mapsto \|\xi(t)\|_{1,h}$ is continuous in the neighborhood of $t = 0$ and $\|\xi(0)\|_{1,h} = 0$. By the broken Sobolev–Poincaré inequality (3.5) and (3.4), for a.e. $t \in (0, t_\star]$, we have that

$$\|f(\,\cdot\,,t,\Pi u) - f(\,\cdot\,,t,u_{\mathrm{DG}})\|_{0,\Omega}^2 \leq C \|\xi(t)\|_{1,h}^2.$$

Inserting this bound into (3.34), integrating from 0 to $t \leq t_\star$, using Lemma 3.7 to deduce that $c_0 \|\xi(t)\|_{1,h}^2 \leq \|\xi(t)\|_B^2$, and noting that $\xi(0) = 0$, we deduce that

$$c_0 \|\xi(t)\|_{1,h}^2 \leq C_1 \int_0^t \left\{ \|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} \mathrm{d}s + C_2 \int_0^t \|\xi(s)\|_{1,h}^2 \, \mathrm{d}s \tag{3.35}$$

for all $t \in [0, t_\star]$. By Lemma 3.8, the first integral on the right-hand side can be bounded in terms of h_κ and p_κ. We define $h = \max_{\kappa \in \mathcal{T}_h} h_\kappa$, $C_3 = C_2/c_0$, and let $h_0 \in (0,1]$ be small enough so that for all $h \leq h_0$ and $t \in [0, t_\star]$ we have

$$C_1 \int_0^t \left\{ \|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} \mathrm{d}s < \frac{c_0}{1+T} \, \mathrm{e}^{-C_3 T}.$$

Thus, for $h \le h_0$ and all $t \in [0, t_\star]$, from (3.35) we have that

$$\|\xi(t)\|_{1,h}^2 < \frac{1}{1+T}\,\mathrm{e}^{-C_3 T} + C_3 \int_0^t \|\xi(s)\|_{1,h}^2 \,\mathrm{d}s;$$

using the Gronwall–Bellman inequality, we deduce that $\|\xi(t)\|_{1,h}^2 < 1$ for all $t \in [0, t_\star]$ with $h \le h_0$. Therefore, by the same continuity argument as in the proof of Lemma 3.10 applied to the mapping $t \mapsto \|\xi(t)\|_{1,h}^2$, we deduce that $t_\star = t_{\star\star} = T$ for all $h \in (0, h_0]$. Taking this into account, assuming that $h \le h_0$, and applying the Gronwall–Bellman inequality to (3.35) gives us the following bound:

$$\|\xi(t)\|_{1,h}^2 \le C \int_0^t \left\{ \|\eta(s)\|_{0,2(\gamma+1),\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} \mathrm{d}s \qquad \forall t \in [0,T], \tag{3.36}$$

where the constant $C > 0$ depends only on the domain Ω, the exponent γ in the growth condition for the function f, the final time T, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$.

Further, by the broken Sobolev–Poincaré inequality (3.5) and (3.4), we have that $\|\eta\|_{0,2(\gamma+1),\Omega}^2 \le C\|\eta\|_{1,h}^2$; employing the triangle inequality, we thus obtain

$$\|(u - u_{\mathrm{DG}})(t)\|_{1,h}^2 \le C \left(\|\eta(t)\|_{1,h}^2 + \int_0^t \left\{ \|\eta(s)\|_{1,h}^2 + \|\eta'(s)\|_{0,\Omega}^2 \right\} \mathrm{d}s \right) \qquad \forall t \in [0,T].$$

Arguing in the same way as in the proof of Theorem 3.11 to bound $\|\eta'(s)\|_{0,\Omega}^2$ and noting that the embedding $\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa)) \hookrightarrow \mathrm{L}^\infty(0,T;\mathrm{H}^{k_\kappa}(\kappa))$ yields (3.32), with $1 \le s_\kappa \le \min\{p_\kappa + 1, k_\kappa\}$, $p_\kappa \ge 1$, for $\kappa \in \mathcal{T}_h$, where the constant $C > 0$ depends only on the domain Ω, the shape-regularity constant of $\mathcal{T}_h$, the final time T, the parameter ρ in (2.2), the exponent γ in the growth condition for the function f, $k = \max_{\kappa\in\mathcal{T}_h} k_\kappa$, and the Lebesgue and Sobolev norms of u over the time interval $(0,T)$. ∎

Let us now prove an error bound in the L^2 norm for the SIP DGFEM.

THEOREM 3.13. *Let $\Omega \subset \mathbb{R}^d$, $d \ge 2$, be an H^2-regular polyhedral domain. Suppose that Γ_{N} is empty, $\{\mathcal{T}_h\}$ is a family of shape-regular subdivisions of Ω, and $\mathbf{p}$ is a polynomial degree vector of bounded local variation. Let each face $e \in \mathcal{E}_{\mathrm{int}} \cup \mathcal{E}_{\mathrm{D}}$ be assigned the positive real number*

$$\sigma_e = C_\sigma \{\!\{p^2\}\!\}_e / h_e,$$

*where h_e is the diameter of e, with the convention that for $e \in \mathcal{E}_{\mathrm{D}}$ the contributions from outside Ω in the definition of σ_e are set to 0, and C_σ is as in (**B**). Suppose that the function f satisfies (**A**). Then, if $u|_\kappa \in \mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))$, $k_\kappa \ge 2$, $\kappa \in \mathcal{T}_h$ and $u \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega) \cap \mathrm{C}(\bar\Omega))$, there exists $h_0 \in (0,1]$ such that for all $h \in (0,h_0]$, $h = \max_{\kappa\in\mathcal{T}_h} h_\kappa$, and $t \in (0,T]$, the solution $u_{\mathrm{DG}}(\,\cdot\,,t) \in S^{\mathbf{p}}(\Omega,\mathcal{T}_h,\mathbf{F})$ of the SIP DGFEM (2.6) satisfies the following error bound:*

$$\|u - u_{\mathrm{DG}}\|_{\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))}^2 \le C \left(\max_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^2}{p_\kappa} \right) \sum_{\kappa\in\mathcal{T}_h} \frac{h_\kappa^{2s_\kappa - 2}}{p_\kappa^{2k_\kappa - 3}} \|u\|_{\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))}^2, \tag{3.37}$$

with $1 \le s_\kappa \le \min\{p_\kappa + 1, k_\kappa\}$, $p_\kappa \ge 1$, for $\kappa \in \mathcal{T}_h$, where C is a positive constant depending only on the domain Ω, the shape-regularity constant of $\mathcal{T}_h$, the final time T, the exponent γ in the growth condition for the function f, the parameter ρ in (2.2), the Lebesgue and Sobolev norms of u over the time interval $(0,T)$, and $k = \max_{\kappa\in\mathcal{T}_h} k_\kappa$.

Proof. By the same argument as in the proof of Lemma 3.10, and with ξ and η defined as in the proof of Theorem 3.12, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_{0,\Omega}^2 + \|\xi\|_B^2 \leq \left|\int_\Omega \{f(x,t,u) - f(x,t,\Pi u)\}\,\xi\,\mathrm{d}x\right| + \left|\int_\Omega \{f(x,t,\Pi u) - f(x,t,u_{\mathrm{DG}})\}\,\xi\,\mathrm{d}x\right| + \left|\int_\Omega \eta'\xi\,\mathrm{d}x\right|$$

for a.e. $t \in [0,T]$. Applying (3.6) and the Cauchy–Schwarz inequality to the right-hand side of the above inequality gives, for a.e. $t \in [0,T]$,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_{0,\Omega}^2 + \|\xi\|_B^2 \leq C\|\eta\|_{0,\Omega}(1 + |||u|||_{1,h}^\gamma + |||\Pi u|||_{1,h}^\gamma)\|\xi\|_{1,h} + C\|\xi\|_{0,\Omega}(1 + |||\Pi u|||_{1,h}^\gamma + |||u_{\mathrm{DG}}|||_{1,h}^\gamma)\|\xi\|_{1,h} + \|\eta'\|_{0,\Omega}\|\xi\|_{0,\Omega}.$$

Let us show that $\operatorname{ess.sup}_{0\leq t\leq T} |||\Pi u(t)|||_{1,h}$ and $\operatorname{ess.sup}_{0\leq t\leq T} |||u_{\mathrm{DG}}(t)|||_{1,h}$ are bounded uniformly with respect to $h \in (0,h_0]$. We have that

$$|||\Pi u(t)|||_{1,h} \leq |||(\Pi u - u)(t)|||_{1,h} + |||u(t)|||_{1,h} = |||\eta(t)|||_{1,h} + |||u(t)|||_{1,h} \leq \|\eta(t)\|_{1,h} + |||u(t)|||_{1,h} \leq \text{Const.} \qquad \forall t \in [0,T],$$

where Const. is a positive constant, independent of the discretization parameters and of $t \in [0,T]$. Here, the last inequality follows from (3.13) on observing that $\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa)) \hookrightarrow \mathrm{L}^\infty(0,T;\mathrm{H}^{k_\kappa}(\kappa))$, recalling the definition of the norm $|||\cdot|||_{1,h}$, and noting that $u \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega)\cap \mathrm{C}(\bar\Omega))$.

By the above and the fact that $\operatorname{ess.sup}_{0\leq t\leq T}\|\xi(t)\|_{1,h}^2 \leq 1$ uniformly in $h \leq h_0$ (see the proof of Theorem 3.12), we have that

$$|||u_{\mathrm{DG}}(t)|||_{1,h} \leq |||(u_{\mathrm{DG}} - \Pi u)(t)|||_{1,h} + |||\Pi u(t)|||_{1,h} = |||\xi(t)|||_{1,h} + |||\Pi u(t)|||_{1,h} \leq \|\xi(t)\|_{1,h} + |||\Pi u(t)|||_{1,h} \leq \|\xi(t)\|_{1,h} + \|\eta(t)\|_{1,h} + |||u(t)|||_{1,h} \leq \text{Const.}$$

for all $t \in [0,T]$ and uniformly in $h \in (0,h_0]$—again by (3.13) on observing that $\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa)) \hookrightarrow \mathrm{L}^\infty(0,T;\mathrm{H}^{k_\kappa}(\kappa))$, the definition of the norm $|||\cdot|||_{1,h}$ and the fact that $u \in \mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega)\cap\mathrm{C}(\bar\Omega))$; once again, Const. denotes a positive constant, independent of the discretization parameters and of $t \in [0,T]$. Hence we deduce that

$$\operatorname*{ess.sup}_{0\leq t\leq T}\left(1 + |||u(t)|||_{1,h}^\gamma + |||\Pi u(t)|||_{1,h}^\gamma\right) \leq \text{Const.},$$

$$\operatorname*{ess.sup}_{0\leq t\leq T}\left(1 + |||\Pi u(t)|||_{1,h}^\gamma + |||u_{\mathrm{DG}}(t)|||_{1,h}^\gamma\right) \leq \text{Const.},$$

uniformly in $h \in (0,h_0]$. Therefore, by Cauchy's inequality,

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_{0,\Omega}^2 + \|\xi\|_{1,h}^2 \leq C\left(\|\eta\|_{0,\Omega}\|\xi\|_{1,h} + \|\xi\|_{0,\Omega}\|\xi\|_{1,h} + \|\eta'\|_{0,\Omega}\|\xi\|_{0,\Omega}\right) \leq \frac{1}{2}\|\xi\|_{1,h}^2 + \frac{1}{2}C^2\left(\|\eta\|_{0,\Omega}^2 + \|\eta'\|_{0,\Omega}^2 + \|\xi\|_{0,\Omega}^2\right),$$

which yields

$$\frac{\mathrm{d}}{\mathrm{d}t}\|\xi\|_{0,\Omega}^2 + \|\xi\|_{1,h}^2 \le C^2\left(\|\eta\|_{0,\Omega}^2 + \|\eta'\|_{0,\Omega}^2 + \|\xi\|_{0,\Omega}^2\right).$$

Upon integrating from 0 to $t \in (0,T]$ and applying the Gronwall–Bellman inequality, we have

$$\|\xi(t)\|_{0,\Omega}^2 + \int_0^t \|\xi(s)\|_{1,h}^2\,\mathrm{d}s \le C\int_0^T \left\{\|\eta(s)\|_{0,\Omega}^2 + \|\eta'(s)\|_{0,\Omega}^2\right\}\,\mathrm{d}s. \tag{3.38}$$

Applying (3.14) to the right-hand side of (3.38), we deduce that

$$\|\xi(t)\|_{0,\Omega}^2 \le C\left(\max_{\kappa\in\mathcal{T}_h}\frac{h_k^2}{p_\kappa}\right)\sum_{\kappa\in\mathcal{T}_h}\frac{h_\kappa^{2s_\kappa-2}}{p_\kappa^{2k_\kappa-3}}\|u\|_{\mathrm{H}^1(0,T;\mathrm{H}^{k_\kappa}(\kappa))}^2 \qquad \forall t\in[0,T].$$

Employing the triangle inequality $\|u(t)-u_{\mathrm{DG}}(t)\|_{0,\Omega} \le \|\eta(t)\|_{0,\Omega} + \|\xi(t)\|_{0,\Omega}$ and applying (3.14) to $\|\eta(t)\|_{0,\Omega}$ once again, we obtain (3.37). ∎

Remark 3.14. Suppose, for example, that $u \in \mathrm{H}^1(0,T;\mathrm{H}^k(\Omega))$, $k \ge 2$, and that $p_\kappa = p$ for all $\kappa \in \mathcal{T}_h$, and let $h = \max_{\kappa\in\mathcal{T}_h} h_\kappa$. Then, (3.32) and (3.37) take the form

$$\frac{h}{p^{3/2}}\|u-u_{\mathrm{DG}}\|_{\mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega,\mathcal{T}_h))} + \frac{1}{p}\|u-u_{\mathrm{DG}}\|_{\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))} \le C\frac{h^s}{p^k}\,\|u\|_{\mathrm{H}^1(0,T;\mathrm{H}^k(\Omega))},$$

where the constant $C>0$ is as above, $1 \le s \le \min\{p+1,k\}$, and $p \ge 1$. Hence the error bounds (3.32) and (3.37) are fully optimal in h; the error bound (3.32) in the broken H^1 norm is suboptimal in p by half a power of p, while the error bound (3.37) in the L^2 norm is suboptimal in p by a single power of the polynomial degree p.

4. Conclusions. We have been concerned with the error analysis of the spatial discretization of semilinear parabolic initial boundary value problems with mixed Dirichlet and Neumann boundary conditions by interior penalty *hp*-DGFEMs. We developed techniques for handling locally Lipschitz-continuous nonlinearities in the error analysis, which allowed us to perform our proofs on the entire time interval of existence of the solution. We showed that the presence of a locally Lipschitz nonlinearity, satisfying a certain growth condition, does not degrade the convergence rates observed in the case of a linear parabolic PDE. The resulting error bounds are optimal in h and slightly suboptimal in p. As we have noted in the introduction, full *hp*-optimality of the error bounds can be restored by hypothesizing piecewise regularity of the solution in augmented Sobolev spaces instead of classical Sobolev spaces, as was done in [17] in the case of linear elliptic equations. To the best of our knowledge, the error bounds derived in the present paper are the first of this kind for semilinear parabolic equations with locally Lipschitz-continuous nonlinearity. The extension of the analysis of our semidiscrete scheme to simple fully discrete schemes, using DGFEM time discretization, say, would proceed along very similar lines and is, therefore, not considered here (see [29], which also includes numerical experiments).

Our error bound for the nonsymmetric version of the method was established in the broken $\mathrm{L}^2(0,T;\mathrm{H}^1(\Omega))$ norm, while for the symmetric version of the method we derived our error bounds in the broken $\mathrm{L}^\infty(0,T;\mathrm{L}^2(\Omega))$ and $\mathrm{L}^\infty(0,T;\mathrm{H}^1(\Omega))$ norms. As we have noted above, all of these bounds are optimal with respect to h. Due to the fact that the bilinear form featured in the nonsymmetric version of the method is *not* adjoint-consistent, one cannot expect to observe a fully optimal bound for NSIP

DGFEM in the $L^\infty(0,T;L^2(\Omega))$ norm with respect to h—at least not for all p; in fact, it is well documented in the literature that, already in the case of Poisson's equation with a homogeneous Dirichlet boundary condition, the (elliptic) NSIP DGFEM is optimally convergent with respect to h only when $p \geq 1$ is an odd integer (see, for example, [20]). Yet, one may nevertheless wonder whether it is possible to derive, instead of the broken $L^2([0,T],H^1(\Omega))$ norm, an optimal error bound for NSIP DGFEM in the broken $L^\infty([0,T],H^1(\Omega))$ norm, as has been done for SIP DGFEM. This is an open problem: the main technical difficulty is that the bilinear form of NSIP DGFEM is nonsymmetric (as well as adjoint-inconsistent), so the usual testing procedure for convergence analysis in the $L^\infty([0,T],H^1(\Omega))$ norm for second-order parabolic equations of the form $\xi' + A\xi = g$, based on taking the $L^2(\Omega)$ inner product of the equation with ξ', fails to deliver a helpful energy estimate.

Finally, we note that by analogous arguments to those presented above all of our results can be extended to the case of a general locally Lipschitz-continuous nonlinearity which, instead of inequality (1.2), satisfies

$$|f(x,t,w) - f(x,t,v)| \leq C(|w|,|v|)\,|w-v| \quad \begin{cases} \forall w,\, v \in \mathbb{R}, \\ \text{a.e. } (x,t) \in \Omega \times (0,T], \end{cases} \tag{4.1}$$

where $C(\cdot,\cdot)$ is a continuous function on $[0,\infty)^2$. Since, this time, no growth condition of the kind $C(|w|,|v|) \leq G_f(1+|w|+|v|)^\gamma$ is assumed, one cannot rely on the broken Sobolev–Poincaré inequality. In fact, the only way to control terms such as $C(\|u(t)\|_{\infty,\Omega}, \|\Pi u(t)\|_{\infty,\Omega})$ and $C(\|u_{\mathrm{DG}}(t)\|_{\infty,\Omega}, \|\Pi u(t)\|_{\infty,\Omega})$, which will arise in the error analysis, is to show that $\max_{t\in[0,T]} \|\Pi u(t)\|_{\infty,\Omega}$ and $\max_{t\in[0,T]} \|u_{\mathrm{DG}}(t)\|_{\infty,\Omega}$ can be bounded, independent of $\mathbf{p}$ and h. For the first of these, we first note that

$$\|\Pi u(t)\|_{\infty,\Omega} \leq \|u(t)\|_{\infty,\Omega} + \|u(t) - \Pi u(t)\|_{\infty,\Omega}.$$

We then show the smallness of the second term by using the smallness of the projection error $\|u(t) - \Pi u(t)\|_{1,h}$, the smallness of $\|u(t) - z_{p_\kappa}^{h_\kappa}(u(t))\|_{\infty,\kappa}$, and an inverse inequality relating $\|\cdot\|_{\infty,\Omega}$ to $\|\cdot\|_{1,h}$. To bound $\|u_{\mathrm{DG}}(t)\|_{\infty,\Omega}$, we note that

$$\|u_{\mathrm{DG}}(t)\|_{\infty,\Omega} \leq \|\xi(t)\|_{\infty,\Omega} + \|\Pi u(t)\|_{\infty,\Omega}$$

and use an inverse inequality to relate the $\|\cdot\|_{\infty,\Omega}$ norm of ξ to its $\|\cdot\|_{1,h}$ norm. In order to accommodate the use of the inverse inequality, one has then to assume in the analysis of *both* SIP DGFEM and NSIP DGFEM that the mesh is quasi-uniform in the sense of (3.20). For our analysis of SIP DGFEM under hypothesis (1.2), this strong mesh-regularity assumption was not required. Thus, and for reasons of brevity, we chose to base this paper on (1.2) rather than on the more general local Lipschitz condition (4.1) which, at the expense of more restrictive hypotheses on the mesh, makes no assumption on the growth rate of $C(|w|,|v|)$ as $|w|,|v| \to \infty$.

Acknowledgments. We are grateful to the Isaac Newton Institute for Mathematical Sciences in Cambridge, UK for its generous support.

REFERENCES

[1] D. N. Arnold, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. Donatella Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] I. Babuška and M. Suri, *The h-p version of the finite element method with quasi-uniform meshes*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 199–238.

[4] D. Braess, *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics* 2nd ed. (translated from the 1992 German edition by Larry L. Schumaker), Cambridge University Press, Cambridge, UK, 2001.

[5] S. C. Brenner, *Poincaré–Friedrichs inequalities for piecewise H^1 functions*, SIAM J. Numer. Anal., 41 (2003), pp. 306–324.

[6] S. C. Brenner and L. Ridgway Scott, *The Mathematical Theory of Finite Element Methods*, 2nd ed., Texts Appl. Math. 15, Springer-Verlag, New York, 2002.

[7] B. Cockburn, *Devising discontinuous Galerkin methods for non-linear hyperbolic conservation laws*, J. Comput. Appl. Math., 128 (2001), pp. 187–204.

[8] B. Cockburn, P.-A. Gremaud, and J. X. Yang, *A priori error estimates for nonlinear scalar conservation laws*, in Hyperbolic Problems: Theory, Numerics, Applications, Vol. I (Zürich, 1998), Internat. Ser. Numer. Math. 129, Birkhäuser, Basel, 1999, pp. 167–176.

[9] B. Cockburn, G. E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lecture Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 3–50.

[10] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws*. II. *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[11] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection P^1-discontinuous-Galerkin finite element method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.

[12] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws*. V. *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[13] M. Dauge, *Elliptic Boundary Value Problems on Corner Domains. Smoothness and Asymptotics of Solutions*, Lecture Notes in Math. 1341, Springer-Verlag, Berlin, 1988.

[14] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*, Grundlehren Math. Wiss. (Fundamental Principles of Mathematical Sciences), 303 Springer-Verlag, Berlin, 1993.

[15] R. S. Falk and G. R. Richter, *Local error estimates for a finite element method for hyperbolic and convection-diffusion equations*, SIAM J. Numer. Anal., 29 (1992), pp. 730–754.

[16] M. F. Wheeler, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

[17] E. H. Georgoulis and E. Süli, *Optimal error estimates for the hp-version interior penalty discontinuous Galerkin finite element method*, IMA J. Numer. Anal., 25 (2005), pp. 214–240.

[18] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Classics Math., Springer-Verlag, Berlin, 2001; reprint of the 1998 edition.

[19] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Monogr. Stud. Math. 24, Pitman (Advanced Publishing Program), Boston, MA, 1985.

[20] K. Harriman, P. Houston, B. Senior, and E. Süli, *hp-version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form*, in Recent Advances in Scientific Computing and Partial Differential Equations (Hong Kong, 2002), Contemp. Math. 330, Amer. Math. Soc., Providence, RI, 2003, pp. 89–119.

[21] R. Hartmann and P. Houston, *Adaptive discontinuous Galerkin finite element methods for nonlinear hyperbolic conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 979–1004.

[22] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2002), pp. 2133–2163.

[23] P. Houston and E. Süli, *hp-adaptive discontinuous Galerkin finite element methods for first-order hyperbolic problems*, SIAM J. Sci. Comput., 23 (2001), pp. 1226–1252.

[24] C. Hu and C.-W. Shu, *A discontinuous Galerkin finite element method for Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (1999), pp. 666–690.

[25] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[26] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp., 46 (1986), pp. 1–26.

[27] O. Karakashian and C. Makridakis, *A space-time finite element method for the nonlinear Schrödinger equation: The discontinuous Galerkin method*, Math. Comp., 67 (1998), pp. 479–499.

[28] P. Lasaint and P.-A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations (Proc. Sympos., Math. Res. Center, Univ. Wisconsin, Madison, WI, 1974), Publication no. 33, Math. Res. Center, Univ. of Wisconsin-Madison, Academic Press, New York, 1974, pp. 89–123.

[29] A. Lasis, *Discontinuous Galerkin Finite Element Approximation of Nonlinear Parabolic Problems*, D.Phil. Thesis, University of Oxford, Oxford, UK, 2005.

[30] A. Lasis and E. Süli, *Poincaré–Type Inequalities for Broken Sobolev Spaces*, Tech. report NA-03/10, Oxford University Computing Laboratory, Oxford, 2003; available online at http://web.comlab.ox.ac.uk/oucl/publications/natr/na-03-10.html.

[31] J. M. Melenk, *HP-Interpolation of Non-Smooth Functions*, Preprint NI03050-CPD, Isaac Newton Institute, Cambridge, UK, 2003; available online at http://www.newton.cam.ac.uk/preprints/NI03050.pdf.

[32] J. Nitsche, *Über ein Variationsprinzip zur Lösung von Dirichlet-Problemen bei Verwendung von Teilräumen, die keinen Randbedingungen unterworfen sind*, in collection of articles dedicated to Lothar Collatz on his sixtieth birthday, Abh. Math. Sem. Univ. Hamburg, 36 (1971), pp. 9–15.

[33] W. H. Reed and T. R. Hill, *Triangular Mesh Methods for the Neutron Transport Equation*, Tech. report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.

[34] G. R. Richter, *An optimal-order error estimate for the discontinuous Galerkin method*, Math. Comp., 50 (1988), pp. 75–88.

[35] G. R. Richter, *The discontinuous Galerkin method with diffusion*, Math. Comp., 58 (1992), pp. 631–643.

[36] B. Rivière and M. F. Wheeler, *A discontinuous Galerkin method applied to nonlinear parabolic equations*, in Discontinuous Galerkin Methods (Newport, RI, 1999), Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 231–244.

[37] C. Schwab, *p- and hp-Finite Element Methods. Theory and Applications in Solid and Fluid Mechanics*. Numer. Math. Sci. Comput. The Clarendon Press Oxford University Press, New York, 1998.

[38] V. Thomée and L. Wahlbin, *On Galerkin methods in semilinear parabolic problems*, SIAM J. Numer. Anal., 12 (1975), pp. 378–389.

[39] J. Warga, *Optimal Control of Differential and Functional Equations*, Academic Press, New York, 1972.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1570–1599

A POSTERIORI ERROR ESTIMATES FOR LOWEST-ORDER MIXED FINITE ELEMENT DISCRETIZATIONS OF CONVECTION-DIFFUSION-REACTION EQUATIONS*

MARTIN VOHRALÍK†

Abstract. We establish residual a posteriori error estimates for lowest-order Raviart–Thomas mixed finite element discretizations of convection-diffusion-reaction equations on simplicial meshes in two or three space dimensions. The upwind-mixed scheme is considered as well, and the emphasis is put on the presence of an inhomogeneous and anisotropic diffusion-dispersion tensor and on a possible convection dominance. Global upper bounds for the approximation error in the energy norm are derived, where in particular all constants are evaluated explicitly, so that the estimators are fully computable. Our estimators give local lower bounds for the error as well, and they hold from the cases where convection or reaction are not present to convection- or reaction-dominated problems; we prove that their local efficiency depends only on local variations in the coefficients and on the local Péclet number. Moreover, the developed general framework allows for asymptotic exactness and full robustness with respect to inhomogeneities and anisotropies. The main idea of the proof is a construction of a locally postprocessed approximate solution using the mean value and the flux in each element, known in the mixed finite element method, and a subsequent use of the abstract framework arising from the primal weak formulation of the continuous problem. Numerical experiments confirm the guaranteed upper bound and excellent efficiency and robustness of the derived estimators.

Key words. convection-diffusion-reaction equation, inhomogeneous and anisotropic diffusion, convection dominance, mixed finite element method, upwind weighting, a posteriori error estimates

AMS subject classifications. 65N15, 65N30, 76S05

DOI. 10.1137/060653184

1. Introduction.

We consider the convection-diffusion-reaction problem

$$-\nabla \cdot (\mathbf{S}\nabla p) + \nabla \cdot (p\mathbf{w}) + rp = f \quad \text{in } \Omega, \tag{1.1a}$$

$$p = 0 \quad \text{on } \partial\Omega, \tag{1.1b}$$

where $\mathbf{S}$ is in general an inhomogeneous and anisotropic (nonconstant full-matrix) diffusion-dispersion tensor, $\mathbf{w}$ is a (dominating) velocity field, r a reaction function, f a source term, and $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a polygonal (polyhedral) domain (open, bounded, and connected set). Our purpose is to derive a posteriori error estimates for the lowest-order Raviart–Thomas mixed finite element discretization of the problem (1.1a)–(1.1b) on simplicial meshes (consisting of triangles if $d = 2$ and of tetrahedra if $d = 3$), as well as for its upwind variant; cf. Douglas and Roberts [17] and Dawson [16].

*Received by the editors February 28, 2006; accepted for publication (in revised form) January 31, 2007; published electronically August 10, 2007. This work was partially supported by the GdR MoMaS project "Numerical Simulations and Mathematical Modeling of Underground Nuclear Waste Disposal," CNRS-2439, ANDRA, BRGM, CEA, EdF, France, and by the Ministry of Education of the Czech Republic, Research Centre "Advanced Remediation Technologies and Processes," no. 1M0554. The main part of this work was carried out during the author's post-doc stay at Laboratoire de Mathématiques, Analyse Numérique et EDP, Université de Paris-Sud and CNRS, Orsay, France, and Department of Process Modeling, Faculty of Mechatronics and Interdisciplinary Engineering Studies, Technical University of Liberec, Czech Republic.

http://www.siam.org/journals/sinum/45-4/65318.html

†Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie (Paris 6), B.C. 187, 4 place Jussieu, 75252 Paris, France (vohralik@ann.jussieu.fr).

A posteriori error estimates, pioneered by Babuška and Rheinboldt [7], are nowadays well established for primal discretizations of second-order elliptic problems involving only a diffusion term; cf., for example, the survey by Verfürth [32] for the conforming finite element method. An approach encompassing all conforming, nonconforming, and discontinuous finite element methods was recently proposed by Ainsworth [3], using a Helmholtz-like decomposition of the error in the numerical solution into its conforming and nonconforming parts in order to give a computable error bound. In most cases the analysis is given only for $\mathbf{S}$ being an identity matrix; an in-depth analysis for the general inhomogeneous and anisotropic diffusion tensor in the framework of the finite element method was presented by Bernardi and Verfürth [9]. Similar results have been obtained by Petzoldt [28], for nonconforming finite elements by Ainsworth [4], and some developments for the finite volume box scheme (in the given case actually equivalent to the lowest-order Raviart–Thomas mixed finite element method) are presented by El Alaoui and Ern [19]. In all these references, a hypothesis of the type "monotonicity around vertices" on the distribution of the inhomogeneities is necessary. In recent years a posteriori error estimates have been extended to convection-diffusion problems as well. We cite in particular Verfürth [33], who derived estimates in the energy norm for the conforming Galerkin method and its stabilized SUPG (streamline upwind Petrov–Galerkin) version. His estimates are both reliable (yielding a global upper bound on the error between the exact and approximate solutions) and locally efficient (giving a local lower bound). Moreover, they are semirobust in the sense that the lower and upper bounds differ by constants whose dependence on the local mesh discretization parameter vanishes as this approaches the ratio of the smallest eigenvalue of $\mathbf{S}$ to the local size of the velocity field (i.e., when the local Péclet number gets sufficiently small). Recently, Verfürth [34] improved his results while giving estimates which are fully robust with respect to convection dominance in a norm incorporating a dual norm of the convective derivative. The new norm is not, however, easily computable, there is no local lower bound, and the estimators do not change with respect to [33], and hence the adaptive strategies will remain the same. Finally, a different approach, yielding an estimate in the L^1-norm, independent of the size of the diffusion tensor, is given by Ohlberger [26] in the framework of the vertex-centered finite volume method.

In comparison with primal methods, the literature on a posteriori error estimates in the mixed finite element method is much less extensive. Most of the results have been obtained for the Poisson equation (i.e., $\mathbf{w} = r = 0$ in (1.1a)–(1.1b)) in two space dimensions: Alonso [5] derived estimates for the error in the flux $\mathbf{u} := -\mathbf{S}\nabla p$ of the scalar variable p and either Raviart–Thomas [29] or Brezzi–Douglas–Marini [11] mixed finite elements. Braess and Verfürth [10] proved estimates for both $\mathbf{u}$ and p for Raviart–Thomas elements, based on mesh-dependent norms and a saturation assumption. Carstensen [13] derived rigorous estimates for various mixed finite element schemes and for both $\mathbf{u}$ and p. Achchab et al. [1] can imbed Raviart–Thomas elements in their hierarchical a posteriori error estimates, whereas Carstensen and Bartels [14] give an upper bound using averaging techniques. Kirby [24] proposed simple residual-based estimates for Raviart–Thomas elements, where, however, the flux estimator is not proved to yield a lower bound and is, moreover, obtained under a saturation assumption. Wheeler and Yotov [39] were able to obtain a posteriori error estimates for the mortar version of all families of mixed finite elements, also including the three-dimensional case; a saturation assumption was, however, necessary for the velocity estimate. Recently, Lovadina and Stenberg [25] employed an idea of postprocessing similar to that used in this paper (with, however, the postprocessed

scalar unknown of one degree lower than the one used here) in order to prove reliable and efficient a posteriori error estimates for both the scalar and flux variables in a mesh-dependent norm. Finally, Hoppe and Wohlmuth [22] treat a diffusion-reaction problem in two space dimensions and use the relation of lowest-order Raviart–Thomas mixed finite elements to nonconforming finite elements derived by Arnold and Brezzi in [6] in order to control, under a saturation assumption, the L^2-norm error in the primal variable p.

To the author's knowledge, no a posteriori estimates for mixed finite element discretizations of convection-diffusion(-reaction) problems have been presented in the literature so far. We do this in section 4 of this paper, after stating the assumptions on the data and formulating the continuous problem in section 2 and after defining the schemes in section 3. The estimates are derived in the energy norm for a new locally (on each element) postprocessed scalar variable $\tilde{p}_h$ such that its flux $-\mathbf{S}\nabla\tilde{p}_h$ is equal to $\mathbf{u}_h$ and such that its mean on each element is equal to p_h. By this construction, we actually have the $L^2(\Omega)$ control over both $\mathbf{u}_h-\mathbf{u}$ and $\tilde{p}_h-p$. Our estimates, in contrast to the usual practice, do not include any undetermined multiplicative constants, so that they are fully (and locally and easily) computable. They represent local lower bounds for the error as well, with efficiency constants of the form $c_1+c_2\min\{\mathrm{Pe},\varrho\}$, where Pe (the local Péclet number) and ϱ are given below by (4.8) and where c_1, c_2 depend only on local variations in $\mathbf{S}$ (i.e., on local inhomogeneities and anisotropies), on local variations in $\mathbf{w}$ and r, on the space dimension, on the polynomial degree of f, and on the shape-regularity parameter of the mesh. They hold from the cases where convection or reaction are not present to convection- or reaction-dominated problems and are in particular semirobust as in [33] with respect to convection dominance. Next, in the pure diffusion case, we can write the general framework for our estimators in a form of an infimum over all $H^1_0(\Omega)$ functions plus a higher-order residual term, which yields asymptotic exactness and full robustness with respect to inhomogeneities and anisotropies, and this without any "monotonicity" hypothesis. Although in numerical experiments we use only local discrete evaluations of the estimators, they remain almost asymptotically exact (the ratio of the estimated and actual error is close to one, and this even in the convection-diffusion-reaction case) and quite robust. Finally, as an interesting consequence of our analysis, we find that in the pure diffusion case with piecewise constant coefficients, the lowest-order mixed finite elements represent an exact three-point scheme in one space dimension, and in two or three space dimensions, the postprocessed approximation is exact with respect to some generalized continuous solution. All these issues are discussed in detail in section 5.

Next, section 6 presents some discrete properties of the schemes and of the postprocessed scalar variable $\tilde{p}_h$. Namely, we show that $\tilde{p}_h$ is nonconforming in the sense that it is not included in $H^1_0(\Omega)$, but we prove that the means of its traces are continuous across interior sides (edges if $d=2$, faces if $d=3$) and equal to zero on exterior sides of the mesh; they are, in fact, shown to equal the Lagrange multipliers from the hybridized forms of the schemes. The actual proofs of our a posteriori error estimates and of their local efficiency are then given in section 7. The key element is Lemma 7.1 which states a primal weak formulation-based abstract framework allowing for the above-discussed asymptotic exactness and asymptotic robustness. The nonconformity of $\tilde{p}_h$ is then treated by the techniques developed in [2, 23, 19]. Neither any additional regularity of the weak solution nor any saturation assumption is needed. Finally, we illustrate the accuracy of the derived estimates in section 8 in several numerical experiments.

In this paper we focus only on lowest-order methods since in practice they are

by far the most commonly used and hence we believe they deserve a special treatment; on the other hand, we do cover the three-dimensional case. Moreover, we have shown in [36] that there exists a local flux-expression formula in lowest-order mixed finite elements and that they can namely be implemented with only one unknown per element, which enables us to significantly decrease their traditional increased computational cost. The extension to higher-order schemes is an ongoing work. Finally, we have also generalized the presented type of a posteriori error estimates to the finite volume method in the forthcoming paper [38]. We treat there among other questions a larger variety of meshes and general inhomogeneous Dirichlet or Neumann boundary conditions. This paper is a detailed description of the results previously announced in [37].

2. Notation, assumptions, and the continuous problem. We introduce here the notation, define admissible triangulations to which the space $W_0(\mathcal{T}_h)$ and the data will be related, and finally give details on the continuous problem (1.1a)–(1.1b).

2.1. Notation. For a domain $S \subset \mathbb{R}^d$, we denote by $L^2(S)$ and $\mathbf{L}^2(S) = [L^2(S)]^d$ the Lebesgue spaces, by $(\cdot,\cdot)_S$ the $L^2(S)$ or $\mathbf{L}^2(S)$ inner product, and by $\|\cdot\|_S$ the associated norm; $|S|$ stands for the Lebesgue measure of S. Next, $H^1(S)$ and $H^1_0(S)$ are the Sobolev spaces of functions with square-integrable weak derivatives, $\mathbf{H}(\mathrm{div}, S) = \{\mathbf{v} \in \mathbf{L}^2(S); \nabla\cdot\mathbf{v} \in L^2(S)\}$ is the space of functions with square-integrable weak divergences, and $\langle\cdot,\cdot\rangle_{\partial S}$ stands for $(d-1)$-dimensional inner product on ∂S or for the duality pairing between $H^{-\frac{1}{2}}(\partial S)$ and $H^{\frac{1}{2}}(\partial S)$. We will also use the "broken Sobolev space" $H^1(\mathcal{T}_h) := \{\varphi \in L^2(\Omega); \varphi|_K \in H^1(K)\ \forall K \in \mathcal{T}_h\}$. In what follows we conceptually denote by C_A, c_A constants dependent only on a quantity A.

2.2. Triangulation, Poincaré and Friedrichs inequalities, and the space $\boldsymbol{W_0(\mathcal{T}_h)}$. We suppose that $\mathcal{T}_h$ for all $h > 0$ consists of closed simplices such that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} K$ and such that if $K, L \in \mathcal{T}_h$, $K \neq L$, then $K \cap L$ is either an empty set or a common face, edge, or vertex of K and L. Let h_K denote the diameter of K and let $h := \max_{K\in\mathcal{T}_h} h_K$. We make the following shape-regularity assumption on the family of triangulations $\{\mathcal{T}_h\}_h$, denoting $\kappa_K := |K|/h_K^d$.

Assumption A (shape-regularity of the meshes). There exists a constant $\kappa_{\mathcal{T}} > 0$ such that $\min_{K\in\mathcal{T}_h} \kappa_K \geq \kappa_{\mathcal{T}}$ for all $h > 0$.

Let ρ_K denote the diameter of the largest ball inscribed in K. Then Assumption A is equivalent to the usual requirement of the existence of a constant $\theta_{\mathcal{T}} > 0$ such that $\max_{K\in\mathcal{T}_h} h_K/\rho_K \leq \theta_{\mathcal{T}}$ for all $h > 0$. We next denote by $\mathcal{E}_h$ the set of all sides of $\mathcal{T}_h$, by $\mathcal{E}_h^{\mathrm{int}}$ the set of interior, by $\mathcal{E}_h^{\mathrm{ext}}$ the set of exterior, and by $\mathcal{E}_K$ the set of all the sides of an element $K \in \mathcal{T}_h$. Finally, h_σ stands for the diameter of $\sigma \in \mathcal{E}_h$.

Let $K \in \mathcal{T}_h$ and $\varphi \in H^1(K)$. Two inequalities play an essential role in our analysis. First, the Poincaré inequality states that

$$\|\varphi - \varphi_K\|_K^2 \leq C_{\mathrm{P},d} h_K^2 \|\nabla\varphi\|_K^2\,, \tag{2.1}$$

where φ_K is the mean of φ over K, $\varphi_K := (\varphi, 1)_K/|K|$, and where the constant $C_{\mathrm{P},d}$ can for a simplex (using its convexity) be evaluated as d/π; cf. [27, 8]. Next, the following generalized Friedrichs inequalities have been proved in [35, Lemma 4.1]:

$$(\varphi_K - \varphi_\sigma)^2 \leq C_{\mathrm{F},d} \frac{h_K^2}{|K|} \|\nabla\varphi\|_K^2\,, \qquad \|\varphi - \varphi_\sigma\|_K^2 \leq C_{\mathrm{F},d} h_K^2 \|\nabla\varphi\|_K^2\,. \tag{2.2}$$

Here φ_σ is the mean of φ over $\sigma \in \mathcal{E}_K$, $\varphi_\sigma := \langle\varphi, 1\rangle_\sigma/|\sigma|$, and $C_{\mathrm{F},d} = 3d$.

We finally define the space $W_0(\mathcal{T}_h)$ of functions with mean values of the traces continuous across interior sides and zero on exterior sides,

$$W_0(\mathcal{T}_h) := \big\{\varphi \in L^2(\Omega)\,;\varphi|_K \in H^1(K)\ \forall K \in \mathcal{T}_h, \\ \langle \varphi|_K - \varphi|_L, 1\rangle_{\sigma_{K,L}} = 0\ \forall \sigma_{K,L} \in \mathcal{E}_h^{\mathrm{int}}\,, \\ \langle \varphi, 1\rangle_\sigma = 0\ \forall \sigma \in \mathcal{E}_h^{\mathrm{ext}}\big\}, \tag{2.3}$$

and recall the discrete Friedrichs inequality

$$\|\varphi\|_\Omega^2 \le C_{\mathrm{DF}} \sum_{K\in\mathcal{T}_h} \|\nabla\varphi\|_K^2 \qquad \forall \varphi \in W_0(\mathcal{T}_h),\ \forall h > 0, \tag{2.4}$$

where C_{DF} depends only on $\kappa_{\mathcal{T}}$ and $\inf_{\mathbf{b}\in\mathbb{R}^d}\{\mathrm{thick}_{\mathbf{b}}(\Omega)\}$; cf. [35, Theorem 5.4].

2.3. Data. We suppose that there exists a basic triangulation $\widetilde{\mathcal{T}}_h$ of Ω such that the data of the problem (1.1a)–(1.1b) are related to $\widetilde{\mathcal{T}}_h$ in the following way.

Assumption B (data).

- (B1) $\mathbf{S}_K := \mathbf{S}|_K$ *is a constant, symmetric, bounded, and uniformly positive definite tensor such that* $c_{\mathbf{S},K}\,\mathbf{v}\cdot\mathbf{v} \le \mathbf{S}_K\mathbf{v}\cdot\mathbf{v} \le C_{\mathbf{S},K}\,\mathbf{v}\cdot\mathbf{v}$, $c_{\mathbf{S},K} > 0$, $C_{\mathbf{S},K} > 0$, *for all* $\mathbf{v}\in\mathbb{R}^d$ *and all* $K \in \widetilde{\mathcal{T}}_h$;
- (B2) $\mathbf{w} \in \mathbf{RTN}^0(\widetilde{\mathcal{T}}_h)$ *satisfies* $\big|\mathbf{w}|_K\big| \le C_{\mathbf{w},K}$, $C_{\mathbf{w},K} \ge 0$, *for all* $K \in \widetilde{\mathcal{T}}_h$;
- (B3) $r_K := r|_K$ *is a constant for all* $K \in \widetilde{\mathcal{T}}_h$;
- (B4) $\frac{1}{2}\nabla\cdot\mathbf{w}|_K + r|_K = c_{\mathbf{w},r,K}$ *and* $\big|\nabla\cdot\mathbf{w}|_K + r_K\big| = C_{\mathbf{w},r,K}$, $c_{\mathbf{w},r,K} \ge 0$, $C_{\mathbf{w},r,K} \ge 0$, *for all* $K \in \widetilde{\mathcal{T}}_h$;
- (B5) $f|_K$ *is a polynomial of degree at most* k *for each* $K \in \widetilde{\mathcal{T}}_h$;
- (B6) *if* $c_{\mathbf{w},r,K} = 0$, *then* $C_{\mathbf{w},r,K} = 0$.

The assumptions that $\mathbf{S}$ and r are piecewise constant on $\widetilde{\mathcal{T}}_h$, that $\mathbf{w} \in \mathbf{RTN}^0(\widetilde{\mathcal{T}}_h)$ (cf. section 3.1 below for the definition of this space), and that f is a piecewise polynomial are made for the sake of simplicity and are usually satisfied in practice. If the functions at hand do not fulfill these requirements, interpolation can be used. Finally, note that Assumption (B6) allows $c_{\mathbf{w},r,K} = 0$ but $\mathbf{w}|_K \neq 0$.

2.4. Continuous problem. Let $\mathcal{T}_h$ be, as throughout the whole paper, a refinement of $\widetilde{\mathcal{T}}_h$. We define a bilinear form $\mathcal{B}$ by

$$\mathcal{B}(p,\varphi) := \sum_{K\in\mathcal{T}_h} \big\{(\mathbf{S}\nabla p, \nabla\varphi)_K + (\nabla\cdot(p\mathbf{w}),\varphi)_K + (rp,\varphi)_K\big\}, \qquad p,\varphi \in H^1(\mathcal{T}_h), \tag{2.5}$$

and the corresponding energy (semi)norm by

$$|||\varphi|||_\Omega^2 := \sum_{K\in\mathcal{T}_h} |||\varphi|||_K^2, \quad |||\varphi|||_K^2 := (\mathbf{S}\nabla\varphi,\nabla\varphi)_K + c_{\mathbf{w},r,K}\|\varphi\|_K^2, \qquad \varphi \in H^1(\mathcal{T}_h). \tag{2.6}$$

In this way $\mathcal{B}(\cdot,\cdot)$ and $|||\cdot|||_\Omega$ are well defined for $p,\varphi \in H^1(\Omega)$ as well as for p,φ that are only piecewise regular. Note also that $|||\cdot|||_\Omega$ is a norm on $W_0(\mathcal{T}_h)$ even if there exists $K \in \mathcal{T}_h$ such that $c_{\mathbf{w},r,K} = 0$ because of the discrete Friedrichs inequality (2.4) and Assumption (B1). The weak formulation of the problem (1.1a)–(1.1b) is then to find $p \in H_0^1(\Omega)$ such that

$$\mathcal{B}(p,\varphi) = (f,\varphi)_\Omega \qquad \forall \varphi \in H_0^1(\Omega). \tag{2.7}$$

Assumption B, the Green theorem, and the Cauchy–Schwarz inequality imply that

$$\mathcal{B}(\varphi,\varphi) = |||\varphi|||_\Omega^2 \qquad \forall \varphi \in H_0^1(\Omega), \tag{2.8}$$

$$\mathcal{B}(\varphi,\varphi) = |||\varphi|||_\Omega^2 + \frac{1}{2}\sum_{K\in\mathcal{T}_h} \langle \varphi^2, \mathbf{w}\cdot\mathbf{n}\rangle_{\partial K} \qquad \forall \varphi \in H^1(\mathcal{T}_h), \tag{2.9}$$

$$\begin{aligned}\mathcal{B}(p,\varphi) \le{}& \max\left\{1, \max_{K\in\mathcal{T}_h}\left\{\frac{C_{\mathbf{w},r,K}}{c_{\mathbf{w},r,K}}\right\}\right\} |||p|||_\Omega |||\varphi|||_\Omega \\ &+ \max_{K\in\mathcal{T}_h}\left\{\frac{C_{\mathbf{w},K}}{\sqrt{c_{\mathbf{S},K}}}\right\} |||p|||_\Omega \|\varphi\|_\Omega \qquad \forall p,\varphi \in H^1(\mathcal{T}_h),\end{aligned} \tag{2.10}$$

and problem (2.7) under Assumption B, in particular, admits a unique solution.

Remark 2.1 (notation). In estimate (2.10), if $c_{\mathbf{w},r,K} = 0$, the term $C_{\mathbf{w},r,K}/c_{\mathbf{w},r,K}$ should be evaluated as zero, since Assumption (B6) in this case gives $C_{\mathbf{w},r,K} = 0$. To simplify notation, we systematically use the convention $0/0 = 0$ throughout the text.

3. Mixed finite element schemes. We define in this section the centered and upwind-weighted mixed finite element schemes.

3.1. Function spaces. Let $\mathbf{RTN}^0_{-1}(\mathcal{T}_h)$ be the space of elementwise linear vector functions $\mathbf{u}_h$ such that, on each $K \in \mathcal{T}_h$, $\mathbf{u}_h|_K = (a_K + d_K x, b_K + d_K y)$ if $d = 2$ and $\mathbf{u}_h|_K = (a_K + d_K x, b_K + d_K y, c_K + d_K z)$ if $d = 3$. The Raviart–Thomas–Nédélec space $\mathbf{RTN}^0(\mathcal{T}_h)$ imposes the continuity of the normal trace across all $\sigma \in \mathcal{E}_h^{\rm int}$ and is given by $\mathbf{RTN}^0(\mathcal{T}_h) := \mathbf{RTN}^0_{-1}(\mathcal{T}_h) \cap \mathbf{H}(\mathrm{div},\Omega)$. There is one basis function $\mathbf{v}_\sigma$ associated with each $\sigma \in \mathcal{E}_h$. For $\sigma_{K,L} \in \mathcal{E}_h^{\rm int}$, $\mathbf{v}_{\sigma_{K,L}}(\mathbf{x}) = \frac{1}{d|K|}(\mathbf{x} - V_K)$, $\mathbf{x} \in K$; $\mathbf{v}_{\sigma_{K,L}}(\mathbf{x}) = \frac{1}{d|L|}(V_L - \mathbf{x})$, $\mathbf{x} \in L$; $\mathbf{v}_{\sigma_{K,L}}(\mathbf{x}) = 0$ otherwise, where V_K is the vertex of K opposite to σ and V_L the vertex of L opposite to σ. We suppose that the orientation of $\mathbf{v}_{\sigma_{K,L}}$, i.e., the order of K and L, is fixed. For a boundary side σ, the support of $\mathbf{v}_\sigma$ consists only of $K \in \mathcal{T}_h$ such that $\sigma \in \mathcal{E}_K$. Next, the space $\Phi(\mathcal{T}_h)$ consists of elementwise constant scalar functions; we denote $p_h|_K = p_K$ for $p_h \in \Phi(\mathcal{T}_h)$. Recall also that $\nabla\cdot\mathbf{u}_h \in \Phi(\mathcal{T}_h)$ for each $\mathbf{u}_h \in \mathbf{RTN}^0_{-1}(\mathcal{T}_h)$.

3.2. Centered scheme. The centered mixed finite element scheme (cf. [17]) reads: find $\mathbf{u}_h \in \mathbf{RTN}^0(\mathcal{T}_h)$ and $p_h \in \Phi(\mathcal{T}_h)$ such that

$$(\mathbf{S}^{-1}\mathbf{u}_h, \mathbf{v}_h)_\Omega - (p_h, \nabla\cdot\mathbf{v}_h)_\Omega = 0 \qquad \forall\, \mathbf{v}_h \in \mathbf{RTN}^0(\mathcal{T}_h), \tag{3.1a}$$

$$(\nabla\cdot\mathbf{u}_h, \phi_h)_\Omega - (\mathbf{S}^{-1}\mathbf{u}_h\cdot\mathbf{w}, \phi_h)_\Omega + ((r + \nabla\cdot\mathbf{w})p_h, \phi_h)_\Omega = (f,\phi_h)_\Omega \qquad \forall \phi_h \in \Phi(\mathcal{T}_h). \tag{3.1b}$$

3.3. Upwind-weighted scheme. The upwind-weighted mixed finite element scheme reads: find $\mathbf{u}_h \in \mathbf{RTN}^0(\mathcal{T}_h)$ and $p_h \in \Phi(\mathcal{T}_h)$ such that

$$(\mathbf{S}^{-1}\mathbf{u}_h, \mathbf{v}_h)_\Omega - (p_h, \nabla\cdot\mathbf{v}_h)_\Omega = 0 \qquad \forall\, \mathbf{v}_h \in \mathbf{RTN}^0(\mathcal{T}_h), \tag{3.2a}$$

$$(\nabla\cdot\mathbf{u}_h, \phi_h)_\Omega + \sum_{K\in\mathcal{T}_h}\sum_{\sigma\in\mathcal{E}_K} \hat{p}_\sigma w_{K,\sigma}\phi_K + (rp_h,\phi_h)_\Omega = (f,\phi_h)_\Omega \qquad \forall \phi_h \in \Phi_h(\mathcal{T}_h), \tag{3.2b}$$

where $w_{K,\sigma} := \langle \mathbf{w}\cdot\mathbf{n}, 1\rangle_\sigma$, $\sigma \in \mathcal{E}_K$, with $\mathbf{n}$ being the unit normal vector of the side σ, outward to K, and where $\hat{p}_\sigma$ is the weighted upwind value defined by

$$\hat{p}_\sigma := \begin{cases} (1-\nu_\sigma)p_K + \nu_\sigma p_L & \text{if } w_{K,\sigma} \geq 0, \\ (1-\nu_\sigma)p_L + \nu_\sigma p_K & \text{if } w_{K,\sigma} < 0, \end{cases} \tag{3.3}$$

if σ is an interior side between elements K and L, and

$$\hat{p}_\sigma := \begin{cases} (1-\nu_\sigma)p_K & \text{if } w_{K,\sigma} \geq 0, \\ \nu_\sigma p_K & \text{if } w_{K,\sigma} < 0, \end{cases} \tag{3.4}$$

if σ is a boundary side. Here, $\nu_\sigma \in [0, 1/2]$ is the coefficient of the amount of upstream weighting which may be, in order to reduce the excessive numerical diffusion added by the full upstream weighting used in [16], chosen as

$$\nu_\sigma := \begin{cases} \min\left\{ c_{\mathbf{S},\sigma}\dfrac{|\sigma|}{h_\sigma |w_{K,\sigma}|}, \dfrac{1}{2}\right\} & \text{if } w_{K,\sigma} \neq 0 \text{ and } \sigma \in \mathcal{E}_h^{\text{int}}, \\ & \text{or if } \sigma \in \mathcal{E}_h^{\text{ext}} \text{ and } w_{K,\sigma} > 0, \\ 0 & \text{if } w_{K,\sigma} = 0 \text{ or if } \sigma \in \mathcal{E}_h^{\text{ext}} \text{ and } w_{K,\sigma} < 0, \end{cases} \tag{3.5}$$

where $c_{\mathbf{S},\sigma}$ is the harmonic average of $c_{\mathbf{S},K}$ and $c_{\mathbf{S},L}$ if $\sigma = \partial K \cap \partial L$ and $c_{\mathbf{S},K}$ otherwise.

4. A posteriori error estimates. We summarize in this section our a posteriori estimates on the error between the weak solution p and a postprocessed variable $\tilde{p}_h$, which we shall define first, along with its modified Oswald interpolate.

4.1. A postprocessed scalar variable $\tilde{p}_h$. In standard mixed finite element theory (see, e.g., Brezzi and Fortin [12] or Roberts and Thomas [31]) the two variables p_h and $\mathbf{u}_h$ are considered as independent. In contrast, the basis for our a posteriori error estimates is a construction of a postprocessed scalar variable $\tilde{p}_h$ which links p_h and $\mathbf{u}_h$ on each simplex in the following way:

$$-\mathbf{S}_K \nabla \tilde{p}_h|_K = \mathbf{u}_h|_K \quad \forall K \in \mathcal{T}_h, \tag{4.1a}$$

$$\frac{(\tilde{p}_h, 1)_K}{|K|} = p_K \quad \forall K \in \mathcal{T}_h. \tag{4.1b}$$

Note that, in particular, if $\mathbf{S} = Id$, $\tilde{p}_h|_K = -d_K/2\,(x^2+y^2) - a_K x - b_K y - e_K$ if $d = 2$ and $\tilde{p}_h|_K = -d_K/2\,(x^2+y^2+z^2) - a_K x - b_K y - c_K z - e_K$ if $d = 3$. Here a_K–d_K are the coefficients from section 3.1, and e_K is given so that (4.1b) was satisfied. If $\mathbf{S} \neq Id$, then $\tilde{p}_h$ verifying (4.1a)–(4.1b) still exists due to the symmetry of $\mathbf{S}$ and is this time a full second-order polynomial on each $K \in \mathcal{T}_h$. The new variable $\tilde{p}_h$ is nonconforming, $\tilde{p}_h \notin H_0^1(\Omega)$, but, by Lemma 6.1 below, $\tilde{p}_h \in W_0(\mathcal{T}_h)$; i.e., its means on interior sides are continuous and its means on exterior sides are equal to zero. In fact, by Lemma 6.4 below, these means coincide with the Lagrange multipliers of hybridized schemes. Moreover, the centered scheme can equivalently be rewritten with the help of $\tilde{p}_h$ (see Lemma 6.2 below), which corresponds to the employment of the Lagrange multipliers in the convection term. Note that the proposed postprocessing is local on each element and its cost is negligible.

4.2. A modified Oswald interpolation operator. Let $\mathbb{P}_l(\mathcal{T}_h)$ denote the space of polynomials of degree at most l on each simplex, not necessary continuous. The Oswald interpolation operator $\mathcal{I}_{\rm Os} : \mathbb{P}_l(\mathcal{T}_h) \to \mathbb{P}_l(\mathcal{T}_h) \cap H^1_0(\Omega)$ has been considered, e.g., in [2, 23, 19]. Given a function $\varphi_h \in \mathbb{P}_l(\mathcal{T}_h)$, $\mathcal{I}_{\rm Os}(\varphi_h)$ is prescribed at the Lagrangian nodes (degrees of freedom; cf. [15, section 2.2]) of $\mathbb{P}_l(\mathcal{T}_h) \cap H^1_0(\Omega)$ by the average of the values of φ_h at this node. We will now construct its modification which preserves the means of $\tilde{p}_h$ over the sides, since this will appear crucial when convection is present.

The modified Oswald interpolation operator $\mathcal{I}_{\rm MO} : \mathbb{P}_2(\mathcal{T}_h) \cap W_0(\mathcal{T}_h) \to \mathbb{P}_d(\mathcal{T}_h) \cap H^1_0(\Omega)$ is defined as follows: at all Lagrangian nodes of $\mathbb{P}_d(\mathcal{T}_h) \cap H^1_0(\Omega)$, except for those lying at the barycenters of the sides, the value of $\mathcal{I}_{\rm MO}(\varphi_h)$ is given by the average of the values of φ_h at this node (as in the standard Oswald interpolation operator). The values at the barycenters of the sides are then established so that the means of $\mathcal{I}_{\rm MO}(\varphi_h)$ over the sides were given by the means of φ_h. (The space $\mathbb{P}_2(\mathcal{T}_h) \cap H^1_0(\Omega)$ in three space dimensions does not have Lagrangian nodes at side barycenters; this is the reason to use $\mathbb{P}_3(\mathcal{T}_h) \cap H^1_0(\Omega)$ in this case.) It is easily verified that, as in the case of the Oswald interpolation operator, $\mathcal{I}_{\rm MO}(\varphi_h)$ is a uniquely defined piecewise polynomial continuous function. Let $[\varphi_h]$ be the jump of a function φ_h across a side σ: if $\sigma = \partial K \cap \partial L$, then $[\varphi_h]$ is the difference of the value of φ_h in K and L (the order of K and L has no influence on what follows), and if $\sigma \in \mathcal{E}_h^{\rm ext}$, then $[\varphi_h] = \varphi_h$. Then the following lemma is an easy modification of [23, Theorem 2.2] ($\sigma \cap K \neq \emptyset$ when σ contains a vertex of K).

LEMMA 4.1 (modified Oswald interpolation operator). *Let $\varphi_h \in \mathbb{P}_2(\mathcal{T}_h) \cap W_0(\mathcal{T}_h)$, and let $\mathcal{I}_{\rm MO}(\varphi_h) \in \mathbb{P}_d(\mathcal{T}_h) \cap H^1_0(\Omega)$ be constructed as described above. Then*

$$\|\nabla(\varphi_h - \mathcal{I}_{\rm MO}(\varphi_h))\|_K^2 \le C_1 \sum_{\sigma;\sigma\cap K\neq\emptyset} h_\sigma^{-1} \|[\varphi_h]\|_\sigma^2 ,$$

where the constant C_1 depends only on d and $\kappa_{\mathcal{T}}$.

4.3. A posteriori error estimates. We now finally state the a posteriori error estimates. Let $K \in \mathcal{T}_h$. Let us first set

$$m_K^2 := \min\left\{ C_{{\rm P},d} \frac{h_K^2}{c_{\mathbf{S},K}}, \frac{1}{c_{\mathbf{w},r,K}} \right\}.$$

We define the *residual estimator* $\eta_{{\rm R},K}$ by

$$\eta_{{\rm R},K} := m_K \|f + \nabla\cdot(\mathbf{S}\nabla\tilde{p}_h) - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h\|_K . \tag{4.2}$$

Next, denote $v := \tilde{p}_h - \mathcal{I}_{\rm MO}(\tilde{p}_h)$. The *nonconformity estimator* $\eta_{{\rm NC},K}$ is given by

$$\eta_{{\rm NC},K} := |||v|||_K \tag{4.3}$$

and the *convection estimator* $\eta_{{\rm C},K}$ by

$$\eta_{{\rm C},K} := \min\left\{ \frac{\|\nabla\cdot(v\mathbf{w}) - \frac{1}{2}v\nabla\cdot\mathbf{w}\|_K}{\sqrt{c_{\mathbf{w},r,K}}}, \left(\frac{C_{{\rm P},d} h_K^2 \|\nabla v\cdot\mathbf{w}\|_K^2}{c_{\mathbf{S},K}} + \frac{9\|v\nabla\cdot\mathbf{w}\|_K^2}{4c_{\mathbf{w},r,K}} \right)^{\frac{1}{2}} \right\}. \tag{4.4}$$

Finally, let

$$m_\sigma^2 := \min\left\{ \max_{K;\,\sigma\in\mathcal{E}_K} \left\{ C_{{\rm F},d} \frac{|\sigma| h_K^2}{|K| c_{\mathbf{S},K}} \right\}, \max_{K;\,\sigma\in\mathcal{E}_K} \left\{ \frac{|\sigma|}{|K| c_{\mathbf{w},r,K}} \right\} \right\} \tag{4.5}$$

for all $\sigma \in \mathcal{E}_h$. We set $\tilde{p}_\sigma := \langle \tilde{p}_h, 1\rangle_\sigma/|\sigma|$, the mean of the postprocessed scalar variable $\tilde{p}_h$ over a side $\sigma \in \mathcal{E}_h$; recall that $\hat{p}_\sigma$ is the upwind value given by (3.3) or (3.4); and define the *upwinding estimator* $\eta_{\mathrm{U},K}$ by

$$\eta_{\mathrm{U},K} := \sum_{\sigma\in\mathcal{E}_K} m_\sigma \|(\hat{p}_\sigma - \tilde{p}_\sigma)\mathbf{w}\cdot\mathbf{n}\|_\sigma \,. \tag{4.6}$$

We have the following a posteriori error estimates.

THEOREM 4.2 (a posteriori error estimate for the centered mixed finite element scheme). *Let p be the weak solution of the problem* (1.1a)–(1.1b) *given by* (2.7)*, and let $\tilde{p}_h$ be the postprocessed solution of the centered mixed finite element scheme* (3.1a)–(3.1b) *given by* (4.1a)–(4.1b)*. Then*

$$|||p - \tilde{p}_h|||_\Omega \le \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2 \right\}^{\frac12} + \left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{C},K})^2 \right\}^{\frac12} . \tag{4.7}$$

THEOREM 4.3 (a posteriori error estimate for the upwind-weighted mixed finite element scheme). *Let p be the weak solution of the problem* (1.1a)–(1.1b) *given by* (2.7)*, and let $\tilde{p}_h$ be the postprocessed solution of the upwind-weighted mixed finite element scheme* (3.2a)–(3.2b) *given by* (4.1a)–(4.1b)*. Then*

$$|||p - \tilde{p}_h|||_\Omega \le \left\{ \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{NC},K}^2 \right\}^{\frac12} + \left\{ \sum_{K\in\mathcal{T}_h} (\eta_{\mathrm{R},K} + \eta_{\mathrm{C},K} + \eta_{\mathrm{U},K})^2 \right\}^{\frac12} .$$

4.4. Local efficiency of the estimates. Let the local Péclet number Pe_K and ϱ_K be given by

$$\mathrm{Pe}_K := h_K \frac{C_{\mathbf{w},K}}{c_{\mathbf{S},K}} , \quad \varrho_K := \frac{C_{\mathbf{w},K}}{\sqrt{c_{\mathbf{w},r,K}}\sqrt{c_{\mathbf{S},K}}} . \tag{4.8}$$

Next, let, for $\varphi \in H^1(K)$,

$$\alpha_{*,K} := c_{\mathbf{S},K}\left(\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}} + 2\varrho_K^2\right) , \quad \beta_{*,K} := c_{\mathbf{w},r,K} + \frac{\left|\nabla\cdot\mathbf{w}|_K\right|^2}{2c_{\mathbf{w},r,K}} ,$$

$$\alpha_{\#,K} := c_{\mathbf{S},K}\left(\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}} + C_{\mathrm{P},d}\mathrm{Pe}_K^2\right) , \quad \beta_{\#,K} := c_{\mathbf{w},r,K} + \frac{9\left|\nabla\cdot\mathbf{w}|_K\right|^2}{4c_{\mathbf{w},r,K}} ,$$

$$|||\varphi|||_{*,K}^2 := \alpha_{*,K}\|\nabla\varphi\|_K^2 + \beta_{*,K}\|\varphi\|_K^2 , \quad |||\varphi|||_{\#,K}^2 := \alpha_{\#,K}\|\nabla\varphi\|_K^2 + \beta_{\#,K}\|\varphi\|_K^2 .$$

Finally, let

$$c_{\mathbf{S},\omega_K} := \min_{L;L\cap K\neq\emptyset} c_{\mathbf{S},L}, \quad c_{\mathbf{w},r,\omega_K} := \min_{L;L\cap K\neq\emptyset} c_{\mathbf{w},r,L}, \quad c_{\mathbf{S},\Omega} := \min_{K\in\mathcal{T}_h} c_{\mathbf{S},K}. \tag{4.9}$$

The theorem below discusses the local efficiency of our a posteriori error estimators.

THEOREM 4.4 (local efficiency of the a posteriori error estimators). *Let p be the weak solution of the problem* (1.1a)–(1.1b) *given by* (2.7)*, and let $\tilde{p}_h$ be the postprocessed solution of the centered mixed finite element scheme* (3.1a)–(3.1b) *or of the*

upwind-weighted mixed finite element scheme (3.2a)–(3.2b) *given by* (4.1a)–(4.1b). *Then, for the residual estimator* $\eta_{\mathrm{R},K}$ *on each* $K \in \mathcal{T}_h$, *there holds*

(4.10)
$$\eta_{\mathrm{R},K} \le C_2 |||p - \tilde{p}_h|||_K \left\{ \sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}}} \max\left\{1, \frac{C_{\mathbf{w},r,K}}{c_{\mathbf{w},r,K}}\right\} + \min\left\{ \mathrm{Pe}_K, \sqrt{\frac{C_{\mathbf{S},K}}{c_{\mathbf{S},K}}}\varrho_K \right\} \right\},$$

where the constant C_2 *depends only on the space dimension* d, *on the shape-regularity parameter* $\kappa_{\mathcal{T}}$, *and on the polynomial degree* k *of* f *(see Lemma* 7.6 *below). Next, for the nonconformity and velocity estimators* $\eta_{\mathrm{NC},K}$ *and* $\eta_{\mathrm{C},K}$ *on each* $K \in \mathcal{T}_h$, *we have*

$$\begin{aligned} \eta_{\mathrm{NC},K}^2 + \eta_{\mathrm{C},K}^2 \le C_3 \min\Bigg\{ & \frac{\alpha_{*,K}}{c_{\mathbf{S},\omega_K}} + \min\left\{ \frac{\beta_{*,K}}{c_{\mathbf{w},r,\omega_K}}, \frac{\beta_{*,K}h_K^2}{c_{\mathbf{S},\omega_K}} \right\}, \\ & \frac{\alpha_{\#,K}}{c_{\mathbf{S},\omega_K}} + \min\left\{ \frac{\beta_{\#,K}}{c_{\mathbf{w},r,\omega_K}}, \frac{\beta_{\#,K}h_K^2}{c_{\mathbf{S},\omega_K}} \right\} \Bigg\} \sum_{L;L\cap K\neq\emptyset} |||p - \tilde{p}_h|||_L^2 \\ & + C_3\beta_{\#,K} \inf_{s_h \in \mathbb{P}_2(\mathcal{T}_h)\cap H_0^1(\Omega)} \sum_{L;L\cap K\neq\emptyset} \|p - s_h\|_L^2, \end{aligned} \tag{4.11}$$

where the constant C_3 *depends only on* d *and* $\kappa_{\mathcal{T}}$ *(see Lemma* 7.7 *below). Finally, the upwinding estimator* $\eta_{\mathrm{U},K}$ *is not efficient and we have only*

$$\sum_{K\in\mathcal{T}_h} \eta_{\mathrm{U},K}^2 \le C_4 \max_{\sigma\in\mathcal{E}_h} \varrho_\sigma \max_{K\in\mathcal{T}_h} \tilde{\varrho}_K \min\left\{ \frac{1}{2} \sum_{K\in\mathcal{T}_h} \frac{\|f\|_K^2}{c_{\mathbf{w},r,K}}, \|f\|_\Omega^2 \frac{C_{\mathrm{DF}}}{c_{\mathbf{S},\Omega}} \right\}, \tag{4.12}$$

where C_{DF} *is the constant from the discrete Friedrichs inequality* (2.4), *the constant* C_4 *depends only on* d *and* $\kappa_{\mathcal{T}}$ *(see Lemma* 7.8 *below), and*

$$\varrho_\sigma := \left(\frac{\max_{K;\sigma\in\mathcal{E}_K} c_{\mathbf{S},K}}{\min_{K;\sigma\in\mathcal{E}_K} c_{\mathbf{S},K}} \right)^2, \quad \tilde{\varrho}_K := \min\left\{ (\mathrm{Pe}_K)^2, (\varrho_K)^2 \frac{\max_{L;L\cap K\in\mathcal{E}_h} c_{\mathbf{w},r,L}}{\min_{L;L\cap K\in\mathcal{E}_h} c_{\mathbf{w},r,L}} \right\}.$$

5. Various remarks. We give several remarks in this section.

5.1. Nature of the estimates. The basis of the a posteriori error estimates derived in this paper is the construction of the postprocessed scalar variable $\tilde{p}_h$ and the consequent application of the abstract framework arising from the primal weak formulation (2.7) of the continuous problem; cf. Lemmas 7.1 and 7.2 below. Compared to Galerkin finite element approximations, the crucial advantage is that $\tilde{p}_h$, an elementwise quadratic polynomial, has the normal traces of its flux $-\mathbf{S}\nabla\tilde{p}_h$ (which is, by (4.1a), nothing else than the mixed finite element vector variable $\mathbf{u}_h$) continuous across interior sides. Hence the side error estimators penalizing the mass balance common in Galerkin finite element methods (cf. [33]) do not appear here at all. This advantage is, however, compensated by the fact that $\tilde{p}_h \notin H_0^1(\Omega)$, so that the estimators known from nonconforming and discontinuous Galerkin finite elements (cf. [19, 23]) appear. Next, whereas in the lowest-order Galerkin finite element method $\nabla\cdot(\mathbf{S}_K\nabla p_h)|_K$ is always equal to zero on all $K \in \mathcal{T}_h$, the element residuals (4.2) give

a very good sense. We also notice that using (2.6), (4.1a), and (2.4),

(5.1)
$$|||p-\tilde{p}_h|||_\Omega^2 = \sum_{K\in\mathcal{T}_h} \left\{ \|\mathbf{S}^{-\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)\|_K^2 + c_{\mathbf{w},r,K}\|p-\tilde{p}_h\|_K^2 \right\}$$
$$\geq \sum_{K\in\mathcal{T}_h} \left\{ \frac{1}{2}\|\mathbf{S}^{-\frac{1}{2}}(\mathbf{u}-\mathbf{u}_h)\|_K^2 + c_{\mathbf{w},r,K}\|p-\tilde{p}_h\|_K^2 \right\} + \frac{c_{\mathbf{S},\Omega}}{2C_{\mathrm{DF}}}\|p-\tilde{p}_h\|_\Omega^2 ,$$

so that we have the usual mixed finite element $L^2(\Omega)$ control over the error in both the scalar and vector unknowns even if $c_{\mathbf{w},r,K}=0$ for some $K\in\mathcal{T}_h$.

5.2. The estimates and their local efficiency with respect to S and w. We discuss here our a posteriori error estimates and their local efficiency that we have been able to prove in Theorem 4.4. For further remarks, see the next section.

The minimum in the definition of the residual estimator $\eta_{\mathrm{R},K}$ (4.2) prevents it from growing to extreme values on coarse elements with a small value $c_{\mathbf{S},K}$ when $c_{\mathbf{w},r,K}>0$. Its local efficiency depends only on anisotropy in its element expressed by the ratio $\sqrt{C_{\mathbf{S},K}/c_{\mathbf{S},K}}$ and there is no dependency on inhomogeneities. Next, under the given assumptions, $C_{\mathbf{w},r,K}/c_{\mathbf{w},r,K}\leq 2$ whenever r_K is nonnegative. Finally, the minimum of the local Péclet number Pe_K and ϱ_K ensures boundedness if $c_{\mathbf{w},r,K}\neq 0$ and if h_K is large and optimal efficiency as Pe_K becomes small.

The minimum in the definition of the convection estimator $\eta_{\mathrm{C},K}$ (4.4) prevents it from exploding when $c_{\mathbf{w},r,K}=0$ but $C_{\mathbf{w},K}\neq 0$. Together with the nonconformity estimator $\eta_{\mathrm{NC},K}$ (4.3), they give local efficiency, up to higher-order terms if $c_{\mathbf{w},r,K}\neq 0$ (the part $\inf_{s_h\in\mathbb{P}_2(\mathcal{T}_h)\cap H_0^1(\Omega)}$), which is shown to be a function of a local (meaning all elements sharing a vertex with the given one) maximal ratio of inhomogeneities (the term $\sqrt{\alpha_{*,K}/c_{\mathbf{S},\omega_K}}$) and of $\sqrt{C_{\mathbf{S},K}/c_{\mathbf{S},K}}$ in each element concerning anisotropy. For further remarks, see the next section. Finally, the efficiency gets into optimal values with respect to convection dominance as Pe_K gets sufficiently small. We note also that the estimate is robust (up to the higher-order term) in the reaction-dominated case as well, since the quantities $C_{\mathbf{w},r,K}/c_{\mathbf{w},r,K}$ and $\sqrt{\beta_{*,K}/c_{\mathbf{w},r,\omega_K}}$ remain well bounded in the limit.

The fact that the upwinding estimator $\eta_{\mathrm{U},K}$ (4.6) cannot in general give a lower bound for the error is quite obvious: it is not difficult to imagine a situation where $p=\tilde{p}_h$, whereas $(\hat{p}_\sigma-\tilde{p}_\sigma)$, the difference of the mean value of $\tilde{p}_h$ on a side σ and of the combination of the mean values of $\tilde{p}_h$ on the elements sharing σ, is generally nonzero. However, we at least show that there is an upper bound for the contributions of this estimator, which moreover decreases with the local Péclet numbers as $O(h)$. It should be noted that this estimator does not change the limit optimality of the schemes and estimates—see section 5.5 below for a remark on this point.

5.3. Asymptotic exactness and asymptotic robustness with respect to inhomogeneities and anisotropies. We show in this remark that the (global asymptotic) efficiency of our estimates is indeed even better than that proved in Theorem 4.4 and discussed in the previous section.

5.3.1. Pure diffusion problems. Let us first consider a pure diffusion problem, i.e., $r=\mathbf{w}=0$ in (1.1a)–(1.1b). Using that in this case $-\nabla\cdot(\mathbf{S}_K\nabla\tilde{p}_h|_K)=\nabla\cdot\mathbf{u}_h|_K=f_K$ for all $K\in\mathcal{T}_h$, where f_K is the mean value of f over K, the analysis for the general

case simplifies to the a posteriori error estimate (4.7) with $\eta_{\mathrm{C},K} = 0$ and

$$\eta_{\mathrm{R},K}^2 := C_{\mathrm{P},d} \frac{h_K^2}{c_{\mathbf{S},K}} \|f - f_K\|_K^2 , \tag{5.2}$$

$$\eta_{\mathrm{NC},K}^2 := \|\mathbf{S}^{\frac{1}{2}} \nabla(\tilde{p}_h - s)\|_K^2 , \tag{5.3}$$

where in particular $s \in H_0^1(\Omega)$ can be chosen arbitrarily (cf. Lemma 7.2 below). Examples are the Oswald or the modified Oswald interpolates of $\tilde{p}_h$—in the pure diffusion case, all the presented results hold similarly for these two operators. Also note that since $\nabla \cdot (\mathbf{u} - \mathbf{u}_h)|_K = f - f_K$ is fully computable for all $K \in \mathcal{T}_h$, the control over $\|\mathbf{u} - \mathbf{u}_h\|_\Omega + \|\nabla \cdot (\mathbf{u} - \mathbf{u}_h)\|_\Omega$ immediately follows using (5.1).

Our main point is, however, that the above developments in fact imply

$$|||p - \tilde{p}_h|||_\Omega \leq \inf_{s \in H_0^1(\Omega)} |||\tilde{p}_h - s|||_\Omega + \left\{ \sum_{K \in \mathcal{T}_h} C_{\mathrm{P},d} \frac{h_K^2}{c_{\mathbf{S},K}} \|f - f_K\|_K^2 \right\}^{\frac{1}{2}} , \tag{5.4}$$

which, in the case where f is piecewise constant, by virtue of

$$\inf_{s \in H_0^1(\Omega)} |||\tilde{p}_h - s|||_\Omega \leq |||\tilde{p}_h - p|||_\Omega,$$

gives asymptotic global efficiency of such an estimator with a constant 1, i.e., asymptotic exactness and asymptotic full robustness with respect to inhomogeneities and anisotropies (asymptotic with respect to the approximation of $\tilde{p}_h$ by some, e.g., polynomial, $s \in H_0^1(\Omega)$ on a fixed grid $\mathcal{T}_h$). In the general case, if, e.g., $f \in H^1(\mathcal{T}_h)$, then $\|f - f_K\|_K^2 \leq C_{\mathrm{P},d} h_K^2 \|\nabla f\|_K^2$, and asymptotic exactness and asymptotic robustness still hold true (this time asymptotic also with respect to $h \to 0$). Although we use only the Oswald or the modified Oswald interpolates of $\tilde{p}_h$ instead of evaluating or approximating the infimum in (5.4), the numerical experiments of section 8.1 below show that estimators of section 4.3 remain almost asymptotically exact and robust with respect to inhomogeneities and anisotropies.

5.3.2. Convection-diffusion-reaction problems. The above considerations roughly extend to the convection-diffusion-reaction case in the following sense: for the centered mixed finite element scheme (3.1a)–(3.1b), one has (7.4) and consequently a superconvergence of the residual estimators $\eta_{\mathrm{R},K}$ (4.2) to zero. Next, for divergence-free velocity fields $\mathbf{w}$, the second arguments of the convection estimators $\eta_{\mathrm{C},K}$ in (4.4) again superconverge to zero since $\tilde{p}_h \in W_0(\mathcal{T}_h)$ (both as $h \to 0$). Hence the estimate will be asymptotically given only by the nonconformity estimators $\eta_{\mathrm{NC},K}$ of (4.3) and thus by the best approximation of $\tilde{p}_h$ by $s \in H_0^1(\Omega)$ such that its means are given by the means of $\tilde{p}_h$. (This property is needed when convection is present; see Lemma 7.4 below.) This asymptotic almost optimal efficiency is again observed below in numerical experiments in section 8.2.

5.4. Pure diffusion problems: Mixed finite elements and a generalized weak solution. Let us in this remark consider $r = \mathbf{w} = 0$ in (1.1a)–(1.1b) and generalize the classical weak solution to a function $\tilde{p} \in W_0(\mathcal{T}_h)$ such that

$$\mathcal{B}(\tilde{p}, \varphi) = (f, \varphi)_\Omega \qquad \forall \varphi \in W_0(\mathcal{T}_h) . \tag{5.5}$$

(In)equalities (2.9) and (2.10) together with the discrete Friedrichs inequality (2.4) ensure the existence of a unique solution of (5.5).

We thus have

$$|||\tilde{p}-\tilde{p}_h|||_\Omega = \frac{\mathcal{B}(\tilde{p}-\tilde{p}_h,\tilde{p}-\tilde{p}_h)}{|||\tilde{p}-\tilde{p}_h|||_\Omega} \le \sup_{\varphi\in W_0(\mathcal{T}_h),\,|||\varphi|||_\Omega=1} \mathcal{B}(\tilde{p}-\tilde{p}_h,\varphi)$$

and develop, similarly as in the proof of Lemma 7.2 below,

$$\begin{aligned}\mathcal{B}(\tilde{p}-\tilde{p}_h,\varphi) &= (f,\varphi)_\Omega + \sum_{K\in\mathcal{T}_h}\left\{(\nabla\cdot(\mathbf{S}\nabla\tilde{p}_h),\varphi)_K - \langle\mathbf{S}\nabla\tilde{p}_h\cdot\mathbf{n},\varphi\rangle_{\partial K}\right\}\\ &= \sum_{K\in\mathcal{T}_h}(f-\nabla\cdot\mathbf{u}_h,\varphi)_K + \sum_{\sigma\in\mathcal{E}_h}\langle\mathbf{u}_h\cdot\mathbf{n},[\varphi]\rangle_\sigma\\ &= \sum_{K\in\mathcal{T}_h}(f-\nabla\cdot\mathbf{u}_h,\varphi)_K = \sum_{K\in\mathcal{T}_h}(f-f_K,\varphi-\varphi_K)_K,\end{aligned}$$

using the bilinearity of $\mathcal{B}(\cdot,\cdot)$, the definition (5.5) of the generalized weak solution $\tilde{p}$, the Green theorem in each $K\in\mathcal{T}_h$, the relation (4.1a) between $\tilde{p}_h$ and $\mathbf{u}_h$, reordering the summation over the boundaries of elements to the summation over the sides, using the continuity of the normal trace of $\mathbf{u}_h$ expressed by $\mathbf{u}_h|_K\cdot\mathbf{n}_K = -\mathbf{u}_h|_L\cdot\mathbf{n}_L$ on $\sigma_{K,L}\in\mathcal{E}_h^{\mathrm{int}}$, the fact that $\mathbf{u}_h\cdot\mathbf{n}$ is constant on all sides $\sigma\in\mathcal{E}_h$ and the definition (2.3) of the space $W_0(\mathcal{T}_h)$, and finally the equation (3.1b) of the definition of the mixed finite element scheme (φ_K is the mean of φ over K). Next, estimate (7.5) given below holds true also in this case, so that finally the Cauchy–Schwarz inequality leads to

$$|||\tilde{p}-\tilde{p}_h|||_\Omega \le \left\{\sum_{K\in\mathcal{T}_h}\eta_{\mathrm{R},K}^2\right\}^{\frac12}$$

with $\eta_{\mathrm{R},K}$ given by (5.2).

First, this is a completely data-dependent a posteriori error estimate, and second, this is in fact an a priori error estimate as well: it shows that the mixed finite element solutions $\tilde{p}_h$ and $\mathbf{u}_h$ (cf. (5.1), which still holds true) converge both as $O(h^2)$ in the $L^2(\Omega)$, $\mathbf{L}^2(\Omega)$, respectively, norms to the generalized weak solution $\tilde{p}$ given by (5.5) and its flux $\tilde{\mathbf{u}}$, $\tilde{\mathbf{u}}|_K := -\mathbf{S}\nabla\tilde{p}|_K$ (for $f\in H^1(\mathcal{T}_h)$). Moreover, as soon as f is piecewise constant, $\tilde{p}_h$ is directly equal to the generalized solution! We emphasize that these results hold true for $\mathbf{S}$ piecewise constant but arbitrarily inhomogeneous and anisotropic; they apparently confirm the observations of a very good behavior of mixed methods in these circumstances. There are also very interesting consequences in one space dimension; cf. section 5.6 below.

5.5. A combination of the centered and upwind-weighted schemes. The scheme (3.2a)–(3.2b) guarantees stability in the convection-dominated case, but the additional upwinding estimator $\eta_{\mathrm{U},K}$ given by (4.6) is unfortunately not efficient. On the other hand, the scheme (3.1a)–(3.1b), however precise if h is sufficiently small, may give completely wrong results on coarse meshes. Hence a good idea may be a smooth transition from the upwind-weighted to the centered scheme under the form

$$(\mathbf{S}^{-1}\mathbf{u}_h,\mathbf{v}_h)_\Omega - (p_h,\nabla\cdot\mathbf{v}_h)_\Omega = 0 \qquad \forall\,\mathbf{v}_h\in\mathbf{RTN}^0(\mathcal{T}_h),$$

$$(\nabla\cdot\mathbf{u}_h,\phi_K)_K + \sum_{\sigma\in\mathcal{E}_K}\left\{(\mu_\sigma\hat{p}_\sigma+(1-\mu_\sigma)\tilde{p}_\sigma)w_{K,\sigma}\phi_K\right\} + (rp_h,\phi_K)_K = (f,\phi_K)_K \qquad \forall K\in\mathcal{T}_h,$$

where $\hat{p}_\sigma$ is the upstream value and μ_σ is set to $1-2\nu_\sigma$ with ν_σ given by (3.5). Notice that such a scheme is fully rewritable in terms of the original unknowns p_h, $\mathbf{u}_h$, using that $\sum_{\sigma\in\mathcal{E}_K}\tilde{p}_\sigma w_{K,\sigma}\phi_K = \langle\tilde{p}_h\mathbf{w}\cdot\mathbf{n},\phi_K\rangle_{\partial K}$ and Lemma 6.2 below.

5.6. The estimates in one space dimension. As the last remark, it appears that the above results have interesting particular consequences in one space dimension, where the two schemes (3.1a)–(3.1b) and (3.2a)–(3.2b) can likewise be defined.

5.6.1. One dimension: No nonconformity. First of all, Lemma 6.1 below reduces in one space dimension to the assertion that the postprocessed variable $\tilde{p}_h$ given by (4.1a)–(4.1b) is continuous, i.e., that in this case $\tilde{p}_h \in H_0^1(\Omega)$. An immediate consequence is that the parts of the a posteriori error estimates of Theorems 4.2–4.3 related to nonconformity disappear.

5.6.2. Lowest-order mixed finite elements: An exact three-point scheme for one-dimensional diffusion problems with piecewise constant coefficients. Another quite interesting consequence is related to the remark of section 5.4 and results of [36]. As there is no nonconformity, the superconvergence $O(h^2)$ of both $\tilde{p}_h$ and $\mathbf{u}_h$ (this time towards the weak solution and its flux, coinciding with the generalized one) always holds true, and, moreover, it appears that in one space dimension, one can always rewrite the schemes with only p_K, $K\in\mathcal{T}_h$, as unknowns. Hence the lowest-order mixed finite elements represent a scheme with a three-point stencil which is exact for one-dimensional pure diffusion problems, where the diffusion tensor $\mathbf{S}$ (this time a scalar function) and the right-hand side f are piecewise constant (and hence possibly arbitrarily discontinuous). This should be compared to the known results for the finite volume/finite difference method. In particular, the (best known?) scheme proposed by Ewing, Iliev, and Lazarov in [21] is exact only when the right-hand side is constant (the diffusion tensor may be piecewise constant); cf. Remark 2.4 in [21].

6. Discrete properties of the schemes. In this section we prove different properties of the schemes (3.1a)–(3.1b) and (3.2a)–(3.2b) and of the postprocessed scalar variable $\tilde{p}_h$ needed in the paper.

LEMMA 6.1 (continuity of the means of traces of $\tilde{p}_h$). *It holds that $\tilde{p}_h\in W_0(\mathcal{T}_h)$; i.e.,*

$$\langle\tilde{p}_h|_K-\tilde{p}_h|_L,1\rangle_{\sigma_{K,L}}=0\qquad\forall\sigma_{K,L}\in\mathcal{E}_h^{\mathrm{int}},$$

$$\langle\tilde{p}_h,1\rangle_\sigma=0\qquad\forall\sigma\in\mathcal{E}_h^{\mathrm{ext}}.$$

Proof. Let us consider a side $\sigma_{K,L}\in\mathcal{E}_h^{\mathrm{int}}$. Then taking $\mathbf{v}_h$ equal to the basis function $\mathbf{v}_{\sigma_{K,L}}$ (cf. section 3.1) in (3.1a) or (3.2a) yields

$$\begin{aligned}0&=-(\nabla\tilde{p}_h,\mathbf{v}_{\sigma_{K,L}})_{K\cup L}-(\tilde{p}_h,\nabla\cdot\mathbf{v}_{\sigma_{K,L}})_{K\cup L}\\&=-\langle\mathbf{v}_{\sigma_{K,L}}\cdot\mathbf{n},\tilde{p}_h\rangle_{\partial K}-\langle\mathbf{v}_{\sigma_{K,L}}\cdot\mathbf{n},\tilde{p}_h\rangle_{\partial L}=\langle\mathbf{v}_{\sigma_{K,L}}\cdot\mathbf{n}_K,\tilde{p}_h|_L-\tilde{p}_h|_K\rangle_{\sigma_{K,L}},\end{aligned}$$

using the definition (4.1a)–(4.1b) of $\tilde{p}_h$, the fact that $\nabla\cdot\mathbf{v}_h$ for $\mathbf{v}_h\in\mathbf{RTN}^0(\mathcal{T}_h)$ is constant in each simplex (which allows us to replace p_h by $\tilde{p}_h$), the Green theorem, and the fact that $\mathbf{v}_{\sigma_{K,L}}$ has a nonzero normal flux only through $\sigma_{K,L}$. The first assertion of the lemma follows by the fact that $\mathbf{v}_h\cdot\mathbf{n}$ for $\mathbf{v}_h\in\mathbf{RTN}^0(\mathcal{T}_h)$ is constant on each side $\sigma\in\mathcal{E}_h$. The proof for boundary sides is completely similar. ∎

LEMMA 6.2 (equivalent form of the centered scheme). *The scheme* (3.1a)–(3.1b) *can be equivalently written: find* $\mathbf{u}_h \in \mathbf{RTN}^0(\mathcal{T}_h)$ *and* $p_h \in \Phi(\mathcal{T}_h)$ *such that*

$$(\mathbf{S}^{-1}\mathbf{u}_h, \mathbf{v}_h)_\Omega - (\tilde{p}_h, \nabla\cdot\mathbf{v}_h)_\Omega = 0 \qquad \forall\, \mathbf{v}_h \in \mathbf{RTN}^0(\mathcal{T}_h), \tag{6.1a}$$

$$(\nabla\cdot\mathbf{u}_h, \phi_K)_K + \langle \tilde{p}_h\mathbf{w}\cdot\mathbf{n}, \phi_K\rangle_{\partial K} + (r\tilde{p}_h, \phi_K)_K = (f, \phi_K)_K \qquad \forall K \in \mathcal{T}_h, \tag{6.1b}$$

where $\tilde{p}_h$ *is defined by* (4.1a)–(4.1b).

Proof. Since $\nabla\cdot\mathbf{v}_h$ for $\mathbf{v}_h \in \mathbf{RTN}^0(\mathcal{T}_h)$ is constant in each simplex and since r was in Assumption (B3) supposed piecewise constant as well, one can replace p_h by $\tilde{p}_h$ in the terms $(p_h, \nabla\cdot\mathbf{v}_h)_\Omega$ and $(rp_h, \phi_K)_K$ using (4.1b). Similarly, using in addition the Green theorem,

$$\begin{aligned}-(\mathbf{S}_K^{-1}\mathbf{u}_h\cdot\mathbf{w}, \phi_K)_K + (p_K\nabla\cdot\mathbf{w}, \phi_K)_K &= (\nabla\tilde{p}_h\cdot\mathbf{w}, \phi_K)_K + (\tilde{p}_h\nabla\cdot\mathbf{w}, \phi_K)_K \\ &= (\nabla\cdot(\tilde{p}_h\mathbf{w}), \phi_K)_K = \langle \tilde{p}_h\mathbf{w}\cdot\mathbf{n}, \phi_K\rangle_{\partial K}. \qquad \square\end{aligned}$$

Remark 6.3 (hybridization of the schemes). Mixed finite element schemes can equivalently be reformulated while relaxing the continuity of the normal trace of $\mathbf{u}_h$ required in the definition of the space $\mathbf{RTN}^0(\mathcal{T}_h)$ and imposing it instead with the help of Lagrange multipliers λ_σ, $\sigma \in \mathcal{E}_h^{\rm int}$; cf. [12, section V.1.2]. The centered scheme (3.1a)–(3.1b), taking into account its equivalent form given by Lemma 6.2, then changes to: find $\mathbf{u}_h \in \mathbf{RTN}^0_{-1}(\mathcal{T}_h)$, $p_h \in \Phi(\mathcal{T}_h)$, and λ_σ, $\sigma \in \mathcal{E}_h^{\rm int}$, with $\tilde{p}_h$ defined by (4.1a)–(4.1b), such that

$$\sum_{K\in\mathcal{T}_h}\left\{(\mathbf{S}^{-1}\mathbf{u}_h, \mathbf{v}_h)_K - (\tilde{p}_h, \nabla\cdot\mathbf{v}_h)_K + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_h^{\rm int}}\langle \mathbf{v}_h\cdot\mathbf{n}, \lambda_\sigma\rangle_\sigma\right\} = 0 \tag{6.2a}$$

$$\forall\, \mathbf{v}_h \in \mathbf{RTN}^0_{-1}(\mathcal{T}_h),$$

$$(\nabla\cdot\mathbf{u}_h, \phi_K)_K + \langle \tilde{p}_h\mathbf{w}\cdot\mathbf{n}, \phi_K\rangle_{\partial K} + (r\tilde{p}_h, \phi_K)_K = (f, \phi_K)_K \qquad \forall K \in \mathcal{T}_h, \tag{6.2b}$$

$$\langle (\mathbf{u}_h\cdot\mathbf{n})|_K + (\mathbf{u}_h\cdot\mathbf{n})|_L, 1\rangle_{\sigma_{K,L}} = 0 \qquad \forall \sigma_{K,L} \in \mathcal{E}_h^{\rm int}, \tag{6.2c}$$

whereas the upwind-weighted scheme (3.2a)–(3.2b) becomes: find $\mathbf{u}_h \in \mathbf{RTN}^0_{-1}(\mathcal{T}_h)$, $p_h \in \Phi(\mathcal{T}_h)$, and λ_σ, $\sigma \in \mathcal{E}_h^{\rm int}$ such that

$$\sum_{K\in\mathcal{T}_h}\left\{(\mathbf{S}^{-1}\mathbf{u}_h, \mathbf{v}_h)_K - (p_h, \nabla\cdot\mathbf{v}_h)_K + \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_h^{\rm int}}\langle \mathbf{v}_h\cdot\mathbf{n}, \lambda_\sigma\rangle_\sigma\right\} = 0 \tag{6.3a}$$

$$\forall\, \mathbf{v}_h \in \mathbf{RTN}^0_{-1}(\mathcal{T}_h),$$

$$(\nabla\cdot\mathbf{u}_h, \phi_K)_K + \sum_{\sigma\in\mathcal{E}_K}\hat{p}_\sigma w_{K,\sigma}\phi_K + (rp_h, \phi_K)_K = (f, \phi_K)_K \qquad \forall K \in \mathcal{T}_h, \tag{6.3b}$$

$$\langle (\mathbf{u}_h\cdot\mathbf{n})|_K + (\mathbf{u}_h\cdot\mathbf{n})|_L, 1\rangle_{\sigma_{K,L}} = 0 \qquad \forall \sigma_{K,L} \in \mathcal{E}_h^{\rm int}. \tag{6.3c}$$

LEMMA 6.4 (relation of $\tilde{p}_h$ to the Lagrange multipliers λ_σ). *It holds that*

$$\lambda_\sigma = \tilde{p}_\sigma = \frac{\langle \tilde{p}_h, 1\rangle_\sigma}{|\sigma|} \qquad \forall \sigma \in \mathcal{E}_h^{\rm int}.$$

Proof. The proof is similar to that of Lemma 6.1. Let $K \in \mathcal{T}_h$ and $\sigma \in \mathcal{E}_K \cap \mathcal{E}_h^{\rm int}$. Then taking $\mathbf{v}_h = \mathbf{v}_\sigma$ in (6.2a) or (6.3a), we have

$$0 = -(\nabla \tilde{p}_h, \mathbf{v}_\sigma)_K - (\tilde{p}_h, \nabla \cdot \mathbf{v}_\sigma)_K + \langle \mathbf{v}_\sigma \cdot \mathbf{n}, \lambda_\sigma \rangle_\sigma = \langle \mathbf{v}_\sigma \cdot \mathbf{n}, \lambda_\sigma - \tilde{p}_h \rangle_\sigma \,,$$

using the definition (4.1a)–(4.1b) of $\tilde{p}_h$, the fact that $\nabla \cdot \mathbf{v}_\sigma$ is constant in each simplex, the fact that $\mathbf{v}_\sigma$ has a nonzero normal flux only through σ, and the Green theorem. The assertion of the lemma follows by the fact that $\mathbf{v}_\sigma \cdot \mathbf{n}$ is constant on σ. ∎

LEMMA 6.5 (a priori estimate for the upwind-weighted scheme). *Let $\mathbf{u}_h$, p_h be the solutions of the upwind-weighted scheme* (3.2a)–(3.2b)*, and let $\tilde{p}_h$ be the postprocessed scalar variable given by* (4.1a)–(4.1b)*. Then*

$$\sum_{K \in \mathcal{T}_h} \left\{ c_{\mathbf{S},K} \|\nabla \tilde{p}_h\|_K^2 + \frac{1}{2} c_{\mathbf{w},r,K} \, \|p_h\|_K^2 \right\} \le \frac{1}{2} \sum_{K \in \mathcal{T}_h} \frac{\|f\|_K^2}{c_{\mathbf{w},r,K}}$$

if $c_{\mathbf{w},r,K} > 0$ for all $K \in \mathcal{T}_h$ and

$$\sum_{K \in \mathcal{T}_h} \left\{ \frac{1}{2} c_{\mathbf{S},K} \|\nabla \tilde{p}_h\|_K^2 + c_{\mathbf{w},r,K} \, \|p_h\|_K^2 \right\} \le \frac{\|f\|_\Omega^2}{2} \frac{C_{\rm DF}}{c_{\mathbf{S},\Omega}},$$

where $c_{\mathbf{S},\Omega}$ is given by (4.9) *and $C_{\rm DF}$ is the constant from the discrete Friedrichs inequality* (2.4).

Proof. Let us set $\phi_h = p_h$ in (3.2b). We then can rewrite the first term of the left-hand side of (3.2b) as

$$\sum_{K \in \mathcal{T}_h} (\nabla \cdot \mathbf{u}_h, p_K)_K = \sum_{K \in \mathcal{T}_h} \left\{ -(\mathbf{u}_h, \nabla \tilde{p}_h)_K + \langle \mathbf{u}_h \cdot \mathbf{n}, \tilde{p}_h \rangle_{\partial K} \right\} = \sum_{K \in \mathcal{T}_h} (\mathbf{S}_K \nabla \tilde{p}_h, \nabla \tilde{p}_h)_K$$

$$+ \sum_{\sigma_{K,L} \in \mathcal{E}_h^{\rm int}} \langle \mathbf{u}_h \cdot \mathbf{n}_K, \tilde{p}_h|_K - \tilde{p}_h|_L \rangle_{\sigma_{K,L}} + \sum_{\sigma \in \mathcal{E}_h^{\rm ext}} \langle \mathbf{u}_h \cdot \mathbf{n}, \tilde{p}_h \rangle_\sigma \ge \sum_{K \in \mathcal{T}_h} c_{\mathbf{S},K} \|\nabla \tilde{p}_h\|_K^2 \,,$$

using the fact that $\nabla \cdot \mathbf{u}_h$ is constant on each $K \in \mathcal{T}_h$ and we thus can replace p_h by $\tilde{p}_h$ employing (4.1b), the Green theorem, (4.1a), the fact that $\mathbf{u}_h \cdot \mathbf{n}$ is constant on each $\sigma \in \mathcal{E}_h$, the continuity of the means of the traces of $\tilde{p}_h$ given by Lemma 6.1, and finally Assumption (B1). Next,

$$\sum_{K \in \mathcal{T}_h} \sum_{\sigma \in \mathcal{E}_K} \hat{p}_\sigma w_{K,\sigma} p_K = \sum_{\sigma_{K,L} \in \mathcal{E}_h^{\rm int}} \{ \hat{p}_\sigma w_{K,\sigma} p_K + \hat{p}_\sigma w_{L,\sigma} p_L \} + \sum_{\sigma_K \in \mathcal{E}_h^{\rm ext}} \hat{p}_\sigma w_{K,\sigma} p_K$$

$$= \sum_{\sigma_{K,L} \in \mathcal{E}_h^{\rm int},\, w_{K,\sigma} \ge 0} w_{K,\sigma} \left(p_K (p_K - p_L) - \nu_\sigma (p_L - p_K)^2 \right) + \sum_{\sigma_K \in \mathcal{E}_h^{\rm ext}} \hat{p}_\sigma w_{K,\sigma} p_K$$

$$= \frac{1}{2} \sum_{\sigma_{K,L} \in \mathcal{E}_h^{\rm int},\, w_{K,\sigma} \ge 0} w_{K,\sigma} (p_K^2 - p_L^2) + \sum_{\sigma_{K,L} \in \mathcal{E}_h^{\rm int}} |w_{K,\sigma}| (p_L - p_K)^2 \left(\frac{1}{2} - \nu_\sigma \right)$$

$$+ \sum_{\sigma_K \in \mathcal{E}_h^{\rm ext}} \left\{ \frac{1}{2} p_K^2 w_{K,\sigma} + |w_{K,\sigma}| p_K^2 \left(\frac{1}{2} - \nu_\sigma \right) \right\} \ge \frac{1}{2} \sum_{K \in \mathcal{T}_h} p_K^2 (\nabla \cdot \mathbf{w}, 1)_K,$$

where we have rewritten the summation over the sides and fixed denotation of $K, L \in \mathcal{T}_h$ sharing a side $\sigma_{K,L} \in \mathcal{E}_h^{\rm int}$ such that $w_{K,\sigma} \ge 0$; used that $w_{K,\sigma} = -w_{L,\sigma}$, the

definition (3.3)–(3.4) of $\hat{p}_\sigma$, and the relation $2a(a-b) = (a-b)^2 + a^2 - b^2$; estimated using $0 \le \nu_\sigma \le 1/2$, which follows from (3.5); rewritten the summation back over the elements and their sides; and finally employed the Green theorem, giving $\sum_{\sigma\in\mathcal{E}_K} w_{K,\sigma} = (\nabla\cdot\mathbf{w},1)_K$. Finally, $(rp_h,p_h)_\Omega = \sum_{K\in\mathcal{T}_h} p_K^2 (r,1)_K$.

The right-hand side of (3.2b) with $\phi_h = p_h$ can be estimated either by

$$(f,p_h)_\Omega \le \sum_{K\in\mathcal{T}_h} \|f\|_K \frac{\sqrt{c_{\mathbf{w},r,K}}}{\sqrt{c_{\mathbf{w},r,K}}} \|p_h\|_K \le \frac{1}{2}\sum_{K\in\mathcal{T}_h} \frac{\|f\|_K^2}{c_{\mathbf{w},r,K}} + \frac{1}{2}\sum_{K\in\mathcal{T}_h} c_{\mathbf{w},r,K}\|p_h\|_K^2$$

or by

$$(f,p_h)_\Omega \le \|f\|_\Omega \|p_h\|_\Omega \le \frac{\|f\|_\Omega^2}{2}\frac{C_{\mathrm{DF}}}{c_{\mathbf{S},\Omega}} + \frac{c_{\mathbf{S},\Omega}}{C_{\mathrm{DF}}}\frac{\|\tilde{p}_h\|_\Omega^2}{2} \le \frac{\|f\|_\Omega^2}{2}\frac{C_{\mathrm{DF}}}{c_{\mathbf{S},\Omega}} + \frac{c_{\mathbf{S},\Omega}}{2}\sum_{K\in\mathcal{T}_h}\|\nabla\tilde{p}_h\|_K^2,$$

using the Cauchy–Schwarz, $ab \le \varepsilon a^2/2 + b^2/(2\varepsilon)$, $\varepsilon > 0$, $\|p_h\|_K \le \|\tilde{p}_h\|_K$, and the discrete Friedrichs (2.4) inequalities. The assertion follows by combining the above estimates. ∎

Remark 6.6 (existence and uniqueness for the upwind-weighted scheme). From Lemma 6.5, existence and uniqueness for the upwind-weighted scheme (3.2a)–(3.2b) easily follows. Indeed, let $f = 0$. Then $p_h = 0$ and $\mathbf{u}_h = -\mathbf{S}\nabla\tilde{p}_h = 0$ for all $K \in \mathcal{T}_h$.

Remark 6.7 (existence and uniqueness for the centered scheme). In contrast with the upwind-weighted scheme, existence and uniqueness for the centered scheme (3.1a)–(3.1b) is in [17] guaranteed only for "h sufficiently small." Alternatively, there exists a unique solution if $C_{\mathbf{w},K} \le 2(1-\mu)\sqrt{c_{\mathbf{S},K}}\sqrt{\tilde{c}_{\mathbf{w},r,K}}$ for some $\mu \in (0,1)$ and all $K \in \mathcal{T}_h$, where $(\nabla\cdot\mathbf{w} + r)|_K = \tilde{c}_{\mathbf{w},r,K} > 0$, which corresponds to the case that is not convection-dominated.

7. Proofs of the a posteriori error estimates and of their local efficiency. We shall prove in this section the a posteriori error estimates stated by Theorems 4.2–4.3, as well as their local efficiency discussed in Theorem 4.4.

7.1. Proofs of the a posteriori error estimates. To begin with, we state the following result, the purpose of which is to give an optimal abstract bound on the error between $p \in H^1(\Omega)$ and $\tilde{p} \in H^1(\mathcal{T}_h)$ in the energy (semi)norm $|||\cdot|||_\Omega$. ($H^1_{\mathrm{D}}(\Omega)$ is the subspace of $H^1(\Omega)$ of functions with traces vanishing on $\Gamma_{\mathrm{D}} \subset \partial\Omega$.)

LEMMA 7.1 (abstract framework). *Let $\Gamma_{\mathrm{D}} \subset \partial\Omega$, $|\Gamma_{\mathrm{D}}| \neq 0$, let $\Gamma_{\mathrm{in}} := \{\mathbf{x} \in \partial\Omega; \mathbf{w}\cdot\mathbf{n} < 0\} \subset \Gamma_{\mathrm{D}}$, let $p, s \in H^1(\Omega)$ be such that $p - s \in H^1_{\mathrm{D}}(\Omega)$, and let $\tilde{p} \in H^1(\mathcal{T}_h)$ be arbitrary. Then*

$$|||p-\tilde{p}|||_\Omega \le |||\tilde{p}-s|||_\Omega + \left| \mathcal{B}\left(p-\tilde{p}, \frac{p-s}{|||p-s|||_\Omega}\right) + \sum_{K\in\mathcal{T}_h}\left(\nabla\cdot((\tilde{p}-s)\mathbf{w}) - \frac{1}{2}(\tilde{p}-s)\nabla\cdot\mathbf{w}, \frac{p-s}{|||p-s|||_\Omega}\right)_K \right|.$$

Proof. Let us set, for $p, \varphi \in H^1(\mathcal{T}_h)$,

$$\mathcal{B}_{\mathrm{S}}(p,\varphi) := \sum_{K\in\mathcal{T}_h}\left\{(\mathbf{S}\nabla p, \nabla\varphi)_K + \left(\left(\frac{1}{2}\nabla\cdot\mathbf{w} + r\right)p, \varphi\right)_K\right\},$$

$$\mathcal{B}_{\mathrm{A}}(p,\varphi) := \sum_{K\in\mathcal{T}_h}\left(\nabla\cdot(p\mathbf{w}) - \frac{1}{2}p\nabla\cdot\mathbf{w}, \varphi\right)_K,$$

so that

$$\mathcal{B}(p,\varphi) = \mathcal{B}_{\mathrm{S}}(p,\varphi) + \mathcal{B}_{\mathrm{A}}(p,\varphi) \qquad \forall p,\varphi \in H^1(\mathcal{T}_h), \tag{7.1}$$

$$\mathcal{B}_{\mathrm{S}}(\varphi,\varphi) = |||\varphi|||_\Omega^2 \qquad \forall \varphi \in H^1(\mathcal{T}_h), \tag{7.2}$$

$$\mathcal{B}_{\mathrm{A}}(\varphi,\varphi) \geq 0 \qquad \forall \varphi \in H^1_{\mathrm{D}}(\Omega), \tag{7.3}$$

using (2.9) and $\sum_{K\in\mathcal{T}_h}\langle\varphi^2, \mathbf{w}\cdot\mathbf{n}\rangle_{\partial K} \geq 0$ for $\varphi \in H^1_{\mathrm{D}}(\Omega)$ in the estimate.

We then have, using that $p - s \in H^1_{\mathrm{D}}(\Omega)$,

$$\begin{aligned}
|||p-s|||_\Omega^2 &\leq \mathcal{B}(p-s,p-s) = \mathcal{B}(p-\tilde{p},p-s) + \mathcal{B}(\tilde{p}-s,p-s)\\
&= \mathcal{B}_{\mathrm{S}}(\tilde{p}-s,p-s) + \mathcal{B}(p-\tilde{p},p-s) + \mathcal{B}_{\mathrm{A}}(\tilde{p}-s,p-s)\\
&\leq |||\tilde{p}-s|||_\Omega |||p-s|||_\Omega + |||p-s|||_\Omega \mathcal{B}\left(p-\tilde{p}, \frac{p-s}{|||p-s|||_\Omega}\right)\\
&\quad + |||p-s|||_\Omega \mathcal{B}_{\mathrm{A}}\left(\tilde{p}-s, \frac{p-s}{|||p-s|||_\Omega}\right),
\end{aligned}$$

employing the Cauchy–Schwarz inequality in the first term. If $|||p-\tilde{p}|||_\Omega \leq |||p-s|||_\Omega$, this concludes the proof. In general, we could use the triangle inequality $|||p-\tilde{p}|||_\Omega \leq |||p-s|||_\Omega + |||s-\tilde{p}|||_\Omega$ and the above bound for $|||p-s|||_\Omega$, but this would lead to an estimate which is not optimal (the term $|||\tilde{p}-s|||_\Omega$ would be replaced by $2|||\tilde{p}-s|||_\Omega$). We thus show below that the same bound holds true also when $|||p-s|||_\Omega \leq |||p-\tilde{p}|||_\Omega$.

We have, using (7.3) and the Cauchy–Schwarz inequality,

$$\begin{aligned}
|||p-\tilde{p}|||_\Omega^2 &= \mathcal{B}_{\mathrm{S}}(p-\tilde{p},p-\tilde{p}) = \mathcal{B}_{\mathrm{S}}(p-\tilde{p},p-s) + \mathcal{B}_{\mathrm{S}}(p-\tilde{p},s-\tilde{p})\\
&= \mathcal{B}_{\mathrm{S}}(p-\tilde{p},s-\tilde{p}) + \mathcal{B}(p-\tilde{p},p-s) - \mathcal{B}_{\mathrm{A}}(p-\tilde{p},p-s)\\
&= \mathcal{B}_{\mathrm{S}}(p-\tilde{p},s-\tilde{p}) + \mathcal{B}(p-\tilde{p},p-s) - \mathcal{B}_{\mathrm{A}}(p-s,p-s) + \mathcal{B}_{\mathrm{A}}(\tilde{p}-s,p-s)\\
&\leq \mathcal{B}_{\mathrm{S}}(p-\tilde{p},s-\tilde{p}) + \mathcal{B}(p-\tilde{p},p-s) + \mathcal{B}_{\mathrm{A}}(\tilde{p}-s,p-s)\\
&\leq |||p-\tilde{p}|||_\Omega |||s-\tilde{p}|||_\Omega + |||p-s|||_\Omega \mathcal{B}\left(p-\tilde{p}, \frac{p-s}{|||p-s|||_\Omega}\right)\\
&\quad + |||p-s|||_\Omega \mathcal{B}_{\mathrm{A}}\left(\tilde{p}-s, \frac{p-s}{|||p-s|||_\Omega}\right),
\end{aligned}$$

which, by virtue of $|||p-s|||_\Omega \leq |||p-\tilde{p}|||_\Omega$ supposed in this second case, concludes the proof. ∎

Consequently, the following bound for the error $|||p-\tilde{p}_h|||_\Omega$ holds.

LEMMA 7.2 (abstract error estimate). *Let p be the weak solution of the problem* (1.1a)–(1.1b) *given by* (2.7), *and let $s \in H^1_0(\Omega)$ be arbitrary. If $\tilde{p}_h$ is the post-processed solution of the centered mixed finite element scheme* (3.1a)–(3.1b) *given by* (4.1a)–(4.1b), *then*

$$|||p-\tilde{p}_h|||_\Omega \leq |||\tilde{p}_h - s|||_\Omega + \sup_{\varphi\in H^1_0(\Omega),\, |||\varphi|||_\Omega = 1} \big\{T_{\mathrm{R}}(\varphi) + T_{\mathrm{C}}(\varphi)\big\},$$

and if $\tilde{p}_h$ is the postprocessed solution of the upwind-weighted mixed finite element scheme (3.2a)–(3.2b), *given by* (4.1a)–(4.1b), *then*

$$|||p-\tilde{p}_h|||_\Omega \le |||\tilde{p}_h - s|||_\Omega + \sup_{\varphi\in H^1_0(\Omega),\, |||\varphi|||_\Omega=1} \big\{T_{\mathrm{R}}(\varphi)+T_{\mathrm{C}}(\varphi)+T_{\mathrm{U}}(\varphi)\big\},$$

where

$$T_{\mathrm{R}}(\varphi) := \sum_{K\in\mathcal{T}_h} \big(f+\nabla\cdot\mathbf{S}\nabla\tilde{p}_h - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h, \varphi-\varphi_K\big)_K,$$

$$T_{\mathrm{C}}(\varphi) := \sum_{K\in\mathcal{T}_h} \left(\nabla\cdot((\tilde{p}_h-s)\mathbf{w}) - \frac{1}{2}(\tilde{p}_h-s)\nabla\cdot\mathbf{w}, \varphi\right)_K,$$

$$T_{\mathrm{U}}(\varphi) := \sum_{K\in\mathcal{T}_h}\sum_{\sigma\in\mathcal{E}_K} \langle(\hat{p}_\sigma-\tilde{p}_h)\mathbf{w}\cdot\mathbf{n}, \varphi_K\rangle_\sigma,$$

and where φ_K is the mean of φ over $K\in\mathcal{T}_h$, $\varphi_K := (\varphi,1)_K/|K|$.

Proof. Let us consider an arbitrary $\varphi\in H^1_0(\Omega)$. We have, using the bilinearity of $\mathcal{B}(\cdot,\cdot)$, the definition (2.7) of the weak solution p, and the Green theorem in each $K\in\mathcal{T}_h$,

$$\begin{aligned}\mathcal{B}(p-\tilde{p}_h,\varphi) &= (f,\varphi)_\Omega - \sum_{K\in\mathcal{T}_h}\big\{(\mathbf{S}\nabla\tilde{p}_h,\nabla\varphi)_K + \big(\nabla\cdot(\tilde{p}_h\mathbf{w}),\varphi\big)_K + (r\tilde{p}_h,\varphi)_K\big\}\\ &= \sum_{K\in\mathcal{T}_h}\big\{\big(f+\nabla\cdot(\mathbf{S}\nabla\tilde{p}_h) - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h,\varphi\big)_K - \langle\mathbf{S}\nabla\tilde{p}_h\cdot\mathbf{n},\varphi\rangle_{\partial K}\big\}\\ &= \sum_{K\in\mathcal{T}_h}\big(f+\nabla\cdot(\mathbf{S}\nabla\tilde{p}_h) - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h,\varphi\big)_K.\end{aligned}$$

Note that we have, in particular, used the continuity of the normal trace of $\mathbf{S}\nabla\tilde{p}_h$ (i.e., by (4.1a), the mixed finite element continuity of the normal trace of $\mathbf{u}_h$), yielding

$$\langle(\mathbf{S}\nabla\tilde{p}_h\cdot\mathbf{n})|_K + (\mathbf{S}\nabla\tilde{p}_h\cdot\mathbf{n})|_L, \varphi\rangle_{\sigma_{K,L}} = \langle 0,\varphi\rangle_{\sigma_{K,L}} = 0 \qquad \forall\sigma_{K,L}\in\mathcal{E}_h^{\mathrm{int}}$$

(the fact that $\langle\mathbf{S}\nabla\tilde{p}_h\cdot\mathbf{n},\varphi\rangle_\sigma = 0$ for $\sigma\in\mathcal{E}_h^{\mathrm{ext}}$ follows by $\varphi\in H^1_0(\Omega)$).

Now the equation (6.1b) of the equivalent form of the centered scheme by the definition of $\tilde{p}_h$ (4.1a)–(4.1b) and by the Green theorem implies that (recall that φ_K is the constant mean of φ over K)

$$\big(f+\nabla\cdot(\mathbf{S}\nabla\tilde{p}_h) - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h,\varphi_K\big)_K = 0 \qquad \forall K\in\mathcal{T}_h. \tag{7.4}$$

Hence in the case of the centered scheme,

$$\mathcal{B}(p-\tilde{p}_h,\varphi) = \sum_{K\in\mathcal{T}_h}\big(f+\nabla\cdot(\mathbf{S}\nabla\tilde{p}_h) - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h,\varphi-\varphi_K\big)_K = T_{\mathrm{R}}(\varphi).$$

For the upwind-weighted scheme, we have

$$\mathcal{B}(p-\tilde{p}_h,\varphi) = T_{\mathrm{R}}(\varphi) + T_{\mathrm{U}}(\varphi).$$

To conclude the proof, if now suffices to use Lemma 7.1. ∎

We now estimate the terms T_{R}, T_{C}, and T_{U} separately, setting $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$ in Lemma 7.2.

LEMMA 7.3 (residual estimate). *Let $\varphi \in H_0^1(\Omega)$ be arbitrary. Then*

$$T_{\mathrm{R}}(\varphi) \le \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K} |||\varphi|||_K,$$

where $\eta_{\mathrm{R},K}$ is given by (4.2).

Proof. The Poincaré inequality (2.1) and the definition of $|||\cdot|||_K$ by (2.6) imply

$$\|\varphi - \varphi_K\|_K^2 \le C_{\mathrm{P},d} h_K^2 \|\nabla\varphi\|_K^2 \le C_{\mathrm{P},d} \frac{h_K^2}{c_{\mathbf{S},K}} |||\varphi|||_K^2 . \tag{7.5}$$

Next, the estimate

$$\|\varphi - \varphi_K\|_K^2 \le \|\varphi\|_K^2 \le \frac{1}{c_{\mathbf{w},r,K}} |||\varphi|||_K^2$$

is obvious using the definition of $|||\cdot|||_K$ by (2.6). Thus the Schwarz inequality implies

$$\begin{aligned} T_{\mathrm{R}}(\varphi) &\le \sum_{K\in\mathcal{T}_h} \left\| f + \nabla\cdot(\mathbf{S}\nabla\tilde{p}_h) - \nabla\cdot(\tilde{p}_h\mathbf{w}) - r\tilde{p}_h \right\|_K \|\varphi - \varphi_K\|_K \\ &\le \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{R},K} |||\varphi|||_K. \qquad \square \end{aligned}$$

LEMMA 7.4 (convection estimate). *Let $\varphi \in H_0^1(\Omega)$ be arbitrary. Then*

$$T_{\mathrm{C}}(\varphi) \le \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{C},K} |||\varphi|||_K,$$

where $\eta_{\mathrm{C},K}$ is given by (4.4).

Proof. Denote $v := \tilde{p}_h - \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$. Then, for each $K \in \mathcal{T}_h$,

$$\left(\nabla\cdot(v\mathbf{w}) - \frac{1}{2} v\nabla\cdot\mathbf{w}, \varphi\right)_K \le \frac{\|\nabla\cdot(v\mathbf{w}) - \frac{1}{2}v\nabla\cdot\mathbf{w}\|_K}{\sqrt{c_{\mathbf{w},r,K}}} |||\varphi|||_K .$$

Note that this estimate is valid for an arbitrary $s \in H_0^1(\Omega)$ instead of $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$.

Next, the fact that the modified Oswald interpolation operator of section 4.2 preserves the means of $\tilde{p}_h$ over the sides and that $\mathbf{w}\cdot\mathbf{n}$ is constant on all sides implies

$$(\nabla\cdot(v\mathbf{w}), \varphi_K)_K = \langle v\mathbf{w}\cdot\mathbf{n}, \varphi_K\rangle_{\partial K} = 0, \tag{7.6}$$

where again $\varphi_K := (\varphi, 1)_K/|K|$. Thus we also have an alternative estimate

$$\begin{aligned} &\left(\nabla\cdot(v\mathbf{w}) - \frac{1}{2}v\nabla\cdot\mathbf{w}, \varphi\right)_K \\ &= (\nabla v\cdot\mathbf{w}, \varphi - \varphi_K)_K + \left(\frac{1}{2}v\nabla\cdot\mathbf{w}, \varphi\right)_K - (v\nabla\cdot\mathbf{w}, \varphi_K)_K \\ &\le \frac{\sqrt{C_{\mathrm{P},d}}h_K\|\nabla v\cdot\mathbf{w}\|_K}{\sqrt{c_{\mathbf{S},K}}} \sqrt{c_{\mathbf{S},K}}\|\nabla\varphi\|_K + \frac{3\|v\nabla\cdot\mathbf{w}\|_K}{2\sqrt{c_{\mathbf{w},r,K}}} \sqrt{c_{\mathbf{w},r,K}}\|\varphi\|_K \\ &\le \left(\frac{C_{\mathrm{P},d}h_K^2\|\nabla v\cdot\mathbf{w}\|_K^2}{c_{\mathbf{S},K}} + \frac{9\|v\nabla\cdot\mathbf{w}\|_K^2}{4c_{\mathbf{w},r,K}}\right)^{\frac{1}{2}} |||\varphi|||_K, \end{aligned}$$

using the Cauchy–Schwarz inequality and the Poincaré inequality (2.1). ∎

Finally, the proof of the following lemma can be found in [38].

LEMMA 7.5 (upwinding estimate). *Let $\varphi \in H^1_0(\Omega)$ be arbitrary. Then*

$$T_{\mathrm{U}}(\varphi) \leq \sum_{K\in\mathcal{T}_h} \eta_{\mathrm{U},K} |||\varphi|||_K,$$

where $\eta_{\mathrm{U},K}$ is given by (4.6).

Lemmas 7.1–7.5 and the Cauchy–Schwarz inequality prove Theorems 4.2–4.3.

7.2. Proofs of the local efficiency of the estimates.

LEMMA 7.6 (local efficiency of the residual estimator). *Let $K \in \mathcal{T}_h$ and let $\eta_{\mathrm{R},K}$ be the residual estimator given by* (4.2). *Then* (4.10) *holds true.*

Proof. The proof follows that given in [33]. Let ψ_K be the bubble function on K, given as the product of the $d+1$ linear functions that take the value 1 at one vertex of K and vanish at the other vertices, and let us denote $v := (f+\nabla\cdot(\mathbf{S}\nabla\tilde{p}_h)-\nabla\cdot(\tilde{p}_h\mathbf{w})-r\tilde{p}_h)$ on a given $K \in \mathcal{T}_h$. Note that v is a polynomial in K by Assumption B. Then the equivalence of norms on finite-dimensional spaces, the inverse inequality (cf., e.g., [15, Theorem 3.2.6]), and the definition of $|||\cdot|||_K$ by (2.6) give

$$\begin{aligned} c\,\|v\|_K^2 &\leq (v,\psi_K v)_K,\\ \|\psi_K v\|_K &\leq \|v\|_K,\\ |||\psi_K v|||_K &\leq C \min\left\{\frac{h_K}{\sqrt{C_{\mathbf{S},K}}}, \frac{1}{\sqrt{c_{\mathbf{w},r,K}}}\right\}^{-1} \|v\|_K, \end{aligned}$$

with the constants c and C depending only on the polynomial degree k of f, d, and κ_K. Next, we immediately have (cf. the proof of Lemma 7.2)

$$\mathcal{B}(p-\tilde{p}_h,\psi_K v) = (v,\psi_K v)_K,$$

and, using (2.10),

$$\begin{aligned} \mathcal{B}(p-\tilde{p}_h,\psi_K v) \leq{}& \max\left\{1, \frac{C_{\mathbf{w},r,K}}{c_{\mathbf{w},r,K}}\right\} |||p-\tilde{p}_h|||_K |||\psi_K v|||_K\\ &+ \frac{C_{\mathbf{w},K}}{\sqrt{c_{\mathbf{S},K}}} |||p-\tilde{p}_h|||_K \|\psi_K v\|_K. \end{aligned}$$

Combining the above estimates, one comes to

$$\begin{aligned} c\|v\|_K^2 \leq{}& |||p-\tilde{p}_h|||_K \|v\|_K\\ &\cdot\left\{\max\left\{1, \frac{C_{\mathbf{w},r,K}}{c_{\mathbf{w},r,K}}\right\} C \min\left\{\frac{h_K}{\sqrt{C_{\mathbf{S},K}}}, \frac{1}{\sqrt{c_{\mathbf{w},r,K}}}\right\}^{-1} + \frac{C_{\mathbf{w},K}}{\sqrt{c_{\mathbf{S},K}}}\right\}. \end{aligned}$$

Considering the definition of $\eta_{\mathrm{R},K}$ by (4.2) and that of Pe_K and ϱ_K by (4.8) concludes the proof. ∎

LEMMA 7.7 (local efficiency of the nonconformity and velocity estimators). *Let $K \in \mathcal{T}_h$ and let $\eta_{\mathrm{NC},K}$ and $\eta_{\mathrm{C},K}$ be the nonconformity and velocity estimators given, respectively, by* (4.3) *and* (4.4). *Then* (4.11) *holds true.*

Proof. One shows easily that (with $|||\cdot|||_{*,K}$ and $|||\cdot|||_{\#,K}$ defined in section 4.4)

$$\eta_{\mathrm{NC},K}^2+\eta_{\mathrm{C},K}^2\le\min\left\{|||\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)|||_{*,K}^2,|||\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)|||_{\#,K}^2\right\}.$$

Throughout the rest of the proof, let C denote a constant depending only on d and on $\kappa_{\mathcal{T}}$, not necessarily the same at each occurrence. We first show that

(7.7)
$$|||\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)|||_{*,K}^2\le C\left(\alpha_{*,K}\sum_{\sigma;\sigma\cap K\neq\emptyset}h_\sigma^{-1}\|[\tilde p_h]\|_\sigma^2+\beta_{*,K}\sum_{\sigma;\sigma\cap K\neq\emptyset}h_\sigma\|[\tilde p_h]\|_\sigma^2\right).$$

The first part of the estimate follows directly from Lemma 4.1 and the definition of $|||\cdot|||_{*,K}$. To estimate $\beta_{*,K}\|\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)\|_K^2$, we notice that the means of $\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)$ over all sides of a simplex $K\in\mathcal{T}_h$ are by the construction of the modified Oswald interpolation operator equal to 0. Hence

$$\|\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)\|_K^2\le C_{\mathrm{F},d}h_K^2\|\nabla(\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h))\|_K^2$$

by the generalized Friedrichs inequality (2.2). The fact that h_K/h_σ for $K\cap\sigma\neq\emptyset$ depends only on $\kappa_{\mathcal{T}}$, which will be used in what follows as well, and another use of Lemma 4.1 proves the second part of the estimate.

We will next use the inequality

$$h_\sigma^{-\frac12}\|[\tilde p_h]\|_\sigma\le C\sum_{L;\sigma\in\mathcal{E}_L}\|\nabla(\tilde p_h-\varphi)\|_L$$

established in [2, Theorem 10] for $\sigma\in\mathcal{E}_h^{\mathrm{int}}$ and an arbitrary $\varphi\in H^1(\Omega)$. It generalizes easily to the case $\sigma\in\mathcal{E}_h^{\mathrm{ext}}$ and $\varphi\in H_0^1(\Omega)$. This inequality implies that

$$h_\sigma^\gamma\|[\tilde p_h]\|_\sigma^2\le C\frac{h_\sigma^{\gamma+1}}{\min_{L;\sigma\in\mathcal{E}_L}c_{\mathbf{S},L}}\sum_{L;\sigma\in\mathcal{E}_L}c_{\mathbf{S},L}\|\nabla(\tilde p_h-p)\|_L^2,\tag{7.8}$$

where we set $\gamma=-1,1$. Next, for an arbitrary $s_h\in\mathbb{P}_2(\mathcal{T}_h)\cap H_0^1(\Omega)$,

$$\begin{aligned}h_\sigma^{\frac12}\|[\tilde p_h]\|_\sigma&\le h_\sigma C\sum_{L;\sigma\in\mathcal{E}_L}\|\nabla(\tilde p_h-s_h)\|_L\le C\sum_{L;\sigma\in\mathcal{E}_L}h_L\|\nabla(\tilde p_h-s_h)\|_L\\&\le C\sum_{L;\sigma\in\mathcal{E}_L}\|\tilde p_h-s_h\|_L\le C\sum_{L;\sigma\in\mathcal{E}_L}\|\tilde p_h-p\|_L+C\sum_{L;\sigma\in\mathcal{E}_L}\|p-s_h\|_L,\end{aligned}$$

by the inverse inequality (cf. [15, Theorem 3.2.6]) and the triangle inequality. Hence

$$h_\sigma\|[\tilde p_h]\|_\sigma^2\le C\frac{1}{\min_{L;\sigma\in\mathcal{E}_L}c_{\mathbf{w},r,L}}\sum_{L;\sigma\in\mathcal{E}_L}c_{\mathbf{w},r,L}\|\tilde p_h-p\|_L^2+C\sum_{L;\sigma\in\mathcal{E}_L}\|p-s_h\|_L^2\tag{7.9}$$

holds as well, which gives a sense when all $c_{\mathbf{w},r,L}$ for L such that $\sigma\in\mathcal{E}_L$ are nonzero. Combining estimates (7.7)–(7.9) while estimating $\min_{L;\sigma\in\mathcal{E}_L}c_L$ for a side σ such that $\sigma\cap K\neq\emptyset$ from below by $\min_{L;L\cap K\neq\emptyset}c_L$ concludes the proof for $|||\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)|||_{*,K}$. The proof for $|||\tilde p_h-\mathcal{I}_{\mathrm{MO}}(\tilde p_h)|||_{\#,K}$ is completely similar. ∎

LEMMA 7.8 ((non)efficiency of the upwinding estimator). *Let $K\in\mathcal{T}_h$ and let $\eta_{\mathrm{U},K}$ be the upwinding estimator given by* (4.6). *Then* (4.12) *holds true.*

Proof. Let $K \in \mathcal{T}_h$, $\varphi \in H^1(K)$, and $\varphi_\sigma := \langle \varphi, 1\rangle_\sigma/|\sigma|$. Let us set $\tilde{\varphi} := \varphi - \varphi_\sigma$ and $\tilde{\varphi}_K := (\tilde{\varphi}, 1)_K/|K|$. We now note that $\tilde{\varphi}_\sigma := \langle \tilde{\varphi}, 1\rangle_\sigma/|\sigma| = 0$ and that $\nabla\tilde{\varphi} = \nabla\varphi$, which allows us to estimate

$$\|\varphi_K - \varphi_\sigma\|_\sigma^2 = \tilde{\varphi}_K^2|\sigma| \le \frac{|\sigma|}{|K|}\|\tilde{\varphi}\|_K^2 \le C_{\mathrm{F},d}\frac{|\sigma|h_K^2}{|K|}\|\nabla\varphi\|_K^2,$$

employing the generalized Friedrichs inequality (2.2). Now using the definition of $\hat{p}_\sigma$ for $\sigma \in \mathcal{E}_h^{\mathrm{int}}$ by (3.3), the fact that $0 \le \nu_\sigma \le 1/2$, (4.1b), and the above estimate,

$$\begin{aligned}\|\hat{p}_\sigma - \tilde{p}_\sigma\|_\sigma &= \|(1-\nu_\sigma)(p_K - \tilde{p}_\sigma) + \nu_\sigma(p_L - \tilde{p}_\sigma)\|_\sigma \\ &\le \max_{M;\sigma\in\mathcal{E}_M}\left\{\frac{C_{\mathrm{F},d}|\sigma|h_M^2}{|M|}\right\}^{\frac{1}{2}}(\|\nabla\tilde{p}_h\|_K + \|\nabla\tilde{p}_h\|_L)\end{aligned}$$

for suitable denotation K, L of the two elements sharing σ. For $\sigma \in \mathcal{E}_h^{\mathrm{ext}}$, a similar estimate holds. The assertion of the lemma follows by using the above estimate, (4.5), (4.6), the definition of κ_K, the estimate $|\sigma| \le h_K^{d-1}/(d-1)$, the Cauchy–Schwarz inequality, and estimating the term $\sum_{K\in\mathcal{T}_h} c_{\mathbf{S},K}\|\nabla\tilde{p}_h\|_K^2$ using Lemma 6.5. ∎

Lemmas 7.6–7.8 together prove Theorem 4.4.

8. Numerical experiments. We test our a posteriori error estimates on two model problems in this section. The first problem contains a strongly inhomogeneous diffusion-dispersion tensor, and the second one is convection-dominated; in both cases, the analytical solution is known. Estimators for inhomogeneous Dirichlet (and Neumann) boundary conditions are adapted from [38].

8.1. Model problem with strongly inhomogeneous diffusion-dispersion tensor. This model problem is taken from [30, 18] and is motivated by the fact that in real-life applications, the diffusion-dispersion tensor $\mathbf{S}$ may be discontinuous and strongly inhomogeneous. We consider in particular $\Omega = (-1,1)\times(-1,1)$ and (1.1a) with $\mathbf{w} = 0$, $r = 0$, and $f = 0$. We suppose that Ω is divided into four subdomains Ω_i corresponding to the axis quadrants (in the counterclockwise direction) and that $\mathbf{S}$ is constant and equal to $s_i\,Id$ in Ω_i. Under such conditions, an analytical solution writing

$$p(r,\theta) = r^\alpha(a_i\sin(\alpha\theta) + b_i\cos(\alpha\theta))$$

in each Ω_i can be found. Here (r,θ) are the polar coordinates in Ω, a_i and b_i are constants depending on Ω_i, and α is a parameter. This solution is continuous across the interfaces, but only the normal component of its flux $\mathbf{u} = -\mathbf{S}\nabla p$ is continuous; it finally exhibits a singularity at the origin. We assume Dirichlet boundary conditions given by this solution and consider two sets of the coefficients, with $s_1 = s_3 = 5$, $s_2 = s_4 = 1$ in the first case and $s_1 = s_3 = 100$, $s_2 = s_4 = 1$ in the second one:

$\alpha = 0.53544095$		$\alpha = 0.12690207$	
$a_1 = 0.44721360$	$b_1 = 1$	$a_1 = 0.1$	$b_1 = 1$
$a_2 = -0.74535599$	$b_2 = 2.33333333$	$a_2 = -9.60396040$	$b_2 = 2.96039604$
$a_3 = -0.94411759$	$b_3 = 0.55555556$	$a_3 = -0.48035487$	$b_3 = -0.88275659$
$a_4 = -2.40170264$	$b_4 = -0.48148148$	$a_4 = 7.70156488$	$b_4 = -6.45646175$

The original grid consisted of 24 right-angled triangles, and we have refined it either uniformly (up to five refinements) or adaptively on the basis of our estimator.

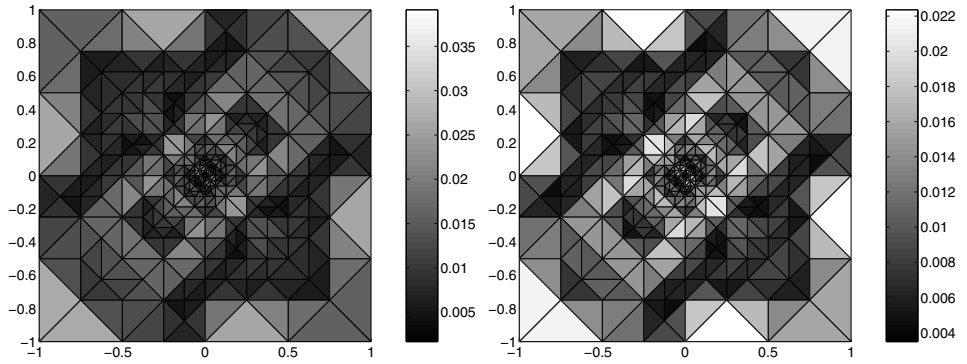


FIG. 8.1. *Estimated (left) and actual (right) error distribution,* $\alpha = 0.53544095$ *(the maximum is attained at the origin).*

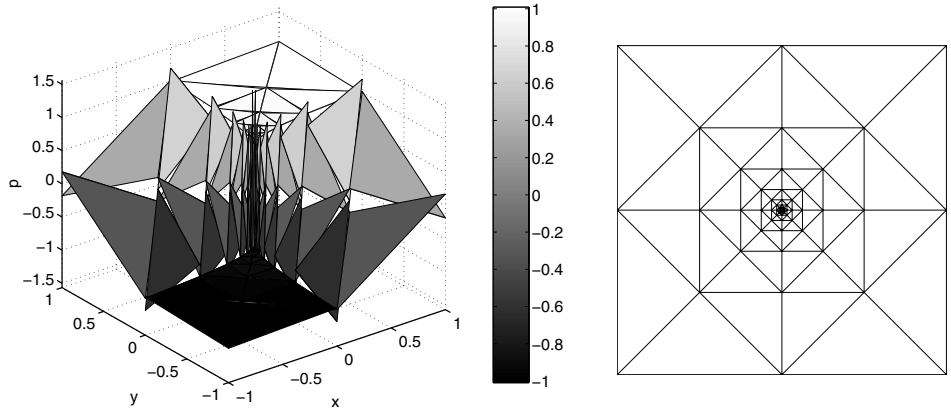


FIG. 8.2. *Approximate solution and the corresponding adaptively refined mesh,* $\alpha = 0.12690207$.

In the latter case, we refine each element where the estimated $|||\cdot|||_\Omega$-error is greater than the half of the maximum of the estimators regularly into four subelements and then use the "longest edge" refinement to recover an admissible mesh. In the given case, the residual estimators $\eta_{\mathrm{R},K}$ of (5.2) are zero for each $K \in \mathcal{T}_h$, and hence the a posteriori error estimate is entirely given by the nonconformity estimators $\eta_{\mathrm{NC},K}$ in (5.3). We have done numerical experiments with two choices, $s = \mathcal{I}_{\mathrm{Os}}(\tilde{p}_h)$ and $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$, and present the results with the first one, which gives a slightly better efficiency.

We can see in Figure 8.1 that the predicted error distribution on an adaptively refined mesh for the first test case is excellent. In particular, even if the solution is smoother, the singularity is well recognized. Next, Figure 8.2 gives an example of the approximate solution on an adaptively refined mesh and this mesh in the second test case. Here, the singularity is much more important, and consequently the grid is highly refined around the origin (for 1800 triangles, the diameter of the smallest ones is 10^{-16}, and 73% of them are contained in the circle of radius 0.1). Figure 8.3 then reports the estimated and actual errors of the numerical solutions on uniformly/adaptively refined grids in the two test cases. The energy norm (2.6) was approximated with a 7-point quadrature formula in each triangle. It can be seen

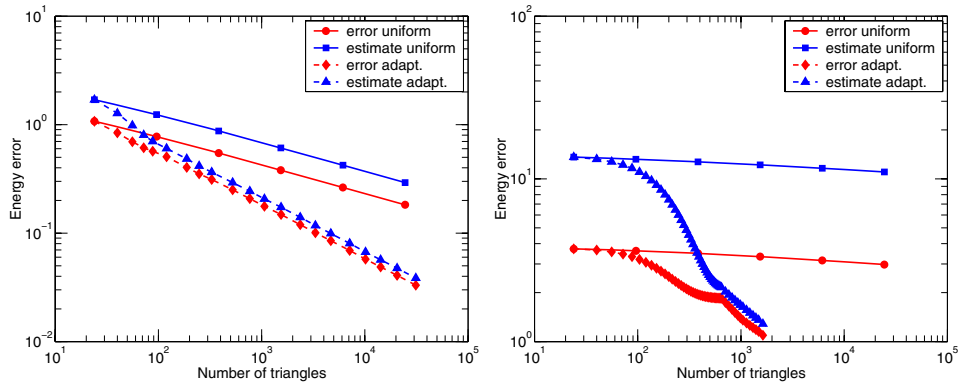


FIG. 8.3. *Estimated and actual error against the number of elements in uniformly/adaptively refined meshes for* $\alpha = 0.53544095$ *(left) and* $\alpha = 0.12690207$ *(right).*

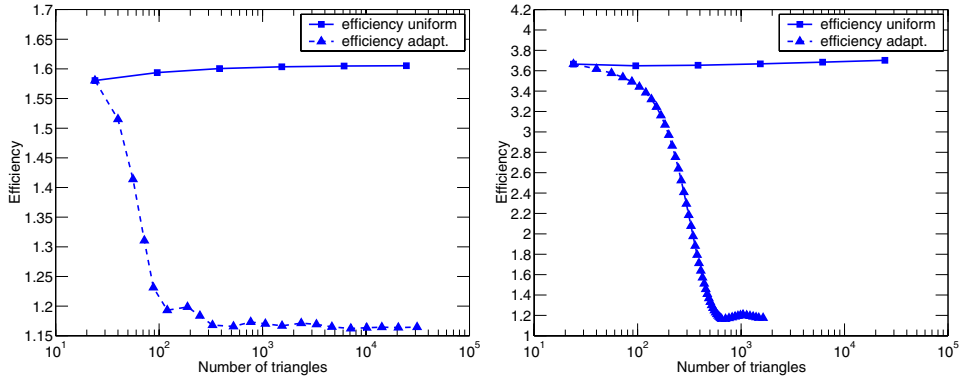


FIG. 8.4. *Overall efficiency of the a posteriori error estimates against the number of elements in uniformly/adaptively refined meshes for* $\alpha = 0.53544095$ *(left) and* $\alpha = 0.12690207$ *(right).*

from these plots that one can substantially reduce the number of unknowns necessary to attain the prescribed precision using the derived a posteriori error estimates and adaptively refined grids. Finally, Figure 8.4 gives the efficiency plots for the two cases, i.e., the ratio of the estimated $|||\cdot|||_\Omega$-error to the actual $|||\cdot|||_\Omega$-error. This quantity simply expresses how many times we have overestimated the error—recall that there are no undetermined multiplicative constants in our estimates. These plots confirm the theoretical results of section 5.3. Even while only using $\mathcal{I}_{\mathrm{Os}}(\tilde{p}_h)$ instead of evaluating the infimum in (5.4), (approximate) asymptotic exactness and robustness with respect to inhomogeneities is confirmed.

8.2. Convection-dominated model problem. This problem is a modification of a problem considered in [20]. We set $\Omega = (0,1) \times (0,1)$, $\mathbf{w} = (0,1)$, and $r = 1$ in (1.1a) and consider three cases with $\mathbf{S} = \varepsilon\, Id$ and ε equal to, respectively, 1, 10^{-2}, and 10^{-4}. The right-hand-side term f, Neumann boundary conditions on the upper side, and Dirichlet boundary conditions elsewhere are chosen so that

$$p(x,y) = 0.5\left(1 - \tanh\left(\frac{0.5 - x}{a}\right)\right)$$

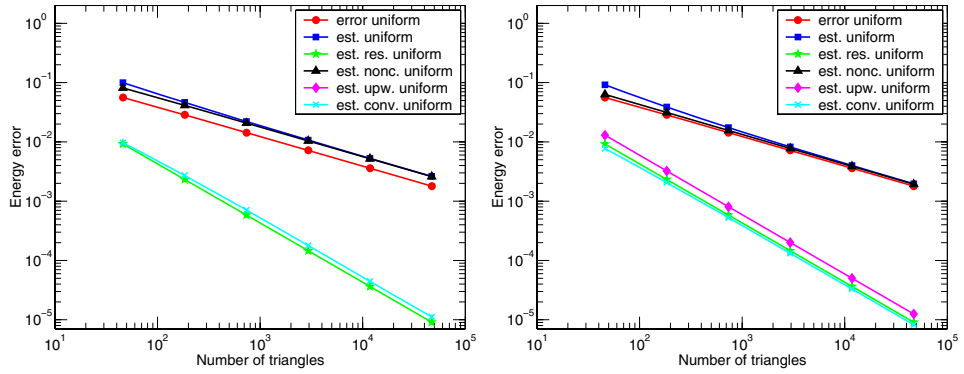


FIG. 8.5. *Estimated and actual error using* $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$ *(left) and* $s = \mathcal{I}_{\mathrm{Os}}(\tilde{p}_h)$ *(right) against the number of elements,* $\varepsilon = 1$, $a = 0.5$.

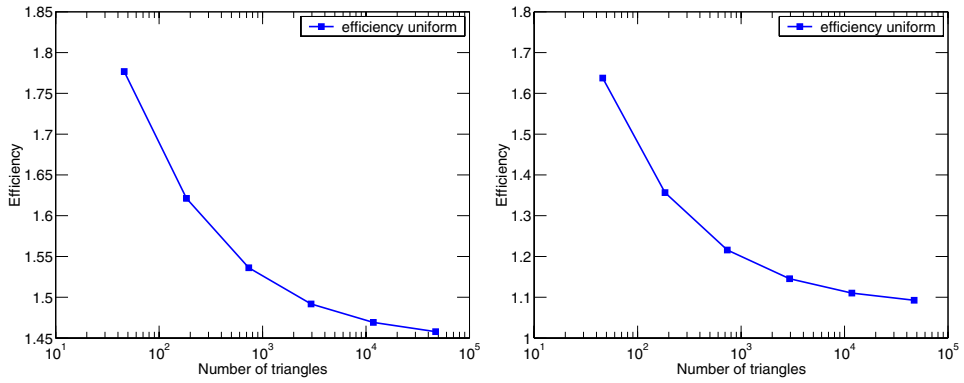


FIG. 8.6. *Overall efficiency using* $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$ *(left) and* $s = \mathcal{I}_{\mathrm{Os}}(\tilde{p}_h)$ *(right) against the number of elements,* $\varepsilon = 1$, $a = 0.5$.

was the exact solution. It is, in fact, one-dimensional and possesses an internal layer of width a which we set, respectively, equal to 0.5, 0.05, and 0.02. We start the computations from an unstructured grid of Ω consisting of 46 triangles and refine it either uniformly (up to five refinements) or adaptively. We use the scheme described in section 5.5.

We first compare, for $\varepsilon = 1$ and $a = 0.5$, the estimates with $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$ as proposed in section 4.3 and a modification with $s = \mathcal{I}_{\mathrm{Os}}(\tilde{p}_h)$, corresponding to the approach chosen in [38, 37], on uniformly refined grids. In the latter case, we no longer have the important property (7.6), and consequently there is an additional term which we associate with the upwinding estimator; it, however, turns out to be of higher order; see Figure 8.5. Note that the (approximate) asymptotic exactness observed in Figure 8.6 is in full correspondence with the theoretical considerations of section 5.3.2. In this case, $s = \mathcal{I}_{\mathrm{Os}}(\tilde{p}_h)$ gives a slightly better efficiency. In the following examples, however, we use $s = \mathcal{I}_{\mathrm{MO}}(\tilde{p}_h)$, since it turns out to be the better choice.

For $\varepsilon = 10^{-2}$ and $a = 0.05$ (convection-dominated regime on coarse meshes and diffusion-dominated regime with progressive refinement), still the distribution of the error is predicted very well; cf. Figure 8.7. Note in particular the correct localization of

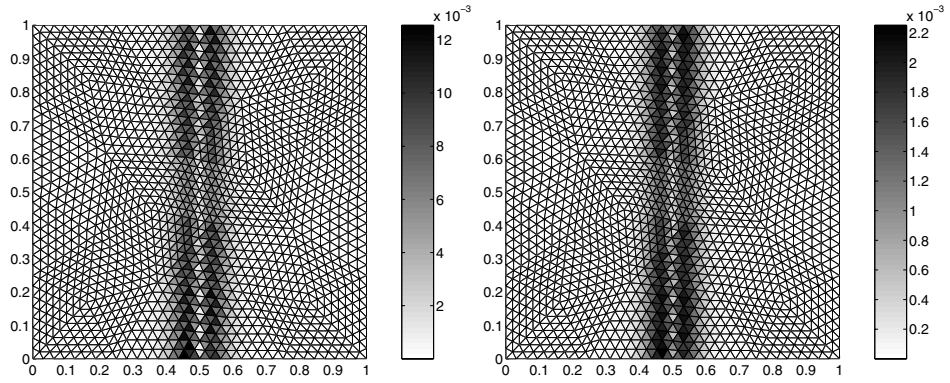


FIG. 8.7. *Estimated (left) and actual (right) error distribution,* $\varepsilon = 10^{-2}$, $a = 0.05$.

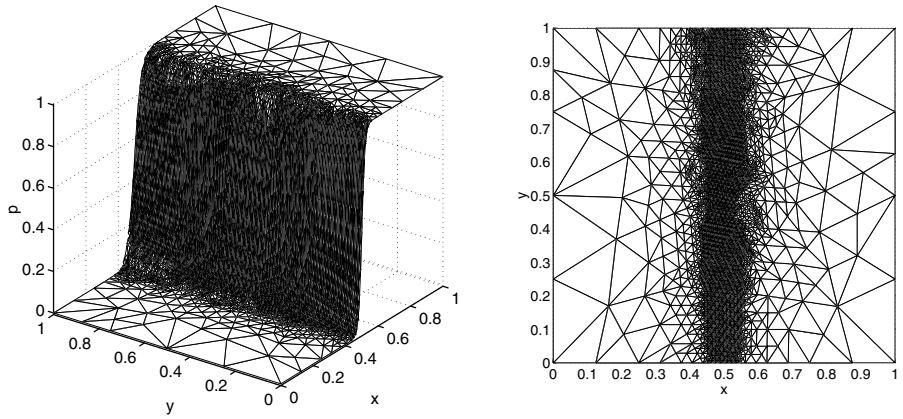


FIG. 8.8. *Approximate solution and the corresponding adaptively refined mesh,* $\varepsilon = 10^{-4}$, $a = 0.02$.

the error away from the center of the shock, as well as the sensitivity of our estimator to the shape of the elements. Next, an example of an adaptively refined mesh and of the corresponding solution for $\varepsilon = 10^{-4}$ and $a = 0.02$ is given in Figure 8.8. For these two test cases, we have used as a refinement criterion 0.2- and 0.05-times the maximum of the estimators, respectively. The estimated and actual errors are plotted against the number of elements in uniformly/adaptively refined meshes in Figure 8.9. Again, one can see that we can substantially reduce the number of unknowns necessary to attain the prescribed precision using the derived estimators and adaptively refined grids. Finally, the efficiency plots are given in Figure 8.10. In the first case, the efficiency is almost optimal for finest grids, whereas in the second one, only the elements in the refined shock region start to leave the convection-dominated regime, and thus the efficiency starts to decrease.

Acknowledgments. The author would like to thank Prof. Alexandre Ern from the CERMICS laboratory of the Ecole Nationale des Ponts et Chaussées, Marne la Vallée, France, for pointing out the compact form of the proof of Lemma 7.1.

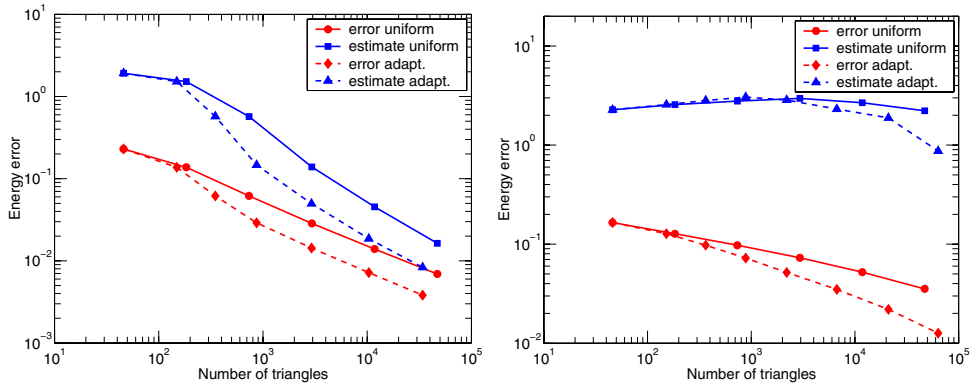


FIG. 8.9. *Estimated and actual error against the number of elements in uniformly/adaptively refined meshes for* $\varepsilon = 10^{-2}$, $a = 0.05$ *(left) and* $\varepsilon = 10^{-4}$, $a = 0.02$ *(right).*

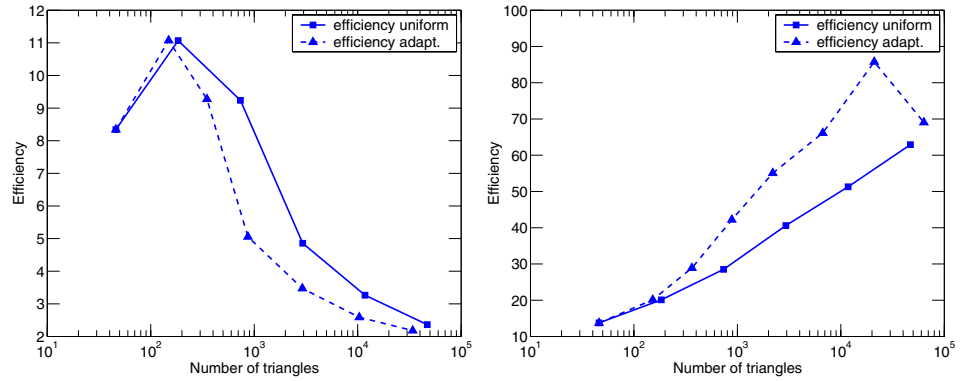


FIG. 8.10. *Overall efficiency of the a posteriori error estimates against the number of elements in uniformly/adaptively refined meshes for* $\varepsilon = 10^{-2}$, $a = 0.05$ *(left) and* $\varepsilon = 10^{-4}$, $a = 0.02$ *(right).*

REFERENCES

[1] B. ACHCHAB, A. AGOUZAL, J. BARANGER, AND J.-F. MAÎTRE, *Estimateur d'erreur a posteriori hiérarchique. Application aux éléments finis mixtes*, Numer. Math., 80 (1998), pp. 159–179.

[2] Y. ACHDOU, C. BERNARDI, AND F. COQUEL, *A priori and a posteriori analysis of finite volume discretizations of Darcy's equations*, Numer. Math., 96 (2003), pp. 17–42.

[3] M. AINSWORTH, *A synthesis of a posteriori error estimation techniques for conforming, nonconforming and discontinuous Galerkin finite element methods*, in Recent Advances in Adaptive Computation, Contemp. Math. 383, AMS, Providence, RI, 2005, pp. 1–14.

[4] M. AINSWORTH, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[5] A. ALONSO, *Error estimators for a mixed method*, Numer. Math., 74 (1996), pp. 385–395.

[6] D. N. ARNOLD AND F. BREZZI, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[7] I. BABUŠKA AND W. C. RHEINBOLDT, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal., 15 (1978), pp. 736–754.

[8] M. BEBENDORF, *A note on the Poincaré inequality for convex domains*, Z. Anal. Anwend., 22 (2003), pp. 751–756.

[9] C. BERNARDI AND R. VERFÜRTH, *Adaptive finite element methods for elliptic equations with non-smooth coefficients*, Numer. Math., 85 (2000), pp. 579–608.

[10] D. BRAESS AND R. VERFÜRTH, *A posteriori error estimators for the Raviart–Thomas element*, SIAM J. Numer. Anal., 33 (1996), pp. 2431–2444.

[11] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), pp. 217–235.
[12] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer-Verlag, New York, 1991.
[13] C. Carstensen, *A posteriori error estimate for the mixed finite element method*, Math. Comp., 66 (1997), pp. 465–476.
[14] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids.* I. *Low order conforming, nonconforming, and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.
[15] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.
[16] C. Dawson, *Analysis of an upwind-mixed finite element method for nonlinear contaminant transport equations*, SIAM J. Numer. Anal., 35 (1998), pp. 1709–1724.
[17] J. Douglas, Jr., and J. E. Roberts, *Global estimates for mixed methods for second order elliptic equations*, Math. Comp., 44 (1985), pp. 39–52.
[18] G. T. Eigestad and R. A. Klausen, *On the convergence of the multi-point flux approximation O-method: Numerical experiments for discontinuous permeability*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 1079–1098.
[19] L. El Alaoui and A. Ern, *Residual and hierarchical a posteriori error estimates for nonconforming mixed finite element methods*, M2AN Math. Model. Numer. Anal., 38 (2004), pp. 903–929.
[20] L. El Alaoui, A. Ern, and E. Burman, *A priori and a posteriori analysis of non-conforming finite elements with face penalty for advection-diffusion equations*, IMA J. Numer. Anal., 27 (2007), pp. 151–171.
[21] R. Ewing, O. Iliev, and R. Lazarov, *A modified finite volume approximation of second-order elliptic equations with discontinuous coefficients*, SIAM J. Sci. Comput., 23 (2001), pp. 1335–1351.
[22] R. H. W. Hoppe and B. Wohlmuth, *Adaptive multilevel techniques for mixed finite element discretizations of elliptic boundary value problems*, SIAM J. Numer. Anal., 34 (1997), pp. 1658–1681.
[23] O. A. Karakashian and F. Pascal, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.
[24] R. Kirby, *Residual a posteriori error estimates for the mixed finite element method*, Comput. Geosci., 7 (2003), pp. 197–214.
[25] C. Lovadina and R. Stenberg, *Energy norm a posteriori error estimates for mixed finite element methods*, Math. Comp., 75 (2006), pp. 1659–1674.
[26] M. Ohlberger, *A posteriori error estimates for vertex centered finite volume approximations of convection-diffusion-reaction equations*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 355–387.
[27] L. E. Payne and H. F. Weinberger, *An optimal Poincaré inequality for convex domains*, Arch. Ration. Mech. Anal., 5 (1960), pp. 286–292.
[28] M. Petzoldt, *A posteriori error estimators for elliptic equations with discontinuous coefficients*, Adv. Comput. Math., 16 (2002), pp. 47–75.
[29] P.-A. Raviart and J.-M. Thomas, *A mixed finite element method for* 2*nd order elliptic problems*, in Mathematical Aspects of Finite Element Methods (Proceedings of the Conference of the C.N.R., Rome, 1975), Lecture Notes in Math. 606, Springer, Berlin, 1977, pp. 292–315.
[30] B. Rivière and M. F. Wheeler, *A posteriori error estimates for a discontinuous Galerkin method applied to elliptic problems*, Comput. Math. Appl., 46 (2003), pp. 141–163.
[31] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North–Holland, Amsterdam, 1991, pp. 523–639.
[32] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, Teubner Wiley, Stuttgart, 1996.
[33] R. Verfürth, *A posteriori error estimators for convection-diffusion equations*, Numer. Math., 80 (1998), pp. 641–663.
[34] R. Verfürth, *Robust a posteriori error estimates for stationary convection-diffusion equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1766–1782.
[35] M. Vohralík, *On the discrete Poincaré–Friedrichs inequalities for nonconforming approximations of the Sobolev space* H^1, Numer. Funct. Anal. Optim., 26 (2005), pp. 925–952.
[36] M. Vohralík, *Equivalence between lowest-order mixed finite element and multi-point finite volume methods on simplicial meshes*, M2AN Math. Model. Numer. Anal., 40 (2006), pp. 367–391.

[37] M. Vohralík, *A posteriori error estimates for finite volume and mixed finite element discretizations of convection–diffusion–reaction equations*, ESAIM Proc., (2007), to appear.
[38] M. Vohralík, *Residual flux-based a posteriori error estimates for finite volume discretizations of inhomogeneous, anisotropic, and convection-dominated problems*, Numer. Math., submitted.
[39] M. F. Wheeler and I. Yotov, *A posteriori error estimates for the mortar mixed finite element method*, SIAM J. Numer. Anal., 43 (2005), pp. 1021–1042.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1600–1621

ERROR ANALYSIS OF IMEX RUNGE–KUTTA METHODS DERIVED FROM DIFFERENTIAL-ALGEBRAIC SYSTEMS*

SEBASTIANO BOSCARINO†

Abstract. In this paper we present an error analysis of the IMEX Runge–Kutta methods when applied to stiff problems containing a nonstiff term and a stiff term, characterized by a small stiffness parameter ε. In this analysis we expand the global error in powers of ε and show that the coefficients of the error are the global errors of the IMEX Runge–Kutta method applied to a differential-algebraic system. Interesting convergence results of these errors and of the remainder of the expansion allow us to determine sharp error bounds for stiff problems. As a representative example of stiff problems we have chosen the van der Pol equation. We illustrate that the theoretical prediction is confirmed by the numerical test. Specifically, an order reduction phenomenon is observed when the problem becomes increasingly stiff. In particular, making several assumptions, we try to improve global error estimates of several IMEX Runge–Kutta methods existing in the literature.

Key words. Runge–Kutta methods, stiff problems, differential-algebraic systems

AMS subject classification. 34E05, 65L06, 65L80

DOI. 10.1137/060656929

1. Introduction.

Several physical phenomena of great importance for applications are described by stiff systems of differential equations in the form

$$U' = F(U) + \frac{1}{\varepsilon}G(U), \tag{1}$$

where $U = U(t) \in R^m$, F,$G : R^m \to R^m$, and $\varepsilon > 0$ is the stiffness parameter.

Systems of such form, with a large number of equations, often arise from the discretization of partial differential equations, such as convection-diffusion problems and hyperbolic systems with relaxation (i.e., discrete kinetic theory of rarefied gases, hydrodynamical models for semiconductors, etc., see [8], [17], [19], [18], [15], [6], [9]), where a method of lines approach is usually used.

In order to be able to treat problems of the form (1), it is important to develop suitable numerical schemes that work in an accurate and efficient way. A general approach to the solution of problem (1) is based on implicit-explicit (IMEX) multistep methods [14], [10], [3] or IMEX Runge–Kutta (R-K) methods [8], [17], [19], [18], [1], [2].

We consider here IMEX R-K methods. An IMEX R-K method consists of applying an implicit discretization for G and an explicit one for F. In general, in order to guarantee simplicity and efficiency in solving the algebraic equations corresponding to the implicit part of the discretization at each step of problem (1), we will consider diagonally implicit R-K (DIRK) methods.

In this paper we show that most of the popular IMEX R-K methods presented in the literature suffer from the phenomenon of order reduction in the stiff regime

*Received by the editors April 10, 2006; accepted for publication (in revised form) February 22, 2007; published electronically August 15, 2007. This research was partially supported by INDAM project "Metodi numerici per lo studio di problemi evolutivi multiscala" and Italian PRIN 2004 project Prot. 2004014411_007.

http://www.siam.org/journals/sinum/45-4/65692.html

†Department of Mathematics and Computer Science, University of Catania, viale A. Doria 6, 95125 Catania, Italy (boscarino@dmi.unict.it).

($\Delta t \gg \varepsilon$) when the classical order is greater than two [8], [17], [18], [1]. To this aim, we investigate this phenomenon and give an answer through a theoretical error analysis using typical techniques of differential-algebraic equations (DAEs) [12], [13], [7], [11].

We observe that system (1) can be written as a system of $2m$ equations in the form

$$\begin{aligned} y' &= f(y,z), \\ \varepsilon z' &= g(y,z) \end{aligned} \tag{2}$$

once we set $U = y+z$, $F(U) = f(y,z)$, and $G(U) = g(y,z)$. On the other hand, system (2) is a particular case of system (1) when $F(U) = (f(y,z),0)$, $G(U) = (0, g(y,z))$. Now, restricting our attention to system (2), such a problem is called a *singular perturbation problem* (SPP). Classical books on this subject are [20] and [16]. These SPPs give us the possibility of studying the dependence of the global error of IMEX R-K methods on the stiffness parameter ε. Then in system (2) we suppose that $0 < \varepsilon \ll 1$ and the functions f and g are sufficiently differentiable, with f, g and the initial values $y(0)$, $z(0)$ that may depend smoothly on ε. For simplicity of notation we suppress this dependence.

When the parameter ε in system (2) is small, the corresponding differential equation is stiff, and when ε tends to zero, the differential equation becomes differential algebraic. A sequence of differential-algebraic systems arises in the study of SPPs. Our analysis is based on the assumption of a smooth solution of system (2) and applies to the stiff case ($\Delta t \gg \varepsilon$).

The paper is organized as follows. In the next section we introduce a description and classification of the different types of IMEX R-K methods present in the literature, based on the structure of the matrix of the implicit part. In section 3 we state our main results, presenting convergence proofs which give sharp error bounds for such methods. On the van der Pol equation, moreover, we provide numerical confirmation of the theoretical analysis and compare the performances of several types of IMEX R-K schemes. Also, these numerical results suggest how we can improve error estimates of some IMEX R-K methods through straightforward assumptions. In section 4 we consider the asymptotic expansion of the exact and numerical solution in terms of the stiffness parameter ε. Sections 5 and 6 are devoted to examining the results obtained when we apply IMEX R-K methods to DAEs of index 1 (zeroth-order expansion) and higher (higher-order expansion). In particular, in section 7 we estimate the remainder of the expansion. Finally, in section 8, conclusions are drawn and work in progress is mentioned.

2. Description and classification of IMEX R-K methods. We consider an IMEX R-K method applied to system (2),

$$\begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \sum_{i=1}^{s} \begin{pmatrix} \tilde{b}_i k_{ni} \\ b_i \ell_{ni} \end{pmatrix}, \tag{3}$$

where

$$\begin{pmatrix} k_{ni} \\ \varepsilon \ell_{ni} \end{pmatrix} = \begin{pmatrix} f(Y_{ni}, Z_{ni}) \\ g(Y_{ni}, Z_{ni}) \end{pmatrix} \tag{4}$$

and the internal stages are given by

$$\begin{pmatrix} Y_{ni} \\ Z_{ni} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} k_{nj} \\ \sum_{j=1}^{i} a_{ij} \ell_{nj} \end{pmatrix}. \tag{5}$$

The matrices $(\tilde{a}_{ij})$, with $\tilde{a}_{ij} = 0$ for $j \geq i$, and (a_{ij}) are $s \times s$ matrices such that the resulting method is explicit in f and implicit in g. We use a diagonally implicit scheme for g, i.e., $a_{ij} = 0$ for $j > i$. This will guarantee that f is always evaluated explicitly.

Such methods are characterized by the coefficient matrices $\tilde{A} = (\tilde{a}_{ij})$, $A = (a_{ij})$ and vectors $\tilde{c} = (\tilde{c}_1, \ldots, \tilde{c}_s)^T$, $\tilde{b} = (\tilde{b}_1, \ldots, \tilde{b}_s)^T$, $c = (c_1, \ldots, c_s)^T$, $b = (b_1, \ldots, b_s)^T$. They can be represented by a double *tableau* in the usual Butcher notation,

$$\begin{array}{c|c} \tilde{c} & \tilde{A} \\ \hline & \tilde{b}^T \end{array} \qquad \begin{array}{c|c} c & A \\ \hline & b^T \end{array}.$$

The coefficients $\tilde{c}$ and c are given by the usual relation,

$$\tilde{c}_i = \sum_{j=1}^{i-1} \tilde{a}_{ij}, \quad c_i = \sum_{j=1}^{i} a_{ij}, \tag{6}$$

which allows the results of our analysis to be extended to nonautonomous systems. We shall use the notation Name(s, σ, p), where this triplet characterizes the number s of the stages of the implicit scheme, the number σ of stages of the explicit scheme and the combined order of the method, p. Now we give some definitions that we will use later.

DEFINITION 2.1. *We call q_i the stage order of the ith stage of an R-K method if and only if for a problem $\dot{y}(t) = f(t, y(t))$, with $0 \leq t \leq T$ and f a smooth function, the intermediate local errors $y(t_n + c_i h) - Y_i = \mathcal{O}(h^{q_i+1})$, where $Y_i = y(t_n) + h \sum_{j=1}^{s} a_{ij} f(t_n + c_j h, Y_j)$ $(1 \leq i \leq s)$.*

Remark. For stiff differential equations the *stage order* q is an essential ingredient. It is defined by the condition $C(q)$ (see [12] and [13, sect. IV.5]), i.e.,

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k} \quad \text{for } k = 1, \ldots, q \quad \text{and all } \ i. \tag{7}$$

For an s-stage DIRK method, the stage order is 1.

DEFINITION 2.2. *Methods that satisfy the condition $a_{sj} = b_j$, $j = 1, \ldots, s$, are called stiffly accurate.*

Remark. In our analysis we indicate $R(\infty) = \lim_{z \to \infty} R(z)$, with $R(z)$ *the stability function* of the implicit scheme, defined by $R(z) = 1 + z b^T (I - zA)^{-1} \mathbf{1}$ (see [13, sect. IV.3]), with $b^T = (b_1, \ldots, b_s)$ and $\mathbf{1} = (1, \ldots, 1)^T$. From the expression of $R(z)$ follows $R(\infty) = 1 - \sum_{i,j=1}^{s} b_i \omega_{ij}$ with ω_{ij} elements of the inverse of (a_{ij}). Moreover, if the implicit method is stiffly accurate and the matrix A is invertible, one has always $R(\infty) = 0$. We shall use the notation $\tilde{q}_i$, with $\tilde{q}_i \geq 1$, to indicate the *stage order* of the ith stage of the explicit part of the IMEX R-K method, and with q_i, $q_i \geq 1$, the *stage order* of the ith stage of the implicit one. IMEX R-K methods present in the literature can be classified in three different types characterized by the structure of the matrix $A = (a_{ij})_{i,j=1}^{s}$ of the implicit scheme.

DEFINITION 2.3. *We call an IMEX R-K method* type A *(see* [18]*) if the matrix $A \in R^{s \times s}$ is invertible.*

DEFINITION 2.4. *We call an IMEX R-K method* type CK *(see* [8]*) if the matrix $A \in R^{s \times s}$ can be written as*

$$A = \begin{pmatrix} 0 & 0 \\ a & \hat{A} \end{pmatrix}$$

with the submatrix $\hat{A} \in R^{(s-1)\times(s-1)}$ *invertible.*

Remark. IMEX R-K methods of type ARS (see [1]) are a special case of type CK with the vector $a = 0$.

3. Main results. Motivated by the procedure first suggested by Hairer, Lubich, and Roche [12] (see also [13]), we extend this analysis to different types of IMEX R-K methods. The main results of this paper are summarized in this section in the form of theorems. The aim of these theorems is to present convergence results of these methods when applied to SPP (2). We suppose that the initial values lie on a suitable manifold that allows smooth solutions even in the limit of infinite stiffness and the step size $h = \Delta t \gg \varepsilon$. In fact, arbitrary initial values introduce in the solution a fast transient. One possible way to overcome this difficulty is simply to ensure that the numerical method resolves the transient phase by taking time step $h \ll \varepsilon$ in the first few steps. Then the following results are obtained assuming that the transient phase is over.

An essential ingredient to obtaining these results is to assume that the system is dissipative. More precisely, we assume that

$$\mu\left(g_z(y,z)\right) \le -1 \tag{8}$$

in an ε-independent neighborhood of the solution, where μ denotes the logarithmic norm with respect to some inner product. Condition (8) guarantees the existence of an ε-expansion of problem (2) (see [13, p. 390]).

The proof of the theorems below will be a consequence of the results of sections 5 to 7. We start by considering the limit case $\varepsilon = 0$ (*the reduced problem* or *problems of index* 1) for problem (2).

THEOREM 3.1 (type A). *Consider the stiff problem* (2), (8) *with initial values* $y(0)$, $z(0)$ *admitting a smooth solution. Apply the type-A IMEX R-K method* (3)–(5) *and let* p *be the order of explicit scheme. Assume that the method with coefficients* b_i *and* a_{ij} *is A-stable, that the stability function satisfies* $|R(\infty)| < 1$, *and that* $a_{ii} > 0$ *for all* i. *Furthermore, assume that the weights satisfy the condition* $\tilde{b}_i = b_i$ *for* $i = 1, \dots, s$.

Then if $\sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1$, *with* ω_{ij} *elements of the inverse matrix of* A, *for any fixed constant* $C > 0$, *the global error satisfies*

$$y_n - y(t_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h^2), \quad z_n - z(t_n) = \mathcal{O}(h^2)$$

for $\varepsilon \le Ch$; *otherwise, we obtain*

$$y_n - y(t_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h), \quad z_n - z(t_n) = \mathcal{O}(h).$$

If in addition $a_{si} = b_i$ *and* $\tilde{a}_{si} = \tilde{b}_i$ *for all* i, *we have* $z_n - z(x_n) = \mathcal{O}(h^p) + \mathcal{O}(\varepsilon h^2)$. *The estimates hold uniformly for* $h \le h_0$ *and* $nh \le$ Const.

THEOREM 3.2 (type CK). *Consider the stiff problem* (2), (8) *with initial values* $y(0)$, $z(0)$ *admitting a smooth solution. Apply the type-CK IMEX R-K method* (3)–(5) *with invertible matrix* $\hat{A}$ *and let* p *be the order of the explicit scheme. Assume that the method, with coefficients* b_i *and* a_{ij}, *is A-stable, that the stability function satisfies* $|R(\infty)| < 1$, *and that* $a_{ii} > 0$ *for all* i. *Assume that the weights satisfy the condition* $\tilde{b}_i = b_i$ *for* $i = 1, \dots, s$ *and that the method is stiffly accurate. Then, for any fixed constant* $C > 0$, *the global error satisfies, for* $\varepsilon \le Ch$,

$$y_n - y(x_n) = \mathcal{O}(h^{\tilde{q}+2} + h^p) + \mathcal{O}(\varepsilon h^2), \quad z_n - z(x_n) = \mathcal{O}(h^{\tilde{q}+1} + h^p) + \mathcal{O}(\varepsilon h) \tag{9}$$

with $\tilde{q} = \min\{\tilde{q}_s,\ \tilde{q}_i + 1$ *for all* $i = 2, \dots, s-1\}$. *The estimates hold uniformly for* $h \le h_0$ *and* $nh \le$ Const.

COROLLARY 3.1 (type ARS). *Under the same assumptions of Theorem* 3.2 *and with* $b_1 = 0$, *the global error satisfies* (9). *These estimates hold uniformly for* $h \le h_0$ *and* $nh \le$ Const.

Remark. Next, we shall show that if the method of type ARS is not *stiffly accurate*, one obtains the following estimates:

$$y_n - y(x_n) = \mathcal{O}(h^p + h^3) + \mathcal{O}(\varepsilon h^2), \quad z_n - z(x_n) = \mathcal{O}(h^p + h^2) + \mathcal{O}(\varepsilon h).$$

3.1. Numerical evidence. Before we provide proof of the main theorems, we present numerical results for the different types of IMEX R-K methods developed in the literature (see, e.g., [8], [1], [18], [19]), which confirm the theoretical prediction. Specifically, we will conduct convergence tests to compare the performance of different types of methods. As an example of a stiff problem (2) we consider one of the simplest nonlinear equations (describing nonlinear oscillations) in the stiff literature, the *van der Pol equation*

$$y' = z, \quad \varepsilon z' = (1 - y^2)z - y \tag{10}$$

with $0 \le \varepsilon \ll 1$. When the stiffness parameter ε is sufficiently small, numerical results confirm order reduction especially for the algebraic z-component. In our experiment, errors are computed by choosing initial values

$$y(0) = 2, \qquad z(0) = -\frac{2}{3} + \frac{10}{81}\varepsilon - \frac{292}{2187}\varepsilon^2 - \frac{1814}{19683}\varepsilon^3 + \mathcal{O}(\varepsilon^4) \tag{11}$$

such that the solution is smooth, and $\varepsilon = 10^{-6}$. In the following figures we have plotted the relative global error at $t_{end} = 0.55139$ as a function of the step size h, which was taken to be a constant over the considered interval $[0, t_{end}]$. We use logarithmic scales in both directions. The relative global error behaves like $C \cdot h^r$, where r is the slope of the straight line and C is a constant. We have indicated this behavior in all figures.

Table 1 shows the different types of IMEX R-K methods together with the global errors predicted by Theorems 3.1 and 3.2 and Corollary 3.1. Several conclusions are drawn from the numerical tests.

3.2. Discussion. (a) In Figures 1–7 we see that whenever p is small or h is very large the $\mathcal{O}(h^p)$ term is dominant in the z-component, whereas the other terms can be seen behaving otherwise. Furthermore the estimates in Table 1 demonstrate order reduction for the algebraic component in every type of method for a sufficiently stiff parameter ($\varepsilon = 10^{-6}$).

(b) An important ingredient, suggested by the analysis, is the condition $\tilde{b}_i = b_i$ for all i. Such a choice provides a significant benefit for the differential y-component. In fact the ARS$(4, 4, 3)$ method does not satisfy this condition, and for the y-component the global error drops to first order for a range of the step h. Note, however, that in Theorems 6.1 and 6.2 a satisfactory theoretical explanation of this fact is given.

In particular, the ARS$(4, 4, 3)$ method satisfies the conditions $\tilde{a}_{si} = \tilde{b}_i$ $\tilde{a}_{si} = \tilde{b}_i$ for all i, and in the next sections we shall observe that as a consequence of the above the z-component has the same estimate of the convergence rate as the y-component, justifying the behavior shown in Figure 3.

TABLE 1
Global errors predicted by theorems for the van der Pol equation.

Method	Stiffly accurate	y-comp.	z-comp.
ARS$(3,4,3)$, [1]	yes	$h^3+\varepsilon h^2$	h^2
MARS$(3,4,3)$	yes	$h^3+\varepsilon h^2$	$h^3+\varepsilon h$
ARS$(4,4,3)$, [1]	yes	$h^3+\varepsilon h$	$h^3+\varepsilon h$
ARK3(2)4L[2]SA, [8]	yes	h^3	h^2
ARK5(4)8L[2]SA, [8]	yes	$h^4+\varepsilon h^2$	$h^3+\varepsilon h$
ARK4(3)6L[2]SA, [8]	yes	h^4	$h^3+\varepsilon h$
MARK3(2)4L[2]SA	yes	h^3	$h^3+\varepsilon h$
IMEX-SSP2$(3,3,2)$, [18]	yes	h^2	h^2
IMEX-SSP3$(3,3,2)$, [18]	no	$h^3+\varepsilon h$	h
IMEX-SSP3$(4,3,3)$, [18]	no	$h^3+\varepsilon h$	h

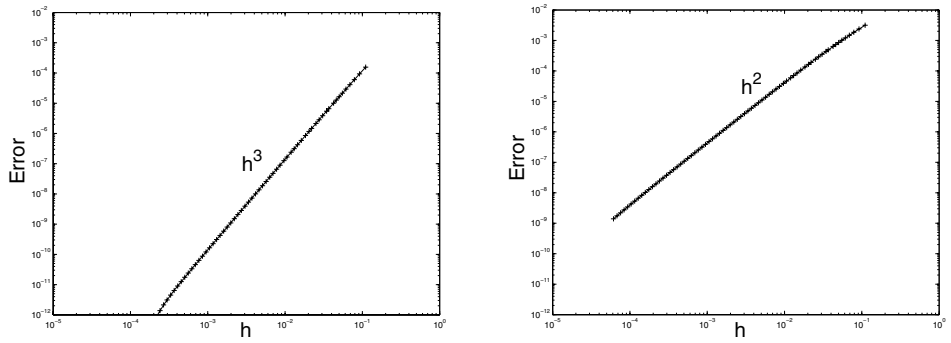


FIG. 1. *Global error versus the step size h for the ARS$(3,4,3)$-IMEX method using the van der Pol equation with $\varepsilon = 10^{-6}$. On the left-hand side is the y-component; on the right-hand side is the z-component.*

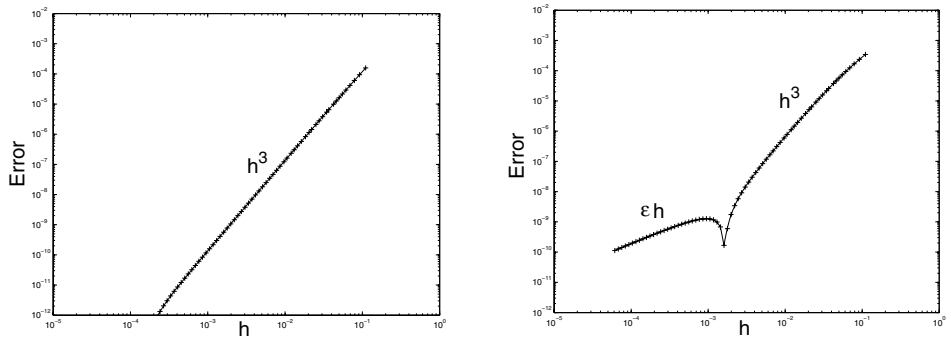


FIG. 2. *Global error versus the step size h for the MARS$(3,4,3)$-IMEX method using the van der Pol equation with $\varepsilon = 10^{-6}$. On the left-hand side is the y-component; on the right-hand side is the z-component.*

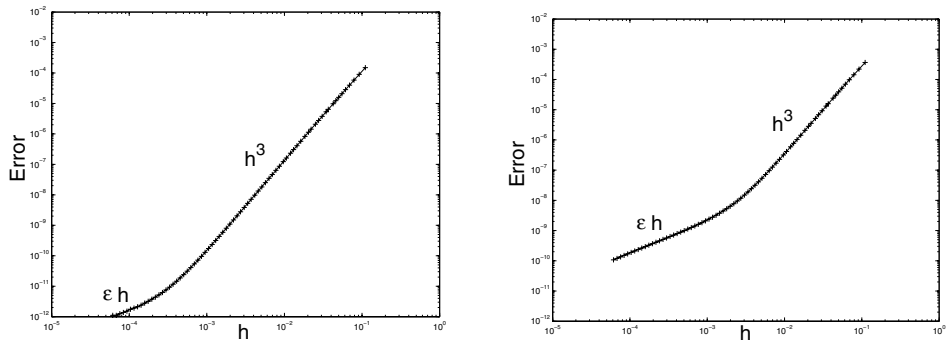


FIG. 3. *Global error versus the step size h for the ARS*$(4,4,3)$*-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$*. On the left-hand side is the y-component; on the right-hand side is the z-component.*

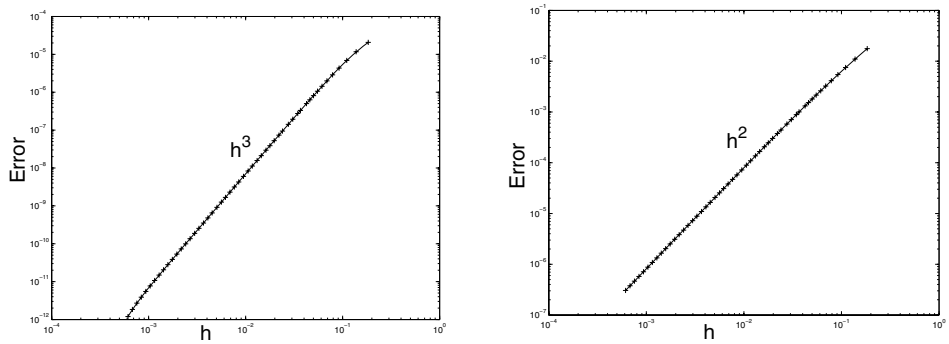


FIG. 4. *Global error versus the step size h for the ARK*$3(2)4L[2]$*SA-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$*. On the left-hand side is the y-component; on the right-hand side is the z-component.*

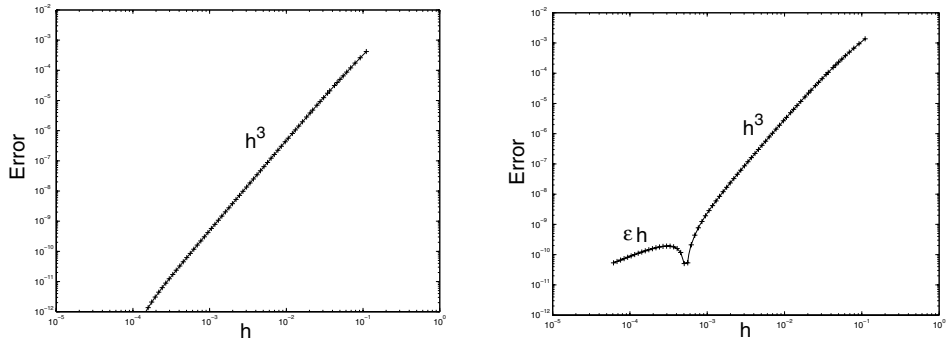


FIG. 5. *Global error versus the step size h for the MARK*$3(2)4L[2]$*SA-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$*. On the left-hand side is the y-component; on the right-hand side is the z-component.*

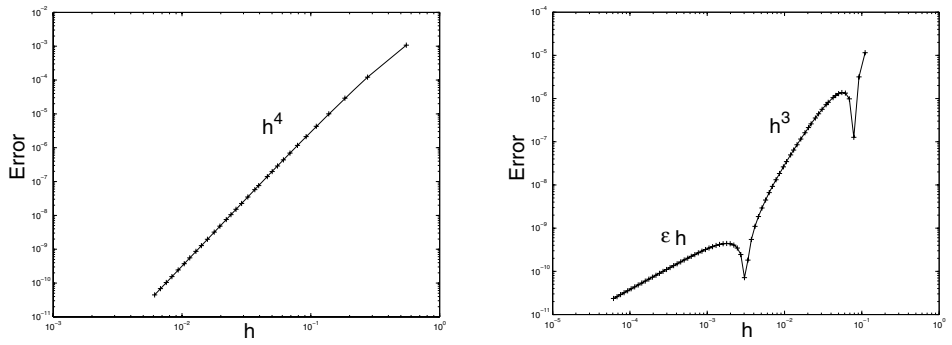


FIG. 6. *Global error versus the step size h for the ARK*4(3)6*L*[2]*SA-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$. *On the left-hand side is the y-component; on the right-hand side is the z-component.*

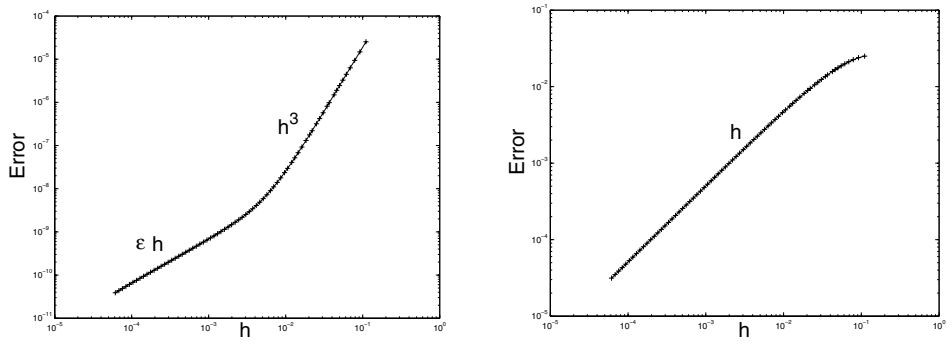


FIG. 7. *Global error versus the step size h for a type A-SSP*$(4,3,3)$*-IMEX method using the van der Pol equation with* $\varepsilon = 10^{-6}$. *On the left-hand side is the y-component; on the right-hand side is the z-component.*

(c) As noted in [8], according to the estimated convergence rates for differential and algebraic variables in [8, Table 12], several IMEX ARK_2 methods confirm the theoretical estimates given in Theorem 3.2. For instance, in order to justify the behavior observed in Figure 6, several pertinent assumptions are satisfied: $b_2 = \tilde{b}_2 = 0$ as well as the formula

$$\sum_{j=1}^{s} \tilde{a}_{ij} c_j = \frac{c_i^2}{2} \tag{12}$$

for $i = 3, \ldots, s$. Thus, using these assumptions, we achieve the estimates in Theorem 3.2.

(d) Finally, it is worth mentioning that the IMEX-SSP3$(4,3,3)$ scheme, as shown in Figure 7, exhibits order reduction both in the differential and algebraic components. Similarly, plots for the IMEX-SSP3$(3,3,2)$ scheme yield similar results. This behavior appears since the IMEX-SSP3$(3,3,2)$ and IMEX-SSP3$(4,3,3)$ schemes don't satisfy the condition $\sum_{ij} b_i \omega_{ij} \tilde{c}_j = 1$ required in Theorem 3.1. On the other hand, the IMEX-SSP2(3,3,2) scheme satisfies this condition so it achieves the anticipated convergence rate.

Improvements of existing schemes. A most relevant point demonstrated by this test is that methods such as *modified* ARK3(2)4L[2]SA (MARK3(2)4L[2]SA) and *modified* ARS$(3,4,3)$ (MARS$(3,4,3)$) produce an estimate for the z-component of the following form:

$$z_n - z(t_n) = \mathcal{O}(h^3) + \mathcal{O}(\varepsilon h) + \mathcal{O}(\varepsilon^2). \tag{13}$$

In this result the term $\mathcal{O}(\varepsilon^2)$ can be neglected since $\varepsilon \ll h$. Furthermore, to illustrate the results shown in Figures 2 and 5, we note that if the step size $h > \varepsilon^{1/2}$, the $\mathcal{O}(h^3)$ term is dominant; otherwise the term $\mathcal{O}(\varepsilon h)$ can be observed. A singularity appears in the neighborhood of $h \approx \varepsilon^{1/2}$ where we have a cancellation of error terms $\mathcal{O}(h^3)$ and $\mathcal{O}(\varepsilon h)$ with error constants of an opposite sign.

Therefore, the modified schemes give an improvement in the error estimate for the z-component when compared to the ARK3(2)4L[2]SA and ARS$(3,4,3)$ methods. In the following sections we will see that the global error estimates of these methods depend on $\tilde{q} = \min\{\tilde{q}_s,\ \tilde{q}_i + 1 \text{ for all } i = 2,\dots,s-1\}$. This fact enables us to construct methods with more accuracy. Notice the following:

(i) For the MARS$(3,4,3)$ method, a natural way to achieve the error estimate (13) is to increase from 1 to 2 the stage order in the sth stage of the explicit scheme so that $\tilde{q} = 2$.

(ii) In order to reach estimate (13) in the case of the MARK3(2)4L[2]SA method, we suggest using formula (12), for $i = 3,\dots,s$, in the explicit scheme accompanied by the assumption $\tilde{b}_2 = 0$. The assumption $\tilde{b}_2 = 0$ is necessary because the assumption (12) cannot be satisfied for $i = 2$; otherwise we would have $c_2 = 0$ and the method would be equivalent to one with fewer stages.

4. Asymptotic expansion. To obtain our main results in a general setting, we start from the ε-expansion of the exact solution of problem (2). Here, in particular, we are interested in smooth solutions which are of the form

$$\begin{aligned} y(t) &= y_0(t) + \varepsilon y_1(t) + \varepsilon^2 y_2(t) + \cdots, \\ z(t) &= z_0(t) + \varepsilon z_1(t) + \varepsilon^2 z_2(t) + \cdots, \end{aligned} \tag{14}$$

where $y_i(t)$ and $z_i(t)$ are ε-independent functions, which are solutions of a sequence of DAEs of arbitrary index.

The aim in this section is to analyze the ε-expansion of the numerical solution for problem (2) and verify how a sequence of differential-algebraic systems arise in the study of such a problem. A general and detailed investigation about the ε-expansion of the exact solution for problem (2) is given in [13] and [16].

We consider the IMEX R-K method (3), (5). We formally expand the quantities Y_{ni}, k_{ni}, y_n, Z_{ni}, ℓ_{ni}, and z_n into powers of ε with ε-independent coefficients:

$$y_n = y_n^0 + \varepsilon y_n^1 + \varepsilon^2 y_n^2 + \cdots, \tag{15a}$$

$$Y_{ni} = Y_{ni}^0 + \varepsilon Y_{ni}^1 + \varepsilon^2 Y_{ni}^2 + \cdots, \tag{15b}$$

$$k_{ni} = k_{ni}^0 + \varepsilon k_{ni}^1 + \varepsilon^2 k_{ni}^2 + \cdots, \tag{15c}$$

$$z_n = z_n^0 + \varepsilon z_n^1 + \varepsilon^2 z_n^2 + \cdots, \tag{15d}$$

$$Z_{ni} = Z_{ni}^0 + \varepsilon Z_{ni}^1 + \varepsilon^2 Z_{ni}^2 + \cdots, \tag{15e}$$

$$\ell_{ni} = \varepsilon^{-1}\ell_{ni}^{-1} + \ell_{ni}^0 + \varepsilon \ell_{ni}^1 + \varepsilon^2 \ell_{ni}^2 + \cdots. \tag{15f}$$

Because of the linearity of relations (3) and (5) we have to order ε^ν, with $\nu = -1$,

$$0 = h\sum_{j=1}^{i} a_{ij}\ell_{nj}^{-1}, \quad 0 = h\sum_{i=1}^{s} b_i\ell_{ni}^{-1} \tag{16}$$

and, for $\nu \geq 0$,

$$\begin{pmatrix} y_{n+1}^\nu \\ z_{n+1}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h\sum_{i=1}^{s} \begin{pmatrix} \tilde{b}_i k_{ni}^\nu \\ b_i \ell_{ni}^\nu \end{pmatrix}, \tag{17}$$

$$\begin{pmatrix} Y_{ni}^\nu \\ Z_{ni}^\nu \end{pmatrix} = \begin{pmatrix} y_n^\nu \\ z_n^\nu \end{pmatrix} + h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} k_{nj}^\nu \\ \sum_{j=1}^{i} a_{ij} \ell_{nj}^\nu \end{pmatrix}. \tag{18}$$

Inserting (15b), (15c), (15e), and (15f) into (4) and comparing equal powers of ε, we obtain

$$\varepsilon^0 : \begin{cases} k_{ni}^0 = f(Y_{ni}^0, Z_{ni}^0), \\ \ell_{ni}^{-1} = g(Y_{ni}^0, Z_{ni}^0), \end{cases} \tag{19a}$$

$$\varepsilon^1 : \begin{cases} k_{ni}^1 = f_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^1 + f_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^1, \\ \ell_{ni}^0 = g_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^1 + g_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^1, \end{cases} \tag{19b}$$

$$\cdots\cdots$$

$$\varepsilon^\nu : \begin{cases} k_{ni}^\nu = f_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^\nu + f_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^\nu + \varphi_\nu(Y_{ni}^0, Z_{ni}^0, \ldots, Y_{ni}^{\nu-1}, Z_{ni}^{\nu-1}), \\ \ell_{ni}^{\nu-1} = g_y(Y_{ni}^0, Z_{ni}^0)Y_{ni}^\nu + g_z(Y_{ni}^0, Z_{ni}^0)Z_{ni}^\nu + \psi_\nu(Y_{ni}^0, Z_{ni}^0, \ldots, Y_{ni}^{\nu-1}, Z_{ni}^{\nu-1}). \end{cases} \tag{19c}$$

Since (4) has a similar form to (2), the formulas (19a), (19b), and (19c) are exactly the same as those of the expansion in powers of ε for the exact solution (see [13] and [12]). In response to this fact, it follows that the coefficients y_n^0, z_n^0, y_n^1, z_n^1, ... represent the numerical solution of an arbitrary IMEX R-K method applied to DAEs of arbitrary index. Finally, subtracting (15a) and (15d) from (14), we get formally

$$y_n - y(t_n) = \sum_{\nu\geq 0} \varepsilon^\nu (y_n^\nu - y_\nu(t_n)), \quad z_n - z(t_n) = \sum_{\nu\geq 0} \varepsilon^\nu (z_n^\nu - z_\nu(t_n)). \tag{20}$$

Hence, the error of the numerical solution possesses an ε-expansion whose coefficients are the errors of the method applied to the differential-algebraic system. Clearly, in order to study this error, one will investigate only the differences $y_n^\nu - y_\nu(t_n)$, $z_n^\nu - z_\nu(t_n)$.

5. Zeroth-order expansion (index 1). From an arbitrary SPP (2) now we want to study the behavior of the global error of different types of IMEX R-K schemes for $\varepsilon \to 0$. In this section we start by studying the limiting case $\varepsilon = 0$. This gives us the corresponding *reduced* problem

$$\begin{aligned} y' &= f(y, z), \\ 0 &= g(y, z). \end{aligned} \tag{21}$$

We assume that $g_z(y,z)$ is invertible in a neighborhood of the solution of (21). This assumption guarantees the solvability of (21) and that the equation $g(y,z)=0$ possesses a locally unique solution (implicit function theorem). Furthermore, the same assumption guarantees that system (21) is a differential-algebraic one of *index* 1 [13]. Therefore, our first goal is to consider the different types of schemes applied to the reduced problem.

Type A. An IMEX R-K method of type A applied to the reduced problem has the form

$$Y_{ni} = y_n + h\sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_{nj}, Z_{nj}), \tag{22a}$$

$$0 = g(Y_{ni}, Z_{ni}), \tag{22b}$$

$$y_{n+1} = y_n + h\sum_{i=1}^{s} \tilde{b}_i f(Y_{ni}, Z_{ni}), \tag{22c}$$

$$z_{n+1} = R(\infty) z_n + \sum_{i,j=1}^{s} b_i \omega_{ij} Z_{nj}. \tag{22d}$$

Remarks. (a) By the implicit function theorem applied to (22b), we have $Z_{ni} = G(Y_{ni})$ for $i=1,\dots,s$. Consequently, by $Y_{ni} = y(t_n + \tilde{c}_i h) + \mathcal{O}(h^{\tilde{q}_i+1})$, it follows that the *internal stages* Z_{ni} depend on the coefficients $\tilde{c}_i$ of the explicit scheme.

(b) Concerning system (21), the y-component can be interpreted as the numerical solution of the ordinary differential equation $y' = f(y, H(y))$ with $z = H(y)$ (implicit function theorem). Therefore, for the method (22a)–(22d) we have

$$y_n - y(t_n) = \mathcal{O}(h^p),$$

because the formulas (22a), (22b), and (22c) are independent of z_n with p the order of the explicit scheme. Thus, we have only to prove a convergence result for the z-component.

Type CK. By Definition 2.4, we assume submatrix $\hat{A}$ is invertible. By (16), we have $0 = ha_{i1}\ell_{n1}^{-1} + h\sum_{j=2}^{i} a_{ij}\ell_{nj}^{-1}$ for $i=2,\dots,s$. Now, by the fact that $\ell_{n1}^{-1} = g(y_n, z_n)$, we obtain

$$\ell_{ni}^{-1} = \alpha_i g(y_n, z_n), \tag{23}$$

where $\alpha_i = -\sum_{i,j=2}^{s} \hat{\omega}_{ij} a_{j1}$ for $i=2,\dots,s$, with $\hat{\omega}_{ij}$ elements of the inverse matrix of $\hat{A}$.

Now, looking at (16), Lemma 5.1 follows from (23) and $\ell_{n1}^{-1} = g(y_n, z_n)$.

LEMMA 5.1. *The condition*

$$b_1 + \sum_{i=2}^{s} b_i \alpha_i = 0 \tag{24}$$

is automatically satisfied if the IMEX R-K method of type CK is stiffly accurate.

Therefore we assume that the method is stiffly accurate in the implicit part. This, moreover, yields $z_{n+1} = Z_{ns}$. Next we will use this lemma.

Then for the reduced problem a type-CK IMEX R-K scheme is defined by

$$Y_{ni} = y_n + h\sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_{nj}, Z_{nj}), \tag{25a}$$

$$y_{n+1} = y_n + h\sum_{i=1}^{s} \tilde{b}_i f(Y_{ni}, Z_{ni}), \tag{25b}$$

$$g(Y_{ni}, Z_{ni}) = \alpha_i g(y_n, z_n), \quad i = 2, \dots, s-1, \tag{25c}$$

$$g(Y_{ns}, z_{n+1}) = \alpha_s g(y_n, z_n). \tag{25d}$$

Type ARS. Since this is a particular case of CK with $a_{i1} = 0$ it follows that $\alpha_i = 0$ and $G(Y_{ni}, Z_{ni}) = 0$ for $i = 2, \dots, s$. As an immediate consequence, we get explicitly $z_{n+1} = R(\infty)z_n + \sum_{i,j=2}^{s} b_j \hat{\omega}_{ij} Z_{nj}$. In particular, if the method is stiffly accurate, $z_{n+1} = Z_{ns}$. In particular more theoretical insight into this type ARS shows that if we have $a_{si} = b_i$ and $\tilde{a}_{si} = \tilde{b}_i$ for $i = 1, \dots, s$, it also follows that $g(y_{n+1}, z_{n+1}) = 0$. Thus if $g_z(y, z)$ is invertible, we may express z_{n+1} as a function of y_{n+1}, and therefore we can declare that the z-component has the same asymptotic error estimate as the y-component.

After having understood the structure of each method, we are now in a position to prove the following results. All the theorems below are built on the assumption that the reduced problem satisfies (8) in a neighborhood of the exact solution $(y(t), z(t))$, and we assume that the initial values are *consistent*, i.e., $g(y_0, z_0) = 0$.

THEOREM 5.1 (type A). *Consider an IMEX R-K method of type A. Let p be the classical order of the explicit R-K method. Assume that the stability function of the implicit scheme satisfies $|R(\infty)| < 1$. Then the numerical solution of* (22a)–(22d) *has global error*

$$z_n - z(t_n) = \begin{cases} \mathcal{O}(h^2) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ \mathcal{O}(h) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j \neq 1. \end{cases} \tag{26}$$

The estimates (26) *hold uniformly for $t_n - t_0 = nh \leq$ Const.*

Proof. We denote the global error by $\Delta z_n = z_n - z(t_n)$ and $R(\infty) = \rho$. By remark (b), we get $Z_{ni} = z(t_n) + \tilde{c}_i h z'(t_n) + \mathcal{O}(h^2)$. Now, inserting it into (22d) and considering $z(t_{n+1}) = z(t_n) + h z'(t_n) + \mathcal{O}(h^2)$, one obtains

$$\Delta z_{n+1} = \rho \Delta z_n + \left(\sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j - 1 \right) h z'(t_n) + \mathcal{O}(h^2),$$

which allows us to conclude that

$$\Delta z_{n+1} = \begin{cases} \rho \Delta z_n + \delta_{n+1} & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ \rho \Delta z_n + \delta_{n+1} & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j \neq 1, \end{cases} \tag{27}$$

where

$$\delta_{n+1} = \begin{cases} O(h^2) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ O(h) & \text{if } \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j \neq 1. \end{cases} \tag{28}$$

Finally, repeated insertion of these formulas gives

$$\Delta z_n = \begin{cases} \sum_{i=1}^{n} \rho^{n-j}\delta_j & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} = 1, \\ \sum_{i=1}^{n} \rho^{n-j}\delta j & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} \neq 1 \end{cases} \tag{29}$$

because $\Delta z_0 = 0$. Thus, by the hypothesis $|\rho| < 1$, we obtain

$$\Delta z_n = \begin{cases} \mathcal{O}(h^2) & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} = 1, \\ \mathcal{O}(h) & \text{if } \sum_{i,j=1}^{s} b_i\omega_{ij}\tilde{c_j} \neq 1. \end{cases} \tag{30}$$

□

Remark. If the IMEX R-K method is stiffly accurate, it follows by (22b) that $z_{n+1} = Z_{ns} = G(Y_{ns})$. By remark (b), since we get $Z_{ns} - z(t_n + \tilde{c}_s h) = G(Y_{ns}) - G(y(t_n + \tilde{c}_s h)) = \mathcal{O}(h^{\tilde{q}_s+1})$, if $\tilde{c}_s = 1$, this proves the following estimate: $z_n - z(t_n) = \mathcal{O}(h^{\tilde{q}_s+1})$. Moreover, if in the explicit part we also have $\tilde{a}_{si} = \tilde{b}_i$ for $i = 1, \dots, s$, this yields $y_{n+1} = Y_{ns}$. Therefore, by $g(y_{n+1}, z_{n+1}) = 0$ and by the implicit function theorem, it follows that $z_{n+1} = G(y_{n+1})$, and in this situation the estimate is $z_n - z(t_n) = \mathcal{O}(h^p)$.

THEOREM 5.2 (type CK). *Consider an IMEX R-K method of type KC stiffly accurate with invertible matrix $\hat{A}$ and weights $\tilde{b}_i = b_i$ for $i = 1, \dots, s$. Let p be the order of explicit scheme. Assume that the stability function of the implicit scheme satisfies $|R(\infty)| < 1$ and $\delta = |\alpha_s| < 1$. Then the numerical solution of* (25a)–(25d) *has global error*

$$y_n - y(t_n) = \mathcal{O}(h^{\tilde{q}+2}) + \mathcal{O}(h^p), \qquad z_n - z(t_n) = \mathcal{O}(h^{\tilde{q}+1}) + \mathcal{O}(h^p) \tag{31}$$

with $\tilde{q} = \min\{\tilde{q}_s, \tilde{q}_i + 1,\ i = 2, \dots, s-1\}$. These estimates holds uniformly for $nh \leq$ Const.

Proof. By the relation (23), it follows that

$$g(Y_{ni}, Z_{ni}) = \alpha_i g(y_n, z_n), \tag{32}$$

so that Z_{ni} is a function of Y_{ni}, y_n, and z_n for $i = 2, \dots, s$. On the other hand, to provide an optimal estimate for the local error of the y-component we introduce the internal stages U_{ni}, V_{ni} that satisfy the relation

$$g(U_{ni}, V_{ni}) = \alpha_i g(y(t_n), z(t_n)) \tag{33}$$

for $i = 2, \dots, s$. Of course, this implies that V_{ni} is a function of U_{ni}, $y(t_n)$, and $z(t_n)$. Also, the internal stage U_{ni} is defined as

$$U_{ni} = y(t_n) + h\left(\tilde{a}_{i1}y'(t_n) + \sum_{j=2}^{i-1} \tilde{a}_{ij} f(U_{nj}, V_{ni})\right), \tag{34}$$

where $y'(t_n) = f(y(t_n), z(t_n))$ is the exact solution of $y(t)$ in t_n.

Next we shall use the abbreviation $g_z(t_n) = g_z(y(t_n), z(t_n))$, $f_y(t_n) = f_y(y(t_n), z(t_n))$ and denote $\Delta y_n = y_n - y(t_n)$ and $\Delta z_n = z_n - z(t_n)$.

Our proof proceeds in two parts, referred to as (a) and (b).

(a) We first estimate the differences $\|Z_{ni} - V_{ni}\|$, $\|Y_{ni} - U_{ni}\|$ of the internal stages. For this, we subtract $Y_{ni} = y_n + h(\tilde{a}_{i1} f(y_n, z_n) + \sum_{j=2}^{i-1} \tilde{a}_{ij} f(Y_{nj}, Z_{nj}))$ from (34) to obtain

$$\|Y_{ni} - U_{ni}\| \leq \|\Delta y_n\| + \mathcal{O}(h\|y_n\| + h\|\Delta z_n\|) + Ch\sum_{j=2}^{i-1} |\tilde{a}_{ij}| \, \|Z_{nj} - V_{nj}\| \tag{35}$$

for $i = 2, \ldots, s$ by the use of a Lipschitz condition for f.

We now linearize (32) and (33). Subtracting the two quantities, by the use of (35) and the condition $g_z^{-1}(t_n + c_i h) g_z(t_n) = I + \mathcal{O}(h)$, we obtain

$$\|Z_{ni} - V_{ni}\| \le |\alpha_i| \, \|\Delta z_n\| + \mathcal{O}(h \, \|\Delta z_n\|) + \mathcal{O}(\|\Delta y_n\|). \tag{36}$$

(b) Our next aim is to prove the recursion

$$\begin{pmatrix} \|\Delta y_{n+1}\| \\ \|\Delta z_{n+1}\| \end{pmatrix} \le \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1) & \delta + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|\Delta y_n\| \\ \|\Delta z_n\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{p+1}) \\ \mathcal{O}(h^{\tilde{q}+1}) \end{pmatrix}. \tag{37}$$

For the verification of the first relation in (37) we again linearize the quantities $y(t_n+h)$ and y_{n+1} to obtain

$$\begin{aligned} \Delta y_{n+1} = \Delta y_n + h\tilde{b}_1 (f_y(t_n) \Delta y_n + f_z(t_n) \Delta z_n) + h \sum_{i=2}^{s} \tilde{b}_i (f_y(t_n) \, (Y_{ni} - U_{ni}) \\ + f_z(t_n) \, (Z_{ni} - V_{ni})) + \mathcal{O}(h^2 \, \|y_n\| + h^2 \, \|z_n\|) + \mathcal{O}(h^{p+1}), \end{aligned} \tag{38}$$

and inserting (35) and (36) into (38), we get

$$\|\Delta y_{n+1}\| \le (1 + C_1 h) \, \|\Delta y_n\| + C_2 h^2 \, \|\Delta z_n\| + \mathcal{O}(h^{p+1}). \tag{39}$$

In (39), we applied the statement of Lemma 5.1.

Now we compute the second relation in (37) from (25d) and its exact expression $g(y(t_n + \tilde{c}_s h), z(t_{n+1})) = \alpha_s g(y(t_n), z(t_n))$. Linearizing and subtracting the two quantities, respectively, we obtain

$$\begin{aligned} \Delta z_{n+1} = -g_z^{-1}(t_n + \tilde{c}_s h) g_y(t_n) \Delta Y_{ns} + \alpha_s g_z^{-1}(t_n + \tilde{c}_s h) g_y(t_n) y_n \\ + \alpha_s g_z^{-1}(t_n + \tilde{c}_s h) g_z(t_n) z_n + \mathcal{O}\left(\|\Delta y_n\|^2 + \|\Delta z_n\|^2 \right). \end{aligned}$$

We now assume that

$$\|\Delta y_n\| \le Ch, \quad \|\Delta z_n\| \le Ch, \tag{40}$$

with some fixed constant C.[1] Therefore, by $g_z^{-1}(t_n + h) g_z(t_n) = I + \mathcal{O}(h)$, it follows that

$$\|\Delta z_{n+1}\| \le |\alpha_s| \, \|\Delta z_n\| + \mathcal{O}\left(\|\Delta y_n\| + h \, \|\Delta z_n\| \right) + \mathcal{O}(\|\Delta Y_{ns}\|) \tag{41}$$

as long as assumption (40) is satisfied. Now, using

$$\Delta Y_{ni} = \Delta y_n + h \sum_{j=1}^{i-1} \tilde{a}_{ij} \Delta k_{nj} + \mathcal{O}(h^{\tilde{q}_i + 1}) \tag{42}$$

and a Lipschitz condition for f gives

$$\|\Delta k_{ni}\| \le M \, \|\Delta Y_{ni}\| + N \, \|\Delta Z_{ni}\| \, . \tag{43}$$

[1] This statement should be interpreted to mean that if h is sufficiently small, the numerical solution will never violate the conditions (40).

In order to find an optimal estimate of (41) we proceed as follows. The linearization of (32) and the exact expression $g(y(t_n+\tilde{c}_ih),z(t_n+\tilde{c}_ih))=\alpha_i g(y(t_n),z(t_n))$ yields

$$\|\Delta Z_{ni}\| \le |\alpha_i|\,\|\Delta z_n\| + \mathcal{O}\left(h\,\|\Delta z_n\| + \|\Delta y_n\|\right) + \mathcal{O}(\|\Delta Y_{ni}\|).$$

Inserted into (43), with the help of (42) after repeated insertions of $\|\Delta Y_{ni}\|$, and setting $i=s$, we obtain

$$\|\Delta Y_{ns}\| \le \|\Delta y_n\| + hC_1\left(\|\Delta y_n\| + \|\Delta z_n\|\right) + hC_2(|\alpha_s|\,\|\Delta z_n\| + \|\Delta y_n\|) + \mathcal{O}(h^{\tilde{q}+1})$$

with $\tilde{q}=\min\{\tilde{q}_s,\ \tilde{q}_i+1,\ i=2,\dots,s-1\}$. Now putting the previous formula into (41), it follows that

$$\|\Delta z_{n+1}\| \le C\,\|\Delta y_n\| + (\delta + Ch)\,\|\Delta z_n\| + \mathcal{O}(h^{\tilde{q}+1}), \tag{44}$$

where $\delta=|\alpha_s|$. This completes the proof of formula (37).

Now applying Lemma 5.2 below to (37) gives the estimates (31) for $nh\le \text{Const}$, completing the proof. □

LEMMA 5.2. *Let $\{u_n\}$ and $\{v_n\}$ be two sequences of nonnegative numbers satisfying (componentwise)*

$$\begin{pmatrix} u_{n+1} \\ v_{n+1} \end{pmatrix} \le \begin{pmatrix} 1+\mathcal{O}(h) & \mathcal{O}(h^2) \\ \mathcal{O}(1) & \delta+\mathcal{O}(h) \end{pmatrix} \begin{pmatrix} u_n \\ v_n \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{p+1}) \\ \mathcal{O}(h^{\tilde{q}+1}) \end{pmatrix} \tag{45}$$

with $0\le\delta<1$. Then the following estimate holds for $nh\le\text{Const}$ and $h\le h_0$:

$$\begin{aligned} u_n &\le C(u_0 + h^2v_0 + h^{\tilde{q}+2} + h^p), \\ v_n &\le C(u_0 + (\delta^n + h)\,v_0 + h^{\tilde{q}+1} + h^p). \end{aligned} \tag{46}$$

The proof is similar to that of Lemma 3.9 in [13].

Remarks. It is worth noting that if $b_i=\tilde{b}_i$ for $i=1,\dots,s$, then inserting (35) and (36) into (38) yields the nonzero quantity $\tilde{b}_1+\sum_{i=2}^s \tilde{b}_i\alpha_i$. For the y-component this implies $y_n-y(t_n)=\mathcal{O}(h^p)+\mathcal{O}(h^{\tilde{q}+1})$.

COROLLARY 5.1 (type ARS). *Suppose that the assumptions of Theorem 5.2 are satisfied and $b_1=0$. Then the numerical solution has global error satisfying* (31).

Remarks. We now suppose that the ARS method is not stiffly accurate. In order to obtain an optimal evaluation of Δz_n, we proceed as follows. Since $g(Y_{ni},Z_{ni})=0$, we get $Z_{ni}=G(Y_{ni})$ for $i=2,\dots,s$. By the Lipschitz condition for G, it follows that $\|\Delta Z_{ni}\|\le C\,\|\Delta Y_{ni}\|$. Using (42) and (43), we get

$$\|\Delta Y_{ni}\| \le \|\Delta y_n\| + h\,|\tilde{a}_{i1}|\,(\|\Delta y_n\| + \|\Delta z_n\|) + \mathcal{O}(h^{\tilde{r}_i+1}) \tag{47}$$

with $\tilde{r}_i=\min\{\tilde{q}_i,\tilde{q}_j+1,\ \ j=1,\dots,i-1\}$. It thus follows from the numerical and exact solution that

$$\|\Delta z_{n+1}\| \le |\rho|\,\|\Delta z_n\| + C\sum_{i,j=2}^{s} |b_i\hat{\omega}_{ij}|\,\|\Delta Y_{nj}\| + \mathcal{O}(h^{q+1}), \tag{48}$$

where $\rho=1-\sum_{i,j=2}^s b_i\hat{\omega}_{ij}$ and $q=\min_{i\le s}q_i$. Now, inserting (47) into (48) yields

$$\|\Delta z_{n+1}\| \le (|\rho|+C_2h)\,\|\Delta z_n\| + C_1\,\|\Delta y_n\| + \mathcal{O}(h^{\tilde{r}+1}) + \mathcal{O}(h^{q+1}), \tag{49}$$

where $\tilde{r} = \min\{\tilde{r}_2, \ldots, \tilde{r}_s\}$ with $|\rho| < 1$. We now solve (39) and (49), applying again Lemma 5.2, thus obtaining for the global error

$$y_n - y(t_n) = \mathcal{O}(h^p) + \mathcal{O}(h^3), \quad z_n - z(t_n) = \mathcal{O}(h^p) + \mathcal{O}(h^2).$$

Observe that the estimates obtained above are given since $q = 1$.

It is interesting to note that if $\tilde{b}_1 \neq 0$, in (39) we get $\|\Delta y_{n+1}\| \leq (1+C_1 h)\,\|\Delta y_n\| + C_2 h\,\|\Delta z_n\| + \mathcal{O}(h^{p+1})$ and the proof follows as above.

6. Higher-order expansion (higher index). Now we study the global error of IMEX R-K methods when applied to the SPP (2). To this end, we are interested in studying the differences $y_n^\nu - y_\nu(t_n)$ and $z_n^\nu - z_\nu(t_n)$ from (20). All the theorems below are built on the assumption that the stability function of the implicit scheme satisfies $|R(\infty)| < 1$ and the weights $\tilde{b}_i = b_i$ for all i. In what follows, when we use the superscript 0 in the quantities Y_{ni}, Z_{ni}, k_{ni}, ℓ_{ni}, y_n, z_n, we are treating the behavior of the numerical solution of the *reduced* problem.

THEOREM 6.1 (type A). *Consider an IMEX R-K method of type A such that (a_{ij}) is invertible. Assume* (8) *holds and the initial values of the differential-algebraic system of index $\nu + 1$ are consistent. Then if $\sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1$, the global error of method* (17)–(19c) *satisfies, for $\nu = 1, 2$,*

$$y_n^\nu - y_\nu(t_n) = \mathcal{O}(h^{3-\nu}), \quad z_n^\nu - z_\nu(t_n) = \mathcal{O}(h^{2-\nu}); \tag{50}$$

otherwise

$$y_n^\nu - y_\nu(t_n) = \mathcal{O}(h^{2-\nu}), \quad z_n^\nu - z_\nu(t_n) = \mathcal{O}(h^{1-\nu}). \tag{51}$$

Proof. Here we emphasize some straightforward differences with respect to Theorem 3.4 in [13].

(a) We begin by denoting the differences to the exact solution values:

$$\begin{aligned} \Delta y_n^\nu &= y_n^\nu - y_\nu(t_n), & \Delta z_n^\nu &= z_n^\nu - z_\nu(t_n), \\ \Delta Y_{ni}^\nu &= Y_{ni}^\nu - y_\nu(t_n + \tilde{c}_i h), & \Delta Z_{ni}^\nu &= Z_{ni}^\nu - Z_\nu(t_n + c_i h), \\ \Delta k_{ni}^\nu &= k_{ni}^\nu - y_\nu'(t_n + \tilde{c}_i h), & \Delta \ell_{ni}^\nu &= \ell_{ni}^\nu - z_\nu'(t_n + c_i h). \end{aligned} \tag{52}$$

Furthermore we have for an IMEX R-K method

$$\begin{pmatrix} \Delta Y_{ni}^\nu \\ \Delta Z_{ni}^\nu \end{pmatrix} = \begin{pmatrix} \Delta y_n^\nu \\ \Delta z_n^\nu \end{pmatrix} + h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} \Delta k_{nj}^\nu \\ \sum_{j=1}^{i} a_{ij} \Delta \ell_{nj}^\nu \end{pmatrix} + \begin{pmatrix} \mathcal{O}(h^{\tilde{q}_i+1}) \\ \mathcal{O}(h^{q_i+1}) \end{pmatrix}. \tag{53}$$

From Theorem 5.1 it follows that

$$\begin{aligned} \Delta y_n^0 &= \mathcal{O}(h^p), & \Delta Y_{ni}^0 &= \mathcal{O}(h^{\tilde{q}_i+1}), \\ \Delta k_{ni}^0 &= \mathcal{O}(h^{\tilde{q}_i+1}), & \Delta Z_{ni}^0 &= \mathcal{O}(h^{\tilde{q}_i+1}), \end{aligned} \tag{54}$$

and

$$\Delta z_n^0 = \begin{cases} \mathcal{O}(h^2) & \text{if} \quad \sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \\ \mathcal{O}(h) & \text{otherwise.} \end{cases} \tag{55}$$

We also have

$$\Delta \ell_{ni}^0 = \begin{cases} \mathcal{O}(h) & \text{if} \quad \sum_{i,j=1}^{s} \omega_{ij} \tilde{c}_j = 1, \\ \mathcal{O}(1) & \text{otherwise.} \end{cases} \tag{56}$$

Here ω_{ij} are the elements of the inverse of matrix A.

(b) We first consider the case $\nu = 1$. In analogy to the proof of Theorem 3.4 in [13], using the estimates (54), we deduce the following expressions:

$$\begin{aligned}
\Delta k^1_{ni} &= f_y(t_n + \tilde{c}_i h)\Delta Y^1_{ni} + f_z(t_n + \tilde{c}_i h)\Delta Z^1_{ni} \\
&\qquad + \mathcal{O}(h^{\tilde{q}_i+1} + h^{\tilde{q}_i+1}\left\|\Delta Y^1_{ni}\right\| + h^{\tilde{q}_i+1}\left\|\Delta Z^1_{ni}\right\|), \\
\Delta \ell^0_{ni} &= g_y(t_n + \tilde{c}_i h)\Delta Y^1_{ni} + g_z(t_n + \tilde{c}_i h)\Delta Z^1_{ni} \\
&\qquad + \mathcal{O}(h^{\tilde{q}_i+1} + h^{\tilde{q}_i+1}\left\|\Delta Y^1_{ni}\right\| + h^{\tilde{q}_i+1}\left\|\Delta Z^1_{ni}\right\|).
\end{aligned} \tag{57}$$

Here we have used the abbreviations $f_y(t) = f_y(y_0(t), z_0(t))$, $g_y(t) = g_y(y_0(t), z_0(t))$. Now, we compute ΔZ^1_{ni} from the second relation in (57). Therefore, inserting it into the first one and using (53), we can eliminate ΔY^1_{ni} and obtain

$$\begin{aligned}
&\Delta k^1_{ni} - (f_z g_z^{-1})(t_n + \tilde{c}_i h)\Delta \ell^0_{ni} = \mathcal{O}(\left\|\Delta y^1_n\right\|) \\
&+ (f_y - f_z g_z^{-1} g_y)(t_n + \tilde{c}_i h) h \sum_{j=1}^{i-1} \left((f_z g_z^{-1})(t_n + \tilde{c}_j h)\tilde{a}_{ij}\Delta \ell^0_{nj} + \mathcal{O}(h^{\tilde{q}_j+1})\right) + \mathcal{O}(h^{\tilde{q}_i+1}).
\end{aligned}$$

By (56), it follows that $\Delta k_{ni} = \mathcal{O}(\left\|\Delta y^1_n\right\|) + \mathcal{O}(h)$ if $\sum_{i,j=1}^{i} b_i \omega_{ij} \tilde{c}_j = 1$; otherwise $\Delta k_{ni} = \mathcal{O}(\left\|\Delta y^1_n\right\|) + \mathcal{O}(1)$. A direct estimation of Δy^1_n proves that $\Delta y^1_n = \mathcal{O}(h)$ if $\sum_{i,j=1}^{s} b_i \omega_{ij} \tilde{c}_j = 1$; otherwise $\Delta y^1_n = \mathcal{O}(1)$. However, these estimations are not optimal.

Following the argument in Theorem 3.4 in [13], we now introduce the new variable

$$\Delta u^1_n = \Delta y^1_n - (f_z g_z^{-1})(t_n)\Delta z^0_n. \tag{58}$$

At this point the only difference is that we have to treat more carefully the quantity $\Delta k^1_{ni} - f_z g_z^{-1}(t_n)\Delta \ell^0_{ni}$ for all i. For details we refer to [4]. Using the hypothesis $\tilde{b}_i = b_i$ for $i = 1, \ldots, s$, we obtain

$$\begin{aligned}
\Delta u^1_{n+1} &= \Delta u^1_n + h \sum_{i=1}^{s} b_i \Big(\mathcal{O}(\left\|\Delta y^1_n\right\|) + \mathcal{O}(h\left\|\Delta \ell^0_{ni}\right\|) \\
&+ (f_y - f_z g_z^{-1} g_y)(t_n + \tilde{c}_i h) h \sum_{j=1}^{i-1} \left((f_y - f_z g_z^{-1} g_y)(t_n + \tilde{c}_j h)\tilde{a}_{ij}\Delta \ell^0_{nj} + \mathcal{O}(h^{\tilde{q}_j+1})\right) \\
&\qquad + \mathcal{O}(h^{\tilde{q}_i+1}) \Big) - ((f_z g_z^{-1})(t_n + h) - (f_z g_z^{-1})(t_n))\Delta z^0_{n+1} + \mathcal{O}(h^{p+1}),
\end{aligned}$$

where $\mathcal{O}(h\left\|\Delta \ell^0_{ni}\right\|) = ((f_z g_z^{-1})(t_n + \tilde{c}_i h) - (f_z g_z^{-1})(t_n))\Delta \ell^0_{ni}$. Consequently, the first relations in (56) and (55) and the fact that $((f_z g_z^{-1})(t_n + h) - (f_z g_z^{-1})(t_n)) = \mathcal{O}(h)$ imply that

$$\left\|\Delta u^1_{n+1}\right\| \le (1 + Ch)\left\|\Delta u^1_n\right\| + \mathcal{O}(h^3). \tag{59}$$

Then we have $\Delta u^1_n = \mathcal{O}(h^2)$ for $nh \le \text{Const}$ (observe that the initial values are assumed to be consistent, i.e., $\Delta u^1_0 = 0$), so that by (58) and (55) we also have $\Delta y^1_n = \mathcal{O}(h^2)$. This implies $\Delta k^1_{ni} = \mathcal{O}(h)$ and $\Delta Y^1_{ni} = \mathcal{O}(h^2)$. The second relation in (57) proves that $\Delta Z^1_{ni} = \mathcal{O}(h)$.

In order to estimate Δz_n^1 we proceed as in Theorem 3.4 in [13] and, because $|R(\infty)| < 1$, we thus obtain $\Delta z_n^1 = \mathcal{O}(h)$. In particular, we emphasize that if we consider the second relation in (56) and (55) in a similar way, we get $\Delta y_n^1 = \mathcal{O}(h)$ with $\Delta k_{ni}^1 = \mathcal{O}(1)$ and, in addition, it follows from $\Delta Z_{ni}^1 = \mathcal{O}(1)$ that $\Delta z_n^1 = \mathcal{O}(1)$.

(c) The proof for general ν is similar to that of Theorem 3.4 in [13]. It is worth commenting that the only difference arises in the quantity $\Delta \ell_{ni}^{\nu-1} = \mathcal{O}(h^{2-\nu})$ if $\sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1$; otherwise $\Delta \ell_{ni}^{\nu-1} = \mathcal{O}(h^{1-\nu})$. Thus the statement follows with

$$\begin{aligned} &\text{if } \sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1, \ \Delta Y_{ni}^{\nu} = \mathcal{O}(h^{3-\nu}), \quad \Delta Z_{ni}^{\nu} = \mathcal{O}(h^{2-\nu}); \\ &\text{otherwise} \qquad\qquad \Delta Y_{ni}^{\nu} = \mathcal{O}(h^{2-\nu}), \quad \Delta Z_{ni}^{\nu} = \mathcal{O}(h^{1-\nu}). \end{aligned} \tag{60}$$

∎

THEOREM 6.2 (type CK). *Consider an IMEX R-K method of type CK which is stiffly accurate and such that* $(\hat{a}_{ij})$ *is invertible. If* (8) *holds and if the initial values of the differential-algebraic system of index* $\nu + 1$ *are consistent, then the global error of method* (17)–(19c) *satisfies, for* $\nu = 1, 2$,

$$y_n^{\nu} - y_{\nu}(t_n) = \mathcal{O}(h^{3-\nu}), \qquad z_n^{\nu} - z_{\nu}(t_n) = \mathcal{O}(h^{2-\nu}). \tag{61}$$

Proof. From Theorem 5.2 it follows that

$$\begin{aligned} &\Delta y_n^0 = \mathcal{O}(h^{\tilde{q}+2} + h^p), \quad \Delta Y_{ni}^0 = \mathcal{O}(h^{\tilde{q}_i+1}), \quad \Delta k_{ni}^0 = \mathcal{O}(h^{\tilde{q}_i+1}), \\ &\Delta z_n^0 = \mathcal{O}(h^{\tilde{q}+1} + h^p), \quad \Delta Z_{ni}^0 = \mathcal{O}(h^{\tilde{q}_i+1}), \end{aligned} \tag{62}$$

with $\tilde{q} = \min\{\tilde{q}_s, \tilde{q}_i + 1, i = 2, \ldots, s-1\}$.

Again we consider the case $\nu = 1$. Here the study of convergence needs further investigation. We start by computing the difference $\Delta \ell_{n1}^0$. From (19b), we have $\ell_{n1}^0 = g_y(y_n^0, z_n^0) y_n^1 + g_z(y_n^0, z_n^0) z_n^1$, and this implies $\|\Delta \ell_{n1}^0\| \le C(\|\Delta y_n^0\| + \|\Delta z_n^0\| + \|\Delta y_n^1\| + \|\Delta z_n^1\|)$. Consequently, using (53), we have

$$\Delta \ell_{ni}^0 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + h^{-1} \hat{\omega}_{i2} (\Delta Z_{n2}^0 - \Delta z_n^0) + \mathcal{O}(h^{\tilde{q}_i}) + \mathcal{O}(h^{q_i}), \tag{63}$$

where $\hat{\omega}_{ij}$ are the elements of the inverse matrix of $\hat{A}$. Therefore, inserting (63) into the quantity $\Delta k_{ni}^1 - (f_z g_z^{-1})(t_n + \tilde{c}_i h) \Delta \ell_{ni}^0$ computed in the previous theorem, we obtain

$$\begin{aligned} \Delta k_{ni}^1 = {} & h^{-1} \hat{\omega}_{i2} (f_z g_z^{-1})(t_n + \tilde{c}_i h)(\Delta Z_{n2}^0 - \Delta z_n^0) \\ & + \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h^{\tilde{q}_i}) + \mathcal{O}(h^{q_i}). \end{aligned}$$

By (62) and $\tilde{q}_2 = 1$, we have $\Delta Y_{n2}^0 = \mathcal{O}(h^2)$ and $\Delta Z_{n2}^0 = \mathcal{O}(h^2)$. Hence, this implies $\Delta k_{ni}^1 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h)$, and a direct estimation of Δy_n^1 leads to

$$\|\Delta y_n^1\| \le (1 + Ch) \|\Delta y_n^1\| + Ch \|\Delta z_n^1\| + \mathcal{O}(h^2). \tag{64}$$

Now, using (53), this gives $\Delta \ell_{ni}^1 = \alpha_i \Delta \ell_{n1}^1 + h^{-1} \sum_{j \ge 2} \hat{\omega}_{ij} (\Delta Z_{nj}^1 - \Delta z_n^1) + \mathcal{O}(h^{q_i})$ for $i = 2, \ldots, s$. Since the method is stiffly accurate and $\tilde{b}_i = b_i$ for $i = 1, \ldots, s$, the statement of Lemma 5.1 is satisfied, and from $\Delta z_{n+1}^1 = \Delta z_n^1 + h(b_1 + \sum_{i \ge 2} b_i \alpha_i) \Delta \ell_{n1}^1 + \sum_{i,j \ge 2} b_i \hat{\omega}_{ij} (\Delta Z_{nj}^1 - \Delta z_n^1) + \mathcal{O}(h^{q_i+1})$ we obtain

$$\|\Delta z_{n+1}^1\| \le |\rho| \, \|\Delta z_n^1\| + \sum_{i,j \ge 2} |b_i \hat{\omega}_{ij}| \, \|\Delta Z_{nj}^1\| + \mathcal{O}(h^{q_i+1}) \tag{65}$$

with $|\rho| = |R(\infty)| < 1$. By (63) and $\Delta Z_{ni}^0 = \mathcal{O}(h^2)$, it follows that $\Delta \ell_{ni}^0 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h)$. Thus, the second relation of (57) proves that $\Delta Z_{ni}^1 = \mathcal{O}(\|\Delta y_n^1\| + \|\Delta z_n^1\|) + \mathcal{O}(h(\|\Delta y_n^1\| + \|\Delta z_n^1\|)) + \mathcal{O}(h)$. Inserting (65), we obtain

$$\left\|\Delta z_{n+1}^1\right\| \le \left\|\Delta y_n^1\right\| + (|\rho| + Ch)\left\|\Delta z_n^1\right\| + \mathcal{O}(h). \tag{66}$$

Now applying Lemma 5.2 to inequalities (64) and (66) gives $\Delta y_n^1 = \mathcal{O}(h)$, $\Delta z_n^1 = \mathcal{O}(h)$. Again, we can conclude that the estimate about Δy_n^1 is not optimal. Therefore, introducing the new variable (58), we obtain

$$\left\|\Delta u_{n+1}^1\right\| \le (1 + Ch)\left\|\Delta u_n^1\right\| + Ch^2\left\|\Delta z_n^1\right\| + \mathcal{O}(h^3). \tag{67}$$

We now apply Lemma 5.2 again, replacing the inequality (64) with (67). Then by (58) and (62) we have $\Delta y_n^1 = \mathcal{O}(h^2)$. Obviously, the proof for general ν is similar to the one presented in Theorem 6.1. This completes the proof of the theorem. ∎

Remark. Of course, concerning type ARS, under the same assumptions as Theorem 6.2 (with also $b_1 = 0$), we again deduce the estimates (61).

7. Estimates on the remainder. In order to estimate the remainder in the expansion (20), we require the same detailed analysis previously developed by Hairer, Lubich, and Roche in [12] (see also [13, sect. VI.3]). The main purpose in this section is to extend the same results presented in [13, sect. VI.3] to the different types of IMEX R-K methods.

Let us introduce existence and local uniqueness of the numerical solution of (4), (5). Next we shall discuss the influence of perturbations in (5) to the numerical solution.

We shall consider two steps in succession. First, we suppose that (y_n, z_n) are known, denoted by (η, ζ), and prove the existence and uniqueness of (y_{n+1}, z_{n+1}). We assume that $g(\eta, \zeta) = \mathcal{O}(h)$, $\mu(g_z(\eta, \zeta)) \le 1$ and that $a_{ii} > 0$ for all i. Thus we have the nonlinear system for the stage values

$$\begin{pmatrix} Y_i - \eta \\ \varepsilon\,(Z_i - \zeta) \end{pmatrix} = h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_j, Z_j) \\ \sum_{j=1}^{i} a_{ij} g(Y_j, Z_j) \end{pmatrix}. \tag{68}$$

It is significant to note that if we restrict ourselves to the use of a particular type of IMEX R-K method, for instance, type A, where the matrix A is invertible, we immediately obtain the statement of Theorem 3.5 in [13]. Instead, for type CK, it is worth commenting that the second equation in (68) becomes

$$\frac{\varepsilon}{h}(Z_i - \zeta) - a_{i1} g(\eta, \zeta) - \sum_{j=2}^{i} \hat{a}_{ij} g(Y_j, Z_j) = 0,$$

whereas for type ARS we have $a_{i1} = 0$ for all i. Therefore, we easily find again the statement of Theorem 3.5 in [13].

We now study the influence of perturbations in (68) to the numerical solution. For the perturbed IMEX R-K method

$$\begin{pmatrix} \widehat{Y}_i - \widehat{\eta} \\ \varepsilon\left(\widehat{Z}_i - \widehat{\zeta}\right) \end{pmatrix} = h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} f(\widehat{Y}_j, \widehat{Z}_j) \\ \sum_{j=1}^{i} a_{ij} g(\widehat{Y}_j, \widehat{Z}_j) \end{pmatrix} + h \begin{pmatrix} \delta_i \\ \theta_i \end{pmatrix} \tag{69}$$

we allow the following remarks.

Remarks. For an IMEX R-K scheme of type A the statement and the proof is similar to that of Theorem 3.6 in [13]. Extra care must to be taken to properly handle type CK. First observe that in addition to the assumptions of Theorem 3.5 in [13] we suppose that $\widehat{\eta} - \eta = \mathcal{O}(h)$, $\widehat{\zeta} - \zeta = \mathcal{O}(h)$, $\delta_i = \mathcal{O}(1)$, and $\theta_i = \mathcal{O}(h)$ for $i = 2, \dots, s$ with $\delta_1 = 0$ and $\theta_1 = 0$. Then we have for $h \le h_0$ the following estimates:

$$\begin{aligned} &\|\widehat{Y}_i - Y_i\| \le C(\|\widehat{\eta} - \eta\| + h\|\widehat{\zeta} - \zeta\|) + hC(\|\delta\| + \|\theta\|), \\ &\|\widehat{Z}_i - Z_i\| \le C\left(\|\widehat{\eta} - \eta\| + \left(\frac{\varepsilon}{h} + h\right)\|\widehat{\zeta} - \zeta\|\right) + C(h\,\|\delta\| + \|\theta\|), \end{aligned} \tag{70}$$

where $\delta = (\delta_1, \dots, \delta_s)^T$ and $\theta = (\theta_1, \dots, \theta_s)^T$. Later, we note that we have to treat the following homotopy more carefully:

$$\begin{pmatrix} Y_i - \eta \\ \varepsilon\,(Z_i - \zeta) \end{pmatrix} - h \begin{pmatrix} \sum_{j=1}^{i-1} \tilde{a}_{ij} f(Y_j, Z_j) \\ \sum_{j=1}^{i} \hat{a}_{ij} g(Y_j, Z_j) \end{pmatrix} = \tau \begin{pmatrix} \widehat{\eta} - \eta + h\delta_i \\ \varepsilon(\widehat{\zeta} - \zeta) + h a_{i1}(g(\widehat{\eta}, \widehat{\zeta}) - g(\eta, \zeta)) + h\theta_i \end{pmatrix},$$

which relates system (68) for $\tau = 0$ to the perturbed system (71) for $\tau = 1$. Furthermore, we denote by $\hat{a}_{ij}$ the elements of the submatrix $\hat{A}$, and, by the Lipschitz condition for g, we have the inequality

$$\|(g(\widehat{\eta}, \widehat{\zeta}) - g(\eta, \zeta))\| \le L\|\widehat{\eta} - \eta\| + L\|\widehat{\zeta} - \zeta\|.$$

Then, in this situation, the same conclusions of Theorem 3.6 in [13] hold. In particular, if $a_{i1} = 0$ for all i, the same also follows for type ARS.

Following [13], we finally estimate the remainder of the expansion (20).

THEOREM 7.1 (type A). *Under the same hypotheses as those of Theorem* 3.1*, for any fixed constant $c > 0$ and $\varepsilon \le ch$, the global error satisfies*

$$\begin{aligned} &y_n - y(t_n) = \Delta y_n^0 + \varepsilon \Delta y_n^1 + \varepsilon^2 \Delta y_n^2 + \mathcal{O}(\varepsilon^3), \\ &z_n - z(t_n) = \Delta z_n^0 + \varepsilon \Delta z_n^1 + \varepsilon^2 \Delta z_n^2 + \mathcal{O}(\varepsilon^3/h), \end{aligned} \tag{71}$$

where $\Delta y_n^0 = y_n^0 - y_0(t_n)$, $\Delta z_n^0 = z_n^0 - z_0(t_n), \dots$ are the global errors of the method applied to differential-algebraic system. The estimates (71) *hold uniformly for $h \le h_0$ and $nh \le$ Const.*

Remark. In order to enable a direct comparison with Theorem 3.8 in [13] (see also [12]), by Theorem 6.1, and by (50) and (60), if $\sum_{ij}^{s} b_i \omega_{ij} \tilde{c}_j = 1$, it suffices to prove the result for $\nu = 2$; otherwise it must be proven for $\nu = 1$. Therefore, the result follows directly by applying Theorem 3.8 in [13].

THEOREM 7.2 (type CK). *Under the same hypotheses as those of Theorem* 3.2*, then, for any fixed constant $c > 0$ and $\varepsilon \le ch$, the global error satisfies the estimates* (71) *uniformly for $h \le h_0$ and $nh \le$ Const.*

Remark. It is interesting, of course, to know how in the proof of Theorem 7.2 several formulas are related to those of Theorem 3.8 in [13]. For instance, by (19a)–(19c) it follows from (60) and $\nu = 2$ that

$$\begin{aligned} &\widehat{k}_{ni} = f(\widehat{Y}_{ni}, \widehat{Z}_{ni}) + \mathcal{O}(\varepsilon^3), \\ &\varepsilon\widehat{\ell}_{ni} = g(\widehat{Y}_{ni}, \widehat{Z}_{ni}) + \varepsilon^3 \ell_{ni}^2 + \mathcal{O}(\varepsilon^3). \end{aligned} \tag{72}$$

Using $Z_{ni}^{\nu} = z_n^{\nu} + ha_{i1}\ell_{n1}^{\nu} + h\sum_{j=2}^{i}\ell_{nj}^{\nu}$, from (61) and

$$\ell_{n1}^{\nu} = g_y(y_n^0, z_n^0)y_n^{\nu+1} + g_z(y_n^0, z_n^0)z_n^{\nu+1} + \psi_{\nu+1}(y_n^0, z_n^0, \dots, y_n^{\nu}, z_n^{\nu}),$$

we get $\ell_{ni}^2 = \mathcal{O}(h^{-1})$. Together with (18), and by (72), it follows that we obtain a perturbed IMEX R-K method which is of the form (69). Therefore, in the case of Theorem 3.8 in [13], this yields

$$\begin{aligned} &\|\Delta Y_{ni}\| \le C(\|\Delta y_n\| + h\|\Delta z_n\|) + \mathcal{O}(\varepsilon^3), \\ &\|\Delta Z_{ni}\| \le C\left(\|\Delta y_n\| + \left(\frac{\varepsilon}{h} + h\right)\|\Delta z_n\|\right) + \mathcal{O}(\varepsilon^3/h), \end{aligned} \tag{73}$$

provided that Δy_n and Δz_n are of size $\mathcal{O}(h)$. The justification of these assumptions follows by induction on n where $\Delta y_0 = \mathcal{O}(\varepsilon^3)$ and $\Delta z_0 = \mathcal{O}(\varepsilon^3)$ and from $\Delta y_n = \mathcal{O}(\varepsilon^3/h)$, $\Delta z_n = \mathcal{O}(\varepsilon^3/h)$, because $\nu = 2$.

Moreover, we prove the recursion

$$\begin{pmatrix} \|\Delta y_{n+1}\| \\ \|\Delta z_{n+1}\| \end{pmatrix} \le \begin{pmatrix} 1 + \mathcal{O}(h) & \mathcal{O}(\varepsilon + h^2) \\ \mathcal{O}(1) & \alpha + \mathcal{O}(h) \end{pmatrix} \begin{pmatrix} \|\Delta y_n\| \\ \|\Delta z_n\| \end{pmatrix} + \begin{pmatrix} \mathcal{O}(\varepsilon^3) \\ \mathcal{O}(\varepsilon^3/h) \end{pmatrix}. \tag{74}$$

The value $\alpha < 1$ is justified in [4] (see also [13]).

Second, in solving the second relation in (74) we use the result of Lemma 5.1 where we emphasize that the method is stiffly accurate and $\tilde{b}_i = b_i$ for all i. Of course, for type ARS, we have again the estimates (71).

Now by combining Theorems 5.1, 6.1, and 7.1, Theorem 3.1 follows. Theorem 3.2 follows from Theorems 5.2, 6.2, and 7.2. Finally, Corollary 3.1 follows from Corollary 5.1 and from the remarks of Theorems 6.2 and 7.2.

8. Conclusions. A study of the global error for different types of IMEX R-K methods has been investigated for a class of singular perturbation problems (SPPs). This asymptotic analysis enables us to obtain convergence results, based on the smoothness of the solution, giving error bounds for several classes of IMEX R-K methods. In particular, the use of DAE techniques, when applied to the stiff case $\Delta t \gg \varepsilon$, was found to give optimal estimates describing the structure of the solutions of SPPs. Concerning the van der Pol equation, numerical results reveal order reduction for all methods in the second (algebraic) component of the solution for small values of the stiffness parameter ε and likewise an order reduction in the first (differential) component when $\tilde{b}_i$ is not equal to b_i for all i. In fact, the hypothesis $\tilde{b}_i = b_i$ represents the only remedy for preserving the classical order for the differential component of the solution. Also, when ε is sufficiently small, and for a given set of suitable assumptions, we obtain numerical results which display improved error estimates in the algebraic component for some IMEX R-K methods appearing in the literature. These results lead us to develop new IMEX R-K methods that work uniformly for a wide range of values of the stiffness parameter ε. In future work we shall introduce new order conditions for the construction of these IMEX R-K methods (see [5]) and study their stability properties.

Acknowledgments. The author wishes to express his gratitude and appreciation to Prof. Ernst Hairer, who generously spent his time to guide him in this work through numerous suggestions and enlightening discussions, during the author's stay at the mathematics department of the University of Geneva. He also thanks the two unknown referees for their critical remarks on the first draft of this paper.

REFERENCES

[1] U. Ascher, S. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time-dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[2] U. Ascher, S. Ruuth, and R. J. Wetton, *Implicit-explicit methods for time dependent PDE's*, Appl. Numer. Math., 32 (1995), pp. 797–823.

[3] J. G. Blom, W. Hundsdorfer, and J. G. Verwer, *An implicit-explicit approach for atmospheric transport-chemistry problems*, Appl. Numer. Math., 20 (1996), pp. 191–209.

[4] S. Boscarino, *On the Uniform Accuracy of Implicit-Explicit Runge-Kutta Methods*, Ph.D. Thesis, Mathematics for the Technology, Department of Mathematics and Computer Science, University of Catania, Italy, 2005.

[5] S. Boscarino, *Uniformly accurate implicit-explicit (IMEX) Runge-Kutta schemes*, submitted.

[6] R. E. Caflisch, S. Jin, and G. Russo, *Uniformly accurate schemes for hyperbolic systems with relaxation*, SIAM J. Numer. Anal., 34 (1997), pp. 246–281.

[7] S. L. Campbell and C. W. Gear, *The index of general nonlinear DAEs*, Numer. Math., 72 (1995), pp. 173–196.

[8] M. H. Carpenter and C. A. Kennedy, *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math., 44 (2003), pp. 139–181.

[9] G. Q. Chen, C. D. Levermore, and T.-P. Liu, *Hyperbolic conservation laws with stiff relaxation terms and entropy*, Comm. Pure Appl. Math., 47 (1994), pp. 787–830.

[10] J. Frank, W. Hundsdorfer, and J. G. Verwer, *On the stability of implicit-explicit linear multistep methods*, Appl. Numer. Math., 25 (1997), pp. 193–205.

[11] C. W. Gear, *Differential algebraic equations, indices, and integral algebraic equation*, SIAM J. Numer. Anal., 27 (1990), pp. 1527–1534.

[12] E. Hairer, Ch. Lubich, and M. Roche, *Error of Runge Kutta methods for stiff problems via differential algebraic equations*, BIT, 28 (1988), pp. 678–700.

[13] E. Hairer and G. Wanner, *Solving Ordinary Differential Equation* II: *Stiff and Differential Algebraic Problems*, 2nd ed., Springer Ser. Comput. Math. 14, Springer-Verlag, New York, 1991, 1996.

[14] W. Hundsdorfer and J. Jaffré, *Implicit-explicit time stepping with spatial discontinuous finite elements*, Appl. Numer. Math., 45 (2003), pp. 231–254.

[15] S. F. Liotta, V. Romano, and G. Russo, *Central schemes for balance laws of relaxation type*, SIAM J. Numer. Anal., 38 (2000), pp. 1337–1356.

[16] R. E. O'Malley, Jr., *Introduction to Singular Perturbations*, Academic Press, New York, 1974.

[17] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes for stiff systems of differential equations*, in Recent Trends in Numerical Analysis, Adv. Theory Comput. Math. 3, Nova Sci. Publ., Huntington, NY, 2001, pp. 269–288.

[18] L Pareschi and G. Russo, *High order asymptotically strong-stability-preserving methods for hyperbolic systems with stiff relaxation*, in Hyperbolic Problems: Theory, Numerics, Applications, Springer-Verlag, Berlin, 2003, pp. 241–251.

[19] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxations*, J. Sci. Comput., 25 (2005), pp. 129–155.

[20] A. N. Tikhonov, A. B. Vasl'eva, and A. G. Sveshnikov, *Differential Equations*, translated from the Russian by A. B. Sossinskij, Springer-Verlag, Berlin, 1985.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1622–1638

AN ENERGY- AND HELICITY-CONSERVING FINITE ELEMENT SCHEME FOR THE NAVIER–STOKES EQUATIONS*

LEO G. REBHOLZ†

Abstract. We present a new finite element scheme for solving the Navier–Stokes equations that exactly conserves both energy ($\int_\Omega u^2$) and helicity ($\int_\Omega u \cdot (\nabla \times u)$) in the absence of viscosity and external force. We prove stability, exact conservation, and convergence for the scheme. Energy and helicity are exactly conserved by using a combination of the usual (convective) form with the rotational form of the nonlinearity and solving for both velocity and a projected vorticity in a trapezoidal time discretization. Numerical results are presented that compare the scheme to the usual trapezoidal schemes.

Key words. Navier–Stokes, conservation, helicity, energy, rotational form, fully discrete

AMS subject classifications. 76D05, 76M10, 76U05, 65M12, 35L65

DOI. 10.1137/060651227

1. Introduction.

It is well known that the Navier–Stokes equations (NSE) conserve energy ($E = \frac{1}{2}\int_\Omega |u|^2$) in the absence of viscosity and external force. Conserving energy in numerical schemes for the NSE not only leads to stability for the scheme, but also is necessary for physical relevance of solutions. In rotational flows, however, other integral invariants are also important. In two dimensions enstrophy ($Ens = \frac{1}{2}\int_\Omega |\nabla \times u|^2$) and in three dimensions helicity ($H = \int_\Omega u \cdot (\nabla \times u)$) are also conserved quantities of the NSE when viscosity and external force are not present [4], [6].

Although the importance of enstrophy in two-dimensional flow has been known for many years, the importance of helicity in understanding three-dimensional turbulent phenomena has only recently been recognized. The inviscid invariance of helicity was discovered by Moreau in 1961 [15], and a famous 1992 review paper [14] of Moffatt and Tsoniber finds helicity to have a status of importance comparable to energy for three-dimensional flows. Hence, accurate helicity treatment, in addition to accurate energy treatment, should be a goal for numerical schemes for three-dimensional rotational flows.

Helicity in true fluid flow is created and dissipated only by viscous and body forces, and thus for a numerical scheme for the NSE to have physical relevance, the nonlinearity in the scheme should not create or dissipate helicity. If a scheme conserves helicity in the inviscid case (even though an NSE scheme is typically designed for viscous flows only), then under viscous conditions the scheme's nonlinearity will not nonphysically create or dissipate helicity.

For two-dimensional flows, schemes such as the classical Arakawa scheme [1] have existed for over forty years which conserve both energy and enstrophy (this and all future references to E/H/Ens conservation implicitly refer to the case of no viscosity or external force). By conserving energy and enstrophy for inviscid flows, the Arakawa

*Received by the editors January 30, 2006; accepted for publication (in revised form) March 9, 2007; published electronically August 15, 2007. This research was partially supported by NSF grants DMS 0508260 and 0207627.

http://www.siam.org/journals/sinum/45-4/65122.html

†Department of Mathematics, University of Pittsburgh, Pittsburgh, PA 15260 (ler6@pitt.edu, http://www.math.pitt.edu/∼ler6).

scheme is accurate over longer time intervals because it ensures only viscosity and external force (and not the nonlinearity) create and dissipate energy and enstrophy. For three-dimensional flows, however, it was not until 2004 that Liu and Wang developed the first scheme that conserves both energy and helicity. In [13], they present an energy- and helicity-preserving scheme for axisymmetric flows and show that this dual conservation eliminates the need for excessive numerical viscosity. It is their work which motivated this article.

In this report, we present a new finite element scheme that conserves both energy and helicity for general, viscous flows. Our development of the scheme herein is for periodic boundaries (and hence we use a box for the domain Ω). The key features that allow the scheme to conserve both energy and helicity are the use of the projection of the vorticity in the scheme and a new variational formulation of the nonlinearity that vanishes when tested against either the velocity or projected vorticity. For nonperiodic boundary conditions, helicity is not necessarily globally conserved. On the other hand, helicity generation and helicity flux are equally important for nonperiodic problems, and a numerical method should not generate spurious helicity through its discretization of the nonlinear term.

Remark 1.1. Helicity is not necessarily globally conserved for more general boundary conditions. Consider the Euler equations on $\Omega = (0, L)^3$. Multiply by the vorticity $w := (\nabla \times u)$ and integrate over the domain:

$$\int_\Omega u_t \cdot w + \int_\Omega \left(\frac{1}{2}\nabla u^2 - u \times w\right) \cdot w + \int_\Omega \nabla p \cdot w = 0. \tag{1}$$

This reduces via integration by parts and vector identities to

$$\int_\Omega u_t \cdot w + \int_{\partial\Omega} \left(p + \frac{1}{2}u^2\right)(w \cdot n) = 0. \tag{2}$$

The boundary integral in (2) can vanish without periodicity (e.g., if $w \cdot n = 0$ is imposed), but the resulting equation, $\int_\Omega u_t \cdot w = 0$, still does not imply the conservation of helicity since integrating by parts with u decomposed as $u := \langle u_1, u_2, u_3\rangle$ shows

$$\begin{aligned} H(T) - H(0) &= \int_0^T \frac{d}{dt}H(t) = \int_0^T \frac{d}{dt}\int_\Omega u \cdot w = \int_0^T \int_\Omega (u_t \cdot w + u \cdot w_t) \\ &= \left(\int_0^L \int_0^L \left(\left(u_1 u_3|_{y=0}^{y=L}\, dz\, dx\right) + \left(u_1 u_2|_{z=0}^{z=L}\, dy\, dx\right) + \left(u_2 u_3|_{x=0}^{x=L}\, dz\, dy\right)\right)\right)|_{t=0}^{t=T}. \end{aligned} \tag{3}$$

Thus we see that helicity is conserved for periodic boundary conditions or for zero (Dirichlet) boundary conditions (with $w \cdot n = 0$ imposed on the boundary), but not necessarily conserved in general.

This article is arranged as follows: We present the energy- and helicity-conserving scheme in section 3, after providing the necessary notation in section 2. Section 4 gives a rigorous numerical analysis for the scheme, section 5 presents numerical results, and section 6 presents conclusions.

2. Notation and preliminaries. $(\cdot, \cdot)$ and $\|\cdot\|$ denote the usual L^2 inner product and norm, respectively, and $\|\cdot\|_k$ the $H^k(\Omega)$ norm. $\|\cdot\|_\infty$ will denote the usual $L^\infty(\Omega)$ norm, and all other norms that appear in this article will be clearly labeled with subscripts. The domain Ω we use is the box $(0, L)^3$.

DEFINITION 2.1. *The Hilbert space $H^1_\#(\Omega)$ will be defined as*

$$H^1_\# := \left(v \in H^1 \ : \ v \text{ periodic on } \Omega, \ \int_\Omega v \ dx = 0 \right).$$

This is the natural velocity space for the NSE with periodic boundary conditions, as discussed in [11] and [12]. Note that velocities in this space automatically conserve momentum ($\int_\Omega u$), i.e., if $u \in H^1_\#$, then $\frac{d}{dt}\int_\Omega u = 0$. This is physically important because the NSE (with periodic boundary conditions) also conserve momentum [6].

Let $T^h = T^h(\Omega)$ be a conforming finite element mesh on Ω. Define the spaces $(X^h, Q^h) \subset (H^1_\#, L^2_0)$ to be conforming velocity, pressure finite element spaces (see, e.g., [3], [5], or [7] for examples) that satisfy the discrete inf-sup condition (also known as the LBB condition)

$$0 < \beta \leq \inf_{q \in Q^h} \sup_{v \in X^h} \frac{(q, \nabla \cdot v)}{\|v\|_1 \|q\|}. \tag{4}$$

Define V^h to be the space of discretely divergence-free, zero-mean, periodic functions.

$$V^h = \{ v \in X^h \ : \ (\nabla \cdot v, q) = 0 \ \forall q \in Q^h \}.$$

Since V^h is a closed subspace of $H^1_\#(\Omega)$, we have also that V^h is a Hilbert space, hence the following result.

LEMMA 2.2. *Let $u^h \in V^h$. Then there exists a unique $w^h \in V^h$ satisfying*

$$(w^h, v) = (\nabla \times u^h, v) \ \forall v \in V^h. \tag{5}$$

Proof. Since $u^h \in V^h \subset H^1(\Omega)$, it follows that $\nabla \times u^h \in L^2(\Omega)$. Since V^h is a closed subset of the Hilbert space $L^2(\Omega)$, the Riesz representation theorem implies the existence and uniqueness of a solution w^h to (5). ∎

The next lemma shows how an elementary property of the cross product can be used for double skew-symmetry of a trilinear term.

LEMMA 2.3. *Let $u^h, w^h \in X^h$. Then*

$$(u^h \times w^h, u^h) = (u^h \times w^h, w^h) = 0.$$

Proof. This follows from an elementary property of the cross product; the cross product of two vectors is perpendicular to each of them. ∎

The significance of this lemma is that in a finite element scheme, the trilinear form $(u^h \times w^h, v^h)$ will vanish when $v^h = u^h$ or w^h. Such a trilinear form has significance in the NSE if the rotational form of the nonlinearity is used (see, e.g., [6, p. 461] or [17]). Our scheme uses this form and exploits the double skew-symmetry to show the scheme conserves both energy and helicity.

The discrete Gronwall lemma will also be an essential tool in the error analysis; we present it now.

LEMMA 2.4 (discrete Gronwall). *Let Δt, H, and a_n, b_n, c_n, d_n (for integers $n \geq 0$) be nonnegative numbers such that*

$$a_l + \Delta t \sum_{n=0}^{l} b_n \leq \Delta t \sum_{n=0}^{l} d_n a_n + \Delta t \sum_{n=0}^{l} c_n + H \text{ for } l \geq 0. \tag{6}$$

Suppose that $\Delta t d_n < 1\ \forall n$. *Then*

$$a_l + \Delta t \sum_{n=0}^{l} b_n \le exp\left(\Delta t \sum_{n=0}^{l} \frac{d_n}{1-\Delta t d_n}\right)\left(\Delta t \sum_{n=0}^{l} c_n + H\right) \text{ for } l \ge 0. \tag{7}$$

Proof. See [9], for example, for the proof of this well-known lemma. □

We end this section with definitions for discrete energy and helicity.

DEFINITION 2.5. *We define the discrete energy* E *and helicity* H *to be, at time* t^k,

$$E_h(t^k) = \frac{1}{2}\|u_h^k\|^2,$$
$$H_h(t^k) = (u_h^k, \nabla \times u_h^k).$$

We are now ready to present the scheme.

3. An energy- and helicity-preserving scheme for periodic flows. The energy- and helicity-preserving finite element scheme we study is composed of a trapezoidal time discretization with a nonlinearity that is doubly skew-symmetric. Let Δt denote the timestep, $t^k = k\Delta t$, $t^{k+1/2} = (k+\frac{1}{2})\Delta t$, and u_h^k the approximation to $u(x,t^k)$. $u_h^{k+1/2}$ will denote

$$u_h^{k+1/2} := \frac{1}{2}(u_h^{k+1} + u_h^k),$$

and $f^{n+1/2}(x) := f(t^{n+1/2},x) \in V^{h,*}$. $T = Nk$ denotes the final time. Given $u_h^0 \in V^h$, define w_h^0 to be the (unique in V^h by Lemma 2.2) solution of $(w_h^0, v) = (\nabla \times u_h^0, v)\ \forall v \in V^h$, and find $(u_h^k; w_h^k; p_h^k) \in X_h \times V_h \times Q_h$ for $k = 1, \ldots, N$, satisfying

$$\frac{1}{\Delta t}(u_h^{n+1}, v) + (u_h^{n+1/2} \times w_h^{n+1/2}, v) - (p_h^{n+1/2}, \nabla \cdot v) + \frac{\nu}{2}(\nabla u_h^{n+1/2}, \nabla v) + \frac{\nu}{2}(w_h^{n+1/2}, \nabla \times v) = (f^{n+1/2}, v) + \frac{1}{\Delta t}(u_h^n, v)\ \forall v \in X^h, \tag{8}$$

$$(\nabla \cdot u_h^{n+1}, q) = 0\ \forall q \in Q^h, \tag{9}$$

$$(w_h^{n+1} - \nabla \times u_h^{n+1}, \chi) = 0\ \forall \chi \in V^h. \tag{10}$$

We now prove the conservation properties of the scheme: energy and helicity are exactly conserved in the absence of viscosity and external force.

LEMMA 3.1. *The scheme* (8)–(10) *conserves energy and helicity in the absence of viscosity and body force, that is,* $E_h(t^n) = E_h(t^0)$ *and* $H_h(t^n) = H_h(t^0)\ \forall n \le N$, *provided* $\nu = f = 0$.

Proof. For the conservation of energy, set $v = u_h^{n+1/2}$ and $\nu = f = 0$ in (8). This gives

$$(u_h^{n+1}, u_h^{n+1/2}) = (u_h^n, u_h^{n+1/2}). \tag{11}$$

By expanding the $u_h^{n+1/2}$ terms in (11), we have

$$\frac{1}{2}\|u_h^{n+1}\|^2 + \frac{1}{2}(u_h^{n+1}, u_h^n) = \frac{1}{2}\|u_h^n\|^2 + \frac{1}{2}(u_h^n, u_h^{n+1}), \tag{12}$$

$$E_h(t^{n+1}) = E_h(t^n), \tag{13}$$

which implies that $E_h(t^n) = E_h(t^0)$.

For helicity conservation, set $v = w_h^{n+1/2}$ in (8). The pressure term vanishes since $w_h^n, w_h^{n+1} \in V^h$, and so after setting $\nu = f = 0$, we are left with

$$\frac{1}{2}(u_h^{n+1}, w_h^{n+1}) + \frac{1}{2}(u_h^{n+1}, w_h^n) = \frac{1}{2}(u_h^n, w_h^n) + \frac{1}{2}(u_h^n, w_h^{n+1}). \tag{14}$$

Using (10) and integrating by parts, we have the following identities for the terms in (14):

$$(u_h^{n+1}, w_h^{n+1}) = (u_h^{n+1}, \nabla \times u_h^{n+1}) = H_h(t^{n+1}), \tag{15}$$

$$(u_h^n, w_h^n) = (u_h^n, \nabla \times u_h^n) = H_h(t^n), \tag{16}$$

$$(u_h^{n+1}, w_h^n) = (u_h^n, w_h^{n+1}). \tag{17}$$

Thus (14) can be rewritten as

$$H_h(t^{n+1}) = H_h(t^n), \tag{18}$$

which implies that $H_h(t^n) = H_h(t^0)$. ∎

Lemma 3.1 shows that only viscous and external forces create and dissipate energy and helicity in the scheme when $\nu > 0$ and nonzero f. This is qualitatively important for the physical relevance of the scheme's solution because this is also true for helicity in the NSE (true fluid flow):

$$H(T) = H(0) + \int_0^T ((f(t), \nabla \times u(t)) + \nu(\nabla u(t), \nabla(\nabla \times u(t)))) \, dt. \tag{19}$$

In the energy- and helicity-conserving scheme, we have that

$$H_h^N = H_h^0 + \sum_{n=0}^{N-1} (f(t^{n+1/2}), w_h^{n+1/2}) + \frac{\nu}{2}(\nabla u_h^{n+1/2}, \nabla w_h^{n+1/2}) + \frac{\nu}{2}(w_h^{n+1/2}, \nabla \times w_h^{n+1/2}). \tag{20}$$

However, schemes that do not conserve helicity will not necessarily share this physical property. For example, in a trapezoidal scheme for the NSE that does not conserve helicity (e.g., usual Crank–Nicholson in rotational form (67)), the nonlinear term will not vanish when the test function is chosen to be the projection of the curl. In this scheme,

$$\begin{aligned} H_h^N = H_h^0 &+ \sum_{n=0}^{N-1} (f(t^{n+1/2}), w_h^{n+1/2}) + \nu(\nabla u_h^{n+1/2}, \nabla w_h^{n+1/2}) \\ &+ (u_h^{n+1/2} \times ((\nabla \times u_h^{n+1/2}) - w_h^{n+1/2}), w_h^{n+1/2}). \end{aligned} \tag{21}$$

It is the last term in (21) that is nonphysical and thus can cause numerical errors and a loss of physical fidelity over long time intervals in the usual trapezoidal scheme.

The following lemma shows that the energy- and helicity-conserving scheme is also bounded by its data.

LEMMA 3.2. *Solutions to the discrete scheme* (8)–(10) *satisfy*

$$\|u_h^N\|^2 + \Delta t \sum_{n=0}^{N-1} \left(\frac{\nu}{2} \|\nabla u_h^{n+1/2}\|^2 + \nu \|w_h^{n+1/2}\|^2 \right) \leq \|u_h^0\|^2 + \frac{2\Delta t}{\nu} \sum_{n=0}^{N-1} \|f^{n+1/2}\|_*^2. \tag{22}$$

Proof. Set $v = u_h^{n+1/2}$ in (8), $q = p_h^{n+1/2}$ in (9), and add the equations. This gives

$$\frac{1}{2\Delta t}\|u_h^{n+1}\|^2 + \frac{1}{2\Delta t}(u_h^{n+1}, u_h^n) + \frac{\nu}{2}\|\nabla u_h^{n+1/2}\|^2 + \frac{\nu}{2}(w_h^{n+1/2}, \nabla \times u^{n+1/2}) = (f^{n+1/2}, u_h^{n+1/2}) + \frac{1}{2\Delta t}\|u_h^n\|^2 + \frac{1}{2\Delta t}(u_h^n, u_h^{n+1}). \tag{23}$$

Note that $(w_h^{n+1/2}, \nabla \times u_h^{n+1/2}) = \|w_h^{n+1/2}\|^2$ since (10) must hold for $(n+1)$ replaced by (n), and thus also for $(n+1)$ replaced by $(n+1/2)$. By making this substitution, (23) reduces to

$$\frac{1}{2\Delta t}\|u_h^{n+1}\|^2 + \frac{\nu}{2}\|\nabla u_h^{n+1/2}\|^2 + \frac{\nu}{2}\|w_h^{n+1/2}\|^2 = (f^{n+1/2}, u_h^{n+1/2}) + \frac{1}{2\Delta t}\|u_h^n\|^2. \tag{24}$$

Next we use the bound $(f^{n+1/2}, u_h^{n+1/2}) \le \frac{\nu}{4}\|\nabla u_h^{n+1/2}\|^2 + \frac{1}{\nu}\|f^{n+1/2}\|_*^2$, and sum from $n = 0, \ldots, (N-1)$, yielding

$$\frac{1}{2\Delta t}\|u_h^N\|^2 + \sum_{n=0}^{N-1}\left(\frac{\nu}{2}\|\nabla u_h^{n+1/2}\|^2 + \nu\|w_h^{n+1/2}\|^2\right) \le \frac{1}{2\Delta t}\|u_h^0\|^2 + \frac{1}{\nu}\sum_{n=0}^{N-1}\|f^{n+1/2}\|_*^2. \tag{25}$$

Now multiplying both sides by $(2\Delta t)$ proves the lemma. ∎

3.1. Existence of solutions for the scheme. Given $u_h^n, w_h^n \in V^h$, a nonlinear system must be solved for the approximations at time level $n+1$. The question arises, Does that system have a solution? In other words, does imposing two integral invariants overdetermine the system for u_h^{n+1}, w_h^{n+1}? The answer is that solutions to (8)–(10) do exist, as we will show in this section.

For clarity, we show existence for the equivalent nonlinear problem: Given $\nu, \Delta t > 0$, $f^{n+1/2} \in V^{h,*}$, and $u_h^n \in V^h$, find $(u_h; w_h) \in V^h \times V^h$ satisfying

$$\frac{2}{\Delta t}(u_h, v) + (u_h \times w_h, v) + \frac{\nu}{2}(\nabla u_h, \nabla v) + \frac{\nu}{2}(w_h, \nabla \times v) = (f^{n+1/2}, v) + \frac{2}{\Delta t}(u_h^n, v)\ \forall v \in V^h, \tag{26}$$

$$(w_h - \nabla \times u_h, \chi) = 0\ \forall \chi \in V^h. \tag{27}$$

This form of the scheme is derived from (8)–(10) by defining $u_h := u_h^{n+1/2}$, $w_h := w_h^{n+1/2}$ and restricting the test functions to V^h. Equations (26)–(27) are equivalent to (8)–(10). To show solutions exist, we formulate (26)–(27) as a fixed point problem, $y = F(y)$, and use the Leray–Schauder fixed point theorem. We will first prove several preliminary lemmas, followed by a theorem which proves that a solution to (26)–(27) exists.

LEMMA 3.3. *For $\nu, \Delta t > 0$, there exists a unique solution (u_h, w_h) to the following: Given $g \in V^{h,*}$, find $(u_h; w_h) \in V^h \times V^h$ satisfying*

$$\frac{2}{\Delta t}(u_h, v) + \frac{\nu}{2}(\nabla u_h, \nabla v) + \frac{\nu}{2}(w_h, \nabla \times v) = (g, v)\ \forall v \in V^h, \tag{28}$$

$$(w_h - \nabla \times u_h, \chi) = 0\ \forall \chi \in V^h. \tag{29}$$

Proof. We will prove uniqueness of solutions to (28)–(29) by showing only that the trivial solution solves the homogeneous problem, which will also imply the existence of solutions to the finite-dimensional problem. Since the space V^h includes only zero-mean functions, functions and operators are uniquely solvable, and thus we need not consider the adjoint problem. Choose $v = u_h$ in (28), $\chi = w_h$ in (29), and substitute (29) into (28). This gives

$$\frac{2}{\Delta t}\|u_h\|^2 + \frac{\nu}{2}\|\nabla u_h\|^2 + \frac{\nu}{2}\|w_h\|^2 = 0, \tag{30}$$

which implies $u_h = w_h = 0$, i.e., uniqueness. ∎

This lemma allows us to define a solution operator to (28)–(29).

DEFINITION 3.4. *We define the solution operator $T : V^{h,*} \to (V^h \times V^h)$ to be the solution operator of* (28)–(29)*: if $g \in V^{h,*}$, then $T(g) = (u_h; w_h)$ solves* (28)–(29).

We have that T is well defined by the previous lemma, and we now prove it is also bounded and linear.

LEMMA 3.5. *The solution operator T is linear, bounded, and continuous.*

Proof. The linearity of T follows from the fact that T is a solution operator to a linear problem. To see that T is bounded (and thus continuous since it is linear), we let $v = u_h,\ \chi = w_h$ in (28)–(29), multiply (29) by $\frac{\nu}{2}$, and add the equations. This gives

$$\frac{2\|u_h\|^2}{\Delta t} + \frac{\nu}{4}\|\nabla u_h\|^2 + \frac{\nu}{2}\|w_h\|^2 \le \frac{1}{\nu}\|g\|_*^2.$$

Then since u_h, w_h are finite-dimensional, $\|u_h, w_h\|_{V^h \times V^h} \le C\|g\|_*$. Hence,

$$\|T\| = \sup_{g \in V_*^h} \frac{\|T(g)\|}{\|g\|_*} = \sup_{g \in V_*^h} \frac{\|u_h, w_h\|_{V^h \times V^h}}{\|g\|_*} \le C. \qquad ∎$$

We next define the operator N. The function F that will be used in the formulation of the fixed point problem will be a composition of T and N.

DEFINITION 3.6. *We define the operator N on $(V^h \times V^h)$ by*

$$N(u_h; w_h) := f^{n+1/2} + \frac{2}{\Delta t}u_h^n + u_h \times w_h.$$

We now prove properties for N necessary for use in Leray–Schauder.

LEMMA 3.7. *For the nonlinear operator N, we have that $N : V^h \times V^h \to V^{h,*}$, N is bounded, and N is continuous.*

Proof. To show N maps as stated, we let $(u_h, w_h) \in V^h \times V^h$ and write

$$\|N(u_h; w_h)\|_* = \sup_{v \in V^h} \frac{(N(u_h; w_h), v)}{\|v\|_1}.$$

From the definition of N, we have that $\frac{(f^{n+1/2}, v) + (2(\Delta t)^{-1}u_h^n, v)}{\|v\|_1} \le \|f\|_* + C_1\|u_h^n\| \le C_2$, and that

$$\frac{(u_h \times w_h, v)}{\|v\|_1} \le \|u_h\|_\infty \|w_h\| \le C_3$$

since u_h and w^h are given to be in V^h, and all norms are equivalent in finite dimension. Hence $\|N(u_h, w_h)\|_* < C$, and so N maps as stated. Note we have also proven that N is bounded.

The equivalence of norms in finite dimension is also key in showing that N is continuous, as

$$\|N(u;w) - N(u_k;w_k)\|_* \le \|u \times (w - w_k)\|_* + \|(u - u_k) \times w_k\|_* \tag{31}$$

$$\le \|u\|_\infty \|w - w_k\| + \|w_k\|_\infty \|u - u_k\|, \tag{32}$$

and thus $\to 0$ as $\|(u;w) - (u_k;w_k)\| \to 0$. ∎

We are now ready to define the operator F, which will formulate (26)–(27) as a fixed point problem.

DEFINITION 3.8. *Define the operator $F : (V^h \times V^h) \to (V^h \times V^h)$ to be the composition of T and N: $F(y) = T(N(Y))$.*

LEMMA 3.9. *F is well defined and compact, and a solution to $y = F(y)$ solves* (26)–(27).

Proof. F is well defined because N and T are. The fact that F is compact follows from the fact that both N and T are continuous and bounded. It can easily be seen that a fixed point of F solves (26)–(27) by expanding F. ∎

We are now ready to prove existence to (26)–(27).

THEOREM 3.10. *Let $y_\lambda = (u_\lambda; w_\lambda) \in V^h \times V^h$ and consider the family of fixed point problems $y_\lambda = \lambda F(y_\lambda)$, $0 \le \lambda \le 1$. A solution y_λ to any of these fixed point problems satisfies $\|y_\lambda\| < K$, independent of λ. Since F is compact, and fixed points of F solve* (26)–(27)*, by the Leray–Schauder theorem there exist solutions to* (26)–(27).

Proof. All we have to show to prove this theorem is that solutions to $y_\lambda = \lambda F(y_\lambda)$ are bounded independent of λ. Using the definition of F and the linearity of T we have that

$$y_\lambda = \lambda F(y_\lambda) = \lambda T(N(y_\lambda)) = T\left(\lambda N\left(y_\lambda\right)\right) = T\left(\lambda\left(f^{n+1/2} + \frac{2}{\Delta t}u_h^n + u_\lambda \times w_\lambda\right)\right),$$

which implies that

$$\frac{2}{\Delta t}(u_\lambda, v) - \lambda(u_\lambda \times w_\lambda, v) + \frac{\nu}{2}(\nabla u_\lambda, \nabla v) + \frac{\nu}{2}(w_\lambda, \nabla \times v) = (\lambda f^{n+1/2}, v) + \frac{2\lambda}{\Delta t}(u_h^n, v)\ \forall v \in V^h, \tag{33}$$

$$(w_\lambda - \nabla \times u_\lambda, \chi) = 0\ \forall \chi \in V^h. \tag{34}$$

Multiply (34) by $\frac{\nu}{2}$, let $\chi = w_\lambda$ in (34), $v = u_\lambda$ in (33), and add the equations. Similarly to the stability estimate, this gives

$$\frac{1}{\Delta t}\|u_\lambda\|^2 + \frac{\nu}{4}\|\nabla u_\lambda\|^2 + \frac{\nu}{2}\|w_\lambda\|^2 \le \lambda^2\left(\frac{1}{\nu}\|f^{n+1/2}\|^2 + \frac{1}{\Delta t}\|u_h^n\|^2\right) \le \left(\frac{1}{\nu}\|f^{n+1/2}\|^2 + \frac{1}{\Delta t}\|u_h^n\|^2\right) \le C, \tag{35}$$

which is a bound independent of λ. Thus the theorem is proven. ∎

We have now shown that the scheme (8)–(10) preserves energy and helicity when $\nu = f = 0$, that it is stable, and that it admits solutions. The final step is an error analysis for the scheme.

4. Error analysis of the scheme. This section presents a theorem for the convergence of the scheme, followed by the proof. The restriction that the theorem places on the timestep is for the use of the discrete Gronwall lemma. Although we found its use necessary in the proof, it is widely believed that it gives a gross underestimate of the largest timestep one can use, and we expect the same asymptotic error. Without the projection step, the proof of the theorem is fairly standard; the smoothness assumptions we make are also fairly standard and are similar to those found in, for example, [10], [18].

THEOREM 4.1. *For* $u \in L^\infty(0,T;W_4^{k+1}) \cap W_2^3(0,T;L^2) \cap W_4^2(O,T,W_2^1)$, $p \in L^4(0,T;W^k) \cap W_2^2(0,T;L^2)$, $f \in L^2(0,T,V^{h,*})$ *satisfying the NSE on the periodic box* $\Omega = (0,L)^3$, $(u_h^n; w_h^n)$ *given by* (8)–(10) *with velocity-pressure spaces chosen as* P_k, P_{k-1} $(k > 1)$, *and timestep* Δt *sufficiently small (for Gronwall's inequality), we have that*

(36)
$$\|u(T) - u_h^N\|^2 + \frac{3\nu\Delta t}{4}\sum_{n=0}^{N-1}\left(\frac{1}{2}\|\nabla(u^{n+1/2} - u_h^{n+1/2})\|^2 + \|w^{n+1/2} - w_h^{n+1/2}\|^2\right) \leq C(u,p,\nu^{-3},\Omega,T)(\Delta t^4 + h^{2k}).$$

Remark 4.2. Under the smoothness assumptions of the theorem, the constant in the error estimate can be prohibitively large for small ν, as the constant contains ν^{-3} terms. This constant can be improved (i.e., to contain ν^{-1} instead of ν^{-3}), but requires the data to be assumed very smooth.

Remark 4.3. Under the assumptions of the theorem, a pressure estimate can also be obtained: $\frac{1}{\Delta t}\sum_{n=0}^{N-1}\|p(t^{n+1/2}) - p_h^{n+1/2}\| \leq C(u,p,\nu^{-3},\Omega,\beta,T)(\Delta t^2 + h^k)$. Enforcing helicity conservation in this setting does not improve this estimate versus the pressure estimate in the usual trapezoidal scheme; that is, this is the expected result and its proof (which is omitted) follows in the standard way (see, e.g., [8], [9]).

Proof. The proof of the theorem is divided into the following parts. We first develop the error equations by subtracting our scheme from the NSE. The error is then split into parts in and out of the finite element spaces. This is followed by bounding the error in the space by interpolation error, and the proof concludes by bounding the total error. Note that we require that the spaces X^h, Q^h satisfy the discrete inf-sup condition; with such spaces, and since $(w_h^0 - \nabla\times u_h^0, v) = 0\ \forall v \in V^h$, the energy- and helicity-conserving scheme is equivalent to finding solutions $u^n, w^n \in V^h$, $n = 0,\ldots,N$, satisfying

(37)
$$\frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v) - (u_h^{n+1/2}\times w_h^{n+1/2}, v) + \frac{\nu}{2}(\nabla u_h^{n+1/2}, \nabla v) + \frac{\nu}{2}(w_h^{n+1/2}, \nabla\times v) = (f^{n+1/2}, v)\ \ \forall v \in V^h,$$

(38)
$$(w_h^{n+1/2} - \nabla\times u_h^{n+1/2}, \chi) = 0\ \ \forall \chi \in V^h.$$

Using the identity $u\cdot\nabla u = \frac{1}{2}\nabla(u^2) - u\times(\nabla\times u)$, and grouping the usual pressure gradient with the $\frac{1}{2}\nabla(u^2)$ term to form the Bernoulli pressure, a periodic solution

$(u;p)$ and $w := \nabla \times u$ of the NSE satisfies

(39)
$$\begin{aligned}
&\frac{1}{\Delta t}(u^{n+1}-u^n,v)-(u^{n+1/2}\times w^{n+1/2},v)+\frac{\nu}{2}(\nabla u^{n+1/2},\nabla v)+\frac{\nu}{2}(w^{n+1/2},\nabla\times v)\\
&-(p(t^{n+1/2}),\nabla\cdot v)=(f^{n+1/2},v)+\left(\frac{u^{n+1}-u^n}{\Delta t}-u_t(t^{n+1/2}),v\right)\\
&-(u^{n+1/2}\times w^{n+1/2}-u(t^{n+1/2})\times w(t^{n+1/2}),v)\\
&+\frac{\nu}{2}(\nabla(u^{n+1/2}-u(t^{n+1/2})),\nabla v)\\
&+\frac{\nu}{2}(w^{n+1/2}-w(t^{n+1/2}),\nabla\times v)\quad\forall v\in V^h.
\end{aligned}$$

Define $e^i := u^i - u_h^i$ and $E^i := w^i - w_h^i$ for $i = n, n+1, n+1/2$, and form the error equations by subtracting the scheme (37), (38) from (39) and $w = \nabla \times u$ to get

(40)
$$\begin{aligned}
&\frac{1}{\Delta t}(e^{n+1}-e^n,v)-(u^{n+1/2}\times E^{n+1/2},v)-(e^{n+1/2}\times w_h^{n+1/2},v)+\frac{\nu}{2}(\nabla e^{n+1/2},\nabla v)\\
&\qquad+\frac{\nu}{2}(E^{n+1/2},\nabla\times v)-(p(t^{n+1/2}),\nabla\cdot v)=IERR(u^n;w^n;v)\quad\forall v\in V^h,
\end{aligned}$$

$$(E^{n+1/2},\chi)-(\nabla\times e^{n+1/2},\chi)=0\quad\forall\chi\in V^h, \tag{41}$$

where the interpolation error in time, $IERR$, is defined by

$$\begin{aligned}
IERR(u^n,w^n,v) &:= \left(\frac{u^{n+1}-u^n}{\Delta t}-u_t(t^{n+1/2}),v\right)\\
&-(u^{n+1/2}\times w^{n+1/2}-u(t^{n+1/2})\times w(t^{n+1/2}),v)\\
&+\frac{\nu}{2}(\nabla(u^{n+1/2}-u(t^{n+1/2})),\nabla v)+\frac{\nu}{2}(w^{n+1/2}-w(t^{n+1/2}),\nabla\times v).
\end{aligned} \tag{42}$$

Next we split the error terms into pieces in and out of V^h. Let U^i and W^i be the projections of u^i and w^i, respectively, into V^h. Then the error terms can be decomposed as

$$e^i=(u^i-U^i)-(u_h^i-U^i)=:\eta^i-\phi_h^i, \tag{43}$$

$$E^i=(w^i-W^i)-(w_h^i-W^i)=:r^i-s_h^i. \tag{44}$$

Note that $(\eta^i, v) = 0$ for $v \in V^h$ by the definition of η^i. Rewriting (40), (41) with this decomposition gives

$$\begin{aligned}
&\frac{1}{\Delta t}(\phi_h^{n+1}-\phi_h^n,v)-(\phi_h^{n+1/2}\times w_h^{n+1/2},v)+\frac{\nu}{2}(\nabla\phi_h^{n+1/2},\nabla v)\\
&+\frac{\nu}{2}(s_h^{n+1/2},\nabla\times v)=(u^{n+1/2}\times s_h^{n+1/2},v)-(u^{n+1/2}\times r^{n+1/2},v)\\
&-(\eta^{n+1/2}\times w_h^{n+1/2},v)+\frac{\nu}{2}(\nabla\eta^{n+1/2},\nabla v)\\
&+\frac{\nu}{2}(r^{n+1/2},\nabla\times v)-(p(t^{n+1/2}),\nabla\cdot v)+IERR(u^n;w^n;v)\quad\forall v\in V^h,
\end{aligned} \tag{45}$$

$$(\nabla\times\phi_h^{n+1},\chi)=(s_h^{n+1/2},\chi)-(r^{n+1/2},\chi)+(\nabla\times\eta^{n+1/2},\chi)=0\quad\forall\chi\in V^h. \tag{46}$$

Let $v = \phi_h^{n+1/2}$ and $\chi = s_h^{n+1/2}$ and combine (45) and (46) to get

$$\begin{aligned}(47)\quad &\frac{1}{2\Delta t}\left(\|\phi_h^{n+1}\|^2 - \|\phi_h^n\|^2\right) + \frac{\nu}{2}\|\nabla\phi_h^{n+1/2}\|^2 + \frac{\nu}{2}\|s_h^{n+1/2}\|^2 \\ &= (u^{n+1/2}\times s_h^{n+1/2}, \phi_h^{n+1/2}) - (u^{n+1/2}\times r^{n+1/2}, \phi_h^{n+1/2}) \\ &\quad - (\eta^{n+1/2}\times w_h^{n+1/2}, \phi_h^{n+1/2}) + \frac{\nu}{2}(\nabla\eta^{n+1/2}, \nabla\phi_h^{n+1/2}) \\ &\quad + \frac{\nu}{2}(r^{n+1/2}, \nabla\times\phi_h^{n+1/2}) + \frac{\nu}{2}(r^{n+1/2}, s_h^{n+1/2}) + \frac{\nu}{2}(\nabla\times\eta^{n+1/2}, s_h^{n+1/2}) \\ &\quad - (p(t^{n+1/2}), \nabla\cdot\phi_h^{n+1/2}) + IERR(u^n; w^n; \phi_h^{n+1/2}).\end{aligned}$$

The terms on the right-hand side of (47) are now majorized in the usual way, using Cauchy–Schwarz and Young's inequalities, and the bound $(u\times w, v) \le C\|a\|_0\|b\|_1\|c\|_{1/2}$. Note that this inequality holds no matter the order of u, w, v (provided the norms exist) due to a well-known vector identity from calculus. We first bound the following right-hand side terms:

$$(48)\quad \frac{\nu}{2}\left|(\nabla\eta^{n+1/2}, \nabla\phi_h^{n+1/2})\right| \le \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2 + C\nu^{-1}\|\nabla\eta^{n+1/2}\|^2,$$

$$(49)\quad \frac{\nu}{2}\left|(r^{n+1/2}, \nabla\times\phi_h^{n+1/2})\right| \le \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2 + C\nu^{-1}\|r^{n+1/2}\|^2,$$

$$(50)\quad \frac{\nu}{2}\left|(r^{n+1/2}, s_h^{n+1/2})\right| \le \frac{\nu}{32}\|s_h^{n+1/2}\|^2 + C\nu^{-1}\|r^{n+1/2}\|^2,$$

$$(51)\quad \frac{\nu}{2}\left|(\nabla\eta^{n+1/2}, s_h^{n+1/2})\right| \le \frac{\nu}{32}\|s_h^{n+1/2}\|^2 + C\nu^{-1}\|\nabla\eta^{n+1/2}\|^2,$$

$$(52)\quad \left|(p(t^{n+1/2}), \nabla\cdot\phi_h^{n+1/2})\right| \le \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2 + C\nu^{-1}\inf_{q\in Q^h}\|p(t^{n+1/2}) - q\|^2.$$

The first of the trilinear terms is bounded by

$$\begin{aligned}(53)\quad \left|(u^{n+1/2}\times s_h^{n+1/2}, \phi_h^{n+1/2})\right| &\le C\|\nabla u^{n+1/2}\|\|s_h^{n+1/2}\|\|\phi_h^{n+1/2}\|^{1/2}\|\nabla\phi_h^{n+1/2}\|^{1/2} \\ &\le \frac{\nu}{32}\|s_h^{n+1/2}\|^2 + \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2 + C\nu^{-3}\|\nabla u^{n+1/2}\|^4\|\phi_h^{n+1/2}\|^2.\end{aligned}$$

Similarly, the second of the trilinear terms is bounded by

$$\begin{aligned}(54)\quad \left|(u^{n+1/2}\times r^{n+1/2}, \phi_h^{n+1/2})\right| &\le \frac{\nu}{32}\|r^{n+1/2}\|^2 + \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2 \\ &\quad + C\nu^{-3}\|\nabla u^{n+1/2}\|^4\|\phi_h^{n+1/2}\|^2.\end{aligned}$$

The third of the trilinear terms is expanded by adding and subtracting $w^{n+1/2}$ to $w_h^{n+1/2}$ to form $w^{n+1/2} - E^{n+1/2}$, followed by decomposing $E^{n+1/2}$, and bounding each of the three resulting trilinear terms to get

$$\begin{aligned}(55)\quad &\left|(\eta^{n+1/2}\times w_h^{n+1/2}, \phi_h^{n+1/2})\right| \le \frac{3\nu}{32}\|\phi_h^{n+1/2}\|^2 + \frac{\nu}{32}\|s_h^{n+1/2}\|^2 \\ &+ C\nu^{-1}\|r^{n+1/2}\|^2\|\nabla\eta^{n+1/2}\|^2 + \frac{1}{2}\|\nabla\eta^{n+1/2}\|^2\|w^{n+1/2}\|^2 \\ &+ C\nu^{-1}\|\phi_h^{n+1/2}\|^2 + C\nu^{-3}\|\nabla\eta^{n+1/2}\|^4\|\phi_h^{n+1/2}\|^2.\end{aligned}$$

Three of the four terms in $IERR(u^n; w^n; \phi_h^{n+1/2})$ are majorized as

$$(56)\quad \left|\left(\frac{u^{n+1}-u^n}{\Delta t} - u_t(t^{n+1/2}), \phi_h^{n+1/2}\right)\right| \le \frac{1}{2}\|\phi_h^{n+1/2}\|^2 + \frac{1}{2}\left\|\frac{u^{n+1}-u^n}{\Delta t} - u_t(t^{n+1/2})\right\|^2,$$

(57)
$$\left|\frac{\nu}{2}(\nabla(u^{n+1/2}-u(t^{n+1/2})),\nabla\phi_h^{n+1/2})\right| \le \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2+C\nu\|\nabla(u^{n+1/2}-u(t^{n+1/2}))\|^2,$$

(58)
$$\left|\frac{\nu}{2}(w^{n+1/2}-w(t^{n+1/2}),\nabla\times\phi_h^{n+1/2})\right| \le \frac{\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2+C\nu\|w^{n+1/2}-w(t^{n+1/2})\|^2,$$

with the remaining term bounded by

(59)
$$\begin{aligned}(u^{n+1/2}&\times w^{n+1/2}-u(t^{n+1/2})\times w(t^{n+1/2}),\phi_h^{n+1/2})\\ &\le \frac{2\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2+C\nu^{-1}\|\nabla u^{n+1/2}\|^2\|w^{n+1/2}-w(t^{n+1/2})\|^2\\ &\quad+C\nu^{-1}\|w(t^{n+1/2})\|^2\|\nabla(u^{n+1/2}-u(t^{n+1/2}))\|^2.\end{aligned}$$

We may now rewrite (47) as

(60)
$$\begin{aligned}&\frac{1}{2\Delta t}(\|\nabla\phi_h^{n+1}\|^2-\|\nabla\phi_h^n\|^2)+\frac{3\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2+\frac{12\nu}{32}\|s_h^{n+1/2}\|^2\\ &\le C\nu^{-1}\|\nabla\eta^{n+1/2}\|^2+C(\nu^{-1}+\nu)\|r^{n+1/2}\|^2+C\nu^{-1}\inf_{q\in Q^h}\|p(t^{n+1/2})-q\|^2\\ &+C\nu^{-3}\|\nabla u^{n+1/2}\|^4\|\phi_h^{n+1/2}\|^2+C\nu^{-1}\|r^{n+1/2}\|^2\|\nabla\eta^{n+1/2}\|^2\\ &+\frac{1}{2}\|\nabla\eta^{n+1/2}\|^2\|w^{n+1/2}\|^2+C\nu^{-3}\|\nabla\eta^{n+1/2}\|^4\|\phi_h^{n+1/2}\|^2\\ &+C\nu^{-1}\|\phi_h^{n+1/2}\|^2+\frac{1}{2}\left\|\frac{u^{n+1}-u^n}{\Delta t}-u_t(t^{n+1/2})\right\|^2+C\nu\|\nabla(u^{n+1/2}-u(t^{n+1/2}))\|^2\\ &+C\nu\|w^{n+1/2}-w(t^{n+1/2})\|^2+C\nu^{-1}\|\nabla u^{n+1/2}\|^2\|w^{n+1/2}-w(t^{n+1/2})\|^2\\ &+C\nu^{-1}\|w(t^{n+1/2})\|^2\|\nabla(u^{n+1/2}-u(t^{n+1/2}))\|^2.\end{aligned}$$

Taylor series can be used to bound the interpolation in time terms, and thus (60) can be reduced to

(61)
$$\begin{aligned}&\frac{1}{2\Delta t}(\|\nabla\phi_h^{n+1}\|^2-\|\nabla\phi_h^n\|^2)+\frac{3\nu}{32}\|\nabla\phi_h^{n+1/2}\|^2+\frac{12\nu}{32}\|s_h^{n+1/2}\|^2\\ &\le C\nu^{-1}\|\nabla\eta^{n+1/2}\|^2+C(\nu^{-1}+\nu)\|r^{n+1/2}\|^2+C\nu^{-1}\inf_{q\in Q^h}\|p^{n+1/2}-q\|^2\\ &+C\nu^{-1}\|r^{n+1/2}\|^2\|\nabla\eta^{n+1/2}\|^2+\frac{1}{2}\|\nabla\eta^{n+1/2}\|^2\|w^{n+1/2}\|^2\\ &+C(\Delta t)^3\int_{t^n}^{t^{n+1}}\|u_{ttt}\|^2\,dt\;+C\nu(\Delta t)^3\int_{t^n}^{t^{n+1}}\|\nabla u_{tt}\|^2\,dt\\ &+C\nu(\Delta t)^3\int_{t^n}^{t^{n+1}}\|w_{tt}\|^2\,dt\;+C\nu^{-1}(\Delta t)^3\|\nabla u^{n+1/2}\|^2\int_{t^n}^{t^{n+1}}\|w_{tt}\|^2\,dt\\ &+C\nu^{-1}(\Delta t)^3\|w(t^{n+1/2})\|^2\int_{t^n}^{t^{n+1}}\|\nabla u_{tt}\|^2\,dt\\ &+C(\nu^{-1}+\nu^{-3}\|\nabla u^{n+1/2}\|^4+\nu^{-3}\|\nabla\eta^{n+1/2}\|^4)\|\phi_h^{n+1/2}\|^2.\end{aligned}$$

Next we sum from $n=0,\dots,N-1$, multiply both sides by $2\Delta t$, recall $\phi_h^0=0$ and the smoothness assumptions, and reduce. With the choice of P_k, P_{k-1} velocity-pressure

spaces, (61) reduces to

$$
\begin{aligned}
(62)\qquad & \|\phi_h^N\|^2 + \sum_{n=0}^{N-1}\left(\frac{3\nu\Delta t}{16}\|\nabla\phi_h^{n+1/2}\|^2 + \frac{3\nu\Delta t}{4}\|s_h^{n+1/2}\|^2\right)\\
&\le C((\Delta t)^4 + \nu^{-1}h^{2k} + (\nu^{-1}+\nu)h^{2k+2} + \nu^{-1}h^{2k} + h^{4k+2})\\
&+ \Delta t\sum_{n=0}^{N-1}\|\nabla\eta^{n+1/2}\|^2\|w^{n+1/2}\|^2 + C\nu^{-1}(\Delta t)^4\sum_{n=0}^{N-1}\|\nabla u^{n+1/2}\|^2\int_{t^n}^{t^{n+1}}\|w_{tt}\|^2\,dt\\
&+ C\nu^{-1}(\Delta t)^4\sum_{n=0}^{N-1}\|w(t^{n+1/2})\|^2\int_{t^n}^{t^{n+1}}\|\nabla u_{tt}\|^2\,dt\\
&+ C\Delta t\sum_{n=0}^{N-1}(\nu^{-1}+\nu^{-3}\|\nabla u^{n+1/2}\|^4 + \nu^{-3}\|\nabla\eta^{n+1/2}\|^4)\|\phi_h^{n+1/2}\|^2.
\end{aligned}
$$

Since $w = \nabla\times u$, we reduce (62) to

$$
\begin{aligned}
(63)\qquad & \|\phi_h^N\|^2 + \sum_{n=0}^{N-1}\left(\frac{3\nu\Delta t}{16}\|\nabla\phi_h^{n+1/2}\|^2 + \frac{3\nu\Delta t}{4}\|s_h^{n+1/2}\|^2\right)\\
&\le C((\Delta t)^4 + \nu^{-1}h^{2k} + (\nu^{-1}+\nu)h^{2k+2} + \nu^{-1}h^{2k} + h^{4k+2})\\
&+ \Delta t\sum_{n=0}^{N-1}\|\nabla\eta^{n+1/2}\|^2\|\nabla u^{n+1/2}\|^2 + C\nu^{-1}(\Delta t)^4\sum_{n=0}^{N-1}\|\nabla u(t^{n+1/2})\|^2\int_{t^n}^{t^{n+1}}\|\nabla u_{tt}\|^2\,dt\\
&+ C\Delta t\sum_{n=0}^{N-1}(\nu^{-1}+\nu^{-3}\|\nabla u^{n+1/2}\|^4 + \nu^{-3}\|\nabla\eta^{n+1/2}\|^4)\|\phi_h^{n+1/2}\|^2.
\end{aligned}
$$

We bound the third and second to last terms with Holder's inequality and the smoothness assumptions, then reduce by assuming $\Delta t,\ \nu\ \le 1$. This yields

$$
\begin{aligned}
(64)\quad & \|\phi_h^N\|^2 + \sum_{n=0}^{N-1}\left(\frac{3\nu\Delta t}{16}\|\nabla\phi_h^{n+1/2}\|^2 + \frac{3\nu\Delta t}{4}\|s_h^{n+1/2}\|^2\right)\\
&\le C((\Delta t)^4+\nu^{-1}h^{2k}) + C\Delta t\sum_{n=0}^{N-1}(\nu^{-1}+\nu^{-3}\|\nabla u^{n+1/2}\|^4+\nu^{-3}\|\nabla\eta^{n+1/2}\|^4)\|\phi_h^{n+1/2}\|^2.
\end{aligned}
$$

Now with Δt chosen sufficiently small, we use the discrete Gronwall inequality to get

(65)
$$
\|\phi_h^N\|^2 + \sum_{n=0}^{N-1}\left(\frac{3\nu\Delta t}{16}\|\nabla\phi_h^{n+1/2}\|^2 + \frac{3\nu\Delta t}{4}\|s_h^{n+1/2}\|^2\right) \le C(u,p,\nu,\Omega)((\Delta t)^4 + h^{2k}).
$$

Using the triangle inequality with (65) completes the proof. ∎

5. Numerical experiments. We now present numerical experiments for the energy- and helicity-conserving scheme. This section makes several comparisons between this scheme and the usual convective form of the trapezoidal (Crank–Nicholson) scheme for the NSE

$$
\begin{aligned}
(66)\quad \frac{1}{\Delta t}(u_h^{n+1}-u_h^n, v) + \frac{1}{2}(u_h^{n+1/2}\cdot\nabla u_h^{n+1/2}, v) - \frac{1}{2}(u_h^{n+1/2}\cdot\nabla v, u_h^{n+1/2})&\\
+ \nu(\nabla u_h^{n+1/2}, \nabla v) = (f^{n+1/2}, v)\ \ \forall v\in V^h,&
\end{aligned}
$$

and the rotational form

$$\frac{1}{\Delta t}(u_h^{n+1} - u_h^n, v) - (u_h^{n+1/2} \times (\nabla \times u_h^{n+1/2}), v) + \nu(\nabla u_h^{n+1/2}, \nabla v) = (f^{n+1/2}, v) \quad \forall v \in V^h. \tag{67}$$

All of the schemes were implemented in MATLAB using Taylor–Hood elements and periodic boundary conditions and uniform meshes on the unit cube. Simple fixed point iterations were used to solve the nonlinear problem in each timestep.

5.1. Computational cost of the schemes. The energy- and helicity-conserving scheme is more computationally expensive than the usual trapezoidal schemes (66) and (67). It solves for velocity and a projected vorticity, both in V^h, and results in linear systems that are double the size of those arising from the usual schemes. Hence, the energy- and helicity-conserving scheme would be more practical if a linearization or decoupling of the system could be found that would still conserve both energy and helicity. At this point, we do not know if such a linearization can be found. It is possible that an (effective and reliable) iteration between decoupled equations could be discovered. Since the energy- and helicity-conserving scheme, when decoupled, will take a form much like that of (67), one may even be able to take advantage of more efficient solvers designed for rotational-form Navier–Stokes schemes such as those described by Benzi and Liu in [2] or Olshanskii in [16].

5.2. Experiment 1: Helicity conservation for $\nu = f = 0$. The first numerical experiment is a comparison of helicity treatment in the three schemes when $\nu = f = 0$. This is the case where helicity is exactly conserved in the true physics, and thus for physical fidelity should also be conserved in the numerical schemes. Using

$$u^0 = \langle \cos(2\pi z), \sin(2\pi z), \sin(2\pi x) \rangle \tag{68}$$

for the initial condition (since it is simple and has nonzero helicity), we set $\nu = f = 0$ in each scheme and computed from $(0, 1]$ on the (periodic) unit cube. The energy- and helicity-conserving scheme was run on an $h = 1/8$ uniform mesh, and the other two schemes were run on $h = 1/8$ and $h = 1/16$ uniform meshes. Timesteps were chosen to be 0.025 and 0.01 for the two meshes, respectively. Figure 1 shows a plot of each solution's helicity on $[0, 1]$, and from here it is clear that the usual trapezoidal schemes do not conserve helicity, and that the energy- and helicity-conserving scheme, as expected, does.

5.3. Experiment 2: Accuracy comparison for a known solution. Given the true solution

$$u = ((2-t)\cos(2\pi z),\ (1+t)\sin(2\pi z),\ (1-t)\sin(2\pi x)),\ p = \sin(2\pi(x+y+t)), \tag{69}$$

we calculated f from u and p and implemented each of the schemes on an $h = 1/8$ mesh with $T = 2$, $\Delta t = 0.025$, and $\nu = 1$. Shown below are the plots of helicity error, L^2 error, and H^1 error vs. time for the three schemes. We see from the plots that the usual trapezoidal schemes (66) and (67) give nearly identical results, and that these schemes have a better H^1 error but worse L^2 error and helicity error than the energy- and helicity-conserving scheme. Similar experiments should be conducted for smaller ν (and thus finer meshes) to verify that the results will hold in a setting with less viscosity.

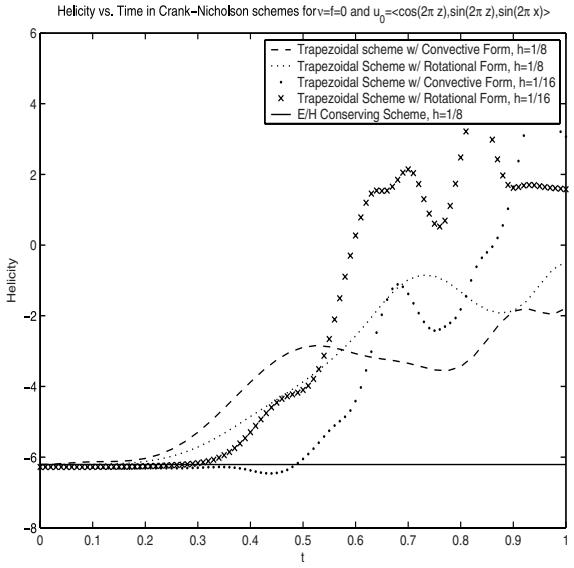


FIG. 1. *Helicity conservation in different trapezoid schemes for the NSE with* $\nu = f = 0$ *and* $u_0 = (\cos(2\pi z), \sin(2\pi z), \sin(2\pi x))$.

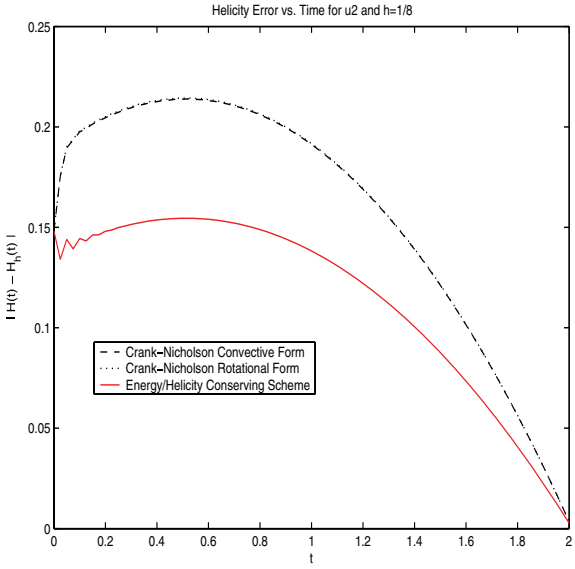


FIG. 2. *Helicity error in the schemes.*

We believe the oscillations near $t = 0$ (seen in Figures 2 and 3) are a result of using Crank–Nicholson schemes, and we offer the following possibility for the more pronounced oscillations in the new scheme: On such a coarse mesh, the different schemes treat the viscous term(s) significantly differently, as the new scheme employs a vorticity projection. We believe the error introduced (by such a coarse mesh projection) may amplify initial oscillations, and thus we expect the difference in magnitude of the schemes' oscillations to decrease with finer meshes.

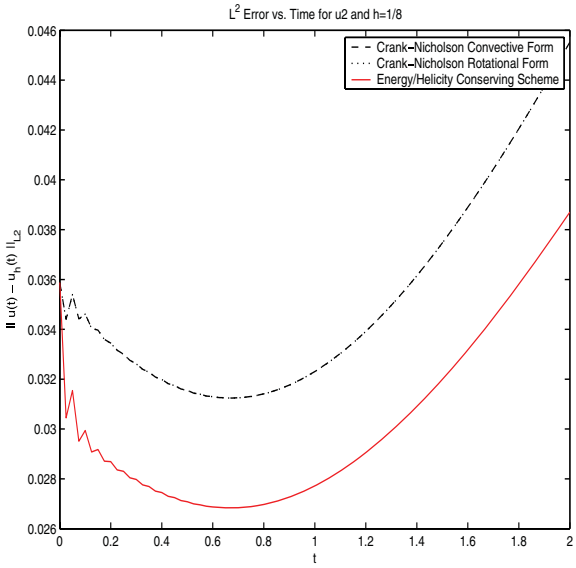


FIG. 3. *L^2 error in the schemes.*

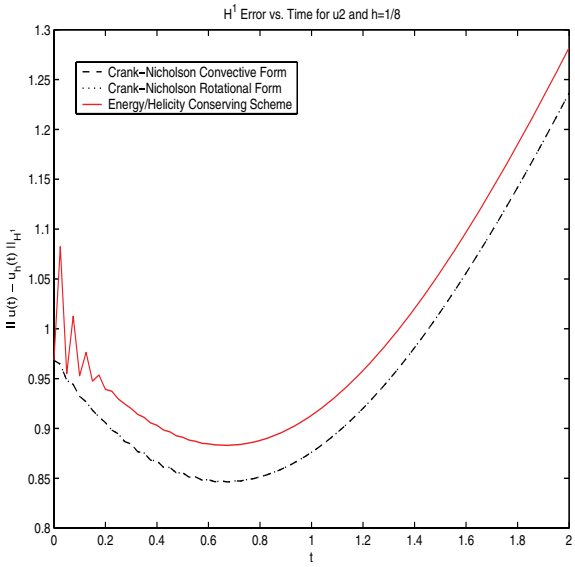


FIG. 4. *H^1 error in the schemes.*

6. Conclusions. In an effort to find more physically relevant solutions to the NSE, we have developed an energy- and helicity-conserving finite element scheme for periodic flows which is second order in time and converges optimally in space. The scheme is able to conserve two inviscid invariants by using the rotational form of the nonlinearity with a projected vorticity. The scheme retains the asymptotic velocity convergence rates of the usual trapezoidal finite element method. Numerical evidence suggests that the scheme can predict helicity more accurately than the usual

trapezoidal scheme. However, each linear system that needs to be solved is double the size of those in usual trapezoidal scheme, and thus further work must be done to make this promising scheme more practical.

REFERENCES

[1] A. Arakawa, *Computational design for long-term numerical integration of the equations of fluid motion: Two dimensional flow, Part* I, J. Comput. Phys., 1 (1966), pp. 119–143.

[2] M. Benzi and J. Liu, *An Efficient Solver for the Navier-Stokes Equations in Rotation Form*, Preprint, Department of Mathematics and Computer Science, Emory University, Atlanta, GA, 2006.

[3] S. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1994.

[4] U. Frisch, *Turbulence*, Cambridge University Press, Cambridge, UK, 1995.

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

[6] P. Gresho and R. Sani, *Incompressible Flow and the Finite Element Method*, Vol. 2, Wiley, New York, 1998.

[7] J. Guermond, *Finite-element-based Faedo-Galerkin weak solutions to the Navier–Stokes equations in the three-dimensional torus are suitable*, J. Math. Pures Appl. (9), 85 (2006), pp. 451–464.

[8] J. G. Heywood and R. Rannacher, *Finite element approximation of the nonstationary Navier–Stokes problem. Part* III: *Smoothing property and higher order error estimates for spatial discretization*, SIAM J. Numer. Anal., 25 (1988), pp. 489–512.

[9] J. G. Heywood and R. Rannacher, *Finite-element approximation of the nonstationary Navier–Stokes problem. Part* IV: *Error analysis for the second-order time discretization*, SIAM J. Numer. Anal., 27 (1990), pp. 353–384.

[10] S. Kaya and B. Riviére, *A discontinuous subgrid eddy viscosity method for the time-dependent Navier–Stokes equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1572–1595.

[11] R. Lewandowski, *Vorticities in a LES model for* 3*D periodic turbulent flows*, J. Math. Fluid Mech., 8 (2006), pp. 398–422.

[12] R. Lewandowski and W. Layton, *On a well-posed turbulence model*, Discrete Contin. Dyn. Syst. Ser. B, 6 (2006), pp. 111–128.

[13] J. Lui and W. Wang, *Energy and helicity preserving schemes for hydro- and magnetohydrodynamics flows with symmetry*, J. Comput. Phys., 200 (2004), pp. 8–33.

[14] H. Moffatt and A. Tsoniber, *Helicity in laminar and turbulent flow*, Ann. Rev. Fluid Mech., 24 (1992), pp. 281–312.

[15] J. Moreau, *Constantes d'un îlot tourbilloinnaire en fluide parfait barotrope*, C.R. Acad. Aci. Paris, 252 (1961), pp. 2810–2812.

[16] M. Olshanskii, *Iterative solver for Oseen problem and numerical solution of incompressible Navier-Stokes equations*, Numer. Linear Algebra Appl., 6 (1999), pp. 353–378.

[17] M. A. Olshanskii and A. Reusken, *Navier–Stokes equations in rotation form: A robust multigrid solver for the velocity problem*, SIAM J. Sci. Comput., 23 (2002), pp. 1683–1706.

[18] R. Temam, *Navier-Stokes Equations: Theory and Numerical Analysis*, North-Holland, New York, 1977.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1639–1653

UNIFIED ANALYSIS OF FINITE VOLUME METHODS FOR SECOND ORDER ELLIPTIC PROBLEMS*

SO-HSIANG CHOU† AND XIU YE‡

Abstract. We establish a general framework for analyzing the class of finite volume methods which employ continuous or totally discontinuous trial functions and piecewise constant test functions. Under the framework, optimal order convergence in the H^1 and L^2 norms can be obtained in a natural and systematic way for classical finite volume methods and new finite volume methods such as discontinuous finite volume methods applied to second order elliptic problems.

Key words. finite element methods, finite volume methods, discontinuous Galerkin methods, finite volume element

AMS subject classifications. Primary, 65N15, 65N30, 76D07; Secondary, 35B45, 35J50

DOI. 10.1137/050643994

1. Introduction.

Due to the local conservation property and other attractive properties such as robustness with unstructured meshes, the finite volume method is widely used in computational fluid dynamics. Numerical analysis of a finite volume method is more difficult than that of a finite element method, since in general a finite volume method uses two different function spaces: one for the trial space and one for the test space. For example, obtaining the optimal L^2 error estimates is a common practice for finite element methods. They are very difficult to obtain for the finite volume methods. Because of this reason, the optimal L^2 estimates have not been derived for the finite volume methods proposed in [8, 9, 10, 13, 25]. The main motivation of this paper is to propose a general framework under which we can systematically give a thorough analysis for finite volume methods to second order elliptic problems and obtain the optimal error estimates in energy norm and L^2 norm.

In recent years, there have appeared different approaches in the convergence and stability analysis of the finite volume method; see, for example, [2, 5, 6, 12, 13, 16, 15, 17, 18, 22], among others. Motivated by the popularity of discontinuous Galerkin methods, Ye [25] proposed a finite volume method with a totally discontinuous trial function space for elliptic problems. Our general framework covers the finite volume methods (continuous or discontinuous) developed in all of the papers mentioned above in a unified way, and previously hard-to-obtain optimal L^2 estimates [8, 10, 9, 13, 25] can now be derived naturally.

For simplicity in this paper we will treat only finite volume methods applied to the self-adjoint elliptic equations. To illustrate the idea, we consider the model problem

$$\mathcal{L}u := -\nabla \cdot \mathcal{A}\nabla u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \tag{1.1}$$

where $\Omega \subset R^2$ is a bounded polygonal domain and $\mathcal{A}$ is in either $W^{1,\infty}$ or $W^{2,\infty}$. A typical finite volume method uses piecewise constant functions as test functions, and,

*Received by the editors November 1, 2005; accepted for publication (in revised form) May 8, 2007; published electronically August 17, 2007.
http://www.siam.org/journals/sinum/45-4/64399.html

†Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH 43403-0221 (chou@bgnet.bgsu.edu). The work of this author was supported in part by the National Center for Theoretical Sciences, Taiwan.

‡Department of Mathematics, University of Arkansas at Little Rock, Little Rock, AK 72204 (xxye@ualr.edu).

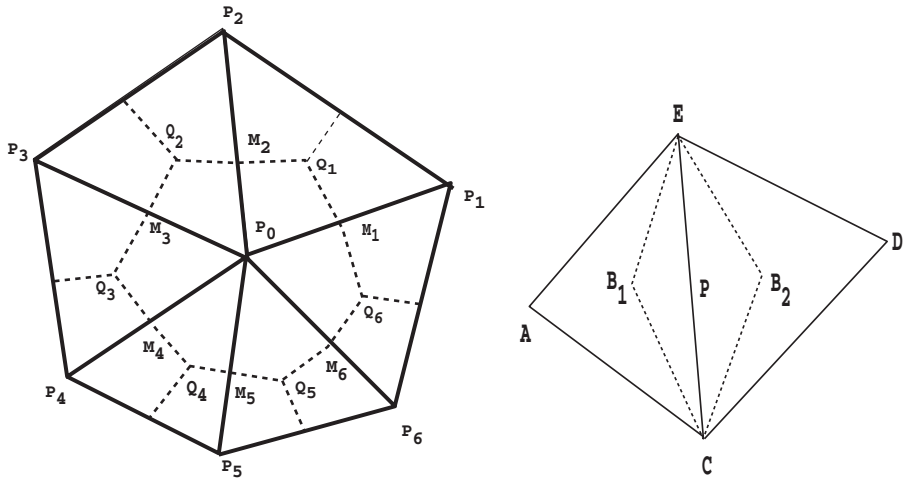


FIG. 1. *Primal and dual grids. Left figure: Conforming finite volume method. Right figure: Nonconforming finite volume method.*

to keep the same dimension for the spaces of the trial functions and test functions, two different partitions of the domain Ω are needed: one called the primal partition is associated with the trial space, and one called the dual partition is associated with the test space. For example, in Figure 1, on the left the primal partition is made up of the standard triangular finite elements, and the dual partition is the usual barycentric subdivision consisting of polygons around P_i's obtained by connecting midpoints M_i's of edges and barycenters Q_i's of the triangles. Thus $M_1Q_1M_2Q_2M_3Q_3M_4Q_4M_5Q_5M_6Q_6$ is a typical dual volume around P_0. On the other hand, in the right figure of Figure 1 we use triangles in the primal partition, and for each midpoint of an edge in the triangles we define a quadrilateral element that serves as an element in the dual partition. So, for example, in Figure 1 the quadrilateral EB_1CB_2 around midpoint P (B_i barycenters of triangles) is in the dual partition.

Figure 2 shows two more possible configurations of primal (solid lines) and dual (dashed lines) partitions. In particular, the partitions in the right figure will be used for the discontinuous finite volume method in section 3.3. Here we use standard triangular elements in the primal partition, and each triangular element then generates three dual triangular volumes (AB_1D and two others) by connecting its barycenter and vertices.

Denote by $\mathcal{T}_h$ the primal triangulation of Ω, by $\mathcal{T}_h^*$ the dual partition of $\mathcal{T}_h$, and by $P_l(T)$ the space of all polynomials on T whose degree is at most l. The finite dimensional trial space V_h associated with $\mathcal{T}_h$ is a subspace of piecewise linears, i.e.,

$$V_h \subset \{v \in V : \ v|_T \in P_1(T) \ \forall T \in \mathcal{T}_h\}, \tag{1.2}$$

where V is either $H_0^1(\Omega)$ or $L^2(\Omega)$ (standard Sobolev spaces notation will be adopted throughout the paper). Examples of such space are continuous P_1 conforming space, the Crouzeix–Raviart P_1 nonconforming space [14] (continuous at midpoints), and totally discontinuous P_1 space to be used in conjunction with the discontinuous finite volume method in section 3.3. The test function space Q_h associated with the dual partition $\mathcal{T}_h^*$ is

$$Q_h = \{q \in L^2(\Omega) : \ q|_K \in P_0(K) \ \ \forall K \in \mathcal{T}_h^*\}. \tag{1.3}$$

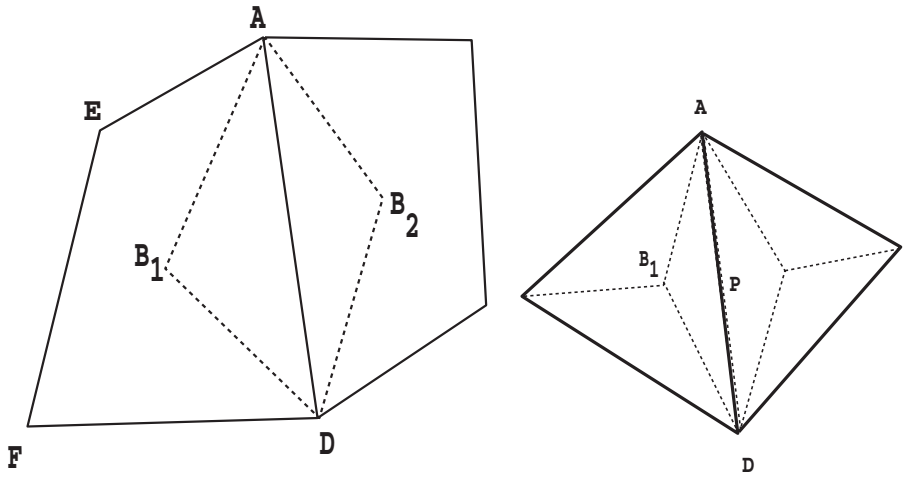


FIG. 2. *Primal and dual grids. Left figure: Nonconforming finite volume method. Right figure: Discontinuous finite volume method.*

We mention in passing that classical finite volume methods adopt piecewise P_0 shape functions, and their applications abound. The present (and newer) finite volume methods using piecewise P_1 shape functions also find many practical applications in heat transfer and fluid flow problems [7, 21] and the references therein. These methods are also natural when combined with the multilevel adaptive methods [19, 20].

Due to the efforts of several authors [6, 12, 15, 17], especially [6, 15, 17], it is now recognized that, for finite volume methods applied to second order elliptic problems on polygonal domains, it is to be expected that, for the exact solution u and approximate solution u_h, the best form of the L^2 estimates is

$$||u - u_h|| \leq Ch^2(||u||_2 + ||f||_1).$$

(We use $||\cdot||_p$ for the standard Sobolev H^p norm and drop the subindex for the L^2 norm.) One notes that this is not the same as assuming u in $H^3(\Omega)$. For example, the solution of the boundary value problem $\Delta u = 1$ on the unit square and $u = 0$ on the boundary belongs to $H^2(\Omega)$ but not to $H^3(\Omega)$. While it is easy and natural to deduce the above error estimates under our present framework, it should be pointed out that there are other ways to view finite volume methods, depending on how one views what the distinctive traits of a finite volume method are. For example, one may consider the so-called mixed finite volume method in which the flux can be recovered by a simple formula [11]. On the other hand, in other finite volume methods the flux itself plays an important role in the derivation of the method. For instance, in [16], finite volume methods are based on considering averages of solutions on the control volumes which coincide with the supports of the test functions in the present paper. The stiffness matrix is calculated from a difference approximation of the fluxes between two neighboring elements. Compactness methods are used to prove the convergence. While this approach can be generalized consistently to convection-diffusion and hyperbolic problems, it shows considerable difficulties when error estimates are to be obtained. Our approach focuses on a narrower elliptic problem class and explores its natural relation to the Galerkin finite element method. Consequently, optimal order error estimates are easier to obtain.

The organization of the paper is as follows. In section 2 we present our general finite volume framework and its stability and convergent analysis. Under this framework, in section 3 we systematically derive for the new as well as the old finite volume methods the optimal H^1 estimates of the usual form and optimal L^2 estimates of the above form.

Let e be an interior edge common to elements T_1 and T_2 in $\mathcal{T}_h$, and let $\mathbf{n}_1$ and $\mathbf{n}_2$ be the unit normal vectors on e exterior to K_1 and K_2, respectively. For a scalar q and a vector $\mathbf{w}$ we define their average $\{\cdot\}$ on e and jump $[\![\cdot]\!]$ across e, respectively, as

$$\{q\} = \frac{1}{2}(q|_{\partial T_1} + q|_{\partial T_2}), \quad [\![q]\!] = q|_{\partial T_1}\mathbf{n}_1 + q|_{\partial T_2}\mathbf{n}_2,$$

$$\{\mathbf{w}\} = \frac{1}{2}(\mathbf{w}|_{\partial T_1} + \mathbf{w}|_{\partial T_2}), \quad [\![\mathbf{w}]\!] = \mathbf{w}|_{\partial T_1} \cdot \mathbf{n}_1 + \mathbf{w}|_{\partial T_2} \cdot \mathbf{n}_2.$$

Note that the jump of a vector is a scalar, whereas the jump of a scalar is a vector. If e is an edge on the boundary of Ω, we define

$$\{q\} = q, \quad [\![\mathbf{w}]\!] = \mathbf{w} \cdot \mathbf{n}.$$

The quantities $[\![q]\!]$ and $\{\mathbf{w}\}$ on boundary edges are defined analogously. Let $\mathcal{E}_h$ denote the union of the boundaries of the triangles T of $\mathcal{T}_h$ and $\mathcal{E}_h^0 := \mathcal{E}_h \backslash \partial\Omega$ the collection of all interior edges.

Following [8, 12], we assume the existence of a transfer operator γ from $V(h) := V_h + H^2(\Omega) \cap H_0^1(\Omega)$ to the test space Q_h. In particular, γ connects the trial space V_h with the test space Q_h. Throughout the paper, the operator γ is required to satisfy the following sets of assumptions.

Assumption 1. Quadraturelike and restriction assumptions for γ:

$$\int_T (v - \gamma v)dx = 0 \quad \forall v \in V_h, \ \ \forall T \in \mathcal{T}_h, \tag{1.4}$$

$$\int_e (v - \gamma v)ds = 0 \quad \forall v \in H^2(\mathcal{T}_h), \ \ \forall e \in \partial T, \ \forall T \in \mathcal{T}_h, \tag{1.5}$$

$$\text{if} \quad [\![v]\!] = 0, \quad \text{then} \ \ [\![\gamma v]\!] = 0, \tag{1.6}$$

where $H^2(\mathcal{T}_h) := \{v \in L^2(\Omega) : v|_T \in H^2(T) \qquad \forall T \in \mathcal{T}_h\}$.

Equations (1.4)–(1.5) have been observed in [12, 13] and perhaps can be viewed as a type of quadrature condition. Equation (1.6) is our new observation in this paper regarding to the jump.

Assumption 2. Approximation property of γ:

$$||\gamma w - w||_{0,T} \le Ch_T|w|_{1,T} \qquad \forall T \in \mathcal{T}_h. \tag{1.7}$$

Then the solution of (1.1) necessarily satisfies

$$\mathcal{L}u = -\nabla \cdot \mathcal{A}\nabla u = f \quad \text{ on } K \quad \forall K \in \mathcal{T}_h^*, \tag{1.8}$$

$$[\![\gamma u]\!]_e = 0 \quad \forall e \in \mathcal{E}_h, \tag{1.9}$$

$$[\![\mathcal{A}\nabla u]\!]_e = 0 \quad \forall e \in \mathcal{E}_h^0. \tag{1.10}$$

2. Finite volume formulation. In this section, we will derive a general formulation for finite volume methods. The formulation is based on enforcing (1.8)–(1.10) by testing with "element" test functions for (1.8) and "edge" test functions for (1.9)

and (1.10). To this end, we further assume the existence of two linear operators $B_1 : V(h) \to L^2(\mathcal{E}_h)$ and $B_2 : V(h) \to L^2(\mathcal{E}_h^0)$ (they will be defined shortly). Testing (1.8), (1.9), and (1.10) by γv, $B_1 v$, and $B_2 v$, respectively, and adding them up, we obtain the "global" equation

$$(\mathcal{L}u, \gamma v)_{\mathcal{T}_h^*} + ([\![\gamma u]\!], B_1 v)_{\mathcal{E}_h} + ([\![\mathcal{A}\nabla u]\!], B_2 v)_{\mathcal{E}_h^0} = (f, \gamma v), \tag{2.1}$$

where each inner product obviously means the sum of its local inner products. *A remark is in order here.* Interpreting PDEs and jump conditions such as (1.8)–(1.10) as residual equations and testing them with test functions of different levels is, of course, quite common in finite element and finite volume methods. However, the fact that summing them up as equal weight relations can lead to fruitful analysis is more recent. In fact, using this technique Brezzi et al. [4] have demonstrated stabilization mechanisms in discontinuous Galerkin methods in a unified way.

Integrating (2.1) by parts and using the fact that γv is constant on K, we have

$$\begin{aligned}
(\mathcal{L}u, \gamma v)_{\mathcal{T}_h^*} &= -\sum_{K\in\mathcal{T}_h^*} \int_K \nabla\cdot\mathcal{A}\nabla u \gamma v dx \\
&= -\sum_{K\in\mathcal{T}_h^*} \int_{\partial K} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds \\
&= \left(-\sum_{K\in\mathcal{T}_h^*} \int_{\partial K} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds + \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds \right) \\
&\quad - \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds,
\end{aligned}$$

where we have added and subtracted the last term to bring in the effect of primal triangulation.

Define the bilinear form $a : V(h)\times V(h) \to R$

$$a(u,v) := -\sum_{K\in\mathcal{T}_h^*} \int_{\partial K} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds + \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds.$$

Recall the following easily derived identity (or see [1]): For all $q \in \prod_{T\in\mathcal{T}_h} L^2(\partial T)$ and for all $\mathbf{v} \in [\prod_{T\in\mathcal{T}_h} L^2(\partial T)]^2$,

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} q\mathbf{v}\cdot\mathbf{n} ds = \int_{\mathcal{E}_h} [\![q]\!]\{\mathbf{v}\} ds + \int_{\mathcal{E}_h^0} \{q\}[\![\mathbf{v}]\!] ds. \tag{2.2}$$

In particular,

$$\sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds = \sum_{e\in\mathcal{E}_h} \int_e [\![\gamma v]\!]\cdot\{\mathcal{A}\nabla u\} ds + \sum_{e\in\mathcal{E}_h^0} \int_e \{\gamma v\}[\![\mathcal{A}\nabla u]\!] ds, \tag{2.3}$$

and hence (2.1) becomes

$$\begin{aligned}
&a(u,v) - ([\![\mathcal{A}\nabla u]\!], \{\gamma v\})_{\mathcal{E}_h^0} - (\{\mathcal{A}\nabla u\}, [\![\gamma v]\!])_{\mathcal{E}_h} \\
&\qquad + ([\![\gamma u]\!], B_1 v)_{\mathcal{E}_h} + ([\![\mathcal{A}\nabla u]\!], B_2 v)_{\mathcal{E}_h^0} = (f, \gamma v).
\end{aligned}$$

The choice of $B_2 v = \{\gamma v\}$ leads to

$$a(u,v) - (\{\mathcal{A}\nabla u\}, [\![\gamma v]\!])_{\mathcal{E}_h} + ([\![\gamma u]\!], B_1 v)_{\mathcal{E}_h} = (f, \gamma v).$$

Furthermore, if we take the common pick of $B_1 v = \alpha h^{-1}[\![\gamma v]\!] + \delta\{\mathcal{A}\nabla v\}$, where α is a positive number and $\delta = 1, -1$, the above equation becomes

$$a(u,v) - (\{\mathcal{A}\nabla u\}, [\![\gamma v]\!])_{\mathcal{E}_h} + \delta([\![\gamma u]\!], \{\mathcal{A}\nabla v\})_{\mathcal{E}_h} + \alpha h^{-1}([\![\gamma u]\!], [\![\gamma v]\!])_{\mathcal{E}_h} = (f, \gamma v).$$

For simplicity, we will fix our choices and take $B_1 v = \alpha h^{-1}[\![\gamma v]\!] + \delta\{\mathcal{A}\nabla v\}$ and $B_2 v = \{\gamma v\}$ in the remaining part of the paper. However, our analysis carries through for other choices in [4] as well.

Let

$$A(u,v) := a(u,v) - (\{\mathcal{A}\nabla u\}, [\![\gamma v]\!])_{\mathcal{E}_h} + \delta([\![\gamma u]\!], \{\mathcal{A}\nabla v\})_{\mathcal{E}_h} + \alpha h^{-1}([\![\gamma u]\!], [\![\gamma v]\!])_{\mathcal{E}_h}, \tag{2.4}$$

and consider the following class of finite volume methods: Find $u_h \in V_h$

$$A(u_h, v) = (f, \gamma v) \qquad \forall v \in V_h. \tag{2.5}$$

The formulation (2.5) is consistent; i.e., the true solution u satisfies

$$A(u,v) = (f, \gamma v) \quad \forall v \in V_h. \tag{2.6}$$

Subtracting (2.5) from (2.6) gives

$$A(u - u_h, v) = 0 \quad \forall v \in V_h. \tag{2.7}$$

We define a norm $|\!|\!| \cdot |\!|\!|$ on $V(h)$ as

$$|\!|\!| v |\!|\!|^2 = |u|_{1,h}^2 + \sum_{e \in \mathcal{E}_h} [\![\gamma v]\!]_e^2 + \sum_{T \in \mathcal{T}_h} h_T^2 |v|_{2,T}^2.$$

We assume the bilinear for $A(\cdot,\cdot)$ is bounded and coercive:

Assumption 3.

$$|A(v,w)| \le C_1 |\!|\!| v |\!|\!| |\!|\!| w |\!|\!| \qquad \forall v, w \in V(h) \times V(h), \tag{2.8}$$

$$A(v,v) \ge C_2 |\!|\!| v |\!|\!|^2 \qquad \forall v \in V_h. \tag{2.9}$$

Then we have the following theorem that is the counterpart of Céa's lemma [3] in the finite element theory.

THEOREM 2.1. *Let u and u_h be the solutions of* (1.1) *and* (2.5). *Then*

$$|\!|\!| u - u_h |\!|\!| \le C \inf_{v \in V_h} |\!|\!| u - v |\!|\!|.$$

Proof. From (2.9) and (2.7), we have that for any $v \in V_h$

$$C_1 |\!|\!| u_h - v |\!|\!|^2 \le A(u_h - v, u_h - v) = A(u - v, u_h - v) \le C_2 |\!|\!| u - v |\!|\!| |\!|\!| u_h - v |\!|\!|.$$

Hence by the triangle inequality we have

$$|\!|\!| u - u_h |\!|\!| \le C \inf_{v \in V_h} |\!|\!| u - v |\!|\!|.$$

This completes the proof. ∎

To obtain the L^2 error estimate for our general finite volume formulation (2.5), we assume that the bilinear form $a(v,w)$ satisfies the following equations.

Assumption 4. For any $v, w \in V(h)$,

$$a(v,w) = (\mathcal{A}\nabla_h v, \nabla_h w) + \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{A}\nabla v \cdot \mathbf{n}(\gamma w - w)ds + \sum_{T\in\mathcal{T}_h} (\nabla\cdot\mathcal{A}\nabla v, w-\gamma w)_T. \tag{2.10}$$

For this reason, we shall take $\delta = -1$ in the following analysis.

THEOREM 2.2. *Let $u \in H^2(\Omega)\cap H_0^1(\Omega)$ and $u_h \in V_h$ be the solutions of* (1.1) *and* (2.5) *with $\delta = -1$, respectively. Assume that $\mathcal{A} \in W^{2,\infty}(\Omega)$ and that* (1.4), (1.5), (1.7), *and* (2.10) *hold. Then*

$$\|u - u_h\| \le Ch(|\!|\!| u - u_h |\!|\!| + h\|f\|_1).$$

Proof. Let $w \in H_0^1(\Omega)\cap H^2(\Omega)$ be the solution of the dual problem

$$-\nabla\cdot\mathcal{A}\nabla w = u - u_h \quad \text{in } \Omega, \tag{2.11}$$

$$w = 0 \quad \text{on } \partial\Omega, \tag{2.12}$$

so that the following estimate holds:

$$\|w\|_2 \le C\|u - u_h\|. \tag{2.13}$$

Let $w_I \in V_h$ be the usual *continuous* piecewise linear Lagrange interpolant of w, so that

$$|\!|\!| w - w_I |\!|\!| \le Ch|w|_2. \tag{2.14}$$

From (2.11) we deduce that

$$\begin{aligned}\|u-u_h\|^2 &= -(u-u_h, \nabla\cdot\mathcal{A}\nabla w)\\ &= (\mathcal{A}\nabla_h(u-u_h), \nabla_h w) - \sum_{T\in\mathcal{T}_h}\int_{\partial T}\mathcal{A}\nabla w\cdot\mathbf{n}(u-u_h)ds\\ &= (\mathcal{A}\nabla_h(u-u_h), \nabla_h w) - \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w\}, [\![u-u_h]\!])_e, \end{aligned} \tag{2.15}$$

where we have used (2.3) and the fact that $[\![\mathcal{A}\nabla w]\!]_e = 0$ on all interior edges e.

On the one hand, (2.10) implies

$$a(u-u_h, w_I) = (\mathcal{A}\nabla_h(u-u_h), \nabla_h w_I) + \sum_{T\in\mathcal{T}_h}(\mathcal{A}\nabla(u-u_h)\cdot\mathbf{n}, \gamma w_I - w_I)_{\partial T} + \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla(u-u_h), w_I - \gamma w_I)_T, \tag{2.16}$$

and, on the other hand, it follows from (2.7) that

$$a(u-u_h, w_I) = \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w_I\}, [\![\gamma(u-u_h)]\!])_e. \tag{2.17}$$

Thus, subtracting (2.16) from the sum of (2.15) and (2.17), we have

$$\begin{aligned}\|u-u_h\|^2 &= (\mathcal{A}\nabla_h(u-u_h),\nabla_h(w-w_I)) - \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla(u-u_h),w_I-\gamma w_I)_T \\ &\quad + \left(\sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w_I\},[\![\gamma(u-u_h)]\!])_e - \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w\},[\![u-u_h]\!])_e\right) \\ &\quad - \sum_{T\in\mathcal{T}_h}(\mathcal{A}\nabla(u-u_h)\cdot\mathbf{n},\gamma w_I-w_I)_{\partial T} \\ &:= I_1+I_2+I_3+I_4. \end{aligned} \tag{2.18}$$

The four I terms can be estimated as follows. Using (2.14) and (2.13), we have

$$\begin{aligned} I_1 &= (\mathcal{A}\nabla_h(u-u_h),\nabla(w-w_I)) \le C|||u-u_h|||w-w_I|_1 \\ &\le Ch\|u-u_h\|\,|||u-u_h|||. \end{aligned}$$

As for the I_2 term, first it follows from (1.1), (1.4), (1.7), and (2.13) that

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla u,w_I-\gamma w_I)_T &= \sum_{T\in\mathcal{T}_h}(\bar f-f,w_I-\gamma w_I)_T \\ &\le Ch^2\|f\|_1\|u-u_h\|, \end{aligned}$$

where $\bar f$ is the average of f over each element. Next,

$$\begin{aligned}\sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla u_h,w_I-\gamma w_I)_T &= \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla u_h-\overline{\nabla\cdot\mathcal{A}\nabla u_h},w_I-\gamma w_I)_T \\ &\le Ch||\mathcal{A}||_{2,\infty}|u_h|_{1,h}\|u-u_h\| \\ &\le Ch||\mathcal{A}||_{2,\infty}(|||u-u_h|||+||f||)\,\|u-u_h\|, \end{aligned} \tag{2.19}$$

where $\overline{\nabla\cdot\mathcal{A}\nabla u_h}$ is the average of $\nabla\cdot\mathcal{A}\nabla u_h$ over each element T.

For the I_3 term, using (1.5) and (2.13), we have

$$\begin{aligned}&\sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w_I\},[\![\gamma(u-u_h)]\!])_e - \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w\},[\![(u-u_h)]\!])_e \\ &\quad= \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w_I\},[\![\gamma(u-u_h)]\!])_e - \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w\},[\![\gamma(u-u_h)]\!])_e \\ &\qquad+ \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w\},[\![\gamma(u-u_h)]\!])_e - \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w\},[\![(u-u_h)]\!])_e \\ &\quad= \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla(w_I-w)\},[\![\gamma(u-u_h)]\!])_e \\ &\qquad- \sum_{e\in\mathcal{E}_h}(\{\mathcal{A}\nabla w-\overline{\mathcal{A}\nabla w}\},[\![(u-u_h)-\gamma(u-u_h)]\!])_e \\ &\quad:= J_1+J_2 \\ &\quad\le Ch|||u-u_h|||\|u-u_h\|, \end{aligned}$$

where $\overline{\mathcal{A}\nabla w}$ is the average of $\mathcal{A}\nabla w$ over each edge and the J terms are estimated as follows. In fact, the J_i terms can be estimated using the following easily derived trace inequality [12]: For $\phi\in H^1(T)$ and for an edge e of T with h_e the length of e,

$$\|\phi\|_e^2 \le C(h_e^{-1}|\phi|_{0,T}^2 + h_e|\phi|_{1,T}^2), \tag{2.20}$$

where C depends on the shape parameter of T such as the minimal angle of T in the triangular case. For instance,

$$\begin{aligned}
J_1 &= \sum_{e\in\mathcal{E}_h} (\{\mathcal{A}\nabla(w_I - w)\}, [\![\gamma(u-u_h)]\!])_e \\
&\le \sum_{e\in\mathcal{E}_h} |\{\mathcal{A}\nabla(w_I - w)\}|_{0,e} |[\![\gamma(u-u_h)]\!]|_{0,e} \\
&= \sum_{e\in\mathcal{E}_h} |\{\mathcal{A}\nabla(w_I - w)\}|_{0,e}\, h_e^{1/2} [\![\gamma(u-u_h)]\!]_e \\
&\le \sum_{e\in\mathcal{E}_h} h_e^{1/2} \left(h_e^{-1/2} |\{\mathcal{A}\nabla(w_I - w)\}|_T + h_e^{1/2} |\{\mathcal{A}\nabla(w_I - w)\}|_{1,T} \right) [\![\gamma(u-u_h)]\!]_e \\
&\le Ch||\mathcal{A}||_{0,\infty} \|u-u_h\| \, |||u-u_h|||,
\end{aligned}$$

where we have used (2.20) in the last inequality. The term J_2 can be handled similarly.

For the I_4 term first observe that, for any matrix-valued function $\mathcal{M}$ such that $\mathcal{M}$ is constant on each $e \in \mathcal{E}_h$,

$$\begin{aligned}
\sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{M}\nabla(u-u_h)\cdot\mathbf{n}(\gamma w_I - w_I)ds &= \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{M}\nabla u\cdot\mathbf{n}(\gamma w_I - w_I)ds \\
&\quad - \sum_{T\in\mathcal{T}_h} \int_{\partial T} \mathcal{M}\nabla u_h\cdot\mathbf{n}(\gamma w_I - w_I)ds \\
&= I_1 + I_2 = 0,
\end{aligned}$$

where $I_1 = 0$ due to $[\![\mathcal{M}\nabla u]\!] = 0$, and $[\![w_I - \gamma w_I]\!] = 0$ and $I_2 = 0$ due to the fact that $\mathcal{M}\nabla u_h \cdot \mathbf{n}$ is a constant on e and (1.5). Now define $\mathcal{M}$ so that on each $e \in \mathcal{E}_h$, $\mathcal{M} = \mathcal{A}(m)$, the value of $\mathcal{A}$ at the midpoint:

$$\begin{aligned}
|I_4| &= \left| \sum_{T\in\mathcal{T}_h} ((\mathcal{A}-\mathcal{M})\nabla(u-u_h)\cdot\mathbf{n}, \gamma w_I - w_I)_{\partial T} \right| \\
&\le Ch||A||_{1,\infty} \sum_{T\in\mathcal{T}_h} (|\nabla(u-u_h)\cdot\mathbf{n}|, |\gamma w_I - w_I|)_{\partial T} \\
&\le Ch||\mathcal{A}||_{1,\infty} |||u-u_h||| \, ||u-u_h||_0, \qquad (2.21)
\end{aligned}$$

where the last inequality was obtained via the trace inequality (2.20) as before.

Combining the above four estimates with (2.18), we obtain

$$\|u-u_h\| \le Ch(|||u-u_h||| + h\|f\|_1).$$

This completes the proof. □

The counterexamples in [15, 17] show that the assumption of $f \in H^1(\Omega)$ is necessary for finite volume methods.

3. Applications to finite volume and discontinuous finite volume methods. In this section, we will illustrate how our general theory can be applied to analyze different finite volume schemes.

3.1. Finite volume method with conforming trial functions. The finite volume discussed in this subsection is the classical finite volume method. For a given regular subdivision $\mathcal{T}_h$ of triangles, its dual partition $\mathcal{T}_h^*$ is the union of the convex hulls. These convex hulls in $\mathcal{T}_h^*$ are obtained by connecting the barycenters of the triangles and the midpoints of the edges of the triangles in $\mathcal{T}_h$ as shown in Figure 1.

The trial function space associated with $\mathcal{T}_h$ for the traditional finite volume method is defined as

$$V_h = \{v \in H_0^1(\Omega) : \ v|_T \in P_1(T) \ \forall T \in \mathcal{T}^h\},$$

with $V = H_0^1(\Omega)$ in (1.2). The test function space is defined as in (1.3).

Let $\mathcal{N}$ be a set containing all of the interior nodal points associated with the partition $\mathcal{T}^h$. The operator $\gamma : V(h) \to Q_h$ is defined by

$$\gamma v(x) \equiv \sum_{P \in \mathcal{N}} v(P)\chi_P(x) \quad \forall x \in \Omega, \tag{3.1}$$

where χ_P is the characteristic function of the dual element K_P^* associated with the node P. It can be easily verified that γ defined in (3.1) satisfies (1.4)–(1.7).

The traditional conforming finite volume method is to find $u_h \in V_h$ such that for any $v \in V_h$

$$a(u_h, v) = (f, \gamma v). \tag{3.2}$$

The bilinear form $A(v, w)$ in (2.5) reduces to $a(v, w)$ and

$$a(u, v) = -\sum_{K \in \mathcal{T}_h^*} \int_{\partial K} \mathcal{A}\nabla u \cdot \mathbf{n}\gamma v ds.$$

LEMMA 3.1. *For any $v, w \in V(h)$,*

$$\begin{aligned} a(v, w) = (\mathcal{A}\nabla v, \nabla w) &+ \sum_{T \in \mathcal{T}_h} \int_{\partial T} \mathcal{A}\nabla v \cdot \mathbf{n}(\gamma w - w) ds \\ &+ \sum_{T \in \mathcal{T}_h} (\nabla \cdot \mathcal{A}\nabla v, w - \gamma w)_T. \end{aligned} \tag{3.3}$$

Proof. Equation (3.3) appeared in [12, 15, 24], and for completeness we include a short proof here. For ease of proof, a typical primal triangle in Figure 1 is isolated and indexed as in Figure 3. For $j = 1, 2, 3$, let $\square_j$ denote the quadrilaterals formed by the four corner nodes Q, M_j, P_{j+1}, M_{j+1} as shown in Figure 3; when out of bound we use $M_4 = M_1$ and $P_4 = P_1$. Using the divergence theorem on each quadrilateral,

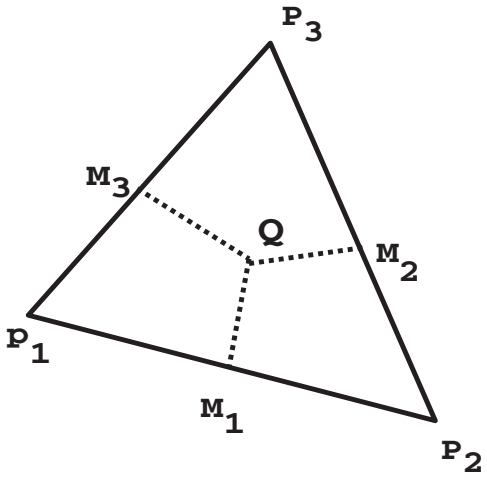


FIG. 3. *Partial primal and dual grids for integration.*

we have

$$
\begin{aligned}
a(v,w) &= -\sum_{T\in\mathcal{T}_h}\sum_{j=1}^{3}\int_{M_{j+1}QM_j}\mathcal{A}\nabla v\cdot\mathbf{n}\gamma w ds\\
&= \sum_{T\in\mathcal{T}_h}\sum_{j=1}^{3}\int_{M_jP_{j+1}M_{j+1}}\mathcal{A}\nabla v\cdot\mathbf{n}\gamma w ds - \sum_{T\in\mathcal{T}_h}\sum_{\square_j}(\nabla\cdot\mathcal{A}\nabla v,\gamma w)\\
&= \sum_{T\in\mathcal{T}_h}\sum_{j=1}^{3}\int_{M_jP_{j+1}M_{j+1}}\mathcal{A}\nabla v\cdot\mathbf{n}(\gamma w - w) ds + \sum_{T\in\mathcal{T}_h}\int_{\partial T} w\mathcal{A}\nabla v\cdot\mathbf{n} ds\\
&\quad - \sum_{T\in\mathcal{T}_h}\sum_{\square_j}(\nabla\cdot\mathcal{A}\nabla v,\gamma w)\\
&= \sum_{T\in\mathcal{T}_h}\int_{\partial T}\mathcal{A}\nabla v\cdot\mathbf{n}(\gamma w - w) ds + \sum_{T\in\mathcal{T}_h}(\mathcal{A}\nabla v,\nabla w)_T + \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla v, w)_T\\
&\quad - \sum_{T\in\mathcal{T}_h}\sum_{\square_j}(\nabla\cdot\mathcal{A}\nabla v,\gamma w)\\
&= (\mathcal{A}\nabla v,\nabla w) + \sum_{T\in\mathcal{T}_h}\int_{\partial T}\mathcal{A}\nabla v\cdot\mathbf{n}(\gamma w - w) ds + \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla v, w-\gamma w)_T. \qquad \square
\end{aligned}
$$

This lemma implies that Assumption 3 holds: The boundedness of $a(v,w)$ is straightforward. For the proof of coercivity (2.9) on V_h, notice the following. First of all, $|\!|\!| v |\!|\!| = |v|_{1,h}$, and so $C|\!|\!| v |\!|\!| \le (\mathcal{A}\nabla v,\nabla v)$ for all $v\in V_h$. The last two terms in the right side of (3.3) are the $O(h|v|_{1,h}^2)$ term when $v=w$. In fact, just as in estimating the I_4 term of (2.21), we have

$$
\sum_{T\in\mathcal{T}_h}\int_{\partial T}\mathcal{A}\nabla v\cdot\mathbf{n}(\gamma v - v) ds \le Ch\|A\|_{1,\infty}|v|_{1,h}^2
$$

and

$$\sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla v, v-\gamma v)_T = \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\cdot\nabla v, v-\gamma v)_T \le Ch\|A\|_{1,\infty}|v|_{1,h}^2, \tag{3.4}$$

where $\nabla\cdot\mathcal{A}$ is the vector obtained by applying the divergence rowwise. Thus for h small enough we have the coercivity. Note that this last term could be handled like (2.19), but this would require $\mathcal{A}$ to be in $W^{2,\infty}$, which is unnecessary.

Applying Theorems 2.1 and 2.2, we have the following results.

THEOREM 3.1. *If $u\in H_0^1(\Omega)\cap H^2(\Omega)$ and $f\in H^1(\Omega)$, then*

$$|||u-u_h||| \le Ch\|u\|_2,$$
$$\|u-u_h\| \le Ch^2(\|u\|_2+\|f\|_1),$$

where the L^2 estimate requires $\mathcal{A}\in W^{2,\infty}(\Omega)$.

The same conclusions hold for the conforming bilinear trial function case [9], and we omit the details.

3.2. Finite volume method with nonconforming trial functions. For a given regular triangulation $\mathcal{T}_h$, its dual partition $\mathcal{T}_h^*$ is the union of quadrilaterals. Each quadrilateral in $\mathcal{T}_h^*$ is made up of two subtriangles which share a common edge (see Figure 1). These subtriangles are formed by connecting the barycenter and the three corners of the triangles.

The trial function space associated with $\mathcal{T}_h$ for the nonconforming finite volume method is defined as

$$\begin{aligned} V_h = \{&v\in L^2(\Omega):\ v|_T\in P_1(T)\ \ \forall T\in\mathcal{T}_h,\\ &\text{is continuous at the midpoint of } e\in\mathcal{E}_h^0\\ &\text{and is zero at the midpoint of boundary edges } e \text{ on } \partial\Omega\}. \end{aligned}$$

The test function space is defined as in (1.3).

Let $\mathcal{M}$ be a set containing all of the midpoints of the interior edges associated with the triangulation $\mathcal{T}^h$. The operator $\gamma: V(h)\to Q_h$ is defined by

$$\gamma v(x) \equiv \sum_{P\in\mathcal{M}} v(P)\chi_P(x) \quad \forall x\in\Omega, \tag{3.5}$$

where χ_P is the characteristic function of dual element K_P^* associated with the node P. The mapping γ satisfies Assumptions 1 and 2 (see [8]). Finite volume methods using the above nonconforming trial functions were considered in [8, 6].

Our version [8] is to find $u_h\in V_h$ such that for any $v\in V_h$

$$a(u_h,v) = (f,\gamma v). \tag{3.6}$$

The bilinear form $A(v,w)$ in (2.5) reduces to $a(v,w)$ and

$$a(u,v) = -\sum_{K\in\mathcal{T}_h^*}\int_{\partial K}\mathcal{A}\nabla u\cdot\mathbf{n}\gamma v ds.$$

LEMMA 3.2. *For any $v,w\in V(h)$,*

$$\begin{aligned} a(v,w) = (\mathcal{A}\nabla_h v,\nabla_h w) &+ \sum_{T\in\mathcal{T}_h}\int_{\partial T}\mathcal{A}\nabla v\cdot\mathbf{n}(\gamma w - w)ds\\ &+ \sum_{T\in\mathcal{T}_h}(\nabla\cdot\mathcal{A}\nabla v, w-\gamma w)_T. \end{aligned}$$

Proof. See Lemma 3.2 in [24]. ∎

Using the above lemma, as before we can prove that (2.8) and (2.9) hold easily. Then we have the following estimates.

THEOREM 3.2. *If $u \in H_0^1(\Omega) \cap H^2(\Omega)$ and $f \in H^1(\Omega)$, then*

$$|||u - u_h||| \le Ch\|u\|_2$$
$$\|u - u_h\| \le Ch^2(\|u\|_2 + \|f\|_1),$$

where the L^2 estimate requires $\mathcal{A} \in W^{2,\infty}$.

The same conclusions hold for the finite volume method [10] using the rotated bilinear trial functions, i.e., the nonconforming Q_1 elements on rectangular grids [23]. We omit the details here.

3.3. Finite volume method with totally discontinuous trial functions. The finite volume method using totally discontinuous trial functions was first proposed in [24].

Let $\mathcal{T}_h$ be a quasiuniform triangulation of Ω. We define the dual partition $\mathcal{T}_h^*$ of $\mathcal{T}_h$ for the test function space as follows. We divide each $T \in \mathcal{T}_h$ into three triangles by connecting the barycenter and the three corners of the triangle as shown in Figure 2. Let $\mathcal{T}_h^*$ consist of all of these triangles T_j, $j = 1, 2, 3$.

We define the finite dimensional space associated with $\mathcal{T}_h$ for the trial functions as

$$V_h = \{v \in L^2(\Omega) : \ v|_T \in P_1(T) \ \forall T \in \mathcal{T}_h\}. \tag{3.7}$$

The test function space is defined as in (1.3). The operator $\gamma : V(h) \to Q_h$ is defined as

$$\gamma v|_T = \frac{1}{h_e}\int_e v|_T ds \quad \forall T \in \mathcal{T}_h, \tag{3.8}$$

where h_e is the length of the edge e. The operator γ satisfies (1.4)–(1.7) (see [25]).

The discontinuous finite volume method is to find $u_h \in V_h$ such that

$$A(u_h, v) = (f, \gamma v) \qquad \forall v \in V_h. \tag{3.9}$$

LEMMA 3.3. *For any $v, w \in V(h)$,*

$$a(v, w) = (\mathcal{A}\nabla_h v, \nabla_h w) + \sum_{T\in\mathcal{T}_h}\int_{\partial T} \mathcal{A}\nabla v \cdot \mathbf{n}(\gamma w - w)ds$$
$$+ \sum_{T\in\mathcal{T}_h} (\nabla \cdot \mathcal{A}\nabla v, w - \gamma w)_T. \tag{3.10}$$

Proof. See Lemma 2.1 in [25]. ∎

Using the above lemma, one can prove coercivity and boundedness.

LEMMA 3.4. *There is a constant C independent of h such that*

$$A(v, v) \ge C|||v|||^2 \qquad \forall v \in V_h \tag{3.11}$$

for any positive α if $\delta = 1$ and for α larger enough if $\delta = -1$.

Proof. See Lemma 2.2 in [25]. ∎

LEMMA 3.5. *For $v, w \in V(h)$, we have*

$$A(v,w) \le C|||v|||\,|||w|||. \tag{3.12}$$

Proof. See Lemma 2.3 in [25]. □

Since all of the conditions for Theorems 2.1 and 2.2 are satisfied, we have the following error estimates for the discontinuous finite volume method.

THEOREM 3.3. *If $u \in H_0^1(\Omega) \cap H^2(\Omega)$ and $f \in H^1(\Omega)$, then*

$$|||u - u_h||| \le Ch\|u\|_2,$$
$$\|u - u_h\| \le Ch^2(\|u\|_2 + \|f\|_1).$$

We point out that the above L^2 estimate was not obtained in [25].

REFERENCES

[1] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] R. E. BANK AND D. J. ROSE, *Some error estimates for the box method*, SIAM J. Numer. Anal., 24 (1987), pp. 777–787.

[3] S. BRENNER AND R. SCOTT, *Mathematical Theory of Finite Element Methods*, Springer, New York, 2002.

[4] F. BREZZI, B. COCKBURN, L. D. MARINI, AND E. SULI, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3293–3310.

[5] Z. CAI, J. MANDEL, AND S. MCCORMICK, *The finite volume element method for diffusion equations on general triangulations*, SIAM J. Numer. Anal., 28 (1991), pp. 392–402.

[6] P. CHATZIPANTELIDIS, *Finite volume methods for elliptic PDE's: A new approach*, Math. Model. Numer. Anal., 36 (2002), pp. 307–324.

[7] Z. CHEN, G. HUAN, AND Y. MA, *Computational Methods for Multiphase Flows in Porous Media*, SIAM, Philadelphia, PA, 2006.

[8] S. H. CHOU, *Analysis and convergence of a covolume method for the generalized Stokes problem*, Math. Comp., 66 (1997), pp. 85–104.

[9] S. H. CHOU AND D. Y. KWAK, *Analysis and convergence of a MAC scheme for the generalized Stokes problem*, Numer. Methods Partial Differential Equations, 13 (1997), pp. 147–162.

[10] S. H. CHOU AND D. Y. KWAK, *A covolume method based on rotated bilinears for the generalized Stokes problem*, SIAM J. Numer. Anal., 35 (1998), pp. 494–507.

[11] S. H. CHOU, D. Y. KWAK, AND K. Y. KIM, *Mixed finite volume methods on non-staggered quadrilateral grids for elliptic problems*, Math. Comp., 72 (2003), pp. 525–539.

[12] S. H. CHOU AND Q. LI, *Error estimates in L^2, H^1, and L^∞ in covolume methods for elliptic and parabolic problems, a unified approach*, Math. Comp., 69 (2000), pp. 103–120.

[13] S. H. CHOU AND P. S. VASSILEVSKI, *A general mixed covolume framework for constructing conservative schemes for elliptic problems*, Math. Comp., 68 (1999), pp. 991–1011.

[14] M. CROUZEIX AND P. A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equation* I, RAIRO Anal. Numer., 7 (1973), pp. 33–76.

[15] R. E. EWING, T. LIN, AND Y. LIN, *On the accuracy of the finite volume element method based on piecewise linear polynomials*, SIAM J. Numer. Anal., 39 (2002), pp. 1865–1888.

[16] R. EYMARD, GALLOUET, AND R. HERBIN, *Finite volume methods*, Handbook of Numerical Analysis VII, P. G. Ciarlet and J. L. Lions, eds., North-Holland, 2000, pp. 713–1020.

[17] J. HUANG AND S. XI, *On the finite volume elment method for general self-adjoint elliptic problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1762–1774.

[18] R. D. LAZAROV, I. D. MISHEV, AND P. S. VASSILEVSKI, *Finite volume methods for convection-diffusion problems*, SIAM J. Numer. Anal., 33 (1996), pp. 31–55.

[19] C. LIU AND S. F. MCCORMICK, *The finite volume-element method for planar cavity flow*, in Proceedings of the 11th International Conference on Numerical Methods in Fluid Dynamics, Williamsburg, 1988, Springer, Berlin, 1988.

[20] S. MCCORMICK, *Multilevel Adaptive Methods for Partial Differential Equations*, SIAM, Philadelphia, 1989.

[21] R. A. Nicholaides, T. A. Porsching, and C. A. Hall, *Covolume methods in computational fluid dynamics*, Comput. Dyn. Rev., M. Hafez and K. Oshima, eds., 1995, Wiley, New York, pp. 279–299.
[22] Z. Y. Chen, R. H. Li, and W. Wu, *Generalized Difference Methods for Differential Equations*, Marcel Dekker, New York, 2000.
[23] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111.
[24] X. Ye, *On the relationship between finite volume and finite element methods applied to the Stokes equations*, Numer. Methods Partial Differential Equations, 17 (2001), pp. 440–453.
[25] X. Ye, *A new discontinuous finite volume method for elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 1062–1072.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1654–1678

THE RESIDUAL-FREE-BUBBLE FINITE ELEMENT METHOD ON ANISOTROPIC PARTITIONS*

ANDREA CANGIANI† AND ENDRE SÜLI‡

Abstract. The subject of this work is the analysis and implementation of stabilized finite element methods on anisotropic meshes. We develop the anisotropic a priori error analysis of the residual-free-bubble (RFB) method applied to elliptic convection-dominated convection-diffusion problems in two dimensions, with finite element spaces of type Q_k, $k \geq 1$. In the case of P_1 finite elements, relying on the equivalence of the RFB method to classical stabilized finite element methods, we propose a new rule, justified through the analysis of the RFB method, for selecting the stabilization parameter in classical stabilized methods on two-dimensional anisotropic triangulations.

Key words. residual-free-bubble finite element method, convection-dominated diffusion problems, stabilized finite element methods

AMS subject classifications. 65N12, 65N39, 76M10

DOI. 10.1137/060658011

1. Introduction.

Elliptic convection-diffusion problems arise in a vast number of applications, and their stable, accurate, and efficient solution is of significant theoretical and practical interest. From the computational point of view, problems of this kind become particularly challenging when convection dominates diffusion in the sense that the Péclet number, which measures the magnitude of the convective vector field over the length scale of the computational domain relative to the size of the diffusion coefficient, is large. Convection-dominated diffusion equations exhibit features which resemble those of the *reduced*, first-order hyperbolic equation arising from the second-order elliptic convection-diffusion equation on neglecting the diffusion term. For example, the solution may contain thin internal layers within the computational domain; also, due to the singular perturbation nature of an elliptic convection-dominated diffusion problem, the solution may exhibit thin boundary layers along sections of the boundary of the computational domain which correspond to the outflow part of the boundary for the reduced problem. As a result of this, on meshes which do not resolve internal and boundary layers, standard Galerkin finite element methods have poor stability and accuracy properties. The difficulties typically manifest themselves as large, maximum-principle-violating, oscillations in the numerical solution which occur predominantly along the characteristics of the reduced problem.

The situation may be remedied by using a classical stabilized finite element method (such as a streamline-diffusion method or a Galerkin least-squares method) or a residual-free-bubble (RFB) finite element method; we refer to the monograph [28] for an extensive survey of the literature. Due to the presence of anisotropic numerical dissipation terms in the direction of the characteristics of the reduced equation whose role is to suppress undesirable numerical oscillations, these methods are capable of delivering accurate numerical solutions even on shape-regular computational meshes

*Received by the editors April 24, 2006; accepted for publication (in revised form) May 9, 2007; published electronically August 17, 2007.
http://www.siam.org/journals/sinum/45-4/65801.html
†Dipartimento di Matematica, Università di Pavia, 27000 Pavia, Italy (andrea.cangiani@unipv.it).
‡Numerical Analysis Group, Oxford University Computing Laboratory, Oxford OX1 3QD, England (endre.suli@comlab.ox.ac.uk).

whose granularity is relatively coarse compared to the thickness of internal and boundary layers. Alternatively, motivated by the fact that internal and boundary layers are highly localized and anisotropic, one may choose to use a standard Galerkin finite element method, albeit on a stretched, anisotropic, or layer-adapted (and, certainly, non-shape-regular) computational mesh (see, for example, the discussion in [28] on Shishkin-type meshes).

In recent years, there have been attempts to employ these remedies simultaneously; see, for example, the work of Apel and Lube [3] and Micheletti, Perotto, and Picasso [25] concerning classical stabilized finite element methods on anisotropic meshes. The developments in the present article are in a similar spirit.

The objective of this paper is twofold. We aim to develop the a priori error analysis of the RFB method for two-dimensional elliptic convection-dominated diffusion equations on anisotropic partitions. Specifically, we aim to bound the error by appropriately weighted norms of directional derivatives of the solution, so as to incorporate the anisotropic nature of the solution into the bounds. On the one hand, our results complement the work in [3, 25] on the a priori error analysis of classical stabilized finite element methods over anisotropic meshes; on the other hand, they extend earlier results by Brezzi, Marini, and Süli [7], Brezzi and Marini [8], and Sangalli [29] on the a priori error analysis of RFB methods on shape-regular triangulations.

Anisotropy also has to be taken into account in the selection of parameters appearing in stabilized finite element methods, such as *streamline-diffusion-type* methods. The second key objective of the paper is to use the stabilizing term derived from the RFB method to redefine the mesh Péclet number and propose a new choice of the streamline-diffusion (SD) parameter that is suitable for use on anisotropic partitions. The proposed choice of the SD parameter improves earlier suggestions based on the a priori analysis of the streamline-diffusion method (cf. [3, 23, 25]).

The paper is structured as follows. The first part of this work is concerned with the analysis of stabilized finite element methods on anisotropic computational meshes: We consider the anisotropic a priori error analysis of the RFB method applied to elliptic convection-dominated convection-diffusion problems in two dimensions. In the second part of the paper, in the case of P_1 finite elements on triangular meshes, appealing to the equivalence of the RFB method to classical stabilized finite element methods, we propose a new rule, justified through the analysis of the RFB method, for selecting the stabilization parameter in classical stabilized methods on two-dimensional anisotropic triangulations; we then relate our work to existing developments on classical stabilized finite element methods on anisotropic meshes, including [3, 23, 25].

2. Statement of the problem. Let $\Omega \subset \mathbb{R}^2$ be a bounded open polygonal domain. We consider the model elliptic boundary-value problem

$$\left\{ \begin{array}{l} \text{find } u \in V = \mathrm{H}_0^1(\Omega) \text{ such that} \\ Lu := -\varepsilon \Delta u + \boldsymbol{a} \cdot \nabla u = f \qquad \text{in } \Omega, \end{array} \right. \tag{2.1}$$

where ε is a positive parameter, $\boldsymbol{a} \in [\mathrm{W}^{1,\infty}(\Omega)]^2$, with $\mathrm{div}(\boldsymbol{a}) \leq 0$ in Ω, and f belongs to $\mathrm{L}^2(\Omega)$. The homogeneous Dirichlet boundary condition $u|_{\partial\Omega} = 0$ has been assumed here only for ease of presentation. We normalize the problem by requiring that $\|\boldsymbol{a}\|_{\mathrm{L}^\infty(\Omega)} \leq 1$. Our focus of interest is the convection-dominated regime, namely, when $0 < \varepsilon \ll 1$; thus we assume, without loss of generality, that $\varepsilon \in (0, 1]$. The extension of the results of this paper to the, more general, convection-diffusion-reaction equation $-\varepsilon\Delta u + \mathbf{a} \cdot \nabla u + cu = f$ in Ω, subject to a homogeneous Dirichlet boundary

condition on $\partial\Omega$, is straightforward, provided that $\mathrm{div}(\mathbf{a}) - 2c \leq 0$ in Ω. Below, we shall briefly comment on the case when $\mathrm{div}(\mathbf{a}) - 2c \leq -2c_0$ in Ω, where c_0 is a positive constant.

The variational formulation of the boundary-value problem (2.1) is

$$\left\{ \begin{array}{l} \text{find } u \in V \text{ such that} \\ \mathcal{L}(u,v) = (f,v) \qquad \forall v \in V, \end{array} \right. \tag{2.2}$$

where

$$\mathcal{L}(w,v) := \varepsilon \int_\Omega \nabla w \cdot \nabla v \, \mathrm{d}\boldsymbol{x} + \int_\Omega (\boldsymbol{a} \cdot \nabla w)\, v \, \mathrm{d}\boldsymbol{x} \tag{2.3}$$

is a continuous and coercive bilinear form on $V \times V$ and $(\cdot,\cdot)$ denotes the L^2 inner product over Ω.

The existence and uniqueness of a solution to (2.2) (that is, of a weak solution to (2.1)) are well-known consequences of the Lax–Milgram lemma; for a more general existence and uniqueness result, see [19, Theorem 8.6].

We consider finite element discretizations of (2.2) over *conforming* partitions $\mathcal{T}_h$ of $\overline{\Omega}$ consisting of affine-equivalent quadrilateral or triangular elements. We shall not assume that the family of partitions $\{\mathcal{T}_h\}_{h>0}$ is *shape-regular*, because we wish to allow anisotropic local refinements in parts of the computational domain where special features of the exact solution, such as thin layers, are detected. Our only assumption will be the existence of a positive constant $c \leq 1$ such that

$$\varepsilon \leq c h_\gamma, \tag{2.4}$$

for all element edges γ in the partition; here h_γ represents the length of γ. This is a reasonable assumption when dealing with the analysis of stabilized finite element methods for convection-dominated diffusion problems such as our model problem, which exhibits boundary layers whose thickness is commensurate with $\varepsilon \ll 1$: For, if we could afford to solve the problem on meshes whose granularity is smaller than ε, then we would not need to use a stabilized method in the first place. Thus, our a priori error bounds, developed under the hypothesis (2.4), will be of a *preasymptotic* nature: Since the lower bound $\varepsilon \ll 1$ on ch_γ is fixed, we will *not* let h_γ tend to zero.

An optimal mesh (in terms of the number of degrees of freedom required to obtain a given accuracy) must mimic the behavior of the solution to (2.1). Such an optimal mesh would, in general, be designed through successive mesh refinements/derefinements. In early stages of the mesh adaptation process, the use of a stabilized finite element method is mandatory, since on coarse meshes classical Galerkin finite element approximations of (2.1) will exhibit large maximum-principle-violating numerical oscillations when $\varepsilon \ll 1$, hence the need for sharp preasymptotic error bounds for stabilized finite element methods. In later stages of the mesh refinement process, when the mesh has been adapted to the solution, the stabilized method could be simplified, for instance, by omitting the stabilization term, as was done in [10].

We denote by λ_1 and λ_2 some characteristic dimensions of a generic element $T \in \mathcal{T}_h$, to be defined on a case-by-case basis; λ_1 and λ_2 are used to group the elements according to the following rule (which defines the subpartitions $\mathcal{T}_1$ and $\mathcal{T}_2$):

1. $T \in \mathcal{T}_1$ if $\lambda_1 \leq \lambda_2$;
2. $T \in \mathcal{T}_2$ if $\lambda_2 < \lambda_1$.

An admissible structured mesh and its subpartitions are shown in Figure 2.1.

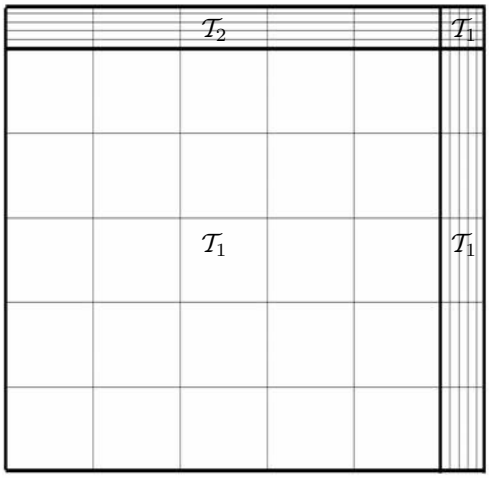


FIG. 2.1. *A locally anisotropic partition and its subpartitions $\mathcal{T}_1$ and $\mathcal{T}_2$.*

Given $k \geq 1$, let $\mathcal{P}_k$ denote the space of algebraic polynomials of degree $\leq k$, and let $\mathcal{Q}_k$ denote the space of algebraic polynomials of degree $\leq k$ with respect to each variable. Further, let $F_T : \widehat{T} \to T$ be the affine transformation mapping the reference element onto $T \in \mathcal{T}_h$.

The *residual-free-bubble* space is defined as follows (see [7]):

$$V_{\mathrm{RFB}} := \left\{ v \in V : v_{|_e} \in \mathcal{P}_k \text{ for each edge } e \text{ of } T \text{ and any element } T \in \mathcal{T}_h \right\}. \tag{2.5}$$

We note that the space V_{RFB} is infinite-dimensional, admitting the representation

$$V_{\mathrm{RFB}} = V_h^k + B_h, \tag{2.6}$$

where V_h^k is the classical finite element space given by

$$V_h^k := \left\{ v_h \in \mathrm{H}_0^1(\Omega) : \left\{ \begin{array}{ll} v_{h|_T} \in \mathcal{P}_k & \text{if } T \text{ is a triangle} \\ v_{h|_T} \circ F_T \in \mathcal{Q}_k & \text{if } T \text{ is a parallelogram} \end{array} \right. \right\},$$

and

$$B_h := \bigoplus_{T \in \mathcal{T}_h} \mathrm{H}_0^1(T) \tag{2.7}$$

is the space of all *bubble functions* in V; i.e., all function with zero trace on the skeleton of the partition $\mathcal{T}_h$.

The RFB approximation of (2.2) is defined as the Galerkin approximation of (2.2) in the space V_{RFB}:

$$\left\{ \begin{array}{l} \text{find } u_{\mathrm{RFB}} \in V_{\mathrm{RFB}} \text{ such that} \\ \mathcal{L}(u_{\mathrm{RFB}}, v) = (f, v) \qquad \forall v \in V_{\mathrm{RFB}}. \end{array} \right. \tag{2.8}$$

Since V_{RFB} is infinite-dimensional, the formulation (2.8) does not represent a numerical method in the classical sense. In fact, a numerical algorithm can be devised from (2.8) through *static condensation* of the bubble component u_b of the solution u_{RFB},

which belongs to the infinite-dimensional space B_h, and then discretizing the resulting infinite-dimensional problem over the finite-dimensional space V_h^k. For instance, if $k \leq 2$, the sum in (2.7) is direct, and hence we then have the following unique decomposition of the RFB solution:

$$u_{\mathrm{RFB}} = u_h + u_b.$$

Consequently, by testing in V_h^k and then in B_h, we can split (2.8) into the following two problems:

$$\mathcal{L}(u_h, v_h) + \mathcal{L}(u_b, v_h) = (f, v_h) \qquad \forall v_h \in V_h^k, \tag{2.9}$$

$$\mathcal{L}(u_h, v_b) + \mathcal{L}(u_b, v_b) = (f, v_b) \qquad \forall v_b \in B_h. \tag{2.10}$$

Equation (2.10) is referred to as a *bubble equation* as it is equivalent to solving, in each element $T \in \mathcal{T}_h$, the boundary-value problem

$$\begin{cases} Lu_b = f - Lu_h & \text{in } T, \\ u_b = 0 & \text{on } \partial T \end{cases} \tag{2.11}$$

for the "fine-scale" bubble component u_b of the approximate solution u_{RFB} in terms of the "coarse-scale" piecewise polynomial component u_h of u_{RFB}. The static condensation procedure corresponds to eliminating u_b from (2.9) in favor of u_h using (2.11). This can be done by numerically solving a finite number of independent local problems such as (2.11); this then leads to a (fully discrete) numerical algorithm. An instance of such a procedure is discussed in section 6 of this paper. For further details, we refer the reader to [9, 7].

The general a priori error analysis of the RFB method on shape-regular partitions is due to Brezzi, Marini, and Süli [7]; it was shown there that if $u \in \mathrm{H}^{k+1}(\Omega)$, then the numerical solution u_{RFB} delivered by the RFB method satisfies the following optimal asymptotic error bound in the energy norm:

$$\varepsilon^{1/2}|u - u_{\mathrm{RFB}}|_{1,\Omega} \leq Ch^{k+1/2}\|u\|_{\mathrm{H}^{k+1}(\Omega)}, \tag{2.12}$$

where h represents the characteristic size of the partition.

The technique used here to extend the a priori error analysis of the RFB method to anisotropic partitions is different from the one employed in [7]. Instead, we follow the approach adopted by Sangalli [29] to subsequently rederive and localize the results presented in [7]. The key idea of Sangalli's approach, and of the analysis below, is to exploit the approximation properties of the space V_{RFB}. To do so, Sangalli explicitly constructs a projector from H^1 onto the RFB space in a certain ε-weighted H^1 norm. A similar approach is followed by Risch in [27].

A second key ingredient of our analysis is the use of anisotropic approximation results. These must be employed in order to derive an a priori error bound in terms of appropriately weighted norms of directional derivatives of the exact solution u.

3. Structured quadrilateral partitions. We begin with the case of axiparallel rectangular elements, leaving the treatment of more general partitions to subsequent sections.

In this case it is natural to define $\lambda_1 = h_1$ and $\lambda_2 = h_2$, where h_1 and h_2 denote the dimensions of the generic element $T \in \mathcal{T}$ in the x_1 and x_2 coordinate directions, respectively.

3.1. Notations and preliminary results. Let $\widehat{T} = (-1,1)^2$ be the master element. Given a function $v \in \mathrm{H}^1(T)$, we consider $\hat{v} \in \mathrm{H}^1(\widehat{T})$, the function associated to v through the affine transformation F_T which maps $\widehat{T}$ into T; hence $\hat{v} := v \circ F_T$. Further, we denote by $i^* = 3 - i$ the complementary index to i with respect to the set $\{1,2\}$.

Since T is a rectangle, the usual scaling properties for functions $v \in \mathrm{H}^1(T)$ yield

$$\|v\|_{0,T}^2 = \frac{1}{4} h_1 h_2 \|\hat{v}\|_{0,\widehat{T}}^2, \tag{3.1}$$

$$\left\| \frac{\partial v}{\partial x_i} \right\|_{0,T}^2 = \frac{h_{i^*}}{h_i} \left\| \frac{\partial \hat{v}}{\partial \widehat{x}_i} \right\|_{0,\widehat{T}}^2, \qquad i \in \{1,2\}. \tag{3.2}$$

We will also need some scaling properties for functions defined over edges of the elements $T \in \mathcal{T}_h$. The trace of a function belonging to the space $\mathrm{H}^1(T) = \mathrm{W}^{1,2}(T)$ and, more generally, to the Sobolev space $\mathrm{W}^{1,p}(T)$, $1 \le p < \infty$, is characterized in terms of the *fractional-order* Sobolev space $\mathrm{W}^{1-1/p,p}(\partial T)$, which, for $p > 1$, can be defined using the real method of function space interpolation; see, e.g., Adams [1].

The space $\mathrm{W}^{s,p}(\partial T)$, $0 < s < 1$, can also be characterized in terms of an intrinsically defined norm. For instance, for every $s \in (0,1)$, the norm $\|\cdot\|_{s,\partial T}$ and seminorm $|\cdot|_{s,\partial T}$ of the Sobolev space $\mathrm{H}^s(\partial T) = \mathrm{W}^{s,2}(\partial T)$ of fractional order s are defined by

$$\begin{aligned} \|v\|_{s,\partial T} &:= \left\{ \|v\|_{0,\partial T}^2 + \int_{\partial T} \int_{\partial T} \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|^2}{|\boldsymbol{x} - \boldsymbol{y}|^{1+2s}} \, \mathrm{d}\sigma(\boldsymbol{x}) \, \mathrm{d}\sigma(\boldsymbol{y}) \right\}^{1/2} \\ &= \left\{ \|v\|_{0,\partial T}^2 + |v|_{s,\partial T}^2 \right\}^{1/2}, \end{aligned} \tag{3.3}$$

where $\mathrm{d}\sigma$ denotes the one-dimensional Hausdorff measure of ∂T. This definition can be extended to portions of ∂T.

The trace theorem (again, see [1]) ensures that the trace of a function $v \in \mathrm{H}^s(T)$ belongs to $\mathrm{H}^{s-1/2}(\partial T)$, $s \in (1/2, 1]$, and that there exists a constant C, independent of v, such that

$$\|v\|_{s-1/2,\partial T} \le C \|v\|_{s,T} \qquad \forall v \in \mathrm{H}^s(T). \tag{3.4}$$

Let γ be an edge of $T \in \mathcal{T}_h$ and $\widehat{\gamma} = F_T^{-1}(\gamma)$ the corresponding edge of $\widehat{T}$. Scaling the Sobolev seminorm $|\cdot|_{s,\gamma}$, $0 \le s \le 1$, from $\widehat{\gamma}$ to γ, we have

$$|v|_{s,\gamma}^2 = \left(\frac{h_\gamma}{2} \right)^{1-2s} |\hat{v}|_{s,\widehat{\gamma}}^2 \qquad \forall v \in \mathrm{H}^s(\gamma), \tag{3.5}$$

where, as before, $h_\gamma = |\gamma|$. The scaling property (3.5) will be used to prove the following anisotropic trace inequalities which are refinements of the usual ones valid for axiparallel domains.

LEMMA 3.1. *Let $v \in \mathrm{H}^1(T)$, where T is an axiparallel rectangle in $\mathbb{R}^2$, and let γ_i be an edge of T parallel to the ith coordinate axis, with $h_i = |\gamma_i|$, $i = 1,2$. The following trace inequalities hold:*

$$\|v\|_{0,\gamma_i}^2 \le \frac{1}{h_{i^*}} \|v\|_{0,T}^2 + 2 \|v\|_{0,T} \|v_{x_{i^*}}\|_{0,T}, \qquad i = 1,2; \tag{3.6}$$

$$|v|_{1/2,\partial T}^2 \le C \left(\frac{1}{h_1 h_2} \|v\|_{0,T}^2 + \frac{h_1}{h_2} \|v_{x_1}\|_{0,T}^2 + \frac{h_2}{h_1} \|v_{x_2}\|_{0,T}^2 \right), \tag{3.7}$$

where the constant C is independent of h_1 and h_2.

Proof. The proof of (3.6) can be found, for instance, in [17]. To prove (3.7), we apply (3.5) with $s = 1/2$ to scale from ∂T to $\partial\widehat{T}$ and the trace inequality (3.4) to shift from $\partial\widehat{T}$ to $\widehat{T}$, and, finally, we use (3.1) and (3.2) to scale back from $\widehat{T}$ to T:

$$\begin{aligned}|v|^2_{1/2,\partial T} &= |\hat{v}|^2_{1/2,\partial\widehat{T}} \leq \|\hat{v}\|^2_{1/2,\partial\widehat{T}} \leq C\,\|\hat{v}\|^2_{1,\widehat{T}}\\ &= C\left(\|\hat{v}\|^2_{0,\widehat{T}} + \|\hat{v}_{x_1}\|^2_{0,\widehat{T}} + \|\hat{v}_{x_2}\|^2_{0,\widehat{T}}\right)\\ &= C\left(\frac{1}{h_1h_2}\,\|v\|^2_{0,T} + \frac{h_1}{h_2}\,\|v_{x_1}\|^2_{0,T} + \frac{h_2}{h_1}\,\|v_{x_2}\|^2_{0,T}\right)\end{aligned}$$

and hence the desired result for any $v \in \mathrm{H}^1(T)$. ∎

We shall also require the following trace-lifting lemma (see, e.g., Sangalli [29]).

LEMMA 3.2. *Given a function $\hat{w}_0 \in \mathrm{H}^{1/2}(\partial\widehat{T})$ and a real parameter t, with $0 < t \leq 1$, there exists $\hat{w} \in \mathrm{H}^1(\widehat{T})$ such that $\hat{w} = \hat{w}_0$ on $\partial\widehat{T}$ and*

$$t\,|\hat{w}|^2_{1,\widehat{T}} + t^{-1}\,\|\hat{w}\|^2_{0,\widehat{T}} \leq C\left(t\,|\hat{w}_0|^2_{1/2,\partial\widehat{T}} + \|\hat{w}_0\|^2_{0,\partial\widehat{T}}\right), \tag{3.8}$$

where the constant C is independent of t and $\hat{w}_0$.

3.2. The projection error. Let us consider the function space $\mathrm{H}^{r_1,r_2}(T)$ of dominant mixed smoothness, defined by

$$\mathrm{H}^{r_1,r_2}(T) := \left\{v \in \mathrm{L}^2(T) : D^{r_1}_{x_1}v,\ D^{r_2}_{x_2}v,\ D^{r_1}_{x_1}D^{r_2}_{x_2} \in \mathrm{L}^2(T)\right\}.$$

It is known that if $r_i > 1/2$, $i = 1, 2$, then $\mathrm{H}^{r_1,r_2}(T)$ is continuously embedded into the space $\mathrm{C}(\overline{T})$ of uniformly continuous functions on $\overline{T}$ (see, for example, [32, Chapter 2, Theorem 2.2.3]). Trivially, $\mathrm{H}^{r+1}(T)$ is continuously embedded into $\mathrm{H}^{1,1}(T)$ for any $r \geq 1$.

We begin by introducing a suitable interpolant from $\mathcal{Q}_k$ of a generic function in $\mathrm{H}^{1,1}(T)$— the tensor-product H^1-projection operator Π_k, as has been defined in [17] (see also [31, 18]), by means of truncated Legendre expansions.

DEFINITION 3.3. *Let L_n denote the Legendre polynomial of degree n on the open interval $I = (-1, 1)$. We define the* L^2-projection operator

$$\tilde{\pi}_k : \mathrm{L}^2(I) \to \mathcal{P}_k(I)$$

by

$$\tilde{\pi}_k v(x) := \sum_{n=0}^{k} a_n L_n(x),$$

where

$$a_n := \frac{2n+1}{2}\int_I v(x)L_n(x)\,\mathrm{d}x.$$

Further, we define the H^1-projection operator

$$\hat{\pi}_k : \mathrm{H}^1(I) \to \mathcal{P}_k(I)$$

by setting, for any $v \in \mathrm{H}^1(I)$,

$$\hat{\pi}_k v(x) := \int_{-1}^{x} \tilde{\pi}_{k-1}(v')(\eta)\,\mathrm{d}\eta + v(-1), \qquad x \in (-1, 1).$$

A convenient feature of the above definition is that it can be easily extended to the multidimensional setting by means of a tensor-product construction; this is achieved at the cost of assuming additional regularity (viz. assuming $\mathrm{H}^{1,1}$-regularity instead of H^1-regularity).

DEFINITION 3.4. *Let $\widehat{T} = (-1,1)^2$. We define the* tensor-product projection operator

$$\widehat{\Pi}_k : \mathrm{H}^{1,1}(\widehat{T}) \to \mathcal{Q}_k(\widehat{T})$$

by

$$\widehat{\Pi}_k := \hat{\pi}_k^{x_1} \circ \hat{\pi}_k^{x_2},$$

where $\hat{\pi}_k^{x_1}$, $\hat{\pi}_k^{x_2}$ denote the one-dimensional H^1-projection operators from Definition 3.3, and the superscripts x_i, $i = 1, 2$, indicate the directions in which the one-dimensional projections are applied.

The above definition is easily extended to a generic axiparallel rectangle T as follows.

DEFINITION 3.5. *Let $T \in \mathcal{T}_h$. We define the* tensor-product projection operator

$$\Pi_k : \mathrm{H}^{1,1}(T) \to \mathcal{Q}_k(T)$$

by setting, for any $v \in \mathrm{H}^{1,1}(T)$,

$$\Pi_k v := \widehat{\Pi}_k \hat{v} \circ F_T^{-1}.$$

By virtue of being of tensor-product type, the projection Π_k admits anisotropic error bounds. As a matter of fact, it is better-behaved than the L^2-projection operator when bounds on the derivatives of the interpolation error are needed. The relevant approximation properties of Π_k are summarized in the next lemma.

LEMMA 3.6. *Suppose that T is an axiparallel rectangle and $v \in \mathrm{H}^{r+1}(T)$, with $1 \le r \le k$—and thereby $v \in \mathrm{H}^{1,1}(T)$. Then, for any s with $0 \le s \le r$, the following error bound holds:*

$$\|v - \Pi_k v\|_{0,T}^2 \le \Phi_2(k,s) \left(\left(\frac{h_1}{2}\right)^{2s+2} \|\partial_{x_1}^{s+1} v\|_{0,T}^2 + \left(\frac{h_2}{2}\right)^{2s+2} \|\partial_{x_2}^{s+1} v\|_{0,T}^2 \right) + \Phi_2(k,s-1) \min_{\substack{i,\, j \,=\, 1,\, 2 \\ i \,\neq\, j}} \left(\frac{h_i}{2}\right)^2 \left(\frac{h_j}{2}\right)^{2s} \|\partial_{x_j}^s \partial_{x_i} v\|_{0,T}^2,$$

and, for any $i = 1, 2$,

$$\|\partial_{x_i}(v - \Pi_k v)\|_{0,T}^2 \le \Phi_1(k,s) \left(\frac{h_i}{2}\right)^{2s} \|\partial_{x_i}^{s+1} v\|_{0,T}^2 + \Phi_2(k,s-1) \left(\frac{h_{i^*}}{2}\right)^{2s} \|\partial_{x_{i^*}}^s \partial_{x_i} v\|_{0,T}^2,$$

where

$$\Phi_1(k,s) := \left(\frac{\Gamma(k-s+1)}{\Gamma(k+s+1)}\right)^{1/2}, \qquad \Phi_2(k,s) := \frac{\Phi_1(k,s)}{\sqrt{k(k+1)}},$$

and Γ is the Gamma function.

The proof of the interpolation error bounds stated in the above lemma has been given by Georgoulis in [17] (see also [18]), where such results are presented in a much more general setting.

Remark. Interpolation error bounds similar to those in Lemma 3.6 are provided, although for a different interpolation operator, by Apel [2, Theorem 2.7]. These, too, are limited to rectangular elements and are obtained as improvements of the general but slightly less sharp bounds presented in earlier sections of [2]; see also section 4 (especially Theorem 4.10) in the recent work of Georgoulis, Hall, and Houston [16] concerning interpolation results on anisotropic nonaxiparallel meshes. For a recent survey of anisotropic mesh adaptivity and anisotropic interpolation error estimates, particularly on triangular meshes, we refer to the work of Huang [21].

3.3. Error bound. Suppose that the bounded polygonal domain $\Omega \subset \mathbb{R}^2$ is a finite union of axiparallel rectangles. We begin the error analysis with the construction of a suitable projector $P : \mathrm{H}^1_0(\Omega) \cap \mathrm{H}^2(\Omega) \to V_{\mathrm{RFB}}$, whose definition is based on the $\mathrm{H}^{1,1}$-projection operator Π_k described above and the trace-lifting lemma, Lemma 3.2.

Given $\hat{v} \in \mathrm{H}^{1,1}(\widehat{T}) \subset \mathrm{H}^1(\widehat{T})$, let $\hat{w} \in \mathrm{H}^1(\widehat{T})$ be the function obtained by applying Lemma 3.2 with

$$\hat{w}_0 = (\hat{v} - \widehat{\Pi}_k \hat{v})_{|_{\partial \widehat{T}}}, \qquad t = \frac{\varepsilon}{h_i}.$$

We note that $t \leq 1$ due to assumption (2.4). We define $P_{\widehat{T}} \hat{v} \in \mathrm{H}^1(\hat{T})$ by

$$P_{\widehat{T}} \hat{v} := \hat{v} - \hat{w}, \tag{3.9}$$

and let $P_T v = P_{\widehat{T}} \hat{v} \circ F_T^{-1}$. Finally, for $v \in \mathrm{H}^1_0(\Omega) \cap \mathrm{H}^2(\Omega)$, we define $Pv \in \mathrm{H}^1_0(\Omega)$ elementwise by $(Pv)|_T = P_T(v|_T)$, $T \in \mathcal{T}_h$; recall that $v|_T \in \mathrm{H}^2(T) \subset \mathrm{H}^{1,1}(T)$, so this definition is meaningful. It is clear from this construction that, for every element $T \in \mathcal{T}_h$, $P_T : \mathrm{H}^{1,1}(T) \to V_{\mathrm{RFB}}|_T$, and $P : \mathrm{H}^1_0(\Omega) \cap \mathrm{H}^2(\Omega) \to V_{\mathrm{RFB}}$.

The main task in the a priori error analysis is to bound the quantity $\mathcal{E}^P_T(v)$ defined for $v \in \mathrm{H}^{1,1}(T)$ by

$$\mathcal{E}^P_T(v) := \varepsilon |v - P_T v|^2_{1,T} + \varepsilon^{-1} \|v - P_T v\|^2_{0,T}. \tag{3.10}$$

To this end, let us assume that $T \in \mathcal{T}_i$, with $i \in \{1, 2\}$. Using (3.1) and (3.2), and noting that for $T \in \mathcal{T}_i$ we have $h_i \leq h_{i^*}$, it follows that

$$\begin{aligned} \mathcal{E}^P_T(v) &= \varepsilon \frac{h_i}{h_{i^*}} \|(\hat{v} - P_{\widehat{T}} \hat{v})_{\hat{x}_{i^*}}\|^2_{0,\widehat{T}} + \varepsilon \frac{h_{i^*}}{h_i} \|(\hat{v} - P_{\widehat{T}} \hat{v})_{\hat{x}_i}\|^2_{0,\widehat{T}} + \frac{\varepsilon^{-1} h_{i^*} h_i}{4} \|\hat{v} - P_{\widehat{T}} \hat{v}\|^2_{0,\widehat{T}} \\ &\leq C h_{i^*} \left(\frac{\varepsilon}{h_i} |\hat{v} - P_{\widehat{T}} \hat{v}|^2_{1,\widehat{T}} + \left(\frac{\varepsilon}{h_i} \right)^{-1} \|\hat{v} - P_{\widehat{T}} \hat{v}\|^2_{0,\widehat{T}} \right). \end{aligned} \tag{3.11}$$

Hence, by applying (3.8) in (3.11) with $\hat{w} = \hat{v} - P_{\widehat{T}} \hat{v}$, we have

$$\mathcal{E}^P_T(v) \leq C \left(\varepsilon \frac{h_{i^*}}{h_i} |\hat{v} - \widehat{\Pi}_k \hat{v}|^2_{1/2,\partial\widehat{T}} + h_{i^*} \|\hat{v} - \widehat{\Pi}_k \hat{v}\|^2_{0,\partial\widehat{T}} \right). \tag{3.12}$$

We are now in a position to prove the following result which justifies our choice of the projector P.

LEMMA 3.7. *Let $T \in \mathcal{T}$ and $v \in \mathrm{H}^{r+1}(T)$, with $1 \le r \le k$, and consider the quantity $\mathcal{E}_T^P(v)$ defined by* (3.10). *If $T \in \mathcal{T}_i$, $i \in \{1,2\}$, then*

$$\begin{aligned}\mathcal{E}_T^P(v) \le & \frac{C}{2^{2r+1}}\left(\Phi_{12}(k,r)\left(h_i^{2r+1}\|\partial_{x_i}^{r+1}v\|_{0,T}^2 + \frac{h_{i^*}^{2r+2}}{h_i}\|\partial_{x_{i^*}}^{r+1}v\|_{0,T}^2\right)\right.\\ &\left. + \frac{5}{2}\Phi_2(k,r-1)\left(h_i^{2r-1}h_{i^*}^2\|\partial_{x_i}^r\partial_{x_{i^*}}v\|_{0,T}^2 + h_ih_{i^*}^{2r}\|\partial_{x_i}\partial_{x_{i^*}}^r v\|_{0,T}^2\right)\right),\end{aligned} \tag{3.13}$$

where $\Phi_{12}(k,r) := 2\Phi_1(k,r) + \Phi_2(k,r)/2$.

Proof. Assume that $T \in \mathcal{T}_i$, $i \in \{1,2\}$, and let $\partial_{x_i}T$ and $\partial_{x_{i^*}}T$ be the collection of the edges of T parallel to the x_i and x_{i^*} coordinate directions, respectively. From (3.12), upon returning to ∂T using (3.5) and applying the trace inequalities of Lemma 3.1, we have

$$\begin{aligned}\mathcal{E}_T^P(v) &\le C\left(\varepsilon\frac{h_{i^*}}{h_i}|v-\Pi_k v|_{1/2,\partial T}^2 + 4\|v-\Pi_k v\|_{0,\partial_{x_{i^*}}T}^2 + 4\frac{h_{i^*}}{h_i}\|v-\Pi_k v\|_{0,\partial_{x_i}T}^2\right)\\ &\le C\left(\left(\frac{\varepsilon}{h_i^2}+\frac{1}{h_i}\right)\|v-\Pi_k v\|_{0,T}^2 + \varepsilon\frac{h_{i^*}^2}{h_i^2}\|(v-\Pi_k v)_{x_{i^*}}\|_{0,T}^2 + \varepsilon\|(v-\Pi_k v)_{x_i}\|_{0,T}^2\right.\\ &\quad\left. + \frac{h_{i^*}}{h_i}\|v-\Pi_k v\|_{0,T}\|(v-\Pi_k v)_{x_{i^*}}\|_{0,T} + \|v-\Pi_k v\|_{0,T}\|(v-\Pi_k v)_{x_i}\|_{0,T}\right)\\ &\le C\left(\left(\frac{\varepsilon}{h_i^2}+\frac{1}{h_i}\right)\|v-\Pi_k v\|_{0,T}^2\right.\\ &\quad\left. + \left(\varepsilon\frac{h_{i^*}^2}{h_i^2}+\frac{h_{i^*}^2}{h_i}\right)\|(v-\Pi_k v)_{x_{i^*}}\|_{0,T}^2 + (\varepsilon+h_i)\|(v-\Pi_k v)_{x_i}\|_{0,T}^2\right).\end{aligned}$$

With assumption (2.4) this bound may be written

$$\mathcal{E}_T^P(v) \le C\left(\frac{1}{h_i}\|v-\Pi_k v\|_{0,T}^2 + \frac{h_{i^*}^2}{h_i}\|(v-\Pi_k v)_{x_{i^*}}\|_{0,T}^2 + h_i\|(v-\Pi_k v)_{x_i}\|_{0,T}^2\right).$$

Thus, we have bounded $\mathcal{E}_T^P(v)$ in terms of the H^1-projection error. The required bound (3.13) follows by applying the projection error bounds from Lemma 3.6. ∎

We are ready to prove the following a priori error bound for the RFB method in the energy norm $\varepsilon^{1/2}|\cdot|_{1,\Omega}$.

THEOREM 3.8. *Let $u \in V$ be the solution of* (2.2) *and $u_{\mathrm{RFB}} \in V_{\mathrm{RFB}}$ the RFB solution defined by* (2.8). *Assume that the partition $\mathcal{T}_h$ consists of axiparallel rectangles and that there exists a constant $c \in (0,1]$ such that, for any $T \in \mathcal{T}_h$, $\varepsilon \le c\min\{h_1,h_2\}$. Finally, let $\mathcal{T}_1$ be the subpartition given by all $T \in \mathcal{T}_h$ such that $h_1 \le h_2$, and let $\mathcal{T}_2 := \mathcal{T}_h \setminus \mathcal{T}_1$.*

If $u \in \mathrm{H}_0^1(\Omega) \cap \mathrm{H}^{k+1}(\Omega)$, then there exists a positive constant C, independent of ε, k and of the mesh dimensions, such that for any $1 \le r \le k$

$$\begin{aligned}\varepsilon^{1/2}|u-u_{\mathrm{RFB}}|_{1,\Omega} \le C\frac{\bar{\Phi}(k,r)}{2^{r+1/2}}\sum_{i=1}^{2}\left(\sum_{T\in\mathcal{T}_i}\left(h_i^{2r+1}\|\partial_{x_i}^{r+1}u\|_{0,T}^2 + \frac{h_{i^*}^{2r+2}}{h_i}\|\partial_{x_{i^*}}^{r+1}u\|_{0,T}^2\right.\right.\\ \left.\left. + h_ih_{i^*}^{2r}\|\partial_{x_i}\partial_{x_{i^*}}^r u\|_{0,T}^2 + h_i^{2r-1}h_{i^*}^2\|\partial_{x_i}^r\partial_{x_{i^*}}u\|_{0,T}^2\right)\right)^{1/2},\end{aligned} \tag{3.14}$$

where $\bar{\Phi}(r,k) := \max\{\Phi_{12}(k,r), \frac{5}{2}\Phi_2(k,r-1)\}$. *The constant* C *depends only on the constant in the trace inequality* (3.7) *and on the constant in Lemma* 3.2.

Proof. We consider the decomposition

$$u - u_{\mathrm{RFB}} = (u - Pu) + (Pu - u_{\mathrm{RFB}}),$$

where P is the approximation operator described in the previous section. By employing the coercivity of $\mathcal{L}$ and the Galerkin orthogonality property, on recalling that $Pu \in V_{\mathrm{RFB}}$, we have that

$$\begin{aligned}\varepsilon|u - u_{\mathrm{RFB}}|^2_{1,\Omega} &\le \mathcal{L}(u - u_{\mathrm{RFB}}, u - u_{\mathrm{RFB}})\\ &= \mathcal{L}(u - u_{\mathrm{RFB}}, u - Pu).\end{aligned}$$

Thus, on applying the Cauchy–Schwarz inequality to $\mathcal{L}(u-u_{\mathrm{RFB}}, u-Pu)$ after rewriting it explicitly using the definition of the bilinear form (2.3), we get

$$\begin{aligned}\varepsilon|u - u_{\mathrm{RFB}}|^2_{1,\Omega} &\le \sum_{T\in\mathcal{T}_h}\left(\varepsilon\int_T \nabla(u - u_{\mathrm{RFB}})\cdot\nabla(u - P_T u)\ \mathrm{d}\boldsymbol{x}\right.\\ &\qquad\left. + \int_T \boldsymbol{a}\cdot\nabla(u - u_{\mathrm{RFB}})(u - P_T u)\ \mathrm{d}\boldsymbol{x}\right)\\ &\le \sum_{T\in\mathcal{T}_h}\left(\varepsilon^{1/2}|u - u_{\mathrm{RFB}}|_{1,T}\right)\left(\varepsilon^{1/2}|u - P_T u|_{1,T} + \varepsilon^{-1/2}\|u - P_T u\|_{0,T}\right)\\ &\le \varepsilon^{1/2}|u - u_{\mathrm{RFB}}|_{1,\Omega}\left(\sum_{T\in\mathcal{T}_h}\left(\varepsilon^{1/2}|u - P_T u|_{1,T} + \varepsilon^{-1/2}\|u - P_T u\|_{0,T}\right)^2\right)^{1/2}.\end{aligned}$$

Next, we split the sum on the right-hand side between the subpartitions $\mathcal{T}_1$ and $\mathcal{T}_2$ to obtain

$$\varepsilon^{1/2}|u - u_{\mathrm{RFB}}|_{1,\Omega} \le C\sum_{i=1,2}\left(\sum_{T\in\mathcal{T}_i}\mathcal{E}_T^P(u)\right)^{1/2},$$

with $\mathcal{E}_T^P(u)$ as in (3.10). The required bound now follows from (3.13). ∎

Remark. When the problem (2.1) is strongly convection-dominated, the solution is highly anisotropic locally. For this reason it is crucial that the error is bounded by appropriately weighted norms of directional derivatives of the solution, as in our error bound (3.14). We also observe that, if the partition is shape-regular, our error bound collapses to the isotropic error estimate (2.12).

We conclude the section with a remark on the extension of the above bound to the case when, in addition to diffusion and convection terms, the equation also contains a reaction term. Suppose therefore that $-\varepsilon\Delta u + \mathbf{a}\cdot\nabla u + cu = f$ in Ω, subject to $u = 0$ on $\partial\Omega$, with $2c - \mathrm{div}(\mathbf{a}) \le -2c_0$ in Ω, where c_0 is a positive constant. Arguing similarly as in the proof above, we then obtain

$$
\begin{aligned}
&\varepsilon|u-u_{\mathrm{RFB}}|_{1,\Omega}^2+c_0\|u-u_{\mathrm{RFB}}\|_{0,\Omega}^2\\
&\quad\le \varepsilon^{1/2}|u-u_{\mathrm{RFB}}|_{1,\Omega}\left(\sum_{T\in\mathcal{T}_h}\left(\varepsilon^{1/2}|u-P_Tu|_{1,T}+\varepsilon^{-1/2}\|u-P_Tu\|_{0,T}\right)^2\right)^{1/2}\\
&\qquad+\|c\|_{\mathrm{L}^\infty(\Omega)}\|u-u_{\mathrm{RFB}}\|_{0,\Omega}\left(\sum_{T\in\mathcal{T}_h}\|u-P_Tu\|_{0,T}^2\right)^{1/2}\\
&\quad\le \left(\varepsilon|u-u_{\mathrm{RFB}}|_{1,\Omega}^2+c_0\|u-u_{\mathrm{RFB}}\|_{0,\Omega}^2\right)^{1/2}\\
&\qquad\times\left(\sum_{T\in\mathcal{T}_h}\left(\varepsilon^{1/2}|u-P_Tu|_{1,T}+\varepsilon^{-1/2}\|u-P_Tu\|_{0,T}\right)^2+\frac{\|c\|_{\mathrm{L}^\infty(\Omega)}^2}{c_0}\|u-P_Tu\|_{0,T}^2\right)^{1/2}.
\end{aligned}
$$

The rest of the argument, based on bounding the second factor on the right-hand side in the final inequality, proceeds as in the proof of Theorem 3.8.

4. Affine partitions. We now discuss the case of partitions $\mathcal{T}_h$ consisting of affine-equivalent (triangular or quadrilateral) elements. As before, our assumptions on the partition are conformity and that (2.4) holds.

The following a priori error analysis is based on Lemma 3.2 and on the technique introduced by Formaggia and Perotto [14] (see also the references therein and Micheletti, Perotto, and Picasso [25]) to prove anisotropic error estimates for the interpolation error. More precisely, we will employ suitable scaling properties derived in [14] in terms of certain characteristic quantities of the affine transformation F_T. A limitation of the approach is that only an a priori error bound in terms of the H^2-seminorm can be obtained, so this analysis applies only in the case when $k=1$. An extension of the bounds presented here to the case when $k\ge 1$ can be carried out using the techniques developed in section 2.2 of the paper of Huang [20].

Let $F_T(\hat{\boldsymbol{x}})=M\hat{\boldsymbol{x}}+\boldsymbol{t}$ (we omit the dependence of M and t on T to simplify the notation). As the matrix M is invertible, it admits a unique *polar decomposition* $M=BZ$, where B is symmetric and positive definite and Z is orthonormal.

Further, B is factorized as $B=R^{\mathrm{T}}\Lambda R$, where Λ is diagonal with positive decreasing entries (the eigenvalues of B) and R is orthonormal (with rows which are the eigenvectors of B). Hence,

$$
\Lambda=\begin{bmatrix}\lambda_1 & 0\\ 0 & \lambda_2\end{bmatrix},\qquad R=\begin{bmatrix}\boldsymbol{r}_1^{\mathrm{T}}\\ \boldsymbol{r}_2^{\mathrm{T}}\end{bmatrix},
$$

where $\lambda_1\ge\lambda_2$ and $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ are the eigenvalues and eigenvectors of B, respectively. The above decomposition corresponds to the singular value decomposition $M=R^{\mathrm{T}}\Lambda Q$, with $Q=RZ$: The reference element $\widehat{T}$ is rotated using Q, stretched by Λ, and then rotated again by R^{T}. The translation $\boldsymbol{t}$ finally gives the correct location of T. The eigenvalues λ_1 and λ_2 of Λ thus give the element dimensions in a rotated orthogonal frame and hence are used to replace h_1 and h_2 from the previous section as the characteristic dimensions of the element T.

With this new notation, we get the following scaling rules, which are the counterparts of (3.1) and (3.2):

$$
\|v\|_{0,T}^2=\lambda_1\lambda_2\|\hat{v}\|_{0,\widehat{T}}^2, \tag{4.1}
$$

$$
|v|_{1,T}^2\le\frac{\lambda_1}{\lambda_2}|\hat{v}|_{1,\widehat{T}}^2. \tag{4.2}
$$

The equality (4.1) is elementary, while (4.2) is proved in [14] as Lemma 2.2.

To scale back from the reference element we shall use the following identity which is Lemma 2.2 in [25] (see also the proof of Lemmas 2.1 and 2.2 in [14]):

$$|\hat{v}|^2_{2,\widehat{T}} = \frac{\lambda_1^3}{\lambda_2} L_{11} v + \frac{\lambda_2^3}{\lambda_1} L_{22} v + 2\lambda_1\lambda_2 L_{12} v, \tag{4.3}$$

where

$$L_{ij} v := \int_T \left(\boldsymbol{r}_i^{\mathrm{T}} H(v) \boldsymbol{r}_j \right)^2 \mathrm{d}\boldsymbol{x}, \qquad \text{with } i,j = 1,2, \tag{4.4}$$

and $H(v)$ is the Hessian matrix associated with the function v; that is,

$$H(v) := \begin{bmatrix} \frac{\partial^2 v}{\partial x_1^2} & \frac{\partial^2 v}{\partial x_1 \partial x_2} \\ \frac{\partial^2 v}{\partial x_1 \partial x_2} & \frac{\partial^2 v}{\partial x_2^2} \end{bmatrix}.$$

THEOREM 4.1. *Let $u \in V$ be the solution of (2.2) and $u_{\mathrm{RFB}} \in V_{\mathrm{RFB}}$ the RFB solution defined by (2.8). Consider a conforming affine-equivalent partition $\mathcal{T}_h$ assuming that there exists a constant $c \in (0,1]$ such that, for every $T \in \mathcal{T}_h$, $\varepsilon \le c\lambda_2$, where $\lambda_1 \ge \lambda_2$ are the characteristic dimensions of T defined above.*

If $u \in \mathrm{H}^1_0(\Omega) \cap \mathrm{H}^2(\Omega)$, then there exists a positive constant C, independent of the mesh dimensions and of ε, such that

$$\varepsilon^{1/2} |u - u_{\mathrm{RFB}}|_{1,\Omega} \le C \left(\sum_{T \in \mathcal{T}_h} \left(\frac{\lambda_1^4}{\lambda_2} L_{11} u + \lambda_2^3 L_{22} u + 2\lambda_1^2 \lambda_2 L_{12} u \right) \right)^{1/2}, \tag{4.5}$$

where the terms L_{ij}, $i,j = 1,2$, are defined elementwise as in (4.4) in terms of the Hessian of the function u.

Proof. Let $T \in \mathcal{T}_h$. As in the previous section, we need to bound the quantity given by (3.10); that is,

$$\mathcal{E}_T^I(v) = \varepsilon |v - P_T v|^2_{1,T} + \varepsilon^{-1} \|v - P_T v\|^2_{0,T},$$

where $v \in \mathrm{H}^1(T)$. As before, we start by scaling $\mathcal{E}_T^I(v)$ to the reference element $\widehat{T}$. Using (4.1) and (4.2) we get

$$\begin{aligned} \mathcal{E}_T^I(v) &\le \varepsilon \frac{\lambda_1}{\lambda_2} |\hat{v} - P_{\widehat{T}} \hat{v}|^2_{1,\widehat{T}} + \varepsilon^{-1} \lambda_1 \lambda_2 \|\hat{v} - P_{\widehat{T}} \hat{v}\|^2_{0,\widehat{T}} \\ &= \lambda_1 \left(\frac{\varepsilon}{\lambda_2} |\hat{v} - P_{\widehat{T}} \hat{v}|^2_{1,\widehat{T}} + \left(\frac{\varepsilon}{\lambda_2} \right)^{-1} \|\hat{v} - P_{\widehat{T}} \hat{v}\|^2_{0,\widehat{T}} \right). \end{aligned} \tag{4.6}$$

We then apply Lemma 3.2, this time with $\hat{w}_0 = (\hat{v} - \hat{\pi}_1 \hat{v})_{|\partial\widehat{T}}$, where $\hat{\pi}_1$ is the standard linear Lagrange interpolant (that is, $\hat{\pi}_k$, with $k = 1$) defined on the reference triangle $\widehat{T}$, and with $t = \varepsilon/\lambda_2$. In this way we get

$$\mathcal{E}_T^I(v) \le C \left(\varepsilon \frac{\lambda_1}{\lambda_2} |\hat{v} - \hat{\pi}_1 \hat{v}|^2_{1/2,\partial\widehat{T}} + \lambda_1 \|\hat{v} - \hat{\pi}_1 \hat{v}\|^2_{0,\partial\widehat{T}} \right).$$

Instead of scaling back to the boundary of the element T as was done previously, we now proceed by applying the trace inequality (3.4) and the standard Lagrange

interpolation error bounds on $\widehat{T}$ (see Ciarlet [13]). Since $\lambda_2 \leq \lambda_1$ and $\varepsilon \leq c\lambda_2$, with $c \in (0,1]$, we get

$$\begin{aligned}\mathcal{E}_T^I(v) &\leq C\left(\varepsilon\frac{\lambda_1}{\lambda_2}+\lambda_1\right)\|\hat{v}-\hat{\pi}_1\hat{v}\|_{1,\widehat{T}}^2\\ &\leq C\lambda_1|\hat{v}|_{2,\widehat{T}}^2\\ &\leq C\left(\frac{\lambda_1^4}{\lambda_2}L_{11}v+\lambda_2^3L_{22}v+2\lambda_1^2\lambda_2L_{12}v\right),\end{aligned}\tag{4.7}$$

the last bound being a consequence of (4.3). The desired error bound now follows by repeating the steps in the proof of Theorem 3.8. ∎

If the partition $\mathcal{T}_h$ is axiparallel, then $\lambda_i = h_i/c_i$, with h_i and c_i, $i = 1, 2$, being the dimensions along the coordinate axes of T and $\widehat{T}$, respectively. In this case Theorem 4.1 collapses to the a priori error bound (3.14), with $r = 1$.

5. Numerical examples. As discussed in section 1, a fully discrete RFB method is obtained after approximating the bubble space. In the following experiment, the local bubble problem on each element is solved using the standard Galerkin finite element method (FEM) on an 8×8 Shishkin partition. This is a piecewise uniform mesh with half of the nodes in each coordinate direction lying in the boundary-layer region of the element; see [24] and references therein. This choice abundantly ensures that the subgrid discretization error is of higher order than the RFB error controlled by our error analysis. In fact, in the case of P_1 shape-regular finite elements, it has been proved by Brezzi and Marini [8] that a subgrid consisting of a single internal node placed inside the boundary layer of the bubble problem is sufficient; see also [4]. This is the fully discrete method that we suggest for practical implementations.

Another possibility, exploited in further experiments presented later on, is to discretize the convection field with piecewise constants and then approximate the solution of each local bubble problem by the solution of the corresponding reduced (hyperbolic) elemental problem [9]. This procedure is computationally inexpensive, as it amounts to the calculation of the volume of a pyramid on each element. Moreover, when the problem is convection-dominated, such an approximation does not compromise the accuracy of the method (a choice that is optimal in all regimes is the link-cutting bubble proposed in [4] for one-dimensional problems). Indeed, the discretization of the bubble functions need not be particularly accurate as long as the elemental average

$$\frac{\int_T b_T \,\mathrm{d}\boldsymbol{x}}{|T|}$$

of the bubble b_T has been sufficiently accurately approximated; the reason, as is shown later on in this paper (see also [5]), is that only the elemental averages of the bubbles enter into the fully discrete method. The behavior of the above term on shape-regular partitions, as a function of the mesh Péclet number $\mathtt{Pe}_T = h_T|\boldsymbol{a}|/\varepsilon$, is analyzed in [5], where it is also shown that the average of the solution of the reduced bubble problem behaves similarly in the convection-dominated regime to the average of the exact bubble b_T. Lemma 6.1 below extends the analysis from [5] to anisotropic partitions, thus suggesting that this simple recipe for full discretization is still viable on anisotropic partitions.

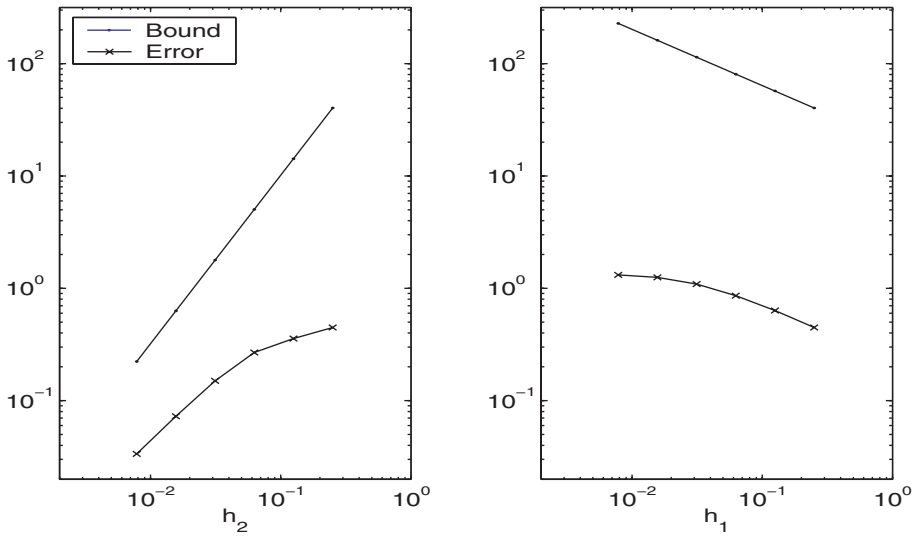


FIG. 5.1. *$\varepsilon^{1/2}$-weighted H^1-seminorm error and error bound under (the correct) h_2-refinement (left) and (the incorrect) h_1-refinement (right); $\varepsilon = 10^{-2}$. In both cases, we start from the 4×4 uniform square mesh.*

We consider the following simple boundary-value problem

$$(5.1)\qquad \begin{cases} -\varepsilon\Delta u + u_{x_2} = 0 & \text{in } \Omega = (0,1)^2, \\ u(x_1,0) = 0; \quad u(x_1,1) = 1, & x_1 \in [0,1], \\ u_{x_1} = 0 & \text{on } \Gamma_{\mathrm{N}} = (\{0\}\times(0,1)) \cup (\{1\}\times(0,1)), \end{cases}$$

whose solution is given by

$$u(x_1,x_2) = \frac{\mathrm{e}^{x_2/\varepsilon} - 1}{\mathrm{e}^{1/\varepsilon} - 1}.$$

We consider discretizations of this problem with respect to axiparallel uniform rectangular grids of dimensions h_1 and h_2 in the respective coordinate directions. For this problem the error bound (3.14) reduces to

$$\varepsilon^{1/2}|u - u_{\mathrm{RFB}}|_{1,\Omega} \le C \begin{cases} h_2^3 ||\partial_{x_2}^2 u||_{0,\Omega}^2 & \text{if } h_2 \le h_1, \\ \dfrac{h_2^4}{h_1} ||\partial_{x_2}^2 u||_{0,\Omega}^2 & \text{if } h_2 > h_1. \end{cases}$$

We verify the validity of the bound by performing the following tests. Starting from the uniform 4×4 mesh, we either

- fix h_1 while halving h_2 (*correct refinement*) or
- fix h_2 while halving h_1 (*incorrect refinement*).

The relevant energy norm errors and error bounds are shown in the log-log plot in Figure 5.1 (left-hand panel) for $\varepsilon = 10^{-2}$.

Performing the correct refinement is, of course, not too different from solving the related sequence of one-dimensional problems. The similarity of the numerical solution of the two-dimensional problem to the numerical solution of the related one-dimensional problem is lost when the incorrect refinement is performed (notice that

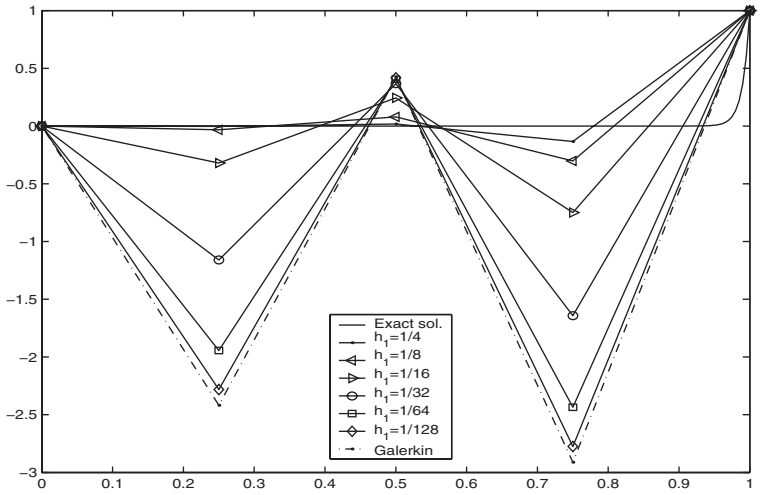


FIG. 5.2. *Profile of the solution along $x_1 = 1/2$ under h_1-refinement (as in the right-hand panel in Figure* 5.1*), while $h_2 = 1/4$. The lowest profile represents the piecewise $\mathcal{Q}_1$ standard Galerkin FEM solution computed on a uniform 4×4 mesh. The exact solution is also plotted for comparison.*

this does not happen when applying the standard Galerkin method with linear elements). As predicted by the error bound, the accuracy of the solution actually deteriorates under the incorrect refinement; see the log-log plot in Figure 5.1 (right-hand panel). This is due to the peculiar definition of the RFB finite element space. Mesh refinement corresponds to a relative impoverishment of the bubble subspace and an enrichment of the piecewise polynomial subspace. If the latter enrichment, as is the case with our incorrect refinement, is ineffective, then the overall approximation properties of V_{RFB} will be worse than on a coarser mesh. The detailed error analysis of the RFB method on shape-regular partitions presented in our recent work [12] aims to clarify the approximation properties of the method in the preasymptotic regime when $\varepsilon \leq ch$. In particular, in [12], we relate the phenomenon just observed to the inadequacy of V_h^k to capture the exponential behavior of the solution along element edges contained in the boundary layer.

In the limit of $h_1 \to 0$, the solution becomes constant along x_1. That is, it tends to the piecewise $\mathcal{Q}_1$ standard Galerkin solution, which is unaffected by the reduction of h_1; see Figure 5.2. Asymptotically, in the case of the incorrect refinement (with $h_1 \to 0$), the error is of order $O(1)$ (cf. Figure 5.1 (right)). In other words, since the bubble part of the solution is forced to tend to zero as $h_1 \to 0$, its stabilizing effect is diminished until, in the limit, it vanishes and the RFB method collapses to the standard Galerkin FEM. This fact shows that the stabilization properties of stabilized FEMs are affected by the anisotropy of the partition.

The use of anisotropic partitions for the solution of highly convection-dominated problems can become mandatory if resolution of thin layers in the solution is paramount. Let us consider, for example, the boundary-value problem

$$\begin{cases} -\varepsilon \Delta u + (2,1)^{\mathrm{T}} \cdot \nabla u = 0 & \text{in } \Omega = (0,1)^2, \\ u(x_1, 0) = u(1, x_2) = 0, & x_1, x_2 \in (0,1), \\ u(x_1, 1) = u(0, x_2) = 1, & x_1, x_2 \in [0,1]. \end{cases} \tag{5.2}$$

The solution of (5.2) exhibits an internal layer emanating from the origin of the coor-

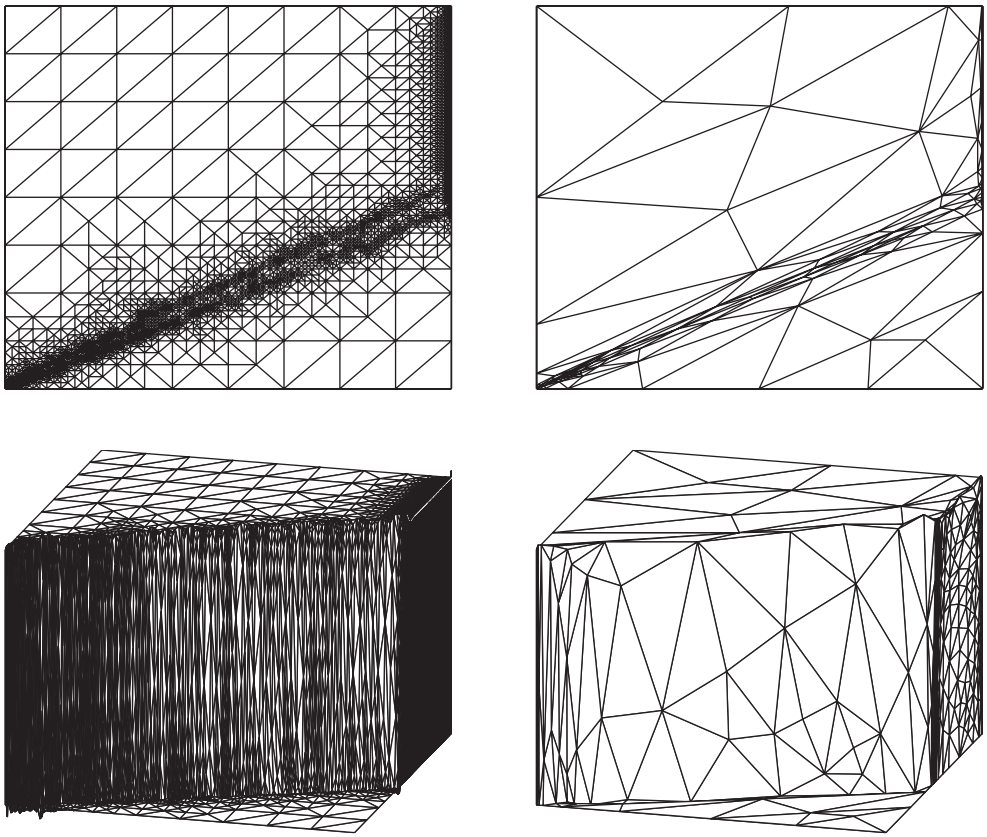


FIG. 5.3. *The solution of* (5.2) *with* $\varepsilon = 10^{-4}$ *on* ad hoc-*refined triangulations. Left: Shape-regular mesh (*23256 *elements,* 12693 *nodes) and the corresponding solution. Right: Anisotropic mesh (*478 *elements,* 263 *nodes) and the corresponding solution.*

dinate system and a boundary layer situated along $x_1 = 1$. The RFB approximation of (5.2) is shown in Figure 5.3. The bubble solution is approximated by the solution of the related reduced (hyperbolic) elemental problem [9]. We compute the RFB solution using, respectively, a shape-regular triangulation (left-hand panels in the figure) and an anisotropic triangulation (right-hand panels in the figure). The anisotropic triangulation has been generated by Picasso [26], by applying a ZZ-type error indicator for the gradient error to the classical stabilized Galerkin least-squares (GLS) method, until the stopping criterion ZZ-indicator $\leq 1/4$ was satisfied in all elements. The triangulation was then used to compute the RFB solution shown in the bottom right-hand panel of Figure 5.3. The computation on the shape-regular triangulation was performed by applying the residual-based L^2-error indicator proposed in [10] for the RFB method. For the sake of consistency, the adaptation was stopped when the error indicator fell below 1/4 in all elements. The RFB solution computed on the anisotropic triangulation is clearly superior, as the triangulation consists of only 263 nodes instead of the 12693 nodes, with comparable accuracy, in the case of the shape-regular partition.

6. Tuning of the SD parameter. The RFB method is closely related to classical stabilized finite element methods (streamline upwind Petrov–Galerkin (SUPG),

GLS, etc.). For instance, in the case of piecewise constant coefficients and linear finite elements, RFB is equivalent to SUPG and GLS (the latter methods coincide in this case with what Johnson, Nävert, and Pitkäranta [22] refer to as the *streamline-diffusion finite element method* (SDFEM)). Here we exploit this identification to obtain a theoretically justified value of the user-selected stabilization parameter in stabilized finite element methods.

We consider the RFB method (2.8), assuming that $\mathcal{T}_h$ consists of triangles, and fix $k = 1$. In this case, $V_{\mathrm{RFB}} = V_h^1 \oplus B_h$, where V_h^1 is the space of linear finite elements.

Let us also assume that $\boldsymbol{a}$ and f are constant on every element of $\mathcal{T}_h$. Then the right-hand side of (2.11) is constant, and the bubble part of the RFB solution is given locally on T by $u_b|_T = (f - Lu_h)|_T b_T$, where $b_T \in \mathrm{H}_0^1(T)$ satisfies

$$-\varepsilon \Delta b_T + \boldsymbol{a} \cdot \nabla b_T = 1. \tag{6.1}$$

Substituting u_b into (2.9) it follows that $u_h \in V_h^1$ is the solution of

$$\mathcal{L}(u_h, v_h) + \sum_{T \in \mathcal{T}_h} \frac{\int_T b_T \,\mathrm{d}\boldsymbol{x}}{|T|} (\boldsymbol{a} \cdot \nabla u_h - f, \boldsymbol{a} \cdot \nabla v_h)_T = (f, v_h) \qquad \forall v_h \in V_h^1. \tag{6.2}$$

The formulation (6.2) coincides with the SDFEM with the particular choice of the SD parameter given by

$$\tau_b := \frac{\int_T b_T \,\mathrm{d}\boldsymbol{x}}{|T|}. \tag{6.3}$$

Thus, as anticipated, the RFB method and the SDFEM are, in this case, equivalent. This well-known fact was first observed by Brezzi and Russo [9].

A numerical method is obtained from the RFB formulation by considering (6.2) where the quantity τ_b has been suitably approximated (examples are given in [9, 15, 6, 8, 4, 30, 11]). As discussed in [5] in the case of shape-regular triangulations, the crucial property is that the approximated value of τ_b scales as τ_b with respect to the mesh size and the coefficients ε and $\mathbf{b}$.

Specifically, let h_a indicate the length of the longest segment parallel to $\boldsymbol{a}$ contained in T. On shape-regular partitions, i.e., assuming that the minimal angle of T is bounded below by a fixed positive constant, we know from [5] that

$$C \frac{h_T}{|\boldsymbol{a}|} \min\left\{ \frac{h|\boldsymbol{a}|}{\varepsilon}, 1 \right\} \le \tau_b \le \frac{h_a}{|\boldsymbol{a}|}. \tag{6.4}$$

In practice, $\tau_b \sim \frac{h_T}{|\boldsymbol{a}|} \min\{\frac{h|\boldsymbol{a}|}{\varepsilon}, 1\}$, which is qualitatively the value of the SD parameter suggested by the a priori error analysis of the SDFEM (see, e.g., [28]).

The situation is less clear when considering anisotropic elements. Attempts have been made to derive the optimal behavior of the SD parameter through a priori analysis; see, e.g., [3, 23, 25]. The outcome of the investigations in these papers is that the stabilization parameter should depend on the smaller dimension of the element.

For instance, assume that T is a right-angled triangle of dimensions h_1, h_2, and let $h_{\min} = \min\{h_1, h_2\}$. Then, according to [25], we should choose the SD parameter as

$$\tau_{\mathrm{sd}} := C \frac{h_{\min}}{2|\boldsymbol{a}|} \min\left\{ \frac{h_{\min}|\boldsymbol{a}|}{6\varepsilon}, 1 \right\}. \tag{6.5}$$

This choice seems less favorable when the mesh is not aligned with the solution (as in the incorrect refinement in our example above). We notice that in this case the a priori analysis does not predict convergence anyway.

By appropriately modifying the argument employed in [5] to derive (6.4), we shall now obtain a new lower bound for τ_b that takes the two characteristic dimensions of T into account. This result is then used to provide a new rule for selecting the SD parameter.

LEMMA 6.1. *Suppose that T is a right-angled triangle, oriented along the coordinate axes, of dimensions h_1, h_2; then the quantity τ_b given by* (6.3)*, where b_T solves* (6.1)*, satisfies*

$$C\frac{h_a}{|\boldsymbol{a}|}\min\{\mathtt{Pe}_T,1\}\le\tau_b\le\frac{h_a}{|\boldsymbol{a}|}, \tag{6.6}$$

with $C=1/45$ and with the following definition of the element Péclet number:

$$\mathtt{Pe}_T:=h_{\min}^2\frac{|\boldsymbol{a}|}{8\varepsilon h_a}. \tag{6.7}$$

Proof. The upper bound is already given in (6.4). Assume that $h_2<h_1$, so that $h_{\min}=h_2$. To prove the lower bound, we map T into the right-angled triangle $\widehat{T}$ with its two orthogonal edges of length h_ah_1/h_2^2 and h_a/h_2 aligned with the positive semiaxes of the coordinate system $(\hat{x}_1,\hat{x}_2)$. The image $\hat{b}$ of $b_T\in\mathrm{H}_0^1(\widehat{T})$ satisfies

$$-\varepsilon\frac{h_a}{h_2^2}\Delta\hat{b}+\boldsymbol{a}\cdot\nabla\hat{b}=\frac{h_2^2}{h_a}\quad\text{in }\widehat{T},$$

and we have

$$\tau_b=\frac{2h_2^3}{h_1h_a^2}\int_{\widehat{T}}\hat{b}\,\mathrm{d}\hat{\boldsymbol{x}}. \tag{6.8}$$

To bound the integral in (6.8) we proceed as in [5]. We let $\hat{\lambda}_1$, $\hat{\lambda}_2$, and $\hat{\lambda}_3$ be the barycentric coordinates on $\widehat{T}$, define $\hat{b}_3:=\hat{\lambda}_1\hat{\lambda}_2\hat{\lambda}_3$, and note that

$$\int_{\widehat{T}}\hat{b}_3\,\mathrm{d}\hat{\boldsymbol{x}}=\frac{h_a^2h_1}{120h_2^3}. \tag{6.9}$$

Since $h_2<h_1$, we have

$$M_\Delta:=\frac{1}{8}\max_{\widehat{T}}|\Delta\hat{b}_3|=\frac{1}{4}\frac{h_2^5}{h_a^3h_1}\max_{\widehat{T}}\left(\frac{\hat{x}_1}{h_2}+\frac{\hat{x}_2}{h_1}\right)=\frac{1}{4}\frac{h_2^2}{h_a^2}, \tag{6.10}$$

the maximum being attained at the vertex $(h_a h_1/h_2^2, 0)$, and

$$\begin{aligned} M_g &:= \frac{1}{|\boldsymbol{a}|}\max_{\widehat{T}}|\boldsymbol{a}\cdot\nabla\hat{b}_3| \\ &= \frac{h_2^3}{|\boldsymbol{a}|h_a^2h_1}\max_{\widehat{T}}\left|a_1\left(\hat{x}_2-2\frac{h_2^2}{h_ah_1}\hat{x}_1\hat{x}_2-\frac{h_2}{h_a}\hat{x}_2^2\right)\right. \\ &\quad \left.+a_2\left(\hat{x}_1-\frac{h_2^2}{h_ah_1}\hat{x}_1^2-2\frac{h_2}{h_a}\hat{x}_1\hat{x}_2\right)\right| \\ &\le \frac{h_2^3}{|\boldsymbol{a}|h_a^2h_1}\left(|a_1|\max_{\widehat{T}}\left|\hat{x}_2-2\frac{h_2^2}{h_ah_1}\hat{x}_1\hat{x}_2-\frac{h_2}{h_a}\hat{x}_2^2\right|\right. \\ &\quad \left.+|a_2|\max_{\widehat{T}}\left|\hat{x}_1-\frac{h_2^2}{h_ah_1}\hat{x}_1^2-2\frac{h_2}{h_a}\hat{x}_1\hat{x}_2\right|\right) \\ &= \frac{h_2^3}{|\boldsymbol{a}|h_a^2h^1}\left(|a_1|\frac{h_a}{4h_2}+|a_2|\frac{h_ah_1}{4h_2^2}\right) \\ &= \frac{1}{4|\boldsymbol{a}|}\left(|a_1|\frac{h_2^2}{h_ah_1}+|a_2|\frac{h_2}{h_a}\right), \end{aligned} \tag{6.11}$$

both maxima being attained at the midpoint of the hypotenuse. We note that if $\operatorname{sign}(a_1)=\operatorname{sign}(a_2)$, the above bound reduces to an equality.

We now define

$$\gamma := \frac{1}{M_\Delta+M_g}\min\left\{\frac{h_2^2}{8\varepsilon h_a},\frac{1}{|\boldsymbol{a}|}\right\}, \qquad \hat{w}:=\gamma\hat{b}_3, \qquad \hat{v}:=\frac{h_a}{h_2^2}\hat{b},$$

and introduce the differential operator

$$\widehat{L}\varphi := -\varepsilon\frac{h_a}{h_2^2}\Delta\varphi+\boldsymbol{a}\cdot\nabla\varphi.$$

By the definition of γ, $\hat{w}$, M_Δ, and M_g, we have

$$|\widehat{L}\hat{w}|\le\gamma\left(\varepsilon\frac{h_a}{h_2^2}M_\Delta+|\boldsymbol{a}|M_g\right)\le 1.$$

Thus, by the definition of $\hat{v}$, we have

$$\widehat{L}(\hat{v}-\hat{w})=\frac{h_a}{h_2^2}\widehat{L}\hat{b}-\widehat{L}\hat{w}=1-\widehat{L}\hat{w}\ge 0,$$

and, since both $\hat{v}$ and $\hat{w}$ vanish on $\partial\widehat{T}$, using the maximum principle, we conclude that $\hat{v}\ge\hat{w}$ in $\widehat{T}$. We are now ready to bound τ_b. Recalling (6.8) and (6.9), we have

$$\tau_b=\frac{2h_2^5}{h_a^3h_1}\int_{\widehat{T}}\hat{v}\,\mathrm{d}\hat{\boldsymbol{x}}\ge\frac{2h_2^5}{h_a^3h_1}\gamma\int_{\widehat{T}}\hat{b}_3\,\mathrm{d}\hat{\boldsymbol{x}}=\frac{h_2^2}{60h_a}\gamma.$$

Further, using the definition of γ, and inserting (6.10) and (6.11), we have

$$\tau_b\ge\frac{1}{15\left(\frac{|\boldsymbol{a}|}{h_a}+\frac{|a_1|}{h_1}+\frac{|a_2|}{h_2}\right)}\min\left\{\frac{|\boldsymbol{a}|h_2^2}{8\varepsilon h_a},1\right\}.$$

We distinguish between the following two cases.

- If $\mathrm{sign}(a_1) = \mathrm{sign}(a_2)$, then h_a is the length of the line segment oriented with $\boldsymbol{a}$ which joins the hypotenuse of T with the opposite vertex. Thus,

$$h_a = \sqrt{\frac{h_2^2}{\left(\frac{a_2}{a_1} + \frac{h_2}{h_1}\right)^2}\left(1 + \frac{a_2^2}{a_1^2}\right)} = \frac{|\boldsymbol{a}|}{\frac{|a_1|}{h_1} + \frac{|a_2|}{h_2}}.$$

  It follows that $|a_1|/h_1 + |a_2|/h_2 = |\boldsymbol{a}|/h_a$.
- If $\mathrm{sign}(a_1) \neq \mathrm{sign}(a_2)$ and $|a_2|/h_2 > |a_1|/h_1$, then h_a is the length of the line segment oriented with $\boldsymbol{a}$ which joins the edge of T parallel to the x_1-axis with the opposite vertex. Thus,

$$h_a = \sqrt{h_2^2 + \frac{a_1^2}{a_2^2}h_2^2} = \frac{h_2|\boldsymbol{a}|}{|a_2|},$$

  and so $|a_2|/h_2 = |\boldsymbol{a}|/h_a$. Similarly, if $|a_2|/h_2 > |a_1|/h_1$, then $|a_1|/h_1 = |\boldsymbol{a}|/h_a$.

It follows that

$$\frac{|\boldsymbol{a}|}{h_a} + \frac{|a_1|}{h_1} + \frac{|a_2|}{h_2} \leq C\frac{|\boldsymbol{a}|}{h_a},$$

with $C = 2$ or 3, depending on the cases listed above, respectively.

Since the above argument can be repeated in the case $h_1 \leq h_2$ by interchanging the role of h_1 and h_2, we conclude that the bound (6.6) holds with $C = 1/45$. ∎

To verify the bound obtained, we compare the behavior of

$$\tau_a := C\frac{h_a}{|\boldsymbol{a}|}\min\left\{\mathtt{Pe}_T, 1\right\},$$

with that of τ_b with respect to the dimensions of T. We let $h_1 = 1$ while halving h_2 starting from $h_2 = 1$. We do this twice in succession, with $\boldsymbol{a} = (1, 0)$ and then with $\boldsymbol{a} = (0, 1)$. The results are shown in Figure 6.1 (τ_b is calculated by solving (6.1) very accurately). The superimposition of the graphs is obtained by renormalizing τ_a (the factor is always around 3) so that its first values coincide with that of τ_b. As we can see in Figure 6.1, τ_a and τ_b are very close to each other.

Figure 6.1 also reports the results obtained with the choice τ_{sd} given by (6.5), which was proposed as a SD parameter in [3, 23, 25]. We notice that the two choices τ_a and τ_{sd} have different turning points, particularly when $\boldsymbol{a}$ is aligned with the longest edge of T. This is due to the fact that our definition of the element Péclet number depends not only on the magnitude of the convective field, but also on its direction. We believe that this should indeed be the case when anisotropic partitions are considered, and hence we propose τ_a as the appropriate SD parameter. The definition of τ_a easily extends to a general element by substituting h_1 and h_2 by the characteristic dimensions λ_1 and λ_2.

We assess experimentally our new choice of the SD parameter τ_a by comparing its performance with that of τ_{sd} on some model problems. From the discussion above we know that the two choices τ_a and τ_{sd} differ the most when the stretching of the element is aligned with the direction of convection. We must also take into account, though, that the magnitude of the SDFEM stabilization term depends on the alignment of the convection with the gradient of the solution; cf. (6.2); see also section 3 in [20]. We therefore consider two test problems: (5.1), whose solution exhibits a boundary layer,

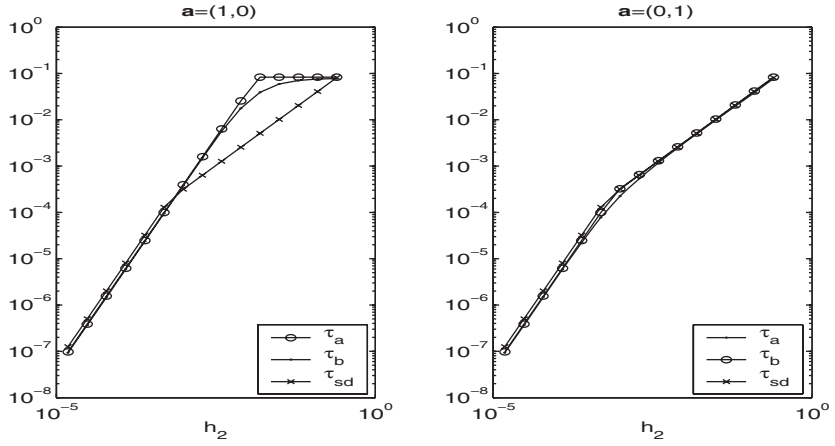


FIG. 6.1. *Comparisons of τ_b with τ_a and $\tau_{\rm sd}$ on a rectangle of dimensions 1 and h_2 for $\varepsilon = 10^{-4}$.*

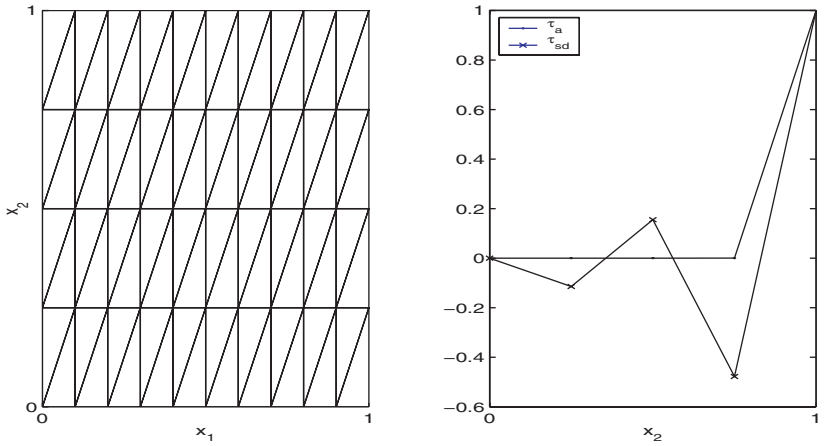


FIG. 6.2. *Mesh and SDFEM solution profile along $x_1 = 1/2$ for the model problem* (5.1) *with $\varepsilon = 10^{-4}$.*

and a modification of (5.2) obtained by imposing a Neumann boundary condition on the outflow boundary, so that the solution of the problem contains an internal layer. In all tests we solved the problem on a slightly stretched uniform partition of aspect ratio 4/10.

We start with (5.1). We compare the two different choices of the SD parameter τ_a and $\tau_{\rm sd}$ by solving the model problem (5.1) with $\varepsilon = 10^{-4}$ by means of the SDFEM. In both cases, the constant factors C in the definitions of the two parameters are tuned by solving the problem on a uniform partition. We apply the SDFEM on the partition depicted in the left-hand panel of Figure 6.2. The solution profile at $x_1 = 1/2$ is shown in the right-hand panel of Figure 6.2. While the solution obtained using τ_a correctly reproduces the exact solution, the one obtained using $\tau_{\rm sd}$ is corrupted by oscillations, indicating that the stabilization parameter $\tau_{\rm sd}$ is too small. The difference is due to the fact that, while $\tau_{\rm sd}$ always depends on $h_{\min}$, the parameter τ_a is linked to $h_{\max}$ as long as $\mathtt{Pe}_T > 1$. Eventually, if the mesh is further stretched in the incorrect direction, the

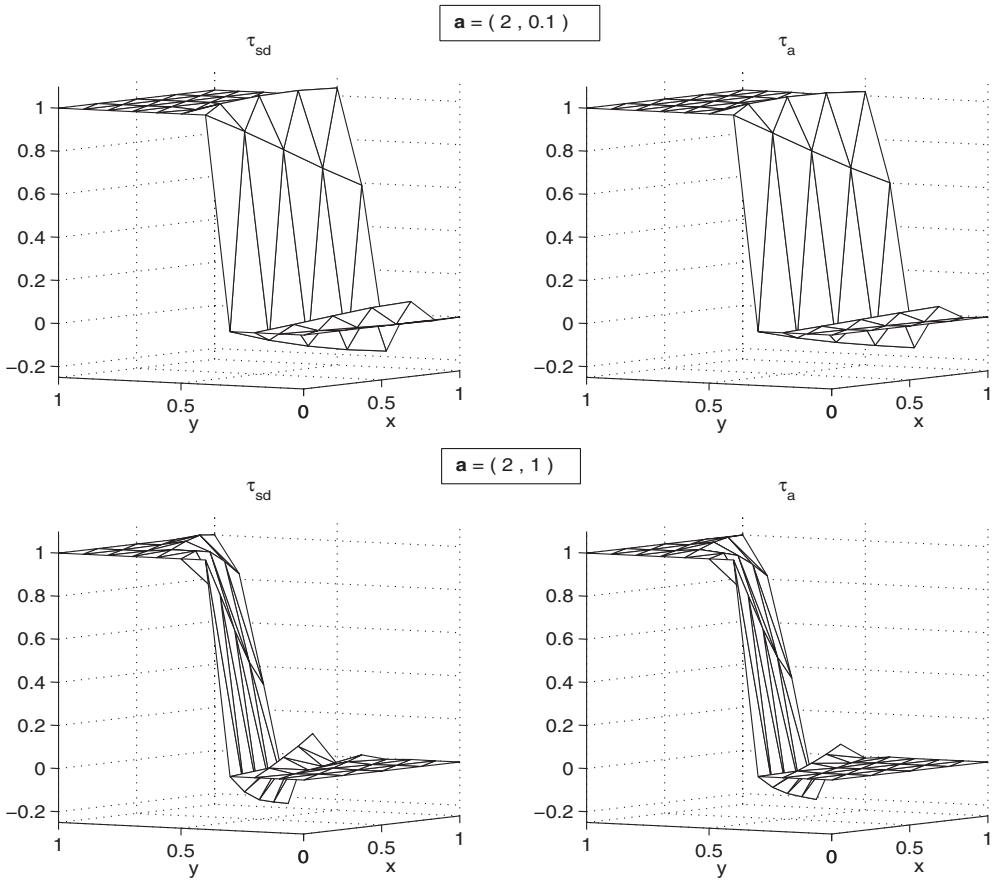


FIG. 6.3. *SDFEM solution of the model problem* (6.12) *with* $\varepsilon = 10^{-4}$.

use of τ_a will also lead to maximum-principle-violating oscillations in the numerical solution, but this happens for partitions with significantly higher aspect ratios than for $\tau_{\rm sd}$; for the present model problem, $\mathtt{Pe}_T > 1$ for $h_{\min} > 2^{5/2}10^{-2} \approx 0.05$, corresponding to an aspect ratio of 1/5. In conclusion, our choice will guarantee stability for any, *not too unreasonably* designed, partition, such as the one used in the experiment.

We finally consider the following boundary-value problem:

$$\begin{cases} -\varepsilon\Delta u + \boldsymbol{a}\cdot\nabla u = 0 & & \text{in } \Omega = (0,1)^2, \\ u(x_1,0) = 0; & u(x_1,1) = 1, & x_1 \in (0,1), \\ u(0,x_2) = \chi_{[1/3,1]}(x_2); & \frac{\partial u}{\partial x_1}(1,x_2) = 0, & x_2 \in [0,1], \end{cases} \tag{6.12}$$

which exhibits an internal layer emanating from the boundary-value discontinuity in $(0,1/3)$ in the direction of $\boldsymbol{a}$. We fix the partition to be a uniform 4×10 partition and test the different choices of the SD parameter as functions of the convection direction by setting $\boldsymbol{a} = (2,1)$ as in (5.2) and then $\boldsymbol{a} = (2,0.1)$, i.e., aligned with the partition. The SDFEM solutions are shown in Figure 6.3. The solutions obtained using τ_a are slightly less oscillatory, particularly in the case $\boldsymbol{a} = (2,1)$, where we observe differences in the solutions at the outflow up to a factor of 1.6. This latter fact may seem counterintuitive, as τ_a and $\tau_{\rm sd}$ differ the most in the case $\boldsymbol{a} = (2,0.1)$

when convection is aligned with the stretching of the partition, but the alignment improves the performance of the method and reduces the need for stabilization.

7. Conclusions. When a convection-diffusion problem is strongly convection-dominated, the solution is often highly anisotropic, exhibiting large gradients in specific directions. In this paper we have developed the a priori error analysis of the RFB method, in the energy norm, on anisotropic partitions. The error is bounded by appropriately weighted norms of directional derivatives of the solution, so as to respect the anisotropic nature of the solution to the problem. The error bound established is an extension of that obtained by Sangalli [29] for shape-regular partitions.

Anisotropy also has to be taken into account in the tuning of the parameters appearing in *streamline-diffusion*-type methods. We have used the stabilizing term derived from the RFB method to redefine the mesh Péclet number and proposed a new choice of the SD parameter which is suitable for use on anisotropic partitions. Our choice improves the choices of the SD parameter presented in previous works based on the a priori analysis of the SD method (cf. [3, 23, 25]).

Acknowledgment. We are grateful to Professor Marco Picasso (Ecole Polytechnique Fédérale de Lausanne) for supplying the anisotropic triangulation that was used to generate the right-hand panels in Figure 5.3.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Pure Appl. Math. (Amst.) 65, Academic Press, New York, 1975.
[2] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.
[3] T. Apel and G. Lube, *Anisotropic mesh refinement in stabilized Galerkin methods*, Numer. Math., 74 (1996), pp. 261–282.
[4] F. Brezzi, G. Hauke, L. D. Marini, and G. Sangalli, *Link-cutting bubbles for the stabilization of convection-diffusion-reaction problems*, Math. Models Methods Appl. Sci., 13 (2003), pp. 445–461.
[5] F. Brezzi, T. J. R. Hughes, L. D. Marini, A. Russo, and E. Süli, *A priori error analysis of residual-free bubbles for advection-diffusion problems*, SIAM J. Numer. Anal., 36 (1999), pp. 1933–1948.
[6] F. Brezzi, D. Marini, and A. Russo, *Applications of the pseudo residual-free bubbles to the stabilization of convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 51–63.
[7] F. Brezzi, D. Marini, and E. Süli, *Residual-free bubbles for advection-diffusion problems: The general error analysis*, Numer. Math., 85 (2000), pp. 31–47.
[8] F. Brezzi and L. D. Marini, *Augmented spaces, two-level methods, and stabilizing subgrids*, Internat. J. Numer. Methods Fluids, 40 (2002), pp. 31–46.
[9] F. Brezzi and A. Russo, *Choosing bubbles for advection-diffusion problems*, Math. Models Methods Appl. Sci., 4 (1994), pp. 571–587.
[10] A. Cangiani and E. Süli, *A-Posteriori Error Estimators and RFB*, Technical report NA-04/22, Computing Laboratory, 2004.
[11] A. Cangiani and E. Süli, *Enhanced residual-free bubble method for convection-diffusion problems*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 1307–1313.
[12] A. Cangiani and E. Süli, *Enhanced RFB method*, Numer. Math., 101 (2005), pp. 273–308.
[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[14] L. Formaggia and S. Perotto, *New anisotropic* a priori *error estimates*, Numer. Math., 89 (2001), pp. 641–667.
[15] L. P. Franca and A. Russo, *Deriving upwinding, mass lumping and selective reduced integration by residual-free bubbles*, Appl. Math. Lett., 9 (1996), pp. 83–88.
[16] E. Georgoulis, E. Hall, and P. Houston, *Discontinuous Galerkin methods for advection-diffusion-reaction problems on anisotopically refined meshes*, SIAM J. Sci. Comput., to appear.

[17] E. H. Georgoulis, *Discontinuous Galerkin Methods on Shape–Regular and Anisotropic Meshes*, D.Phil. thesis, University of Oxford, 2003.
[18] E. H. Georgoulis and E. Süli, *Optimal error estimates for the hp-version interior penalty discontinuous Galerkin finite element method*, IMA J. Numer. Anal., 25 (2005), pp. 205–220.
[19] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Classics Math., Springer-Verlag, Berlin, 2001.
[20] W. Huang, *Measuring mesh qualities and application to variational mesh adaptation*, SIAM J. Sci. Comput., 26 (2005), pp. 1643–1666.
[21] W. Huang, *Mathematical principles of anisotropic mesh adaptation*, Commun. Comput. Phys., 1 (2006), pp. 276–310.
[22] C. Johnson, U. Nävert, and J. Pitkäranta, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.
[23] G. Kunert, *Robust a posteriori error estimation for a singularly perturbed reaction-diffusion equation on anisotropic tetrahedral meshes*, Adv. Comput. Math., 15 (2001), pp. 237–259.
[24] N. Madden and M. Stynes, *Efficient generation of Shishkin meshes in solving convection-diffusion problems*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 565–576.
[25] S. Micheletti, S. Perotto, and M. Picasso, *Stabilized finite elements on anisotropic meshes: A priori error estimates for the advection-diffusion and the Stokes problems*, SIAM J. Numer. Anal., 41 (2003), pp. 1131–1162.
[26] M. Picasso, *An anisotropic error indicator based on Zienkiewicz-Zhu error estimator: Application to elliptic and parabolic problems*, SIAM J. Sci. Comput., 24 (2003), pp. 1328–1355.
[27] U. Risch, *Convergence analysis of the residual free bubble method for bilinear elements*, SIAM J. Numer. Anal., 39 (2001), pp. 1366–1379.
[28] H.-G. Roos, M. Stynes, and L. Tobiska, *Numerical Methods for Singularly Perturbed Differential Equations*, Springer-Verlag, Berlin, 1996.
[29] G. Sangalli, *Global and local error analysis for the residual-free bubbles method applied to advection–dominated problems*, SIAM J. Numer. Anal., 38 (2000), pp. 1496–1522.
[30] G. Sangalli, *A discontinuous residual-free bubble method for advection-diffusion problems*, J. Engrg. Math., 49 (2004), pp. 149–162.
[31] C. Schwab, *p- and hp-Finite Element Methods*, Oxford University Press, New York, 1998.
[32] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, second ed., Johann Ambrosius Barth, Heidelberg, 1995.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1679–1700

FINITE ELEMENT DISCRETIZATION ERROR ANALYSIS OF A SURFACE TENSION FORCE IN TWO-PHASE INCOMPRESSIBLE FLOWS*

SVEN GROSS† AND ARNOLD REUSKEN†

Abstract. We consider a standard model for a stationary two-phase incompressible flow with surface tension. In the variational formulation of the model a linear functional which describes the surface tension force occurs. This functional depends on the location and the curvature of the interface. In a finite element discretization method the functional has to be approximated. For an approximation method based on a Laplace–Beltrami representation of the curvature we derive sharp bounds for the approximation error. A new modified approximation method with a significantly smaller error is introduced.

Key words. two-phase flow, continuum surface force technique, interface, Laplace–Beltrami operator, finite elements

AMS subject classifications. 65M60, 65N15, 65N30, 76D45, 76T99

DOI. 10.1137/060667530

1. Introduction.

Let $\Omega \subset \mathbb{R}^3$ be a polyhedral domain that contains a flow of two different immiscible incompressible newtonian phases (fluid-fluid or fluid-gas). At the interface between the two phases there are surface tension forces that are significant and cannot be neglected. An example is a (rising) liquid drop contained in a surrounding fluid. The standard model to describe such a flow problem consists of instationary Navier–Stokes equations with certain coupling conditions at the interface which describe the effect of surface tension. In this paper we analyze errors that are due to the discretization of the surface tension force that occurs in the continuous model. To simplify the presentation and the analysis we assume a *stationary* flow.

The domains which contain the phases are denoted by Ω_1 and Ω_2 with $\overline{\Omega}_1 \cup \overline{\Omega}_2 = \overline{\Omega}$ and $\partial\Omega_1 \cap \partial\Omega = \emptyset$. The interface between the two phases $(\partial\Omega_1 \cap \partial\Omega_2)$ is denoted by Γ. To model the forces at the interface we make the standard assumption that the surface tension balances the jump of the normal stress on the interface; i.e., we have an interface condition

$$[\boldsymbol{\sigma}\mathbf{n}]_\Gamma = \tau\mathcal{K}\mathbf{n},$$

with $\mathbf{n} = \mathbf{n}_\Gamma$ the unit normal at the interface (pointing from Ω_1 in Ω_2), τ the surface tension coefficient (material parameter), $\mathcal{K}$ the curvature of Γ, and $\boldsymbol{\sigma}$ the stress tensor, i.e.,

$$\boldsymbol{\sigma} = -p\mathbf{I} + \mu\mathbf{D}(\mathbf{u}), \quad \mathbf{D}(\mathbf{u}) = \nabla\mathbf{u} + (\nabla\mathbf{u})^T,$$

with $p = p(x,t)$ the pressure, $\mathbf{u} = \mathbf{u}(x,t)$ the velocity vector, and μ the viscosity. We assume continuity of the velocity across the interface. In combination with the

*Received by the editors August 15, 2006; accepted for publication (in revised form) January 17, 2007; published electronically August 22, 2007. This work was supported by the German Research Foundation through SFB 540.

http://www.siam.org/journals/sinum/45-4/66753.html

†Institut für Geometrie und Praktische Mathematik, RWTH Aachen, D-52056 Aachen, Germany (gross@igpm.rwth-aachen.de, reusken@igpm.rwth-aachen.de).

conservation laws of mass and momentum this yields the following standard model (cf., for example, [23, 22, 26, 25]):

$$\begin{cases} -\operatorname{div}(\mu_i \mathbf{D}(\mathbf{u})) + \rho_i(\mathbf{u}\cdot\nabla)\mathbf{u} - \nabla p = \rho_i \mathbf{g} & \text{in } \Omega_i \\ \operatorname{div}\mathbf{u} = 0 & \text{in } \Omega_i \end{cases} \quad \text{for } i = 1,2, \tag{1.1}$$

$$[\boldsymbol{\sigma}\mathbf{n}]_\Gamma = \tau\mathcal{K}\mathbf{n}, \quad [\mathbf{u}]_\Gamma = 0. \tag{1.2}$$

The vector $\mathbf{g}$ is a known external force (gravity). In addition we need boundary conditions for $\mathbf{u}$ at $\partial\Omega$. For simplicity we take homogeneous Dirichlet boundary conditions. The *two* Navier–Stokes equations in (1.1) and the coupling conditions at the interface in (1.2) can be reformulated into *one* Navier–Stokes equation in the whole domain in which the effect of the surface tension is expressed in terms of a localized force at the interface; cf. the so-called continuum surface force (CSF) model [5, 6]. We consider this alternative formulation in a standard weak form (as in [12, 27, 28, 29, 30]) in the spaces

$$\mathbf{V} := H_0^1(\Omega)^3, \quad Q := L_0^2(\Omega) = \left\{ q \in L^2(\Omega) \mid \int_\Omega q\,dx = 0 \right\}.$$

For the L^2 scalar product we use the notation $(f,g) := \int_\Omega fg\,dx$ (and similarly for vector functions). The standard norm in $\mathbf{V}$ is denoted by $\|\cdot\|_1$. The weak formulation is as follows: Determine $(\mathbf{u},p) \in \mathbf{V}\times Q$ such that

$$\begin{aligned} &\int_\Omega \frac{\mu}{2}\,\mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v})\,dx + (\rho\,\mathbf{u}\cdot\nabla\mathbf{u},\mathbf{v}) + (\operatorname{div}\mathbf{v},p) = (\rho\,\mathbf{g},\mathbf{v}) + f_\Gamma(\mathbf{v}) \quad \text{for all } \mathbf{v}\in\mathbf{V}, \\ &(\operatorname{div}\mathbf{u},q) = 0 \quad \text{for all } q \in Q, \end{aligned} \tag{1.3}$$

with

$$f_\Gamma(\mathbf{v}) = \tau \int_\Gamma \mathcal{K}\mathbf{n}_\Gamma \cdot \mathbf{v}\,ds, \tag{1.4}$$

and $\mathbf{D}(\mathbf{u}) : \mathbf{D}(\mathbf{v}) = \operatorname{tr}\big(\mathbf{D}(\mathbf{u})\mathbf{D}(\mathbf{v})\big)$. The functions μ and ρ are strictly positive and piecewise constant in $\Omega_i,\ i=1,2$. For Γ sufficiently smooth we have $\sup_{x\in\Gamma}|\mathcal{K}(x)| \le c < \infty$, and thus

$$|f_\Gamma(\mathbf{v})| \le c\,\tau \int_\Gamma |\mathbf{n}_\Gamma\cdot\mathbf{v}|\,ds \le c\,\|\mathbf{v}\|_{L^2(\Gamma)} \le c\|\mathbf{v}\|_1 \quad \text{for all } \ \mathbf{v}\in\mathbf{V}. \tag{1.5}$$

Here and in the remainder we use the notation c for a generic constant. From (1.5) we see that $f_\Gamma \in \mathbf{V}'$, and thus under the usual assumptions (cf. [13]) the stationary Navier–Stokes equations (1.3) have a unique solution. We emphasize that the location of the interface is in general unknown and has to be determined (approximated) before the Navier–Stokes equations (1.3) can be solved. In this paper we assume that the unknown interface is captured using a level set technique. For a discussion of level set methods in incompressible two-phase flow problems we refer to the literature [6, 14, 21, 24]. We assume that the interface Γ is characterized as the zero level of the level set function d, which locally (close to the interface) is a signed distance function.

We now turn to the discretization of (1.3). We assume that $\mathcal{S}$ is a triangulation of Ω consisting of tetrahedra. With this triangulation we associate a mesh size parameter H. Let $\mathbf{V}_H \subset \mathbf{V}$, $Q_H \subset Q$ be standard polynomial finite element spaces corresponding to the triangulation $\mathcal{S}$, for example, the Hood–Taylor P_2-P_1 pair. In practice, the triangulation $\mathcal{S}$ is locally refined close to the interface Γ but *not* aligned with this interface; cf. Figures 2.1 and 6.1. The Galerkin discretization is as follows: Determine $(\mathbf{u}_H, p_H) \in \mathbf{V}_H \times Q_H$ such that

$$
\begin{aligned}
&\int_\Omega \frac{\mu}{2}\, \mathbf{D}(\mathbf{u}_H) : \mathbf{D}(\mathbf{v}_H)\, dx + (\rho\, \mathbf{u}_H \cdot \nabla \mathbf{u}_H, \mathbf{v}_H) + (\operatorname{div} \mathbf{v}_H, p_H) \\
&= (\rho\, \mathbf{g}, \mathbf{v}_H) + f_\Gamma(\mathbf{v}_H) \quad \text{for all } \ \mathbf{v}_H \in \mathbf{V}_H, \\
&(\operatorname{div} \mathbf{u}_H, q_H) = 0 \quad \text{for all } \ q_H \in Q_H.
\end{aligned} \tag{1.6}
$$

For this discrete problem, many important theoretical issues are still unsolved. For example, regarding iterative solvers there is the issue of robustness w.r.t. large jumps in the density and viscosity coefficients (results for Stokes equations are given in [20, 19, 18]). A second example is the effect of errors in the approximation of $f_\Gamma(\mathbf{v}_H)$ on the accuracy of the flow variables. In this paper we treat the latter topic.

As mentioned above, the interface Γ has to be approximated. Furthermore, to evaluate the integral in (1.4) the curvature of Γ has to be approximated and a quadrature rule may be needed. Thus the term $f_\Gamma(\mathbf{v}_H)$ on the right-hand side in (1.6) will be replaced by an approximation $\tilde{f}(\mathbf{v}_H)$. For the effect of the surface tension force approximation error on the accuracy of the velocity and pressure variables, the quantity

$$
\sup_{\mathbf{v} \in \mathbf{V}_H} \frac{f_\Gamma(\mathbf{v}_H) - \tilde{f}(\mathbf{v}_H)}{\|\mathbf{v}_H\|_1} \tag{1.7}
$$

is crucial (Strang lemma). The two main ingredients in the approximation method that we use are the following. First, a Laplace–Beltrami characterization of the curvature is used. This technique has been applied in mean curvature flows (cf. [7]) and in flows with a free capillary surface (cf. [3, 4]). Application of this technique in two-phase incompressible flows can be found in [12, 11, 14, 17]. Second, the unknown interface Γ (zero level of d) is approximated as the zero level Γ_h of a finite element approximation d_h of d. The approximate interface Γ_h consists of triangular faces. The parameter h is the maximal diameter of these faces and is not necessarily of the same order of magnitude as H. For this approximation technique we derive a sharp bound for the quantity in (1.7). The main result of this paper is the $\mathcal{O}(\sqrt{h})$ bound given in Corollary 4.8. We do not know of any literature in which, for this technique or for any other technique for approximating $f_\Gamma(\mathbf{v}_H)$, rigorous bounds for the quantity in (1.7) are derived. A numerical experiment (given in section 6) indicates that the $\mathcal{O}(\sqrt{h})$ is sharp. Our analysis reveals how the approximation method can be improved. A modified new approach, resulting in an $\mathcal{O}(h)$ bound, is presented in section 5.

2. Approximation of the surface tension force $f_\Gamma(\mathbf{v}_H)$. In this section we explain how the localized surface tension force term, $f_\Gamma(\mathbf{v}_H)$ in (1.6), is approximated. For this we first need some notions from differential geometry.

Let U be an open subset in $\mathbb{R}^3$ and Γ a connected C^2 compact hypersurface contained in U. For a sufficiently smooth function $g : U \to \mathbb{R}$ the tangential derivative

(along Γ) is defined by projecting the derivative on the tangent space of Γ, i.e.,

$$\nabla_\Gamma g = \nabla g - \nabla g \cdot \mathbf{n}_\Gamma \, \mathbf{n}_\Gamma. \tag{2.1}$$

The *Laplace–Beltrami operator* of g on Γ is defined by

$$\Delta_\Gamma g := \nabla_\Gamma \cdot \nabla_\Gamma g.$$

It can be shown that $\nabla_\Gamma g$ and $\Delta_\Gamma g$ depend only on values of g on Γ. For vector valued functions f, $g : \Gamma \to \mathbb{R}^3$ we define

$$\Delta_\Gamma f := (\Delta_\Gamma f_1, \Delta_\Gamma f_2, \Delta_\Gamma f_3)^T, \qquad \nabla_\Gamma f \cdot \nabla_\Gamma g := \sum_{i=1}^{3} \nabla_\Gamma f_i \cdot \nabla_\Gamma g_i.$$

We recall the following basic result from differential geometry.

THEOREM 2.1. *Let* $\mathrm{id}_\Gamma : \Gamma \to \mathbb{R}^3$ *be the identity on* Γ *and* $\mathcal{K} = \kappa_1 + \kappa_2$ *the sum of the principal curvatures. For all sufficiently smooth vector functions* $\mathbf{v}$ *on* Γ *the following holds:*

$$\int_\Gamma \mathcal{K}\, \mathbf{n}_\Gamma \cdot \mathbf{v}\, ds = -\int_\Gamma (\Delta_\Gamma \,\mathrm{id}_\Gamma) \cdot \mathbf{v}\, ds = \int_\Gamma \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma \mathbf{v}\, ds. \tag{2.2}$$

In a finite element setting (which is based on a weak formulation) it is natural to use the expression on the right-hand side of (2.2) as a starting point for the discretization. This idea is used in, for example, [10, 4, 12, 14]. In this discretization we use *an approximation* Γ_h *of* Γ.

For the formulation of assumptions on the approximate interface Γ_h it is convenient to introduce the signed distance function $d : U \to \mathbb{R}$, $|d(x)| := \mathrm{dist}(x, \Gamma)$ for all $x \in U$. Thus Γ is the zero level set of d. We assume $d < 0$ on the interior of Γ (that is, in Ω_1) and $d > 0$ on the exterior. Note that $\mathbf{n}_\Gamma = \nabla d$ on Γ. We define $\mathbf{n}(x) := \nabla d(x)$ for all $x \in U$. Thus $\mathbf{n} = \mathbf{n}_\Gamma$ on Γ and $\|\mathbf{n}(x)\| = 1$ for all $x \in U$. Here and in the remainder $\|\cdot\|$ denotes the Euclidean norm. The Hessian of d is denoted by $\mathbf{H}$:

$$\mathbf{H}(x) = D^2 d(x) \in \mathbb{R}^{3\times 3} \quad \text{for all} \quad x \in U. \tag{2.3}$$

The eigenvalues of $\mathbf{H}(x)$ are denoted by $\kappa_1(x), \kappa_2(x)$, and 0. For $x \in \Gamma$ the eigenvalues $\kappa_i(x)$, $i = 1, 2$, are the principal curvatures.

We will need the orthogonal projection

$$\mathbf{P}(x) = \mathbf{I} - \mathbf{n}(x)\mathbf{n}(x)^T \quad \text{for} \quad x \in U.$$

Note that the tangential derivative can be written as $\nabla_\Gamma g = \mathbf{P}\nabla g$.

Using the distance function d we now introduce assumptions on the approximate interface Γ_h. In Remark 2 below we indicate how in practice an approximate interface Γ_h can be constructed which satisfies these assumptions. Let $\{\Gamma_h\}_{h>0}$ be a family of polygonal approximations of Γ. Each Γ_h is contained in U and consists of a set $\mathcal{F}_h$ of triangular faces: $\Gamma_h = \cup_{T\in\mathcal{F}_h} T$. For $T_1, T_2 \in \mathcal{F}_h$ with $T_1 \neq T_2$ we assume that $T_1 \cap T_2$ is either empty or a common edge or a common vertex. The parameter h denotes the maximal diameter of the triangles in $\mathcal{F}_h$: $h = \max_{T\in\mathcal{F}_h} \mathrm{diam}(T)$. By $\mathbf{n}_h$ we denote the outward pointing unit normal on Γ_h. This normal is piecewise constant with possible discontinuities at the edges of the triangles in $\mathcal{F}_h$.

The approximation Γ_h is assumed to be close to Γ in the following sense:

$$|d(x)| \leq ch^2 \quad \text{for all} \quad x \in \Gamma_h, \tag{2.4}$$

$$\operatorname{ess\,inf}_{x\in\Gamma_h} \mathbf{n}(x)^T \mathbf{n}_h(x) \geq c > 0, \tag{2.5}$$

$$\operatorname{ess\,sup}_{x\in\Gamma_h} \|\mathbf{P}(x)\mathbf{n}_h(x)\| \leq ch. \tag{2.6}$$

Here c denotes a generic constant independent of h.

Remark 1. The conditions (2.5), (2.6) are satisfied if

$$\operatorname{ess\,sup}_{x\in\Gamma_h} \|\mathbf{n}(x) - \mathbf{n}_h(x)\| \leq \min\{c_0, ch\} \quad \text{with} \quad c_0 < \sqrt{2} \tag{2.7}$$

holds. This easily follows from

$$\|\mathbf{n}(x) - \mathbf{n}_h(x)\|^2 = 2\big(1 - \mathbf{n}(x)^T \mathbf{n}_h(x)\big)$$

and

$$\|\mathbf{P}(x)\mathbf{n}_h(x)\| = \|\mathbf{P}(x)\big(\mathbf{n}(x) - \mathbf{n}_h(x)\big)\| \leq \|\mathbf{n}(x) - \mathbf{n}_h(x)\|.$$

Remark 2. We briefly explain the approach that is used in [14] (cf. also [9]) for computing Γ_h. Let $\mathcal{S}$ be the (locally refined) triangulation of Ω, consisting of tetrahedra, that is used for the discretization of the flow variables with finite elements; cf. (1.6) (in our approach we use the Hood–Taylor P_2-P_1 pair). The level set equation for d is discretized with continuous piecewise quadratic finite elements on a triangulation $\mathcal{T}$. This triangulation is either equal to $\mathcal{S}$ or obtained from one or a few regular refinements of $\mathcal{S}$ (the subdivision of each tetrahedron in eight child tetrahedra). The piecewise *quadratic* finite element approximation of d on $\mathcal{T}$ is denoted by d_h. We now introduce one further regular refinement of $\mathcal{T}$, resulting in $\mathcal{T}'$. Let $I(d_h)$ be the continuous piecewise *linear* function on $\mathcal{T}'$ which interpolates d_h at all vertices of all tetrahedra in $\mathcal{T}'$. The approximation of the interface Γ is defined by

$$\Gamma_h := \{\, x \in \Omega \mid I(d_h)(x) = 0 \,\}$$

which consists of piecewise planar segments. The mesh size parameter h is the maximal diameter of these segments. This (maximal) diameter is approximately the (maximal) diameter of the tetrahedra in $\mathcal{T}'$ that contain the discrete interface; i.e., h is approximately the maximal diameter of the tetrahedra in $\mathcal{T}'$ *that are close to the interface*. In Figure 2.1 we illustrate this construction for the two-dimensional case.

Each of the planar segments of Γ_h is either a triangle or a quadrilateral. The quadrilaterals can (formally) be divided into two triangles. Thus Γ_h consists of a set $\mathcal{F}_h$ of triangular faces. For the example considered in section 6, in which Γ is a sphere, the resulting polygonal approximations Γ_h for $h = \frac{1}{5}$ and $h = \frac{1}{10}$, resp., are shown in Figure 2.2.

We note the following related to the assumptions (2.4)–(2.6). If we assume $|I(d_h)(x) - d(x)| \leq c\,h^2$ for all x in a neighborhood of Γ, which is reasonable for a smooth d and piecewise quadratic d_h, then for $x \in \Gamma_h$ we have $|d(x)| = |d(x) - I(d_h)(x)| \leq c\,h^2$, and thus (2.4) is satisfied. Instead of (2.5), (2.6) we consider the sufficient condition (2.7). We assume $\|\nabla d(x) - \nabla I(d_h)(x)\| \leq c\,h$ for all x in a neighborhood of Γ (x not on an edge), which again is reasonable for a smooth d and

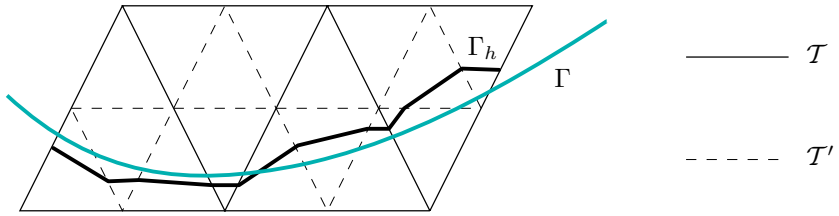


FIG. 2.1. *Construction of approximate interface for the two-dimensional case.*

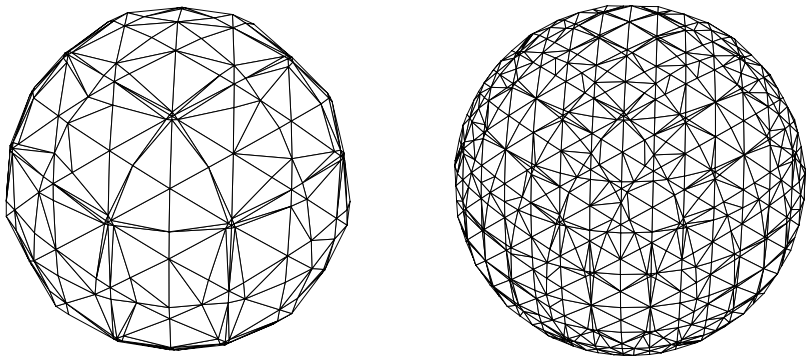


FIG. 2.2. *Approximate interface Γ_h for the example from section 6 on a coarse grid (left) and after one refinement (right).*

piecewise quadratic d_h. Due to $\|\nabla d\| = 1$ we then also have $\|\nabla I(d_h)(x)\| = 1 + \mathcal{O}(h)$ in a neighborhood of Γ. For $x \in \Gamma_h$ (not on an edge) we obtain

$$\begin{aligned}\|\mathbf{n}_h(x) - \mathbf{n}(x)\| &= \Big\| \frac{\nabla I(d_h)(x)}{\|\nabla I(d_h)(x)\|} - \nabla d(x) \Big\| \\ &\leq \Big| \frac{1}{\|\nabla I(d_h)(x)\|} - 1 \Big| \cdot \|\nabla I(d_h)(x)\| + \|\nabla I(d_h)(x) - \nabla d(x)\| \leq c\, h,\end{aligned}$$

and thus (2.7) is satisfied (for h sufficiently small).

Given an approximate interface Γ_h, the localized force term $f_\Gamma(\mathbf{v}_H)$ is approximated by

$$\tilde{f}(\mathbf{v}_H) = f_{\Gamma_h}(\mathbf{v}_H) := \tau \int_{\Gamma_h} \nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h} \cdot \nabla_{\Gamma_h} \mathbf{v}_H \, ds, \qquad \mathbf{v}_H \in \mathbf{V}_H. \tag{2.8}$$

Under the assumptions (2.4)–(2.6) on the family $\{\Gamma_h\}_{h>0}$ in section 4 we will derive a bound for the approximation error

$$\sup_{\mathbf{v}_H \in \mathbf{V}_H} \frac{f_\Gamma(\mathbf{v}_H) - f_{\Gamma_h}(\mathbf{v}_H)}{\|\mathbf{v}_H\|_1} \quad \text{with} \quad f_{\Gamma_h}(\mathbf{v}_H) \quad \text{as in (2.8)}. \tag{2.9}$$

Remark 3. From Theorem 2.1, the fact that $f_\Gamma(\mathbf{v}) = \tau \int_\Gamma \mathcal{K}\mathbf{v} \cdot \mathbf{n}\, ds$ is a bounded linear functional on $\mathbf{V}$, and a density argument, it follows that the linear functional

$$f_\Gamma : \ \mathbf{v} \to \tau \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma \mathbf{v} \, ds \ , \qquad \mathbf{v} \in \left(C_0^\infty(\Omega)\right)^3, \tag{2.10}$$

has a unique bounded extension to $\mathbf{V}$. Therefore, for $f_\Gamma : \mathbf{V} \to \mathbb{R}$ we can use both the representation in (1.4) and the one in (2.10) (these are the same on a dense subset). This, however, is *not* the case for f_{Γ_h}. Because Γ_h is not sufficiently smooth, a partial integration result as in Theorem 2.1 does not hold. The linear functional

$$\mathbf{v} \to \tau \int_{\Gamma_h} \nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h} \cdot \nabla_{\Gamma_h} \mathbf{v}\, ds$$

is *not* necessarily bounded on $\mathbf{V}$. For this reason the restriction to $\mathbf{v}_H$ from the finite element space $\mathbf{V}_H$ in (2.9) is essential.

Remark 4. At many places in this section, for example in (2.2), (2.3) and (implicitly) in (2.4), and also in the analysis presented in the next section, the assumption that Γ is a C^2 smooth interface plays a crucial role. We do not know of any literature in which smoothness properties of the interface are analyzed for a Navier–Stokes incompressible two-phase flow problem with surface tension. In [2] and [1] a two-phase Stokes flow problem without surface tension, in which the evolution is driven by the gravity force, is analyzed. In [2] it is proved that if the initial configuration has a C^2 smooth interface $\Gamma = \Gamma(0)$, then for arbitrary finite time $t > 0$ the interface $\Gamma(t)$ is a surface of class $C^{2-\varepsilon}$ for arbitrary $\varepsilon \in (0, 2]$. In [1] it is shown that if $\Gamma(0)$ is a $C^{2+\ell}$ smooth surface, with $\ell > 0$, then $\Gamma(t)$ is of class $C^{2+\ell}$, too, for all $t \in [0, T]$ and $T > 0$ sufficiently small.

3. Preliminaries. In this section we collect some results that will be used in the analysis in section 4. The techniques that we use come from the paper [8]. For proofs of certain results we will refer to that paper.

We introduce a locally (in a neighborhood of Γ) orthogonal coordinate system by using the projection $\mathbf{p} : U \to \Gamma$:

$$\mathbf{p}(x) = x - d(x)\mathbf{n}(x) \quad \text{for all} \quad x \in U.$$

We assume that the decomposition $x = \mathbf{p}(x) + d(x)\mathbf{n}(x)$ is unique for all $x \in U$. Note that

$$\mathbf{n}(x) = \mathbf{n}(\mathbf{p}(x)) \quad \text{for all} \quad x \in U.$$

We use an extension operator defined as follows. For a (scalar) function v defined on Γ we define

$$v_\Gamma^e(x) := v(x - d(x)\mathbf{n}(x)) = v(\mathbf{p}(x)) \quad \text{for all} \quad x \in U;$$

i.e., v is extended along normals on Γ. We will also need extensions of functions defined on Γ_h to U. This is done again by extending along normals $\mathbf{n}(x)$. For v defined on Γ_h we define, for $x \in \Gamma_h$,

$$v_{\Gamma_h}^e(x + \alpha\mathbf{n}(x)) := v(x) \quad \text{for all} \quad \alpha \in \mathbb{R} \quad \text{with} \quad x + \alpha\mathbf{n}(x) \in U. \tag{3.1}$$

The projection $\mathbf{p}$ and the extensions v_Γ^e, $v_{\Gamma_h}^e$ are illustrated in Figure 3.1.

We define a discrete analogue of the orthogonal projection $\mathbf{P}$:

$$\mathbf{P}_h(x) := \mathbf{I} - \mathbf{n}_h(x)\mathbf{n}_h(x)^T \quad \text{for} \quad x \in \Gamma_h,\ x \text{ not on an edge.}$$

The tangential derivative along Γ_h can be written as $\nabla_{\Gamma_h} g = \mathbf{P}_h \nabla g$. In the analysis a further technical assumption is used, namely that the neighborhood U of Γ is

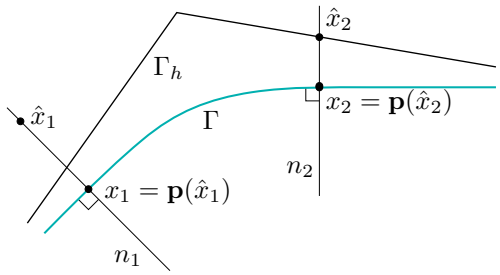


FIG. 3.1. *Example for projection* $\mathbf{p}$ *and construction of extension operators.* n_1 *and* n_2 *are straight lines perpendicular to* Γ. *For* v *defined on* Γ *we have* $v_\Gamma^e \equiv v(x_1)$ *on* n_1. *For* v_h *defined on* Γ_h *we have* $v_{\Gamma_h}^e \equiv v_h(\hat{x}_2)$ *on* n_2.

sufficiently small in the following sense. We assume that U is a strip of width $\delta > 0$ with

$$\delta^{-1} > \max_{i=1,2} \|\kappa_i(x)\|_{L^\infty(\Gamma)}. \tag{3.2}$$

Assumption 1. In the remainder of the paper we assume that (2.4), (2.5), (2.6), and (3.2) hold.

We present two lemmas from [8]. Proofs are elementary and can be found in [8].

LEMMA 3.1. *For the projection operator* $\mathbf{P}$ *and the Hessian* $\mathbf{H}$ *the relation*

$$\mathbf{P}(x)\mathbf{H}(x) = \mathbf{H}(x)\mathbf{P}(x) = \mathbf{H}(x) \quad \text{for all} \quad x \in U$$

holds. For v *defined on* Γ *and sufficiently smooth the following holds:*

$$\nabla_{\Gamma_h} v_\Gamma^e(x) = \mathbf{P}_h(x)\big(\mathbf{I} - d(x)\mathbf{H}(x)\big)\mathbf{P}(x)\nabla_\Gamma v(\mathbf{p}(x)) \quad \text{a.e. on} \quad \Gamma_h. \tag{3.3}$$

Proof. A proof is given in section 2.3 in [8]. ∎

In (3.3) (and also below) we have the result "a.e. on Γ_h" because quantities (derivatives, $\mathbf{P}_h$, etc.) are not well defined on the edges of the triangulation Γ_h.

LEMMA 3.2. *For* $x \in \Gamma_h$ *(not on an edge) define*

$$\mu(x) = \big[\Pi_{i=1}^2 (1 - d(x)\kappa_i(x))\big]\mathbf{n}(x)^T\mathbf{n}_h(x), \tag{3.4}$$

$$\mathbf{A}(x) = \frac{1}{\mu(x)}\mathbf{P}(x)\big[\mathbf{I} - d(x)\mathbf{H}(x)\big]\mathbf{P}_h(x)\big[\mathbf{I} - d(x)\mathbf{H}(x)\big]\mathbf{P}(x). \tag{3.5}$$

Let $\mathbf{A}_{\Gamma_h}^e$ *be the extension of* $\mathbf{A}$ *as in* (3.1). *The following identity holds for functions* v *and* ψ *that are defined on* Γ_h *and sufficiently smooth:*

$$\int_{\Gamma_h} \nabla_{\Gamma_h} v \cdot \nabla_{\Gamma_h} \psi \, ds = \int_\Gamma \mathbf{A}_{\Gamma_h}^e \nabla_\Gamma v_{\Gamma_h}^e \cdot \nabla_\Gamma \psi_{\Gamma_h}^e \, ds. \tag{3.6}$$

Proof. A proof is given in section 2.3 in [8]. ∎

Due to the assumptions in (2.5) and (3.2) we have $\operatorname{ess\,inf}_{x\in\Gamma_h} \mu(x) > 0$, and thus $\mathbf{A}(x)$ is well defined.

We now derive two further results that are needed in the analysis in section 4.

LEMMA 3.3. *There exists a constant* c *independent of* h *such that*

$$\|\nabla_\Gamma v_{\Gamma_h}^e\|_{L^2(\Gamma)} \le c\,\|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \quad \text{for all} \quad v \in H^1(\Gamma_h) \cap C(\Gamma_h).$$

Proof. Due to Lemma 3.2 we have

$$\|\nabla_{\Gamma_h} v\|^2_{L^2(\Gamma_h)} = \int_\Gamma \mathbf{A}^e_{\Gamma_h}(y)\nabla_\Gamma v^e_{\Gamma_h}(y)\cdot\nabla_\Gamma v^e_{\Gamma_h}(y)\,ds(y)$$

with $ds(y)$ the surface measure on Γ. Take $x\in\Gamma_h$ with $\mathbf{p}(x)=y$. If x does not lie on an edge, we have $\mathbf{A}^e_{\Gamma_h}(y)=\mathbf{A}(x)$ with $\mathbf{A}(x)$ as in (3.5). We drop the symbol x in the notation and write $\mathbf{A}(x)=\mathbf{A}=\frac{1}{\mu}\mathbf{P}(\mathbf{I}-d\mathbf{H})\mathbf{P}_h(\mathbf{I}-d\mathbf{H})\mathbf{P}$. Decompose $\mathbf{n}_h$ as $\mathbf{n}_h=\alpha\mathbf{n}+\beta\mathbf{n}^\perp$ with $\|\mathbf{n}^\perp\|=1$ and $\mathbf{n}^T\mathbf{n}^\perp=0$. From (2.5) is follows that $\alpha\geq c>0$ and thus $\beta^2\leq 1-c^2<1$. Take $z\in\operatorname{range}(\mathbf{P})$ with $\|z\|=1$. We then have $\|\mathbf{P}_h z\|\geq\|z\|-|z^T\mathbf{n}_h|=\|z\|-|\beta||z^T\mathbf{n}^\perp|\geq(1-|\beta|)\|z\|$. Hence, there is a constant $c>0$ such that

$$\|\mathbf{P}_h\mathbf{P}w\|\geq c\,\|\mathbf{P}w\|\quad\text{for all}\ \ w\in\mathbb{R}^3.$$

Using (3.2) it follows that there is a constant $c>0$ such that $\|(\mathbf{I}-d\mathbf{H})w\|\geq c\,\|w\|$ for all $w\in\mathbb{R}^3$. Note that $\mu=\mu(x)\geq c>0$ holds. From these results we obtain, using $\mathbf{PH}=\mathbf{HP}$ (Lemma 3.1),

$$\begin{aligned} w^T\mathbf{A}w &= \frac{1}{\mu}w^T\mathbf{P}(\mathbf{I}-d\mathbf{H})\mathbf{P}_h(\mathbf{I}-d\mathbf{H})\mathbf{P}w\\ &= \frac{1}{\mu}\|\mathbf{P}_h\mathbf{P}(\mathbf{I}-d\mathbf{H})w\|^2\geq c\,\|\mathbf{P}w\|^2\quad\text{for all}\ \ w\in\mathbb{R}^3, \end{aligned}$$

with a constant $c>0$. This yields, using $\mathbf{n}(x)=\mathbf{n}(\mathbf{p}(x))=\mathbf{n}(y)$,

$$\mathbf{A}^e_{\Gamma_h}(y)w\cdot w = w^T\mathbf{A}(x)w\geq c\,\|\mathbf{P}(x)w\|^2=\|\mathbf{P}(y)w\|^2,$$

with $c>0$. For $w=\nabla_\Gamma v^e_{\Gamma_h}(y)$ we have $\mathbf{P}(y)w=\mathbf{P}(y)\nabla_\Gamma v^e_{\Gamma_h}(y)=\nabla_\Gamma v^e_{\Gamma_h}(y)$, and thus we get

$$\begin{aligned} \|\nabla_{\Gamma_h} v\|^2_{L^2(\Gamma_h)} &= \int_\Gamma \mathbf{A}^e_{\Gamma_h}(y)\nabla_\Gamma v^e_{\Gamma_h}(y)\cdot\nabla_\Gamma v^e_{\Gamma_h}(y)\,ds(y)\\ &\geq c\int_\Gamma \nabla_\Gamma v^e_{\Gamma_h}(y)\cdot\nabla_\Gamma v^e_{\Gamma_h}(y)\,ds(y)=c\,\|\nabla_\Gamma v^e_{\Gamma_h}\|^2_{L^2(\Gamma)}, \end{aligned}$$

with a constant $c>0$. ∎

LEMMA 3.4. *The following holds:*

$$\operatorname{ess\,sup}_{y\in\Gamma}\|\big(\mathbf{A}^e_{\Gamma_h}(y)-\mathbf{I}\big)\mathbf{P}(y)\|\leq ch^2.$$

Proof. Take $y\in\Gamma$ and a corresponding $x\in\Gamma_h$ such that $\mathbf{p}(x)=y$. Assume that x does not lie on an edge of the triangulation Γ_h, which is true for almost all $y\in\Gamma$. Then we have

$$(\mathbf{A}^e_{\Gamma_h}(y)-\mathbf{I})\mathbf{P}(y)=(\mathbf{A}(x)-\mathbf{I})\mathbf{P}(x).$$

We drop the symbol x in the notation and write $\mathbf{A}(x)=\mathbf{A}=\frac{1}{\mu}\mathbf{P}(\mathbf{I}-d\mathbf{H})\mathbf{P}_h(\mathbf{I}-d\mathbf{H})\mathbf{P}$. Note that $|\mu|=\mu(x)\geq c>0$ holds. Decompose $\mathbf{n}_h$ as $\mathbf{n}_h=\alpha\mathbf{n}+\beta\mathbf{n}^\perp$ with $\|\mathbf{n}^\perp\|=1$ and $\mathbf{n}^T\mathbf{n}^\perp=0$. Due to (2.5) we have $\alpha=\mathbf{n}^T\mathbf{n}_h\geq c>0$. From (2.6) we get $\|\mathbf{P}\mathbf{n}_h\|=|\beta|\leq ch$. Hence,

$$|\mathbf{n}^T\mathbf{n}_h-1|=1-\alpha=\frac{1-\alpha^2}{1+\alpha}\leq 1-\alpha^2=\beta^2\leq c\,h^2. \tag{3.7}$$

Using this and $|d(x)| \leq ch^2,\ |\kappa_i(x)| \leq c$, we obtain $|\mu - 1| \leq c\,h^2$. Thus

$$\left|\frac{1}{\mu} - 1\right| = \frac{|\mu - 1|}{\mu} \leq c\,h^2 \tag{3.8}$$

holds. We have

$$\begin{aligned}(\mathbf{A} - \mathbf{I})\mathbf{P} &= \frac{1}{\mu}\mathbf{P}(\mathbf{I} - d\mathbf{H})\mathbf{P}_h(\mathbf{I} - d\mathbf{H})\mathbf{P} - \mathbf{P} \\ &= \left[\left(\frac{1}{\mu} - 1\right)\mathbf{P}(\mathbf{I} - d\mathbf{H})\mathbf{P}_h(\mathbf{I} - d\mathbf{H})\mathbf{P}\right] + \left[\mathbf{P}(\mathbf{I} - d\mathbf{H})\mathbf{P}_h(\mathbf{I} - d\mathbf{H})\mathbf{P} - \mathbf{P}\right]\end{aligned}$$

and consider the two terms on the right-hand side separately. For the first term we get, using (3.8),

$$\left\|\left(\frac{1}{\mu} - 1\right)\mathbf{P}(\mathbf{I} - d\mathbf{H})\mathbf{P}_h(\mathbf{I} - d\mathbf{H})\mathbf{P}\right\| \leq \left|\frac{1}{\mu} - 1\right|(1 + c\,h^2)(1 + c\,h^2) \leq c\,h^2.$$

For the second term we obtain, using (2.6),

$$\begin{aligned}\|\mathbf{P}(\mathbf{I} - d\mathbf{H})\mathbf{P}_h(\mathbf{I} - d\mathbf{H})\mathbf{P} - \mathbf{P}\| &\leq \|\mathbf{P}\mathbf{P}_h\mathbf{P} - \mathbf{P}\| + c\,h^2 \\ &= \|\mathbf{P}\mathbf{n}_h\mathbf{n}_h^T\mathbf{P}\| + c\,h^2 = \|\mathbf{P}\mathbf{n}_h\|^2 + c\,h^2 \leq c\,h^2.\end{aligned}$$

Combination of these bounds completes the proof. ∎

4. Approximation error analysis. We are interested in the difference between the terms

$$\tau \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma \mathbf{v}_H\, ds \quad \text{and} \quad \tau \int_{\Gamma_h} \nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h} \cdot \nabla_{\Gamma_h} \mathbf{v}_H\, ds \quad \text{for} \quad \mathbf{v}_H \in \mathbf{V}_H.$$

Since $\nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma \mathbf{v}_H = \sum_{i=1}^3 \nabla_\Gamma(\mathrm{id}_\Gamma)_i \cdot \nabla_\Gamma(\mathbf{v}_H)_i$ we consider only one term in this sum, say the ith. We write id_Γ and v for the *scalar* functions $(\mathrm{id}_\Gamma)_i$ and $(\mathbf{v}_H)_i$, respectively. We write id_{Γ_h} for $(\mathrm{id}_{\Gamma_h})_i$. Note that

$$\nabla_\Gamma \mathrm{id}_\Gamma = \mathbf{P}\nabla \mathrm{id}_\Gamma = \mathbf{P}e_i, \quad \nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h} = \mathbf{P}_h \nabla \mathrm{id}_{\Gamma_h} = \mathbf{P}_h e_i,$$

with e_i the ith basis vector in $\mathbb{R}^3$. We introduce scalar versions of the functionals f_Γ and f_{Γ_h} defined in (2.10) and (2.8) (without loss of generality we can take $\tau := 1$):

$$g(v) := \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma v\, ds, \quad g_h(v) := \int_{\Gamma_h} \nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h} \cdot \nabla_{\Gamma_h} v\, ds.$$

As noted in Remark 3, g is a bounded linear functional on $H^1(U)$. To guarantee that g_h and the extension operator in (3.1) are well defined we assume $v \in H^1(\Gamma_h) \cap C(\Gamma_h)$. Therefore, in the analysis in this section we use the subspace W of $H^1(U)$ consisting of functions whose restriction to Γ_h belongs to $H^1(\Gamma_h) \cap C(\Gamma_h)$.

Remark 5. If we use a Hood–Taylor pair $\mathbf{V}_H \times Q_H$ in the discretization of the Navier–Stokes equations, then the ith component $v \in V_H$ of $\mathbf{v}_H \in \mathbf{V}_H = (V_H)^3$ is continuous and piecewise polynomial (on the tetrahedral triangulation $\mathcal{S}$). Thus $v \in W$ holds.

In this section we first derive, for $v \in W$, a bound for $|g(v) - g_h(v)|$ in terms of $\|v\|_{1,U} := \|v\|_{H^1(U)}$ and $\|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)}$. This bound is given in Corollary 4.4. Using this bound we then derive a bound for

$$\sup_{v \in V_H} \frac{g(v) - g_h(v)}{\|v\|_1};$$

cf. Theorem 4.7. This immediately implies a bound for the approximation error as in (2.9); cf. Corollary 4.8.

The analysis is based on the following splitting:

$$\begin{aligned}
&g(v) - g_h(v) \\
&= \int_\Gamma \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v \, ds - \int_{\Gamma_h} \nabla_{\Gamma_h} \,\mathrm{id}_\Gamma^e \cdot \nabla_{\Gamma_h} v \, ds + \int_{\Gamma_h} \nabla_{\Gamma_h} (\mathrm{id}_\Gamma^e - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v \, ds \\
&\overset{(3.6)}{=} \int_\Gamma \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v \, ds - \int_\Gamma \mathbf{A}_{\Gamma_h}^e \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v_{\Gamma_h}^e \, ds + \int_{\Gamma_h} \nabla_{\Gamma_h} (\mathrm{id}_\Gamma^e - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v \, ds \\
&= \int_\Gamma \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma (v - v_{\Gamma_h}^e) \, ds + \int_\Gamma (\mathbf{I} - \mathbf{A}_{\Gamma_h}^e) \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v_{\Gamma_h}^e \, ds \\
&\qquad + \int_{\Gamma_h} \nabla_{\Gamma_h} (\mathrm{id}_\Gamma^e - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v \, ds. \qquad (4.1)
\end{aligned}$$

In the lemmas below we derive bounds for the three terms in (4.1). Note that the first two terms do *not* involve id_{Γ_h}.

LEMMA 4.1. *The following holds:*

$$\left| \int_\Gamma (\mathbf{I} - \mathbf{A}_{\Gamma_h}^e) \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v_{\Gamma_h}^e \, ds \right| \leq c\, h^2 \, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \quad \textit{for all} \quad v \in W.$$

Proof. Using the Cauchy–Schwarz inequality and the results in Lemmas 3.3 and 3.4 we obtain

$$\begin{aligned}
&\left| \int_\Gamma (\mathbf{I} - \mathbf{A}_{\Gamma_h}^e) \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v_{\Gamma_h}^e \, ds \right| = \left| \int_\Gamma (\mathbf{I} - \mathbf{A}_{\Gamma_h}^e) \mathbf{P} \nabla_\Gamma \,\mathrm{id}_\Gamma \cdot \nabla_\Gamma v_{\Gamma_h}^e \, ds \right| \\
&\leq \operatorname{ess\,sup}_{y \in \Gamma} \|(\mathbf{I} - \mathbf{A}_{\Gamma_h}^e(y)) \mathbf{P}(y)\| \, \|\nabla_\Gamma \,\mathrm{id}_\Gamma\|_{L^2(\Gamma)} \, \|\nabla_\Gamma v_{\Gamma_h}^e\|_{L^2(\Gamma)} \\
&\leq c\, h^2 \, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)},
\end{aligned}$$

and thus the result holds. ∎

LEMMA 4.2. *The following holds:*

$$\left| \int_{\Gamma_h} \nabla_{\Gamma_h} (\mathrm{id}_\Gamma^e - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v \, ds \right| \leq c\, h \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \quad \textit{for all} \quad v \in W.$$

Proof. From Lemma 3.1 we get for $x \in \Gamma_h$ (not on an edge),

$$\begin{aligned}
\nabla_{\Gamma_h} \,\mathrm{id}_\Gamma^e(x) &= \mathbf{P}_h(x) \big(\mathbf{I} - d(x) \mathbf{H}(x)\big) \mathbf{P}(x) \nabla_\Gamma \,\mathrm{id}_\Gamma(\mathbf{p}(x)) \\
&= \mathbf{P}_h(x) \big(\mathbf{I} - d(x) \mathbf{H}(x)\big) \mathbf{P}(x) e_i.
\end{aligned}$$

We also have $\nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h} = \mathbf{P}_h \nabla \mathrm{id}_{\Gamma_h} = \mathbf{P}_h e_i$. Hence,

$$\left| \int_{\Gamma_h} \nabla_{\Gamma_h} (\mathrm{id}_\Gamma^e - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v \, ds \right| \tag{4.2}$$

$$= \left| \int_{\Gamma_h} \big(\mathbf{P}_h (\mathbf{I} - d\mathbf{H}) \mathbf{P} e_i - \mathbf{P}_h e_i \big) \cdot \nabla_{\Gamma_h} v \, ds \right|$$

$$\leq c \, \mathrm{ess\,sup}_{x \in \Gamma_h} \| \mathbf{P}_h(x) \big(\mathbf{I} - d(x) \mathbf{H}(x) \big) \mathbf{P}(x) - \mathbf{P}_h(x) \| \, \| \nabla_{\Gamma_h} v \|_{L^2(\Gamma_h)}$$

$$\leq c \, \mathrm{ess\,sup}_{x \in \Gamma_h} \big(\| \mathbf{P}_h(x) \mathbf{P}(x) - \mathbf{P}_h(x) \| \tag{4.3}$$

$$+ |d(x)| \| \mathbf{P}_h(x) \mathbf{H}(x) \mathbf{P}(x) \| \big) \| \nabla_{\Gamma_h} v \|_{L^2(\Gamma_h)}. \tag{4.4}$$

Note that $|d(x)| \leq c\, h^2$ for $x \in \Gamma_h$, and

$$\mathrm{ess\,sup}_{x \in \Gamma_h} \| \mathbf{P}_h(x) \mathbf{H}(x) \mathbf{P}(x) \| \leq \mathrm{ess\,sup}_{x \in \Gamma_h} \| \mathbf{H}(x) \| \leq c.$$

For the term in (4.3) we have (we drop x in the notation)

$$\| \mathbf{P}_h \mathbf{P} - \mathbf{P}_h \| = \| \mathbf{P}_h \mathbf{n} \mathbf{n}^T \| \leq \| \mathbf{P}_h \mathbf{n} \| \leq \| \mathbf{P}_h \mathbf{n} + \mathbf{P} \mathbf{n}_h \| + \| \mathbf{P} \mathbf{n}_h \|.$$

For the first term we get, using (3.7),

$$\| \mathbf{P}_h \mathbf{n} + \mathbf{P} \mathbf{n}_h \| = \big\| (1 - \mathbf{n}^T \mathbf{n}_h)(\mathbf{n} + \mathbf{n}_h) \big\| \leq 2 |1 - \mathbf{n}^T \mathbf{n}_h| \leq c\, h^2.$$

From (2.6) we get $\| \mathbf{P} \mathbf{n}_h \| \leq c\, h$ (a.e. on Γ_h). Thus $\| \mathbf{P}_h(x) \mathbf{P}(x) - \mathbf{P}_h(x) \| \leq c\, h$ holds a.e. on Γ_h. As an upper bound for (4.2) we obtain $ch \, \| \nabla_{\Gamma_h} v \|_{L^2(\Gamma_h)}$. ∎

LEMMA 4.3. *The following holds:*

$$\left| \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma (v - v^e_{\Gamma_h}) \, ds \right| \leq ch \| v \|_{1,U} \quad \textit{for all} \;\; v \in W.$$

Proof. We take $v \in C^1(U)$. For $y \in \Gamma$ we have $v^e_{\Gamma_h}(y) = v(y \pm \delta(y) \mathbf{n}(y))$ with a unique $\delta(y) \geq 0$ such that $y \pm \delta(y) \mathbf{n}(y) \in \Gamma_h$. Note that $\delta(y) \leq c\, h^2$ holds. Let $U_m \subset U$ be a strip around Γ that contains Γ_h and has width $m \leq c\, h^2$. We now have

$$\left| \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma (v - v^e_{\Gamma_h}) \, ds(y) \right| = \left| \int_\Gamma \Delta_\Gamma \mathrm{id}_\Gamma \, \big(v(y) - v(y \pm \delta(y) \mathbf{n}(y)) \big) \, ds(y) \right|$$

$$\leq \int_\Gamma |\Delta_\Gamma \mathrm{id}_\Gamma| \left| \int_0^{\delta(y)} \frac{\partial v}{\partial t}(y \pm t \mathbf{n}(y)) \, dt \right| ds(y)$$

$$\leq c \int_\Gamma \int_0^{\delta(y)} \left| \frac{\partial v}{\partial t}(y \pm t \mathbf{n}(y)) \right| dt \, ds(y).$$

For $x = y \pm t \mathbf{n}(y)$ with $0 \leq t \leq \delta(y)$ we use $\mathbf{n}(x) = \mathbf{n}(\mathbf{p}(x)) = \mathbf{n}(y)$ and obtain

$$\left| \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma (v - v^e_{\Gamma_h}) \, ds(y) \right| \leq c \int_{U_m} |\mathbf{n}(x) \cdot \nabla v(x)| \, dx$$

$$\leq c \left(\int_{U_m} 1 \, dx \right)^{\frac{1}{2}} \left(\int_{U_m} (\nabla v)^2 \, dx \right)^{\frac{1}{2}} \leq c\, h \, \| v \|_{1,U}.$$

A density argument yields the same bound for all $v \in W$. ∎

A direct consequence of the previous three lemmas is the following corollary.

COROLLARY 4.4. *The three terms in* (4.1) *can be bounded by*

$$\left| \int_\Gamma \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma (v - v^e_{\Gamma_h})\, ds \right| \leq c\, h\, \|v\|_{1,U}, \tag{4.5}$$

$$\left| \int_\Gamma (\mathbf{I} - \mathbf{A}^e_{\Gamma_h}) \nabla_\Gamma \mathrm{id}_\Gamma \cdot \nabla_\Gamma v^e_{\Gamma_h}\, ds \right| \leq c\, h^2\, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)}, \tag{4.6}$$

$$\left| \int_{\Gamma_h} \nabla_{\Gamma_h} (\mathrm{id}^e_\Gamma - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v\, ds \right| \leq c\, h\, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)}, \tag{4.7}$$

and thus

$$|g(v) - g_h(v)| \leq c\, h\, \|v\|_{1,U} + c\, h^2\, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} + c\, h\, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \quad \textit{for all} \;\; v \in W$$

holds.

In view of Corollary 4.4 and the error measure in (2.9), we want to derive a bound for $\|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)}$ in terms of $\|v\|_1$ for v from the scalar finite element space V_H. An obvious approach is to apply an inverse inequality combined with a trace theorem, resulting in

$$\|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \leq c\, h_{\min}^{-1} \|v\|_{L^2(\Gamma_h)} \leq c\, h_{\min}^{-1} \|v\|_1 \quad \text{for all} \;\; v \in V_H. \tag{4.8}$$

This, however, is too crude (cf. the bound in Corollary 4.4). To derive a better bound than the one in (4.8) we have to introduce some further assumptions related to the family of triangulations $\{\Gamma_h\}_{h>0}$. We assume that to each triangulation $\Gamma_h = \cup_{T \in \mathcal{F}_h} T$ there can be associated a set of tetrahedra $\mathcal{S}_h$ with the following properties:

(4.9) For each $T \in \mathcal{F}_h$ there is a corresponding $S_T \in \mathcal{S}_h$ with $T \subset S_T$.

(4.10) For $T_1, T_2 \in \mathcal{F}_h$ with $T_1 \neq T_2$ we have $\mathrm{meas}_3(S_{T_1} \cap S_{T_2}) = 0$.

(4.11) The family $\{\mathcal{S}_h\}_{h>0}$ is shape-regular.

(4.12) $c_0 h \leq \mathrm{diam}(S_T) \leq ch$ for all $T \in \mathcal{F}_h$ with $c_0 > 0$ (quasi-uniformity).

(4.13) For each $S_T \in \mathcal{S}_h$ there is a tetrahedron $S \in \mathcal{S}$ such that $S_T \subset S$.

Recall that $\mathcal{S}$ is the (fixed) tetrahedral triangulation that is used in the finite element discretization of the Navier–Stokes problem in (1.6). Note that the set of tetrahedra $\mathcal{S}_h$ has to be defined only close to the approximate interface Γ_h and that this set does not necessarily form a regular tetrahedral triangulation of Ω. Furthermore, it is *not* assumed that the family $\{\Gamma_h\}_{h>0}$is shape-regular or quasi-uniform.

Remark 6. Consider the construction of $\{\Gamma_h\}_{h>0}$ as in Remark 2. The approximate interface Γ_h is the zero level of the function $I(d_h)$, which is continuous piecewise linear on the tetrahedral triangulation $\mathcal{T}'$:

$$\Gamma_h = \cup_T T.$$

Each T is a triangle or a quadrilateral. To each T there can be associated a tetrahedron $S_T \in \mathcal{T}'$ such that $T \subset S_T$. If T is a quadrilateral, then we can subdivide T and

S_T in two disjoint triangles T_1, T_2 and two disjoint tetrahedra S_{T_1}, S_{T_2}, respectively, such that $T_i \subset S_{T_i} \subset S_T$ for $i = 1, 2$. One can check that this construction results in a family $\{\mathcal{S}_h\}_{h>0}$ that satisfies conditions (4.9)–(4.13).

In the following lemma we consider a standard affine mapping between a tetrahedron $S_T \in \mathcal{S}_h$ and the reference unit tetrahedron and apply it to the triangle $T \subset S_T$.

LEMMA 4.5. *Assume that the family $\{\Gamma_h\}_{h>0}$ is such that for the associated family of sets of tetrahedra $\{\mathcal{S}_h\}_{h>0}$ conditions* (4.9)–(4.13) *are satisfied. Take $T \in \mathcal{F}_h$ and the corresponding $S_T \in \mathcal{S}_h$. Let $\hat{S}$ be the reference unit tetrahedron and $F(x) = \mathbf{J}x + \mathbf{b}$ an affine mapping such that $F(\hat{S}) = S_T$. Define $\hat{T} := F^{-1}(T)$. The following holds:*

$$\|\mathbf{J}\|^2 \frac{\mathrm{meas}_3(\hat{S})}{\mathrm{meas}_3(S_T)} \leq c\, h^{-1}, \tag{4.14}$$

$$\|\mathbf{J}^{-1}\|^2 \frac{\mathrm{meas}_2(T)}{\mathrm{meas}_2(\hat{T})} \leq c, \tag{4.15}$$

with constants c independent of T and h.

Proof. Let $\rho(S_T)$ be the diameter of the maximal ball contained in S_T and similarly for $\rho(\hat{S})$. From standard finite element theory we have

$$\|\mathbf{J}\| \leq \frac{\mathrm{diam}(S_T)}{\rho(\hat{S})}, \quad \|\mathbf{J}^{-1}\| \leq \frac{\mathrm{diam}(\hat{S})}{\rho(S_T)}.$$

Using (4.11) and (4.12) we then get

$$\|\mathbf{J}\|^2 \frac{\mathrm{meas}_3(\hat{S})}{\mathrm{meas}_3(S_T)} \leq c\, \frac{\mathrm{diam}(S_T)^2}{\mathrm{meas}_3(S_T)} \leq c\ \mathrm{diam}(S_T)^{-1} \leq c\, h^{-1},$$

and thus the result in (4.14) holds.

The vertices of $\hat{T} = F^{-1}(T)$ are denoted by $\hat{V}_i$, $i = 1, 2, 3$. Let $\hat{V}_1\hat{V}_2$ be a longest edge of $\hat{T}$ and $\hat{M}$ the point on this edge such that $\hat{M}\hat{V}_3$ is perpendicular to $\hat{V}_1\hat{V}_2$. Define $V_i := F(\hat{V}_i)$, $i = 1, 2, 3$, and $M := F(\hat{M})$. Then V_i, $i = 1, 2, 3$, are the vertices of T and M lies on the edge V_1V_2. We then have

$$\mathrm{meas}_2(\hat{T}) = \frac{1}{2}\|\hat{V}_1 - \hat{V}_2\|\|\hat{V}_3 - \hat{M}\| = \frac{1}{2}\|\mathbf{J}^{-1}(V_1 - V_2)\|\|\mathbf{J}^{-1}(V_3 - M)\|$$

$$\geq \frac{1}{2}\|\mathbf{J}\|^{-2}\|V_1 - V_2\|\|V_3 - M\| \geq c\, \frac{\rho(\hat{S})^2}{\mathrm{diam}(S_T)^2}\, \mathrm{meas}_2(T),$$

with a constant $c > 0$. Thus we obtain

$$\|\mathbf{J}^{-1}\|^2 \frac{\mathrm{meas}_2(T)}{\mathrm{meas}_2(\hat{T})} \leq c\, \frac{\mathrm{diam}(\hat{S})^2}{\rho(S_T)^2} \frac{\mathrm{diam}(S_T)^2}{\rho(\hat{S})^2} \leq c,$$

which completes the proof. ∎

THEOREM 4.6. *Assume that the family $\{\Gamma_h\}_{h>0}$ is such that for the associated family of sets of tetrahedra $\{\mathcal{S}_h\}_{h>0}$ conditions* (4.9)–(4.13) *are satisfied. The following holds:*

$$\|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \leq c\, h^{-\frac{1}{2}} \|v\|_1 \quad \textit{for all} \quad v \in V_H.$$

Proof. Note that

$$\|\nabla_{\Gamma_h} v\|^2_{L^2(\Gamma_h)} = \sum_{T\in\mathcal{F}_h} \|\nabla_T v\|^2_{L^2(T)}.$$

Take $T \in \mathcal{F}_h$ and let S_T be the associated tetrahedron as explained above. Let $\hat{S}$ be the reference unit tetrahedron and $F : \hat{S} \to S_T$ as in Lemma 4.5. Define $\hat{v} := v \circ F$. Using standard transformation rules and Lemma 4.5 we get

$$\begin{aligned}
\|\nabla_T v\|^2_{L^2(T)} = \|\mathbf{P}_h \nabla v\|^2_{L^2(T)} &\le \|\nabla v\|^2_{L^2(T)} = \sum_{|\alpha|=1} \|\partial^\alpha v\|^2_{L^2(T)} \\
&\le c\,\|\mathbf{J}^{-1}\|^2 \sum_{|\alpha|=1} \|(\partial^\alpha \hat{v}) \circ F^{-1}\|^2_{L^2(T)} \\
&\le c\,\|\mathbf{J}^{-1}\|^2 \frac{\mathrm{meas}_2(T)}{\mathrm{meas}_2(\hat{T})} \sum_{|\alpha|=1} \|\partial^\alpha \hat{v}\|^2_{L^2(\hat{T})} \le c \sum_{|\alpha|=1} \|\partial^\alpha \hat{v}\|^2_{L^2(\hat{T})} \\
&\le c \sum_{|\alpha|=1} \max_{x\in\hat{T}} \left|\partial^\alpha \hat{v}(x)\right|^2 \le c \sum_{|\alpha|=1} \max_{x\in\hat{S}} \left|\partial^\alpha \hat{v}(x)\right|^2,
\end{aligned}$$

with a constant c independent of T. From (4.13) it follows that $\hat{v}$ is a polynomial on $\hat{S}$ of maximal degree k, where k depends only on the choice of the finite element space $\mathbf{V}_H$. On $P_k^* := \{\, p \in P_k \mid p(0) = 0 \,\}$ we have, due to equivalence of norms,

$$\sum_{|\alpha|=1} \max_{x\in\hat{S}} \left|\partial^\alpha \hat{v}(x)\right|^2 \le c \sum_{|\alpha|=1} \|\partial^\alpha \hat{v}\|^2_{L^2(\hat{S})} \quad \text{for all} \;\; \hat{v} \in P_k^*.$$

Because $\partial^\alpha \hat{v}$ is independent of $\hat{v}(0)$ for $\hat{v} \in P_k$ and $|\alpha| = 1$, the same inequality holds for all $\hat{v} \in P_k$. Thus we get

$$\begin{aligned}
\|\nabla_T v\|^2_{L^2(T)} &\le c \sum_{|\alpha|=1} \|\partial^\alpha \hat{v}\|^2_{L^2(\hat{S})} \le c\,\|\mathbf{J}\|^2 \sum_{|\alpha|=1} \|(\partial^\alpha v) \circ F\|^2_{L^2(\hat{S})} \\
&= c\,\|\mathbf{J}\|^2 \frac{\mathrm{meas}_3(\hat{S})}{\mathrm{meas}_3(S_T)} \sum_{|\alpha|=1} \|\partial^\alpha v\|^2_{L^2(S_T)} \le c\,h^{-1} \|\nabla v\|^2_{L^2(S_T)},
\end{aligned}$$

with a constant c independent of T and h. Using (4.10) we finally obtain

$$\begin{aligned}
\|\nabla_{\Gamma_h} v\|^2_{L^2(\Gamma_h)} &\le c\,h^{-1} \sum_{T\in\mathcal{F}_h} \|\nabla v\|^2_{L^2(S_T)} \\
&\le c\,h^{-1} \int_\Omega (\nabla v)^2\,dx \le c\,h^{-1} \|v\|^2_1,
\end{aligned}$$

which proves the result. ∎

We now present the main result of this paper.

THEOREM 4.7. *Let the assumptions be as in Theorem* 4.6. *The following holds:*

$$\sup_{v\in V_H} \frac{g(v) - g_h(v)}{\|v\|_1} \le c\sqrt{h}.$$

Proof. Combine the result in Corollary 4.4 with that in Theorem 4.6. ∎

As a direct consequence we obtain the following.

COROLLARY 4.8. *Let the assumptions be as in Theorem* 4.6. *For* f_Γ *and* f_{Γ_h} *as defined in section* 2 *the following holds:*

$$\sup_{\mathbf{v}\in\mathbf{V}_H} \frac{f_\Gamma(\mathbf{v}_H) - f_{\Gamma_h}(\mathbf{v}_H)}{\|\mathbf{v}_H\|_1} \leq \tau c \sqrt{h}.$$

Proof. Note that

$$f_\Gamma(\mathbf{v}_H) - f_{\Gamma_h}(\mathbf{v}_H)$$
$$= \tau \sum_{i=1}^{3} \left(\int_\Gamma \nabla_\Gamma(\mathrm{id}_\Gamma)_i \cdot \nabla_\Gamma(\mathbf{v}_H)_i \, ds - \int_{\Gamma_h} \nabla_{\Gamma_h}(\mathrm{id}_{\Gamma_h})_i \cdot \nabla_{\Gamma_h}(\mathbf{v}_H)_i \, ds \right),$$

and use the result in Theorem 4.7. ∎

An upper bound $\mathcal{O}(\sqrt{h})$ as in Corollary 4.8 for the error in the approximation of the localized force term may seem rather pessimistic, because Γ_h is an $\mathcal{O}(h^2)$ accurate approximation of Γ. Numerical experiments in section 6, however, indicate that the bound is sharp.

5. Improved approximation of the localized force term $f_\Gamma(\mathbf{v}_h)$. In this section we show how the approximation of the localized force term can be improved, resulting in an improved error bound of the form $\mathcal{O}(h)$ (instead of $\mathcal{O}(\sqrt{h})$). From Corollary 4.4 and Theorem 4.6 we see that the $\sqrt{h}$ behavior is caused by the estimate in (4.7):

$$\left| \int_{\Gamma_h} \nabla_{\Gamma_h}(\mathrm{id}_\Gamma^e - \mathrm{id}_{\Gamma_h}) \cdot \nabla_{\Gamma_h} v \, ds \right| \leq c\, h \, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)}. \tag{5.1}$$

The term $\nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h}$ that is used in $g_h(v)$ occurs in (5.1) but not in the other two terms of the splitting; cf. (4.5), (4.6). We consider

$$\tilde{g}_h(v) = \int_{\Gamma_h} m_h \cdot \nabla_{\Gamma_h} v \, ds$$

and try to find a function $m_h = m_h(x)$ such that $\tilde{g}_h(v)$ remains easily computable and the bound in (5.1) is improved if we use m_h instead of $\nabla_{\Gamma_h} \mathrm{id}_{\Gamma_h}$. The latter condition is trivially satisfied for $m_h = \nabla_{\Gamma_h} \mathrm{id}_\Gamma^e$ (leading to a bound 0 in (5.1)). This choice, however, does not satisfy the first condition, because Γ is not known. We now discuss another possibility that is used in the experiments in section 6.

Due to $|d(x)| \leq ch^2$ we get from Lemma 3.1, for $x \in \Gamma_h$:

$$\nabla_{\Gamma_h} \mathrm{id}_\Gamma^e(x) = \mathbf{P}_h(x)\mathbf{P}(x)\nabla_\Gamma \mathrm{id}_\Gamma(\mathbf{p}(x)) + \mathcal{O}(h^2) = \mathbf{P}_h(x)\mathbf{P}(x)e_i + \mathcal{O}(h^2).$$

In the construction of the interface Γ_h (cf. Remark 2), we have available a piecewise *quadratic* function $d_h \approx d$. Define

$$\tilde{\mathbf{n}}_h(x) := \frac{\nabla d_h(x)}{\|\nabla d_h(x)\|}, \qquad \tilde{\mathbf{P}}_h(x); = \mathbf{I} - \tilde{\mathbf{n}}_h(x)\tilde{\mathbf{n}}_h(x)^T, \quad x \in \Gamma_h.$$

Thus an obvious modification is based on the choice $m_h(x) = \mathbf{P}_h(x)\tilde{\mathbf{P}}_h(x)e_i$, i.e.,

$$\tilde{g}_h(v) := \int_{\Gamma_h} \mathbf{P}_h(x)\tilde{\mathbf{P}}_h(x)e_i \cdot \nabla_{\Gamma_h} v \, ds = \int_{\Gamma_h} \tilde{\mathbf{P}}_h(x)e_i \cdot \nabla_{\Gamma_h} v \, ds. \tag{5.2}$$

In this approach the approximate interface Γ_h is *not* changed (piecewise planar). For piecewise quadratics d_h and v, the function $\nabla_{\Gamma_h} v = \mathbf{P}_h \nabla v$ is piecewise linear and $\tilde{\mathbf{P}}_h e_i$ is piecewise (very) smooth on the segments of Γ_h. Hence, the functional in (5.2) can be evaluated easily.

Under reasonable assumptions the modified functional indeed yields a better error bound, as the following lemma shows.

LEMMA 5.1. *Assume that there exists $p > 0$ such that*

$$\|\nabla d_h(x) - \nabla d(x)\| \leq c\, h^p \quad \textit{for} \;\; x \in \Gamma_h. \tag{5.3}$$

Then the following holds:

$$\left| \int_{\Gamma_h} \left(\nabla_{\Gamma_h} \mathrm{id}_\Gamma^e - \mathbf{P}_h \tilde{\mathbf{P}}_h e_i \right) \cdot \nabla_{\Gamma_h} v \, ds \right| \leq c\, h^{\min\{p,2\}} \, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \quad \textit{for all} \;\; v \in W.$$

Proof. Using $\|\nabla d\| = 1$ it follows that $\|\nabla d_h\| = 1 + \mathcal{O}(h^p)$ holds. We can use the same line of reasoning as in the proof of Lemma 4.2. The term in (4.4) remains the same. Instead of the term in (4.3) we now get $\|\mathbf{P}_h(x)\mathbf{P}(x) - \mathbf{P}_h(x)\tilde{\mathbf{P}}_h(x)\|$. We drop x in the notation, and using the assumption we obtain

$$\begin{aligned}
\|\mathbf{P}_h\mathbf{P} - \mathbf{P}_h\tilde{\mathbf{P}}_h\| = \|\mathbf{P}_h(\mathbf{P} - \tilde{\mathbf{P}}_h)\| &\leq \|\mathbf{n}\mathbf{n}^T - \tilde{\mathbf{n}}_h\tilde{\mathbf{n}}_h^T\| \\
&\leq \|(\mathbf{n} - \tilde{\mathbf{n}}_h)\mathbf{n}^T\| + \|\tilde{\mathbf{n}}_h(\mathbf{n} - \tilde{\mathbf{n}}_h)^T\| = 2\|\mathbf{n} - \tilde{\mathbf{n}}_h\| \\
&= 2 \left\| \nabla d - \frac{\nabla d_h}{\|\nabla d_h\|} \right\| \\
&\leq 2\big|1 - \|\nabla d_h\|^{-1}\big| \|\nabla d_h\| + 2\|\nabla d - \nabla d_h\| \leq c\, h^p.
\end{aligned}$$

Thus we get an estimate $\|\mathbf{P}_h\mathbf{P} - \mathbf{P}_h\tilde{\mathbf{P}}_h\| \leq c\, h^p$. Combined with the inequality $|d(x)|\|\mathbf{P}_h(x)\mathbf{H}(x)\mathbf{P}(x)\| \leq c\, h^2$ for the term in (4.4) this proves the result. ∎

If we assume that the condition in (5.3) is satisfied for $p = 2$, which is reasonable for a piecewise quadratic approximation d_h of d, we get the following improvement due to the modified functional $\tilde{g}_h$ (cf. Corollary 4.4):

$$|g(v) - \tilde{g}_h(v)| \leq c\, h \, \|v\|_{1,U} + c\, h^2 \, \|\nabla_{\Gamma_h} v\|_{L^2(\Gamma_h)} \quad \text{for all} \;\; v \in W.$$

Combining this with the result in Theorem 4.6 yields (under the assumption as in Theorem 4.6)

$$|g(v) - \tilde{g}_h(v)| \leq c\, h \, \|v\|_{1,U} + c\, h^{\frac{3}{2}} \, \|v\|_1 \quad \text{for all} \;\; v \in V_H.$$

Hence, using this modified functional $\tilde{g}_h$ we have an $\mathcal{O}(h)$ error bound. This significant improvement (compared to the $\mathcal{O}(\sqrt{h})$ error bound for the functional g_h) is confirmed by the numerical experiments in the next section.

6. Numerical experiments. In this section we present results of a numerical experiment which indicates that the $\mathcal{O}(\sqrt{h})$ bound in Corollary 4.8 is sharp. Furthermore, for the improved approximation described in section 5 the $\mathcal{O}(h)$ bound will be confirmed numerically.

We consider the domain $\Omega := [-1, 1]^3$, where the ball $\Omega_1 := \{\mathbf{x} \in \Omega \mid \|\mathbf{x}\| < R\}$ is located in the center of the domain. In our experiments we take $R = \frac{1}{2}$.

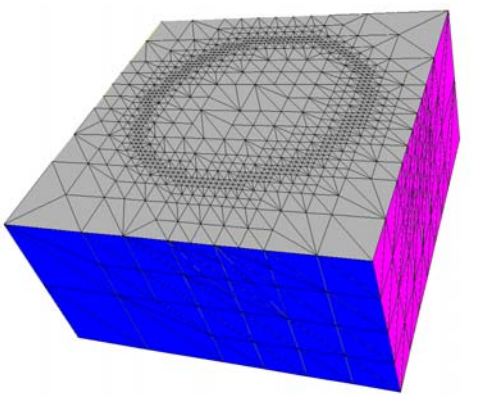


FIG. 6.1. *Lower half of the four times refined mesh* $\mathcal{T}_4$.

For the discretization a uniform tetrahedral mesh $\mathcal{T}_0$ is used, where the vertices form a $6 \times 6 \times 6$ lattice; hence $h_0 = \frac{1}{5}$. This coarse mesh $\mathcal{T}_0$ is locally refined in the vicinity of $\Gamma = \partial\Omega_1$ using an adaptive refinement algorithm presented in [15]. This repeated refinement process yields the gradually refined meshes $\mathcal{T}_1, \mathcal{T}_2, \ldots$ with *local* (i.e., close to the interface) mesh sizes $h_i = \frac{1}{5} \cdot 2^{-i}, i = 1, 2, \ldots$. Part of the tetrahedral triangulation $\mathcal{T}_4$ is shown in Figure 6.1. The corresponding finite element spaces $\mathbf{V}_i := \mathbf{V}_{h_i} = (V_{h_i})^3$ consist of vector functions where each component is a continuous piecewise quadratic function on $\mathcal{T}_i$.

The interface $\Gamma = \partial\Omega_1$ is a sphere, and thus the curvature $\mathcal{K} = \frac{2}{R}$ is constant. If we discretize the flow problem using $\mathbf{V}_i$ as a discrete velocity space, we have to approximate the surface tension force

$$f_\Gamma(\mathbf{v}) = \frac{2\tau}{R} \int_\Gamma \mathbf{n}_\Gamma \cdot \mathbf{v}\, ds = \tau \int_\Gamma \nabla_\Gamma \operatorname{id}_\Gamma \cdot \nabla_\Gamma \mathbf{v}\, ds, \qquad \mathbf{v} \in \mathbf{V}_i. \tag{6.1}$$

To simplify notation, we take a fixed $i \geq 0$, and the corresponding local mesh size parameter is denoted by $h = h_i$. For the approximation of the interface we use the following approach (cf. Remark 2). The interface Γ is the zero level of the signed distance function d. In this test problem, d is known. For the finite element approximation $d_h \in V_h$ of d we take the continuous piecewise *quadratic* function on $\mathcal{T}_i$ that interpolates d at the vertices and midpoints of edges. Then $I(d_h)$ is the continuous piecewise *linear* function on $\mathcal{T}_i'$ that interpolates d_h at the vertices of all tetrahedra in $\mathcal{T}_i'$; cf. Remark 2 (note that in this test problem, d_h also can be computed by piecewise linear interpolation of d on $\mathcal{T}_i'$). The approximation of Γ is defined by

$$\Gamma_h = \{\, x \in \Omega \mid I(d_h)(x) = 0 \,\}$$

and is illustrated in Figure 2.2. The discrete approximation of the surface tension force is

$$f_{\Gamma_h}(\mathbf{v}) = \tau \int_{\Gamma_h} \nabla_{\Gamma_h} \operatorname{id}_{\Gamma_h} \cdot \nabla_{\Gamma_h} \mathbf{v}\, ds, \qquad \mathbf{v} \in \mathbf{V}_i.$$

We are interested in (cf. Corollary 4.8)

$$\|f_\Gamma - f_{\Gamma_h}\|_{\mathbf{V}_i'} := \sup_{\mathbf{v} \in \mathbf{V}_i} \frac{f_\Gamma(\mathbf{v}) - f_{\Gamma_h}(\mathbf{v})}{\|\mathbf{v}\|_1}. \tag{6.2}$$

The evaluation of $f_\Gamma(\mathbf{v})$, for $\mathbf{v} \in \mathbf{V}_i$, requires the computation of integrals on curved triangles or quadrilaterals $\Gamma \cap S$, where S is a tetrahedron from the mesh $\mathcal{T}_i$. We are not able to compute these exactly. Therefore, we introduce an artificial force term which, in this model problem with a known constant curvature, is computable and sufficiently close to f_Γ.

LEMMA 6.1. *For* $\mathbf{v} \in \mathbf{V} = (H_0^1(\Omega))^3$ *define*

$$\hat{f}_{\Gamma_h}(\mathbf{v}) := \frac{2\tau}{R} \int_{\Gamma_h} \mathbf{n}_h \cdot \mathbf{v}\, ds$$

($\mathbf{n}_h$ *denotes the piecewise constant outward unit normal on* Γ_h*). Then the following inequality holds:*

$$\|f_\Gamma - \hat{f}_{\Gamma_h}\|_{\mathbf{V}'} \le ch. \tag{6.3}$$

Proof. Let $\Omega_{1,h} \subset \Omega$ be the domain enclosed by Γ_h, i.e., $\partial\Omega_{1,h} = \Gamma_h$. We define $D_h^+ := \Omega_1 \setminus \Omega_{1,h}$, $D_h^- := \Omega_{1,h} \setminus \Omega_1$, and $D_h := D_h^+ \cup D_h^-$. Due to the Stokes theorem, for $\mathbf{v} \in \mathbf{V}$ we have

$$|f_\Gamma(\mathbf{v}) - \hat{f}_{\Gamma_h}(\mathbf{v})| = \frac{2\tau}{R} \left| \int_{\Omega_1} \operatorname{div} \mathbf{v}\, dx - \int_{\Omega_{1,h}} \operatorname{div} \mathbf{v}\, dx \right| \tag{6.4}$$

$$= \frac{2\tau}{R} \left| \int_{D_h^+} \operatorname{div} \mathbf{v}\, d\mathbf{x} - \int_{D_h^-} \operatorname{div} \mathbf{v}\, dx \right| \tag{6.5}$$

$$\le \frac{2\tau}{R} \int_{D_h} |\operatorname{div} \mathbf{v}|\, dx. \tag{6.6}$$

Using the Cauchy–Schwarz inequality, we get the estimate

$$|f_\Gamma(\mathbf{v}) - \hat{f}_{\Gamma_h}(\mathbf{v})| \le c\sqrt{|D_h|}\|\mathbf{v}\|_1 \quad \text{for all} \quad \mathbf{v} \in \mathbf{V}.$$

For the piecewise planar approximation Γ_h of the interface Γ we have $|D_h| = \mathcal{O}(h^2)$, and thus (6.3) holds. ∎

From Lemma 6.1 we obtain $\|f_\Gamma - \hat{f}_{\Gamma_h}\|_{\mathbf{V}_j'} \le c\,h$ with a constant c independent of j. Thus we have

$$\|\hat{f}_{\Gamma_h} - f_{\Gamma_h}\|_{\mathbf{V}_i'} - ch \le \|f_\Gamma - f_{\Gamma_h}\|_{\mathbf{V}_i'} \le \|\hat{f}_{\Gamma_h} - f_{\Gamma_h}\|_{\mathbf{V}_i'} + ch. \tag{6.7}$$

The term $\|\hat{f}_{\Gamma_h} - f_{\Gamma_h}\|_{\mathbf{V}_i'}$ can be evaluated as follows. Since Γ_h is piecewise planar and $\mathbf{v} \in \mathbf{V}_i$ is a piecewise quadratic function, for $\mathbf{v} \in \mathbf{V}_i$, both $\hat{f}_{\Gamma_h}(\mathbf{v})$ and $f_{\Gamma_h}(\mathbf{v})$ can be computed exactly (up to machine accuracy) using suitable quadrature rules. For the evaluation of the dual norm $\|\cdot\|_{\mathbf{V}_i'}$ we proceed as follows. Let $\{\boldsymbol{\phi}_j\}_{j=1,\dots,n}$ be the standard nodal basis in $\mathbf{V}_i$ and $J : \mathbb{R}^n \to \mathbf{V}_i$ the isomorphism $J\vec{x} = \sum_{k=1}^n x_k \boldsymbol{\phi}_k$. Let M_h be the mass matrix and A_h the stiffness matrix of the Laplacian:

$$\begin{aligned} (M_h)_{i,j} &:= \int_\Omega \boldsymbol{\phi}_i \cdot \boldsymbol{\phi}_j\, dx, \\ (A_h)_{i,j} &:= \int_\Omega \nabla\boldsymbol{\phi}_i \cdot \nabla\boldsymbol{\phi}_j\, dx. \end{aligned} \qquad 1 \le i, j \le n.$$

TABLE 6.1
Error norms and numerical order of convergence for different refinement levels.

i	$\|\hat{f}_{\Gamma_h} - f_{\Gamma_h}\|_{\mathbf{V}_i'}$	order	$\|\hat{f}_{\Gamma_h} - \tilde{f}_{\Gamma_h}\|_{\mathbf{V}_i'}$	order
0	1.79 E-1	–	1.32 E-1	–
1	1.40 E-1	0.35	4.43 E-2	1.57
2	1.03 E-1	0.45	1.46 E-2	1.61
3	7.22 E-2	0.51	5.06 E-3	1.52
4	5.02 E-2	0.53	1.78 E-3	1.51

Define $C_h = A_h + M_h$. Note that for $\mathbf{v} = J\vec{x} \in \mathbf{V}_i$ we have $\|\mathbf{v}\|_1^2 = \langle C_h\vec{x}, \vec{x}\rangle$. Take $e \in \mathbf{V}_i'$ and define $\vec{e} \in \mathbb{R}^n$ by $e_j := e(\boldsymbol{\phi}_j)$, $j = 1, \ldots, n$. Due to

$$\|e\|_{\mathbf{V}_i'} = \sup_{\mathbf{v}\in\mathbf{V}_i} \frac{|e(\mathbf{v})|}{\|\mathbf{v}\|_1} = \sup_{\vec{x}\in\mathbb{R}^n} \frac{|\sum_{j=1}^n x_j e(\boldsymbol{\phi}_j)|}{\sqrt{\langle C_h\vec{x}, \vec{x}\rangle}}$$

we obtain

$$\|e\|_{\mathbf{V}_i'} = \sup_{\vec{x}\in\mathbb{R}^n} \frac{\langle \vec{x}, \vec{e}\rangle}{\sqrt{\langle C_h\vec{x}, \vec{x}\rangle}} = \|C_h^{-1/2}\vec{e}\| = \sqrt{\langle C_h^{-1}\vec{e}, \vec{e}\rangle}. \tag{6.8}$$

Thus for the computation of $\|e\|_{\mathbf{V}_i'}$ we proceed in the following way:

1. Compute $\vec{e} = \big(e(\boldsymbol{\phi}_j)\big)_{j=1}^n$.
2. Solve the linear system $C_h\vec{z} = \vec{e}$ up to machine accuracy.
3. Compute $\|e\|_{\mathbf{V}_i'} = \sqrt{\langle \vec{z}, \vec{e}\rangle}$.

We applied this strategy to $e := \hat{f}_{\Gamma_h} - f_{\Gamma_h}$. The results are given in the second column in Table 6.1. The numerical order of convergence in the third column of this table clearly indicates an $\mathcal{O}(\sqrt{h})$ behavior. Due to (6.7) this implies the same $\mathcal{O}(\sqrt{h})$ convergence behavior for $\|f_\Gamma - f_{\Gamma_h}\|_{\mathbf{V}_i'}$. This indicates that the $\mathcal{O}(\sqrt{h})$ bound in Corollary 4.8 is sharp.

The same procedure can be applied with f_{Γ_h} replaced by the modified (improved) approximate surface tension force

$$\tilde{f}_{\Gamma_h}(\mathbf{v}) = \tau \sum_{i=1}^3 \tilde{g}_{h,i}(v_i)$$

with $\tilde{g}_{h,i}$ as defined in (5.2). This yields the results in the fourth column in Table 6.1. For this modification the numerical order of convergence is significantly better, namely, at least first order in h. From (6.7) it follows that for $\|f_\Gamma - \tilde{f}_{\Gamma_h}\|_{\mathbf{V}_i'}$ we can expect $\mathcal{O}(h^p)$ with $p \geq 1$.

Summarizing, we conclude that the results of these numerical experiments confirm the theoretical $\mathcal{O}(\sqrt{h})$ error bound derived in the analysis in section 4 and show that the modified approximation indeed leads to (much) better results.

Results of numerical experiments for a Stokes two-phase flow problem using both f_{Γ_h} and $\tilde{f}_{\Gamma_h}$ are presented in [16].

REFERENCES

[1] S. N. ANTONTSEV, A. M. MEIRMANOV, AND V. A. SOLONNIKOV, *Smooth interface in a two-component Stokes flow*, Ann. Univ. Ferrara Sez. VII (N.S.), 47 (2001), pp. 269–284.

[2] S. N. Antontsev, A. M. Meirmanov, and B. V. Yurinsky, *A free-boundary problem for Stokes equations: Classical solutions*, Interfaces Free Bound., 2 (2000), pp. 413–424.
[3] E. Bänsch, *Numerical Methods for the Instationary Navier-Stokes Equations with a Free Capillary Surface*, Habilitation thesis, University of Freiburg, Freiburg, Germany, 1998.
[4] E. Bänsch, *Finite element discretization of the Navier-Stokes equations with a free capillary surface*, Numer. Math., 88 (2001), pp. 203–235.
[5] J. U. Brackbill, D. B. Kothe, and C. Zemach, *A continuum method for modeling surface tension*, J. Comput. Phys., 100 (1992), pp. 335–354.
[6] Y. C. Chang, T. Y. Hou, B. Merriman, and S. Osher, *A level set formulation of Eulerian interface capturing methods for incompressible fluid flows*, J. Comput. Phys., 124 (1996), pp. 449–464.
[7] K. Deckelnick and G. Dziuk, *Mean curvature flow and related topics*, in Frontiers in Numerical Analysis (Durham 2002), J. F. Blowey, A. W. Craig, and T. Shardlow, eds., Springer, Berlin, 2003, pp. 63–108.
[8] A. Demlow and G. Dziuk, *An adaptive finite element method for the Laplace–Beltrami operator on implicitly defined surfaces*, SIAM J. Numer. Anal., 45 (2007), pp. 421–442.
[9] *DROPS package*, http://www.igpm.rwth-aachen.de/DROPS/ (2006).
[10] G. Dziuk, *An algorithm for evolutionary surfaces*, Numer. Math., 58 (1991), pp. 603–611.
[11] S. Ganesan, G. Matthies, and L. Tobiska, *On Spurious Velocities in Incompressible Flow Problems with Interfaces*, Preprint 05-35, Department of Mathematics, University of Magdeburg, Magdeburg, Germany, 2005.
[12] S. Ganesan and L. Tobiska, *Finite element simulation of a droplet impinging a horizontal surface*, in Proceedings of ALGORITMY 2005, pp. 1–11. Available online at http://pc2iam.fmph.uniba.sk/amuc/_contributed/algo2005/ganesan-tobiska.pdf
[13] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer, Berlin, 1986.
[14] S. Gross, V. Reichelt, and A. Reusken, *A finite element based level set method for two-phase incompressible flows*, Comput. Vis. Sci., 9 (2006), pp. 239–257.
[15] S. Gross and A. Reusken, *Parallel multilevel tetrahedral grid refinement*, SIAM J. Sci. Comput., 26 (2005), pp. 1261–1288.
[16] S. Gross and A. Reusken, *An extended pressure finite element space for two-phase incompressible flows with surface tension*, J. Comput. Phys., 224 (2007), pp. 40–58.
[17] S. Hysing, *A new implicit surface tension implementation for interfacial flows*, Internat. J. Numer. Methods Fluids, 51 (2006), pp. 659–672.
[18] M. A. Olshanskii, J. Peters, and A. Reusken, *Uniform preconditioners for a parameter dependent saddle point problem with application to generalized Stokes interface equations*, Numer. Math., 105 (2006), pp. 159–191.
[19] M. A. Olshanskii and A. Reusken, *A Stokes interface problem: Stability, finite element analysis and a robust solver*, in Proceedings of the European Congress on Computational Methods in Applied Sciences and Engineering (ECCOMAS 2004), P. Neittaanmäki et al., eds., University of Jyväskylä, Jyväskylä, Finland. Available online at http://www.mit.jyu.fi/eccomas2004/proceedings/pdf/344.pdf, 2004.
[20] M. A. Olshanskii and A. Reusken, *Analysis of a Stokes interface problem*, Numer. Math., 103 (2006), pp. 129–149.
[21] S. Osher and R. P. Fedkiw, *Level set methods: An overview and some recent results*, J. Comput. Phys., 169 (2001), pp. 463–502.
[22] S. P. van der Pijl, A. Segal, C. Vuik, and P. Wesseling, *A mass-conserving level-set method for modelling of multi-phase flows*, Internat. J. Numer. Methods Fluids, 47 (2005), pp. 339–361.
[23] S. B. Pillapakkam and P. Singh, *A level-set method for computing solutions to viscoelastic two-phase flow*, J. Comput. Phys., 174 (2001), pp. 552–578.
[24] J. A. Sethian, *Level Set Methods and Fast Marching Methods*, Cambridge University Press, Cambridge, UK, 1999.
[25] M. Sussman, A. S. Almgren, J. B. Bell, Ph. Colella, L. H. Howell, and M. L. Welcome, *An adaptive level set approach for incompressible two-phase flows*, J. Comput. Phys., 148 (1999), pp. 81–124.
[26] M. Sussman, P. Smereka, and S. Osher, *A level set approach for computing solutions to incompressible two-phase flow*, J. Comput. Phys., 114 (1994), pp. 146–159.
[27] A.-K. Tornberg, *Interface Tracking Methods with Application to Multiphase Flows*, Doctoral thesis, Department of Numerical Analysis and Computing Science, Royal Institute of Technology, Stockholm, 2000.

[28] A.-K. Tornberg and B. Engquist, *A finite element based level-set method for multiphase flow applications*, Comput. Vis. Sci., 3 (2000), pp. 93–101.

[29] X. Yang, A. J. James, J. Lowengrub, X. Zheng, and V. Cristini, *An adaptive coupled level-set/volume-of-fluid interface capturing method for unstructured triangular grids*, J. Comput. Phys., 217 (2006), pp. 364–394.

[30] X. Zheng, A. Anderson, J. Lowengrub, and V. Cristini, *Adaptive unstructured volume remeshing*. II. *Application to two- and three-dimensional level-set simulations of multiphase flow*, J. Comput. Phys., 208 (2005), pp. 626–650.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1701–1722

CAN WE HAVE SUPERCONVERGENT GRADIENT RECOVERY UNDER ADAPTIVE MESHES?*

HAIJUN WU† AND ZHIMIN ZHANG‡

Abstract. We study adaptive finite element methods for elliptic problems with domain corner singularities. Our model problem is the two-dimensional Poisson equation. Results of this paper are twofold. First, we prove that there exists an adaptive mesh (gauged by a discrete mesh density function) under which the recovered gradient by the polynomial preserving recovery (PPR) is superconvergent. Second, we demonstrate by numerical examples that an adaptive procedure with an a posteriori error estimator based on PPR does produce adaptive meshes that asatisfy our mesh density assumption, and the recovered gradient by PPR is indeed superconvergent in the adaptive process.

Key words. finite element method, adaptive, superconvergence, gradient recovery

AMS subject classifications. 65N30, 65N15, 45K20

DOI. 10.1137/060661430

1. Introduction.

Let $\Omega \subset R^2$ be a bounded polygon with boundary $\partial\Omega$. Consider the following Dirichlet boundary problem: Find $u \in H^1(\Omega)$ such that $u = g$ on $\partial\Omega$ and

$$A(u,v) = \int_\Omega \nabla u \cdot \nabla v = f(v) \quad \forall v \in H_0^1(\Omega), \tag{1.1}$$

where $f \in H^{-1}(\Omega)$.

It is well known that the solution u may have singularities at corners of Ω. Since the treatment of multiple singular points is no different from a simple one, without loss of generality we assume that the solution u has a singularity at the origin O and can be decomposed as a sum of a singular part and a smooth part:

$$u = v + w, \tag{1.2}$$

where

$$\left|\frac{\partial^m v}{\partial x^i \partial y^{m-i}}\right| \lesssim r^{\delta - m} \text{ and } \left|\frac{\partial^m w}{\partial x^i \partial y^{m-i}}\right| \lesssim 1, \quad m = 1, \ldots, k+2,\ i = 0, \ldots, m, \tag{1.3}$$

where $r = \sqrt{x^2 + y^2}$ and $0 < \delta < k + 1$ is a constant. Here $k = 1$ for linear finite element methods and $k = 2$ for quadratic finite element methods.

Next, we briefly explain the rationale of the above regularity assumption. When Ω is a polygonal domain, the solution of the Poisson equation with the Dirichlet

*Received by the editors May 30, 2006; accepted for publication (in revised form) January 16, 2007; published electronically August 22, 2007.
http://www.siam.org/journals/sinum/45-4/66143.html

†Department of Mathematics, Nanjing University, Jiangsu, 210093, People's Republic of China, and Department of Mathematics, Wayne State University, Detroit, MI 48202 (hjw@nju.edu.cn). This author was supported in part by China NSF under grant 10401016 and by the National Basic Research Program under grant 2005CB321701.

‡Department of Mathematics, Wayne State University, Detroit, MI 48202 (zzhang@math.wayne.edu). This author was supported in part by U.S. National Science Foundation grants DMS-0311807 and DMS-0622908.

boundary condition,

$$-\Delta u = f \quad \text{in} \quad \Omega, \qquad u|_{\partial\Omega} = g,$$

with sufficiently smooth data f and g, has the following decomposition (see, e.g., [3] and [11]), at a corner with angle ω:

$$u(r,\theta) = \sum_{j=1}^{J} c_j r^{\alpha_j} \ln^{s_j} r \sin \alpha_j \theta + w, \quad \alpha_j = \frac{j\pi}{\omega},$$

where w is smoother than the terms in the sum, and

$$s_j = \begin{cases} 1 & \alpha_j \text{ is an integer,} \\ 0 & \text{otherwise.} \end{cases}$$

Especially for the L-shaped domain, $\omega = 3\pi/2$ at the re-entrance corner, and the expansion is

$$u = c_1 r^{2/3} \sin \frac{2}{3}\theta + c_2 r^{4/3} \sin \frac{4}{3}\theta + c_3 r^2 \ln r \sin 2\theta + c_4 r^{8/3} \sin \frac{8}{3}\theta + w,$$

with $w \in W^3_\infty(\Omega)$. For a cracked domain, $\omega = 2\pi$ at the crack tip and the expansion is

$$\begin{aligned} u = {} & c_1 r^{1/2} \sin \frac{1}{2}\theta + c_2 2r \ln r \sin \theta + c_3 r^{3/2} \sin \frac{3}{2}\theta \\ & + c_4 r^2 \ln r \sin 2\theta + c_5 r^{5/2} \sin \frac{5}{2}\theta + c_6 r^3 \ln r \sin 3\theta + w, \end{aligned}$$

with $w \in W^3_\infty(\Omega)$. More terms are needed in the expansion if we want higher regularity on w. These are the two cases we shall test numerically in the last section.

Let $\mathcal{M}_h$ be a regular triangulation of the domain Ω, $\mathcal{E}_h$ be the set of all interior edges, and $\mathcal{N}_h$ be the set of all nodal points. Assume that the origin $O \in \mathcal{N}_h$. Remember that any triangle $\tau \in \mathcal{M}_h$ is considered to be closed. Let

$$V_h^k = \left\{ v_h : \ v_h \in H^1(\Omega), v_h|_\tau \in P_k(\tau) \right\}, \quad k = 1, 2,$$

be the conforming finite element space associated with $\mathcal{M}_h$, and let $\overset{\circ}{V}{}_h^k = V_h^k \bigcap H_0^1(\Omega)$. Here P_k denotes the set of polynomials with degree $\leq k$. Denote by $I_h^k : C(\bar{\Omega}) \to V_h^k$ the standard finite element interpolation operator. The finite element solution $u_h \in V_h^k$ satisfies $u_h = I_h^k u$ on $\partial\Omega$ and

$$A(u_h, v_h) = \int_\Omega \nabla u_h \cdot \nabla v_h = f(v_h) \quad \forall v_h \in \overset{\circ}{V}{}_h^k. \tag{1.4}$$

In adaptive finite element methods, the convergence rate is measured by the total number of degrees of freedom N, since the mesh is not quasi-uniform. For a two-dimensional second-order elliptic equation, the optimal convergence rates are

$$\|\nabla(u - u_h)\|_{L^2(\Omega)} \lesssim \begin{cases} N^{-1/2}, & k = 1, \\ N^{-1}, & k = 2, \end{cases} \tag{1.5}$$

where $k = 1$ for the linear element and $k = 2$ for the quadratic.

The theoretical development of residual-type error estimates is now in its maturity. For the early literature, readers are referred to [1, 4, 10, 21] and references therein. Starting from the fundamental work of [9], in the last decade the convergence proof of residual-based adaptive finite element method has been well established; see, e.g., [2, 8, 17, 19]. On the contrary, there is no convergence proof for using recovery-based error estimators. Nevertheless, by shifting the error estimator from residual based to recovery based, we have obtained the same numerical convergence rate following the same mark-up and refinement procedure for two model problems—the Poisson equation on the L-shaped domain and cracked square. Theoretically, we are able to prove that there exists an adaptive mesh satisfying a discrete mesh density condition such that the convergence rate (1.5) can be established. Moreover, under the same mesh density condition, the recovered gradient $G_h u_h$ is superconvergent in the sense that

$$\|\nabla u - G_h u_h\|_{L^2(\Omega)} \lesssim \begin{cases} N^{-1/2-\rho}, & k = 1, \\ N^{-1-\rho}, & k = 2, \end{cases} \tag{1.6}$$

where $\rho > 0$ is a constant, which depends on the quality of the adaptive mesh, and $G_h : V_h^k \to V_h^k \times V_h^k$ is the recovery operator. Now the question is: Is the condition required by our theory practical? We demonstrate that the meshes generated by the standard adaptive procedure in both of our model problems indeed satisfies the mesh density condition.

In recent years there have been some superconvergence results for a recovered gradient [5, 15, 8, 16, 20, 23, 24, 25, 26, 28]. All of them assumed at least $u \in H^3(\Omega) \cap W_\infty^2(\Omega)$ (a condition that rules out domains with a re-entrant corner) and required some stronger (than we required here) mesh conditions. Our current work fills in this gap. To the best of our knowledge, this is the first theoretical superconvergence proof for real-life adaptive meshes.

Some further theoretical results about recovery techniques and recovery-type error estimators can be found in [1, 7, 22, 13].

Based on the estimate (1.6), we suggest that, even for residual-type adaptive method, a gradient recovery procedure at the very last mesh would dramatically improve the numerical gradient.

Throughout the paper, we use the notation $A \lesssim B$ to represent the inequality $A \leq \textit{constant} \times B$, where the *constant* may depend only on the minimum angle of the triangles in the mesh $\mathcal{M}_h$, the constant δ, and the domain Ω. The notation $A \eqsim B$ is equivalent to the statement $A \lesssim B$ and $B \lesssim A$.

2. Preliminaries. Following the discussion in [8], we consider in Figure 2.1 an edge e, two elements τ and τ' sharing e, and $\Omega_e = \tau \cup \tau'$ the patch of e. For an element $\tau \subset \Omega_e$, θ_e denotes the angle opposite of the edge e, h_e, h_{e+1}, and h_{e-1} denote the lengths of the three edges of τ. The subscript $e+1$ or $e-1$ is for orientation. All triangles in the triangulation are orientated counterclockwise. $\boldsymbol{t}_e$ is the unit tangent vector of e with counterclockwise orientation and $\boldsymbol{n}_e$ is the unit outward normal vector. An index $'$ is added for the corresponding quantities in τ'. Notice that $\boldsymbol{t}_e = -\boldsymbol{t}'_e$ and $\boldsymbol{n}_e = -\boldsymbol{n}'_e$ because of the orientation. For any $\tau \in \mathcal{M}_h$, we denote by h_τ its diameter and by r_τ the distance from the origin to the barycenter of τ, and by $|\tau|$ the area of the triangle τ. For any $e \in \mathcal{E}_h$, let r_e be the distance from the origin O to the midpoint of e.

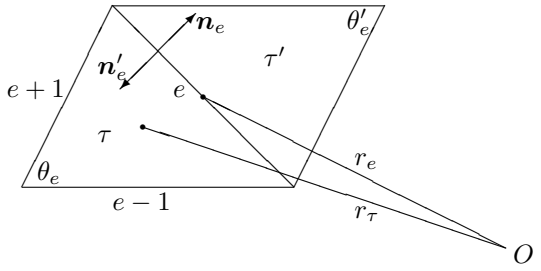


FIG. 2.1. *Notation in the patch* Ω_e.

Let $e \in \mathcal{E}_h$ be an interior edge. Recall that Ω_e, the patch of e, consists of two adjacent triangles sharing e. We say that Ω_e is an ε approximate parallelogram if the lengths of any two opposite edges differ by at most ε.

Definition. *The triangulation* $\mathcal{M}_h$ *is said to satisfy* Condition (α, σ, μ) *if there exist constants* $\alpha > 0$, $\sigma \geq 0$, *and* $\mu > 0$ *such that the interior edges can be separated into two parts* $\mathcal{E}_h = \mathcal{E}_{1,h} \oplus \mathcal{E}_{2,h}$: Ω_e *forms an* $O\big(h_e^{1+\alpha}/r_e^{\alpha+\mu(1-\alpha)}\big)$ *parallelogram for* $e \in \mathcal{E}_{1,h}$ *and the number of edges in* $\mathcal{E}_{2,h}$ *satisfies* $\#\mathcal{E}_{2,h} \lesssim N^\sigma$.

Remark 2.1. The meaning of *Condition* (α, σ, μ) is the following. The edges can be grouped into "good" ($\mathcal{E}_{1,h}$) and "bad" ($\mathcal{E}_{2,h}$), where the number of bad edges is much smaller than that of good edges. The ratio is

$$\frac{\#\mathcal{E}_{2,h}}{\#\mathcal{E}_{1,h}} \lesssim \frac{N^\sigma}{N} = \frac{1}{N^{1-\sigma}}.$$

When $r_e = O(1)$, i.e., an edge e is far away from the singular point O, more restrictions are put on the adjacent triangles with the common edge e. This condition requires that they form an $O(h_e^{1+\alpha})$ parallelogram, which is the same as in previous works [20, 23, 25, 26]. When e is in a neighborhood of O, where $r_e^{1+\mu(1-\alpha)/\alpha} \lesssim h_e$, the condition $O(h_e)$ implies $O\big(h_e^{1+\alpha}/r_e^{\alpha+\mu(1-\alpha)}\big)$. In other words, two adjacent triangles that share e are allow to distort $O(h_e)$ from a parallelogram, which implies no restriction on them. Roughly speaking, the number of edges in $\mathcal{E}_{1,h}$ that have no restriction imposed is $O(N^{1-\alpha})$ if $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^\mu$ for any $\tau \in \mathcal{M}_h$. Here $\underline{h}$ and μ are positive constants. An explanation is given below after Lemma 2.1.

We see from the above discussion that the closer we are to the singular point, the less restriction is imposed on the mesh. Indeed, for an adaptively refined mesh, the closer we are to the singular point, the worse the mesh quality is in terms of forming parallelogram triangular pairs.

LEMMA 2.1. *Assume that* $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^\mu$ *for any* $\tau \in \mathcal{M}_h$, *where* $\underline{h}$ *and* μ *are positive constants. Then the total number of degrees of freedom* N *of the finite element equation* (1.4) *satisfies*

$$N \eqsim \frac{1}{\underline{h}^{2\mu}}. \tag{2.1}$$

Proof.

$$N \eqsim \sum_{\tau\in\mathcal{M}_h} \frac{h_\tau^2}{h_\tau^2} \eqsim \frac{1}{\underline{h}^{2\mu}} \sum_{\tau\in\mathcal{M}_h} \frac{1}{r_\tau^{2-2\mu}} \cdot |\tau|$$

$$\eqsim \frac{1}{\underline{h}^{2\mu}} \int_\Omega \frac{1}{r^{2-2\mu}} \eqsim \frac{1}{\underline{h}^{2\mu}} \int_0^1 \frac{1}{r^{2-2\mu}} \cdot r \,\mathrm{d}r \eqsim \frac{1}{\underline{h}^{2\mu}}.$$

This completes the proof of the lemma. ∎

Remark 2.2. For the linear element, $\mu = \delta/2$, $N \eqsim 1/\underline{h}^{\delta}$, and for the quadratic element $\mu = \delta/3$, $N \eqsim 1/\underline{h}^{2\delta/3}$. The condition $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^{\mu}$ can be viewed as a discrete mesh density function. The positive number $\underline{h} \eqsim \min_{\tau\in\mathcal{M}_h} h_\tau$, is the size of the minimum element because for an element τ neighboring O, $r_\tau \eqsim h_\tau$ and the condition $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^{\mu}$ implies that $h_\tau \eqsim \underline{h}$. It is clear that the condition $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^{\mu}$ for any $\tau \in \mathcal{M}_h$ is equivalent to the condition $h_e \eqsim r_e^{1-\mu}\underline{h}^{\mu}$ for any $e \in \mathcal{E}_h$. We recall that *Condition* (α, σ, μ) means no restriction on Ω_e if $r_e^{1+\mu(1-\alpha)/\alpha} \lesssim h_e$. Furthermore, if $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^{\mu}$, i.e., $h_e \eqsim r_e^{1-\mu}\underline{h}^{\mu}$, then $r_e \lesssim \underline{h}^{\alpha}$. Therefore if the mesh $\mathcal{M}_h$ satisfies *Condition* (α, σ, μ) and $h_\tau \eqsim r_\tau^{1-\mu}\underline{h}^{\mu}$, then no restriction is imposed on edges within the ball of radius $R \lesssim \underline{h}^{\alpha}$. The number of edges in the ball is $O(N^{1-\alpha})$ by an argument similar to the proof of Lemma 2.1.

3. Superconvergence between the finite element solution and linear interpolant. We now define a quadratic interpolant of ϕ based on moment conditions on edges. Let $\phi_Q = \Pi_Q\phi$ be a quadratic element defined by

$$(\Pi_Q\phi)(z) = \phi(z), \quad \text{and} \quad \int_e \Pi_Q\phi = \int_e \phi \quad \forall z \in \mathcal{N}_h, e \in \mathcal{E}_h. \tag{3.1}$$

The following fundamental identity is proved in [8] for $v_h \in P_1(\tau)$:

$$\int_\tau \nabla(\phi - \phi_I)\cdot\nabla v_h = \sum_{e\subset\partial\tau}\left(\beta_e \int_e \frac{\partial^2\phi_Q}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e \int_e \frac{\partial^2\phi_Q}{\partial \boldsymbol{t}_e \partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}\right), \tag{3.2}$$

where

$$\beta_e = \frac{1}{12}\cot\theta_e(h_{e+1}^2 - h_{e-1}^2), \quad \gamma_e = \frac{1}{3}\cot\theta_e\,|\tau|\,, \tag{3.3}$$

and $\phi_I \in P_1(\tau)$ is the linear interpolant of ϕ on τ. The following lemma is a simple modification of [8, Lemma 2.13].

LEMMA 3.1. *Let $\boldsymbol{m}_e$ denote $\boldsymbol{t}_e$ or $\boldsymbol{n}_e$. Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha, \sigma, \delta/2)$ *with $0 < \alpha \le 1$ and $0 \le \sigma < 1$. For any interior edge $e \in \mathcal{M}_h$ and two elements τ, $\tau' \subset \Omega_e$, we have*

$$|\beta_e| + |\beta_e'| \lesssim h_e^2, \quad |\gamma_e| + |\gamma_e'| \lesssim h_e^2 \quad \forall e \in \mathcal{E}_h; \tag{3.4}$$

$$|\beta_e - \beta_e'| \lesssim h_e^{2+\alpha}/r_e^{\alpha+\delta(1-\alpha)/2}, \quad |\gamma_e - \gamma_e'| \lesssim h_e^{2+\alpha}/r_e^{\alpha+\delta(1-\alpha)/2} \quad \forall e \in \mathcal{E}_{1,h}; \tag{3.5}$$

$$\int_e \frac{\partial^2\phi}{\partial \boldsymbol{t}_e\partial \boldsymbol{m}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e} \lesssim |\phi|_{W^{2,\infty}(e)}\,\|\nabla v_h\|_{L^2(\tau)}\,; \tag{3.6}$$

$$\int_e \frac{\partial^2(\phi - \phi_Q)}{\partial \boldsymbol{t}_e\partial \boldsymbol{m}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e} \lesssim |\phi|_{H^3(\tau)}\,\|\nabla v_h\|_{L^2(\tau)}\,. \tag{3.7}$$

Proof. The arguments for (3.4), (3.5), and (3.6) are trivial, and that for (3.7) follows from the trace theorem and the standard error estimate $|\phi - \phi_Q|_{H^2(\tau)} \lesssim h_\tau\,|\phi|_{H^3(\tau)}$. ∎

To deal with the singularity at the origin O we introduce the following lemma. Recall that v is the singular part of the decomposition $u = v + w$.

LEMMA 3.2. *Let $\mathcal{M}^O = \{\tau \in \mathcal{M}_h : \text{the origin } O \in \partial\tau\}$ be the set of elements with one vertex at O. Then,*

$$\|\nabla v - \nabla v_I\|_{L^2(\tau)} \lesssim h_\tau^\delta \quad \forall \tau \in \mathcal{M}^O,$$

where $v_I = I_h^1 v$ is the linear interpolant of v.

Proof.

$$\|\nabla v - \nabla v_I\|_{L^2(\tau)} \lesssim \|\nabla v\|_{L^2(\tau)} + \|\nabla v_I\|_{L^2(\tau)}. \tag{3.8}$$

It follows from (1.3) that

$$\|\nabla v\|_{L^2(\tau)} = \left(\int_\tau |\nabla v|^2\right)^{1/2} \lesssim \left(\int_\tau r^{2\delta-2}\right)^{1/2} \lesssim \left(\int_0^{h_\tau} r^{2\delta-2} r\,\mathrm{d}r\right)^{1/2} \lesssim h_\tau^\delta. \tag{3.9}$$

Since $\nabla C = 0$, for any constant C, we have,

$$\begin{aligned}\|\nabla v_I\|_{L^2(\tau)} = \|\nabla(v_I - v(O))\|_{L^2(\tau)} &\lesssim h_\tau \max_{z\in\mathcal{N}_h\cap\tau} \big|\nabla\big(v_I - v(O)\big)(z)\big| \\ &\lesssim h_\tau \frac{1}{h_\tau} \max_{z\in\mathcal{N}_h\cap\tau} |v(z) - v(O)| \\ &= \max_{z\in\mathcal{N}_h\cap\tau} \left|\int_0^1 \frac{d}{dt} v(zt)\mathrm{d}t\right| = \max_{z\in\mathcal{N}_h\cap\tau} \left|\int_0^1 z\cdot\nabla v(zt)\mathrm{d}t\right|.\end{aligned}$$

Noting that $|z| \lesssim h_\tau$ for $\tau \in \mathcal{M}^O$, it follows from assumption (1.3) that

$$\|\nabla v_I\|_{L^2(\tau)} \lesssim \int_0^1 h_\tau \cdot (h_\tau t)^{\delta-1}\mathrm{d}t \lesssim h_\tau^\delta. \tag{3.10}$$

The proof is completed by combining (3.8)–(3.10). ∎

LEMMA 3.3. *Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha, \sigma, \delta/2)$ *with $0 < \alpha \le 1$ and $0 \le \sigma < 1$, and that $h_\tau \eqsim r_\tau^{1-\delta/2}\underline{h}^{\delta/2}$ for any $\tau \in \mathcal{M}_h$. Then for any $v_h \in \mathring{V}_h^1$,*

$$\left|\int_\Omega \nabla(u - u_I)\cdot\nabla v_h\right| \lesssim \frac{1 + (\ln N)^{1/2}}{N^{1/2+\rho}} \|\nabla v_h\|_{L^2(\Omega)}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right), \tag{3.11}$$

where $u_I = I_h^1 u \in V_h^1$ is the piecewise linear interpolant of u.

Proof. From the decomposition $u = v + w$,

$$\int_\Omega \nabla(u - u_I)\cdot\nabla v_h = \int_\Omega \nabla(v - v_I)\cdot\nabla v_h + \int_\Omega \nabla(w - w_I)\cdot\nabla v_h, \tag{3.12}$$

where $v_I = I_h^1 v$ and $w_I = I_h^1 w$ are the linear interpolants of v and w, respectively.

We first estimate $\int_\Omega \nabla(v - v_I)\cdot\nabla v_h$. Let $\mathcal{E}^O = \{e \in \mathcal{E}_h : e \subset \partial\tau \text{ the origin } O \in \tau\}$ and $\partial\mathcal{E}^O = \big\{e \in \mathcal{E}^O : O \notin e\big\}$. Recall that $\mathcal{M}^O$ is the set of elements with one vertex at O. Applying (3.2),

$$\begin{aligned}\int_\Omega \nabla(v - v_I)\cdot\nabla v_h &= \sum_{\tau\in\mathcal{M}_h} \int_\tau \nabla(v - v_I)\cdot\nabla v_h = \sum_{\tau\in\mathcal{M}^O} \int_\tau \nabla(v - v_I)\cdot\nabla v_h \\ &\quad + \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^O} \sum_{e\subset\partial\tau} \left(\beta_e \int_e \frac{\partial^2 v_Q}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e \int_e \frac{\partial^2 v_Q}{\partial \boldsymbol{t}_e \partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}\right) \\ &= I_1 + I_2 + I_3 + I_4,\end{aligned} \tag{3.13}$$

where

$$I_j = \sum_{e\in\mathcal{E}_{j,h}\backslash\mathcal{E}^O}\Big[(\beta_e-\beta_e')\int_e \frac{\partial^2 v}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + (\gamma_e-\gamma_e')\int_e \frac{\partial^2 v}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}$$
$$+\beta_e\int_e \frac{\partial^2 (v_Q-v)}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e\int_e \frac{\partial^2 (v_Q-v)}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}$$
$$+\beta_e'\int_e \frac{\partial^2 (v-v_Q)}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e'\int_e \frac{\partial^2 (v-v_Q)}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}\Big], \quad j=1,2,$$

$$I_3 = \sum_{\tau\in\mathcal{M}^O}\int_\tau \nabla(v-v_I)\cdot\nabla v_h,$$

$$I_4 = \sum_{e\in\partial\mathcal{E}^O}\left(\beta_e\int_e \frac{\partial^2 v_Q}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e\int_e \frac{\partial^2 v_Q}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}\right).$$

First, I_3 can be estimated by Lemma 3.2 and the fact that $h_\tau \eqsim \underline{h}$ for $\tau\in\mathcal{M}^O$:

$$|I_3| \lesssim \underline{h}^\delta \sum_{\tau\in\mathcal{M}^O}\|\nabla v_h\|_{L^2(\tau)} \lesssim \underline{h}^\delta \|\nabla v_h\|_{L^2(\Omega)}. \tag{3.14}$$

Second, I_4 can be estimated by Lemma 3.1, assumption (1.3), and the fact that $h_e \eqsim r_e \eqsim \underline{h}$ for $e\in\partial\mathcal{E}^O$:

$$\begin{aligned}|I_4| &\lesssim \sum_{e\in\partial\mathcal{E}^O} h_e^2\left(|v|_{W^{2,\infty}(e)} + |v|_{H^3(\tau:\,\tau\in\Omega_e,\tau\notin\mathcal{M}^O)}\right)\|\nabla v_h\|_{L^2(\tau:\,\tau\in\Omega_e,\tau\notin\mathcal{M}^O)}\\ &\lesssim \sum_{e\in\partial\mathcal{E}^O} h_e^2\big(r_e^{\delta-2}+h_e r_e^{\delta-3}\big)\|\nabla v_h\|_{L^2(\tau:\,\tau\in\Omega_e,\tau\notin\mathcal{M}^O)} \qquad (3.15)\\ &\lesssim \underline{h}^\delta \sum_{e\in\partial\mathcal{E}^O}\|\nabla v_h\|_{L^2(\tau:\,\tau\in\Omega_e,\tau\notin\mathcal{M}^O)} \lesssim \underline{h}^\delta\|\nabla v_h\|_{L^2(\Omega)}.\end{aligned}$$

Next we estimate I_1. Notice that $h_e \eqsim h_\tau$ and $r_e \eqsim r_\tau$ for $\tau\subset\Omega_e$ and $e\in\mathcal{E}_{1,h}\backslash\mathcal{E}^O$. It follows from Lemma 3.1 and assumption (1.3) that

$$\begin{aligned}|I_1| &\lesssim \sum_{e\in\mathcal{E}_{1,h}\backslash\mathcal{E}^O}\left[\frac{h_e^{2+\alpha}}{r_e^{\alpha+\delta(1-\alpha)/2}}r_e^{\delta-2} + h_e^2 h_\tau r_\tau^{\delta-3}\right]\|\nabla v_h\|_{L^2(\tau:\,\tau\in\Omega_e)}\\ &\lesssim \sum_{e\in\mathcal{E}_{1,h}\backslash\mathcal{E}^O}\left[h_e^{2+\alpha}r_e^{\delta-2-\alpha-\delta(1-\alpha)/2} + h_e^3 r_e^{\delta-3}\right]\|\nabla v_h\|_{L^2(\tau:\,\tau\in\Omega_e)}\\ &\lesssim \left\{\sum_{e\in\mathcal{E}_{1,h}\backslash\mathcal{E}^O}\left[h_e^2 h_e^{2+2\alpha}r_e^{2\delta-4-2\alpha-\delta(1-\alpha)} + h_e^2 h_e^4 r_e^{2\delta-6}\right]\right\}^{1/2}\|\nabla v_h\|_{L^2(\Omega)}\\ &\lesssim \left\{\sum_{e\in\mathcal{E}_{1,h}\backslash\mathcal{E}^O}\left[h_e^2\underline{h}^{\delta(1+\alpha)}r_e^{(2-\delta)(1+\alpha)}r_e^{2\delta-4-2\alpha-\delta(1-\alpha)} + h_e^2\underline{h}^{2\delta}r_e^{4-2\delta}r_e^{2\delta-6}\right]\right\}^{1/2}\\ &\quad\times\|\nabla v_h\|_{L^2(\Omega)}.\end{aligned}$$

Here we have used $h_e \eqsim r_e^{1-\delta/2}\underline{h}^{\delta/2}$ to derive the last inequality. Therefore

(3.16)
$$\begin{aligned} I_1 &\lesssim \left\{\underline{h}^{\delta(1+\alpha)} \sum_{e\in\mathcal{E}_{1,h}\setminus\mathcal{E}^O} h_e^2 r_e^{-2}\right\}^{1/2} \|\nabla v_h\|_{L^2(\Omega)} \lesssim \left\{\underline{h}^{\delta(1+\alpha)} \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^O} h_\tau^2 r_\tau^{-2}\right\}^{1/2} \\ &\quad \times \|\nabla v_h\|_{L^2(\Omega)} \\ &\lesssim \left\{\underline{h}^{\delta(1+\alpha)} \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^O} \int_\tau r^{-2}\right\}^{1/2} \|\nabla v_h\|_{L^2(\Omega)} \lesssim \left\{\underline{h}^{\delta(1+\alpha)} \int_{\underline{h}}^1 r^{-1}\,\mathrm{d}r\right\}^{1/2} \\ &\quad \times \|\nabla v_h\|_{L^2(\Omega)} \\ &\lesssim \underline{h}^{\delta(1+\alpha)/2}(|\ln \underline{h}|^{1/2}) \|\nabla v_h\|_{L^2(\Omega)}. \end{aligned}$$

Finally, we estimate I_2. Notice that $h_e \lesssim r_e$ for $e \notin \mathcal{E}^O$. It follows from Lemma 3.1 and assumption (1.3) that

$$\begin{aligned} |I_2| &\lesssim \sum_{e\in\mathcal{E}_{2,h}\setminus\mathcal{E}^O} \left[h_e^2 r_e^{\delta-2} + h_e^2 h_\tau r_\tau^{\delta-3}\right] \|v_h\|_{L^2(\tau:\,\tau\in\Omega_e)} \\ &\lesssim \sum_{e\in\mathcal{E}_{2,h}\setminus\mathcal{E}^O} h_e^2 r_e^{\delta-2} \|v_h\|_{L^2(\tau:\,\tau\in\Omega_e)} \\ &\lesssim \left\{\sum_{e\in\mathcal{E}_{2,h}\setminus\mathcal{E}^O} h_e^4 r_e^{2\delta-4}\right\}^{1/2} \|\nabla v_h\|_{L^2(\Omega)} \\ &\lesssim \underline{h}^\delta \left\{\sum_{e\in\mathcal{E}_{2,h}\setminus\mathcal{E}^O} 1\right\}^{1/2} \|\nabla v_h\|_{L^2(\Omega)}. \end{aligned}$$

Here we have used $h_e \eqsim r_e^{1-\delta/2}\underline{h}^{\delta/2}$ to derive the last inequality. Therefore

$$|I_2| \lesssim \underline{h}^\delta \left\{\#\mathcal{E}_{2,h}\right\}^{1/2} \|\nabla v_h\|_{L^2(\Omega)} \lesssim \underline{h}^\delta \left\{N^\sigma\right\}^{1/2} \|\nabla v_h\|_{L^2(\Omega)}. \tag{3.17}$$

From Lemma 2.1, $\underline{h}^\delta \eqsim 1/N$, $|\ln \underline{h}| \eqsim \ln N$. Combining (3.13)–(3.17) we have

$$\begin{aligned} \left|\int_\Omega \nabla(v - v_I)\cdot\nabla v_h\right| &\lesssim \left(\underline{h}^{\delta(1+\alpha)/2}(|\ln \underline{h}|^{1/2}) + \underline{h}^\delta \left\{N^\sigma\right\}^{1/2}\right) \|\nabla v_h\|_{L^2(\Omega)} \\ &\lesssim \frac{1+(\ln N)^{1/2}}{N^{1/2+\rho}} \|\nabla v_h\|_{L^2(\Omega)}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right). \end{aligned} \tag{3.18}$$

Now we turn to the estimate for $\int_\Omega \nabla(w - w_I)\cdot\nabla v_h$. Since w is smooth, we do not exclude the point O. From (3.2),

$$\int_\Omega \nabla(w - w_I)\cdot\nabla v_h = \sum_{\tau\in\mathcal{M}_h} \int_\tau \nabla(w - w_I)\cdot\nabla v_h = J_1 + J_2, \tag{3.19}$$

where

$$J_j = \sum_{e\in\mathcal{E}_{j,h}} \Big[(\beta_e-\beta_e')\int_e \frac{\partial^2 w}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + (\gamma_e-\gamma_e')\int_e \frac{\partial^2 w}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}$$
$$+\beta_e\int_e \frac{\partial^2 (w_Q-w)}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e\int_e \frac{\partial^2 (w_Q-w)}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}$$
$$+\beta_e'\int_e \frac{\partial^2 (w-w_Q)}{\partial \boldsymbol{t}_e^2}\frac{\partial v_h}{\partial \boldsymbol{t}_e} + \gamma_e'\int_e \frac{\partial^2 (w-w_Q)}{\partial \boldsymbol{t}_e\partial \boldsymbol{n}_e}\frac{\partial v_h}{\partial \boldsymbol{t}_e}\Big], \quad j=1,2.$$

By a similar argument as for I_1 and I_2, we can prove that

$$\left|\int_\Omega \nabla(w-w_I)\cdot\nabla v_h\right| \lesssim \frac{1}{N^{1/2+\rho}}\left\|\nabla v_h\right\|_{L^2(\Omega)}. \tag{3.20}$$

Now, the proof of the lemma follows from (3.12), (3.18), and (3.20). ∎

Applying Lemma 3.3 we obtain the following superconvergence result between the finite element solution u_h and the linear interpolant u_I of the solution of (1.1).

THEOREM 3.4. *Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha,\sigma,\delta/2)$ *with $0<\alpha\le 1$ and $0\le\sigma<1$ and that $h_\tau \eqsim r_\tau^{1-\delta/2}\underline{h}^{\delta/2}$ for any $\tau\in\mathcal{M}_h$. Then*

$$\left\|\nabla(u_h-u_I)\right\|_{L^2(\Omega)} \lesssim \frac{1+(\ln N)^{1/2}}{N^{1/2+\rho}}, \quad \rho=\min\left(\frac{\alpha}{2},\frac{1-\sigma}{2}\right). \tag{3.21}$$

Proof. Taking $v_h=u_h-u_I$ in Lemma 3.3 we have

$$\left\|\nabla(u_h-u_I)\right\|_{L^2(\Omega)}^2 = A(u_h-u_I,v_h) = A(u-u_I,v_h) = \int_\Omega \nabla(u-u_I)\cdot\nabla v_h$$
$$\lesssim \frac{1+(\ln N)^{1/2}}{N^{1/2+\rho}}\left\|\nabla v_h\right\|_{L^2(\Omega)} = \frac{1+(\ln N)^{1/2}}{N^{1/2+\rho}}\left\|\nabla(u_h-u_I)\right\|_{L^2(\Omega)}.$$

The proof is completed by canceling $\left\|\nabla(u_h-u_I)\right\|_{L^2(\Omega)}$ on both sides of the inequality. ∎

4. Superconvergence between the finite element solution and quadratic interpolation. Most parts of the proof are similar to those for linear elements and therefore are omitted. We emphasize only the differing parts. In this section u_h is the solution of (1.4) with $k=2$, that is, the quadratic finite element approximation of u.

We first introduce some estimates over triangles from [14]. Recall that $\phi_Q=\Pi_Q\phi$ is the quadratic interpolant defined in (3.1) based on the moment conditions.

LEMMA 4.1. *Assume that $\phi\in H^4(\tau)$; then there holds*

$$\int_\tau \nabla(\phi-\Pi_Q\phi)\cdot\nabla v_h = \sum_{e\subset\partial\tau}\sum_{s=0}^{3}\left(a_e^s(\tau)\frac{|\tau|}{h_e}+b_e^s(\tau)\right)\int_e \frac{\partial^3\phi}{\partial \boldsymbol{n}_e^s\partial \boldsymbol{t}_e^{3-s}}\frac{\partial^2 v_h}{\partial \boldsymbol{t}_e^2}$$
$$+O(h_\tau^3)\left|\phi\right|_{H^4(\tau)}\left\|v_h\right\|_{H^1(\tau)} \quad \forall v_h\in P_2(\tau), \tag{4.1}$$

where for $s = 0, 1, 2, 3$,

(4.2)

$$|a_e^s(\tau)| + |a_e^s(\tau')| \lesssim h_e^3, \quad |b_e^s(\tau)| + |b_e^s(\tau')| \lesssim h_e^4 \quad \text{if } e \in \mathcal{E}_h;$$

(4.3)

$$|a_e^s(\tau)|\tau| - a_e^s(\tau')|\tau'|| \lesssim h_e^{5+\alpha}/r_e^{\alpha+\delta(1-\alpha)/3}, |b_e^s(\tau) - b_e^s(\tau')| \lesssim h_e^{4+\alpha}/r_e^{\alpha+\delta(1-\alpha)/3}$$

if $\mathcal{M}_h$ *satisfies* Condition $(\alpha, \sigma, \delta/3)$ *with* $0 < \alpha \le 1$ *and* $0 \le \sigma < 1$, *and if* $e \in \mathcal{E}_{1,h}$.

To obtain the superconvergence of $\left\|\nabla(u_h - I_h^2 u)\right\|_{L^2(\Omega)}$, we estimate the difference between two quadratic interpolation operators Π_Q and I_h^2. It is easy to check that [27]

$$\Pi_Q p - I_h^2 p = 0 \quad \forall p \in P_3.$$

By the Bramble–Hilbert lemma, we have

$$\int_\tau (\nabla \Pi_Q \phi - \nabla I_h^2 \phi) \cdot \nabla v_h \lesssim h_\tau^3 \left|\phi\right|_{H^4(\tau)} \left\|\nabla v_h\right\|_{L^2(\tau)}.$$

Therefore we have the following lemma from (4.1).

LEMMA 4.2. *Assume that* $\phi \in H^4(\tau)$, *then there holds*

$$\int_\tau \nabla(\phi - I_h^2\phi) \cdot \nabla v_h = \sum_{e \subset \partial\tau} \sum_{s=0}^{3} \left(a_e^s(\tau) \frac{|\tau|}{h_e} + b_e^s(\tau) \right) \int_e \frac{\partial^3 \phi}{\partial \boldsymbol{n}_e^s \partial \boldsymbol{t}_e^{3-s}} \frac{\partial^2 v_h}{\partial \boldsymbol{t}_e^2} + O(h_\tau^3) \left|\phi\right|_{H^4(\tau)} \left\|v_h\right\|_{H^1(\tau)} \quad \text{for } v_h \in P_2(\tau). \tag{4.4}$$

Recall from Lemma 2.1 that, in the quadratic case, if $h_\tau \eqsim r_\tau^{1-\delta/3} \underline{h}^{\delta/3}$ for any $\tau \in \mathcal{M}_h$, then the total number of degrees of freedom N of the finite element equation (1.4) satisfies

$$N \eqsim \frac{1}{\underline{h}^{2\delta/3}}. \tag{4.5}$$

The following lemma is analogous to Lemma 3.2. We omit the proof.

LEMMA 4.3. *For* v *in decomposition* (1.2),

$$\left\|\nabla v - \nabla I_h^2 v\right\|_{L^2(\tau)} \lesssim h_\tau^\delta \quad \forall \tau \in \mathcal{M}^O.$$

The following lemma is the counterpart of Lemma 3.3 for the quadratic case.

LEMMA 4.4. *Assume that* $\mathcal{M}_h$ *satisfies* Condition $(\alpha, \sigma, \delta/3)$ *with* $0 < \alpha \le 1$ *and* $0 \le \sigma < 1$, *and that* $h_\tau \eqsim r_\tau^{1-\delta/3} \underline{h}^{\delta/3}$ *for any* $\tau \in \mathcal{M}_h$. *Then for any* $v_h \in \overset{\circ}{V}{}_h^2$,

$$\left| \int_\Omega \nabla(u - I_h^2 u) \cdot \nabla v_h \right| \lesssim \frac{1 + (\ln N)^{1/2}}{N^{1+\rho}} \left\|\nabla v_h\right\|_{L^2(\Omega)}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right). \tag{4.6}$$

Proof. From the decomposition $u = v + w$,

$$\int_\Omega \nabla(u - I_h^2 u) \cdot \nabla v_h = \int_\Omega \nabla(v - I_h^2 v) \cdot \nabla v_h + \int_\Omega \nabla(w - I_h^2 w) \cdot \nabla v_h. \tag{4.7}$$

We first estimate the term $\int_\Omega \nabla(v - I_h^2 v)\cdot\nabla v_h$. It follows from Lemma 4.2 that

$$\int_\Omega \nabla(v - I_h^2 v)\cdot\nabla v_h = \sum_{\tau\in\mathcal{M}_h}\int_\tau \nabla(v - I_h^2 v)\cdot\nabla v_h = I_1 + I_2 + I_3 + I_4, \tag{4.8}$$

where

$$I_j = \sum_{e=\tau\cap\tau'\in\mathcal{E}_{j,h}\setminus\mathcal{E}^O}\left\{\sum_{s=0}^{3}\left\{\frac{a_e^s(\tau)\,|\tau| - a_e^s(\tau')\,|\tau'|}{h_e} + \left[b_e^s(\tau) - b_e^s(\tau')\right]\right\}\int_e \frac{\partial^3 v}{\partial \boldsymbol{n}_e^s \partial \boldsymbol{t}_e^{3-s}}\frac{\partial^2 v_h}{\partial \boldsymbol{t}_e^2}\right.$$
$$\left. + O(h_e^3)\,|v|_{H^4(\Omega_e)}\,\|v_h\|_{H^1(\Omega_e)}\right\},\quad j = 1, 2,$$

$$I_3 = \sum_{\tau\in\mathcal{M}^O}\int_\tau \nabla(v - I_h^2 v)\cdot\nabla v_h,$$

$$I_4 = \sum_{e\in\partial\mathcal{E}^O}\left[\sum_{s=0}^{3}\left(a_e^s(\tau)\frac{|\tau|}{h_e} + b_e^s(\tau)\right)\int_e \frac{\partial^3 v}{\partial \boldsymbol{n}_e^s \partial \boldsymbol{t}_e^{3-s}}\frac{\partial^2 v_h}{\partial \boldsymbol{t}_e^2} + O(h_\tau^3)\,|v|_{H^4(\tau)}\,\|v_h\|_{H^1(\tau)}\right].$$

Notice that the τ in I_4 is not in $\mathcal{M}^O$.

From Lemma 4.3,

$$|I_3| \lesssim \underline{h}^\delta\,\|\nabla v_h\|_{L^2(\Omega)}. \tag{4.9}$$

It follows from (4.2) and assumption (1.3) that

$$\begin{aligned} |I_4| &\lesssim \sum_{e\in\partial\mathcal{E}^O}\left(h_e^5 r_e^{\delta-3}\,|v_h|_{W^{2,\infty}(\tau)} + h_\tau^3 h_\tau r_e^{\delta-4}\,\|v_h\|_{H^1(\tau)}\right)\\ &\lesssim \sum_{e\in\partial\mathcal{E}^O}\left(h_e^3 r_e^{\delta-3}\,\|v_h\|_{H^1(\tau)}\right) + h_e^4 r_e^{\delta-4}\,\|v_h\|_{H^1(\tau)} \lesssim \underline{h}^\delta\,\|v_h\|_{H^1(\Omega)}. \end{aligned} \tag{4.10}$$

Here we have used the inverse estimate $|v_h|_{W^{2,\infty}(\tau)} \lesssim h_e^{-2}\,\|v_h\|_{H^1(\tau)}$ and the fact that $h_e \eqsim r_e \eqsim \underline{h}$ for $e\in\partial\mathcal{E}^O$.

Next we estimate I_1. It follows from Lemma 4.1 and assumption (1.3) that

$$\begin{aligned} |I_1| &\lesssim \sum_{e\in\mathcal{E}_{1,h}\setminus\mathcal{E}^O}\left[\frac{h_e^{5+\alpha}}{r_e^{\alpha+\delta(1-\alpha)/3}} r_e^{\delta-3}\,|v_h|_{W^{2,\infty}(\tau)} + h_e^3 h_\tau r_\tau^{\delta-4}\,\|v_h\|_{H^1(\Omega_e)}\right]\\ &\lesssim \sum_{e\in\mathcal{E}_{1,h}\setminus\mathcal{E}^O}\left[h_e^{3+\alpha} r_e^{\delta-3-\alpha-\delta(1-\alpha)/3} + h_e^4 r_e^{\delta-4}\right]\|v_h\|_{H^1(\Omega_e)}\\ &\lesssim \left\{\sum_{e\in\mathcal{E}_{1,h}\setminus\mathcal{E}^O}\left[h_e^2 h_e^{4+2\alpha} r_e^{2\delta-6-2\alpha-2\delta(1-\alpha)/3} + h_e^2 h_e^6 r_e^{2\delta-8}\right]\right\}^{1/2}\|v_h\|_{H^1(\Omega)}\\ &\lesssim \left\{\sum_{e\in\mathcal{E}_{1,h}\setminus\mathcal{E}^O}\left[h_e^2 \underline{h}^{2\delta(2+\alpha)/3} r_e^{(4+2\alpha)(1-\delta/3)} r_e^{2\delta-6-2\alpha-2\delta(1-\alpha)/3}\right.\right.\\ &\qquad\left.\left. + h_e^2 \underline{h}^{2\delta} r_e^{6-2\delta} r_e^{2\delta-8}\right]\right\}^{1/2}\|v_h\|_{H^1(\Omega)}. \end{aligned}$$

Here we have used $h_e \eqsim r_e^{1-\delta/3}\underline{h}^{\delta/3}$ to derive the last inequality. Therefore

(4.11)

$$I_1 \lesssim \left\{\underline{h}^{2\delta(2+\alpha)/3}\sum_{e\in\mathcal{E}_{1,h}\backslash\mathcal{E}^O} h_e^2 r_e^{-2}\right\}^{1/2}\|v_h\|_{H^1(\Omega)} \lesssim \left\{\underline{h}^{2\delta(2+\alpha)/3}\sum_{\tau\in\mathcal{M}_h\backslash\mathcal{M}^O} h_\tau^2 r_\tau^{-2}\right\}^{1/2}$$
$$\times \|v_h\|_{H^1(\Omega)}$$
$$\lesssim \left\{\underline{h}^{2\delta(2+\alpha)/3}\sum_{\tau\in\mathcal{M}_h\backslash\mathcal{M}^O}\int_\tau r^{-2}\right\}^{1/2}\|v_h\|_{H^1(\Omega)} \lesssim \underline{h}^{\delta(2+\alpha)/3}\left|\ln\underline{h}\right|^{1/2}\|v_h\|_{H^1(\Omega)}.$$

By a similar argument for (3.17) we can show that

$$|I_2| \lesssim \underline{h}^{\delta}\left\{\#\mathcal{E}_{2,h}\right\}^{1/2}\|v_h\|_{H^1(\Omega)} \lesssim \underline{h}^{\delta}\left\{N^\sigma\right\}^{1/2}\|v_h\|_{H^1(\Omega)}. \tag{4.12}$$

Notice that $\|v_h\|_{H^1(\Omega)} \lesssim \|\nabla v_h\|_{L^2(\Omega)}$ from Poincaré's inequality. Combining (4.8)–(4.12), we have

$$\begin{aligned}\left|\int_\Omega \nabla(v-v_I)\cdot\nabla v_h\right| &\lesssim \left(\underline{h}^{\delta(2+\alpha)/3}\left|\ln\underline{h}\right|^{1/2} + \underline{h}^{\delta}\left\{N^\sigma\right\}^{1/2}\right)\|\nabla v_h\|_{L^2(\Omega)}\\ &\lesssim \frac{1+(\ln N)^{1/2}}{N^{1+\rho}}\|\nabla v_h\|_{L^2(\Omega)}, \quad \rho=\min\left(\frac{\alpha}{2},\frac{1-\sigma}{2}\right).\end{aligned} \tag{4.13}$$

The estimate for the term $\int_\Omega \nabla(w-I_h^2 w)\cdot\nabla v_h$ is similar to (4.13). It follows from Lemma 4.2 that

$$\int_\Omega \nabla(w-I_h^2 w)\cdot\nabla v_h = \sum_{\tau\in\mathcal{M}_h}\int_\tau \nabla(w-I_h^2 w)\cdot\nabla v_h = J_1+J_2, \tag{4.14}$$

where

$$J_j = \sum_{e=\tau\cap\tau'\in\mathcal{E}_{j,h}}\left\{\sum_{s=0}^{3}\left\{\frac{a_e^s(\tau)\,|\tau| - a_e^s(\tau')\,|\tau'|}{h_e} + \left[b_e^s(\tau)-b_e^s(\tau')\right]\right\}\int_e \frac{\partial^3 w}{\partial \boldsymbol{n}_e^s\partial \boldsymbol{t}_e^{3-s}}\frac{\partial^2 v_h}{\partial \boldsymbol{t}_e^2}\right.$$
$$\left. + O(h_e^3)\,|w|_{H^4(\Omega_e)}\|v_h\|_{H^1(\Omega_e)}\right\}, \quad j=1,2.$$

There holds

$$\left|\int_\Omega \nabla(w-w_I)\cdot\nabla v_h\right| \lesssim \frac{1}{N^{1+\rho}}\|\nabla v_h\|_{L^2(\Omega)}. \tag{4.15}$$

Now, the conclusion follows from (4.7), (4.13), and (4.15). ∎

Applying Lemma 4.4 we obtain the following superconvergence result between the quadratic finite element approximation u_h and the quadratic interpolant $I_h^2 u$ of the solution of problem (1.1).

THEOREM 4.5. *Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha,\sigma,\delta/3)$ *with $0<\alpha\le 1$ and $0\le\sigma<1$, and that $h_\tau \eqsim r_\tau^{1-\delta/3}\underline{h}^{\delta/3}$ for any $\tau\in\mathcal{M}_h$. Then*

$$\left\|\nabla(u_h - I_h^2 u)\right\|_{L^2(\Omega)} \lesssim \frac{1+(\ln N)^{1/2}}{N^{1+\rho}}, \quad \rho=\min\left(\frac{\alpha}{2},\frac{1-\sigma}{2}\right). \tag{4.16}$$

5. The asymptotically exact a posteriori error estimators. In this section, we apply a newly developed gradient recovery operator, called polynomial preserving recovery (PPR) [20, 26, 28], to define an a posteriori error estimator. We further prove some superconvergence properties of the recovery operator. As a consequence, the error estimator based on PPR is asymptotically exact under a mesh density assumption.

5.1. The gradient recovery operator G_h and its superconvergence. Given a node $z \in \mathcal{N}_h$, we select $n \geq m = (k+2)(k+3)/2$ sampling points $z_j \in \mathcal{N}_h$, $j = 1, 2, \ldots, n$, in an element patch ω_z containing z (z is one of z_j) and fit a polynomial of degree $k+1$, in the least squares sense, with values of u_h at those sampling points. In other words, we are looking for $p_{k+1} \in \mathcal{P}_{k+1}$ such that

$$\sum_{j=1}^{n} (p_{k+1} - u_h)^2(z_j) = \min_{q \in \mathcal{P}_{k+1}} \sum_{j=1}^{n} (q - u_h)^2(z_j). \tag{5.1}$$

The recovered gradient at of z is then defined as

$$G_h u_h(z) = (\nabla p_{k+1})(z). \tag{5.2}$$

It was proved in [20] that the above least squares fitting procedure has a unique solution as long as those n sampling points are not on the same conic curve for the linear element. Conditions for higher order elements were given as well. Furthermore, the gradient recovery operator $G_h : C(\Omega) \mapsto V_h^k \times V_h^k$, $k = 1$ or 2, has the following properties:

(i) $\|G_h v_h\|_{L^2(\Omega)} \lesssim \|\nabla v_h\|_{L^2(\Omega)} \quad \forall v_h \in V_h^k$.
(ii) For any nodal point z, $(G_h p)(z) = \nabla p(z)$ if $p \in P_{k+1}(\omega_z)$.
(iii) $|(G_h\phi)(z)| \lesssim \frac{1}{h_\tau} \max_{z' \in \mathcal{N}_h \cap \omega_z} |\phi(z')|$ for any node z in an element $\tau \in \mathcal{M}_h$.
(iv) $G_h\phi = G_h I_h^k \phi$.

Since $I_h^k\phi$ and ϕ have the same nodal values and G_h uses only nodal values, (iv) is clear. The polynomial preserving property (ii) can be established easily by the least squares procedure [28]. A key observation is that G_h provides a finite difference scheme at each node $z \in \mathcal{N}_h$; therefore, (iii) is obvious. Under a very mild mesh condition, "the sum of any two adjacent angles in $\mathcal{M}_h$ is at most π," the boundedness property (i) can be proved, though it is not trivial. The reader is referred to [20, 26, 28] for more details.

We first consider the case of linear finite elements and then state the corresponding results for quadratic elements since the proofs are similar. We have from (i),

$$\begin{aligned} \|G_h u_h - \nabla u\|_{L^2(\Omega)} &\leq \|G_h u_h - G_h u_I\|_{L^2(\Omega)} + \|G_h u_I - \nabla u\|_{L^2(\Omega)} \\ &\lesssim \|\nabla(u_h - u_I)\|_{L^2(\Omega)} + \|G_h u_I - \nabla u\|_{L^2(\Omega)}. \end{aligned} \tag{5.3}$$

Here u_I is the linear interpolant of u. The estimate for the first term of the right hand side of the inequality (5.3) is given in Theorem 3.4. To estimate the second term we need the following lemma.

LEMMA 5.1. *Under properties* (ii)–(iii)*, for any element $\tau \in \mathcal{M}_h$ and any function $\phi \in W^{3,\infty}(\tilde{\tau})$,*

$$\|G_h\phi_I - \nabla\phi\|_{L^2(\tau)} \lesssim h_\tau^3 |\phi|_{W^{3,\infty}(\tilde{\tau})},$$

where $\tilde{\tau} = \bigcup \{\omega_z : z \in \mathcal{N}_h \cap \tau\}$ and ϕ_I is the linear interpolant of ϕ.

Proof. Let $(\nabla\phi)_I$ be the linear interpolant of $\nabla\phi$. Then

$$\|G_h\phi_I - \nabla\phi\|_{L^2(\tau)} \le \|G_h\phi_I - (\nabla\phi)_I\|_{L^2(\tau)} + \|(\nabla\phi)_I - \nabla\phi\|_{L^2(\tau)}\,. \tag{5.4}$$

The standard theory of finite element interpolation estimates says that [6]

$$\|(\nabla\phi)_I - \nabla\phi\|_{L^2(\tau)} \lesssim h_\tau^2\,|\phi|_{H^3(\tau)} \lesssim h_\tau^3\,|\phi|_{W^{3,\infty}(\tilde\tau)}\,. \tag{5.5}$$

For a node $z \in \tau$, let $\phi_2(x,y)$ be the 2nd-degree Taylor expansion of ϕ at the point z. It is clear that

$$|\phi(x,y) - \phi_2(x,y)| \lesssim h_\tau^3\,|\phi|_{W^{3,\infty}(\tilde\tau)} \quad \forall (x,y) \in \tilde\tau.$$

By properties (ii) and (iii),

$$\begin{aligned}\big|\big(G_h\phi_I - (\nabla\phi)_I\big)(z)\big| &= \big|\big(G_h\phi_I - \nabla\phi\big)(z)\big| = \big|\big(G_h(\phi_I - \phi_2) - (\nabla\phi - \nabla\phi_2)\big)(z)\big| \\ &= \big|\big(G_h(\phi_I - \phi_2)\big)(z)\big| \lesssim \frac{1}{h_\tau}\max_{z'\in\mathcal{N}_h\cap\omega_z}|(\phi-\phi_2)(z')| \\ &\lesssim h_\tau^2\,|\phi|_{W^{3,\infty}(\omega_z)}\,.\end{aligned}$$

Therefore

$$\|G_h\phi_I - (\nabla\phi)_I\|_{L^2(\tau)} \lesssim h_\tau \max_{z\in\mathcal{N}_h\cap\tau}\big|\big(G_h\phi_I - (\nabla\phi)_I\big)(z)\big| \lesssim h_\tau^3\,|\phi|_{W^{3,\infty}(\tilde\tau)}\,. \tag{5.6}$$

The proof of the lemma is completed by combining (5.4)–(5.6). ∎

The following theorem is devoted to the estimate of the second term of (5.3).

THEOREM 5.2. *Assume that* $h_\tau \eqsim r_\tau^{1-\delta/2}\underline{h}^{\delta/2}$ *for any* $\tau \in \mathcal{M}_h$. *Then*

$$\|G_hu_I - \nabla u\|_{L^2(\Omega)} \lesssim \frac{1+(\ln N)^{1/2}}{N}. \tag{5.7}$$

Proof. Recall the decomposition (1.2) $u = v + w$, we have, by the triangular inequality,

$$\|G_hu_I - \nabla u\|_{L^2(\Omega)} \le \|G_hv_I - \nabla v\|_{L^2(\Omega)} + \|G_hw_I - \nabla w\|_{L^2(\Omega)}\,, \tag{5.8}$$

where $v_I = I_h^1 v$ and $w_I = I_h^1 w$ are the linear interpolants of v and w, respectively.

We first estimate the singular part $\|G_hv_I - \nabla v\|_{L^2(\Omega)}$. Introduce the set of triangles $\mathcal{M}^{\tilde O} = \{\tau \in \mathcal{M}_h :$ the origin $O \in \tilde\tau\}$. For any $\tau \in \mathcal{M}^{\tilde O}$,

$$\|G_hv_I - \nabla v\|_{L^2(\tau)} \le \|G_hv_I\|_{L^2(\tau)} + \|\nabla v\|_{L^2(\tau)}\,. \tag{5.9}$$

By property (ii), $G_hC = 0$ for any constant C. Thus, from property (iii),

$$\begin{aligned}\|G_hv_I\|_{L^2(\tau)} = \|G_h(v_I - v(O))\|_{L^2(\tau)} &\lesssim h_\tau \max_{z\in\mathcal{N}_h\cap\tau}\big|G_h\big(v_I - v(O)\big)(z)\big| \\ &\lesssim h_\tau\frac{1}{h_\tau}\max_{z'\in\mathcal{N}_h\cap\tilde\tau}|v(z') - v(O)| \\ &= \max_{z'\in\mathcal{N}_h\cap\tilde\tau}\left|\int_0^1 \frac{d}{dt}v(z't)\mathrm{d}t\right| = \max_{z'\in\mathcal{N}_h\cap\tilde\tau}\left|\int_0^1 z'\cdot\nabla v(z't)\mathrm{d}t\right|.\end{aligned}$$

Since $\tau \in \mathcal{M}^{\widetilde{O}}$, $|z'| \lesssim \underline{h}$. It follows from assumption (1.3) that

$$\|G_h v_I\|_{L^2(\tau)} \lesssim \int_0^1 \underline{h} \cdot (\underline{h}t)^{\delta-1} \mathrm{d}t \lesssim \underline{h}^{\delta}. \tag{5.10}$$

On the other hand,

$$\|\nabla v\|_{L^2(\tau)} \lesssim \left(\int_\tau |\nabla v|^2\right)^{1/2} \lesssim \left(\int_\tau r^{2\delta-2}\right)^{1/2} \lesssim \left(\int_0^{c\underline{h}} r^{2\delta-2} r \,\mathrm{d}r\right)^{1/2} \lesssim \underline{h}^{\delta}. \tag{5.11}$$

Here $c\underline{h}$ is the diameter of $\tilde{\tau}$. Combining (5.9), (5.10), and (5.11), we obtain

$$\|G_h v_I - \nabla v\|_{L^2(\tau)} \lesssim \underline{h}^{\delta} \quad \text{for } \tau \in \mathcal{M}^{\widetilde{O}}. \tag{5.12}$$

It follows from Lemma 5.1 and (1.3) that

$$\|G_h v_I - \nabla v\|_{L^2(\tau)} \lesssim h_\tau^3 \, |v|_{W^{3,\infty}(\tilde{\tau})} \lesssim h_\tau^3 r_\tau^{\delta-3} \quad \text{for } \tau \in \mathcal{M}_h \setminus \mathcal{M}^{\widetilde{O}}, \tag{5.13}$$

where r_τ is the distance form O to the barycenter of τ. Therefore from $h_\tau \eqsim r_\tau^{1-\delta/2} \underline{h}^{\delta/2}$,

$$\begin{aligned}
\|G_h v_I - \nabla v\|_{L^2(\Omega)}^2 &= \sum_{\tau\in\mathcal{M}_h} \|G_h v_I - \nabla v\|_{L^2(\tau)}^2 \lesssim \underline{h}^{2\delta} + \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^{\widetilde{O}}} h_\tau^6 r_\tau^{2\delta-6} \\
&\lesssim \underline{h}^{2\delta} + \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^{\widetilde{O}}} h_\tau^2 r_\tau^{4-2\delta} \underline{h}^{2\delta} r_\tau^{2\delta-6} \lesssim \underline{h}^{2\delta} + \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^{\widetilde{O}}} \underline{h}^{2\delta} h_\tau^2 r_\tau^{-2} \\
&\lesssim \underline{h}^{2\delta} + \underline{h}^{2\delta} \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^{\widetilde{O}}} \int_\tau r^{-2} \lesssim \underline{h}^{2\delta} + \underline{h}^{2\delta} \int_{\underline{h}}^1 r^{-1} \,\mathrm{d}r \lesssim \underline{h}^{2\delta} \\
&\quad + \underline{h}^{2\delta} \left|\ln \underline{h}\right|.
\end{aligned}$$

Therefore Lemma 2.1 implies that

$$\|G_h v_I - \nabla v\|_{L^2(\Omega)} \lesssim \underline{h}^{\delta} \left(1 + |\ln \underline{h}|^{1/2}\right) \lesssim \frac{1 + (\ln N)^{1/2}}{N}. \tag{5.14}$$

Next we estimate the term $\|G_h w_I - \nabla w\|_{L^2(\Omega)}$ in (5.8). Since w is smooth, we do not have to divide $\mathcal{M}_h$ into two parts as above. From Lemma 5.1 and assumption (1.3),

$$\begin{aligned}
\|G_h w_I - \nabla w\|_{L^2(\Omega)} &\lesssim \left(\sum_{\tau\in\mathcal{M}_h} \|G_h w_I - \nabla w\|_{L^2(\tau)}^2\right)^{1/2} \lesssim \left(\sum_{\tau\in\mathcal{M}_h} h_\tau^6\right)^{1/2} \\
&\lesssim \left(\sum_{\tau\in\mathcal{M}_h} h_\tau^2 r_\tau^{4-2\delta} \underline{h}^{2\delta}\right)^{1/2} \lesssim \underline{h}^{\delta} \left(\int_\Omega r^{4-2\delta}\right)^{1/2} \lesssim \underline{h}^{\delta} \lesssim \frac{1}{N}.
\end{aligned} \tag{5.15}$$

The proof of the theorem is completed by inserting estimates (5.14) and (5.15) into inequality (5.8). ∎

The following superconvergence result of the gradient recovery operator G_h can be proved by combining (5.3), Theorem 3.4, and Theorem 5.2.

THEOREM 5.3. *Let u_h be the linear finite element approximation of u. Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha, \sigma, \delta/2)$ *with $0 < \alpha \leq 1$ and $0 \leq \sigma < 1$, and that $h_\tau \eqsim r_\tau^{1-\delta/2} \underline{h}^{\delta/2}$ for any $\tau \in \mathcal{M}_h$. Then*

$$\|G_h u_h - \nabla u\|_{L^2(\Omega)} \lesssim \frac{1 + (\ln N)^{1/2}}{N^{1/2+\rho}}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right). \tag{5.16}$$

We remark that the result of Theorem 5.3 is a superconvergence result since the asymptotically optimal convergence rate of $\|\nabla(u - u_h)\|_{L^2(\Omega)}$ is $O(1/N^{1/2})$.

Next we state the results for quadratic finite elements. The following theorem provides the estimate for the gradient recovery operator G_h. The proof is similar to that of Theorem 5.2 and therefore is omitted.

THEOREM 5.4. *Assume that $h_\tau \eqsim r_\tau^{1-\delta/3} \underline{h}^{\delta/3}$ for any $\tau \in \mathcal{M}_h$. Then*

$$\left\|G_h I_h^2 u - \nabla u\right\|_{L^2(\Omega)} \lesssim \frac{1 + (\ln N)^{1/2}}{N^{3/2}}. \tag{5.17}$$

The superconvergence of the gradient recovery operator G_h is presented in the following theorem which is parallel to Theorem 5.3.

THEOREM 5.5. *Let u_h be the quadratic finite element approximation of u. Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha, \sigma, \delta/3)$ *with $0 < \alpha \leq 1$ and $0 \leq \sigma < 1$ and that $h_\tau \eqsim r_\tau^{1-\delta/3} \underline{h}^{\delta/3}$ for any $\tau \in \mathcal{M}_h$. Then*

$$\|G_h u_h - \nabla u\|_{L^2(\Omega)} \lesssim \frac{1 + (\ln N)^{1/2}}{N^{1+\rho}}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right). \tag{5.18}$$

5.2. The a posteriori error estimators. With preparation from the previous subsections, it is now straightforward to prove the asymptotic exactness of error estimators based on the recovery operator G_h. The global error estimator is naturally defined by

$$\eta_h = \|G_h u_h - \nabla u_h\|_{L^2(\Omega)}. \tag{5.19}$$

THEOREM 5.6. *Let u_h be the linear finite element approximation of u. Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha, \sigma, \delta/2)$ *with $0 < \alpha \leq 1$ and $0 \leq \sigma < 1$, and that $h_\tau \eqsim r_\tau^{1-\delta/2} \underline{h}^{\delta/2}$ for any $\tau \in \mathcal{M}_h$. Furthermore, assume that*

$$\frac{1}{N^{1/2}} \lesssim \|\nabla(u - u_h)\|_{L^2(\Omega)}. \tag{5.20}$$

Then

$$\left|\frac{\eta_h}{\|\nabla(u - u_h)\|_{L^2(\Omega)}} - 1\right| \lesssim \frac{1 + (\ln N)^{1/2}}{N^{\rho}}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right). \tag{5.21}$$

The following lemma says that $\|\nabla(u - u_h)\|_{L^2(\Omega)}$ is the asymptotically optimal on the mesh $\mathcal{M}_h$ satisfying $h_\tau \eqsim r_\tau^{1-\delta/2} \underline{h}^{\delta/2}$ as the total number of degrees of freedom $N \to \infty$.

LEMMA 5.7. *Let u_h be the linear finite element approximation of u. Assume that $h_\tau \eqsim r_\tau^{1-\delta/2}\underline{h}^{\delta/2}$ for any $\tau \in \mathcal{M}_h$. Then*

$$\|\nabla(u-u_I)\|_{L^2(\Omega)} \lesssim \frac{1}{N^{1/2}} \text{ and hence } \|\nabla(u-u_h)\|_{L^2(\Omega)} \lesssim \frac{1}{N^{1/2}}.$$

Proof. Recall that u is decomposed as $u = v + w$ satisfying (1.3). Noticing that

$$\|\nabla(v-v_I)\|_{L^2(\tau)} \lesssim h_\tau |v|_{H^2(\tau)} \lesssim h_\tau^2 r_\tau^{\delta-2} \quad \forall \tau \in \mathcal{M}_h \setminus \mathcal{M}^O,$$

and that

$$\|\nabla(w-w_I)\|_{L^2(\tau)} \lesssim h_\tau |w|_{H^2(\tau)} \lesssim h_\tau^2 \quad \forall \tau \in \mathcal{M}_h,$$

we have, by Lemma 3.2,

$$\begin{aligned}
\|\nabla(u-u_I)\|_{L^2(\Omega)}^2 &\lesssim \|\nabla(v-v_I)\|_{L^2(\Omega)}^2 + \|\nabla(w-w_I)\|_{L^2(\Omega)}^2 \\
&= \sum_{\tau\in\mathcal{M}_h} \left(\|\nabla(v-v_I)\|_{L^2(\tau)}^2 + \|\nabla(w-w_I)\|_{L^2(\tau)}^2 \right) \\
&\lesssim \underline{h}^{2\delta} + \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^O} h_\tau^4 r_\tau^{2\delta-4} \lesssim \underline{h}^{2\delta} + \sum_{\tau\in\mathcal{M}_h\setminus\mathcal{M}^O} h_\tau^2 r_\tau^{2-\delta}\underline{h}^{\delta} r_\tau^{2\delta-4} \\
&\lesssim \underline{h}^{2\delta} + \underline{h}^{\delta}\int_\Omega r^{\delta-2} \lesssim \underline{h}^{2\delta} + \underline{h}^{\delta}.
\end{aligned}$$

In light of Lemma 2.1, we obtain

$$\|\nabla(u-u_I)\|_{L^2(\Omega)}^2 \lesssim \frac{1}{N^2} + \frac{1}{N},$$

which completes the proof of the lemma. ∎

The following lemma says that, for the quadratic finite element approximation u_h, $\|\nabla(u-u_h)\|_{L^2(\Omega)}$ is asymptotically optimal on the mesh $\mathcal{M}_h$ satisfying $h_\tau \eqsim r_\tau^{1-\delta/3}\underline{h}^{\delta/3}$ as the total number of degrees of freedom $N \to \infty$.

LEMMA 5.8. *Let u_h be the quadratic finite element approximation of u. Assume that $h_\tau \eqsim r_\tau^{1-\delta/3}\underline{h}^{\delta/3}$ for any $\tau \in \mathcal{M}_h$. Then*

$$\left\|\nabla(u-I_h^2 u)\right\|_{L^2(\Omega)} \lesssim \frac{1}{N} \text{ and hence } \|\nabla(u-u_h)\|_{L^2(\Omega)} \lesssim \frac{1}{N}.$$

By Theorem 5.5, we can prove the asymptotic exactness of error estimators based on the recovery operator G_h for quadratic elements.

THEOREM 5.9. *Let u_h be the quadratic finite element approximation of u. Assume that $\mathcal{M}_h$ satisfies* Condition $(\alpha, \sigma, \delta/3)$ *with $0 < \alpha \le 1$ and $0 \le \sigma < 1$ and that $h_\tau \eqsim r_\tau^{1-\delta/3}\underline{h}^{\delta/3}$ for any $\tau \in \mathcal{M}_h$. Furthermore, assume that*

$$\frac{1}{N} \lesssim \|\nabla(u-u_h)\|_{L^2(\Omega)}. \tag{5.22}$$

Then

$$\left| \frac{\eta_h}{\|\nabla(u-u_h)\|_{L^2(\Omega)}} - 1 \right| \lesssim \frac{1+(\ln N)^{1/2}}{N^\rho}, \quad \rho = \min\left(\frac{\alpha}{2}, \frac{1-\sigma}{2}\right). \tag{5.23}$$

6. Implementation and numerical examples. In this section we present some numerical examples to verify the asymptotic exactness of the error estimator η_h based on the recovery operator G_h using quadratic finite elements. For examples on linear elements we refer to [12].

Implementation of the adaptive algorithm in this section is based on FEMLAB.[1] We define the local a posteriori error estimator on element τ as

$$\eta_\tau = \|G_h u_h - \nabla u_h\|_{L^2(\tau)},$$

and the global error estimator as

$$\eta_h = \left(\sum_{\tau \in \mathcal{M}_h} \eta_\tau^2 \right)^{1/2}.$$

Now we describe the adaptive algorithm used in this paper.

ALGORITHM. Given the tolerance $\texttt{TOL} > 0$,

- generate an initial mesh $\mathcal{M}_h$ over Ω;
- while $\eta_h > \texttt{TOL}$ do
  - choose a set of elements $\widehat{\mathcal{M}}_h \subset \mathcal{M}_h$ such that

$$\left(\sum_{\tau \in \widehat{\mathcal{M}}_h} \eta_\tau^2 \right)^{1/2} > 0.7 \left(\sum_{\tau \in \mathcal{M}_h} \eta_\tau^2 \right)^{1/2},$$

    then refine the elements in $\widehat{\mathcal{M}}_h$. Update the mesh $\mathcal{M}_h$.
  - solve the discrete problem (1.4) on $\mathcal{M}_h$.
  - compute error estimators on $\mathcal{M}_h$.

  end while

Remark 6.1. The marking strategy, that is, the method of how to choose $\widehat{\mathcal{M}}_h$ for refinements used in our algorithm, is well known in the adaptive finite element community. Actually, it was used, e.g., in [9, 18] to design convergent finite element algorithms. In our implementation of the algorithm, the elements in $\widehat{\mathcal{M}}_h$ are chosen from the elements which have larger local a posteriori error estimators η_τ.

Example 1. The Laplace equation on the L-shaped domain of Figure 6.1 with the Dirichlet boundary condition is chosen so that the true solution is $r^{2/3}\sin(2\theta/3)$ in polar coordinates.

Figure 6.1 plots the initial mesh and the adaptively refined mesh of 3565 elements after 15 adaptive iterations. Figure 6.2 demonstrates asymptotic exactness of the error estimator $\eta_h = \|G_h u_h - \nabla u_h\|_{L^2(\Omega)}$ for the Laplace equation on the L-shaped domain. We see that

$$\|\nabla u_h - \nabla u\|_{L^2(\Omega)} \approx O(N^{-1}), \quad \|G_h u_h - \nabla u\|_{L^2(\Omega)} \approx O(N^{-1.2}),$$

and

$$\|G_h u_h - \nabla u_h\|_{L^2(\Omega)} \Big/ \|\nabla u - \nabla u_h\|_{L^2(\Omega)} \approx 1 + O(N^{-0.5}).$$

[1] http://ecs.rutgers.edu/eitlab/femlab.php.

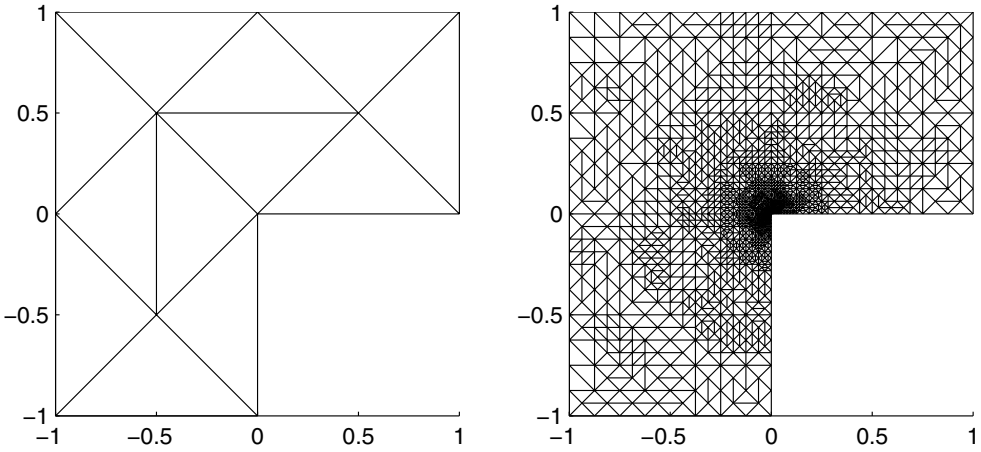


FIG. 6.1. *The initial mesh (left) and the adaptively refined mesh (right) of* 3565 *elements after* 15 *adaptive iterations for the Laplace equation on the L-shaped domain.*

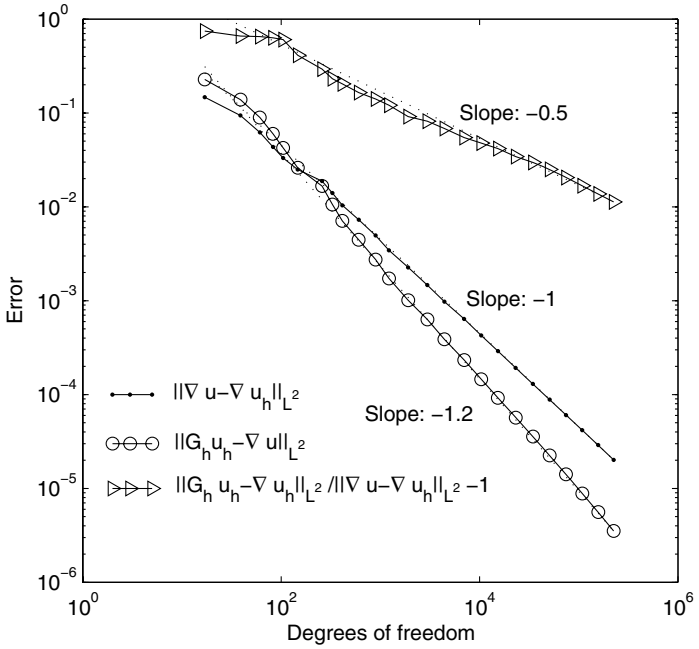


FIG. 6.2. $\|\nabla u - \nabla u_h\|_{L^2(\Omega)}$, $\|\nabla u - G_h u_h\|_{L^2(\Omega)}$, *and* $\|G_h u_h - \nabla u_h\|_{L^2(\Omega)} / \|\nabla u - \nabla u_h\|_{L^2(\Omega)} - 1$ *versus the total number of degrees of freedom for the Laplace equation on the L-shaped domain. Dotted lines give reference slopes.*

Notice that the decay of $\|\nabla u_h - \nabla u\|_{L^2(\Omega)}$ is quasi-optimal, $\|G_h u_h - \nabla u\|_{L^2(\Omega)}$ is superconvergent with order $O(N^{-1.2})$, and $\eta_h / \|\nabla u - \nabla u_h\|_{L^2(\Omega)}$ approaches 1 at the rate of $O(N^{-0.5})$. In this paper, the L^2 norms are calculated by the six points Gauss quadrature rule over triangles.

Let us have a close look at the mesh density assumption $h_\tau \approx r_\tau^{1-\delta/3} \underline{h}^{\delta/3} = r_\tau^{7/9} \underline{h}^{2/9}$ for $\delta = 2/3$. We shall verify this on the final mesh, which has 112880

elements after 24 adaptive iterations. We choose $\underline{h} = \min_{\tau\in\mathcal{M}_h} h_\tau \approx 5.96 \times 10^{-8}$ and have

$$0.44 \le \frac{h_\tau}{r_\tau^{7/9}\underline{h}^{2/9}} \le 2.35$$

for all elements $\tau \in \mathcal{M}_h$. Note that the ratio between the upper and lower bounds is less than 6. This fact indicates that all elements in the final mesh satisfy the mesh density assumption.

Next, let us examine the condition (α, σ, μ) on the final mesh. Here $\mu = \delta/3 = 2/9$. It is shown that, for every $e \in \mathcal{E}_h$, Ω_e is a $3.92 \times h_e^{1+0.4}/r_e^{0.4+\mu(1-0.4)}$ approximate parallelogram. That is, the final mesh satisfies Condition $(0.4, 0, 2/9)$.

Example 2. Let $\Omega = \{(x_1, x_2) : |x_1|, |x_2| < 0.5\}\backslash\{(x_1, 0) : 0 \le x_1 < 0.5\}$ be the domain with a crack. We consider the Poisson equation

$$-\Delta u = 1$$

with a Dirichlet boundary condition chosen so that the true solution is $r^{1/2}\sin(\theta/2) - \frac{1}{4}r^2$ in polar coordinates.

Figure 6.3 plots the initial mesh and the adaptively refined mesh of 3353 elements after 16 adaptive iterations. Figure 6.4 shows asymptotic exactness of the error estimator $\eta_h = \|G_h u_h - \nabla u_h\|_{L^2(\Omega)}$ for the crack problem. We see that

$$\|\nabla u_h - \nabla u\|_{L^2(\Omega)} \approx O(N^{-1}), \quad \|G_h u_h - \nabla u\|_{L^2(\Omega)} \approx O(N^{-1.1}),$$

and

$$\|G_h u_h - \nabla u_h\|_{L^2(\Omega)} \Big/ \|\nabla u - \nabla u_h\|_{L^2(\Omega)} \approx 1 + O(N^{-0.3}).$$

Notice that the decay of $\|\nabla u_h - \nabla u\|_{L^2(\Omega)}$ is quasi-optimal, $\|G_h u_h - \nabla u\|_{L^2(\Omega)}$ is superconvergent at an order $O(N^{-1.1})$, and $\eta_h / \|\nabla u - \nabla u_h\|_{L^2(\Omega)}$ approaches 1 at the rate of $O(N^{-0.3})$.

Let us take a close look at the mesh density assumption $h_\tau \eqsim r_\tau^{1-\delta/3}\underline{h}^{\delta/3} = r_\tau^{5/6}\underline{h}^{1/6}$ for $\delta = 1/2$. We verify this on the final mesh, which has 110563 elements after 27 adaptive iterations. We choose $\underline{h} = \min_{\tau\in\mathcal{M}_h} h_\tau \approx 3.67 \times 10^{-9}$ and have

$$0.32 < \frac{h_\tau}{r_\tau^{5/6}\underline{h}^{1/6}} < 1.92$$

for all elements $\tau \in \mathcal{M}_h$. Note that the ratio between the upper and lower bounds is 6. This fact indicates that all elements in the final mesh satisfy the mesh density assumption.

Next, let us examine the condition (α, σ, μ) on the final mesh. Here $\mu = \delta/3 = 1/6$. It is shown that, for every $e \in \mathcal{E}_h$, Ω_e is a $1.49 \times h_e^{1+0.2}/r_e^{0.2+\mu(1-0.2)}$ approximate parallelogram. That is, the final mesh satisfies Condition $(0.2, 0, 1/6)$.

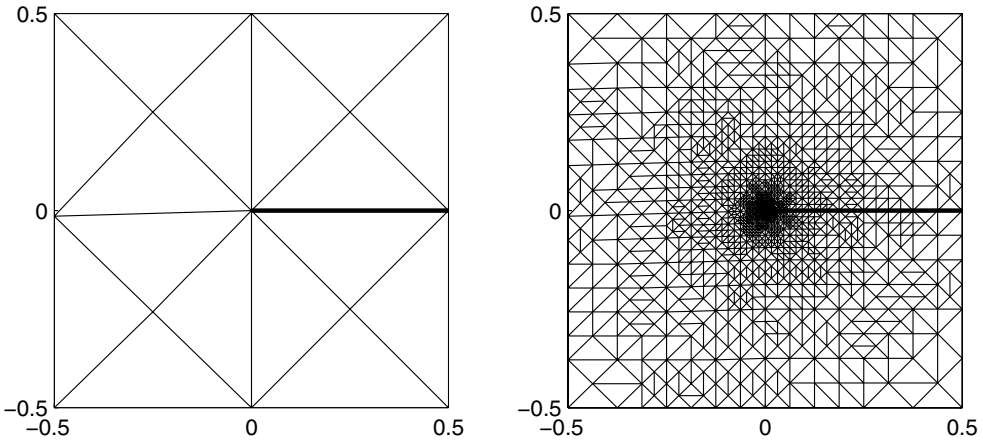


FIG. 6.3. *The initial mesh (left) and the adaptively refined mesh (right) of* 3353 *elements after* 16 *adaptive iterations for the crack problem.*

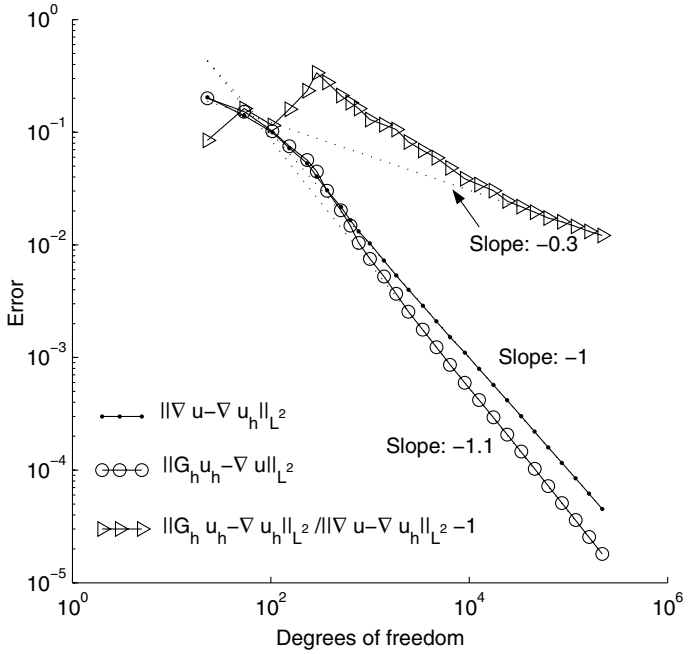


FIG. 6.4. $\|\nabla u - \nabla u_h\|_{L^2(\Omega)}$, $\|\nabla u - G_h u_h\|_{L^2(\Omega)}$, *and* $\|G_h u_h - \nabla u_h\|_{L^2(\Omega)} / \|\nabla u - \nabla u_h\|_{L^2(\Omega)} - 1$ *versus the total number of degrees of freedom for the crack problem. Dotted lines give reference slopes.*

Acknowledgments. The authors would like to thank a referee and the editor for many constructive and valuable suggestions, which considerably improved the presentation of the paper.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley Interscience, New York, 2000.

[2] W. Babgerth and R. Rannacher, *Adaptive Finite Element Methods for Differential Equations*, Birkhäuser, Basel, 2003.

[3] I. Babuška and M. Suri, *The p and h-p versions of the finite element method, basic principles and properties*, SIAM Rev., 36 (1994), pp. 578–632.

[4] R. E. Bank, *Hierarchical bases and the finite element method*, Acta Numer., (1996), pp. 1–43.

[5] R. E. Bank and J. Xu, *Asymptotically exact a posteriori error estimators, Part* I: *Grid with superconvergence*, SIAM J. Numer. Anal., 41 (2003), pp. 2294–2312.

[6] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 2002.

[7] C. Carstensen and S. Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids.* I. *Low order conforming, nonconforming, and mixed FEM*, Math. Comp., 71 (2002), pp. 945–969.

[8] L. Chen and J. Xu, *Topics on adaptive finite element methods*, in Adaptive Computations: Theory and Algorithms, T. Tang and J. Xu, eds., Science Press, Beijing, 2007.

[9] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[10] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numer., (1995), pp. 105–158.

[11] P. Grisvard, *Singularities in Boundary Value Problems*, Springer-Verlag, Berlin, 1992.

[12] F. Fierro and A. Veeser, *A posteriori error estimators, gradient recovery by averaging, and superconvergence*, Numer. Math., 103 (2006), pp. 267–298.

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

[14] Y. Huang and J. Xu, *Superconvergence of quadratic finite elements on mildly structured grids*, Math. Comp.

[15] A. M. Lakhany, I. Marek, and J. R. Whiteman, *Superconvergence results on mildly structured triangulations*, Comput. Methods Appl. Mech. Engrg., 189 (2000), pp. 1–75.

[16] B. Li and Z. Zhang, *Analysis of a class of superconvergence patch recovery techniques for linear and bilinear finite elements*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 151–167.

[17] K. Mekchay and R. H. Nochetto, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827.

[18] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.

[19] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.

[20] A. Naga and Z. Zhang, *A posteriori error estimates based on the polynomial preserving recovery*, SIAM J. Numer. Anal., 42 (2004), pp. 1780–1800.

[21] R. Verfürth, *A Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Teubner Skripten zur Numerik, B.G. Teubner, Stuttgart, 1995.

[22] J. Wang, *A superconvergence analysis for finite element solutions by the least squares surface fitting on irregular meshes for smooth problems*, J. Math. Stud., 33 (2000), pp. 229–243.

[23] J. Xu and Z. Zhang, *Analysis of recovery type a posteriori error estimators for mildly structured grids*, Math. Comp., 73 (2003), pp. 1139–1152.

[24] N. Yan and A. Zhou, *Gradient recovery type a posteriori error estimates for finite element approximations on irregular meshes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4289–4299.

[25] Z. Zhang, *Polynomial preserving gradient recovery and a posteriori estimate for bilinear element on irregular quadrilaterals*, Internat. J. Numer. Anal. Model., 1 (2004), pp. 1–24.

[26] Z. Zhang, *Polynomial preserving recovery for anisotropic and irregular grids*, J. Comput. Math., 22 (2004), pp. 331–340.

[27] Z. Zhang and R. Lin, *Ultraconvergence of zz patch recovery at mesh symmetry points*, Numer. Math., 95 (2003), pp. 781–801.

[28] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1723–1741

MESHLESS COLLOCATION: ERROR ESTIMATES WITH APPLICATION TO DYNAMICAL SYSTEMS*

PETER GIESL† AND HOLGER WENDLAND‡

Abstract. In this paper, we derive error estimates for generalized interpolation, in particular collocation, in Sobolev spaces. We employ our estimates in collocation problems using radial basis functions and extend and improve previously known results for elliptic problems. Finally, we use meshless collocation to approximate Lyapunov functions for dynamical systems.

Key words. partial differential equation, radial basis function, error estimates, Lyapunov function

AMS subject classifications. 65N15, 65N35, 37B25

DOI. 10.1137/060658813

1. Introduction. Meshless collocation methods for the numerical solution of partial differential equations have recently become more and more popular. They provide a greater flexibility when it comes to adaptivity and time-dependent changes of the underlying region.

Radial basis functions or, more generally, (conditionally) positive definite kernels are one of the mainstream methods in the field of meshless collocation. There are, in principle, two different approaches to collocation using radial basis functions. The *unsymmetric* approach by Kansa [14, 15] has the advantage that less derivatives have to be formed but has the drawback of an unsymmetric collocation matrix, which can even be singular [13]. Despite this drawback unsymmetric collocation has been used frequently and successfully in several applications.

In this paper, however, we will concentrate on *symmetric* collocation methods based on radial basis functions, as they have been introduced in the context of *generalized interpolation* in [28, 17] and used for elliptic problems in [4, 5, 7, 6].

Radial basis functions, in general, are a powerful tool for reconstruction processes from scattered data (see, for example, [3, 26]).

In this paper, we study a general linear partial differential equation of the form

$$Lu = f \text{ on } \Omega, \tag{1}$$

where Ω is a domain in $\mathbb{R}^n$ and L is a linear differential operator of the form

$$Lu(x) = \sum_{|\alpha| \le m} c_\alpha(x) D^\alpha u(x), \tag{2}$$

where the coefficients have a certain smoothness $c_\alpha \in C^\sigma(\overline{\Omega}, \mathbb{R})$; i.e., the derivatives of order β with $|\beta| \le \sigma$ exist and are continuous on $\overline{\Omega}$.

*Received by the editors May 3, 2006; accepted for publication (in revised form) January 22, 2007; published electronically August 24, 2007.
http://www.siam.org/journals/sinum/45-4/65881.html
†Zentrum Mathematik, TU München, Boltzmannstr. 3, D-85747 Garching bei München, Germany. (giesl@ma.tum.de).
‡Department of Mathematics, University of Sussex, Brighton BN1 9RF, UK (H.Wendland@sussex.ac.uk).

Moreover, we consider boundary value problems, where in addition to (1), u is required to satisfy the following boundary condition:

$$u(x) = F(x) \text{ for } x \in \partial\Omega. \tag{3}$$

The numerical solution of such boundary value problems by collocation using radial basis functions has been studied by several authors. First error estimates have been given in [7, 6]. However, despite following a rather general approach, the authors of those papers show that the problems are well-posed and provide error estimates only for differential operators with *constant coefficients* c_α. A generalization to nonconstant coefficients without zeros, including also a more thorough discussion of the boundary estimates, can be found in [26]. However, in that book the approximation orders are, to a certain extent, not optimal. Moreover, the restriction to nonzero coefficients is not sufficient for our applications in dynamical systems.

Our goals in this paper are to investigate well-posedness of the collocation problem for the differential operator (2) with *nonconstant* coefficients and to state error estimates with optimal orders in Sobolev spaces. To this end we will put the setting in the general framework of generalized interpolation in reproducing kernel Hilbert spaces and then use a recent result [18] on error estimates in Sobolev spaces for *arbitrary* scattered data reconstruction methods.

Next, we will apply the general estimates to derive error estimates in Sobolev spaces for elliptic partial differential equations. Another major and new application will be the approximation of Lyapunov functions in dynamical systems. Here, the differential operator is given by the *orbital derivative* of a function u with respect to the ordinary differential equation $\dot{x} = g(x)$, i.e., by

$$Lu(x) := \langle \nabla u(x), g(x) \rangle = \sum_{j=1}^{n} g_j(x) \partial_j u(x).$$

This operator L is a first-order differential operator of the form (2) with $c_{e^j}(x) = g_j(x)$. The approximation of the orbital derivative for Lyapunov functions has been studied in [11, 8, 9, 10]. However, the approximation orders of those results can be improved significantly with the results of this paper.

This paper is organized as follows. In the rest of this section we will introduce notation which is necessary throughout the paper. Section 2 deals with generalized interpolation and is mainly a collection of known results, which will be helpful in this paper. In section 3 we investigate collocation by radial basis functions, derive our new estimates, and apply these results to elliptic problems. The final section deals with applications to dynamical systems. In particular, we describe a method to calculate Lyapunov functions and thus to calculate the basin of attraction of an equilibrium.

1.1. Notation. We will need to work with a variety of Sobolev spaces. Let $\Omega \subseteq \mathbb{R}^n$ be a domain. For $k \in \mathbb{N}_0$, and $1 \le p < \infty$, the Sobolev spaces $W_p^k(\Omega)$ consist, as usual, of all u with weak derivatives $D^\alpha u \in L_p(\Omega)$, $|\alpha| \le k$. Associated with these spaces are the (semi-)norms

$$|u|_{W_p^k(\Omega)} = \left(\sum_{|\alpha|=k} \|D^\alpha u\|_{L_p(\Omega)}^p \right)^{1/p} \text{ and } \|u\|_{W_p^k(\Omega)} = \left(\sum_{|\alpha|\le k} \|D^\alpha u\|_{L_p(\Omega)}^p \right)^{1/p}.$$

The case $p = \infty$ is defined in the obvious way:

$$|u|_{W_\infty^k(\Omega)} = \sup_{|\alpha|=k} \|D^\alpha u\|_{L_\infty(\Omega)} \quad \text{and} \quad \|u\|_{W_\infty^k(\Omega)} = \sup_{|\alpha|\le k} \|D^\alpha u\|_{L_\infty(\Omega)}.$$

We also need fractional order Sobolev spaces. However, it will not be necessary to discuss them in detail. We just remind the reader that there are different ways of introducing fractional order Sobolev spaces. For our purposes, interpolation theory in Sobolev spaces as it has, for example, been discussed in [1, 23, 2] will be sufficient.

Let $X := \{x_1, \dots, x_N\}$ be a finite, discrete subset of Ω, which we now assume to be bounded. There are two quantities that we associate with X: the *separation radius* and the *mesh norm* or *fill distance*. Respectively, these are given by

$$q_X := \frac{1}{2} \min_{j \neq k} \|x_j - x_k\|_2, \quad h_{X,\Omega} := \sup_{x \in \Omega} \min_{x_j \in X} \|x - x_j\|_2,$$

where $\|\cdot\|_2$ denotes the Euclidean distance in $\mathbb{R}^n$.

The first is half the smallest distance between points in X, and the second measures the maximum distance which a point in Ω can be from any point in X. Frequently, when it is clear from the context what the set Ω (or X) is, we will drop subscripts and write h_X or h. Other notation will be introduced along the way.

2. Generalized interpolation.

2.1. Reproducing kernel Hilbert spaces. Let $H \subseteq C(\Omega)$ be a Hilbert space of functions $f : \Omega \to \mathbb{R}$ and let H^* be its dual. We consider a generalized interpolation problem of the following form.

DEFINITION 2.1. *Given N linearly independent functionals $\lambda_1, \dots, \lambda_N \in H^*$ and N function values $f_1, \dots, f_N \in \mathbb{R}$, a generalized interpolant is a function $s \in H$ satisfying $\lambda_j(s) = f_j$, $1 \leq j \leq N$. The norm-minimal interpolant s^* is the interpolant that, in addition, minimizes the norm of the Hilbert space; i.e., s^* is the solution of*

$$\min\{\|s\|_H : \lambda_j(s) = f_j, 1 \leq j \leq N\}. \tag{4}$$

It is well known that the norm-minimal generalized interpolant is a linear combination of the Riesz representers of the functionals and that the coefficients can be computed by solving a linear system. Such problems can be best solved if H is a reproducing kernel Hilbert space (RKHS), i.e., if there exists a unique kernel $\Phi : \Omega \times \Omega \to \mathbb{R}$, satisfying

1. $\Phi(\cdot, x) \in H$ for all $x \in \Omega$,
2. $f(x) = (f, \Phi(\cdot, x))_H$ for all $x \in \Omega$ and all $f \in H$.

Here, the Riesz representer of a functional $\lambda \in H^*$ is simply given by applying it to one argument of the kernel, i.e., by $\lambda^y \Phi(\cdot, y)$.

LEMMA 2.2 (see [26, Theorem 16.1]). *If H is a reproducing kernel Hilbert space, then the solution s^* of (4) is given by*

$$s^* = \sum_{j=1}^{N} \alpha_j \lambda_j^y \Phi(\cdot, y),$$

where $\alpha \in \mathbb{R}^N$ is the solution of the linear system $A_{\Lambda,\Phi}\alpha = f$ with $A_{\Lambda,\Phi} = (\lambda_i^x \lambda_j^y \Phi(x, y))$ and $f = (f_j)$.

Note that the matrix $A_{\Lambda,\Phi} = (a_{ij})$ is a Gramian matrix because of

$$a_{ij} = \lambda_i^x \lambda_j^y \Phi(x, y) = (\lambda_i^x \Phi(\cdot, x), \lambda_j^y \Phi(\cdot, y))_H = (\lambda_i, \lambda_j)_{H^*}$$

and hence is positive semidefinite. Since the functionals are assumed to be linearly independent the matrix is even positive definite.

Looking at point evaluations $\lambda_j(f) = \delta_{x_j}(f) = f(x_j)$ alone, we see that the kernel of a reproducing kernel Hilbert space is *positive definite* in the sense that all the matrices

$$(\Phi(x_i, x_j))_{1\le i,j\le N}$$

are positive definite, provided that point evaluation functionals are linearly independent.

Now, it is easy to see that the kernel of a reproducing kernel Hilbert space is uniquely determined. On the other hand, the Hilbert space is also uniquely determined by the kernel. Moreover, every positive definite kernel generates a unique Hilbert space to which it is the reproducing kernel. More details about this fact and the construction of such *native* function spaces can be found in [26]. Here, the only thing that matters is that two *different* kernels can generate the *same* function Hilbert space H but with different, yet equivalent, inner products. In such a situation we will say that both kernels are reproducing kernels of H, thus relaxing the definition of a reproducing kernel Hilbert space. Moreover, it will be helpful to consider kernels defined on all $\mathbb{R}^n$ instead of on only $\Omega \subseteq \mathbb{R}^n$. Such kernels are often *translation-invariant*, meaning $\Phi(x, y) = \Phi(x - y)$, and are often even *radial*, meaning $\Phi(x, y) = \Phi(\|x - y\|_2)$.

This will be very useful when it comes to Sobolev spaces. Remember that the Sobolev embedding theorem states that $W_2^\tau(\mathbb{R}^n)$ can be embedded into $C(\mathbb{R}^n)$ provided that $\tau > n/2$. Hence, in this situation $W_2^\tau(\mathbb{R}^n)$ is a reproducing kernel Hilbert space. Unfortunately, the reproducing kernel involves some modified Bessel functions of the third kind.

However, it is well known that other reproducing kernels of $W_2^\tau(\mathbb{R}^n)$ can be characterized by their Fourier transform

$$\widehat{\Phi}(\omega) = (2\pi)^{-n} \int_{\mathbb{R}^n} \Phi(x) e^{-ix^T\omega} dx.$$

To be more precise, the following result holds.

LEMMA 2.3 (see [26, Corollary 10.13]). *Let $\tau > n/2$. Suppose the Fourier transform of an integrable function $\Phi : \mathbb{R}^n \to \mathbb{R}$ satisfies*

$$c_1(1 + \|\omega\|_2^2)^{-\tau} \le \widehat{\Phi}(\omega) \le c_2(1 + \|\omega\|_2^2)^{-\tau}, \qquad \omega \in \mathbb{R}^n, \tag{5}$$

with two constants $c_2 \ge c_1 > 0$. Then, the kernel Φ is also a reproducing kernel of $W_2^\tau(\mathbb{R}^n)$, and the inner product defined by

$$(f, g) := \int_{\mathbb{R}^n} \frac{\widehat{f}(\omega)\overline{\widehat{g}(\omega)}}{\widehat{\Phi}(\omega)} d\omega$$

is equivalent to the usual inner product on $W_2^\tau(\mathbb{R}^n)$.

The following observation will be of use. It follows directly from the Fourier inversion theorem.

Remark 2.4. If $\Phi \in L_1(\mathbb{R}^n)$ satisfies (5) with $\tau > m + n/2$, then $\Phi \in C^{2m}(\mathbb{R}^n)$.

The most prominent examples of kernels satisfying (5) are the Wendland functions [24, 25]. They are positive definite and radial functions with compact support. On its support, each function can be represented by a univariate polynomial. Here it is mainly important that they satisfy (5) with $\tau = k + (n + 1)/2$, where k is a given smoothness index. Hence, they belong to $C^{2k}(\mathbb{R}^n)$ and generate integer order Sobolev

spaces in odd space dimensions, while for even space dimensions the order is integer plus one half.

Although most kernels which generate Sobolev spaces are radial, there exist also kernels which are not even translation invariant; cf. [21, 20]. Our results will hold regardless of whether the kernels are translation invariant or not.

We end this section by citing a general convergence result from [18] in its improved form (see the remarks in [19]) using also the fact that a region with a Lipschitz boundary automatically satisfies a cone condition (see [27]).

THEOREM 2.5. *Let $\Omega \subseteq \mathbb{R}^n$ be a bounded domain with a Lipschitz continuous boundary. Let $1 \leq p < \infty$, $1 \leq q \leq \infty$, and let $m \in \mathbb{N}_0$ and $\tau \in \mathbb{R}$ satisfying $\lfloor \tau \rfloor > m + n/p$ if $p > 1$, or $\lfloor \tau \rfloor \geq m + n$ if $p = 1$. Also, let $X \subseteq \Omega$ be a discrete set with a sufficiently small mesh norm h. If $u \in W_p^\tau(\Omega)$ satisfies $u|X = 0$, then*

$$|u|_{W_q^m(\Omega)} \leq C h^{\tau - m - n(1/p - 1/q)_+} |u|_{W_p^\tau(\Omega)}, \tag{6}$$

where $(x)_+ = \max\{x, 0\}$.

3. Partial differential equations (PDEs).

3.1. General PDE operators. It is now time to look at specific collocation problems. We start with the PDE (1). Following the general approach of the previous section, we define functionals

$$\lambda_j(u) := \delta_{x_j} \circ L(u) = (Lu)(x_j)$$

with scattered points $X = \{x_1, \ldots, x_N\} \subseteq \Omega$. Hence, employing a sufficiently smooth kernel $\Phi : \Omega \times \Omega \to \mathbb{R}$ results in the approximating function

$$s = \sum_{k=1}^{N} \alpha_k (\delta_{x_k} \circ L)^y \Phi(\cdot, y). \tag{7}$$

Applying the interpolation conditions yields the following interpolation problem.

DEFINITION 3.1 (interpolation problem, operator). *Let $X = \{x_1, \ldots, x_N\}$ be a set of pairwise distinct points in $\Omega \subseteq \mathbb{R}^n$ and $u : \Omega \to \mathbb{R}$. Let L be a linear differential operator. Then, the* reconstruction s of u with respect to the set X and the operator L *is given by* (7)*, where the coefficient vector α is the solution of $A\alpha = f = (f_j)$ with the interpolation matrix $A = (a_{jk})_{j,k=1,\ldots,N}$ given by*

$$a_{jk} = (\delta_{x_j} \circ L)^x (\delta_{x_k} \circ L)^y \Phi(x, y) \tag{8}$$

and $f_j = (\delta_{x_j} \circ L)^x u(x) = Lu(x_j)$.

According to Lemma 2.2, the generalized interpolation matrix is positive definite, provided that the involved functionals are linearly independent.

DEFINITION 3.2 (singular points of L). *The point $x \in \mathbb{R}^n$ is called a* singular point of L *if $\delta_x \circ L = 0$, i.e., $c_\alpha(x) = 0$ for all $|\alpha| \leq m$.*

PROPOSITION 3.3. *Suppose $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a reproducing kernel of $W_2^\tau(\mathbb{R}^n)$ with $\tau > m + n/2$. Let L be a linear differential operator of degree m. Let $X = \{x_1, \ldots, x_N\}$ be a set of pairwise distinct points, which are not singular points of L. Then, the functionals $\lambda_j = \delta_{x_j} \circ L$ are linearly independent over $W_2^\tau(\mathbb{R}^n)$.*

Proof. First of all note that, according to Remark 2.4, (8) is well defined for reproducing kernels of $W_2^\tau(\mathbb{R}^n)$ even with $\tau > m + n/2$. Moreover, the functionals are indeed in the dual space to $W_2^\tau(\mathbb{R}^n)$.

Next, suppose that

$$\sum_{k=1}^{N} d_k \lambda_k = 0 \tag{9}$$

on $W_2^\tau(\mathbb{R}^n)$ with certain coefficients $d_1, \dots, d_N$.

Then we choose a flat bump function $g \in C_0^\infty(\mathbb{R}^n)$, i.e., a nonnegative, compactly supported function with support $B(0,1) = \{x \in \mathbb{R}^n : \|x\|_2 \le 1\}$ which is nonvanishing and satisfies $g(x) = 1$ on $B(0,1/2)$. Fix $1 \le j \le N$. Since x_j is not a singular point of L, there exists a $\beta \in \mathbb{N}_0^n$ with minimal $|\beta| \le m$ such that $c_\beta(x_j) \neq 0$. Employing the separation radius q_X, the function

$$g_j(x) = \frac{1}{\beta!}(x - x_j)^\beta g\left(\frac{x - x_j}{q_X}\right)$$

then satisfies $D^\alpha g_j(x_k) = 0$ for all $|\alpha| \le m$ and $x_k \neq x_j$. Furthermore, we have $D^\alpha g_j(x_j) = 0$ if $\alpha \neq \beta$ and $D^\beta g_j(x_j) = 1$. Hence, (9) gives, in particular,

$$0 = \sum_{k=1}^{N} d_k \lambda_k(g_j) = \sum_{|\alpha| \le m} \sum_{k=1}^{N} d_k c_\alpha(x_k) D^\alpha g_j(x_k) = d_j c_\beta(x_j),$$

which implies $d_j = 0$. Since j was chosen arbitrarily, this shows that the functionals are linearly independent. ∎

This proposition is a generalization of the results in [6], where only constant coefficients have been allowed, and of the results in [26], where also variable coefficients without zeros were treated.

Note also that the reproducing kernel Hilbert space does not have to be a Sobolev space at all. It is necessary only that the Hilbert space contains bump functions of the described form. Hence, the results remain true, if, for example, function spaces associated with Gaussians or (inverse) multiquadrics are considered.

Next we turn to error estimates. We need a simple auxiliary result.

LEMMA 3.4. *Fix $\tau \in \mathbb{R}$ with $k = \lfloor \tau \rfloor > n/2 + m$, where m is the order of the differential operator L. Suppose that the coefficients c_α of the differential operator L belong to $W_\infty^{k-m+1}(\Omega)$. Then L is a bounded operator from $W_2^\tau(\Omega)$ to $W_2^{\tau-m}(\Omega)$, i.e.,*

$$\|Lu\|_{W_2^{\tau-m}(\Omega)} \le C \|u\|_{W_2^\tau(\Omega)}, \qquad u \in W_2^\tau(\Omega).$$

Proof. Take a multi-index $\alpha \in \mathbb{N}_0^n$ with $|\alpha| \le k + 1 - m$. Then

$$\begin{aligned} |D^\alpha(Lu)| &= \left| \sum_{|\beta| \le m} \sum_{\gamma \le \alpha} \binom{\alpha}{\gamma} (D^{\alpha-\gamma} c_\beta)(D^{\gamma+\beta} u) \right| \\ &\le C \sum_{|\beta| \le m} \sum_{\gamma \le \alpha} |D^{\gamma+\beta} u|, \end{aligned}$$

where we used the boundedness of the derivatives of the coefficients. This shows that

$$\|D^\alpha(Lu)\|_{L_2(\Omega)} \le C \|u\|_{W_2^{m+|\alpha|}(\Omega)},$$

and hence

$$\|Lu\|_{W_2^{k-m}(\Omega)} \le C \|u\|_{W_2^k(\Omega)}, \qquad \|Lu\|_{W_2^{k+1-m}(\Omega)} \le C \|u\|_{W_2^{k+1}(\Omega)}.$$

From this, the result for fractional order Sobolev spaces $W_2^\tau(\Omega)$ follows by interpolation theory. ∎

THEOREM 3.5. *Suppose Φ is a reproducing kernel of $W_2^\tau(\mathbb{R}^n)$ with $k := \lfloor \tau \rfloor > m + n/2$. Let $\Omega \subseteq \mathbb{R}^n$ be a bounded domain having a Lipschitz boundary. Let L be a linear differential operator of order m with coefficients c_α in $W_\infty^{k-m+1}(\Omega)$. Finally, let s be the generalized interpolant to $u \in W_2^\tau(\Omega)$ from Definition 3.1. If $X \subseteq \Omega$ has a sufficiently small mesh norm h_X, then for $1 \le p \le \infty$, the error estimate*

$$\|Lu - Ls\|_{L_p(\Omega)} \le C h_X^{\tau - m - n(1/2 - 1/p)_+} \|u\|_{W_2^\tau(\Omega)}$$

is satisfied.

Proof. Note that $u \in W_2^\tau(\Omega) \subseteq C^m(\mathbb{R}^n)$ by assumption, while $s \in C^m(\mathbb{R}^n)$ by Remark 2.4. Hence, application of L is feasible.

Since $Lu|X = Ls|X$ by definition, we can apply Theorem 2.5 to derive

$$\begin{aligned}\|Lu - Ls\|_{L_p(\Omega)} &\le C h_X^{\tau - m - n(1/2 - 1/p)_+} \|Lu - Ls\|_{W_2^{\tau - m}(\Omega)} \\ &\le C h_X^{\tau - m - n(1/2 - 1/p)_+} \|u - s\|_{W_2^\tau(\Omega)},\end{aligned}$$

where we have also used Lemma 3.4.

Next, we follow the ideas in [18]. Our assumptions on the region Ω allow us to extend the function $u \in W_2^\tau(\Omega)$ to a function $Eu \in W_2^\tau(\mathbb{R}^n)$. Moreover, since $X \subseteq \Omega$ and $Eu|\Omega = u|\Omega$, the generalized interpolant $s = s_u$ to u coincides with the generalized interpolant s_{Eu} to Eu on Ω. Finally, the Sobolev space norm on $W_2^\tau(\mathbb{R}^n)$ is equivalent to the norm induced by the kernel Φ on $W_2^\tau(\mathbb{R}^n)$ (Lemma 2.3) and the generalized interpolant is norm-minimal (Lemma 2.2). This all gives

$$\begin{aligned}\|u - s\|_{W_2^\tau(\Omega)} = \|Eu - s_{Eu}\|_{W_2^\tau(\Omega)} &\le \|Eu - s_{Eu}\|_{W_2^\tau(\mathbb{R}^n)} \\ &\le C \|Eu\|_{W_2^\tau(\mathbb{R}^n)} \le C \|u\|_{W_2^\tau(\Omega)},\end{aligned}$$

which establishes the stated error estimate. ∎

The most important choices of $p = 2$ and $p = \infty$ yield

$$\|Lu - Ls\|_{L_2(\Omega)} \le C h_X^{\tau - m} \|u\|_{W_2^\tau(\Omega)},$$

$$\|Lu - Ls\|_{L_\infty(\Omega)} \le C h_X^{\tau - m - n/2} \|u\|_{W_2^\tau(\Omega)}.$$

As a consequence, using Wendland's compactly supported functions, we have to set $\tau = k + (n+1)/2$, where k is the smoothness index of the compactly supported functions, i.e., $\Phi = \psi_{\ell,k}(c\|\cdot\|_2) \in C^{2k}(\mathbb{R}^n)$. Note that this k is different from the k in Theorem 3.5. As a matter of fact the k in that theorem is given by $\lfloor \tau \rfloor = k + \lfloor (n+1)/2 \rfloor$.

COROLLARY 3.6. *Denote by k the smoothness index of the compactly supported Wendland function. Let $k > m - \frac{1}{2}$ if n is odd or $k > m$ if n is even. Let $c_\alpha \in W_\infty^{k-m+1+\lfloor \frac{n+1}{2} \rfloor}$. Suppose $u \in W_2^{k+(n+1)/2}(\Omega)$. Then, employing this basis function yields*

$$\|Lu - Ls\|_{L_\infty(\Omega)} \le C h_X^{k-m+1/2} \|u\|_{W_2^{k+(n+1)/2}(\Omega)}.$$

3.2. Boundary value problems. The collocation problem of the previous section will already be useful in its form in our application to dynamical systems; however, boundary value problems also will occur; cf. section 4. Furthermore, for applications such as solving elliptic PDEs, incorporating boundary values is crucial.

In order to solve a boundary value problem of the form (1) and (3), we need two linear operators L and $L^0 = \mathrm{id}$, the values of which are given on Ω, $\partial\Omega$, respectively. The ansatz for the approximating function s reflects this. We choose two sets of points, $X_1 := \{x_1, \dots, x_N\} \subseteq \Omega$ and $X_2 := \{x_{N+1}, \dots, x_{N+M}\} \subseteq \partial\Omega$, and define the functionals by

$$\lambda_j = \begin{cases} \delta_{x_j} \circ L & \text{for } 1 \le j \le N, \\ \delta_{x_j} \circ L^0 & \text{for } N+1 \le j \le N+M. \end{cases} \tag{10}$$

The mixed ansatz for the approximant s of the function u is then given by

$$\begin{aligned} s(x) &= \sum_{k=1}^{N+M} \alpha_k \lambda_k^y \Phi(x,y) \\ &= \sum_{k=1}^{N} \alpha_k (\delta_{x_k} \circ L)^y \Phi(x,y) + \sum_{k=N+1}^{N+M} \alpha_k (\delta_{x_k} \circ L^0)^y \Phi(x,y), \end{aligned} \tag{11}$$

where we will assume that $L^0 = \mathrm{id}$. The coefficient vector $\alpha \in \mathbb{R}^{N+M}$ is determined by the interpolation conditions

$$(\delta_{x_j} \circ L)(s) = (\delta_{x_j} \circ L)(u) = f(x_j), \quad 1 \le j \le N, \tag{12}$$

$$(\delta_{x_j} \circ L^0)(s) = (\delta_{x_j} \circ L^0)(u) = F(x_j), \quad N+1 \le j \le N+M. \tag{13}$$

Plugging the ansatz (11) into both (12) and (13) gives the following.

DEFINITION 3.7 (mixed interpolation problem). *Let $u\colon \Omega \to \mathbb{R}$ be the solution of* (1) *and* (3). *Let $X_1 = \{x_1, \dots, x_N\} \subseteq \Omega$ and $X_2 := \{x_{N+1}, \dots, x_{N+M}\} \subseteq \partial\Omega$ be two sets of pairwise distinct points. Then the* collocation reconstruction *s of u based upon X_1 and X_2 and the kernel Φ is given by* (11), *where the coefficient vector is determined by solving the linear system $\widetilde{A}\alpha = \beta$, with the interpolation matrix*

$$\widetilde{A} := \begin{pmatrix} A & C \\ C^T & A^0 \end{pmatrix} \in \mathbb{R}^{(N+M)\times(N+M)} \tag{14}$$

having submatrices $A = (a_{ij}) \in \mathbb{R}^{N\times N}$, $C = (c_{ij}) \in \mathbb{R}^{N\times M}$, and $A^0 = (a^0_{ij}) \in \mathbb{R}^{M\times M}$ with elements

$$\begin{aligned} a_{i,j} &= (\delta_{x_i} \circ L)^x (\delta_{x_j} \circ L)^y \Phi(x,y), \\ c_{i,\ell-N} &= (\delta_{x_i} \circ L)^x (\delta_{x_\ell} \circ L^0)^y \Phi(x,y), \\ a^0_{k-N,\ell-N} &= (\delta_{x_k} \circ L^0)^x (\delta_{x_\ell} \circ L^0)^y \Phi(x,y) \end{aligned}$$

for $1 \le i, j \le N$, $N+1 \le k, \ell \le N+M$.

The right-hand side of the linear system is determined by $\beta_j = f(x_j)$ for $1 \le j \le N$ and $\beta_j = F(x_j)$ for $N+1 \le j \le N+M$, respectively.

As in the case of one operator, it is easy to show that the functionals λ_j, this time defined by (10), are linearly independent.

PROPOSITION 3.8. *Suppose $\Phi : \mathbb{R}^n \to \mathbb{R}$ is a reproducing kernel of $W_2^\tau(\mathbb{R}^n)$ with $\tau > m + n/2$. Let L be a linear differential operator of degree m. Let $X_1 = \{x_1, \dots, x_N\} \subseteq \Omega$ and $X_2 = \{x_{N+1}, \dots, x_{N+M}\} \subseteq \partial\Omega$ be two sets of pairwise distinct points such that X_1 contains no singular point of L. Then, the functionals $\Lambda = \{\lambda_1, \dots, \lambda_{N+M}\}$ with $\lambda_j = \delta_{x_j} \circ L$ for $1 \le j \le N$ and $\lambda_j = \delta_{x_j}$ for $N+1 \le j \le N+M$ are linearly independent over $W_2^\tau(\mathbb{R}^n)$.*

Next we turn to error estimates. To this end we have to make certain further assumptions on the boundary.

We will assume that the bounded region $\Omega \subseteq \mathbb{R}^n$ has a $C^{k,s}$-boundary $\partial\Omega$, where $\tau = k + s$ with $k \in \mathbb{N}_0$ and $s \in [0, 1)$. This means, in particular, that $\partial\Omega$ is an $n-1$ dimensional $C^{k,s}$-submanifold of $\mathbb{R}^n$. It also means that Ω is Lipschitz continuous and satisfies the cone condition. For details, we refer the reader to [27].

We will represent the boundary $\partial\Omega$ by a finite atlas consisting of $C^{k,s}$-diffeomorphisms with a slight abuse of terminology. To be more precise, we assume that $\partial\Omega \subseteq \cup_{j=1}^K V_j$, where $V_j \subseteq \mathbb{R}^n$ are open sets. Moreover, the sets V_j are images of $C^{k,s}$-diffeomorphisms

$$\varphi_j : B \to V_j,$$

where $B = B(0, 1)$ denotes the unit ball in $\mathbb{R}^{n-1}$. Finally, suppose $\{w_j\}$ is a partition of unity with respect to $\{V_j\}$. Then the Sobolev norms on $\partial\Omega$ can be defined via

$$\|u\|^p_{W_p^\mu(\partial\Omega)} = \sum_{j=1}^K \|(uw_j) \circ \varphi_j\|^p_{W_p^\mu(B)}.$$

It is well known that this norm is independent of the chosen atlas $\{V_j, \varphi_j\}$, but this is of less importance here since we will assume that the atlas is fixed. For us, the next well known result will play a crucial role.

LEMMA 3.9 (trace theorem [27, Theorem 8.7]). *Suppose $\Omega \subseteq \mathbb{R}^n$ is a bounded region with a $C^{k,s}$-boundary $\partial\Omega$. Then, the restriction of $u \in W_2^\tau(\Omega)$ with $\tau = k + s$ to $\partial\Omega$ is well defined, belongs to $W_2^{\tau-1/2}(\partial\Omega)$, and satisfies*

$$\|u\|_{W_2^{\tau-1/2}(\partial\Omega)} \le C\|u\|_{W_2^\tau(\Omega)}.$$

Moreover, we now have two different mesh norms, $h_{X_1,\Omega}$ for the domain part and $h_{X_2,\partial\Omega}$ for the boundary part. Using the atlas $\{V_j, \varphi_j\}$, we simply define the latter to be

$$h_{X_2,\partial\Omega} := \max_{1 \le j \le K} h_{T_j,B}$$

with $T_j = \varphi_j^{-1}(X_2 \cap V_j) \subseteq B$. As mentioned before, we will assume the atlas is fixed and hence will not be concerned about the dependence of $h_{X_2,\partial\Omega}$ on the atlas.

THEOREM 3.10. *Suppose Φ is the reproducing kernel of $W_2^\tau(\mathbb{R}^n)$ with $k := \lfloor\tau\rfloor > m + n/2$. Let $\Omega \subseteq \mathbb{R}^n$ be a bounded domain having a $C^{k,s}$-boundary. Let L be a linear differential operator of order m with coefficients c_α in $W_\infty^{k-m+1}(\Omega)$. Finally, let s be the generalized interpolant to $u \in W_2^\tau(\Omega)$ from Definition 3.7. If the data sets have sufficiently small mesh norms, then for $1 \le p \le \infty$, the error estimates*

$$\|Lu - Ls\|_{L_p(\Omega)} \le Ch^{\tau-m-n(1/2-1/p)_+}_{X_1,\Omega}\|u\|_{W_2^\tau(\Omega)}, \tag{15}$$

$$\|u - s\|_{L_p(\partial\Omega)} \le Ch^{\tau-1/2-(n-1)(1/2-1/p)_+}_{X_2,\partial\Omega}\|u\|_{W_2^\tau(\Omega)} \tag{16}$$

are satisfied.

Proof. Estimate (15) follows as in Theorem 3.5. For the second estimate, note that the functions $u_j = ((u-s)w_j) \circ \varphi_j$ belong to $W_2^{\tau-1/2}(B)$ and vanish on T_j. Hence, using the definition of the Sobolev norm on $\partial\Omega$ and Theorem 2.5 yields

$$\begin{aligned}\|u-s\|_{L_p(\partial\Omega)}^p &= \sum_{j=1}^K \|u_j\|_{L_p(B)}^p \\ &\le C \sum_{j=1}^K h_{T_j,B}^{p(\tau-1/2-(n-1)(1/2-1/p)_+)} \|u_j\|_{W_2^{\tau-1/2}(B)}^p \\ &\le C h_{X_2,\partial\Omega}^{p(\tau-1/2-(n-1)(1/2-1/p)_+)} \|u-s\|_{W_2^{\tau-1/2}(\partial\Omega)}^p \\ &\le C h_{X_2,\partial\Omega}^{p(\tau-1/2-(n-1)(1/2-1/p)_+)} \|u-s\|_{W_2^{\tau}(\Omega)}^p\end{aligned}$$

for $1 \le p < \infty$, and the case $p = \infty$ is treated in the same fashion. Finally, since s is a norm-minimal interpolant, the norm in the last expression can again be bounded by the norm of u. ∎

The two most important estimates for the boundary part are hence

$$\|u-s\|_{L_\infty(\partial\Omega)} \le C h_{X_2,\partial\Omega}^{\tau-n/2} \|u\|_{W_2^\tau(\Omega)},$$

$$\|u-s\|_{L_2(\partial\Omega)} \le C h_{X_2,\partial\Omega}^{\tau-1/2} \|u\|_{W_2^\tau(\Omega)}.$$

The proof of Theorem 3.10 shows that the following alternative version of Theorem 3.10 is also true.

COROLLARY 3.11. *Suppose $\Gamma \subseteq \partial\Omega$ is a part of the boundary satisfying*

$$\Gamma = \bigcup_{j=1}^J (V_j \cap \partial\Omega). \tag{17}$$

This means, that the first J charts $\{V_j, \varphi_j\}_{j=1}^J$ are exclusive for Γ, or that, for $1 \le j \le J$, $V_j \cap (\partial\Omega \setminus \Gamma) = 0$. Suppose further that the boundary collocation points X_2 are chosen only on Γ, while the interior points are still chosen in Ω; then estimate (15) *remains valid and* (16) *becomes*

$$\|u-s\|_{L_p(\Gamma)} \le C h_{X_2,\Gamma}^{\tau-1/2-(n-1)(1/2-1/p)_+} \|u\|_{W_2^\tau(\Omega)}, \tag{18}$$

where $h_{X_2,\Gamma} = \max_{1\le j\le L} h_{T_j,B}$ with T_j defined as before.

As a matter of fact, neither condition (17) nor the fact that $X_2 \subseteq \Gamma$ are necessary to derive (18). But if (17) is not satisfied, the fill distance $h_{X_2,\Gamma}$ might be larger than necessary if X_2 is chosen only from Γ. On the other hand, if X_2 is dense on all of $\partial\Omega$, then, of course, (16) implies (18).

Considering again the compactly supported functions $\Phi = \psi_{\ell,k}(\|\cdot\|_2)$, i.e., choosing $\tau = k + (n+1)/2$, gives this time the following corollary.

COROLLARY 3.12. *Let $k > m - 1/2$ if n is odd or $k > m$ if n is even. Let $c_\alpha \in W_\infty^{k-m+1+\lfloor \frac{n+1}{2} \rfloor}$. Suppose $u \in W_2^{k+(n+1)/2}(\Omega)$. Then, employing Wendland's*

compactly supported basis functions yields

$$\|Lu - Ls\|_{L_\infty(\Omega)} \le C h_{X_1,\Omega}^{k-m+1/2} \|u\|_{W_2^{k+(n+1)/2}(\Omega)}, \tag{19}$$

$$\|u - s\|_{L_\infty(\partial\Omega)} \le C h_{X_2,\partial\Omega}^{k+1/2} \|u\|_{W_2^{k+(n+1)/2}(\Omega)}. \tag{20}$$

A similar statement holds also for $\Gamma \subset \partial\Omega$; cf. Corollary 3.11.

3.3. Elliptic PDEs. We now consider the following elliptic operator of second order in a bounded domain $\Omega \subset \mathbb{R}^n$ with a sufficiently smooth boundary

$$Lu(x) := \sum_{i,j=1}^{n} a_{ij}(x)\partial_{i,j}u(x) + \sum_{i=1}^{n} b_i(x)\partial_i u(x) + c(x)u(x), \tag{21}$$

where a, b, and c are bounded, $a_{ij}(x) = a_{ji}(x)$ (symmetry), and $c(x) \le 0$ holds for all $x \in \Omega$. Moreover, let L be strictly elliptic; i.e., there is a constant $\lambda > 0$ such that

$$\lambda \|\xi\|_2^2 \le \sum_{i,j=1}^{n} a_{ij}(x)\xi_i\xi_j$$

for all $x \in \Omega$ and $\xi \in \mathbb{R}^n$. Then, if $u \in C^0(\overline{\Omega}) \cap C^2(\Omega)$ is the solution of (1) and (3), it enjoys the following estimate (see [12, Theorem 3.7]):

$$\|u\|_{L_\infty(\Omega)} \le \|F\|_{L_\infty(\partial\Omega)} + \frac{C}{\lambda}\|f\|_{L_\infty(\Omega)}, \tag{22}$$

where the constant C depends on the diameter of Ω and on $\|b\|_{L_\infty(\Omega)}/\lambda$. This, together with Theorem 3.10, immediately yields the next result.

COROLLARY 3.13. *Assume that the solution u belongs to $W_2^\tau(\Omega)$ with $\lfloor\tau\rfloor > 2+n/2$. Then, the error between u and its collocation approximation s can be bounded by*

$$\begin{aligned}\|u - s\|_{L_\infty(\Omega)} &\le C\left(h_{X_1,\Omega}^{\tau-2-n/2} + h_{X_2,\partial\Omega}^{\tau-n/2}\right)\|u\|_{W_2^\tau(\Omega)} \\ &\le C h_X^{\tau-2-n/2}\|u\|_{W_2^\tau(\Omega)},\end{aligned}$$

where $h_X = \max\{h_{X_1,\Omega}, h_{X_2,\partial\Omega}\}$.

Note that this result unfortunately means that we have to choose a higher data density in the interior than on the boundary.

The result for the compactly supported functions is

$$\|u - s\|_{L_\infty(\Omega)} \le C\left(h_{X_1,\Omega}^{k-3/2} + h_{X_2,\partial\Omega}^{k+1/2}\right)\|u\|_{W_2^{k+(n+1)/2}(\Omega)}.$$

In the case of constant coefficients, i.e., $a_{ij}(x) = a_{ij}$, $b_i(x) = b_i$ and $c(x) = c$ for all $x \in \Omega$, this result was obtained in [6] using a transformation theorem. Our result, however, also holds for nonconstant coefficients and is mainly a simple application of Theorem 3.10.

4. Dynamical systems.

4.1. A short introduction. Consider the ordinary differential equation

$$\dot{x} = \frac{dx}{dt} = g(x), \tag{23}$$

where $g \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$, $\sigma \geq 1$, and $x(t) \in \mathbb{R}^n$. We search for solutions $x(t)$, $t \geq 0$, of the initial value problem (23), $x(0) = \xi$. We denote these solutions also by $S_t\xi := x(t)$. Since g is at least C^1, we have existence and uniqueness of solutions of this initial value problem locally in time.

Since one cannot calculate the solutions of (23) in general, from the viewpoint of dynamical systems theory we are interested in the qualitative long-time behavior of solutions. Therefore, one studies simple solutions such as equilibria, i.e., solutions which are constant in time.

DEFINITION 4.1. *$x_0 \in \mathbb{R}^n$ is called an equilibrium for* (23) *if $g(x_0) = 0$. Then $S_t x_0 = x_0$ for all $t \geq 0$; i.e., the constant function $x(t) = x_0$ is a solution of* (23).

The concept of stability describes the behavior of solutions near the equilibrium x_0. Stability can be analyzed using the linearization of g at x_0.

PROPOSITION 4.2. *Let $x_0 \in \mathbb{R}^n$ be an equilibrium for* (23). *If all eigenvalues of the Jacobian $Dg(x_0)$ have a negative real part, then x_0 is asymptotically stable.*

For the rest of this section we assume that x_0 is an equilibrium such that all eigenvalues of $Dg(x_0)$ have a negative real part. For such an asymptotically stable equilibrium x_0 we can define the basin of attraction $A(x_0)$. Note that $A(x_0) \neq \varnothing$ and $A(x_0)$ is open.

DEFINITION 4.3. *Let $x_0 \in \mathbb{R}^n$ be an asymptotically stable equilibrium for* (23). *Then we define the basin of attraction as $A(x_0) := \{\xi \in \mathbb{R}^n \mid \lim_{t\to\infty} S_t\xi = x_0\}$.*

A method to determine subsets of the basin of attraction is the method of Lyapunov functions; cf. [16]. The main characteristic of a Lyapunov function $V \in C^1(\mathbb{R}^n, \mathbb{R})$ is that its orbital derivative $V'(x)$ is negative.

DEFINITION 4.4. *Given a function $V \in C^1(\mathbb{R}^n, \mathbb{R})$ its orbital derivative with respect to* (23) *is defined as $V'(x) := \langle \nabla V(x), g(x)\rangle = \sum_{j=1}^n \partial_j V(x) g_j(x)$.*

The orbital derivative is the derivative along a solution of (23) due to the chain rule:

$$\frac{d}{dt} V(x(t)) = \langle \nabla V(x(t)), \dot{x}(t)\rangle = \sum_{j=1}^n (\partial_j V)(x(t)) g_j(x(t)) = V'(x(t)).$$

Note that the orbital derivative is a linear differential operator of first order of the form (2):

$$LV(x) = V'(x) = \sum_{i=1}^n g_i(x) \partial_i V(x).$$

Here, the singular points, i.e., those points where $(\delta_x \circ L) = 0$, are simply the equilibrium points, i.e., those points satisfying $g(x) = 0$.

The following theorem explains the use of Lyapunov functions for the determination of the basin of attraction.

THEOREM 4.5 (see [11, Theorem 2.24]). *Let $s \in C^1(\mathbb{R}^n, \mathbb{R})$ and $K \subset \mathbb{R}^n$ be a compact set with neighborhood B such that $x_0 \in \overset{\circ}{K}$. Furthermore, let*

1. *$K = \{x \in B \mid s(x) \leq R\}$ with an $R \in \mathbb{R}$; i.e., K is a sublevel set of s.*

2. *$s'(x) < 0$ for all $x \in K \setminus \{x_0\}$; i.e., s is decreasing along solutions in $K \setminus \{x_0\}$.*

Then $K \subset A(x_0)$.

Hence, a Lyapunov function provides information on the basin of attraction through its sublevel sets. However, it is not easy to find a Lyapunov function for a general system (23). Although existence of several types of Lyapunov functions is known, their construction is not easy.

However, for linear differential equations, i.e., $g(x)$ is linear, one can easily calculate a Lyapunov function. For a nonlinear system we consider the linearized system at the equilibrium point, namely, $\dot{x} = Dg(x_0)(x - x_0)$. This is a linear system and, thus, one can easily calculate a Lyapunov function of the form $v(x) = (x - x_0)^T C(x - x_0)$, where the positive definite matrix C is the unique solution of the matrix equation $Dg(x_0)^T C + CDg(x_0) = -I$; cf. [22]. The function v is a Lyapunov function not only for the linearized system but also for the nonlinear system in a neighborhood of x_0; for details, cf. [11].

LEMMA 4.6 (local Lyapunov function). *Let x_0 be an equilibrium of $\dot{x} = g(x)$ such that all eigenvalues of $Dg(x_0)$ have a negative real part. Denote by $C \in \mathbb{R}^{n \times n}$ the unique solution of the matrix equation $Dg(x_0)^T C + CDg(x_0) = -I$ and define the local Lyapunov function*

$$v(x) = (x - x_0)^T C(x - x_0).$$

Then there is a compact set K with a neighborhood B such that $x_0 \in \overset{\circ}{K}$. Moreover, $v'(x) < 0$ holds for all $x \in K \setminus \{x_0\}$ and $K = \{x \in B \mid v(x) \leq R\}$ with $R > 0$.

We return to Lyapunov functions which have a negative orbital derivative for all $x \in A(x_0) \setminus \{x_0\}$. We consider special Lyapunov functions satisfying certain equations for their orbital derivatives. In the first part of Theorem 4.8 below, a feasible candidate is given by $p(x) = \|x - x_0\|_2^2$. For the second part we need the following definition.

DEFINITION 4.7 (noncharacteristic hypersurface [11, Definition 2.36]). *Let $h \in C^\sigma(\mathbb{R}^n, \mathbb{R})$. The set $\Gamma \subset \mathbb{R}^n$ is called a* noncharacteristic hypersurface *if*

- *Γ is compact,*
- *$h(x) = 0$ holds if and only if $x \in \Gamma$,*
- *$h'(x) < 0$ holds for all $x \in \Gamma$, and*
- *for each $x \in A(x_0) \setminus \{x_0\}$ there is a time $\theta(x) \in \mathbb{R}$ such that $S_{\theta(x)} x \in \Gamma$.*

An example of a noncharacteristic hypersurface is a level set of the local Lyapunov function; cf. Lemma 4.6.

THEOREM 4.8 (see [11, Theorems 2.38 and 2.46]). *Consider (23) with $g \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$ and let x_0 be an equilibrium such that all eigenvalues of $Dg(x_0)$ have a negative real part.*

1. *Let $p(x) \in C^\sigma(\mathbb{R}^n, \mathbb{R})$ satisfy the following conditions:*
   (a) *$p(x) > 0$ for $x \neq x_0$.*
   (b) *$p(x) = O(\|x - x_0\|_2^\eta)$ with $\eta > 0$ for $x \to x_0$.*
   (c) *For all $\epsilon > 0$, p has a lower positive bound on $\mathbb{R}^n \setminus B(x_0, \epsilon)$.*

   Then there exists a Lyapunov function $V_1 \in C^\sigma(A(x_0), \mathbb{R})$ such that $V_1(x_0) = 0$ and

$$LV_1(x) = f_1(x) := -p(x) \text{ for all } x \in A(x_0).$$

2. *Let $c > 0$, let Γ be a noncharacteristic hypersurface (see Definition 4.7), and let $F \in C^\sigma(\Gamma, \mathbb{R})$. Then there is a Lyapunov function $V_2 \in C^\sigma(A(x_0) \setminus \{x_0\}, \mathbb{R})$ such that*

$$LV_2(x) = f_2(x) := -c \text{ for all } x \in A(x_0) \setminus \{x_0\},$$

$$V_2(x) = F(x) \text{ for all } x \in \Gamma.$$

4.2. Approximating Lyapunov functions. Theorem 4.8 shows two possibilities for approximating Lyapunov functions. We can use the first part to approximate V_1 by solving the problem

$$Ls_1(x) = LV_1(x) = -p(x), \qquad x \in A(x_0).$$

This is an example of an operator problem of type (1), and our theory from section 3.1 applies.

On the other hand, the second part of Theorem 4.8 implies to solve the boundary value problem

$$Ls_2(x) = f_2(x) = -c, \qquad x \in A(x_0) \setminus \{x_0\},$$

$$s_2(x) = F(x), \qquad x \in \Gamma,$$

such that we can use our theory from section 3.2.

However, in both cases the application of our error estimates now has a different character. An error bound of the form $|LV(x) - Ls(x)| = |V'(x) - s'(x)| < \epsilon$ leads to $s'(x) \leq V'(x) + \epsilon < 0$, provided that ϵ is sufficiently small. Remember that V, as a Lyapunov function, satisfies $V'(x) < 0$. Hence, in this case s is itself a Lyapunov function.

However, for the specific choices of Lyapunov functions from Theorem 4.8 we have a problem if x is close to x_0. In the first case, $V_1'(x) = f_1(x) = -p(x)$ and $p(x) \to 0$ as $x \to x_0$. Hence, this estimate will not hold near x_0 and thus s_1' may be positive near x_0. The same problem arises for the approximation s_2 of V_2, since V_2 is not defined in x_0. Fortunately, locally it is easy to determine the basin of attraction by linearization; cf. Lemma 4.6.

Before we can apply the results of this paper to the calculation of Lyapunov functions, we need some information about the level sets of Lyapunov functions. We assume that g is bounded in $A(x_0)$. This can easily be achieved by considering the system $\dot{x} = h(x) := \frac{g(x)}{1+\|g(x)\|^2}$. Note that $\|h(x)\| \leq \frac{1}{2}$. This system has the same equilibria and basins of attraction as system (23) since $h(x)$ is obtained by multiplication of $g(x)$ by a positive, scalar factor; i.e., the orbits of both systems are the same, but the velocities are different.

THEOREM 4.9 (see [11, Corollary 2.43, Proposition 2.44, and Theorem 2.46]). *Let x_0 be an equilibrium of $\dot{x} = g(x)$, $g \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$, $\sigma \geq 1$, and let the maximal real part of all eigenvalues of $Dg(x_0)$ be negative. Let g be bounded in $A(x_0)$ and let $V = V_i$, $i = 1, 2$, be one of the functions of Theorem 4.8.*

Then for all $r > 0$ the set $\{x \in A(x_0) \setminus \{x_0\} \mid V(x) \leq r\} \cup \{x_0\}$ is compact. Moreover, there is a C^σ-diffeomorphism

$$\phi \in C^\sigma(S^{n-1}, \{x \in A(x_0) \mid V(x) = r\}),$$

where $S^{n-1} = \{x \in \mathbb{R}^n \mid \|x\|_2 = 1\}$. For V_2 we have $\lim_{x \to x_0} V_2(x) = -\infty$.

In the second case V_2, one first has to link the function V_2 to a local Lyapunov function to obtain the above theorem. For details, see [11].

In order to apply the results of section 3 to approximate the functions V_1, V_2 of Theorem 4.8, we have to choose a set Ω in an appropriate way such that Ω has a smooth boundary.

THEOREM 4.10. *Let $k := \lfloor \tau \rfloor > 1 + n/2$ and $\sigma := \lceil \tau \rceil$. Consider the dynamical system defined by the ordinary differential equation $\dot{x} = g(x)$, where $g \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$. Let $x_0 \in \mathbb{R}^n$ be an equilibrium such that the real parts of all eigenvalues of $Dg(x_0)$ are negative. Let g be bounded in $A(x_0)$ and denote by $V_1 \in W_2^\tau(A(x_0), \mathbb{R})$, $V_2 \in W_2^\tau(A(x_0) \setminus \{x_0\}, \mathbb{R})$ the Lyapunov functions of Theorem* 4.8.

1. *The reconstruction s_1 of the Lyapunov function V_1 with respect to the operator $Lu(x) = \langle \nabla u(x), g(x) \rangle$ and a set $X \subseteq \Omega := \{x \in A(x_0) \mid V_1(x) \le r\} \setminus \{x_0\}$, $r > 0$, satisfies*

$$\|s_1' - V_1'\|_{L_\infty(\Omega)} = \|s_1' + p\|_{L_\infty(\Omega)} \le C h_X^{\tau - 1 - n/2} \|V_1\|_{W_2^\tau(\Omega)}.$$

2. *Let $\Gamma = \{x \in A(x_0) \setminus \{x_0\} \mid h(x) = 0\}$ be a noncharacteristic hypersurface and set $\Omega = \{x \in A(x_0) \setminus \{x_0\} \mid V_2(x) \le r$ and $h(x) \ge 0\}$, where $r > 0$ is large enough such that $\{x \in A(x_0) \setminus \{x_0\} \mid V_2(x) = r\} \cap \Gamma = \varnothing$. The reconstruction s_2 of V_2 with respect to the boundary value problem $Lu(x) = \langle \nabla u(x), g(x) \rangle$, $u(x) = 0 = F(x)$ for Γ and the data sites $X_1 \subset \Omega$ and $X_2 \subset \Gamma$ satisfies*

$$\|s_2' - V_2'\|_{L_\infty(\Omega)} = \|s_2' + c\|_{L_\infty(\Omega)} \le C h_{X_1,\Omega}^{\tau - 1 - n/2} \|V_2\|_{W_2^\tau(\Omega)},$$

$$\|s_2 - V_2\|_{L_\infty(\Gamma)} = \|s_2\|_{L_\infty(\Gamma)} \le C h_{X_2,\Gamma}^{\tau - n/2} \|V_2\|_{W_2^\tau(\Omega)}.$$

Proof. Note that the data sites x_j, $1 \le j \le N$, are no singular points, i.e., $g(x_j) \ne 0$ or equilibria in this case, since there are no equilibria in $A(x_0) \setminus \{x_0\}$.

1. We apply Theorem 3.5 with $m = 1$. The set Ω is bounded and has a smooth boundary by Theorem 4.9 and thus satisfies the conditions of Theorem 3.5; cf. [27]. The functions c_α are $g_j \in C^\sigma(\mathbb{R}^n, \mathbb{R})$ and thus are in $W_\infty^k(\Omega)$.
2. We apply Corollary 3.11 with $m = 1$. The sets Ω and $\Gamma \subset \partial\Omega$ are bounded and Ω has a smooth boundary by Theorem 4.9 (see also [27]). Thus the conditions of Corollary 3.11 are satisfied. The functions c_α are $g_j \in C^\sigma(\mathbb{R}^n, \mathbb{R})$ and thus are in $W_\infty^k(\Omega)$. ∎

The calculation of the interpolation matrix A in Definition 3.1 can easily be achieved for radial basis functions, in particular for Wendland's compactly supported ones; cf. [11, Proposition 3.5 and Table 3.1].

COROLLARY 4.11. *Denote by k the smoothness index of the compactly supported Wendland function. Let $k > \frac{1}{2}$ if n is odd or $k > 1$ if n is even. Set $\tau = k + (n+1)/2$ and $\sigma = \lceil \tau \rceil$. Consider the dynamical system defined by the ordinary differential equation $\dot{x} = g(x)$, where $g \in C^\sigma(\mathbb{R}^n, \mathbb{R}^n)$. Let $x_0 \in \mathbb{R}^n$ be an equilibrium such that all eigenvalues of $Dg(x_0)$ have a negative real part. Let g be bounded in $A(x_0)$ and denote by $V_1 \in W_2^\tau(A(x_0), \mathbb{R})$ and $V_2 \in W_2^\tau(A(x_0) \setminus \{x_0\}, \mathbb{R})$ the Lyapunov functions of Theorem* 4.8.

1. *The reconstruction s_1 of the Lyapunov function V_1 with respect to the operator $Lu(x) = \langle \nabla u(x), g(x) \rangle$ and a set $X \subseteq \Omega := \{x \in A(x_0) \mid V_1(x) \le r\} \setminus \{x_0\}$, $r > 0$, satisfies*

$$\|s_1' - V_1'\|_{L_\infty(\Omega)} = \|s_1' + p\|_{L_\infty(\Omega)} \le C h_X^{k - \frac{1}{2}} \|V_1\|_{W_2^{k+(n+1)/2}(\Omega)}. \tag{24}$$

2. *Let* $\Gamma = \{x \in A(x_0)\setminus\{x_0\} \mid h(x) = 0\}$ *be a noncharacteristic hypersurface and set* $\Omega = \{x \in A(x_0) \setminus \{x_0\} \mid V_2(x) \le r$ *and* $h(x) \ge 0\}$, *where* $r > 0$ *is large enough such that* $\{x \in A(x_0) \setminus \{x_0\} \mid V_2(x) = r\} \cap \Gamma = \varnothing$. *The reconstruction* s_2 *of* V_2 *with respect to the boundary value problem* $Lu(x) = \langle \nabla u(x), g(x)\rangle$, $u(x) = 0 = F(x)$, *for* Γ *and the sets of data sites* $X_1 \subset \Omega$ *and* $X_2 \subset \Gamma$ *satisfies*

$$\|s_2' - V_2'\|_{L_\infty(\Omega)} \le C h_{X_1,\Omega}^{k-\frac{1}{2}} \|V_2\|_{W_2^{k+(n+1)/2}(\Omega)}, \tag{25}$$

$$\|s_2 - V_2\|_{L_\infty(\Gamma)} \le C h_{X_2,\Gamma}^{k+\frac{1}{2}} \|V_2\|_{W_2^{k+(n+1)/2}(\Omega)}. \tag{26}$$

Proof. Apply Corollaries 3.6, 3.11, and 3.12, respectively, with $m = 1$. ∎

The method described in this paper has already been used in [8, 9, 10, 11]. However, the approximation orders derived in those papers were based on Taylor approximation of first order, and hence the results in those papers were significantly worse than the results of Corollary 4.11.

The theorems and corollaries of this section, in particular (24) and (25), ensure that the approximation of the Lyapunov functions V_1 and V_2 produces functions s_1, s_2, respectively, with negative orbital derivatives in Ω if the data sites are dense enough. For the remaining neighborhood of the equilibrium x_0 we use a local Lyapunov function; cf. Lemma 4.6. We can combine the approximated function s and the local Lyapunov function v to a new Lyapunov function $\tilde{s}$ such that $\tilde{s}'(x) < 0$ holds for all $x \in \Omega \setminus \{x_0\}$ and such that level sets of s are level sets of $\tilde{s}$.

However, since Theorem 4.5 requires a sublevel set of s within the region where $s'(x) < 0$, we need information about the level sets of the approximants s. Here we make use of the estimate for s_2 on Γ; cf. (26). The following theorem shows that we can cover each compact subset $\tilde{K}$ of the basin of attraction with a sublevel set of s, and thus the approximation method finds every compact subset of the basin of attraction, provided that the sets Ω and Γ are chosen appropriately and the data sites are dense enough.

THEOREM 4.12 (see [11, Theorems 5.1 and 5.3]).

1. *Let* $\tilde{K}$ *be a compact set with* $x_0 \in \overset{\circ}{\tilde{K}} \subset \tilde{K} \subset A(x_0)$. *Let* s_1 *be an approximation of* V_1 *as in Corollary* 4.11 *with* $\Omega := \{x \in A(x_0) \mid V_1(x) \le r\} \setminus \{x_0\}$, *where* $r > 0$ *is large enough and* h_X *is small enough.*
   Then there is a $\rho \in \mathbb{R}$ *with* $\tilde{K} \subset \{x \in \Omega \mid s_1(x) \le \rho\}$.
2. *Let* $\tilde{K}$ *be a compact set with* $x_0 \in \overset{\circ}{\tilde{K}} \subset \tilde{K} \subset A(x_0)$. *Let* s_2 *be an approximation of* V_2 *as in Corollary* 4.11 *with* $\Omega = \{x \in A(x_0) \setminus \{x_0\} \mid V_2(x) \le r$ *and* $h(x) \ge 0\}$, *where* $r > 0$ *is large enough and* h_{X_1} *and* h_{X_2} *are small enough. Set* $U = \{x \in A(x_0) \mid h(x) \le 0\}$.
   Then there is a $\rho \in \mathbb{R}$ *with* $\tilde{K} \subset U \cup \{x \in \Omega \mid s_2(x) \le \rho\}$.

The proof of 2 compares level sets of s_2 with level sets of V_2 using estimate (26) on Γ and (25) along solutions. For 1 we can derive an estimate near x_0 since V_1 is defined and smooth at x_0; then we use the estimate (24) along solutions.

4.3. Example. As an example we consider the dynamical system given by

$$\begin{cases} \dot{x} &= -x - 2y + x^3, \\ \dot{y} &= -y + \frac{1}{2}x^2 y + x^3 \end{cases}$$

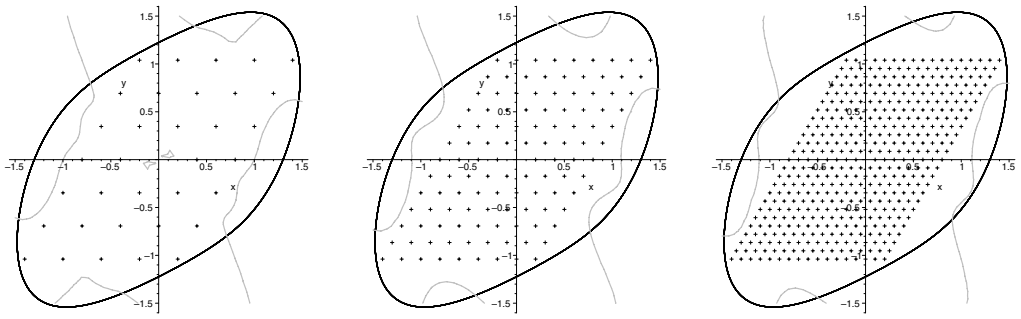


FIG. 1. *The grid X_N (plus signs), the basin of attraction bounded by the black periodic orbit, and the set $\{(x,y) \in \mathbb{R}^2 \mid s'(x,y) = 0\}$ (grey dotted lines) with the approximation s of the function V, where $V'(x,y) = -x^2 - y^2$ with the Wendland function $\psi_{4,2}(2/3\|x\|_2)$ and the grid distance α. Left: $\alpha = 0.4$; middle: $\alpha = 0.2$; right: $\alpha = 0.1$.*

and denote the right-hand side by $g(x,y)$. The system has an asymptotically stable equilibrium at $(0,0)$ with Jacobian

$$Dg(0,0) = \begin{pmatrix} -1 & -2 \\ 0 & -1 \end{pmatrix}.$$

For a local Lyapunov function (cf. Lemma 4.6), we calculate the unique solution C of the matrix equation $Dg(0,0)^T C + CDg(0,0) = -I$, which is given by

$$C = \begin{pmatrix} \frac{1}{2} & -\frac{1}{2} \\ -\frac{1}{2} & \frac{3}{2} \end{pmatrix}.$$

The basin of attraction $A(0,0)$ is bounded by an unstable periodic orbit which we have calculated numerically. We approximate the function V_1 satisfying $V_1'(x,y) = -x^2 - y^2$. We use a hexagonal grid of the form $\alpha[j\,(1,0)^T + k(\frac{1}{2}, \frac{\sqrt{3}}{2})^T]$ for the data sites. Then the mesh norm is $h = \alpha/2$. Since we have to avoid singular points we must exclude the origin. We use three different grids with parameters $\alpha_1 = 0.1$, $\alpha_2 = 0.2$, and $\alpha_3 = 0.4$ and two different Wendland functions as radial basis functions $\Phi(x) = \psi_{k,l}(c\|x\|_2)$ with $c = 2/3$ and $k = 2, 3$; cf. Figures 1 and 2.

We consider the grid $0.1[j\,(1,0)^T + k(\frac{1}{2}, \frac{\sqrt{3}}{2})^T + (\frac{3}{4}, \frac{\sqrt{3}}{4})^T]$. These grid points are in between the grid points of the smallest grid above. We calculate the maximal error on this grid. By our error analysis the errors $e_{k,\alpha}$ and $e_{k,2\alpha}$ should behave as

$$\frac{e_{k,2\alpha}}{e_{k,\alpha}} \approx \frac{(2\alpha)^{k-1/2}}{(\alpha)^{k-1/2}} = 2^{k-1/2}$$

(cf. (24)), which is approximately reflected in our numerical results; see Table 1.

For the basin of attraction, however, the level sets of s are also important. Even if the set, where s' is negative, is large, a subset of the basin of attraction is given only by a sublevel set of s within this region. For one example we have calculated such a sublevel set and have compared it to the sublevel set of the local Lyapunov function; see Figure 3. If the function g is bounded in the basin of attraction, then one can cover each given compact set in $A(x_0)$ with a sublevel set of s, where the data sites are dense enough; see Theorem 4.12.

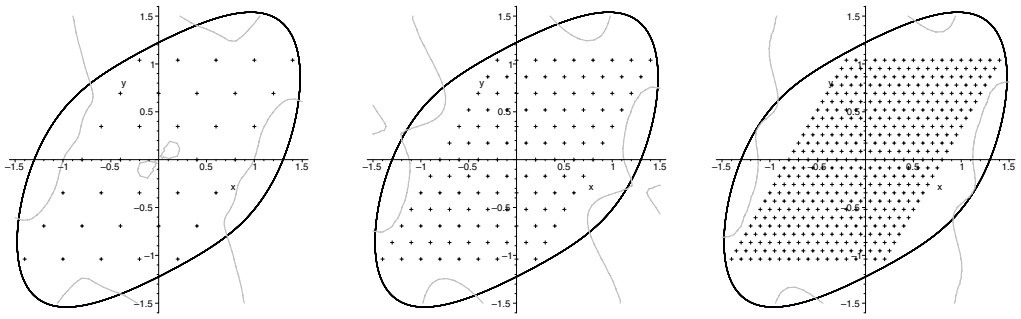


FIG. 2. *The grid X_N (plus signs), the basin of attraction bounded by the black periodic orbit, and the set $\{(x,y) \in \mathbb{R}^2 \mid s'(x,y) = 0\}$ (grey dotted lines) with the approximation s of the function V, where $V'(x,y) = -x^2 - y^2$ with the Wendland function $\psi_{5,3}(2/3\|x\|_2)$ and the grid distance α. Left: $\alpha = 0.4$; middle: $\alpha = 0.2$; right: $\alpha = 0.1$.*

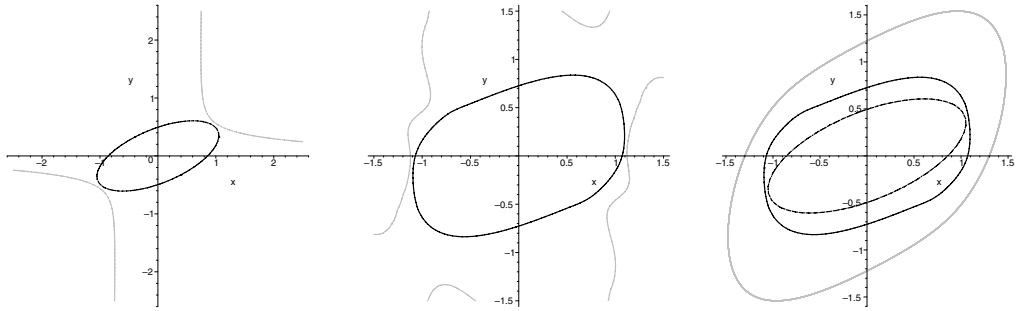


FIG. 3. *Left: The local Lyapunov function $v(x) = x^T Cx$: level set $v'(x) = 0$ (grey dotted lines) and a sublevel set $\{x \in \mathbb{R}^2 \mid v(x) \leq 0.37\}$ which is a subset of the basin of attraction. Middle: The calculated Lyapunov function s ($k = 3$, $\alpha = 0.1$): level set $s'(x) = 0$ (grey dotted lines) and a sublevel set $\{x \in \mathbb{R}^2 \mid s(x) \leq -0.5\}$ which is a subset of the basin of attraction. Right: Comparison of the subsets obtained by the local Lyapunov function v (small black ellipse), the calculated Lyapunov function s (large black set), and the whole basin of attraction (grey dotted lines).*

TABLE 1

The approximation error $e_\alpha = \max_{x \in X_3} \|s_1'(x) - V_1'(x)\|_2$, where X_3 is a dense grid for different Wendland functions $\psi_{k+2,k}$ and different grids with mesh norm α for the example discussed in this section. The ratio of the errors e_α is compared to the theoretical bound $2^{k-1/2}$ of Corollary 4.11, (24).

k / α	0.4	0.2	0.1	$e_{0.4}/e_{0.2}$	$e_{0.2}/e_{0.1}$	$2^{k-1/2}$
2	0.8862	0.4641	0.1814	1.9094	2.5592	2.8284
3	1.1308	0.4265	0.1041	2.6516	4.0960	5.6569

REFERENCES

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

[2] S. BRENNER AND L. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[3] M. D. BUHMANN, *Radial Basis Functions*, Cambridge Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 2003.

[4] G. E. FASSHAUER, *Solving partial differential equations by collocation with radial basis functions*, in Surface Fitting and Multiresolution Methods, A. L. Méhauté, C. Rabut, and L. L. Schumaker, eds., Vanderbilt University Press, Nashville, 1997, pp. 131–138.

[5] G. E. Fasshauer, *On the numerical solution of differential equations with radial basis functions*, in Boundary Element Technology XIII, C. S. Chen, C. A. Brebbia, and D. W. Pepper, eds., WIT Press, Southampton, 1999, pp. 291–300.
[6] C. Franke and R. Schaback, *Convergence order estimates of meshless collocation methods using radial basis functions*, Adv. Comput. Math., 8 (1998), pp. 381–399.
[7] C. Franke and R. Schaback, *Solving partial differential equations by collocation using radial basis functions*, Appl. Math. Comput., 93 (1998), pp. 73–82.
[8] P. Giesl, *Approximation of domains of attraction and Lyapunov functions using radial basis functions*, in Proceedings of the 6th IFAC Symposium on Nonlinear Control Systems (NOLCOS) vol. II, Stuttgart, Germany, 2004, pp. 865–870.
[9] P. Giesl, *Stepwise calculation of the basin of attraction in dynamical systems using radial basis functions*, in Algorithms for Approximation, A. Iske and J. Levesly, eds., Springer-Verlag, Heidelberg, 2007, pp. 113–122.
[10] P. Giesl, *Construction of a global Lyapunov function using radial basis functions with a single operator*, Discrete Contin. Dyn. Syst. Ser. B, 7 (2007), pp. 101–124.
[11] P. Giesl, *Construction of Global Lyapunov Functions Using Radial Basis Functions*, Lecture Notes in Math. 1904, Springer, Heidelberg, 2007.
[12] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Grundlehren Math. Wiss. 224, Springer, Berlin, 1977.
[13] Y. C. Hon and R. Schaback, *On unsymmetric collocation by radial basis functions*, J. Appl. Math. Comput., 119 (2001), pp. 177–186.
[14] E. J. Kansa, *Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics*—I: *Surface approximations and partial derivative estimates*, Comput. Math. Appl., 19 (1990), pp. 127–145.
[15] E. J. Kansa, *Multiquadrics—A scattered data approximation scheme with applications to computational fluid-dynamics*—II: *Solutions to parabolic, hyperbolic and elliptic partial differential equations*, Comput. Math. Appl., 19 (1990), pp. 147–161.
[16] A. M. Lyapunov, *Problème général de la stabilité du mouvement*, Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys. (2), 9 (1907), pp. 203–474. Translation of the original 1892 Russian version published in Comm. Soc. Math. Kharkow. Also available in book form as Ann. of Math. Stud. 17, Princeton University Press, Princeton, NJ, 1949 (in French).
[17] F. J. Narcowich and J. D. Ward, *Generalized Hermite interpolation via matrix-valued conditionally positive definite functions*, Math. Comput., 63 (1994), pp. 661–687.
[18] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev bounds on functions with scattered zeros, with applications to radial basis function surface fitting*, Math. Comput., 74 (2005), pp. 643–763.
[19] F. J. Narcowich, J. D. Ward, and H. Wendland, *Sobolev error estimates and a Bernstein inequality for scattered data interpolation via radial basis functions*, Constr. Approx., 24 (2006), pp. 175–186.
[20] R. Opfer, *Multiscale kernels*, Adv. Comput. Math., 25 (2006), pp. 357–380.
[21] R. Opfer, *Tight frame expansions of multiscale reproducing kernels in Sobolev spaces*, Appl. Comput. Harmon. Anal., 20 (2006), pp. 357–274.
[22] E. Sontag, *Mathematical Control Theory*, 2nd ed., Springer, New York, 1998.
[23] H. Triebel, *Interpolation Theory, Function Spaces, Differential Operators*, North–Holland Publishing Company, Amsterdam, 1978.
[24] H. Wendland, *Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree*, Adv. Comput. Math., 4 (1995), pp. 389–396.
[25] H. Wendland, *Error estimates for interpolation by compactly supported radial basis functions of minimal degree*, J. Approx. Theory, 93 (1998), pp. 258–272.
[26] H. Wendland, *Scattered Data Approximation*, Cambridge Monogr. Appl. Comput. Math., Cambridge University Press, Cambridge, UK, 2005.
[27] J. Wloka, *Partial Differential Equations*, Cambridge University Press, Cambridge, UK, 1987.
[28] Z. Wu, *Hermite-Birkhoff interpolation of scattered data by radial basis functions*, Approx. Theory Appl., 8 (1992), pp. 1–10.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1742–1776

LOCALLY CONSERVATIVE FLUXES FOR THE CONTINUOUS GALERKIN METHOD*

BERNARDO COCKBURN†, JAYADEEP GOPALAKRISHNAN‡, AND HAIYING WANG†

Abstract. The standard continuous Galerkin (CG) finite element method for second order elliptic problems suffers from its inability to provide conservative flux approximations, a much needed quantity in many applications. We show how to overcome this shortcoming by using a two-step postprocessing. The first step is the computation of a numerical flux trace defined on element interfaces and is motivated by the structure of the numerical traces of discontinuous Galerkin methods. This computation is nonlocal in that it requires the solution of a symmetric positive definite system, but the system is well conditioned independently of mesh size, so it can be solved at asymptotically optimal cost. The second step is a local element-by-element postprocessing of the CG solution incorporating the result of the first step. This leads to a conservative flux approximation with continuous normal components. This postprocessing applies for the CG method in its standard form or for a hybridized version of it. We present the hybridized version since it allows easy handling of variable-degree polynomials and hanging nodes. Furthermore, we provide an a priori analysis of the error in the postprocessed flux approximation and display numerical evidence suggesting that the approximation is competitive with the approximation provided by the Raviart–Thomas mixed method of corresponding degree.

Key words. continuous Galerkin methods, conforming finite element method, hybridization, elliptic problems, conservation

AMS subject classification. 65M60, 65N30, 35L65

DOI. 10.1137/060666305

1. Introduction.

In this paper, we revisit the classical finite element method [13, 20], otherwise known as the continuous Galerkin (CG) method, for second order elliptic problems, with the intention of showing how to overcome what is perhaps its main disadvantage, namely, the discontinuity of the normal component of the approximate flux across element interfaces. We show how to achieve this by means of an efficient postprocessing of the approximate solution provided by the CG method. We also show that the postprocessed flux is competitive with the flux provided by the Raviart–Thomas mixed method of corresponding degree.

We illustrate our technique in the framework of the model second order elliptic boundary value problem

$$-\nabla \cdot (a\nabla u) = f \qquad \text{on } \Omega, \tag{1.1a}$$

$$u = g \qquad \text{on } \Gamma_D, \tag{1.1b}$$

$$-a\nabla u \cdot \boldsymbol{n} = \mathsf{q}_{\mathsf{N}} \qquad \text{on } \Gamma_N. \tag{1.1c}$$

Here $\Omega \subset \mathbb{R}^N$ is a polyhedral domain ($N \geq 2$) with boundary $\partial\Omega$, $f \in L^2(\Omega)$, and $a = a(\boldsymbol{x})$ is a symmetric $N \times N$ matrix function that is uniformly positive definite

*Received by the editors July 28, 2006; accepted for publication (in revised form) February 16, 2007; published electronically August 24, 2007.
http://www.siam.org/journals/sinum/45-4/66630.html

†School of Mathematics, University of Minnesota, Vincent Hall, Minneapolis, MN 55455 (cockburn@math.umn.edu, hywang@math.umn.edu). The research of the first author was supported in part by the National Science Foundation under grant DMS-0411254 and by the University of Minnesota Supercomputing Institute.

‡Department of Mathematics, University of Florida, Gainesville, FL 32611 (jayg@math.ufl.edu). This author's research was supported in part by the National Science Foundation under grants DMS-0410030 and SCREMS-0619080.

on Ω with components in $L^\infty(\Omega)$. The boundary conditions are given by functions g and $\mathsf{q_N}$ on disjoint subsets Γ_D and Γ_N of $\partial\Omega$, upon which further assumptions will be placed shortly. Here and elsewhere we use $\boldsymbol{n}$ to denote the unit outward normal on the boundary of some domain—the domain will be clear from the context, e.g., in (1.1c) it is Ω. As is well known, this boundary value problem models a wide range of problems of practical interest from electromagnetics to heat dissipation and flow in porous media.

To facilitate the discussion of the results, let us introduce our notation for the CG method right away. Let $\mathcal{T}_h$ denote a triangulation of the domain Ω, which for simplicity we assume consists of simplices. Define the space

$$\mathsf{V}_h = \left\{ v \in \mathcal{C}^0(\Omega) : v|_K \in \mathcal{P}_k(K) \text{ for } K \in \mathcal{T}_h \right\}, \tag{1.2}$$

where $\mathcal{C}^0(D)$ denotes the space of continuous functions on a domain D. We assume that Γ_D is the union of some mesh faces (edges if $N = 2$) lying on $\partial\Omega$ and that $\Gamma_N = \partial\Omega \setminus \Gamma_D$. We assume that g is in the space of traces on Γ_D of functions in V_h and set $\mathsf{V}_h(g) = \{v \in \mathsf{V}_h : v = g \text{ on } \Gamma_D\}$. If a Dirichlet data that is not polynomial is given, one can proceed by approximating it as usual, but we shall not consider this case. As is well known, the approximate solution u_h of the CG method is the function in $\mathsf{V}_h(g)$ determined by

$$(a\nabla u_h, \nabla v)_\Omega = (f,\ v)_\Omega - \langle \mathsf{q_N}\,,\, v\rangle_{\Gamma_N} \quad \text{for all } v \in \mathsf{V}_h(0). \tag{1.3}$$

Here we have used common notation for innerproducts: For scalar functions w and v on some domain $\mathcal{D} \subset \mathbb{R}^N$, $(w, v)_\mathcal{D} = \int_\mathcal{D} w\, v\, dx$; for vector functions $(\boldsymbol{p}, \boldsymbol{q})_\mathcal{D} = \int_\mathcal{D} \boldsymbol{p} \cdot \boldsymbol{q}\, dx$; and for functions on domains B formed by lower-dimensional objects like union of a few mesh faces, $\langle \eta, \zeta\rangle_B = \int_B \eta\, \zeta\, d\gamma$.

It is well known that the CG approximation given by $-a\nabla u_h$ to the flux $\boldsymbol{q} = -a\nabla u$ is not conservative. The root of the problem is evident once we write (1.1a) in conservation form as $\operatorname{div} \boldsymbol{q} = f$. While the flux approximations from mixed and discontinuous Galerkin (DG) methods satisfy a discrete analogue of this equation, the CG flux $-a\nabla u_h$ does not. We say that a discrete flux $\boldsymbol{q}_h$ approximating the exact flux $\boldsymbol{q}$ is *conservative* if the total outward flux across any "discrete subdomain" as measured by $\boldsymbol{q}$ and $\boldsymbol{q}_h$ coincides, or more precisely,

$$\int_{\partial D_h} \boldsymbol{q} \cdot \boldsymbol{n}\, ds = \int_{\partial D_h} \boldsymbol{q}_h \cdot \boldsymbol{n}\, ds \tag{1.4}$$

for any domain D_h formed by the union of some mesh elements in $\mathcal{T}_h$ (where $\boldsymbol{n}$ is unit outward normal on the boundary of D_h). Conservative flux approximations are very important in many applications, e.g., in oil recovery simulations, more generally in flows through porous media, and indeed in computational fluid dynamics in general. The same is true in computational structural mechanics, where mixed and hybrid methods were devised to cope with its absence in the so-called one-field displacement method for linear elasticity (which is the CG method for elasticity); see, e.g., the first paragraph of section 3.3 in [31].

Many researchers have attempted to overcome the lack of conservativity of the CG flux by generating a better flux through postprocessing. However, a conservative $\boldsymbol{H}(\operatorname{div}, \Omega)$-conforming flux approximation has eluded their efforts for more than three decades. Let us briefly review what has been achieved to date. In [33], J. Wheeler showed how to postprocess the CG solution to obtain approximations to the normal component of $\boldsymbol{q}$ at the boundary of the computational domain. In one space

dimension, this procedure can be extended to compute approximations to $\boldsymbol{q}$ at all the nodes. In fact, such approximations were proven by M. Wheeler in [34] to superconverge with order $2\,k$ when using polynomial approximations of degree k. This solves the problem in the one-dimensional case. In the multidimensional case, however, the situation is rather different and no $\boldsymbol{H}(\mathrm{div},\Omega)$-conforming approximation of $\boldsymbol{q}$ has been constructed so far. Moreover, there are only a few theoretical and numerical studies of the approximation given by J. Wheeler's procedure. In [23], it was shown that such a procedure provides an approximation that superconverges in the $L^2(\partial\Omega)$-norm with order $k+1$ for $a\equiv 1$, and with order $k+1/2$ when a is smooth (under the assumption that Ω is a square endowed with a Cartesian mesh). In [3], the integral of the normal component of the flux on the whole boundary was proven to superconverge with order $k+1$ when Ω is a curved domain and isoparametric elements are used, and with order $2k$ when it is a polyhedron. For numerical studies, see the references cited in [25]. More importantly, in [25] the CG method was argued to have the property of local conservativity; see also [26] for an extension of this approach to the advection-diffusion and incompressible Navier–Stokes equations. In [9], the so-called *superconvergent integral flux postprocessing formula* was revisited. The conservation property was proven and a relation to a Lagrange multiplier mixed formulation and the associated consistency implications were established. See also [10] for further work on conservative projections involving multipliers in a different context. However, none of the approaches used in [25, 9] can be employed to construct an $\boldsymbol{H}(\mathrm{div},\Omega)$-conforming approximation of the flux, rendering the CG method locally conservative. (The precise relation between this approach and ours is displayed right before section 3.2.) In [7], this approach was used (for $a=1$ and $N=2$) to obtain an approximation of the integral of the normal component of the flux on an internal boundary which splits the domain in two; an order of convergence of $2\,k$ was proven for such an approximation.

In this paper, we show how to obtain a conservative flux approximation $\boldsymbol{q}_h$ in $\boldsymbol{H}(\mathrm{div},\Omega)$ that renders the CG method locally conservative. This is done by postprocessing the CG solution u_h in two steps. The objective of the first is to compute a numerical trace $\widehat{\boldsymbol{q}}_h$ of the flux whose normal component is single-valued on the interelement boundaries and *renders locally conservative the CG method*, that is, it satisfies

$$-\sum_{K\in\mathcal{T}_h}(a\,\nabla u_h,\nabla v)_K+\sum_{K\in\mathcal{T}_h}\langle\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n},v\rangle_{\partial K}=(f,v)_\Omega$$

for all v such that $v|_K\in\mathcal{P}_\ell(K)$ for all $K\in\mathcal{T}_h$ for some $\ell\le k$. The form of this numerical trace is similar to that of the corresponding numerical traces of the DG methods. However, unlike the DG numerical traces, the crucial *stabilization* term cannot have the form of a parameter times the jump of the u_h, since in our case such a jump is identically equal to zero. Instead, it is a quantity that belongs to a certain nonstandard space of *jumps* and that depends globally on the CG approximation u_h. While the need for this term is far from obvious when approaching from the standard CG formulation, it becomes clearer from the hybridized form of the CG method, which uses a space of discontinuous functions that generate the above-mentioned space of jumps on mesh faces. Because of this, we now face difficulties not encountered in DG methods: the computation of $\widehat{\boldsymbol{q}}_h$ requires (i) a local basis representation of the space of jumps, and (ii) the solution of a global system in that space. We are able to overcome the former difficulty by extending some techniques developed in [17, 18].

Although the latter difficulty persists, it turns out that the stiffness matrix of the global system is symmetric, positive definite, and well conditioned. In particular, we prove that its condition number is bounded independently of mesh size, so it can be solved iteratively at asymptotically optimal cost. In [27] a similar but different way of computing a numerical trace has been proposed; see the discussion before section 3.3.

The second step in the postprocessing is the local element-by-element recovery of a conservative flux approximation $\boldsymbol{q}_h$ throughout the computational domain by a variation of the so-called Raviart–Thomas (RT) projection [29]. Similar techniques have been used by [4] in the framework of DG methods for Darcy's law and by [19] in the context of DG methods for the Navier–Stokes equations. The flux approximation $\boldsymbol{q}_h$ *coincides* with the numerical trace $\widehat{\boldsymbol{q}}_h$ on element boundaries supplied by the previous step and is lifted to the interior of each element by using the $a\,\nabla u_h$ in such a way that

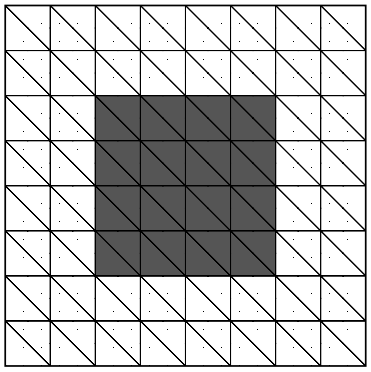
$$-\sum_{K\in\mathcal{T}_h}(a\,\nabla u_h,\nabla v)_K+\sum_{K\in\mathcal{T}_h}\langle\widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n},v\rangle_{\partial K}=\sum_{K\in\mathcal{T}_h}(\nabla\cdot\boldsymbol{q}_h,v)_K=(f,v)_\Omega$$

for all v such that $v|_K\in\mathcal{P}_\ell(K)$ for all $K\in\mathcal{T}_h$ for some $\ell\le k$. We prove that the resulting approximation $\boldsymbol{q}_h$ converges to the exact flux $\boldsymbol{q}$ at the *same* order of convergence as the approximation provided by the RT mixed method of corresponding order. Moreover, since the computation of the CG solution requires solving a system that is smaller in size than the corresponding RT system, our flux computation becomes a competitive alternative.

In [32], a technique is proposed for computing a locally conservative flux approximation in the domain Ω from its exact divergence in Ω and an approximation of its normal component on the interelement boundaries. It also proceeds in two steps. In the first, a locally conservative approximation to the normal component is obtained by solving a global constrained minimization problem. Then, on each element, the data on the border is lifted to the interior to obtain the desired flux; a local mixed element method is used to achieve this. The application of this technique to the CG method differs from ours in several respects. First of all, the resulting numerical trace does not render locally conservative the CG method, in the sense defined above. Moreover, to obtain it, a global constrained minimization problem is to be solved; this has to be contrasted with our unconstrained minimization problem whose stiffness matrix has a condition number bounded independently of the mesh size. Finally, to obtain what we call $\boldsymbol{q}_h$, the approximation u_h given by the CG method is not used.

Let us compare our flux $\boldsymbol{q}_h$ with the RT flux obtained for the model problem (1.1) with $f=0$, $\Omega=(0,1)\times(0,1)$ and boundary conditions as indicated in Figure 1. Here $a=0.001\mathrm{Id}$ in the region $(.25,.75)\times(.25,.75)$ and $a=\mathrm{Id}$ elsewhere (Id denotes the identity matrix); see Figure 1. We can think of this problem as modeling the steady state flow of a fluid through a porous medium with permeability given by a. In Figure 2 we display the streamlines of the approximations to the velocity field $-a\nabla u$ for the approximation given by the RT mixed method of order 1 (left) as well as that given by a postprocessing of the CG method of order 2 (right). The results are very similar. Notice that the singularity of the flow around the corners of the low permeability region $(.25,.25)\times(.75,.75)$ makes this a hard test problem.

We discuss the postprocessing procedure for a hybridized version of the CG method, although it can be applied directly to the standard CG formulation. This is not only because it is easier to understand the first step of the postprocessing using the hybridized formulation (as mentioned previously), but also because the hybridized



Boundary conditions:

$$u = 0 \quad \text{on } \{(1,y) : y \in [0,1]\},$$

$$\boldsymbol{q}\cdot\boldsymbol{n} = \begin{cases} -1 & \{(0,y) : y \in [0,.5)\}, \\ 0 & \{(0,y) : y \in [.5,1]\}, \\ 0 & \{(x,1) : x \in [0,1]\}, \\ 0 & \{(x,0) : x \in [0,1]\}. \end{cases}$$

FIG. 1. *The computational domain* $\Omega = (0,1)^2$ *with a uniform* 8×8 *mesh. The region of low permeability is indicated in dark gray.*

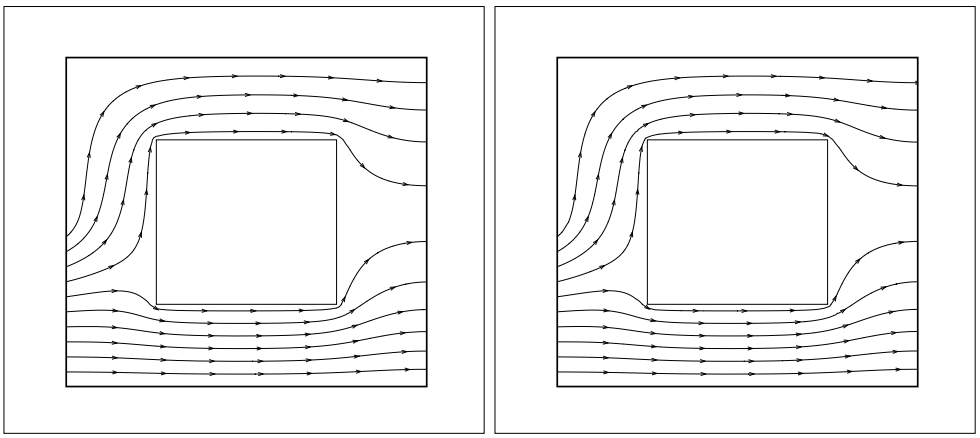


FIG. 2. *Streamlines of the approximate fluxes for the* RT_1 *method (left) and the* RT_1*-postprocessed* CG_2 *method (right) obtained using a uniform* 32×32 *mesh.*

method has interesting features in its own right. The hybridized CG method is obtained as a natural extension of the new perspective introduced in [15] for hybridizing mixed methods. It can be briefly described in two steps. First, we express the approximate solution of the CG method u_h in terms of the data components (g and f) and a *Lagrange multiplier* λ_h. It turns out that for the CG method, λ_h is nothing but the restriction of u_h to the faces of the elements of the triangulation. The second step consists in showing that λ_h can be characterized as the only element of certain set $M_h(g)$ satisfying a weak formulation of the form

$$a_h(\lambda_h, \mu) = b_h(\mu) \qquad \text{for all } \mu \in M_h(0). \tag{1.5}$$

This formulation was also obtained in [5] with the purpose of devising efficient substructuring preconditioners for the CG method.

Hybridization in the context of mixed methods is different from what goes by the name of static condensation in the engineering literature, because the former gives extra information through the Lagrange multiplier, a solution component absent in

static condensation. However, in the hybridized CG case, the fact that the Lagrange multiplier λ_h equals u_h on the element interfaces implies that hybridization and static condensation coincide, except when we have variable degree elements and hanging nodes. In the static condensation approach, the degrees of freedom of the approximate solution u_h must be very carefully chosen in order to ensure the required continuity across interelement boundaries. The data structures needed to enforce such continuity for variable-degree approximations and hanging nodes have attained a high degree of sophistication; see, for example, [21, 22]. On the other hand, if we use the hybridized version of the CG method (1.5), there is no need to enforce any continuity constraint at all. We apply CG on each element without caring about continuity restrictions, as the continuity is automatically enforced by the equations of the method, provided we pick a suitable Lagrange multiplier space $M_h(0)$.

The paper is organized as follows. In section 2, we present the hybridized CG method and briefly discuss the result characterizing λ_h as the unique solution of (1.5). We also discuss extensions to the variable-degree case and hanging nodes. In section 3, we describe the construction of $\boldsymbol{H}(\mathrm{div}, \Omega)$-conforming approximation to the flux. We state the error estimates of the flux approximation and the results on the relationship between our method and the corresponding RT mixed method. We explain how to explicitly construct a local basis for the space required to compute a single-valued numerical flux trace. An estimate of the conditioning of the global system that arises also appears in this section. In section 4, we give all the proofs of the theorems. A numerical study of the approximation properties of these approximations is presented in section 5. We end with some concluding remarks in section 6.

2. Characterization of the Lagrange multiplier. We begin this section by hybridizing the CG method. We then state, discuss, and prove the main result of this section, Theorem 2.1, which characterizes the Lagrange multiplier.

2.1. The hybridized CG method. To hybridize the CG method, we relax the continuity restriction and impose it back through suitably chosen new equations. Since the continuity restriction is enforced in the sets $\mathsf{V}_h(\cdot)$, to relax it means to work instead with the space

$$V_h = \left\{ v \in L^2(\Omega) : v|_K \in \mathcal{P}_k(K) \quad \text{for all} \quad K \in \mathcal{T}_h \right\}. \tag{2.1a}$$

The new approximation U_h in V_h must, however, coincide with u_h, which means, in particular, that it has to be continuous. To enforce the continuity of U_h across interelement boundaries, we force U_h to be equal to the Lagrange multiplier λ_h, which we take in

$$M_h(g) = \{\mu \in \mathcal{C}^0(\mathcal{E}_h) : \ \mu|e \in \mathcal{P}_k(e) \quad \text{for all } e \in \mathcal{E}_h, \mu = g \text{ on } \Gamma_D\}, \tag{2.1b}$$

where

$$\mathcal{E}_h = \{e : e \text{ is a face of } K \text{ for all } K \in \mathcal{T}_h\}. \tag{2.1c}$$

Notice that we are implicitly assuming that the triangulation $\mathcal{T}_h$ does not have hanging nodes. To ensure that $U_h = u_h$, we are going to use an auxiliary variable which approximates $\boldsymbol{q} \cdot \boldsymbol{n} = -a\nabla u \cdot \boldsymbol{n}$ on ∂K for each element K. This additional variable is denoted by $q_{n,h}$ and will be taken in the space

$$W_h = \left\{ p \in L^2(\{\partial K : \ K \in \mathcal{T}_h\}) : \ p|_{\partial K} = v|_{\partial K} \quad \text{for } v \in V_h \right\}. \tag{2.1d}$$

Note that $p \in W_h$ is double-valued in the interior faces of the elements $K \in \mathcal{T}_h$. Thus the hybridized method seeks an approximation to $(u|_{K\in\mathcal{T}_h}, u|_{\mathcal{E}_h}, \boldsymbol{q}\cdot\boldsymbol{n}|_{\partial K, K\in\mathcal{T}_h})$, $(U_h, \lambda_h, q_{n,h})$ in the space $V_h \times M_h(g) \times W_h$. It is defined by

$$\sum_{K\in\mathcal{T}_h} (a\nabla U_h, \nabla v)_K + \sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} = (f, v)_\Omega \qquad \text{for all } v \in V_h, \tag{2.2a}$$

$$U_h = \lambda_h \qquad \text{on } \mathcal{E}_h, \tag{2.2b}$$

$$\sum_{K\in\mathcal{T}_h} \langle q_{n,h}, \mu\rangle_{\partial K} = \langle \mathsf{q_N}\,,\, \mu\rangle_{\Gamma_N} \qquad \text{for all } \mu \in M_h(0). \tag{2.2c}$$

Notice that, by the definition of the space $M_h(0)$, (2.1b), $\mu = v|_{\mathcal{E}_h}$ belongs to $M_h(0)$ whenever $v \in \mathsf{V}_h(0)$. This implies that the last equation can be rewritten as

$$\langle [\![q_{n,h}]\!], v\rangle_{\mathcal{E}_h} = \langle \mathsf{q_N}\,,\, v\rangle_{\Gamma_N} \qquad \text{for all } v \in \mathsf{V}_h(0),$$

where the *jump* of the approximate normal component of the flux is

$$[\![q_{n,h}]\!] := \begin{cases} q_{n,h}|_{\partial K^+} + q_{n,h}|_{\partial K^-} & \text{on the face } e = \partial K^+ \cap \partial K^-, \\ q_{n,h} & \text{on the face } e = \partial K \cap \partial\Omega. \end{cases}$$

We thus see that it enforces a weak continuity of the interelement boundary of the jump of this variable; this is why we call it the *jump condition*. Next, we see that this condition ensures that $U_h = u_h$.

PROPOSITION 2.1. *There exists a unique function $(U_h, \lambda_h, q_{n,h})$ in the space $V_h \times M_h(g) \times W_h$ satisfying the formulation* (2.2)*. Moreover,*

$$U_h = u_h \text{ on } \Omega \quad \text{and} \quad \lambda_h = u_h \text{ on } \mathcal{E}_h.$$

Proof. Since $\lambda_h \in M_h(g)$ and $U_h \in V_h$, we have that $U_h \in \mathsf{V}_h(g)$. Moreover, since $\mathsf{V}_h(0) \subset V_h$, by (2.2a) we have

$$(a\nabla U_h, \nabla v)_\Omega + \sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} = (f, v)_\Omega \qquad \text{for all } v \in \mathsf{V}_h(0),$$

and, by the jump condition (2.2c),

$$(a\nabla U_h, \nabla v)_\Omega = (f, v)_\Omega - \langle \mathsf{q_N}\,,\, v\rangle_{\Gamma_N} \qquad \text{for all } v \in \mathsf{V}_h(0).$$

By the uniqueness of the approximate of the CG method, we immediately obtain that $U_h = u_h$ on Ω and, as a consequence, that $\lambda_h = u_h$ on $\mathcal{E}_h$.

It only remains to prove that the function $q_{n,h}$ exists and is unique. This is equivalent to proving that the trivial solution is the only solution of

$$\sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} = 0 \qquad \text{for all } v \in V_h\,.$$

Since $q_{n,h} \in W_h$, there is a $w \in V_h$ such that $q_{n,h} = w$. Taking $v = w$ in the above equation, we conclude that $q_{n,h} \equiv 0$, as desired. This completes the proof. ∎

2.2. Characterization of the Lagrange multiplier $\boldsymbol{\lambda_h}$. Next, we show how to eliminate the unknowns U_h and $q_{n,h}$ from (2.2) and obtain a formulation only for λ_h. The discussion here is a straightforward generalization of well-known results in domain decomposition [5] to the case when subdomains are reduced to elements. Analogous to the discrete harmonic extensions of [5], we now define a local lifting.

The lifting associates to each $\mathsf{m} \in M_h(\cdot)$ the pair of functions $(\,\mathcal{U}\mathsf{m}\,, \boldsymbol{\mathcal{Q}_n}\mathsf{m}\,) \in \mathcal{P}_k(K) \times \{v|_{\partial K} : v \in \mathcal{P}_k(K)\}$ defined by requiring that

$$(a\nabla\,\mathcal{U}\mathsf{m}\,, \nabla v)_K + \langle \boldsymbol{\mathcal{Q}_n}\mathsf{m}\,, v\rangle_{\partial K} = 0 \qquad \text{for all } v \in \mathcal{P}_k(K), \tag{2.3a}$$

$$\mathcal{U}\mathsf{m} = \mathsf{m} \qquad \text{on } \partial K. \tag{2.3b}$$

In addition, we define a second local mapping that associates to the function $f \in L^2(\Omega)$ the pair of functions $(\,\mathcal{U}f\,, \boldsymbol{\mathcal{Q}_n}f\,) \in \mathcal{P}_k(K) \times \{v|_{\partial K},\ v \in \mathcal{P}_k(K)\}$ defined by

$$(a\nabla\,\mathcal{U}f\,, \nabla v)_K + \langle \boldsymbol{\mathcal{Q}_n}f\,, v\rangle_{\partial K} = (f, v)_K \qquad \text{for all } v \in \mathcal{P}_k(K), \tag{2.4a}$$

$$\mathcal{U}f = 0 \qquad \text{on } \partial K. \tag{2.4b}$$

Notice that $(\,\mathcal{U}\mathsf{m}\,, \boldsymbol{\mathcal{Q}_n}\mathsf{m}\,)$ and $(\,\mathcal{U}f\,, \boldsymbol{\mathcal{Q}_n}f\,)$ are approximations to the solutions of

$$-\operatorname{div}(a\nabla u) = 0, \qquad -\operatorname{div}(a\nabla u) = f \qquad \text{on } K, \tag{2.5a}$$

$$u = \mathsf{m}, \qquad u = 0 \qquad \text{on } \partial K. \tag{2.5b}$$

We are now ready to state the characterization of the CG solution in terms of the Lagrange multiplier, whose proof is at the end of this section.

THEOREM 2.1. *Let $(U_h, \lambda_h, q_{n,h})$ be the solution of the hybridized version of the* CG *method. Then*

$$U_h = \mathcal{U}\lambda_h + \mathcal{U}f \quad \textit{and} \quad q_{n,h} = \boldsymbol{\mathcal{Q}_n}\lambda_h + \boldsymbol{\mathcal{Q}_n}f\,.$$

Moreover, the Lagrange multiplier $\lambda_h \in M_h(g)$ is the unique solution of

$$\sum_{K\in\mathcal{T}_h} (a\,\nabla\,\mathcal{U}\lambda_h, \nabla\,\mathcal{U}\mu\,)_K = (f,\,\mathcal{U}\mu\,)_\Omega - \langle \mathsf{q_N}\,,\,\mu\rangle_{\Gamma_N} \qquad \textit{for all } \mu \in M_h(0).$$

Like other hybridized formulations, the utility of such a result lies in its ease of computation of a "stiffness matrix" for the Lagrange multiplier. Furthermore, once λ_h has been obtained, U_h and $q_{n,h}$ can be easily computed element by element using the local mappings (2.3) and (2.4).

It is interesting to note that $q_{n,h}|_{\partial K}$ is strongly related to what was denoted by $H^h(K)$ in [25]; in fact, when the element K does not have a face lying on the boundary, these two quantities are identical. However, in [25] they are used to uncover a local conservativity property of the CG method, whereas here we use them as an auxiliary means to hybridize it.

Finally, notice that Theorem 2.1 states that the functions $q_{n,h}|_{\partial K}$ need *not* be actually computed to construct the matrix equations for the multiplier λ_h. Indeed, from the definition of the lifting (2.3), we see that we can independently compute $\mathcal{U}\mathsf{m}$ on the element K by solving

$$(a\nabla\,\mathcal{U}\mathsf{m}\,, \nabla v)_K = 0 \qquad \text{for all } v \in \mathcal{P}_k(K) \text{ such that } v = 0 \text{ on } \partial K,$$

$$\mathcal{U}\mathsf{m} = \mathsf{m} \qquad \text{on } \partial K.$$

This implies that $\mathcal{U}\mathsf{m}$ can be written as a linear combination of

$$\dim \mathcal{P}_k(K) - \dim \mathcal{P}_{k-3}(K) = \binom{k+N}{N} - \binom{k-3+N}{N}$$

basis functions, when $k \geq 3$, of course. In two space dimensions ($N = 2$), this means that instead of working with a basis of $(k+2)(k+1)/2$ functions, we can work with a basis of only $3\,k$ functions. In three space dimensions, it means that instead of working with $(k+3)(k+2)(k+1)/6$ basis functions, we only have to work with $(3k^2+3k+2)/2$. Thus, the computation of $\mathcal{U}\mathsf{m}$ can be rendered extremely efficient, especially for high polynomial degrees k. This is especially true if the exact solution is harmonic, that is, if $f = 0$.

2.3. Variable-degree approximations and hanging nodes. The hybridized CG formulation is particularly attractive for variable-degree approximate spaces and meshes with hanging nodes.

We begin by briefly showing how to extend our previous results to the variable-degree case, that is, to the case in which the approximate solution u_h belongs to

$$\mathsf{V}_h(s) = \left\{ v \in \mathcal{C}^0(\Omega) : \ v|_K \in \mathcal{P}_{k(K)}(K), \ v = s \text{ on } \Gamma_D \right\},$$

where the polynomial degree $k(K)$ now varies with as K varies within $\mathcal{T}_h$. We can then hybridize the resulting CG method, just as we hybridized the uniform-degree CG method, if we take

$$\begin{aligned} M_h(g) &= \left\{ \mu \in \mathcal{C}^0(\mathcal{E}_h) : \ \mu|e \in \mathcal{P}_{k(e)}(e) \quad \text{for all } e \in \mathcal{E}_h, \mu = g \text{ on } \Gamma_D \right\}, \\ W_h &= \left\{ w \in L^2(\mathcal{E}_h) : w = v|_{\partial K}, \ \ v \in \mathcal{P}_{k(K)}(K) \quad \text{for all } K \in \mathcal{T}_h \right\}, \\ V_h &= \left\{ v \in L^2(\mathcal{T}_h) : v|_K \in \mathcal{P}_{k(K)}(K) \quad \text{for all } K \in \mathcal{T}_h \right\}. \end{aligned} \tag{2.6}$$

With this, the burden of enforcing the continuity constraint is automatically dealt with by the local mappings which are defined *exactly* as before with k replaced by $k(K)$. While the current practice for implementing variable-degree methods is via transitional basis functions and the minimum degree rule [21], the above hybridization approach removes the continuity matching considerations from the design of shape functions.

To end this subsection, let us briefly address the case of hanging nodes, which is also surprisingly simple to handle by hybridization, even in three dimensions. We only have to define the multiplier space $M_h(g)$ in a suitable way. In fact, we can continue to define $M_h(g)$ by (2.6) provided we redefine the set $\mathcal{E}_h$ there. To do this, we need to introduce the notion of a *maximal face*. A face e of an element $K \in \mathcal{T}_h$ is said to be a maximal face of the triangulation $\mathcal{T}_h$ if it lies on $\partial\Omega$ or whenever there is another element $K' \in \mathcal{T}_h$ such that $e \cap \partial K'$ has nonzero $(N-1)$-Lebesgue measure, $e \cap \partial K'$ is a face of K'. An illustration is given in Figure 3. The new definition of $\mathcal{E}_h$ is simply

$$\mathcal{E}_h = \{ e : e \text{ is a maximal face of the triangulation } \mathcal{T}_h \}. \tag{2.7}$$

2.4. Proof of Theorem 2.1. To prove this result, we need the following lemma.

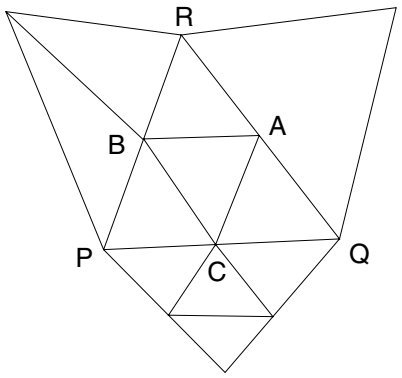


FIG. 3. *Detail of a triangulation. The faces RQ, BR, and AB are maximal, whereas the faces RA and AQ are not.*

LEMMA 2.2 (elementary identities). *We have, for any* $\mathsf{m} \in M_h(\cdot)$, $\mu \in M_h(0)$, *and* $f \in L^2(\Omega)$,

$$\text{(i)} \quad -\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n}\mathsf{m}\,,\mu\rangle_{\partial K} = \sum_{K\in\mathcal{T}_h} (a\nabla\,\mathcal{U}\mathsf{m}\,,\nabla\,\mathcal{U}\mu\,)_K\,,$$

$$\text{(ii)} \quad -\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n} f\,,\mu\rangle_{\partial K} = -\,(f,\,\mathcal{U}\mu\,)_\Omega\,.$$

Proof. We have

$$\begin{aligned} -\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n}\mathsf{m}\,,\mu\rangle_{\partial K} &= -\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n}\mathsf{m}\,,\,\mathcal{U}\mu\,\rangle_{\partial K} \qquad \text{by (2.3b)},\\ &= \sum_{K\in\mathcal{T}_h} (a\nabla\,\mathcal{U}\mathsf{m}\,,\nabla\,\mathcal{U}\mu\,)_K \qquad \text{by (2.3a)}. \end{aligned}$$

This proves the first identity.

Let us prove the second identity. We have

$$\begin{aligned} -\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n} f\,,\mu\rangle_{\partial K} &= -\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n} f\,,\,\mathcal{U}\mu\,\rangle_{\partial K} && \text{by (2.3b)},\\ &= -\,(f,\,\mathcal{U}\mu\,)_\Omega + \sum_{K\in\mathcal{T}_h} (a\nabla\,\mathcal{U}f\,,\nabla\,\mathcal{U}\mu\,)_K && \text{by (2.4a)},\\ &= -\,(f,\,\mathcal{U}\mu\,)_\Omega + \sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n}\mu\,,\,\mathcal{U}f\,\rangle_{\partial K} && \text{by (2.3a)},\\ &= -\,(f,\,\mathcal{U}\mu\,)_\Omega && \text{by (2.4b)}. \end{aligned}$$

This completes the proof. ∎

Proof of Theorem 2.1. By the definition of the local mappings, we have that

$$U_h = \mathcal{U}\lambda_h + \mathcal{U}f \quad \text{and} \quad q_{n,h} = \boldsymbol{\Omega_n}\lambda_h + \boldsymbol{\Omega_n} f\,.$$

This implies that the third equation in the definition of the hybridized version of the CG method (2.2c) can be rewritten as

$$\sum_{K\in\mathcal{T}_h} \langle \boldsymbol{\Omega_n}\lambda_h + \boldsymbol{\Omega_n} f\,,\mu\rangle_{\partial K} = \langle \mathsf{q_N}\,,\,v\rangle_{\Gamma_N} \quad \text{for all } \mu \in M_h(0),$$

or, by Lemma 2.2, as

$$-\sum_{K\in\mathcal{T}_h} (a\nabla\,\mathcal{U}\lambda_h\,,\nabla\,\mathcal{U}\mu\,)_K + (f,\,\mathcal{U}\mu\,)_\Omega = \langle \mathsf{q_N}\,,\,v\rangle_{\Gamma_N}.$$

This completes the proof. □.

3. An $\boldsymbol{H}(\mathrm{div},\Omega)$-conforming approximation of the flux. In this section, we define an $\boldsymbol{H}(\mathrm{div},\Omega)$-conforming approximation, $\boldsymbol{q}_h$, to the flux $\boldsymbol{q} = -a\nabla u$. Then we state, discuss, and prove a theorem about the quality of the resulting approximation as well as the complexity of the algorithm needed to compute it. Although all considerations in this section hold for the variable degree case, for simplicity we restrict ourselves to the uniform degree case spaces defined in (2.1) with no hanging nodes.

3.1. The new approximation to the flux. The key step in the construction of an $\boldsymbol{H}(\mathrm{div},\Omega)$-conforming approximation $\boldsymbol{q}_h$ is the definition of its normal component on the element interfaces. The function $q_{n,h}$ represents an approximation to the normal component of the flux, but unfortunately it is not a single-valued function in general. Notice, however, that by (2.2a), we have

$$\sum_{K\in\mathcal{T}_h} (a\nabla U_h,\nabla v)_K + \sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} = (f,v)_\Omega \quad \text{for all } v\in V_h,$$

so the possibility of constructing a single-valued function $\widehat{\boldsymbol{q}}_h$ satisfying

$$\sum_{K\in\mathcal{T}_h} \langle \widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n},\, v\rangle_{\partial K} = \sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} \quad \text{for all } v\in V_h \tag{3.1}$$

opens up. If such a $\widehat{\boldsymbol{q}}_h$ could be constructed, we could then define the approximate flux $\boldsymbol{q}_h$ as follows: On any simplicial element K, we can set $\boldsymbol{q}_h$ in the RT space

$$V_{RT_\ell}(K) := \mathcal{P}_\ell(K)^N + \boldsymbol{x}\,\mathcal{P}_\ell(K) \tag{3.2a}$$

by requiring that

$$\langle \boldsymbol{q}_h\cdot\boldsymbol{n}, v\rangle_e = \langle \widehat{\boldsymbol{q}}_h\cdot\boldsymbol{n}, v\rangle_e \qquad \text{for all } v\in\mathcal{P}_\ell(e) \text{ for any face } e\subset\partial K, \tag{3.2b}$$

$$(\boldsymbol{q}_h,\boldsymbol{v})_K = -(a\nabla U_h,\boldsymbol{v})_K \quad \text{for all } \boldsymbol{v}\in\mathcal{P}_{\ell-1}(K)^N. \tag{3.2c}$$

Note that the definition (3.2) is a modification of the well-known RT projection, (see (3.12) later). A similar projection was suggested in [4] in the framework of the interior penalty method for Darcy's law and in [19] in the framework of local discontinuous Galerkin methods for the Navier–Stokes equations. It is not difficult to show that a $\boldsymbol{q}_h$ constructed by (3.2b) belongs to $\boldsymbol{H}(\mathrm{div},\Omega)$, thanks to the single-valuedness of the normal component of the numerical trace $\widehat{\boldsymbol{q}}_h$.

Such a construction will yield a flux $\boldsymbol{q}_h$ that is conservative whenever $\ell\le k$. Indeed, we can rewrite (2.2a) as

$$-\sum_{K\in\mathcal{T}_h} (\boldsymbol{q}_h,\nabla v)_K + \sum_{K\in\mathcal{T}_h} \langle \boldsymbol{q}_h\cdot\boldsymbol{n}, v\rangle_{\partial K} = (f,v)_\Omega$$

for all v such that $v|_K\in\mathcal{P}_\ell(K)$ for all $K\in\mathcal{T}_h$. Hence, if we take v to be the characteristic function of a discrete subdomain D_h formed by the union of some elements $K\in\mathcal{T}_h$, we obtain

$$\langle \boldsymbol{q}_h\cdot\boldsymbol{n}, 1\rangle_{\partial D_h} = (f,1)_{D_h},$$

which is the same as the exact conservation property (1.4).

It is interesting to see that there is an extremely simple relation between the normal component of $\boldsymbol{q}_h$ and the approximation $H^h(\cdot)$ defined in [25] or, equivalently, what is called $\tilde{\sigma}_h$ in [9]. Indeed, if D is any union of elements $K \in \mathcal{T}_h$, then from the definition of $\boldsymbol{q}_h$ and that of $H^h(D)$ (see equations (47) and (57) in [25]), we have that

$$\langle \boldsymbol{q}_h \cdot \boldsymbol{n} - H^h(D)\,,\, v\rangle_{\partial D} = 0 \quad \text{for all } v \in V_{h,D} := \{v \in V_h \cap \mathcal{C}^0(D)\}.$$

Since $H^h(D)|_{\partial D}$ belongs to the space of traces on ∂D of the functions in $V_{h,D}$, we see that $H^h(D)$ is the L^2-projection of $\boldsymbol{q}_h \cdot \boldsymbol{n}$ into such space.

3.2. The numerical trace $\widehat{\boldsymbol{q}}_h$. It remains to find the numerical trace $\widehat{\boldsymbol{q}}_h$. To do that, we first notice that if $\widehat{\boldsymbol{q}}_h$ is single valued, then (3.1) takes the form

$$\langle \widehat{\boldsymbol{q}}_h\,,\, [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h} = \sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} \quad \text{for all } v \in V_h.$$

Since the flux on Γ_N is given to be $\mathsf{q_N}$, incorporating this information into the above equation, we get

$$\langle \widehat{\boldsymbol{q}}_h\,,\, [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\Gamma_N} = \sum_{K\in\mathcal{T}_h} \langle q_{n,h}, v\rangle_{\partial K} - \langle \mathsf{q_N}\,,\, v\rangle_{\Gamma_N} \quad \text{for all } v \in V_h. \tag{3.3}$$

In the one-dimensional case $\Omega = (0,1)$, this equation can be readily solved. Indeed, we have that

$$\widehat{\boldsymbol{q}}_h(x_i) = \begin{cases} q_{n,h}(1^-) & \text{if } x_i = 1,\\ q_{n,h}(x_i^-) = -q_{n,h}(x_i^+) & \text{if } x_i \text{ is an interior node},\\ -q_{n,h}(0^+) & \text{if } x_i = 0, \end{cases}$$

where we have used the fact that, by (2.2c), $q_{n,h}(x_i^-) + q_{n,h}(x_i^+) = 0$ on all interior nodes x_i. Let us find expressions for $q_{n,h}$ in terms of the data f and u_h. By (2.2c), $q_{n,h} = \mathsf{q_N}$ on Γ_N, and we get that

$$\widehat{\boldsymbol{q}}_h = \mathsf{q_N} \quad \text{on } \Gamma_N.$$

To find $q_{n,h}$ in the remaining nodes, we simply use (2.2a). Thus, if we let x_i be any node not lying on Γ_N, and let φ_i^+ (resp., φ_i^-) be the linear function with support the interval $I_i^+ = (x_i, x_{i+1})$ (resp., $I_i^- = (x_{i-1}, x_i)$) such that $\varphi_i^+(x_i) = 1$ and $\varphi_i^+(x_{i+1}) = 0$ (resp., $\varphi_i^-(x_i) = 1$ and $\varphi_i^-(x_{i-1}) = 0$), we obtain that

$$\widehat{\boldsymbol{q}}_h(x_i) = \mp\left(a\frac{d}{dx}u_h, \frac{d}{dx}\varphi_i^{\pm}\right)_{I_i^{\pm}} \pm (f, \varphi_i^{\pm})_{I_i^{\pm}}.$$

These expressions have been known for a long time; see the work by J. Wheeler [33] and M. Wheeler [34]. Moreover, in [34], it was shown that the approximation $\widehat{\boldsymbol{q}}_h$ superconverges with order $2\,k$ if the CG method uses polynomials of degree k and is exact, that is,

$$\widehat{\boldsymbol{q}}_h(x_i) = -a\frac{d}{dx}u(x_i),$$

whenever a is a constant.

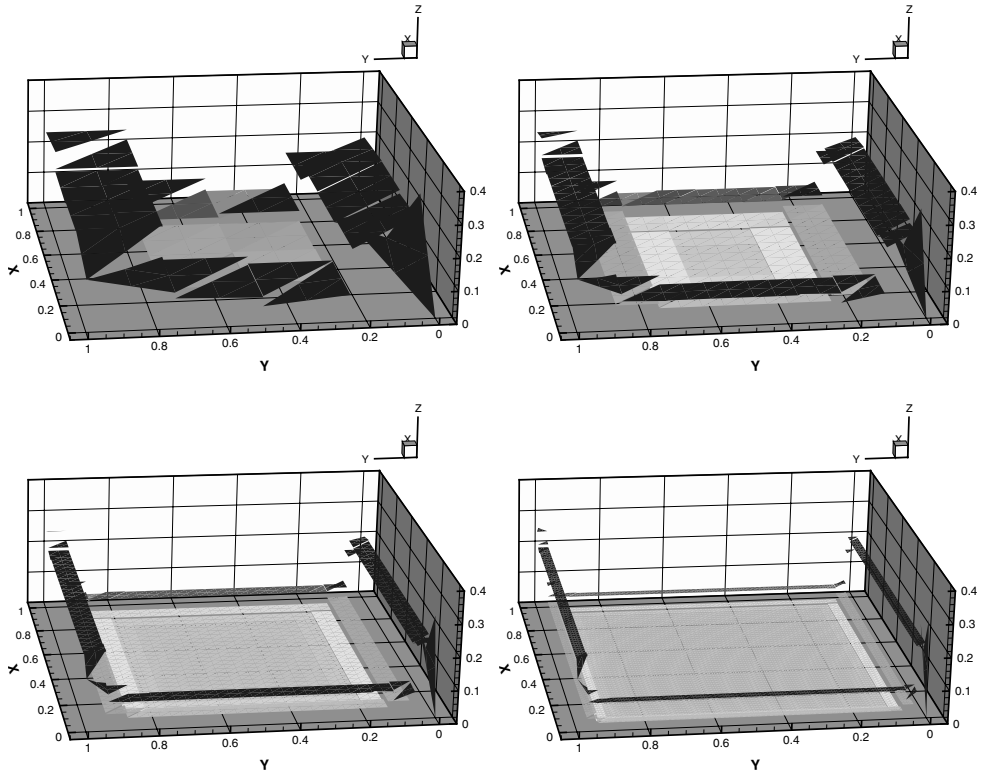


FIG. 4. *Plots of the error* $|\boldsymbol{q} - \boldsymbol{q}_h|$ *with the wrong flux trace choice on successively refined meshes. Computational details: Here* $\boldsymbol{q}_h$ *is obtained by* (3.2) *with* $\widehat{\boldsymbol{q}}_h$ *chosen as the unique function in* $\mathfrak{J}_h$ *satisfying* (3.3). *The parameters are* $a = \mathrm{Id}$, $f = 0$, $\Omega = (0,1) \times (0,1)$, $\Gamma_D = \{0\} \times (0,1)$, *the polynomial degrees are* $k = 1$ *and* $\ell = k - 1$ *(for postprocessing), and the boundary conditions are set in such a way that the exact solution is* $u(x,y) = 1 + x$. *We see that while the error is small far from the boundary, near the boundary the error remains of order one. Therefore, we expect to see an order of convergence of* $1/2$ *in the* L^2*-norm. This is confirmed in Table* 1.

Extensions of the above approach to the multidimensional case for obtaining approximations to the normal component of $\widehat{\boldsymbol{q}}_h$ have been explored by many authors. See [9] for an overview and recent developments, [8] for early computational tricks, [11] for a fully developed technique, and [14, 28] for rigorous error estimates.

Here, we do not use this approach. Instead, we begin by noting that from the formulation (3.3) it is clear that we can only obtain a projection of $\widehat{\boldsymbol{q}}_h$ into the space of jumps

$$\mathfrak{J}_h = \{ [\![w\boldsymbol{n}]\!] |_{\mathcal{E}_h \setminus \Gamma_N} : w \in V_h \}. \tag{3.4}$$

This may seem to suggest choosing $\widehat{\boldsymbol{q}}_h$ in $\mathfrak{J}_h$. We have experimented with such a choice. The results of one such experiment are reported in Figure 4 and Table 1. We found that such a flux approximation is often reasonable away from the boundary, but near $\partial\Omega$ the degradation of the approximation is clearly evident for some problems. Furthermore, from a theoretical standpoint, such a choice appears dubious as the space $\mathfrak{J}_h$ does not contain the constant function. For these reasons, we do *not* advocate it.

The solution we found practically acceptable as well as theoretically sound proceeds by borrowing ideas from the development of the DG method. We select the

TABLE 1
The L^2-norm of the error $\boldsymbol{q} - \boldsymbol{q}_h$ when the wrong flux trace is used. The parameters are the same as those described in Figure 4.

	$k=1$		$k=2$		$k=3$	
h	Error	Order	Error	Order	Error	Order
1/8	0.11E+00	0.46	0.62E-01	0.41	0.40E-01	0.46
1/16	0.77E-01	0.48	0.45E-01	0.46	0.29E-01	0.48
1/32	0.55E-01	0.49	0.32E-01	0.48	0.21E-01	0.49
1/64	0.39E-01	0.50	0.23E-01	0.49	0.15E-01	0.50

following form for the numerical trace:

$$\widehat{\boldsymbol{q}}_h = \begin{cases} \mathsf{q}_{\mathsf{N}} \boldsymbol{n} & \text{on } \Gamma_N, \\ -a\nabla U_h + \alpha\, \boldsymbol{J}_h & \text{on } \Gamma_D, \\ -\{\!\{a\nabla U_h\}\!\} - \boldsymbol{\beta}\, [\![a\nabla U_h \cdot \boldsymbol{n}]\!] + \alpha\, \boldsymbol{J}_h & \text{on } \mathcal{E}_h \setminus \partial\Omega, \end{cases} \tag{3.5}$$

where α and $\boldsymbol{\beta}$ are single-valued bounded (resp., scalar and vector) functions on $\mathcal{E}_h \setminus \partial\Omega$, $\alpha > 0$, and $\boldsymbol{J}_h$ is an element of the space of jumps $\mathfrak{J}_h$ to be determined. A typical choice of the parameters that we have found adequate in our numerical experiments (on uniform meshes) is $\boldsymbol{\beta} \equiv \boldsymbol{0}$ and $\alpha \equiv 1$ (also see Theorem 3.2 for better choices of α on highly nonuniform meshes). Here, we have used the now standard DG notation (cf., e.g., [2]),

$$\{\!\{v\}\!\} = \begin{cases} \frac{1}{2}\,(v^+ + v^-) & \text{on } \mathcal{E}_h^\circ, \\ v & \text{on } \partial\Omega \end{cases} \tag{3.6a}$$

and

$$[\![v\,\boldsymbol{n}]\!] = \begin{cases} v^+\,\boldsymbol{n}^+ + v^-\,\boldsymbol{n}^- & \text{on } \mathcal{E}_h^\circ, \\ v\,\boldsymbol{n} & \text{on } \partial\Omega, \end{cases} \tag{3.6b}$$

where for a piecewise smooth function v, the traces from either side of a mesh face (edge) e are denoted by $v^\pm(\boldsymbol{x}) = \lim_{\epsilon \downarrow 0} v(\boldsymbol{x} - \epsilon\,\boldsymbol{n}^\pm)$ for all $\boldsymbol{x}$ in e (and $\boldsymbol{n}^\pm$ denotes the corresponding unit outward normal on e from either side).

Next, we insert the expression we have selected in (3.5) for the numerical flux $\widehat{\boldsymbol{q}}_h$ into (3.3). This gives us an equation for $\boldsymbol{J}_h$:

$$\begin{aligned} \langle \alpha\, \boldsymbol{J}_h,\ [\![v\boldsymbol{n}]\!] \rangle_{\mathcal{E}_h \setminus \Gamma_N} = {} & \langle\, \{\!\{a\nabla U_h\}\!\} + \boldsymbol{\beta}\, [\![a\nabla U_h \cdot \boldsymbol{n}]\!],\ [\![v\boldsymbol{n}]\!] \rangle_{\mathcal{E}_h \setminus \partial\Omega} \\ & + \langle a\nabla U_h,\, v\boldsymbol{n} \rangle_{\Gamma_D} + \sum_{K \in \mathcal{T}_h} \langle q_{n,h}, v \rangle_{\partial K} - \langle \mathsf{q}_{\mathsf{N}}\,,\, v \rangle_{\Gamma_N}. \end{aligned} \tag{3.7}$$

The computation of $\boldsymbol{J}_h$ from this equation requires solving a global, but well-conditioned, system. The details involved are discussed in the next subsection. For the moment, observe that if we are using the hybridized form of the CG method and have already computed U_h and $q_{n,h}$, the right-hand side of (3.7) can be computed using integrations only on element boundaries.

On the other hand, if we have computed U_h using a standard CG implementation without hybridization (and so do not have access to $q_{n,h}$), we can still use the above

postprocessing. Indeed, by using (2.2a), we can transform (3.7) into an equation that is more convenient for this case:

(3.8)
$$\langle \alpha \boldsymbol{J}_h,\ [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\Gamma_N} = \langle\, \{\!\{a\nabla U_h\}\!\} + \boldsymbol{\beta}\, [\![a\nabla U_h\cdot\boldsymbol{n}]\!],\ [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\partial\Omega} + \langle a\nabla U_h,\, v\boldsymbol{n}\rangle_{\Gamma_D} - \sum_{K\in\mathcal{T}_h} (a\nabla U_h, \nabla v)_K + (f,v)_\Omega - \langle \mathsf{q_N}\,,\, v\rangle_{\Gamma_N}.$$

Observe that whenever $[\![v\boldsymbol{n}]\!] = 0$ on $\mathcal{E}_h\setminus\Gamma_N$, i.e., whenever $v\in \mathsf{V}_h(0)$, the right-hand side of the above equation is equal to zero by the definition of the CG method, by (1.3), and by Proposition 2.1. Therefore, this equation defines $\boldsymbol{J}_h$ uniquely. When using spaces of high polynomial degrees, it is preferable to use (3.7) instead of (3.8), as the former involves faster quadratures.

This completes the definition of the numerical trace $\widehat{\boldsymbol{q}}_h$. To summarize, $\widehat{\boldsymbol{q}}_h$ is defined by (3.3), wherein $\boldsymbol{J}_h$ is the unique function in $\boldsymbol{\mathfrak{J}}_h$ satisfying (3.7) or (3.8). Let us point out that this definition of the numerical trace reproduces constant fluxes. More precisely, if $-a\nabla U_h$ is a constant vector, say $\boldsymbol{c}$, then $\widehat{\boldsymbol{q}}_h$ is also $\boldsymbol{c}$. To see this, note that in this case we must have $\mathsf{q_N} = \boldsymbol{c}\cdot\boldsymbol{n}$ and $f=0$, so that (3.8) becomes

$$\langle \alpha \boldsymbol{J}_h,\ [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\Gamma_N} = -\langle \boldsymbol{c},\ [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\Gamma_N} + \sum_{K\in\mathcal{T}_h} (\boldsymbol{c},\nabla v)_K - \langle \boldsymbol{c}\cdot\boldsymbol{n}\,,\, v\rangle_{\Gamma_N} = 0.$$

This implies $\boldsymbol{J}_h\equiv\boldsymbol{0}$, and hence $\widehat{\boldsymbol{q}}_h = \boldsymbol{c}$, as claimed.

Let us end this subsection by relating our approach to compute $\widehat{\boldsymbol{q}}_h$ to that proposed in [27]. In such an approach, the numerical trace $\widehat{\boldsymbol{q}}_h$ is taken as in (3.5) with $\beta=0$ and $\alpha = 1/h$, where $\boldsymbol{J}_h$ is taken in the space

$$\boldsymbol{\mathfrak{J}}_{h,0} = \{\, [\![w\boldsymbol{n}]\!]|_{\mathcal{E}_h\setminus\Gamma_N} : w|_K\in\mathcal{P}_0(K)\ \text{ for all } K\in\mathcal{T}_h\}$$

and is defined by requiring that

$$\langle \alpha \boldsymbol{J}_h,\ [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\Gamma_N} = \langle\, \{\!\{a\nabla U_h\}\!\},\ [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\partial\Omega} + \langle a\nabla U_h,\, v\boldsymbol{n}\rangle_{\Gamma_D} + (f,v)_\Omega - \langle \mathsf{q_N}\,,\, v\rangle_{\Gamma_N}$$

be satisfied for all $v\in\boldsymbol{\mathfrak{J}}_{h,0}$. Note that our $\boldsymbol{J}_h$ also satisfies this formulation, in the case in which $\beta=0$ and $\alpha=1/h$, since the formulation (3.8) reduces to the one under consideration when $v\in\boldsymbol{\mathfrak{J}}_h$ is restricted to $v\in\boldsymbol{\mathfrak{J}}_{h,0}$.

3.3. The computation of $\boldsymbol{J_h}$. Next, we discuss the computation of $\boldsymbol{J}_h$ through solution of (3.7) or (3.8). First, in order to represent $\boldsymbol{J}_h$ in computations we need a basis for the space $\boldsymbol{\mathfrak{J}}_h$ of jumps. We construct a local basis for $\boldsymbol{\mathfrak{J}}_h$ extending a similar construction carried out in [17, 18] in the context of Stokes flow. Second, we need to solve for $\boldsymbol{J}_h$ from (3.7) or (3.8). We show that this can be accomplished by solving a square system whose matrix is well conditioned. Thus, we conclude that the computational complexity needed to solve for $\boldsymbol{J}_h$ is negligible with respect to that required to solve for the multiplier λ_h. Proofs of all results here are given in section 4.

The basis is easiest to see in the lowest order case (i.e., when $k=1$). In two dimensions, this basis is closely related to the "wedge" basis functions obtained in [17]. However, in three dimensions, it is different from that given in [18], so let us begin by describing our lowest order basis in three dimensions. For a mesh vertex $\boldsymbol{z}$ and a mesh element K having $\boldsymbol{z}$ as a vertex, let $\lambda_{\boldsymbol{z},K}$ denote the linear function on K which equals one on $\boldsymbol{z}$ and zero at all other vertices of K and let $\phi_{\boldsymbol{z},K}$ denote its extension

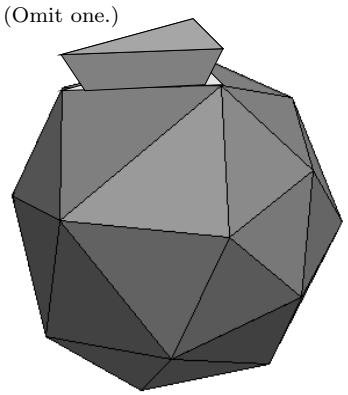


FIG. 5. *Illustration of the elements connected to one vertex* $\boldsymbol{z}$ *and the omission of one* $\phi_{z,K}$ *to construct a basis.*

by zero from K to all Ω. Then clearly the restrictions of $[\![\phi_{\boldsymbol{z},K}\boldsymbol{n}]\!]$ on $\mathcal{E}_h\backslash\Gamma_N$ are in $\mathfrak{J}_h$. However, they are not all linearly independent, because for any vertex $\boldsymbol{z}$ not on Γ_D, the sum of the functions $[\![\phi_{\boldsymbol{z},K}\boldsymbol{n}]\!]|_{\mathcal{E}_h\backslash\Gamma_N}$ over all K sharing $\boldsymbol{z}$ vanishes. Therefore, we must omit one function per vertex to get a basis: For each $\boldsymbol{z}$, we define $V_{\boldsymbol{z}}$ as the set of functions $\phi_{\boldsymbol{z},K}$ for all K having $\boldsymbol{z}$ as a vertex. Then for vertices $\boldsymbol{z}$ not on $\overline{\Gamma}_D$, we define $V_{\boldsymbol{z}}^\star$ as the set obtained by omitting (any) one member of $V_{\boldsymbol{z}}$ (see Figure 5), while for vertices on $\overline{\Gamma}_D$, we define $V_{\boldsymbol{z}}^\star = V_{\boldsymbol{z}}$. Then, by a straightforward generalization of the arguments in [17, Proposition 4.2], one can prove that the set

$$\mathfrak{B}^1 = \{\, [\![\phi\,\boldsymbol{n}]\!]\big|_{\mathcal{E}_h\backslash\Gamma_N} :\ \phi \in V_{\boldsymbol{z}}^\star \text{ for all mesh vertices } \boldsymbol{z}\}$$

is linearly independent, so it forms a basis for $\mathfrak{J}_h$.

Next, we describe one possible extension of this basis construction to the higher order case. For any given simplex $S \in \mathbb{R}^N$ with vertices $\boldsymbol{x}_{i,S}$, $i = 1, \ldots, N+1$, we define the points in its principal lattice (of order k) [13] by

$$\boldsymbol{x}_{\boldsymbol{\alpha},S} = \sum_{j=1}^{N+1} \alpha_j\, \boldsymbol{x}_{j,S},$$

where $\boldsymbol{\alpha}$ is taken in $\mathcal{A}_N^k = \{(\alpha_1, \ldots, \alpha_{N+1}) :\ k\,\alpha_j \in \{0, 1, \ldots, k\} \text{ and } \sum_{j=1}^{N+1} \alpha_j = 1\}$. We associate to each point $\boldsymbol{x}_{\boldsymbol{\alpha},S}$ the standard Lagrange finite element basis function $v_{\boldsymbol{\alpha},S}$ defined as the unique function in $\mathcal{P}_k(S)$ satisfying

$$v_{\boldsymbol{\alpha},S}(\boldsymbol{x}_{\boldsymbol{\beta},S}) = \begin{cases} 1 & \text{if } \boldsymbol{\alpha} = \boldsymbol{\beta}, \\ 0 & \text{otherwise} \end{cases}$$

for all $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in $\mathcal{A}_N^k$. Let $\phi_{\boldsymbol{\alpha},S}$ be the extension by zero to Ω of $v_{\boldsymbol{\alpha},S}$. Since the basis will be constructed using the jumps of these functions across element interfaces, we will need to separate the functions associated to the points on element interfaces, which we collect in

$$\mathcal{G}_h^k = \{\boldsymbol{x}_{\boldsymbol{\alpha},S} \in \partial K : K \in \mathcal{T}_h, \boldsymbol{\alpha} \in \mathcal{A}_N^k\}. \tag{3.9}$$

To any $\boldsymbol{z}$ in $\mathcal{G}_h^k$, we associate more than one $\phi_{\boldsymbol{\alpha},K}$ if more than one simplex shares $\boldsymbol{z}$. We collect these functions in $V_{\boldsymbol{z}} = \{\phi_{\boldsymbol{\alpha},K} : \boldsymbol{x}_{\boldsymbol{\alpha},K} = \boldsymbol{z}\}$ and define

$$V_{\boldsymbol{z}}^\star = \begin{cases} V_{\boldsymbol{z}} & \text{if } \boldsymbol{z} \in \overline{\Gamma}_D, \\ V_{\boldsymbol{z}} \setminus \{\phi_{\boldsymbol{\alpha}^\star,K^\star} \text{ for some } \boldsymbol{x}_{\boldsymbol{\alpha}^\star,K^\star} = \boldsymbol{z}\} & \text{otherwise,} \end{cases}$$

where, as in the lowest order case, we have selected (arbitrarily) one degree of freedom (represented by the multi-index $\boldsymbol{\alpha}^\star$ and $K^\star$) for every $\boldsymbol{z}$ in $\mathcal{G}_h^k \setminus \overline{\Gamma}_D$ and omitted the corresponding Lagrange function $\phi_{\boldsymbol{\alpha}^\star,K^\star}$. With this notation, we have the following result, whose proof follows by generalizing the above-mentioned arguments for the lowest order case.

THEOREM 3.1. *The set*

$$\mathfrak{B}^k = \{ [\![\phi\, \boldsymbol{n}]\!]\big|_{\mathcal{E}_h \setminus \Gamma_N} : \phi \in V_{\boldsymbol{z}}^\star \text{ for all } \boldsymbol{z} \in \mathcal{G}_h^k\}$$

is a basis for $\mathfrak{J}_h$.

Now that a local basis of the jump space $\mathfrak{J}_h$ has been constructed, we can compute the representation of the jump function $\boldsymbol{J}_h$ in the basis $\mathfrak{B}^k$ by using (3.7) or (3.8). For example, to solve for $\boldsymbol{J}_h$ using (3.7), we begin by introducing an extension operator T_h from the space of jumps $\mathfrak{J}_h$ to the space V_h, constructed in such a way that we have

$$[\![T_h(\boldsymbol{J}_h)\, \boldsymbol{n}]\!] = \boldsymbol{J}_h \qquad \text{on } \mathcal{E}_h \setminus \Gamma_N, \tag{3.10a}$$

$$T_h(\boldsymbol{J}_h)\big|_{K^\star,\, \boldsymbol{x}_{\boldsymbol{\alpha}^\star,K^\star}} = 0. \tag{3.10b}$$

Here and elsewhere, we use the notation $w|_{K,\boldsymbol{r}}$ to denote the limit of the function $w(\boldsymbol{x})$ as $\boldsymbol{x}$ approaches $\boldsymbol{r}$ from within K. One can easily verify that the choice

$$T_h(\boldsymbol{J}_h) = \sum_{\boldsymbol{z} \in \mathcal{G}_h^k} \sum_{\phi \in V_{\boldsymbol{z}}^\star} c_\phi\, \phi \quad \text{whenever} \quad \boldsymbol{J}_h = \sum_{\boldsymbol{z} \in \mathcal{G}_h^k} \sum_{\phi \in V_{\boldsymbol{z}}^\star} c_\phi\, [\![\phi\, \boldsymbol{n}]\!] \tag{3.11}$$

satisfies both properties of (3.10). Then, for any $\boldsymbol{Y}_h \in \mathfrak{J}_h$, setting $v = T_h(\boldsymbol{Y}_h)$ in (3.7) and using (3.10a), we get that $\boldsymbol{J}_h$ satisfies

$$\begin{aligned} \langle \alpha\, \boldsymbol{J}_h,\, \boldsymbol{Y}_h \rangle_{\mathcal{E}_h \setminus \Gamma_N} &= \langle\, \{\!\{ a \nabla U_h \}\!\} + \boldsymbol{\beta}\, [\![a \nabla U_h \cdot \boldsymbol{n}]\!],\ [\![T_h(\boldsymbol{Y}_h) \boldsymbol{n}]\!] \rangle_{\mathcal{E}_h \setminus \partial\Omega} \\ &\quad + \langle a \nabla U_h,\, T_h(\boldsymbol{Y}_h) \boldsymbol{n} \rangle_{\Gamma_D} + \sum_{K \in \mathcal{T}_h} \langle q_{n,h}, T_h(\boldsymbol{Y}_h) \rangle_{\partial K} - \langle \mathsf{q}_{\mathsf{N}}\, ,\, T_h(\boldsymbol{Y}_h) \rangle_{\Gamma_N} \\ &\quad \equiv F(\boldsymbol{Y}_h). \end{aligned}$$

This shows that $\boldsymbol{J}_h$ is the unique solution of a square system.

The next result shows that this square system is well conditioned. Let $[\boldsymbol{J}_h]$ denote the vector of coefficients in the expansion of $\boldsymbol{J}_h$ in the basis $\mathfrak{B}^k$. We place some minimal assumptions on the mesh from now on. As per standard terminology, we say that the mesh $\mathcal{T}_h$ is shape regular if, letting ρ_K be the diameter of the largest ball contained in K, the ratios $\gamma_K = \operatorname{diam}(K)/\rho_K$ are uniformly bounded by some fixed constant γ for all K. If we use the parameter α on every mesh face e to scale by the measure of the face, namely, $|e|$, then we obtain a well-conditioned matrix as stated in the following theorem.

THEOREM 3.2. *Let M be the matrix defined by*

$$[\boldsymbol{Y}_h]^t M [\boldsymbol{Z}_h] = \langle \alpha\, \boldsymbol{Z}_h, \boldsymbol{Y}_h \rangle_{\mathcal{E}_h \setminus \Gamma_N} \qquad \textit{for all } \boldsymbol{Y}_h \textit{ and } \boldsymbol{Z}_h \in \mathfrak{J}_h\,.$$

TABLE 2
Condition number of M with $\alpha \equiv 1$ for different mesh levels using different polynomial approximations.

h	1	1/2	1/4	1/8	1/16	1/32
$k=1$	3.4	5.7	6.8	7.3	7.5	7.5
$k=2$	2.0	3.4	5.9	7.5	8.7	10.4
$k=3$	2.1	5.8	5.7	8.6	11.8	10.9

Then, whenever

$$\alpha|_e = \frac{\zeta|_e}{|e|} \quad \text{on } e \in \mathcal{E}_h \setminus \Gamma_N$$

for some piecewise constant function ζ on $\mathcal{E}_h \setminus \Gamma_N$ satisfying $0 < \zeta_\star \le \zeta \le \zeta^\star$, the spectral condition number of M, namely, κ_M, is uniformly bounded by

$$\kappa_M \le C_0\, \zeta^\star/\zeta_\star,$$

where $C_0 > 0$ is independent of the number of mesh elements (but depends on the space dimension N, the polynomial degree k, and the shape regularity constant γ).

This theorem implies that to compute the solution $\boldsymbol{J}_h$ of (3.7) by the method of conjugate gradients, we need a number of iterations that is independent of the number of unknowns. In Table 2, we numerically verify this fact for $k = 1, 2$, and 3 on a sequence of uniform meshes. Since the meshes are uniform, we have simply taken $\alpha \equiv 1$ for this computation. Notice that, as expected, the condition numbers observed do not vary significantly as the mesh size h is reduced. For practical computations, one often uses the method of conjugate gradients to solve for the Lagrange multiplier λ_h. Since the system for λ_h has condition number $O(h^{-2})$ (cf. [24]) without any preconditioner, it is clear that the cost of computing $\boldsymbol{J}_h$ is a negligible addition to the cost of solving for λ_h.

3.4. Error analysis. In this subsection, we give a priori error estimates for our new postprocessed flux approximation $\boldsymbol{q}_h$. Recall that $\boldsymbol{q}_h$ is computed by the following steps:

1. Compute the CG solution U_h.
2. Using this U_h in (3.7), compute the unique function $\boldsymbol{J}_h$ in $\mathfrak{J}_h$ satisfying (3.7).
3. Set the flux trace $\widehat{\boldsymbol{q}}_h$ by substituting the $\boldsymbol{J}_h$ computed above in (3.3).
4. Solve for $\boldsymbol{q}_h$ element by element using (3.2), with the data set by the U_h and $\widehat{\boldsymbol{q}}_h$ computed above.

Notice that the last step involves equations very similar to the well-known RT projection defined as follows: We denote by $\pi_\ell\, \boldsymbol{q}$ the RT projection [29] of the function $\boldsymbol{q}$, which is the unique function in

$$V_{RT_\ell}(K) := \mathcal{P}_\ell(K)^N + \boldsymbol{x}\, \mathcal{P}_\ell(K), \tag{3.12a}$$

satisfying

$$\langle \pi_\ell\, \boldsymbol{q} \cdot \boldsymbol{n}, v\rangle_e = \langle \boldsymbol{q} \cdot \boldsymbol{n}, v\rangle_e \text{ for all } v \in \mathcal{P}_\ell(e) \text{ for any face } e \subset \partial K, \tag{3.12b}$$

$$(\pi_\ell\, \boldsymbol{q}, \boldsymbol{v})_K = (\boldsymbol{q}, \boldsymbol{v})_K \quad \text{for all } \boldsymbol{v} \in \mathcal{P}_{\ell-1}(K)^N. \tag{3.12c}$$

It is well known [6] that the domain of definition of π_ℓ is slightly smaller than $\boldsymbol{H}(\mathrm{div},\Omega)$. We shall tacitly assume that the exact flux $\boldsymbol{q}$ is smooth enough so that π_ℓ can be applied to it (e.g., $\boldsymbol{q}$ in $\boldsymbol{H}(\mathrm{div},\Omega)\cap L^p(\Omega)$ with $p>2$ is enough when $N=2$). Because of the similarity of (3.12) to (3.2), we shall refer to our flux approximation $\boldsymbol{q}_h$ as the *RT_ℓ-postprocessed CG_k flux.*

To describe our error estimates for this flux approximation, we need the following notation. We set $\mathrm{diam}(K)=h_K$, and h is the maximum of h_K over all K in $\mathcal{T}_h$. For Sobolev norms, we denote by $\|\cdot\|_{\ell,\mathcal{D}}$ and $|\cdot|_{\ell,\mathcal{D}}$ the H^ℓ-norm and seminorm, respectively, on $\mathcal{D}$. We also set

(3.13a)
$$\boldsymbol{\mathcal{V}}^0_{h,\ell} := \{\boldsymbol{v}\in\boldsymbol{H}(\mathrm{div},\Omega):\ \nabla\cdot\boldsymbol{v}=0, \boldsymbol{v}|_K\in\mathcal{P}_\ell(K)^N \text{ for all } K\in\mathcal{T}_h, \boldsymbol{v}\cdot\boldsymbol{n}|_{\Gamma_N}=0\},$$

and denote by $\mathbb{P}_\ell$ the weighted L^2-projection into $\boldsymbol{\mathcal{V}}^0_{h,\ell}$ defined by

(3.13b)
$$(a^{-1}(\mathbb{P}_\ell\,\boldsymbol{q}-\boldsymbol{q}),\boldsymbol{v})_\Omega=0 \qquad \text{for all } \boldsymbol{v}\in\boldsymbol{\mathcal{V}}^0_{h,\ell}.$$

Finally, we set $\alpha^\star := \max_{e\in\mathcal{E}_h\setminus\Gamma_N}\alpha|_e$ and $\alpha_\star := \min_{e\in\mathcal{E}_h\setminus\Gamma_N}\alpha|_e$. With these notations, we have the following result.

THEOREM 3.3. *Let $\boldsymbol{q}_h$ be the RT_ℓ-postprocessed CG_k flux for an integer $0\le\ell\le k$. Then the following statements hold:*

1. *The RT_ℓ-postprocessed CG_k flux $\boldsymbol{q}_h$ is in $\boldsymbol{H}(\mathrm{div},\Omega)$ and satisfies*
$$\mathrm{div}(\pi_\ell\,\boldsymbol{q}-\boldsymbol{q}_h)=0.$$
In particular, it satisfies the exact conservativity property (1.4).
2. *If $\ell>1$ and $a(\boldsymbol{x})$ is constant on each mesh element, $\mathbb{P}_{\ell-1}(\pi_\ell\,\boldsymbol{q}-\boldsymbol{q}_h)=0$.*
3. *The divergence of the flux approximation satisfies*
$$\|\nabla\cdot(\boldsymbol{q}-\boldsymbol{q}_h)\|_{0,\Omega}\le C_1 h^{\min(\ell,s)+1}\,|\,f\,|_{s+1,\Omega}.$$
4. *If $a(\boldsymbol{x})|_K$ is in $W^{1,\infty}(K)$ for all mesh elements K, and the mesh $\mathcal{T}_h$ is quasi uniform, then the following error estimate holds:*
$$\|\boldsymbol{q}-\boldsymbol{q}_h\|_{0,\Omega}\ \le\ C_2 h^{\min(k,\,\ell+1,s)}\big(|\,\boldsymbol{q}\,|_{s,\Omega}+|u|_{s,\Omega}\big).$$

In the inequalities above, C_1 and $C_2/(1+\alpha^\star/\alpha_\star)$ are independent of $\boldsymbol{q}$ and h (but dependent on k, N, $\boldsymbol{\beta}$, a, and γ).

The first identity of the theorem can be interpreted as a superconvergence property for the divergence. Indeed, if the load f is a piecewise polynomial satisfying $f|_K\in\mathcal{P}_\ell(K)$, then the exact and discrete divergences coincide, i.e., $\mathrm{div}(\boldsymbol{q}-\boldsymbol{q}_h)=0$, because of a well-known commutativity property of π_ℓ. In one space dimension, it states that the difference between $\pi_\ell\boldsymbol{q}$ and $\boldsymbol{q}_h$ is just a constant; moreover, if the Neumann boundary is not empty, $\pi_\ell\boldsymbol{q}$ and $\boldsymbol{q}_h$ are *identical*. This implies that at each node x_i we have that

$$\boldsymbol{q}_h(x_i)=\boldsymbol{q}(x_i),$$

by definition of the projection π_ℓ. The fact that this holds independently of how we chose the parameter $\boldsymbol{\beta}$ in (3.5) is remarkable, although this fits very well with similar results obtained in [12]. Notice also that when the Neumann boundary is empty but

a is piecewise constant, the second identity states that $\pi_\ell \boldsymbol{q}$ and $\boldsymbol{q}_h$ are also identical, provided $\ell \geq 1$.

Next, we compare our approximation $\boldsymbol{q}_h$ to the corresponding RT approximation. We begin by recalling the standard result that *all statements of Theorem* 3.3 *continue to hold if we replace* $\boldsymbol{q}$ *by the approximation to the flux given by the* RT_ℓ *method;* i.e., both the standard RT_{k-1} method and our new RT_{k-1}-postprocessed CG_k method produce $H(\mathrm{div})$-conforming approximations to the flux $\boldsymbol{q}$ that converge at the same order. This indicates that the RT_{k-1}-postprocessed CG_k method is competitive with the RT_{k-1} method. Indeed, to compare their computational complexities, we recall the earlier observation that the cost of the computation of our approximation $\boldsymbol{q}_h$ is negligible compared to that of solving for the Lagrange multiplier λ_h. The condition number of the Lagrange multiplier system in the CG case as well as the RT case [24] is $O(h^{-2})$, so in both cases, the cost of solving for λ_h dominates the cost of the computation of $\boldsymbol{q}_h$. Thus the relative size of the stiffness matrices for the Lagrange multiplier becomes the deciding factor.

It is not difficult to see that this matrix for the CG_k method has smaller size than that of the RT_{k-1} method. Let us show this in the case of a two-dimensional simply connected domain Ω. We denote the number of mesh vertices, edges, and triangles by n_e, n_v, and n_t, respectively. The number of degrees of freedom of the Lagrange multipliers for the CG_k method is $(k-1)n_e + n_v$, whereas it is kn_e for the RT_{k-1} method. Since $n_v - n_e + n_t = 1$, we see that the Lagrange multipliers of the CG_k method have $(n_t - 1)$ fewer degrees of freedom than RT_{k-1}. This is a significant difference in practice. The numerical experiments of section 5 show that for the same mesh, the approximations given by the RT_{k-1}-postprocessed CG_k method and the RT_{k-1} method are very similar. This shows that the former method may be better than the latter. A final point reinforcing this conclusion is obtained by comparing the approximation to u given by both methods. The u_h of the CG_k method converges in the L^2-norm with order $k+1$ when the exact solution is smooth. However, the approximation to u given by the RT_{k-1} method converges only with order k. Of course, following [1], we can use the Lagrange multipliers to obtain a locally postprocessed approximation that also converges with order $k+1$, but such postprocessing is not available for arbitrary values of k. Moreover, our numerical results show that for $k \in \{0, 1, 2\}$, the CG_k and the postprocessed RT_{k-1} methods produce roughly similar approximations to u.

4. Proofs.

4.1. Norm equivalences. In this subsection, we will prove the condition number estimate of Theorem 3.2 using certain norm equivalences. Recall that the extension operator T_h is defined in (3.10) and that $[\boldsymbol{J}_h]$ is the vector representation of the function $\boldsymbol{J}_h$ in the basis $\mathfrak{B}^k$. Define the norms

$$\|\boldsymbol{J}_h\|_S = \left(\sum_{e \in \mathcal{E}_h \setminus \Gamma_N} \frac{1}{|e|} \|\boldsymbol{J}_h\|_{0,e}^2 \right)^{1/2} \quad \text{and} \quad \|\boldsymbol{J}_h\|_T = \left(\sum_{K \in \mathcal{T}_h} \frac{1}{|K|} \|T(\boldsymbol{J}_h)\|_{0,K}^2 \right)^{1/2},$$

where $|K|$ and $|e|$ denote the measures (in their respective dimensions) of an element K and face e, respectively. The following lemma shows that the three norms $\|\boldsymbol{J}_h\|_S$, $\|\boldsymbol{J}_h\|_T$, and $\|\,[\boldsymbol{J}_h]\,\|_{\ell^2}^2$ are equivalent.

LEMMA 4.1. *There is a constant C independent of the mesh size (but depending on the degree k, dimension N, and the shape regularity constant γ) such that*

$$\frac{1}{C}\,\|\boldsymbol{J}_h\|_T^2 \;\le\; \|\boldsymbol{J}_h\|_S^2 \;\le\; C\,\|\boldsymbol{J}_h\|_T^2 \tag{4.1}$$

and

$$\frac{1}{C}\,\|\boldsymbol{J}_h\|_S^2 \;\le\; \big\|\,[\boldsymbol{J}_h]\,\big\|_{\ell^2}^2 \;\le\; C\,\|\boldsymbol{J}_h\|_S^2 \tag{4.2}$$

for all $\boldsymbol{J}_h \in \mathfrak{J}_h$.

Proof. First, let us prove the upper bound of (4.1). Recall that by (3.10a), $\boldsymbol{J}_h$ and $[\![T_h(\boldsymbol{J}_h)\,\boldsymbol{n}]\!]$ coincide for any $\boldsymbol{J}_h$ in $\mathfrak{J}_h$, so by standard trace inequalities,

$$\frac{1}{|e|}\|\boldsymbol{J}_h\|_{0,e}^2 \le C\sum_{K\in\mathcal{K}_e}\frac{1}{|K|}\big\|T_h(\boldsymbol{J}_h)\big\|_{0,K}^2, \tag{4.3}$$

where $\mathcal{K}_e$ denotes the set of elements $K \in \mathcal{T}_h$ such that e is a face of K. Since $\mathcal{K}_e$ has at most two elements for any mesh face e, summing over all edges in $\mathcal{E}_h \setminus \Gamma_N$, we obtain the upper bound in (4.1).

Next, let us prove the lower bound of (4.1). By standard scaling arguments using the principal lattice $\mathcal{A}_N^k$ on any mesh element K, we have

$$\begin{aligned}\frac{C}{|K|}\|T_h(\boldsymbol{J}_h)\|_{0,K}^2 &\le \sum_{\boldsymbol{\alpha}\in\mathcal{A}_N^k}\big|(T_h(\boldsymbol{J}_h))(\boldsymbol{x}_{\boldsymbol{\alpha},K})\big|^2\\ &= \sum_{\boldsymbol{z}\in\mathcal{G}_h^k\cap\partial K} T_h(\boldsymbol{J}_h)\big|_{K,\boldsymbol{z}}^2,\end{aligned} \tag{4.4}$$

where, as before, $w|_{K,\boldsymbol{z}}$ denote the limit of $w(\boldsymbol{x})$ as $\boldsymbol{x}$ approaches $\boldsymbol{z}$ from within K.

We need to bound each of the terms in (4.4) using norms of $\boldsymbol{J}_h$. Let us first consider the case when $\boldsymbol{z}$ is not on Γ_D. For such a $\boldsymbol{z}$, recalling the way we constructed the basis of Theorem 3.1, note that there is a mesh element $K^\star$ such that $\boldsymbol{z} = \boldsymbol{x}_{\boldsymbol{\alpha}^\star,K^\star}$, where the limit $T_h(\boldsymbol{J}_h)|_{K^\star,\,\boldsymbol{z}}$ is zero; see (3.10b). Using this fact, it is easy to see that we can write $T_h(\boldsymbol{J}_h)\big|_{K,\boldsymbol{z}}$ as the telescoping sum

$$T_h(\boldsymbol{J}_h)\big|_{K,\boldsymbol{z}} = \sum_{i=1}^{m}\Big[T_h(\boldsymbol{J}_h)\big|_{K_i,\boldsymbol{z}} - T_h(\boldsymbol{J}_h)\big|_{K_{i+1},\boldsymbol{z}}\Big] \tag{4.5}$$

for some collection of mesh elements K_i such that $\boldsymbol{z}$ is in $\bar{K}_i$, $K_1 = K$, $K_{m+1} = K^\star$, and $K_i \cap K_{i+1}$ is a mesh face in $\mathcal{E}_h \setminus \Gamma_N$. If $\boldsymbol{z}$ lies on Γ_D, we can still write a similar sum as long as we omit the last term (as there is no $K^\star$ for such $\boldsymbol{z}$) and choose K_m such that it has a face on $\partial\Omega$. By (3.10a), the absolute value of the ith summand inside the square brackets in (4.5) equals the magnitude of the limit of $\boldsymbol{J}_h$ as we approach $\boldsymbol{z}$ from within the mesh face $K_i \cap K_{i+1}$. Expressing each of the terms in the sum in (4.4) in terms of $\boldsymbol{J}_h$ this way, we obtain

$$\frac{C}{|K|}\|T_h(\boldsymbol{J}_h)\|_{0,K}^2 \;\le\; \sum_{\boldsymbol{z}\in\mathcal{G}_h^k\cap\partial K}\;\sum_{e\in\mathcal{F}_{\boldsymbol{z}}}\frac{1}{|e|}\|\boldsymbol{J}_h\|_{0,e}^2, \tag{4.6}$$

where $\mathcal{F}_{\boldsymbol{z}}$ denotes the set of all mesh faces e in $\mathcal{E}_h \setminus \Gamma_N$ such that $\boldsymbol{z} \in \bar{e}$. Note that in obtaining the above inequality, we have used the fact that for every $\boldsymbol{z}$, the number

m in (4.5) can be bounded uniformly in terms of the shape regularity constants. Summing over all mesh elements K, we obtain the lower bound of (4.1).

It now remains to prove (4.2). Recall that a standard norm equivalence asserts the existence of a constant C (depending on the shape regularity of K, but otherwise independent of K) such that for all $w \in \mathcal{P}_k(K)$,

$$\frac{1}{C|K|}\|w\|_{0,K}^2 \le \sum_{\boldsymbol{\alpha}\in\mathcal{A}_N^k} |w(\boldsymbol{x}_{\boldsymbol{\alpha},K})|^2 \le \frac{C}{|K|}\|w\|_{0,K}^2.$$

Applying this with $w = T_h(\boldsymbol{J}_h)|_K$, and observing that in the expansion for $T_h(\boldsymbol{J}_h)$ in (3.11), the coefficients $\{c_\phi\}$ are the nonzero values of $T_h(\boldsymbol{J}_h)$ at the points $\boldsymbol{x}_{\boldsymbol{\alpha},K}$, we obtain

$$\frac{1}{C|K|}\|T_h(\boldsymbol{J}_h)\|_{0,K}^2 \;\le \sum_{\boldsymbol{z}\in\mathcal{G}_h^k\cap\partial K}\sum_{\phi\in V_{\boldsymbol{z}}^\star} c_\phi^2 \;\le\; \frac{C}{|K|}\|T_h(\boldsymbol{J}_h)\|_{0,K}^2. \tag{4.7}$$

The upper inequality above implies

$$\begin{aligned}\sum_{\boldsymbol{z}\in\mathcal{G}_h^k\cap\partial K}\sum_{\phi\in V_{\boldsymbol{z}}^\star} c_\phi^2 \;&\le \frac{C}{|K|}\|T_h(\boldsymbol{J}_h)\|_{0,K}^2 && \text{by (4.7)}\\ &\le\; C\sum_{\boldsymbol{z}\in\mathcal{G}_h^k\cap\partial K}\sum_{e\in\mathcal{F}_{\boldsymbol{z}}}\frac{1}{|e|}\|\boldsymbol{J}_h\|_{0,e}^2 && \text{by (4.6).}\end{aligned}$$

If we sum this inequality over all mesh elements K, the resulting left-hand side dominates $\|\,[\boldsymbol{J}_h]\,\|_{\ell^2}^2$. Hence we have proven that

$$\big\|\,[\boldsymbol{J}_h]\,\big\|_{\ell^2}^2 \le\; C\sum_{e\in\mathcal{E}_h\setminus\Gamma_N}\frac{1}{|e|}\|\boldsymbol{J}_h\|_{0,e}^2. \tag{4.8}$$

Returning to (4.7) and using its lower inequality, we also have

$$\begin{aligned}\frac{1}{|e|}\|\boldsymbol{J}_h\|_{0,e}^2 &\le C\sum_{K\in\mathcal{K}_e}\frac{1}{|K|}\|T_h(\boldsymbol{J}_h)\|_{0,K}^2 && \text{by (4.3)}\\ &\le C\sum_{K\in\mathcal{K}_e}\sum_{\boldsymbol{z}\in\mathcal{G}_h^k\cap\partial K}\sum_{\phi\in V_{\boldsymbol{z}}^\star} c_\phi^2 && \text{by (4.7).}\end{aligned}$$

Summing this inequality over all edges e in $\mathcal{E}_h \setminus \Gamma_N$ and noting that the resulting number of repetitions in c_ϕ^2 can be uniformly bounded, we obtain

$$\sum_{e\in\mathcal{E}_h\setminus\Gamma_N}\frac{1}{|e|}\|\boldsymbol{J}_h\|_{0,e}^2 \le C\big\|\,[\boldsymbol{J}_h]\,\big\|_{\ell^2}^2. \tag{4.9}$$

Combining (4.9) and (4.8), the proof of (4.2) is finished. ∎

Proof of Theorem 3.2. Since the matrix M is symmetric and positive definite, its spectral condition number κ_M is given by

$$\kappa_M = \left(\max_{\boldsymbol{Y}_h\in\mathfrak{Z}_h}\frac{\langle\alpha\,\boldsymbol{Y}_h\,,\,\boldsymbol{Y}_h\rangle_{\mathcal{E}_h\setminus\Gamma_N}}{\|\,[\boldsymbol{Y}_h]\,\|_{\ell^2}^2}\right)\Big/\left(\min_{\boldsymbol{Y}_h\in\mathfrak{Z}_h}\frac{\langle\alpha\,\boldsymbol{Y}_h\,,\,\boldsymbol{Y}_h\rangle_{\mathcal{E}_h\setminus\Gamma_N}}{\|\,[\boldsymbol{Y}_h]\,\|_{\ell^2}^2}\right). \tag{4.10}$$

By the assumptions in the theorem on α,

$$\zeta_\star\Big(\sum_{e\in\mathcal{E}_h\setminus\Gamma_N}\frac{1}{|e|}\|\boldsymbol{Y}_h\|_{0,e}^2\Big)\le\langle\alpha\,\boldsymbol{Y}_h\,,\,\boldsymbol{Y}_h\rangle_{\mathcal{E}_h\setminus\Gamma_N}\le\zeta^\star\Big(\sum_{e\in\mathcal{E}_h\setminus\Gamma_N}\frac{1}{|e|}\|\boldsymbol{Y}_h\|_{0,e}^2\Big).$$

Applying (4.2) of Lemma 4.1 to the above inequality, we obtain

$$\frac{\zeta_\star}{C}\,\|\,[\boldsymbol{Y}_h]\,\|^2\le\langle\alpha\,\boldsymbol{Y}_h\,,\,\boldsymbol{Y}_h\rangle_{\mathcal{E}_h\setminus\Gamma_N}\le\;C\zeta^\star\|\,[\boldsymbol{Y}_h]\,\|^2.$$

Using this in (4.10), we find that $\kappa_M\le C^2\zeta^\star/\zeta_\star$. ∎

4.2. Proof of the flux error estimates. This subsection is devoted to proving the error estimates of Theorem 3.3. The error in the divergence is easy to analyze, but the proof of the L^2-estimate is more involved. Proceeding as in [16] in the analysis of the hybridized RT method, we start with the error equations.

Proof of Theorem 3.3. We divide this proof into seven steps.

Step 1. *Obtaining the error equations.* If we set $\boldsymbol{q}_{\nabla,h}:=-a\nabla U_h$, on each element $K\in\mathcal{T}_h$, from (2.2a) and (2.2b) defining the hybridized continuous Galerkin method and from (3.1) relating $q_{n,h}$ with the numerical trace $\widehat{\boldsymbol{q}}_h$, it follows that

$$\begin{aligned}&\big(a^{-1}\boldsymbol{q}_{\nabla,h},\,\boldsymbol{v}\big)_K-(U_h,\,\nabla\cdot\boldsymbol{v})_K=-\langle\lambda_h,\,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial K},\\&-\big(\boldsymbol{q}_{\nabla,h},\,\nabla w\big)_K+\langle\widehat{\boldsymbol{q}}_h,\,w\boldsymbol{n}\rangle_{\partial K}=(f,\,w)_K\end{aligned}$$

for any $(\boldsymbol{v},w)\in\mathcal{P}_k(K)^N\times\mathcal{P}_k(K)$. As a consequence, by the definition of $\boldsymbol{q}_h$ given by (3.2), we obtain that for $\ell\le k$,

$$\big(a^{-1}\boldsymbol{q}_h,\,\boldsymbol{v}\big)_K-(U_h,\,\nabla\cdot\boldsymbol{v})_K=-\langle\lambda_h,\,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial K}+\big(a^{-1}(\boldsymbol{q}_h-\boldsymbol{q}_{\nabla,h}),\,\boldsymbol{v}\big)_K\,,\tag{4.11a}$$

$$-\,(\boldsymbol{q}_h,\,\nabla w)_K+\langle\boldsymbol{q}_h,\,w\boldsymbol{n}\rangle_{\partial K}=(f,\,w)_K\tag{4.11b}$$

for any $(\boldsymbol{v},w)\in\mathcal{P}_k(K)^N\times\mathcal{P}_\ell(K)$. The error equations are derived by comparing these equations to the equations satisfied by the exact solution $(\boldsymbol{q},u)$, namely,

$$\begin{aligned}&\big(a^{-1}\boldsymbol{q},\,\boldsymbol{v}\big)_K-(u,\,\nabla\cdot\boldsymbol{v})_K=-\langle u,\,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial K},\\&-\,(\boldsymbol{q},\,\nabla w)_K+\langle\boldsymbol{q},\,w\boldsymbol{n}\rangle_{\partial K}=(f,\,w)_K\end{aligned}$$

for any $(\boldsymbol{v},w)\in V_{RT_\ell}(K)\times\mathcal{P}_\ell(K)$. They imply, as a consequence of the definition of the RT projection π_ℓ given in (3.12), that

$$\big(a^{-1}\pi_\ell\,\boldsymbol{q},\,\boldsymbol{v}\big)_K-(u,\,\nabla\cdot\boldsymbol{v})_K=-\langle u,\,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial K}+\big(a^{-1}(\pi_\ell\,\boldsymbol{q}-\boldsymbol{q}),\,\boldsymbol{v}\big)_K\,,\tag{4.12a}$$

$$-\,(\pi_\ell\,\boldsymbol{q},\,\nabla w)_K+\langle\pi_\ell\,\boldsymbol{q},\,w\boldsymbol{n}\rangle_{\partial K}=(f,\,w)_K\tag{4.12b}$$

for any $(\boldsymbol{v},w)\in V_{RT_\ell}(K)\times\mathcal{P}_\ell(K)$.

Thus, if we define the errors of the approximation as

$$\boldsymbol{e}_{\boldsymbol{q}}=\pi_\ell\,\boldsymbol{q}-\boldsymbol{q}_h,\quad e_u=u-U_h,\quad e_\lambda=u-\lambda_h,$$

we see, after subtracting (4.11) from (4.12), that they satisfy

$$\begin{aligned}\big(a^{-1}\boldsymbol{e}_{\boldsymbol{q}},\,\boldsymbol{v}\big)_K-(e_u,\,\nabla\cdot\boldsymbol{v})_K=&-\langle e_\lambda,\,\boldsymbol{v}\cdot\boldsymbol{n}\rangle_{\partial K}\\&-\big(a^{-1}(\boldsymbol{q}_h-\boldsymbol{q}_{\nabla,h}),\,\boldsymbol{v}\big)_K\\&+\big(a^{-1}(\pi_\ell\,\boldsymbol{q}-\boldsymbol{q}),\,\boldsymbol{v}\big)_K\,,\end{aligned}\tag{4.13a}$$

$$-(\boldsymbol{e}_{\boldsymbol{q}},\,\nabla w)_K+\langle\boldsymbol{e}_{\boldsymbol{q}}\cdot\boldsymbol{n},\,w\rangle_{\partial K}=0\tag{4.13b}$$

for any $(\boldsymbol{v}, w) \in (V_{RT_\ell}(K) \cap \mathcal{P}_k(K)^N) \times \mathcal{P}_\ell(K)$ for $\ell \leq k$.

Step 2. *Analyzing errors in the divergence of the flux.* Integrating (4.13b) by parts, we obtain

$$(\nabla \cdot \boldsymbol{e_q},\, w)_K = 0 \quad \text{for all } w \in \mathcal{P}_\ell(K).$$

Since $\nabla \cdot \boldsymbol{e_q} \in \mathcal{P}_\ell(K)$, we immediately get that

$$\nabla \cdot \boldsymbol{e_q} \equiv 0 \quad \text{on } K,$$

which is the first identity of Theorem 3.3. (It is obvious that $\boldsymbol{e_q}$ is in $\boldsymbol{H}(\mathrm{div}, \Omega)$.)

The first inequality of Theorem 3.3 follows from the fact that

$$\nabla \cdot (\boldsymbol{q} - \boldsymbol{q}_h) = \nabla \cdot (\boldsymbol{q} - \pi_\ell\, \boldsymbol{q}) = (\mathrm{Id} - \mathsf{P}_\ell)\, f,$$

where P_ℓ is the L^2-projection into the space of functions w such that $w_K \in \mathcal{P}_\ell(K)$ for all $K \in \mathcal{T}_h$. Notice that in the last step, we used the commutativity property $\nabla \cdot \pi_\ell = \mathrm{P}_\ell \nabla \cdot$ (see, e.g., [6, 16, 22]).

Step 3. *Establishing the second identity.* If in the error equation (4.13a) we select $\boldsymbol{v} \in \mathcal{P}_{\ell-1}^N(K)$, we find that whenever $a(\boldsymbol{x})$ is constant on K,

$$\left(a^{-1}\boldsymbol{e_q},\, \boldsymbol{v}\right)_K - (e_u,\, \nabla \cdot \boldsymbol{v})_K = -\langle e_\lambda,\, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\partial K},$$

where we used (3.2c) of the definition of $\boldsymbol{q}_h$ and (3.12c) of the definition of π_ℓ. This readily implies that

$$\left(a^{-1}\boldsymbol{e_q},\, \boldsymbol{v}\right)_\Omega - \sum_{K \in \mathcal{T}_h} (e_u,\, \nabla \cdot \boldsymbol{v})_K = - \sum_{e \in \mathcal{E}_h} \langle e_\lambda,\, [\![\boldsymbol{v} \cdot \boldsymbol{n}]\!] \rangle_e,$$

and so

$$\left(a^{-1}\boldsymbol{e_q},\, \boldsymbol{v}\right)_\Omega = -\langle e_\lambda\,,\, \boldsymbol{v} \cdot \boldsymbol{n} \rangle_{\Gamma_D} = 0 \quad \text{for all } \boldsymbol{v} \in \boldsymbol{\mathcal{V}}_{h,\ell-1}^0,$$

since we are assuming that $g|_e \in \mathcal{P}_k(e)$ for each face e on Γ_D. The second identity of Theorem 3.3 immediately follows from this and the definition of the projection $\mathbb{P}_{\ell-1}$ given by (3.13).

Step 4. *Splitting errors in the flux.* It remains to prove the second inequality of Theorem 3.3. To do this, we begin by noting that since $\boldsymbol{e_q} \in \boldsymbol{\mathcal{V}}_{h,\ell}^0$, we can choose $\boldsymbol{v} = \boldsymbol{e_q}$ in the error equation (4.13a). Doing this and summing over all the elements $K \in \mathcal{T}_h$, we obtain

$$(a^{-1}e_{\boldsymbol{q}}, e_{\boldsymbol{q}})_\Omega = -\left(a^{-1}(\boldsymbol{q}_h - \boldsymbol{q}_{\nabla,h}),\, \boldsymbol{e_q}\right)_\Omega + \left(a^{-1}(\pi_\ell\, \boldsymbol{q} - \boldsymbol{q}),\, \boldsymbol{e_q}\right)_\Omega.$$

Introducing $\pi_\ell \boldsymbol{q}_{\nabla,h}$ into the right-hand side,

$$\begin{aligned}(a^{-1}e_{\boldsymbol{q}}, e_{\boldsymbol{q}})_\Omega &= -\left(a^{-1}(\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h}),\, \boldsymbol{e_q}\right)_\Omega - \left(a^{-1}(\pi_\ell \boldsymbol{q}_{\nabla,h} - \boldsymbol{q}_{\nabla,h}),\, \boldsymbol{e_q}\right)_\Omega \\ &\quad + \left(a^{-1}(\pi_\ell\, \boldsymbol{q} - \boldsymbol{q}),\, \boldsymbol{e_q}\right)_\Omega \\ &= -\left(a^{-1}(\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h}),\, \boldsymbol{e_q}\right)_\Omega - \left(a^{-1}(Id - \pi_\ell)(\boldsymbol{q} - \boldsymbol{q}_{\nabla,h}),\, \boldsymbol{e_q}\right)_\Omega.\end{aligned}$$

Applying the Cauchy–Schwarz inequality,

$$C\|\boldsymbol{e_q}\|_{0,\Omega} \leq \|\boldsymbol{q}_h - \pi_\ell\, \boldsymbol{q}_{\nabla,h}\|_{0,\Omega} + \|(\mathrm{Id} - \pi_\ell)(\boldsymbol{q} - \boldsymbol{q}_{\nabla,\mathrm{h}})\|_{0,\Omega}. \tag{4.14}$$

We now estimate each of the terms on the right-hand side separately in the next two steps.

Step 5. *Estimating* $\|\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h}\|_{0,\Omega}$. In order to estimate this term, we rewrite (3.2b) and (3.2c) defining $\boldsymbol{q}_h$ as

$$\begin{aligned}
&(\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h},\, \boldsymbol{v})_K = 0 && \text{for all } \boldsymbol{v} \in \mathcal{P}_{\ell-1}(K)^N,\\
&\langle (\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h})\cdot \boldsymbol{n},\, w\rangle_e = \langle (\widehat{\boldsymbol{q}}_h - \boldsymbol{q}_{\nabla,h})\cdot \boldsymbol{n},\, w\rangle_e && \text{for all } w \in \mathcal{P}_\ell(e) \text{ and all faces } e \subset \partial K.
\end{aligned}$$

Then a standard scaling argument gives

$$\|\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h}\|_{0,K}^2 \le C h_K \|(\widehat{\boldsymbol{q}}_h - \boldsymbol{q}_{\nabla,h})\cdot \boldsymbol{n}\|_{0,\partial K}^2.$$

Summing over all mesh elements and using the definition of the numerical trace $\widehat{\boldsymbol{q}}_h$ given (3.5), we obtain

$$\begin{aligned}
\|\boldsymbol{q}_h - \pi_\ell \boldsymbol{q}_{\nabla,h}\|_{0,\Omega}^2 \le C \sum_{K\in\mathcal{T}_h} h_K \Bigg(&\left\| \left(\boldsymbol{\beta}\cdot\boldsymbol{n} - \frac{1}{2}\right) [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!] + \alpha\, \boldsymbol{J}_h\cdot\boldsymbol{n} \right\|_{0,\partial K\setminus\,\partial\Omega}^2 \\
&+ \left\|\alpha\, \boldsymbol{J}_h\cdot\boldsymbol{n}\right\|_{0,\partial K\cap\Gamma_D}^2 + \left\|\mathsf{q}_{\mathsf{N}} - \boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}\right\|_{0,\partial K\cap\Gamma_N}^2 \Bigg)\\
&\le C\, h\, (T_1 + T_2), \qquad (4.15)
\end{aligned}$$

where

$$T_1 := \left\| [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n} - \pi_\ell \boldsymbol{q}\cdot\boldsymbol{n}]\!] \right\|_{0,\mathcal{E}_h\setminus\Gamma_D}^2 \quad \text{and} \quad T_2 := \|\alpha\, \boldsymbol{J}_h\cdot\boldsymbol{n}\|_{0,\mathcal{E}_h\setminus\Gamma_N}^2.$$

The term T_1 can be easily estimated by an inverse inequality:

$$T_1 \le C h^{-1} \|\boldsymbol{q}_{\nabla,h} - \pi_\ell \boldsymbol{q}\|_{0,\Omega}^2 \le C h^{-1} \big(|u - U_h|_{1,\Omega}^2 + \|\boldsymbol{q} - \pi_\ell \boldsymbol{q}\|_{0,\Omega}^2\big). \qquad (4.16)$$

The other term T_2 requires more work.

To estimate T_2, we rewrite the definition of the jump $\boldsymbol{J}_h$, namely, (3.8), as

$$\begin{aligned}
\langle \alpha\, \boldsymbol{J}_h,\, [\![v\boldsymbol{n}]\!]\rangle_{\mathcal{E}_h\setminus\Gamma_N} &= \big\langle [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!],\, \{\!\{v\}\!\} - \boldsymbol{\beta}\cdot [\![v\boldsymbol{n}]\!]\big\rangle_{\mathcal{E}_h\setminus\partial\Omega}\\
&\quad + \sum_{K\in\mathcal{T}_h} (f - \nabla\cdot\boldsymbol{q}_{\nabla,h}, v)_K\\
&= \big\langle [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!],\, \{\!\{v\}\!\} - \boldsymbol{\beta}\cdot [\![v\boldsymbol{n}]\!]\big\rangle_{\mathcal{E}_h\setminus\partial\Omega}\\
&\quad + \sum_{K\in\mathcal{T}_h} (\nabla\cdot(\pi_k\boldsymbol{q} - \boldsymbol{q}_{\nabla,h}), v)_K
\end{aligned}$$

for all $v \in V_h$. Choosing $v = T_h(\boldsymbol{J}_h)$ and using the property (3.10a) of the operator T_h, we get

$$\begin{aligned}
\langle \alpha\, \boldsymbol{J}_h\,,\, \boldsymbol{J}_h\rangle_{\mathcal{E}_h\setminus\Gamma_N} = \langle [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!],\, \{\!\{T_h(\boldsymbol{J}_h)\}\!\}\rangle_{\mathcal{E}_h^\circ} - \langle [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!],\, \boldsymbol{\beta}\cdot\boldsymbol{J}_h)\rangle_{\mathcal{E}_h^\circ}\\
+ \sum_{K\in\mathcal{T}_h} (\nabla\cdot(\pi_k\boldsymbol{q} - \boldsymbol{q}_{\nabla,h}),\, T_h(\boldsymbol{J}_h))_K.
\end{aligned}$$

Using (4.1) of Lemma 4.1 after applying suitable inverse inequalities, we obtain

$$\langle \alpha\, \boldsymbol{J}_h\,,\, \boldsymbol{J}_h\rangle_{\mathcal{E}_h\setminus\Gamma_N} \le C \left(\left\| [\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!] \right\|_{0,\mathcal{E}_h^\circ} + h^{1/2} \|\nabla\cdot(\pi_k\boldsymbol{q} - \boldsymbol{q}_{\nabla,h})\|_{0,\Omega} \right) \|\boldsymbol{J}_h\|_{0,\mathcal{E}_h\setminus\Gamma_N}.$$

This implies that

$$T_2 \leq C\,\frac{\alpha^\star}{\alpha_\star}\,\Big(\big\|\,[\![\boldsymbol{q}_{\nabla,h}\cdot\boldsymbol{n}]\!]\,\big\|^2_{0,\mathcal{E}_h^\circ} + h\,\|\nabla\cdot(\pi_k\boldsymbol{q}-\boldsymbol{q}_{\nabla,h})\|^2_{0,\Omega}\Big).$$

Treating the first term inside the parentheses above as in the proof of (4.16), and applying an inverse inequality to the second, we get

$$T_2 \leq C\,h^{-1}\big(|u-U_h|^2_{1,\Omega} + \|\boldsymbol{q}-\pi_\ell\boldsymbol{q}\|^2_{0,\Omega} + \|\boldsymbol{q}-\pi_k\boldsymbol{q}\|^2_{0,\Omega}\big).$$

Using the estimates for T_1 and T_2 in (4.15), we conclude that

$$C\,\|\boldsymbol{q}_h-\pi_\ell\boldsymbol{q}_{\nabla,h}\|^2_{0,\Omega} \leq |u-U_h|^2_{1,\Omega} + \|\boldsymbol{q}-\pi_\ell\boldsymbol{q}\|^2_{0,\Omega} + \|\boldsymbol{q}-\pi_k\boldsymbol{q}\|^2_{0,\Omega}. \tag{4.17}$$

Step 6. *Estimating* $\|(Id-\pi_\ell)(\boldsymbol{q}-\boldsymbol{q}_{\nabla,h})\|_{0,\Omega}$. On an element K, using the well-known approximation property of π_ℓ [6, 29], we obtain

$$\begin{aligned}\|(Id-\pi_\ell)(\boldsymbol{q}-\boldsymbol{q}_{\nabla,h})\|_{0,K} &\leq Ch_K|\boldsymbol{q}-\boldsymbol{q}_{\nabla,h}|_{1,K}\\ &\leq Ch_K\|a\|_{W^1_\infty(K)}\|\nabla u-\nabla U_h\|_{1,K}.\end{aligned}$$

Now using any projector Π_K well defined on $L^2(K)^N$ with standard approximation properties, e.g., the one constructed in [30], we have

$$\begin{aligned}\|(Id-\pi_\ell)(\boldsymbol{q}-\boldsymbol{q}_{\nabla,h})\|_{0,K} &\leq Ch_K\big(\|\nabla u-\Pi_K\nabla u\|_{1,K} + \|\Pi_K(\nabla u-\nabla U_h)\|_{1,K}\big)\\ &\leq C\big(h_K\|\nabla u-\Pi_K\nabla u\|_{1,K} + \|\Pi_K(\nabla u-\nabla U_h)\|_{0,K}\big)\\ &\leq C\big(h_K\|\nabla u-\Pi_K\nabla u\|_{1,K} + \|\nabla u-\nabla U_h\|_{0,K}\big).\end{aligned}$$

Thus, we obtain

$$\|(\mathrm{Id}-\pi_\ell)(\boldsymbol{q}-\boldsymbol{q}_{\nabla,\mathrm{h}})\|_{0,\Omega} \leq \mathrm{Ch}^{\min(\mathrm{s},\mathrm{k})}|\mathrm{u}|_{\mathrm{s}+1,\Omega}. \tag{4.18}$$

Step 7. *Completing the proof of Theorem* 3.3. Now we use the results of the previous two steps, namely, (4.17) and (4.18), in the splitting (4.14) of the error term. Then we obtain

$$C\|\boldsymbol{e}_{\boldsymbol{q}}\|_{0,\Omega} \leq |u-U_h|_{1,\Omega} + \|\boldsymbol{q}-\pi_\ell\,\boldsymbol{q}\|_{0,\Omega} + \|\boldsymbol{q}-\pi_k\boldsymbol{q}\|_{0,\Omega} + h^{\min(s,k)}|u|_{s+1,\Omega},$$

and the estimate of $\boldsymbol{q}_h$ immediately follows from the standard approximation results:

$$\begin{aligned}|u-U_h|_{1,\Omega} &\leq Ch^{\min\{k,s\}}|u|_{s+1,\Omega},\\ \|\boldsymbol{q}-\pi_m\boldsymbol{q}\|_{0,\Omega} &\leq Ch^{\min\{m+1,s\}}|\boldsymbol{q}|_{s,\Omega}.\end{aligned}$$

This concludes the proof. ∎

5. Numerical results. In this section, we carry out some numerical experiments to verify the theoretical results when the exact solution is smooth (Test 1) and to test the performance of the method when the exact solution has a singularity (Test 2). For the sake of simplicity, we use uniform meshes and pick $\boldsymbol{\beta}=\mathbf{0}$ in the definition of the numerical trace $\widehat{\boldsymbol{q}}_h$, (3.5).

In what follows, by the approximation given by the "RT$_\ell$ method" we mean the pair $(\boldsymbol{q}_h, U_h)$ obtained as follows. The function $(\boldsymbol{q}_h, u_h, \lambda_h)$ is the solution of the hybridized RT method whose Lagrange multipliers are piecewise polynomials of

TABLE 3
Comparison of the history of convergence of the RT_0 and the postprocessed CG_1 methods.

Grid	$\|e_{U_h}\|_0$		$\|e_{\mathrm{div}\mathbf{q}_h}\|_0$		$\|\boldsymbol{e}_{\mathbf{q}_h}\|_0$		$\|e_{a\nabla U_h}\|_0$	
level	Error	Order	Error	Order	Error	Order	Error	Order
RT_0 method								
1	.42e+0	–	.18e+2	–	.35e+1	–	.44e+1	–
2	.12e+0	1.77	.11e+2	0.69	.27e+1	0.39	.29e+1	0.59
3	.35e-1	1.81	.58e+1	0.95	.14e+1	0.96	.15e+1	0.95
4	.90e-2	1.95	.30e+1	0.99	.69e+0	0.99	.76e+0	0.99
5	.23e-2	1.99	.15e+1	1.00	.35e+0	1.00	.38e+0	1.00
6	.57e-3	2.00	.74e+0	1.00	.17e+0	1.00	.19e+0	1.00
7	.14e-3	2.00	.37e+0	1.00	.87e-1	1.00	.95e-1	1.00
RT_0-postprocessed CG_1 method								
1	.50e+0	–	.18e+2	–	.38e+1	–	.61e+1	–
2	.27e+0	0.89	.11e+2	0.69	.30e+1	0.33	.42e+1	0.54
3	.94e-1	1.52	.58e+1	0.95	.16e+1	0.96	.25e+1	0.76
4	.26e-1	1.83	.30e+1	0.99	.74e+0	1.09	.13e+0	0.92
5	.69e-2	1.95	.15e+1	1.00	.36e+0	1.05	.66e+0	0.98
6	.17e-2	1.98	.74e+0	1.00	.18e+0	1.02	.33e+0	0.99
7	.43e-3	2.00	.37e+0	1.00	.87e-1	1.01	.17e+0	1.00
RT_1-postprocessed CG_1 method								
1	.50e+0	–	.90e+1	–	.83e+1	–	.61e+1	–
2	.27e+0	0.89	.28e+1	1.70	.60e+1	0.47	.42e+1	0.54
3	.94e-1	1.52	.73e+0	1.93	.37e+1	0.67	.25e+1	0.76
4	.26e-1	1.83	.19e+0	1.98	.20e+1	0.91	.13e+1	0.92
5	.69e-2	1.95	.46e-1	2.00	.10e+1	0.98	.66e+0	0.98
6	.17e-2	1.98	.12e-1	2.00	.51e+0	1.00	.33e+0	0.99
7	.43e-3	2.00	.29e-2	2.00	.25e+0	1.00	.17e+0	1.00

degree ℓ. The function U_h is obtained from (u_h, λ_h) by using the local postprocessing described in [1]. The resulting pair $(\boldsymbol{q}_h, U_h)$ is then compared to the solution of our RT_ℓ-postprocessed CG_k method, for which $\boldsymbol{q}_h$ is the RT_ℓ-postprocessed CG_k flux and U_h is the solution of the CG method with piecewise polynomials of degree k.

Test 1. We take

$$a = \begin{pmatrix} x+2 & x+y \\ x+y & y+2 \end{pmatrix}$$

and then g and f so that the exact solution is

$$u(x,y) = \sin(\pi x)\,\sin(\pi y).$$

The history of convergence of the approximations given by the RT_{k-1} and the RT_ℓ-postprocessed CG_k methods, for $\ell \in \{k-1, k\}$, are displayed in Tables 3, 4, and 5 for $k = 1, 2,$ and 3, respectively. Plots of these results are also displayed in Figure 6 for an easier comparison.

We see that the approximation given by the RT_ℓ-postprocessed CG_k method converges with the orders predicted by Theorem 3.3. Observe that the errors of the

TABLE 4
Comparison of the history of convergence of the RT_1 and the postprocessed CG_2 methods.

Grid	$\|e_{U_h}\|_0$		$\|e_{\mathrm{div}\mathbf{q}_h}\|_0$		$\|\boldsymbol{e}_{\mathbf{q}_h}\|_0$		$\|e_{a\nabla U_h}\|_0$	
level	Error	Order	Error	Order	Error	Order	Error	Order
RT_1-method								
1	.16e+0	–	.90e+1	–	0.23e+1	–	.42e+1	–
2	.28e-1	2.49	.28e+1	1.70	0.50e+0	2.17	.12e+1	1.87
3	.37e-2	2.94	.73e+0	1.93	0.13e+0	1.94	.30e+0	1.95
4	.47e-3	2.96	.19e+0	1.98	0.33e-1	1.98	.76e-1	1.99
5	.60e-4	2.98	.46e-1	2.00	0.83e-2	1.99	.19e-1	2.00
6	.75e-5	2.99	.12e-1	2.00	0.21e-2	2.00	.47e-2	2.00
7	.94e-6	3.00	.29e-2	2.00	0.52e-3	2.00	.12e-2	2.00
RT_1-postprocessed CG_2 method								
1	.24e+0	–	.90e+1	–	.25e+1	–	.40e+1	–
2	.35e-1	2.77	.28e+1	1.70	.76e+0	1.73	.13e+1	1.64
3	.46e-2	2.93	.73e+0	1.93	.16e+0	2.25	.37e+0	1.79
4	.56e-3	3.04	.19e+0	1.98	.35e-1	2.17	.97e-1	1.93
5	.69e-4	3.02	.46e-1	2.00	.85e-2	2.07	.25e-1	1.98
6	.86e-5	3.01	.12e-1	2.00	.21e-2	2.02	.62e-2	1.99
7	.11e-5	3.00	.29e-2	2.00	.52e-3	2.00	.15e-2	2.00
RT_2-postprocessed CG_2 method								
1	.24e+0	–	.29e+1	–	.66e+1	–	.40e+1	–
2	.35e-1	2.77	.54e+0	2.40	.21e+1	1.65	.13e+1	1.64
3	.46e-2	2.93	.72e-1	2.92	.67e+0	1.64	.37e+0	1.79
4	.56e-3	3.04	.91e-2	2.98	.18e+0	1.89	.97e-1	1.93
5	.69e-4	3.02	.11e-2	2.99	.46e-1	1.97	.25e-1	1.98
6	.86e-5	3.01	.14e-3	3.00	.12e-1	1.99	.62e-2	1.99
7	.11e-5	3.00	.18e-4	3.00	.29e-2	2.00	.15e-2	2.00

divergence between the RT_{k-1} and the RT_{k-1}-postprocessed CG_k methods are exactly the same, as predicted by the theory. Moreover, as can be clearly seen from Figure 6, the approximations of the RT_{k-1} and the RT_{k-1}-postprocessed CG_k methods are comparable in accuracy. We also see that the approximate flux provided by the RT_{k-1}-postprocessed CG_k is better than the approximation $-a\nabla U_h$ provided by the CG_k method. Finally, note that if we increase ℓ by one more degree than $k-1$ in RT_ℓ-postprocessing, there is no improvement—in fact, the approximate flux given by the RT_k-postprocessed CG_k method produces an approximate flux that is worse than that provided by the RT_{k-1}-postprocessed CG_k method.

Test 2. Now we work on a problem in which the solution has singularities produced by drastic changes in the permeability a; see Figure 1 in the introduction. We compare the streamlines of the approximate flux obtained by the RT_{k-1}-postprocessed CG_k method and that of the RT_{k-1} method around the upper left corner of the rock in Figures 7, 8, and 9. We see that the presence of the singularity at the corner induces small distortions in the streamlines. However, even in this hard case, the flux produced by the RT_{k-1} method and the solution given by the RT_{k-1}-postprocessed CG_k method are remarkably similar.

TABLE 5
Comparison of the history of convergence of the RT_2 and the postprocessed CG_3 methods.

Grid	$\|e_{U_h}\|_0$		$\|e_{\mathrm{div}\mathbf{q}_h}\|_0$		$\|\boldsymbol{e}_{\mathbf{q}_h}\|_0$		$\|e_{a\nabla U_h}\|_0$	
level	Error	Order	Error	Order	Error	Order	Error	Order
RT_2-method								
1	.12e+0	–	.29e+1	–	0.47e+0	–	.35e+1	–
2	.85e-2	3.84	.54e+0	2.40	0.85e-1	2.46	.50e+0	2.78
3	.60e-3	3.84	.72e-1	2.92	0.11e-1	2.92	.72e-1	2.81
4	.39e-4	3.94	.91e-2	2.98	0.14e-2	2.98	.94e-2	2.94
5	.25e-5	3.98	.11e-2	2.99	0.18e-3	2.99	.12e-2	2.98
6	.16e-6	3.99	.14e-3	3.00	0.22e-4	3.00	.15e-3	2.99
7	.97e-8	4.00	.18e-4	3.00	0.28e-5	3.00	.19e-4	3.00
RT_2-postprocessed CG_3 method								
1	.96e-1	–	.29e+1	–	.14e+1	–	.20e+1	–
2	.63e-2	3.93	.54e+0	2.40	.20e+0	2.82	.27e+0	2.90
3	.35e-3	4.15	.72e-1	2.92	.21e-1	3.23	.35e-1	2.94
4	.20e-4	4.12	.91e-2	2.98	.22e-2	3.29	.43e-2	3.00
5	.12e-5	4.06	.11e-2	2.99	.23e-3	3.23	.54e-3	3.01
6	.75e-7	4.02	.14e-3	3.00	.26e-4	3.15	.67e-4	3.01
7	.47e-8	4.01	.18e-4	3.00	.31e-5	3.09	.83e-5	3.00
RT_3-postprocessed CG_3 method								
1	.96e-1	–	.16e+1	–	.31e+1	–	.20e+1	–
2	.63e-2	3.93	.90e-1	4.12	.48e+0	2.71	.27e+0	2.90
3	.35e-3	4.15	.59e-2	3.93	.63e-1	2.93	.35e-1	2.94
4	.20e-4	4.12	.37e-3	3.98	.79e-2	3.00	.43e-2	3.00
5	.12e-5	4.06	.24e-4	4.99	.98e-3	3.01	.54e-3	3.01
6	.75e-7	4.02	.15e-5	4.00	.12e-3	3.01	.67e-4	3.01
7	.47e-8	4.01	.93e-7	3.98	.15e-4	3.00	.83e-5	3.00

6. Concluding remarks. We have shown that a new postprocessing of the CG_k solution gives rise to an $H(\mathrm{div})$-conforming approximation to the flux which renders the CG method locally conservative. The postprocessing belongs to the RT space of degree $k-1$ and displays convergence properties similar to the approximation given by the RT method of degree $k-1$ itself. By counting the degrees of freedom we have established that the computational effort needed to obtain the new postprocessed flux is less than that of the RT method.

We have also shown how to hybridize the CG method, making it easier to treat variable degree approximation spaces and hanging nodes.

The study of the effect of the numerical trace parameter $\boldsymbol{\beta}$ on the quality of the approximation and the extension of this approach to linear elasticity are subjects of ongoing research.

Acknowledgments. The authors would like to thank the reviewers for bringing to their attention the papers [7] and [26]. They would also like to thank Clint Dawson and Graham F. Carey for bringing to their attention the papers [27, 32] and [9, 28, 10, 14, 11, 8], respectively.

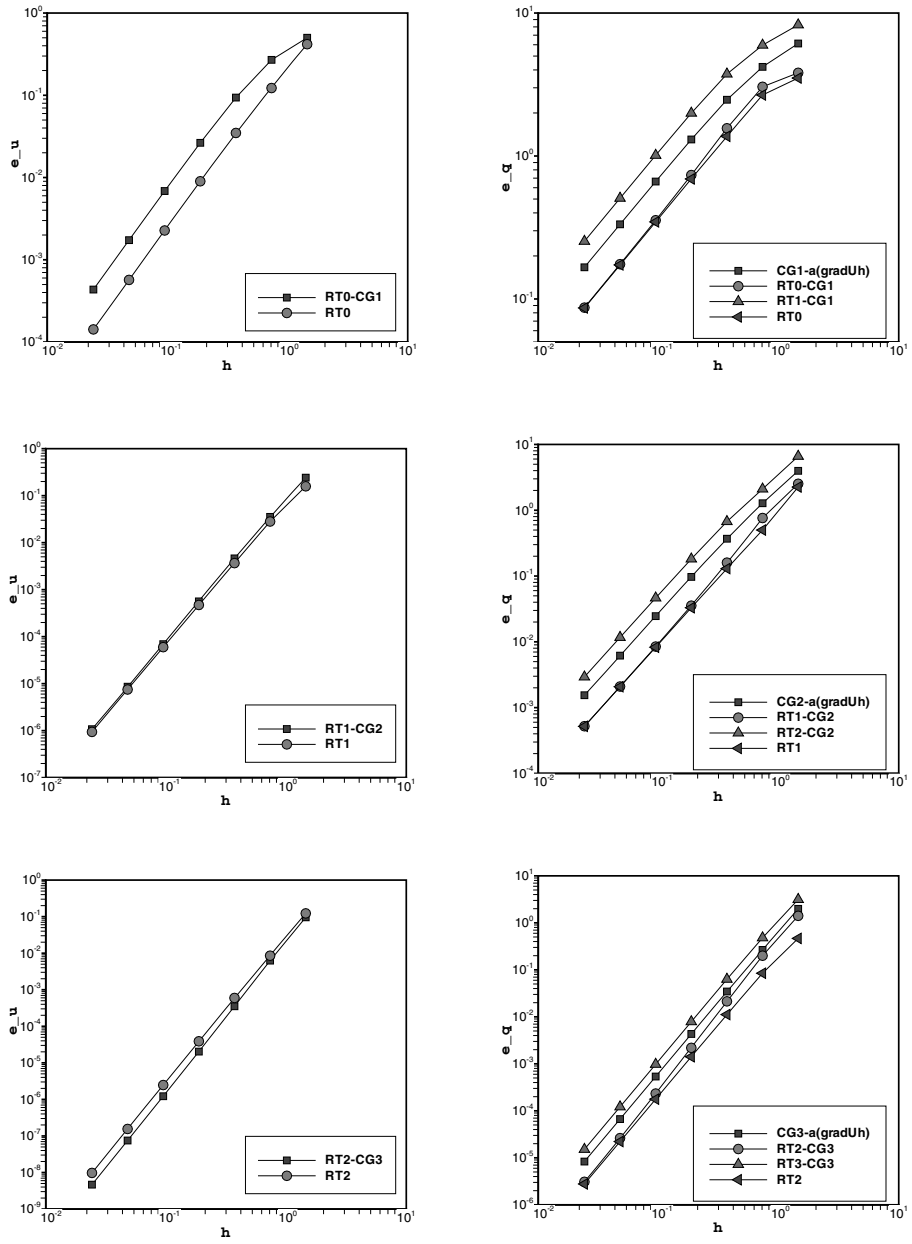


FIG. 6. *History of convergence for u_h (left) and $\boldsymbol{q}_h$ (right) for $k = 1$ (first row), $k = 2$ (second row), and $k = 3$ (third row). Note:* CGk-a(gradUh) *is $a\nabla U_h$ from CG_k, $k = 1, 2,$ or 3.*

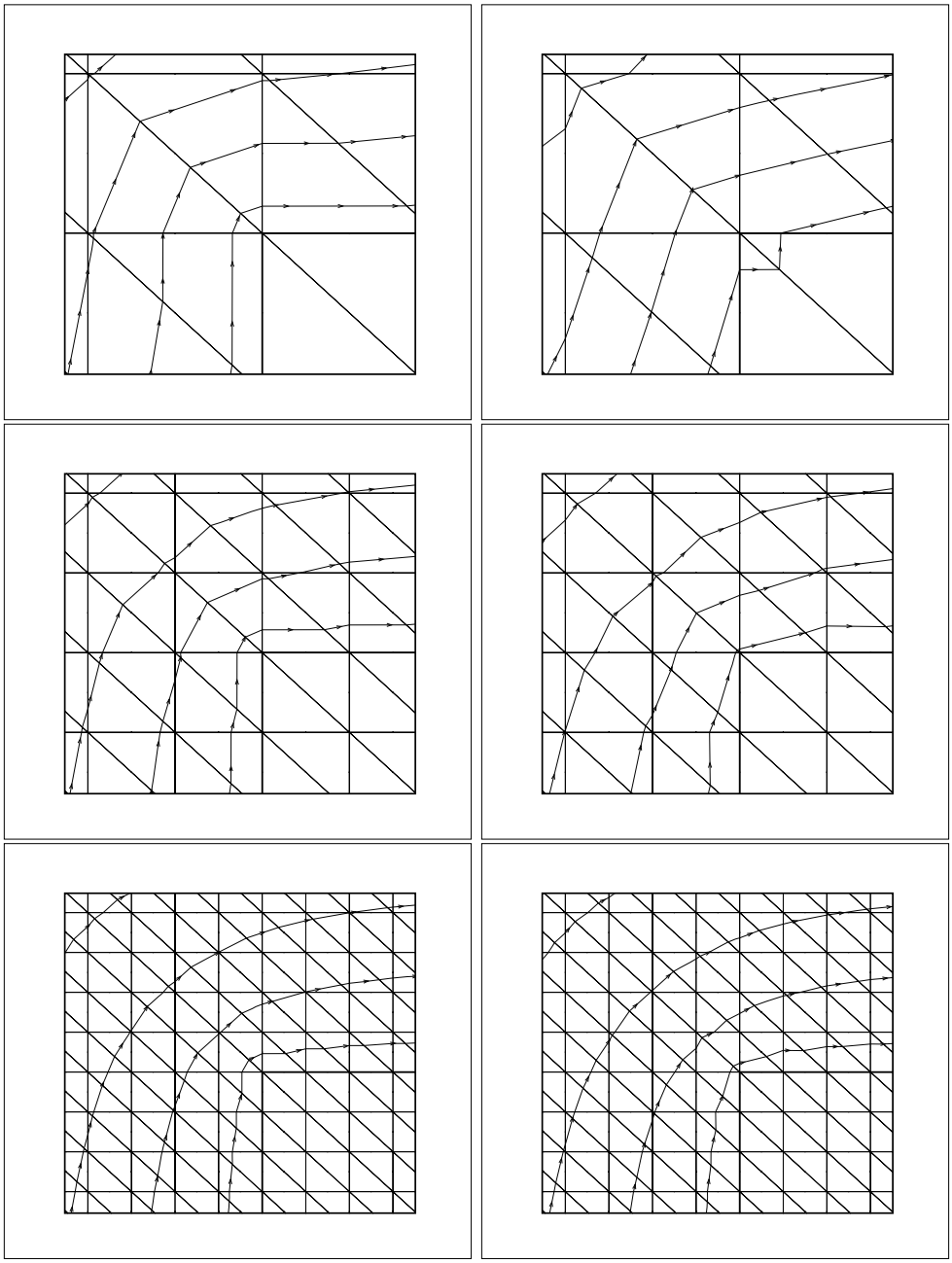


FIG. 7. *Streamlines in the upper left corner with $k = 1$. On the left column is the solution given by the RT_k method and on the right column that of the $RT_{(k-1)}$-postprocessed CG_k method. From top to bottom, mesh size $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$.*

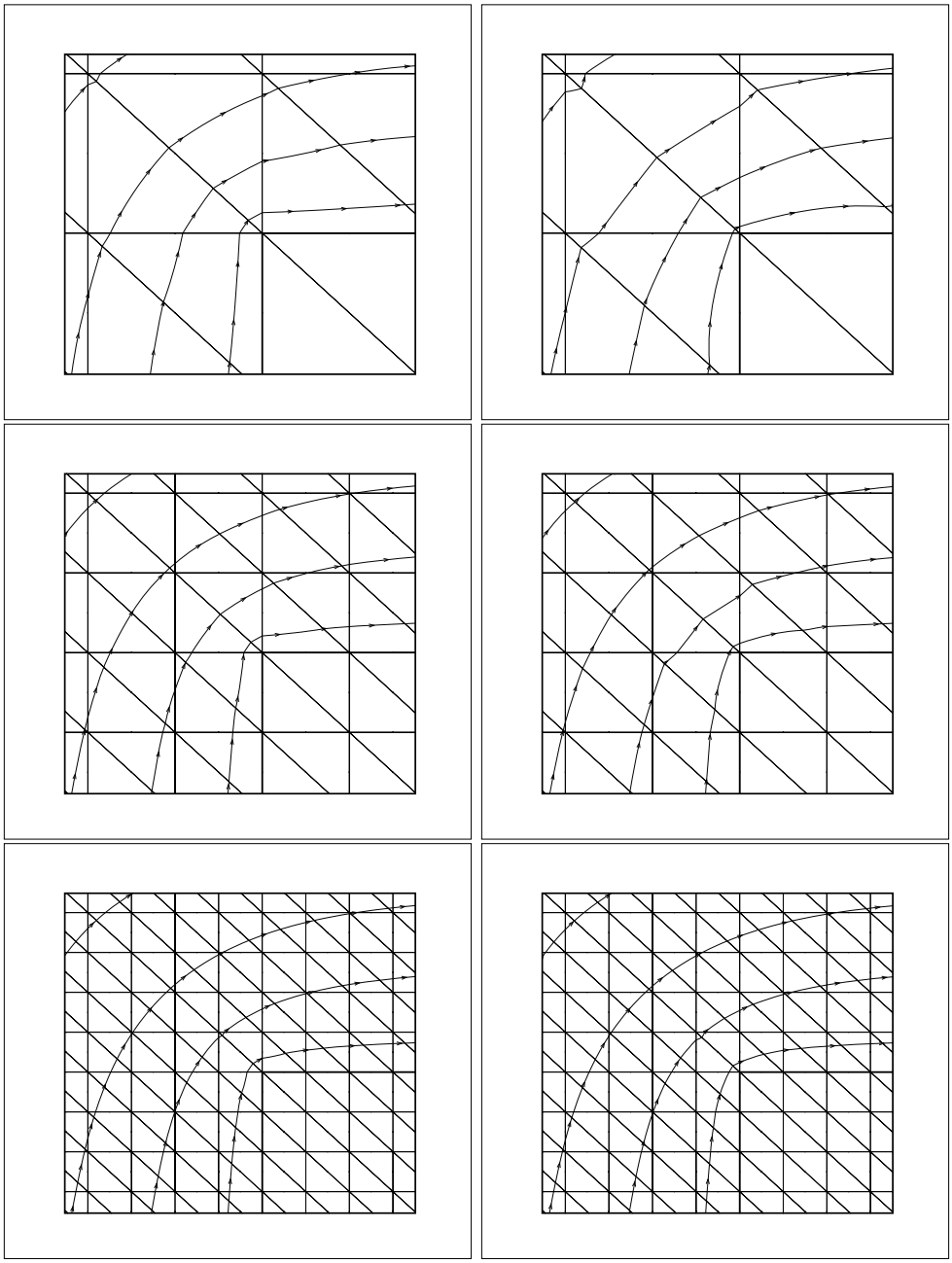


FIG. 8. *Streamlines in the upper left corner with* $k = 2$. *On the left column is the solution given by the* RT_k *method and on the right column that of the* $RT_{(k-1)}$*-postprocessed* CG_k *method. From top to bottom, mesh size* $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$.

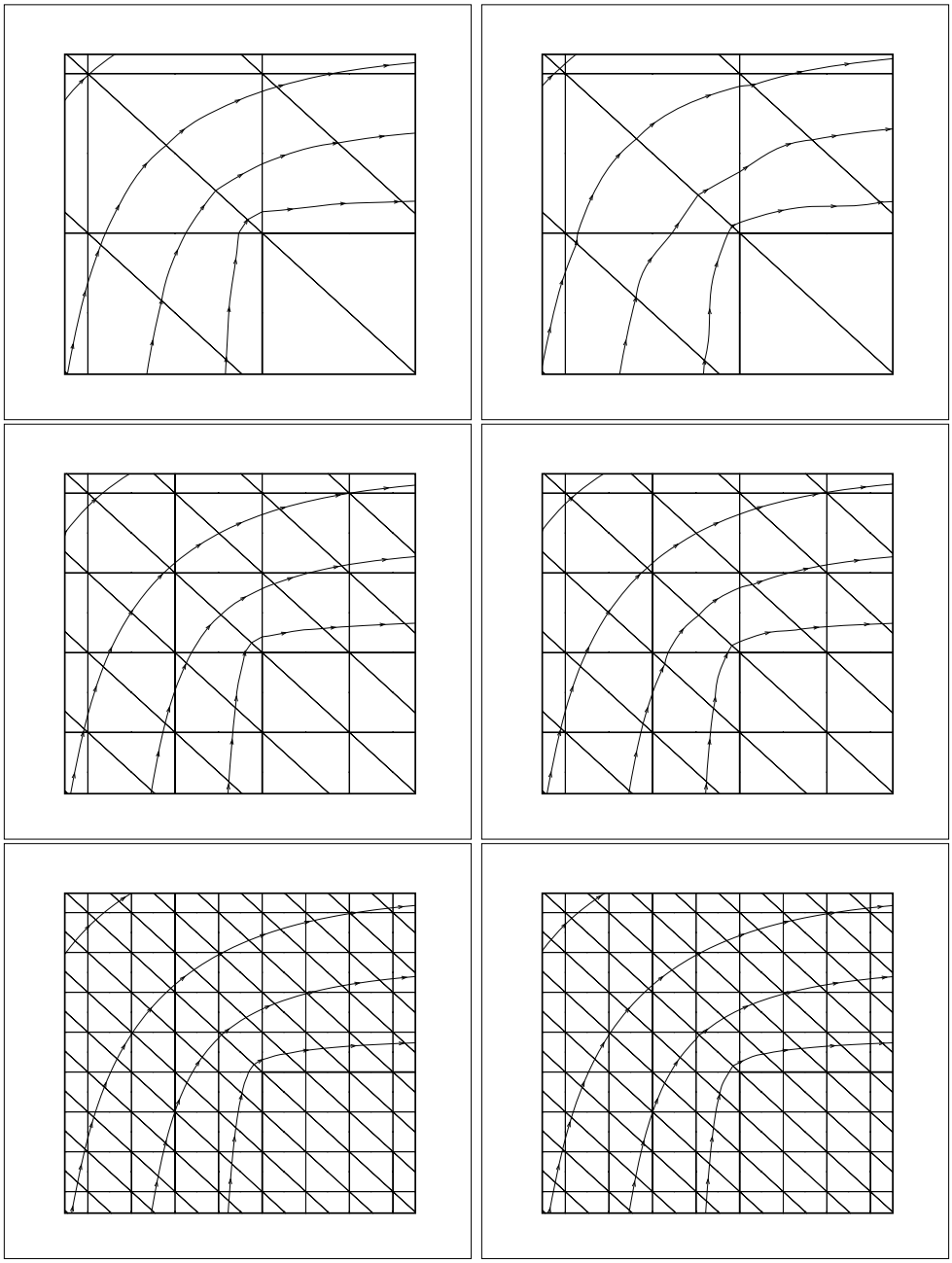


FIG. 9. *Streamlines in the upper left corner with $k = 3$. On the left column is the solution given by the RT_k method and on the right column that of the $RT_{(k-1)}$-postprocessed CG_k method. From top to bottom, mesh size $h = \frac{1}{8}, \frac{1}{16}, \frac{1}{32}$.*

REFERENCES

[1] D. N. Arnold and F. Brezzi, *Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates*, RAIRO Modél. Math. Anal. Numér., 19 (1985), pp. 7–32.

[2] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[3] J. Barrett and C. Elliott, *Total flux estimates for a finite-element approximation of elliptic equations*, IMA J. Numer. Anal., 7 (1987), pp. 129–148.

[4] P. Bastian and B. Rivière, *Superconvergence and $H(\mathrm{div})$ projection for discontinuous Galerkin methods*, Internat. J. Numer. Methods Fluids, 42 (2003), pp. 1043–1057.

[5] J. H. Bramble, J. E. Pasciak, and A. H. Schatz, *The construction of preconditioners for elliptic problems by substructuring.* I, Math. Comp., 47 (1986).

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

[7] F. Brezzi, T. J. R. Hughes, and E. Süli, *Variational approximation of flux in conforming finite element methods for elliptic partial differential equations: A model problem*, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 12 (2001), pp. 159–166.

[8] G. Carey, *Derivative calculation from finite element solutions*, Comput. Methods Appl. Mech. Engrg., 35 (1982), pp. 1–14.

[9] G. Carey, *Some further properties of the superconvergent flux projection*, Comput. Methods Appl. Mech. Engrg., 18 (2002), pp. 241–250.

[10] G. F. Carey, G. Bicken, V. Carey, C. Berger, and J. Sanchez, *Locally constrained projections on grids*, Internat. J. Numer. Methods Engrg., 50 (2001), pp. 549–577.

[11] G. F. Carey, S.-S. Chow, and M. K. Seager, *Approximate boundary-flux calculations*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 107–120.

[12] F. Celiker and B. Cockburn, *Superconvergence of the numerical traces of discontinuous Galerkin and hybridized mixed methods for convection-diffusion problems in one space dimension*, Math. Comp., 67 (2007), pp. 67–96.

[13] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[14] S.-S. Chow, G. F. Carey, and R. D. Lazarov, *Natural and postprocessed superconvergence in semilinear problems*, Numer. Methods Partial Differential Equations, 7 (1991), pp. 245–259.

[15] B. Cockburn and J. Gopalakrishnan, *A characterization of hybridized mixed methods for second order elliptic problems*, SIAM J. Numer. Anal., 42 (2004), pp. 283–301.

[16] B. Cockburn and J. Gopalakrishnan, *Error analysis of variable degree mixed methods for elliptic problems via hybridization*, Math. Comp., 74 (2005), pp. 1653–1677.

[17] B. Cockburn and J. Gopalakrishnan, *Incompressible finite elements via hybridization. Part* I: *The Stokes system in two space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1627–1650.

[18] B. Cockburn and J. Gopalakrishnan, *Incompressible finite elements via hybridization. Part* II: *The Stokes system in three space dimensions*, SIAM J. Numer. Anal., 43 (2005), pp. 1651–1672.

[19] B. Cockburn, G. Kanschat, and D. Schötzau, *A locally conservative LDG method for the incompressible Navier-Stokes equations*, Math. Comp., 74 (2005), pp. 1067–1095.

[20] R. Courant, *Variational methods for the solution of problems of equilibrium and vibrations*, Bull. Amer. Math. Soc., 49 (1943), pp. 1–23.

[21] L. Demkowicz, *2D hp-adaptive finite element package (2Dhp90). Version* 2.0, Technical Report 02–06, Texas Institute for Computational and Applied Mathematics, Austin, TX, 2002.

[22] L. Demkowicz and A. Buffa, *H^1, $H(\mathrm{curl})$ and $H(\mathrm{div})$-conforming projection-based interpolation in three dimensions. Quasi-optimal p-interpolation estimates*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 267–296.

[23] J. Douglas, Jr., T. Dupont, and M. F. Wheeler, *A Galerkin procedure for approximating the flux on the boundary for elliptic and parabolic boundary value problems*, RAIRO Modél. Math. Anal. Numér., 8 (1974), pp. 47–59.

[24] J. Gopalakrishnan, *A Schwarz preconditioner for a hybridized mixed method*, Comput. Methods Appl. Math., 3 (2003), pp. 116–134.

[25] J. T. R. Hughes, G. Engel, L. Mazzei, and M. Larson, *The continuous Galerkin method is locally conservative*, J. Comput. Phys., 163 (2000), pp. 467–488.

[26] J. T. R. Hughes and G. N. Wells, *Conservation properties for the Galerkin and stabilised*

forms of the advection-diffusion and incompressible Navier-Stokes equations, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 1141–1159.

[27] M. Larson and A. Niklasson, *A conservative flux for the continuous Galerkin method based on discontinuous enrichment*, Calcolo, 41 (2004), pp. 65–76.

[28] A. I. Pehlivanov, R. D. Lazarov, G. F. Carey, and S.-S. Chow, *Superconvergence analysis of approximate boundary-flux calculations*, Numer. Math., 63 (1992), pp. 483–501.

[29] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, in Mathematical Aspects of Finite Element Method, I. Galligani and E. Magenes, eds., Lecture Notes in Math. 606, Springer-Verlag, New York, 1977, pp. 292–315.

[30] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[31] E. Stein and M. Rütter, *Finite element methods for elasticity with error-controlled discretization and model adaptivity*, in Encyclopedia of Computational Mechanics, Vol. 2, R. de Borst, E. Stein, and T. Hughes, eds., John Wiley & Sons, London, 2004, pp. 5–58.

[32] S. Sun and M. Wheeler, *Projections of velocity data for the compatibility with transport*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 653–673.

[33] J. A. Wheeler, *Simulation of heat transfer from a warm pipe buried in permafrost*, in Proceedings of the 74th National Meeting of the American Institute of Chemical Engineers, New Orleans, 1973, p. 43.

[34] M. F. Wheeler, *A Galerkin procedure for estimating the flux for two-point boundary value problems*, SIAM J. Numer. Anal., 11 (1974), pp. 764–768.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1777–1798

A POSTERIORI ERROR ESTIMATION FOR DISCONTINUOUS GALERKIN FINITE ELEMENT APPROXIMATION*

MARK AINSWORTH†

Abstract. It is shown that the interelement discontinuities in a discontinuous Galerkin finite element approximation are subordinate to the error measured in the broken H^1-seminorm. One consequence is that the DG-norm of the error is equivalent to the broken energy seminorm. Computable a posteriori error bounds are obtained for the error measured in both the DG-norm and the broken energy seminorm and are shown to be efficient in the sense that they also provide lower bounds up to a constant and higher order data oscillation terms. The estimators are completely free of unknown constants and provide guaranteed numerical bounds for the error.

Key words. a posteriori error estimation, discontinuous Galerkin method, computable error bounds

AMS subject classifications. Primary, 65N50; Secondary, 65N15, 65N50, 76S05

DOI. 10.1137/060665993

1. Introduction.

Discontinuous Galerkin finite element methods for the approximation of elliptic problems were pioneered in the late 1970s [4, 6, 18] but lay virtually dormant until recently, when they became the subject of intense research activity. The reader is referred to [5, 9] for an overview of developments in the formulation of the methods and their a priori error analysis.

The theory of a posteriori error bounds for discontinuous Galerkin methods in energy-type norms is considerably less developed in comparison with the huge literature on such methods for conforming finite element schemes. Explicit a posteriori estimators for the error in the discontinuous Galerkin approximation measured in mesh-dependent energy-type norms were developed in [7, 12, 14, 15] and later in [11, 16, 19]. Explicit estimators take the form of norms of residuals and jump discontinuities in the finite element approximation, weighted in terms of the local mesh-size and involving unknown generic constants. As noted in [14], the presence of such unknown constants means that one does not actually have numerical bounds on the error, and this limits the practical usage of the estimators to error indication for the purposes of, say, adaptive mesh refinement, as opposed to actual quantitative error control.

In the present work, we wish to derive a posteriori error estimators that are free of any unknown constants, provide actual error bounds, and give local error indicators suitable for driving adaptive refinement procedures. Of course, if one aims to provide error bounds, then the question of the choice of norm in which the error should be bounded naturally arises. Traditionally, a priori error analysis of discontinuous Galerkin methods has been carried out in a mesh-dependent "DG-norm" defined by

$$|||e|||_{DG}^2 = \sum_{K\in\mathcal{P}} \|a^{1/2}\,\mathbf{grad}\, e\|_K^2 + \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \frac{\kappa}{|\gamma|}\|[e]\|_\gamma^2$$

*Received by the editors July 25, 2006; accepted for publication (in revised form) February 27, 2007; published electronically August 24, 2007. This work was partially supported by the Engineering and Physical Sciences Research Council of Great Britain under grant GR/S35103.
http://www.siam.org/journals/sinum/45-4/66599.html
†Mathematics Department, Strathclyde University, 26 Richmond Street, Glasgow G1 1XH, Scotland (M.Ainsworth@strath.ac.uk).

using standard notation (see (11)), which incorporates a term involving the jump discontinuity in the error $[e]$ across element interfaces weighted with the interior penalty parameter κ. However, when it comes to a posteriori error estimation, the practical relevance of a bound for the error measured in a parameter-dependent DG-norm is less clear cut. Of course, the jump term has to be included to ensure that $|||\cdot|||_{DG}$ defines a norm on the discontinuous finite element space, but one may hope that if the jumps have been appropriately penalized, then their contribution should be dominated by the first term; i.e., for κ "sufficiently large" one may hope that

$$\sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D}\frac{\kappa}{|\gamma|}\|[e]\|_\gamma^2\le C\sum_{K\in\mathcal{P}}\|a^{1/2}\,\mathbf{grad}\,e\|_K^2$$

for a positive constant C depending on κ. At first sight, the validity of such a bound hardly seems credible since the bounding term vanishes if e is piecewise constant, while the jump term does not. Nevertheless, in Theorem 3 we show that such a bound does indeed hold (modulo data oscillation which we ignore in this introductory discussion). Moreover, we quantify precisely how "sufficiently large" κ must be and show that the threshold is consistent with the value of κ that must be chosen to ensure the discrete problem itself is uniquely solvable. This result seems to be of some interest in its own right—quite apart from its significance for a posteriori error estimation. In particular, the following equivalence holds:

$$\sum_{K\in\mathcal{P}}\|a^{1/2}\,\mathbf{grad}\,e\|_K^2\le|||e|||_{DG}^2\le C\sum_{K\in\mathcal{P}}\|a^{1/2}\,\mathbf{grad}\,e\|_K^2.$$

As a consequence, we see that it makes sense to estimate and control the error in the broken energy seminorm safe in the knowledge that the jumps are controlled implicitly.

Finally, returning to the issue of a posteriori error bounds, we derive computable upper bounds for the error in the symmetric interior penalty discontinuous Galerkin approximation in both the DG- and broken energy norms. In addition, we obtain local lower bounds which imply that our upper bounds are also lower bounds up to a positive constant. We restrict our attention to the symmetric interior penalty discontinuous Galerkin scheme, although we believe the extension of our results to other variants of discontinuous Galerkin is possible. Some of the results in the present work were announced previously in [2].

2. Preliminaries.

2.1. Model problem.

Consider the model problem

$$\left.\begin{aligned}-\mathbf{div}\,\boldsymbol{\sigma}(u)&=f\in L_2(\Omega)\\ \boldsymbol{\sigma}(u)-a\,\mathbf{grad}\,u&=0\end{aligned}\right\}\ \text{in }\Omega \tag{1}$$

subject to $u=u_D$ on Γ_D and $\sigma_\nu(u)=\boldsymbol{\nu}\cdot\boldsymbol{\sigma}(u)=g\in L_2(\Gamma_N)$ on Γ_N, where $\boldsymbol{\nu}$ is the unit outward normal. The domain Ω is assumed to be a plane polygon, and the disjoint sets Γ_D and Γ_N form a partitioning of the boundary $\Gamma=\partial\Omega$ of the domain. The datum a is assumed to be strictly positive and, for simplicity, is assumed piecewise constant on subdomains of Ω, while u_D is assumed to be continuous, piecewise linear on Γ_D.

The standard variational formulation of the problem consists of seeking $u \in H^1(\Omega)$ such that $u = u_D$ on Γ_D, with

$$(a\,\mathbf{grad}\,u, \mathbf{grad}\,v) = (f, v) + \int_{\Gamma_N} gv\,\mathrm{d}s \quad \forall v \in H^1_E(\Omega), \tag{2}$$

where $H^1_E(\Omega)$ denotes the space $\{v \in H^1(\Omega) : v = 0 \text{ on } \Gamma_D\}$. Throughout, we use the notation $(\cdot,\cdot)_\omega$ to denote the integral inner product over a region ω, and omit the subscript in the case where ω is the physical domain Ω.

2.2. Discontinuous Galerkin formulation. We consider a family of partitions $\{\mathcal{P}\}$ of the domain Ω into the union of nonoverlapping, triangular elements such that the nonempty intersection of a distinct pair of elements is a single common node or single common edge. The family of partitions is assumed to be locally quasi-uniform in the sense that the ratio of the diameters of any pair of neighboring elements is uniformly bounded above and below over the whole family. Furthermore, it is assumed that the partitioning is compatible with the data so that a is piecewise constant on each element and u_D is piecewise linear on an element edge.

The set of all edges of the elements is denoted by $\mathcal{E}$, which we partition into subsets $\mathcal{E}_D$, $\mathcal{E}_N$, and $\mathcal{E}_I$ consisting of edges lying on the Dirichlet boundary Γ_D, the Neumann boundary Γ_N, and the interior edges, respectively. Likewise, the corresponding quantities relative to an individual element K are denoted by $\mathcal{E}(K)$, $\mathcal{E}_D(K)$, $\mathcal{E}_N(K)$, and $\mathcal{E}_I(K)$, respectively. The set of element nodes is denoted by $\mathcal{N}$, while the nodes on a particular element K or edge γ are denoted by $\mathcal{N}(K)$ or $\mathcal{N}(\gamma)$, respectively. For each edge $\gamma \in \mathcal{E}$, the set $\widetilde{\gamma}$ consists of those elements for which γ is an edge,

$$\widetilde{\gamma} = \{K' \in \mathcal{P} : \gamma \in \mathcal{E}(K')\}, \tag{3}$$

while for each element $K \in \mathcal{P}$, the set $\widetilde{K}$ consists of those elements having an edge in common with K,

$$\widetilde{K} = \{K' \in \mathcal{P} : \mathcal{E}(K) \cap \mathcal{E}(K') \text{ is nonempty}\}. \tag{4}$$

Let $X_{\mathcal{P}}$ denote the finite-dimensional space relative to the partition defined by

$$X_{\mathcal{P}} = \{v \in L_2(\Omega) : v_{|K} \in \mathbb{P}_1(K) \quad \forall v \in \mathcal{P}\},$$

where $\mathbb{P}_1(K)$ denotes the set of polynomials of degree at most one in each variable. We now describe the so-called *discontinuous Galerkin finite element method* for the approximation of the solution of the model problem (2) using the space $X_{\mathcal{P}}$. Observe that membership of the space $X_{\mathcal{P}}$ carries no interelement continuity constraints or boundary conditions. Instead, these will be enforced indirectly in the variational scheme. For this purpose, we shall need some notation and conventions to describe jumps and averages of functions associated with the space $X_{\mathcal{P}}$ across interelement edges. For each element $K \in \mathcal{P}$, we let $\mu_K : \partial K \to \{+1, -1\}$ denote a sign function that is piecewise constant on the edges of element K and chosen such that $\mu_K + \mu_{K'} = 0$ on $\partial K \cap \partial K'$ and $\mu_K = 1$ on $\partial K \cap \partial\Omega$. For $v \in X_{\mathcal{P}}$, we define the jump and average values of v on the edges $\mathcal{E}$ by

$$[v] = \begin{cases} \mu_K v_K + \mu_{K'} v_{K'} & \text{on } \gamma = \partial K \cap \partial K', \\ v_K & \text{on } \gamma = \partial K \cap \Gamma_D \end{cases} \tag{5}$$

and

$$\langle v\rangle = \begin{cases} \frac{1}{2}(v_K + v_{K'}) & \text{on } \gamma = \partial K \cap \partial K', \\ v_K & \text{on } \gamma = \partial K \cap \Gamma_D. \end{cases} \tag{6}$$

The sign functions $\{\mu_K\}$ may be used to define a unique, unit normal vector $\boldsymbol{\nu}$ on any given edge $\gamma \in \mathcal{E}$ according to the formula $\boldsymbol{\nu} = \mu_K \boldsymbol{\nu}_K$, $\gamma \subset \partial K$, where $\boldsymbol{\nu}_K$ denotes the unit outward normal relative to element K. This definition is independent of the choice of element K sharing the γ. The jump and average in the flux of a function $v \in X_{\mathcal{P}}$ on individual edges may then be defined by $[\sigma_\nu(v)] = \boldsymbol{\nu} \cdot [\boldsymbol{\sigma}(v)]$ and $\langle \sigma_\nu(v)\rangle = \boldsymbol{\nu} \cdot \langle \boldsymbol{\sigma}(v)\rangle$.

For a given positive constant κ to be specified later, we define the bilinear form $\mathcal{B}_{\mathcal{P}} : X_{\mathcal{P}} \times X_{\mathcal{P}} \to \mathbb{R}$ by the rule

$$\begin{aligned} \mathcal{B}_{\mathcal{P}}(v,w) &= \sum_{K\in\mathcal{P}} (a\,\mathbf{grad}\, v, \mathbf{grad}\, w)_K \\ &\quad - \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \int_\gamma \left(\langle\sigma_\nu(v)\rangle\,[w] + [v]\,\langle\sigma_\nu(w)\rangle\right) \mathrm{d}s + \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \frac{\kappa}{|\gamma|}\int_\gamma [v]\,[w]\,\mathrm{d}s \end{aligned} \tag{7}$$

and the linear form $\mathcal{L}_{\mathcal{P}} : X_{\mathcal{P}} \to \mathbb{R}$ by

$$\begin{aligned} \mathcal{L}_{\mathcal{P}}(w) &= \sum_{K\in\mathcal{P}} (f, w_K)_K \\ &\quad + \sum_{\gamma\in\mathcal{E}_N} \int_\gamma g w\,\mathrm{d}s - \sum_{\gamma\in\mathcal{E}_D}\int_\gamma u_D\,\langle\sigma_\nu(w)\rangle\,\mathrm{d}s + \sum_{\gamma\in\mathcal{E}_D}\frac{\kappa}{|\gamma|}\int_\gamma u_D w\,\mathrm{d}s, \end{aligned} \tag{8}$$

where $|\gamma|$ is used to denote the length of an edge γ.

An approximation of the true solution u is obtained by seeking $U_{\mathcal{P}} \in X_{\mathcal{P}}$ such that

$$\mathcal{B}_{\mathcal{P}}(U_{\mathcal{P}}, v) = \mathcal{L}_{\mathcal{P}}(v) \quad \forall v \in X_{\mathcal{P}}. \tag{9}$$

This type of scheme is often referred to as the *symmetric interior penalty discontinuous Galerkin finite element method*. The quantity $|||\cdot|||$ defined by

$$|||v||| = \left\{ \sum_{K\in\mathcal{P}} (a\,\mathbf{grad}\, v, \mathbf{grad}\, v)_K \right\}^{1/2} \tag{10}$$

is sometimes dubbed the *broken energy or H^1-seminorm* (note that $|||v|||$ vanishes whenever v is a piecewise constant function), while the quantity $|||\cdot|||_{DG}$ given by

$$|||v|||_{DG} = \left\{ |||v|||^2 + \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \frac{\kappa}{|\gamma|}\int_\gamma [v]^2\,\mathrm{d}s \right\}^{1/2} \tag{11}$$

is generally referred to as the *DG-norm* and plays an important role in the a priori error analysis of the method.

For a sufficiently large choice of parameter κ, it may be shown (see, for example, section 4.2 of [5]) that (9) has a unique solution. In view of the developments in section 4, where bounds on the residuals are obtained under an assumption on the size of the interior penalty parameter κ, it is of practical importance that the bound assumed there is consistent with the values of interior penalty for which the method itself is well-posed. Therefore, we shall give a simple quantitative bound on the choice of the interior penalty parameter κ.

The bound on the interior penalty parameter is stated in terms of the spectral radius of the element stiffness matrix $\boldsymbol{S}_K$ obtained using the standard barycentric coordinates $\{\lambda_n : n \in \mathcal{N}(K)\}$ on element K, i.e.,

$$[\boldsymbol{S}_K]_{mn} = (a\,\mathbf{grad}\,\lambda_m, \mathbf{grad}\,\lambda_n)_K. \tag{12}$$

Obviously, $\boldsymbol{S}_K$ is positive semidefinite and has a largest eigenvalue $\varrho(\boldsymbol{S}_K)$ that depends on the shape of the element but not on the mesh-size.

LEMMA 1. *Suppose that the interior penalty parameter appearing in the bilinear form $\mathcal{B}_{\mathcal{P}}(\cdot,\cdot)$ is chosen so that*

$$\kappa > 4 \max_{K\in\mathcal{P}} \varrho(\boldsymbol{S}_K). \tag{13}$$

Then there exists a unique $U_{\mathcal{P}} \in X_{\mathcal{P}}$ such that

$$\mathcal{B}_{\mathcal{P}}(U_{\mathcal{P}}, v) = \mathcal{L}_{\mathcal{P}}(v) \quad \forall v \in X_{\mathcal{P}}. \tag{14}$$

Proof. Thanks to the finite dimensionality, it suffices to show that the only solution to the homogeneous problem vanishes. Let $\delta > 0$ be a constant to be determined and let $v \in X_{\mathcal{P}}$ be arbitrary. Then, for each edge $\gamma \in \mathcal{E}_I \cup \mathcal{E}_D$,

$$2\int_\gamma \langle \sigma_\nu(v)\rangle\,[v]\;\mathrm{d}s \le \delta \int_\gamma |\gamma|\,\langle \sigma_\nu(v)\rangle^2\;\mathrm{d}s + \delta^{-1}\int_\gamma |\gamma|^{-1}\,[v]^2\;\mathrm{d}s.$$

If $\gamma \in \mathcal{E}_I$ is an interior edge shared by elements K and K', then

$$\int_\gamma |\gamma|\,\langle \sigma_\nu(v)\rangle^2\;\mathrm{d}s \le \frac{1}{2}|\gamma|^2\left\{(\boldsymbol{\nu}_K\cdot\boldsymbol{\sigma}_K(v)_{|\gamma})^2 + (\boldsymbol{\nu}_{K'}\cdot\boldsymbol{\sigma}_{K'}(v)_{|\gamma})^2\right\},$$

where we have exploited the fact that $\boldsymbol{\sigma}(v)$ is constant on each element. Likewise, if $\gamma \in \mathcal{E}_D$ is an edge of element K, then

$$\int_\gamma |\gamma|\,\langle \sigma_\nu(v)\rangle^2\;\mathrm{d}s \le |\gamma|^2(\boldsymbol{\nu}_K\cdot\boldsymbol{\sigma}_K(v)_{|\gamma})^2.$$

Hence, by including nonnegative contributions from the edges on the Neumann boundary, we deduce that

$$\sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D}\int_\gamma |\gamma|\,\langle \sigma_\nu(v)\rangle^2\;\mathrm{d}s \le \sum_{K\in\mathcal{P}}\sum_{\gamma\subset\partial K}|\gamma|^2(\boldsymbol{\nu}_K\cdot\boldsymbol{\sigma}_K(v))^2_{|\gamma}.$$

Making use of the relationship $\mathbf{grad}\,\lambda_{n(\gamma)} = -|\gamma|\boldsymbol{\nu}_K/2|K|$ between the barycentric coordinate $\lambda_{n(\gamma)}$ associated with the vertex opposite edge γ and the unit outward normal $\boldsymbol{\nu}_K$ on the edge, we deduce that

$$\sum_{\gamma\subset\partial K}|\gamma|^2(\boldsymbol{\nu}_K\cdot\boldsymbol{\sigma}_K(v))^2_{|\gamma} = 4\vec{v}_K^{\top}\boldsymbol{S}_K^2\vec{v}_K,$$

where $\vec{v}_K$ is the vector of values of v at the vertices of element K, and hence

$$|\gamma|^2(\boldsymbol{\nu}_K \cdot \boldsymbol{\sigma}_K(v))^2_{|\gamma} \le 4\varrho(\boldsymbol{S}_K)\vec{v}_K^{\top}\boldsymbol{S}_K\vec{v}_K = 4\varrho(\boldsymbol{S}_K)(a\,\mathbf{grad}\,v, \mathbf{grad}\,v)_K.$$

These estimates imply that

$$\mathcal{B}_{\mathcal{P}}(v,v) \ge \sum_{K\in\mathcal{P}} (1 - 4\delta\varrho(\boldsymbol{S}_K))(a\,\mathbf{grad}\,v, \mathbf{grad}\,v)_K$$

$$+ \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} (\kappa - \delta^{-1})|\gamma|^{-1}\int_\gamma [v]^2\,\mathrm{d}s.$$

By selecting δ such that $\kappa > \delta^{-1} > 4\max_{K\in\mathcal{P}}\varrho(\boldsymbol{S}_K)$, we ensure that $1 - 4\delta\varrho(\boldsymbol{S}_K)$ and $\kappa - \delta^{-1}$ are positive. Consequently, in the case of homogeneous data $\mathcal{L}_{\mathcal{P}}(\cdot)$, any solution $U_{\mathcal{P}}$ of (14) satisfies $(a\,\mathbf{grad}\,U_{\mathcal{P}}, \mathbf{grad}\,U_{\mathcal{P}})_K = 0$ on every element K, and $\int_\gamma [U_{\mathcal{P}}]^2\,\mathrm{d}s = 0$ on the interior and Dirichlet edges. Consequently, $U_{\mathcal{P}}$ must vanish identically. This completes the proof. □

3. A posteriori error analysis. Various a priori error bounds are available for the approximation scheme described above [5]. However, the issues that we wish to focus on in the present work are (i) *can one obtain computable a posteriori estimates for the error $u - U_{\mathcal{P}}$, and if so,* (ii) *in what norms?* One complication is caused by the fact that the error $e = u - U_{\mathcal{P}}$ in the finite element approximation generally fails to belong to the natural space $H^1_E(\Omega)$ for the original variational problem (2). The following result, due to Dari et al. [10], will be useful in this respect, where, for a discontinuous function v, we define the broken gradient by the rule $\mathbf{grad}_{\mathcal{P}}\,v = (\mathbf{grad}\,v)_{|K}$ and the broken flux by the rule $\boldsymbol{\sigma}_{\mathcal{P}}(v) = (a\,\mathbf{grad}\,v)_{|K}$ on element K.

THEOREM 1. *Let $\mathcal{H}$ denote the space*

$$\mathcal{H} = \{w \in H^1(\Omega) : \partial w/\partial s = 0 \text{ on } \Gamma_N\}. \tag{15}$$

Then the error in the flux may be decomposed into the form

$$\boldsymbol{\sigma}_{\mathcal{P}}(e) = \boldsymbol{\sigma}(\chi) + \mathbf{curl}\,\psi, \tag{16}$$

where $\chi \in H^1_E(\Omega)$ satisfies

$$(a\,\mathbf{grad}\,\chi, \mathbf{grad}\,v) = (a\,\mathbf{grad}_{\mathcal{P}}\,e, \mathbf{grad}\,v) \quad \forall v \in H^1_E(\Omega) \tag{17}$$

and $\psi \in \mathcal{H}$ satisfies

$$(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,w) = (a^{-1}\boldsymbol{\sigma}_{\mathcal{P}}(e), \mathbf{curl}\,w) = (\mathbf{grad}_{\mathcal{P}}\,e, \mathbf{curl}\,w) \quad \forall w \in \mathcal{H}. \tag{18}$$

This splitting is orthogonal in the sense that

$$(a^{-1}\boldsymbol{\sigma}_{\mathcal{P}}(e), \boldsymbol{\sigma}_{\mathcal{P}}(e)) = (a^{-1}\boldsymbol{\sigma}(\chi), \boldsymbol{\sigma}(\chi)) + (a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi). \tag{19}$$

A proof of this result can be found in [10], while a proof of a slightly more general version is given in [1]. The function χ constitutes the orthogonal projection of the total error onto the conforming space $H^1_E(\Omega)$ and is referred to as the *conforming error*, while the remaining part ψ is referred to as the *nonconforming error*. The orthogonality of the splitting means that it suffices to estimate the contribution of each part to the total error independently.

3.1. Statement of the a posteriori error bounds. The treatment of the nonconforming part of the error forms the subject of section 6, where it is shown in Lemma 9 that

$$(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi) \leq \sum_{K\in\mathcal{P}} \eta_{\mathsf{NC},K}^2,$$

where $\eta_{\mathsf{NC},K} = \|a^{1/2}\,\mathbf{grad}_{\mathcal{P}}(U_{\mathcal{P}}^* - U_{\mathcal{P}})\|_K$ and $U_{\mathcal{P}}^*$ is defined in section 6. The conforming part of the error is dealt with in section 5, where it is shown in Lemma 8 that

$$(a\,\mathbf{grad}\,\chi, \mathbf{grad}\,\chi) \leq \sum_{K\in\mathcal{P}} \eta_{\mathsf{CF},K}(\beta)^2,$$

where

$$\eta_{\mathsf{CF},K}(\beta) = a_K^{-1/2}\sqrt{\|\boldsymbol{\rho}_K\|_K^2 - \mathcal{C}_K^\star(\beta)^2|K|\|\mathrm{curl}\,\boldsymbol{\rho}_K\|_K^2}$$
$$+ \mathcal{C}_p(K)a_K^{-1/2}\|f - \overline{f}_K\|_K + \sum_{\gamma\in\mathcal{E}_N(K)} \mathcal{C}_t(K,\gamma)a_K^{-1/2}\|g - \overline{g}_\gamma\|_\gamma,$$

a_K denotes the restriction of the data a to element K, and $\overline{f}_K = |K|^{-1}\int_K f\,\mathrm{d}\boldsymbol{x}$ and $\overline{g}_\gamma = |\gamma|^{-1}\int_\gamma g\,\mathrm{d}s$ denote average values of the data over an element or edge, respectively.

We emphasize that *all* of these quantities are given explicitly and are fully computable in terms of the discontinuous Galerkin approximation $U_{\mathcal{P}}$, the data f and g_N, and geometrical information on the element. The final two terms often represent higher order contributions measuring the oscillation in the data f and g_N but are nevertheless included in the estimator, using values for the multiplicative constants $\mathcal{C}_p$ and $\mathcal{C}_t$ given explicitly in (61) and (64), so that one has a guaranteed upper bound on the error even if the terms turn out not to be of higher order. The principal part of the estimator involves the function $\boldsymbol{\rho}_K$ defined in Lemma 6, where an easily computable closed form expression for its norm is given. The quantity $\mathcal{C}_K^\star(\beta)$ defined in (49) involves an arbitrarily chosen "bubble" function $\beta \in H_0^1(K)$. One possibility is to choose $\beta = 0$. The best choice, in terms of maximizing $\mathcal{C}_K^\star$, satisfies $-\Delta\beta = 1$ in K and vanishes on the element boundary. In general such a function is not available in closed form and in practice a simple cubic polynomial approximation of the function is found to be sufficient, for which one may show that

$$\mathcal{C}_K^\star(\beta)^2 = \frac{a_K}{20\,\mathrm{trace}(\boldsymbol{S}_K)}, \tag{20}$$

where $\boldsymbol{S}_K$ is the element stiffness matrix in terms of the barycentric coordinates defined earlier.

Lemmas 8 and 9 also assert that $\eta_{\mathsf{CF},K}(\beta)$ and $\eta_{\mathsf{NC},K}$ provide local lower bounds for the conforming and nonconforming parts of the error up to data oscillation. Consequently, by combining Theorem 1 and the above lemmas we obtain computable bounds for the error measured in both the broken energy seminorm and DG-norm, along with corresponding lower bounds.

THEOREM 2. *Let $e = U_{\mathcal{P}} - u$ denote the error in the discontinuous Galerkin approximation. Then*

$$|||e|||^2 \leq \sum_{K\in\mathcal{P}} (\eta_{\mathsf{CF},K}(\beta)^2 + \eta_{\mathsf{NC},K}^2) \tag{21}$$

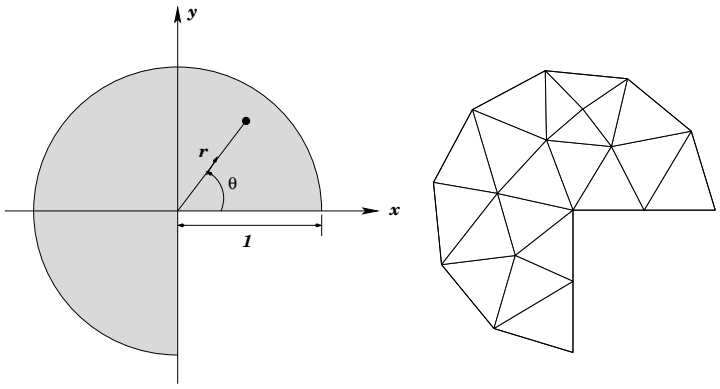


FIG. 1. (a) *Domain* Ω *and* (b) *initial mesh used in numerical example.*

and

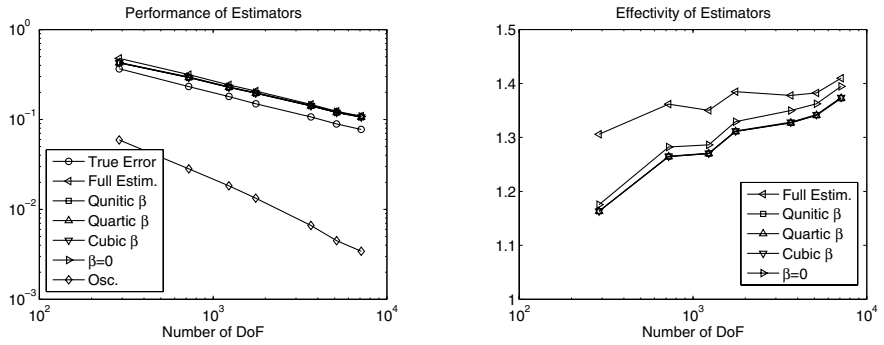
$$\|\!|e|\!\|_{DG}^2 \leq \sum_{K\in\mathcal{P}} (\eta_{\mathsf{CF},K}(\beta)^2 + \eta_{\mathsf{NC},K}^2) + \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \frac{\kappa}{|\gamma|}\|[U_{\mathcal{P}}]\|_\gamma^2, \tag{22}$$

where $\eta_{\mathsf{CF},K}(\beta)$ *and* $\eta_{\mathsf{NC},K}$ *are defined in Lemmas* 8 *and* 9*, respectively. Moreover, if the interior penalty parameter* $\kappa > 4\max_{K\in\mathcal{P}} \varrho(\boldsymbol{S}_K)$*, then the above bounds are efficient up to data oscillation and a positive constant independent of any mesh-size and the solution.*

3.2. Numerical example. In order to illustrate the theoretical results, we consider a simple Poisson problem with homogeneous Dirichlet data on the domain Ω as shown in Figure 1, with the source term f chosen so that the true solution is given by $u(r,\theta) = (r^3 - r^{2/3})\sin(2\theta/3)$. The problem is approximated using discontinuous Galerkin with an interior penalty parameter $\kappa = 10$, which was observed to be consistent with the bound given in Lemma 1, using a sequence of adaptively refined meshes obtained starting with the mesh shown in Figure 1 and selecting those elements for which the "full" local error indicator exceeds 30% of the value of the largest local error indicator. In Figure 2 we compare the values of the a posteriori error estimator corresponding to the choice $\beta = 0$, or to choosing β to be the cubic, quartic, or quintic polynomial that maximizes the value of $\mathcal{C}_K^\star(\beta)$. In order to illustrate the influence of the choice of the bubble more clearly, the data oscillation term in (57) is not included but rather is shown separately, where it is seen to be of higher order. The "full" estimator, with optimal quintic bubble and including data oscillation, is also shown separately.

The results obtained in this particular example are typical. In particular, we observe that the use of a cubic bubble β provides a marked improvement over the raw estimator ($\beta = 0$), while the use of higher order bubbles provides marginal further improvement at best. Therefore, the use of a cubic bubble seems to be merited in general practical computations, particularly in view of the ease with which it may be implemented.

4. Jumps in DGFEM are subordinate to the error in the broken H^1-seminorm. As usual, we define the oscillation of the data $f \in L_2(\Omega)$ over a collection



(a) (b)

FIG. 2. (a) *Values of a posteriori estimators and data oscillation, and* (b) *ratio of estimated error to true error obtained for numerical example.*

$\mathcal{P}'$ of elements by

$$\operatorname{osc}(f,\mathcal{P}')^2 = \sum_{K\subset\mathcal{P}'} |K| \|f - \overline{f}_K\|_K^2, \tag{23}$$

where $\overline{f}_K$ is the average value of f over element K. Likewise, the oscillation of the Neumann data g over a collection $\mathcal{E}' \subset \mathcal{E}_N$ of edges is defined by

$$\operatorname{osc}(g,\mathcal{E}')^2 = \sum_{\gamma\subset\mathcal{E}'} |\gamma| \|g - \overline{g}_\gamma\|_\gamma^2, \tag{24}$$

where $\overline{g}_\gamma$ denotes the average value of g on the edge γ. The main result of this section may be stated as follows.

THEOREM 3. *Suppose that* $\kappa > 4\max_{K\in\mathcal{P}} \varrho(\boldsymbol{S}_K)$, *and* $e = U_\mathcal{P} - u$ *denote the error in the discontinuous Galerkin approximation. Then*

$$\begin{aligned}\sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \frac{\kappa}{|\gamma|}\int_\gamma [e]^2 \,\mathrm{d}s \\ \le C\kappa \left[\frac{\kappa}{\kappa - 4\max_{K\in\mathcal{P}}\varrho(\boldsymbol{S}_K)}\right]^2 \left[\sum_{K\in\mathcal{P}} \|a^{1/2}\,\mathbf{grad}\, e\|_K^2 + \operatorname{osc}(f,\mathcal{P})^2 + \operatorname{osc}(g,\mathcal{E}_N)^2\right],\end{aligned} \tag{25}$$

where C *is a positive constant independent of any mesh-size.*

This result shows, up to terms involving oscillation of the data, that the contributions from the jump terms to the value of $\|\!|e|\!\|_{DG}$ may be bounded in terms of the contributions from the broken H^1-seminorm $\|\!|e|\!\|$. At first glance, this result is somewhat surprising in the sense that it does not hold for an arbitrary function. Note, for example, that (ignoring oscillation terms) the expression appearing in the bound vanishes when applied to a piecewise constant function, while the left-hand side is nonzero for such a function. There is of course no paradox here, since the estimate is claimed to hold only in the case of the error in a discontinuous Galerkin approximation.

This type of estimate seems to be of wider significance for discontinuous Galerkin finite element schemes, confirming that the jump discontinuities in the approximation

are properly controlled through the interior penalty term and, as a result, play a subordinate role to that of the error measured in the broken H^1-seminorm.

The proof of Theorem 3 is postponed until the end of this section. We begin by deriving estimates for the element volume residuals and jumps in function values and fluxes across interelement edges in terms of the error in a neighborhood of the entity.

The bounds given in the following result for the volume residuals and interelement fluxes are more or less standard and we confine ourselves to a statement of the results.

LEMMA 2. *Let $\chi \in H^1_E(\Omega)$ denote the conforming part of the error defined in* (17). *Then there exist positive constants c, independent of any mesh-size, such that on every element $K \in \mathcal{P}$,*

$$ch_K\|f\|_K \le \|a^{1/2}\,\mathbf{grad}\,\chi\|_K + \operatorname{osc}(f,K); \tag{26}$$

on every edge $\gamma \in \mathcal{E}_I$,

$$c|\gamma|^{1/2}\|[\sigma_\nu(U_{\mathcal{P}})]\|_\gamma \le \|a^{1/2}\,\mathbf{grad}\,\chi\|_{\widetilde{\gamma}} + \operatorname{osc}(f,\widetilde{\gamma}); \tag{27}$$

and, on every edge $\gamma \in \mathcal{E}_N$,

$$c|\gamma|^{1/2}\|g - \sigma_\nu(U_{\mathcal{P}})\|_\gamma \le \|a^{1/2}\,\mathbf{grad}\,\chi\|_{\widetilde{\gamma}} + \operatorname{osc}(f,\widetilde{\gamma}) + \operatorname{osc}(g,\gamma). \tag{28}$$

Let $\psi \in \mathcal{H}$ denote the nonconforming part of the error defined in (18). *Then, on every edge $\gamma \in \mathcal{E}_I$,*

$$c|\gamma|^{1/2}\|[\partial U_{\mathcal{P}}/\partial s]\|_\gamma \le \|a^{-1/2}\,\mathbf{curl}\,\psi\|_{\widetilde{\gamma}}, \tag{29}$$

and on every edge $\gamma \in \mathcal{E}_D$,

$$c|\gamma|^{1/2}\|\partial(u_D - U_{\mathcal{P}})/\partial s\|_\gamma \le \|a^{-1/2}\,\mathbf{curl}\,\psi\|_{\widetilde{\gamma}}. \tag{30}$$

Proof. The first three estimates are obtained by applying a standard "bubble" function argument to (17) (see, e.g., [3, section 2.3] or [17]). The remaining estimates are obtained in a similar fashion (see, e.g., [10]) using (18). ∎

The above estimates give only bounds on the jumps in the gradient of the approximation across element boundaries. Our objective here is to obtain bounds on the jumps in actual values. The next result gives bounds on the average value of the jump in the approximation across edges provided that the choice of the penalty parameter κ is consistent with the bound obtained in Lemma 1 that was shown to be sufficient to ensure that the discrete scheme is well-posed.

LEMMA 3. *Suppose that $\kappa > 4\varrho(\boldsymbol{S}_K)$. Then*

$$\begin{aligned}&\sum_{\gamma\in\mathcal{E}_I(K)}\left||\gamma|^{-1}\int_\gamma [U_{\mathcal{P}}]\,\mathrm{d}s\right| + \sum_{\gamma\in\mathcal{E}_D(K)}\left||\gamma|^{-1}\int_\gamma (U_{\mathcal{P}} - u_D)\,\mathrm{d}s\right| \\ &\le C(\kappa - 4\varrho(\boldsymbol{S}_K))^{-1}\Big[\|a^{1/2}\,\mathbf{grad}\,\chi\|_{\widetilde{K}} + \operatorname{osc}(f,\widetilde{K}) \\ &\qquad + \kappa\|a^{-1/2}\,\mathbf{curl}\,\psi\|_{\widetilde{K}} + \operatorname{osc}(g,\mathcal{E}_N(K))\Big],\end{aligned} \tag{31}$$

where C is a positive constant independent of any mesh-size.

Proof. Fix an element $K \in \mathcal{P}$ and, for each edge $\gamma \in \mathcal{E}_I(K) \cup \mathcal{E}_D(K)$, define

$$\alpha_\gamma = \begin{cases} \dfrac{1}{|\gamma|} \displaystyle\int_\gamma \mu_K \, [U_\mathcal{P}] \, \mathrm{d}s, & \gamma \in \mathcal{E}_I(K), \\ \dfrac{1}{|\gamma|} \displaystyle\int_\gamma (U_\mathcal{P} - u_D) \, \mathrm{d}s, & \gamma \in \mathcal{E}_D(K). \end{cases}$$

Let $\gamma^* \in \mathcal{E}_I(K) \cup \mathcal{E}_D(K)$ and define $\varphi_{\gamma^*} \in \mathbb{P}_1(K)$ by the rule $\varphi_{\gamma^*} = 1 - 2\lambda_*$, where λ_* is the barycentric coordinate associated with the vertex in element K that is opposite to the edge γ^*. Observe that φ_{γ^*} has constant value (unity) on the edge γ^* and satisfies

$$\frac{1}{|\gamma|} \int_\gamma \varphi_{\gamma^*} \, \mathrm{d}s = \delta_{\gamma^*\gamma}, \quad \gamma \in \mathcal{E}(K). \tag{32}$$

The lack of any interelement continuity requirements on the test functions means that the function φ_{γ^*}, extended onto the remaining elements by zero, is admissible.

Inserting the test function into the variational statement (9), integrating the volumetric term in the bilinear form (7) by parts, and simplifying, we arrive at the identity

$$\begin{aligned} &(f, \varphi_{\gamma^*})_K + \sum_{\gamma \in \mathcal{E}_N(K)} \int_\gamma (g - \sigma_\nu(U_\mathcal{P})) \varphi_{\gamma^*} \, \mathrm{d}s - \frac{1}{2} \sum_{\gamma \in \mathcal{E}_I(K)} \int_\gamma [\sigma_\nu(U_\mathcal{P})] \, \varphi_{\gamma^*} \, \mathrm{d}s \\ &= \sum_{\gamma \in \mathcal{E}_D(K)} \frac{\kappa}{|\gamma|} \int_\gamma (U_\mathcal{P} - u_D) \varphi_{\gamma^*} \, \mathrm{d}s + \sum_{\gamma \in \mathcal{E}_I(K)} \frac{\kappa}{|\gamma|} \int_\gamma \mu_K \, [U_\mathcal{P}] \, \varphi_{\gamma^*} \, \mathrm{d}s \\ &\quad - \sum_{\gamma \in \mathcal{E}_D(K)} \int_\gamma (U_\mathcal{P} - u_D) \sigma_{\nu_K}(\varphi_{\gamma^*}) \, \mathrm{d}s - \frac{1}{2} \sum_{\gamma \in \mathcal{E}_I(K)} \int_\gamma \mu_K \, [U_\mathcal{P}] \, \sigma_{\nu_K}(\varphi_{\gamma^*}) \, \mathrm{d}s. \end{aligned} \tag{33}$$

It will be useful to express several of the terms appearing on the right-hand side in an alternative form. In view of (32), for $\gamma \in \mathcal{E}_I(K)$ there holds

$$\frac{\kappa}{|\gamma|} \int_\gamma \mu_K \, [U_\mathcal{P}] \, \varphi_{\gamma^*} \, \mathrm{d}s = \begin{cases} \kappa \alpha_{\gamma^*} & \text{if } \gamma^* = \gamma, \\ \dfrac{\kappa}{|\gamma|} \displaystyle\int_\gamma \mu_K ([U_\mathcal{P}] - c_\gamma) \varphi_{\gamma^*} \, \mathrm{d}s & \text{otherwise,} \end{cases}$$

while for $\gamma \in \mathcal{E}_D(K)$,

$$\frac{\kappa}{|\gamma|} \int_\gamma (U_\mathcal{P} - u_D) \varphi_{\gamma^*} \, \mathrm{d}s = \begin{cases} \kappa \alpha_{\gamma^*} & \text{if } \gamma^* = \gamma, \\ \dfrac{\kappa}{|\gamma|} \displaystyle\int_\gamma (U_\mathcal{P} - u_D - c_\gamma) \varphi_{\gamma^*} \, \mathrm{d}s & \text{otherwise,} \end{cases}$$

where c_γ is an arbitrary constant (to be chosen later).

By observing that $\sigma_{\nu_K}(\varphi_{\gamma^*})$ is piecewise constant on the element boundary and again using property (32), we see that

$$|\gamma| \, \sigma_{\nu_K}(\varphi_{\gamma^*})|_\gamma = \int_{\partial K} \varphi_\gamma \sigma_{\nu_K}(\varphi_{\gamma^*}) \, \mathrm{d}s = (a \, \mathbf{grad} \, \varphi_{\gamma^*}, \mathbf{grad} \, \varphi_\gamma)_K$$

for any edge $\gamma \subset \partial K$. By further exploiting the fact that $\sigma_{\nu_K}(\varphi_{\gamma^*})$ is piecewise constant and by the definition of α_γ, we deduce that

$$\int_\gamma (U_{\mathcal{P}} - u_D)\sigma_{\nu_K}(\varphi_{\gamma^*})\,\mathrm{d}s = (a\,\mathbf{grad}\,\varphi_{\gamma^*}, \mathbf{grad}\,\varphi_\gamma)_K\ \alpha_\gamma, \quad \gamma \in \mathcal{E}_D(K),$$

and

$$\int_\gamma \mu_K\,[U_{\mathcal{P}}]\,\sigma_{\nu_K}(\varphi_{\gamma^*})\,\mathrm{d}s = (a\,\mathbf{grad}\,\varphi_{\gamma^*}, \mathbf{grad}\,\varphi_\gamma)_K\ \alpha_\gamma, \quad \gamma \in \mathcal{E}_I(K).$$

It is convenient to express these quantities in terms of entries in the element stiffness matrix $\boldsymbol{S}_K$ relative to the barycentric coordinates defined earlier:

$$(a\,\mathbf{grad}\,\varphi_\gamma, \mathbf{grad}\,\varphi_{\gamma'})_K = 4[\boldsymbol{S}_K]_{\gamma\gamma'}.$$

Inserting the above alternative forms into expression (33) and rearranging, we obtain the following relation for each edge $\gamma^* \in \mathcal{E}_I(K) \cup \mathcal{E}_D(K)$:

$$\kappa\alpha_{\gamma^*} - 4\sum_{\gamma\in\mathcal{E}_D(K)} [\boldsymbol{S}_K]_{\gamma^*\gamma}\ \alpha_\gamma - 2\sum_{\gamma\in\mathcal{E}_I(K)} [\boldsymbol{S}_K]_{\gamma^*\gamma}\ \alpha_\gamma = r^K_{\gamma^*},$$

where

$$\begin{aligned} r^K_{\gamma^*} &= (f, \varphi_{\gamma^*})_K \\ &+ \sum_{\gamma\in\mathcal{E}_N(K)} \int_\gamma (g - \sigma_\nu(U_{\mathcal{P}}))\varphi_{\gamma^*}\,\mathrm{d}s - \frac{1}{2}\sum_{\gamma\in\mathcal{E}_I(K)} \int_\gamma [\sigma_\nu(U_{\mathcal{P}})]\,\varphi_{\gamma^*}\,\mathrm{d}s \\ &- \sum_{\gamma\in\mathcal{E}_D(K)\setminus\gamma^*} \frac{\kappa}{|\gamma|}\int_\gamma (U_{\mathcal{P}} - u_D - c_\gamma)\varphi_{\gamma^*}\,\mathrm{d}s \\ &- \sum_{\gamma\in\mathcal{E}_I(K)\setminus\gamma^*} \frac{\kappa}{|\gamma|}\int_\gamma \mu_K([U_{\mathcal{P}}] - c_\gamma)\varphi_{\gamma^*}\,\mathrm{d}s. \end{aligned}$$

Introducing a diagonal matrix $\boldsymbol{\Lambda}_K \in \mathbb{R}^{m\times m}$, where m is the number of Dirichlet and internal edges on K, with entries

$$\Lambda_{\gamma\gamma} = \begin{cases} 1 & \text{if } \gamma \in \mathcal{E}_D(K), \\ \frac{1}{2} & \text{if } \gamma \in \mathcal{E}_I(K), \end{cases}$$

we may write the above equations in the form

$$\kappa\boldsymbol{\alpha}_K - 4\boldsymbol{S}_K\boldsymbol{\Lambda}_K\boldsymbol{\alpha}_K = \boldsymbol{r}_K,$$

where $[\boldsymbol{\alpha}_K]_\gamma = \alpha_\gamma$ and $[\boldsymbol{r}_K]_\gamma = r^K_\gamma$. Observing that

$$(\boldsymbol{\Lambda}_K\boldsymbol{\alpha}_K)^\top \boldsymbol{S}_K\boldsymbol{\Lambda}_K\boldsymbol{\alpha}_K \le \varrho(\boldsymbol{S}_K)\boldsymbol{\alpha}_K^\top\boldsymbol{\Lambda}_K^2\boldsymbol{\alpha}_K \le \varrho(\boldsymbol{S}_K)\boldsymbol{\alpha}_K^\top\boldsymbol{\Lambda}_K\boldsymbol{\alpha}_K,$$

we obtain

$$(\kappa - 4\varrho(\boldsymbol{S}_K))\boldsymbol{\alpha}_K^\top\boldsymbol{\Lambda}_K\boldsymbol{\alpha}_K \le \boldsymbol{\alpha}_K^\top\boldsymbol{\Lambda}_K\boldsymbol{r}_K$$

and hence, with the aid of a Cauchy–Schwarz inequality,

$$(\kappa - 4\varrho(\boldsymbol{S}_K))^2 \boldsymbol{\alpha}_K^\top \boldsymbol{\Lambda}_K \boldsymbol{\alpha}_K \leq \boldsymbol{r}_K^\top \boldsymbol{\Lambda}_K \boldsymbol{r}_K.$$

It remains to bound the terms appearing in the entries $r_{\gamma^*}^K$ of the vector $\boldsymbol{r}_K$. Observing that φ_{γ^*} satisfies $\|\varphi_{\gamma^*}\|_K \sim h_K$ and $\|\varphi_{\gamma^*}\|_\gamma \sim |\gamma|^{1/2}$, we obtain

$$\begin{aligned} c|r_{\gamma^*}^K| &\leq h_K \|f\|_K \\ &+ \sum_{\gamma \in \mathcal{E}_N(K)} |\gamma|^{1/2} \|g - \sigma_\nu(U_{\mathcal{P}})\|_\gamma + \frac{1}{2} \sum_{\gamma \in \mathcal{E}_I(K)} |\gamma|^{1/2} \|[\sigma_\nu(U_{\mathcal{P}})]\|_\gamma \\ &+ \sum_{\gamma \in \mathcal{E}_D(K) \setminus \gamma^*} \kappa |\gamma|^{-1/2} \|U_{\mathcal{P}} - u_D - c_\gamma\|_\gamma \\ &+ \sum_{\gamma \in \mathcal{E}_I(K) \setminus \gamma^*} \kappa |\gamma|^{-1/2} \|[U_{\mathcal{P}}] - c_\gamma\|_\gamma, \end{aligned}$$

where c is a positive constant independent of κ and any mesh-size. The above estimate holds for all choices of constants $\{c_\gamma\}$ and hence, taking c_γ to be appropriate averages and using a scaling argument, we may arrange that

$$\|U_{\mathcal{P}} - u_D - c_\gamma\|_\gamma \leq C|\gamma| \|\partial(U_{\mathcal{P}} - u_D)/\partial s\|_\gamma, \quad \gamma \in \mathcal{E}_D(K),$$

and

$$\|[\sigma_\nu(U_{\mathcal{P}})] - c_\gamma\|_\gamma \leq C|\gamma| \|[\partial U_{\mathcal{P}}/\partial s]\|_\gamma, \quad \gamma \in \mathcal{E}_I(K).$$

In view of the estimates in Lemma 2, we conclude that

$$c|\boldsymbol{r}_K| \leq \|a^{1/2} \operatorname{\mathbf{grad}} \chi\|_{\widetilde{K}} + \kappa \|a^{-1/2} \operatorname{\mathbf{curl}} \psi\|_{\widetilde{K}} + \operatorname{osc}(f, \widetilde{K}) + \operatorname{osc}(g, \mathcal{E}_N(K)),$$

and the result then follows as claimed. ∎

We are now in a position to give bounds on the norms of the jumps in the discontinuous Galerkin approximation in terms of the local conforming and nonconforming parts of the error.

LEMMA 4. *Suppose that $\kappa > 4\varrho(\boldsymbol{S}_K)$. Then*

$$\begin{aligned} |\gamma|^{-1/2} \|[U_{\mathcal{P}}]\|_\gamma \leq & C(\kappa - 4\varrho(\boldsymbol{S}_K))^{-1} \Big[\|a^{1/2} \operatorname{\mathbf{grad}} \chi\|_{\widetilde{K}} + \operatorname{osc}(f, \widetilde{K}) \\ & + \kappa \|a^{-1/2} \operatorname{\mathbf{curl}} \psi\|_{\widetilde{K}} + \operatorname{osc}(g, \mathcal{E}_N(K)) \Big] \end{aligned} \tag{34}$$

for $\gamma \in \mathcal{E}_I(K)$, and

$$\begin{aligned} |\gamma|^{-1/2} \|U_{\mathcal{P}} - u_D\|_\gamma \leq & C(\kappa - 4\varrho(\boldsymbol{S}_K))^{-1} \Big[\|a^{1/2} \operatorname{\mathbf{grad}} \chi\|_{\widetilde{K}} + \operatorname{osc}(f, \widetilde{K}) \\ & + \kappa \|a^{-1/2} \operatorname{\mathbf{curl}} \psi\|_{\widetilde{K}} + \operatorname{osc}(g, \mathcal{E}_N(K)) \Big] \end{aligned} \tag{35}$$

for $\gamma \in \mathcal{E}_D(K)$, where C is a positive constant independent of any mesh-size.

Proof. It is not difficult to show that for any function $F \in H^1(0, h)$ we have

$$h^{-1} \|F - \overline{F}\|_{L_2(0,h)}^2 \leq Ch \|F'\|_{L_2(0,h)}^2,$$

where the prime denotes the derivative, and $\overline{F}$ is the average value of F. Observing that the left-hand side may be rewritten as

$$h^{-1}\|F\|^2_{L_2(0,h)} - |\overline{F}|^2,$$

we deduce that

$$h^{-1}\|F\|^2_{L_2(0,h)} \le |\overline{F}|^2 + Ch\|F'\|^2_{L_2(0,h)}.$$

Choosing $F = [U_{\mathcal{P}}]$, we obtain

$$|\gamma|^{-1}\|[U_{\mathcal{P}}]\|^2_\gamma \le \left||\gamma|^{-1}\int_\gamma [U_{\mathcal{P}}]\ \mathrm{d}s\right|^2 + C|\gamma|\|[\partial U_{\mathcal{P}}/ds]\|^2_\gamma,$$

while choosing $F = U_{\mathcal{P}} - u_D$, we obtain

$$|\gamma|^{-1}\|U_{\mathcal{P}} - u_D\|^2_\gamma \le \left||\gamma|^{-1}\int_\gamma (U_{\mathcal{P}} - u_D)\,\mathrm{d}s\right|^2 + C|\gamma|\|\partial(U_{\mathcal{P}} - u_D)/ds\|^2_\gamma.$$

The assertions then follow at once from Lemmas 2 and 3. ∎

Finally, we come to the proof of Theorem 3.

Proof. The result follows at once from Theorem 1 and Lemma 4 on observing that $[e]$ coincides with $[U_{\mathcal{P}}]$ on an interior edge and with $U_{\mathcal{P}} - u_D$ on an edge $\gamma \in \mathcal{E}_D$. ∎

5. Estimation of conforming error.

5.1. Equilibrated fluxes. Given the finite element approximation $U_{\mathcal{P}}$, we introduce a set of piecewise linear *flux functions* $\{g_K : K \in \mathcal{P}\}$ on the element boundaries $g_K : \partial K \to \mathbb{R}$ as follows:

$$g_{K|\gamma} = \begin{cases} \mu_K\left(\langle\sigma_\nu(U_{\mathcal{P}})\rangle - \kappa|\gamma|^{-1}\,[U_{\mathcal{P}}]\right) & \text{on } \gamma \in \mathcal{E}_I(K), \\ \sigma_\nu(U_{\mathcal{P}}) - \kappa|\gamma|^{-1}(U_{\mathcal{P}} - u_D) & \text{on } \gamma \in \mathcal{E}_D(K), \\ \overline{g}_\gamma & \text{on } \gamma \in \mathcal{E}_N(K), \end{cases} \tag{36}$$

where $\overline{g}_\gamma$ denotes the average value of g on edge γ.

The fluxes $\{g_K\}$ have the following useful properties, referred to as *equilibration conditions* in [3].

LEMMA 5. *Let $\{g_K : K \in \mathcal{P}\}$ be defined as in* (36). *Then*

$$\sum_{K\in\mathcal{P}}\int_{\partial K} g_K v\,\mathrm{d}s = \int_{\Gamma_N} gv\,\mathrm{d}s - \sum_{\gamma\in\mathcal{E}_N}\int_\gamma (g - \overline{g}_\gamma)v\,\mathrm{d}s \quad \forall v \in H^1_E(\Omega) \tag{37}$$

and

$$\int_{\partial K} g_K\,\mathrm{d}s + \int_K f\,\mathrm{d}\boldsymbol{x} = 0 \tag{38}$$

for each element $K \in \mathcal{P}$.

Proof. The presence of the sign function μ_K in definition (36) means that, on any given interior edge $\gamma \in \mathcal{E}_I(K) \cap \mathcal{E}_I(K')$, the fluxes satisfy $g_K + g_{K'} = 0$. The first identity is an easy consequence of this fact, the definition of the fluxes on Γ_N, and the vanishing of the trace of the test function v on Γ_D. Let χ_K denote the piecewise constant function supported on element K, where it takes the value unity. The second

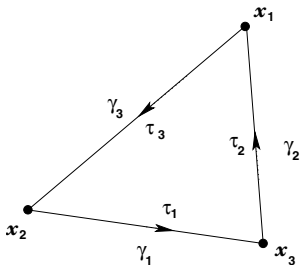


FIG. 3. *Locations of vertices and associated tangent vectors on triangle.*

assertion follows by inserting the expression for the flux function and simplifying to obtain

$$\begin{aligned}
&\int_{\partial K} g_K \,\mathrm{d}s + \int_K f \,\mathrm{d}\boldsymbol{x} \\
&= \sum_{\gamma\in\mathcal{E}_N} \int_\gamma g\chi_K \,\mathrm{d}s + (f,\chi_K) + \sum_{\gamma\in\mathcal{E}_D} \kappa|\gamma|^{-1} \int_\gamma u_D\chi_K \,\mathrm{d}s \\
&\quad + \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \int_\gamma \langle\sigma_\nu(U_\mathcal{P})\rangle\,[\chi_K] - \sum_{\gamma\in\mathcal{E}_I\cup\mathcal{E}_D} \kappa|\gamma|^{-1} \int_\gamma [U_\mathcal{P}]\,[\chi_K] \\
&= \mathcal{L}_\mathcal{P}(\chi_K) - \mathcal{B}_\mathcal{P}(U_\mathcal{P},\chi_K) = 0,
\end{aligned}$$

where we have used the facts that $[\chi_K] = \mu_K\chi_K$, that $\mathbf{grad}_\mathcal{P}\,\chi_K$ and $\langle\sigma_\nu(\chi_K)\rangle$ both vanish, and that $U_\mathcal{P}$ satisfies (9). ∎

The data appearing in the definition of the conforming part of the error may be rewritten, using (2), as

$$(a\,\mathbf{grad}\,\chi, \mathbf{grad}\,v) = (f,v) + \int_{\Gamma_N} gv \,\mathrm{d}s - (a\,\mathbf{grad}_\mathcal{P}\,U_\mathcal{P}, \mathbf{grad}\,v) \quad \forall v \in H^1_E(\Omega). \tag{39}$$

Then, exploiting property (37), we may split the right-hand side into contributions from individual elements, arriving at the identity

$$\begin{aligned}
&(a\,\mathbf{grad}\,\chi, \mathbf{grad}\,v) \\
&= \sum_{K\in\mathcal{P}} \left\{ (\overline{f}_K, v)_K + \int_{\partial K} g_K v \,\mathrm{d}s - (a\,\mathbf{grad}\,U_\mathcal{P}, \mathbf{grad}\,v)_K \right\} \\
&\quad + \sum_{K\in\mathcal{P}} (f-\overline{f}_K, v)_K + \sum_{\gamma\in\mathcal{E}_N} \int_\gamma (g-\overline{g}_\gamma) v \,\mathrm{d}s \quad \forall v \in H^1_E(\Omega).
\end{aligned} \tag{40}$$

The next step is to construct a computable local representation of the functionals appearing in parentheses.

5.2. Construction of local representer. Let $K \in \mathcal{P}$ be any element and, without loss of generality, assume that the vertices and edges are enumerated as shown in Figure 3, where $\boldsymbol{\tau}_n$ are the nonnormalized tangent vectors shown. The following result gives an explicit construction for the local representer and a closed form for its norm.

LEMMA 6. *For $K \in \mathcal{P}$, let*

$$\boldsymbol{\rho}_K = \frac{1}{2|K|} \sum_{n \in \mathcal{N}(K)} \boldsymbol{\rho}_n^{(K)} \lambda_n, \tag{41}$$

where $|K|$ is the area of the element,

$$\begin{aligned} \boldsymbol{\rho}_1^{(K)} &= |\gamma_3| \Delta_3^{(K)}(\boldsymbol{x}_1)\boldsymbol{\tau}_2 - |\gamma_2| \Delta_2^{(K)}(\boldsymbol{x}_1)\boldsymbol{\tau}_3, \\ \boldsymbol{\rho}_2^{(K)} &= |\gamma_1| \Delta_1^{(K)}(\boldsymbol{x}_2)\boldsymbol{\tau}_3 - |\gamma_3| \Delta_3^{(K)}(\boldsymbol{x}_2)\boldsymbol{\tau}_1, \\ \boldsymbol{\rho}_3^{(K)} &= |\gamma_2| \Delta_2^{(K)}(\boldsymbol{x}_3)\boldsymbol{\tau}_1 - |\gamma_1| \Delta_1^{(K)}(\boldsymbol{x}_3)\boldsymbol{\tau}_2, \end{aligned} \tag{42}$$

and

$$\Delta_n^{(K)} = (g_K - \sigma_{\nu_K}(U_{\mathcal{P}}))_{|\gamma_n}, \quad \gamma_n \subset \mathcal{E}(K). \tag{43}$$

Then

$$(\boldsymbol{\rho}_K, \mathbf{grad}\, v)_K = (\overline{f}_K, v)_K + \int_{\partial K} g_K v \, \mathrm{d}s - (a \,\mathbf{grad}\, U_{\mathcal{P}}, \mathbf{grad}\, v)_K \quad \forall v \in H^1_E(K) \tag{44}$$

and

$$\|\boldsymbol{\rho}_K\|_K^2 = \frac{1}{48|K|} \left[\left| \sum_{n \in \mathcal{N}(K)} \boldsymbol{\rho}_n^{(K)} \right|^2 + \sum_{n \in \mathcal{N}(K)} \left| \boldsymbol{\rho}_n^{(K)} \right|^2 \right]. \tag{45}$$

Proof. Let $\boldsymbol{\nu}_1$ denote the unit normal on edge γ_1. Then

$$\boldsymbol{\nu}_1 \cdot \boldsymbol{\rho}_K = \frac{1}{2|K|} \left[\boldsymbol{\nu}_1 \cdot \boldsymbol{\rho}_2^{(K)} \lambda_2 + \boldsymbol{\nu}_1 \cdot \boldsymbol{\rho}_3^{(K)} \lambda_3 \right] \quad \text{on } \gamma_1.$$

Elementary algebra reveals that $\boldsymbol{\nu}_1 \cdot \boldsymbol{\rho}_n^{(K)} = 2|K| \Delta_1^{(K)}(\boldsymbol{x}_n)$ for $n = 2$ and 3, and hence

$$\boldsymbol{\nu}_1 \cdot \boldsymbol{\rho}_K = \Delta_1^{(K)}(\boldsymbol{x}_2)\lambda_2 + \Delta_1^{(K)}(\boldsymbol{x}_3)\lambda_3 \quad \text{on } \gamma_1.$$

Since $\Delta^{(K)}$ is piecewise linear on the element edges, we have $\boldsymbol{\nu}_1 \cdot \boldsymbol{\rho}_K = \Delta_1^{(K)}$ on γ_1. The same argument applies equally well to the remaining edges, and we conclude that

$$\boldsymbol{\nu}_K \cdot \boldsymbol{\rho}_K = g_K - \sigma_{\nu_K}(U_{\mathcal{P}}) \quad \text{on } \partial K.$$

Furthermore, since $\mathbf{div}\, \boldsymbol{\rho}_K$ is constant, we have

$$\mathbf{div}\, \boldsymbol{\rho}_K = \frac{1}{|K|} \int_K \mathbf{div}\, \boldsymbol{\rho}_K \, \mathrm{d}\boldsymbol{x} = \frac{1}{|K|} \int_{\partial K} \boldsymbol{\nu}_K \cdot \boldsymbol{\rho}_K \, \mathrm{d}s.$$

Then inserting the expression for the normal components and integrating by parts, we obtain

$$\mathbf{div}\, \boldsymbol{\rho}_K = \frac{1}{|K|} \int_{\partial K} g_K \, \mathrm{d}s - \frac{1}{|K|} \int_K \mathbf{div}\, \boldsymbol{\sigma}(U_{\mathcal{P}}) \, \mathrm{d}\boldsymbol{x},$$

and, thanks to (38), we deduce that

$$-\mathbf{div}\, \boldsymbol{\rho}_K = \overline{f}_K + \mathbf{div}\, \boldsymbol{\sigma}(U_{\mathcal{P}}) \quad \text{on } K.$$

Given $v \in H^1(K)$, we have

$$(\boldsymbol{\rho}_K, \mathbf{grad}\, v)_K = \int_{\partial K} v \boldsymbol{\nu}_K \cdot \boldsymbol{\rho}_K \,\mathrm{d}s - (v, \mathbf{div}\, \boldsymbol{\rho}_K)_K,$$

and by again inserting the expressions for the normal components and divergence of $\boldsymbol{\rho}_K$, integrating by parts, and simplifying, we arrive at the identity (44). Finally, recalling that

$$\int_K \lambda_m \lambda_n \,\mathrm{d}\boldsymbol{x} = \frac{1}{12}|K|(1+\delta_{mn}), \quad m, n \in \mathcal{N}(K),$$

we have

$$\|\boldsymbol{\rho}_K\|_K^2 = \frac{1}{48|K|} \sum_{m\in\mathcal{N}(K)} \sum_{n\in\mathcal{N}(K)} (1+\delta_{mn}) \boldsymbol{\rho}_m^{(K)} \cdot \boldsymbol{\rho}_n^{(K)},$$

which simplifies to give (45). ∎

The following result gives an alternative representer having in general a smaller norm than the previous one.

LEMMA 7. *Let $K \in \mathcal{P}$ and define $\boldsymbol{\rho}_K$ as in Lemma 6. Let $\beta \in H_0^1(K)$ be arbitrary. Then*

$$\boldsymbol{\rho}_K^\star = \boldsymbol{\rho}_K - c_K^\star \,\mathbf{curl}\, \beta, \tag{46}$$

where

$$c_K^\star = \frac{\int_K \beta \,\mathrm{d}\boldsymbol{x}}{\|\mathbf{curl}\, \beta\|_K^2} \operatorname{curl} \boldsymbol{\rho}_K \tag{47}$$

satisfies (44) *(with $\boldsymbol{\rho}_K$ replaced by $\boldsymbol{\rho}_K^\star$) and*

$$\|\boldsymbol{\rho}_K^\star\|_K^2 = \|\boldsymbol{\rho}_K\|_K^2 - \mathcal{C}_K^\star(\beta)^2 |K| \|\operatorname{curl} \boldsymbol{\rho}_K\|_K^2, \tag{48}$$

where

$$\mathcal{C}_K^\star(\beta) = \frac{\int_K \beta \,\mathrm{d}\boldsymbol{x}}{|K| \|\mathbf{curl}\, \beta\|_K}. \tag{49}$$

Proof. The fact that $\boldsymbol{\rho}_K^\star$ satisfies the identity follows at once from Lemma 6 after noting that $\operatorname{curl} \boldsymbol{\rho}_K$ is constant and then observing that $(\mathbf{curl}\, \beta, \mathbf{grad}\, v)_K$ vanishes for $v \in H^1(K)$. Equally well, again using the fact that $\operatorname{curl} \boldsymbol{\rho}_K$ is constant, we find

$$\operatorname{curl} \boldsymbol{\rho}_K \int_K \beta \,\mathrm{d}\boldsymbol{x} = \int_K \beta \operatorname{curl} \boldsymbol{\rho}_K \,\mathrm{d}\boldsymbol{x} = (\boldsymbol{\rho}_K, \mathbf{curl}\, \beta)_K, \tag{50}$$

and so $c_K^\star = (\boldsymbol{\rho}_K, \mathbf{curl}\, \beta)_K / \|\mathbf{curl}\, \beta\|_K^2$. Direct calculation then gives

$$\|\boldsymbol{\rho}_K^\star\|_K^2 = \|\boldsymbol{\rho}_K\|_K^2 - \left[\frac{(\boldsymbol{\rho}_K, \mathbf{curl}\, \beta)_K}{\|\mathbf{curl}\, \beta\|_K} \right]^2$$

and the result then follows thanks to (50). ∎

Note that the quantity $\mathcal{C}_K^\star(\beta)$ depends on the shape of the element K but not on its size. The choice of the bubble function β was discussed in section 3.

5.3. Bounds on the conforming error. With the above representation result in hand, we resume the argument following (40) by inserting the representation (44) for the terms in parentheses, giving

$$\begin{aligned}&(a\,\mathbf{grad}\,\chi,\mathbf{grad}\,v)\\ &=\sum_{K\in\mathcal{P}}(\boldsymbol{\rho}_K^\star,\mathbf{grad}\,v)_K+\sum_{K\in\mathcal{P}}(f-\overline{f}_K,v)_K+\sum_{\gamma\in\mathcal{E}_N}(g-\overline{g}_\gamma,v)_\gamma\end{aligned}\tag{51}$$

for all $v\in H^1_E(\Omega)$. The following result gives a computable upper bound on the conforming error and a local lower bound up to a positive constant independent of any mesh-size.

LEMMA 8. *Let $K\in\mathcal{P}$ and define $\boldsymbol{\rho}_K$ as above. Then*

$$(a\,\mathbf{grad}\,\chi,\mathbf{grad}\,\chi)\le\sum_{K\in\mathcal{P}}\eta_{\mathsf{CF},K}(\beta)^2,\tag{52}$$

where

$$\begin{aligned}\eta_{\mathsf{CF},K}(\beta)&=a_K^{-1/2}\sqrt{\|\boldsymbol{\rho}_K\|_K^2-\mathcal{C}_K^\star(\beta)^2|K|\|\mathrm{curl}\,\boldsymbol{\rho}_K\|_K^2}\\ &+\mathcal{C}_p(K)a_K^{-1/2}\|f-\overline{f}_K\|_K+\sum_{\gamma\in\mathcal{E}_N(K)}\mathcal{C}_t(K,\gamma)a_K^{-1/2}\|g-\overline{g}_\gamma\|_\gamma,\end{aligned}\tag{53}$$

with $\mathcal{C}_K^\star$, $\mathcal{C}_p$, and $\mathcal{C}_t$ as defined in Lemma 7, Theorem 4, and Lemma 11, respectively. Moreover,

$$c\eta_{\mathsf{CF},K}(\beta)\le\|a^{1/2}\,\mathbf{grad}\,\chi\|_{\widetilde{K}}+\|a^{-1/2}\,\mathbf{curl}\,\psi\|_{\widetilde{K}}+\mathrm{osc}(g,\{\gamma\in\mathcal{E}_N(K)\})+\mathrm{osc}(f,\widetilde{K}),\tag{54}$$

where c is a positive constant independent of any mesh-size.

Proof. The first term in (51) is simply estimated using

$$(\boldsymbol{\rho}_K^\star,\mathbf{grad}\,v)_K\le\|\boldsymbol{\rho}_K^\star\|_K\|\mathbf{grad}\,v\|_K,$$

while the remaining terms are estimated using Theorem 4 and Lemma 11 to obtain for each element $K\in\mathcal{P}$

$$(f-\overline{f}_K,v)_K=\inf_{c\in\mathbb{R}}(f-\overline{f}_K,v-c)_K\le\mathcal{C}_p(K)\|f-\overline{f}_K\|_K\|\mathbf{grad}\,v\|_K$$

and

$$(g-\overline{g}_\gamma,v)_\gamma=\inf_{c\in\mathbb{R}}(g-\overline{g}_K,v-c)_\gamma\le\mathcal{C}_t(K,\gamma)\|g-\overline{g}_\gamma\|_\gamma\|\mathbf{grad}\,v\|_K\quad\forall\gamma\in\mathcal{E}_N(K).$$

Choosing $v=\chi$ and applying the Cauchy–Schwarz inequality results in the claimed upper bound.

It suffices to prove the lower bound for the choice $\beta=0$. Let $K\in\mathcal{P}$ and note that

$$\|\boldsymbol{\rho}_K\|_K^2\le C|K|^{-1}\sum_{n\in\mathcal{N}(K)}|\boldsymbol{\rho}_n^{(K)}|^2$$

and

$$|\boldsymbol{\rho}_1^{(K)}|\le|\gamma_2||\gamma_3|[|\Delta_2^{(K)}(\boldsymbol{x}_1)|+|\Delta_3^{(K)}(\boldsymbol{x}_1)|]\le C|K|[\Delta_2^{(K)}(\boldsymbol{x}_1)^2+\Delta_3^{(K)}(\boldsymbol{x}_1)^2]^{1/2},$$

with similar bounds for $\boldsymbol{\rho}_2^{(K)}$ and $\boldsymbol{\rho}_3^{(K)}$. Since the flux $\Delta^{(K)}$ is piecewise linear, it follows that

$$\|\boldsymbol{\rho}_K\|_K^2 \leq C|K| \sum_{\gamma\in\mathcal{E}(K)} |\gamma|^{-1}\|\Delta^{(K)}\|_\gamma^2 \leq C \sum_{\gamma\in\mathcal{E}(K)} |\gamma|\|\Delta^{(K)}\|_\gamma^2.$$

Then

$$\Delta^{(K)} = g_K - \sigma_{\nu_K}(U_{\mathcal{P}}) = \begin{cases} -\frac{1}{2}\left[\sigma_\nu(U_{\mathcal{P}})\right] - \kappa|\gamma|^{-1}\mu_K\left[U_{\mathcal{P}}\right], & \gamma\in\mathcal{E}_I(K), \\ -\kappa|\gamma|^{-1}(U_{\mathcal{P}} - u_D), & \gamma\in\mathcal{E}_D(K), \\ g - \sigma_\nu(U_{\mathcal{P}}), & \gamma\in\mathcal{E}_N(K), \end{cases}$$

and hence, applying the triangle inequality and the estimates of Lemma 4, we may bound the quantity $|\gamma|^{1/2}\|g_K - \sigma_{\nu_K}\|_\gamma$ by

$$|\gamma|^{1/2}\|[\sigma_\nu(U_{\mathcal{P}})]\|_\gamma + \kappa\left||\gamma|^{-1}\int_\gamma [U_{\mathcal{P}}]\ \mathrm{d}s\right| + \kappa|\gamma|^{1/2}\|[\partial U_{\mathcal{P}}/ds]\|_\gamma$$

for edges $\gamma\in\mathcal{E}_I(K)$, by

$$\kappa\left||\gamma|^{-1}\int_\gamma (U_{\mathcal{P}} - u_D)\,\mathrm{d}s\right| + \kappa|\gamma|^{1/2}\|[\partial(U_{\mathcal{P}} - u_D)/ds]\|_\gamma$$

for edges $\gamma\in\mathcal{E}_D(K)$, and by

$$|\gamma|^{1/2}\|g - \sigma_\nu(U_{\mathcal{P}})\|_\gamma$$

for edges $\gamma\in\mathcal{E}_N(K)$. Squaring and summing these bounds over all edges γ of the element, then using the estimates in Lemmas 2 and 3, we deduce that

$$c\eta_{\mathsf{CF},K}(0) \leq \|a^{1/2}\,\mathbf{grad}\,\chi\|_{\widetilde{K}} + \|a^{-1/2}\,\mathbf{curl}\,\psi\|_{\widetilde{K}} + \mathrm{osc}(g,\{\gamma\in\mathcal{E}_N(K)\}) + \mathrm{osc}(f,\widetilde{K}),$$

where c is a positive constant that is independent of any mesh-size, since the shape regularity of the elements means that the diameter of the element and the lengths of its edges means that all such dimensions are equivalent up to constants that depend only on the shape of the element. ∎

6. Estimation of nonconforming error. The estimation of the nonconforming part of the error is based on the following identity:

$$(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi) = \min_{u^*\in H^1_{u_D}(\Omega)} (a\,\mathbf{grad}_{\mathcal{P}}(u^* - U_{\mathcal{P}}), \mathbf{grad}_{\mathcal{P}}(u^* - U_{\mathcal{P}})), \tag{55}$$

where $H^1_{u_D}(\Omega) = \{v\in H^1(\Omega) : v = u_D \text{ on } \Gamma_D\}$. A proof of this can be found in [1]. In order to make use of the bound, we construct an admissible function $u^*\in H^1_{u_D}(\Omega)$ by smoothing the finite element approximation $U_{\mathcal{P}}$. Specifically, take u^* to be a piecewise affine function $U^*_{\mathcal{P}}$ on $\mathcal{P}$ with nodal values given by

$$U^*_{\mathcal{P}}(\boldsymbol{x}_n) = \begin{cases} \dfrac{1}{\#\Omega_n}\displaystyle\sum_{K\subset\Omega_n} U_{\mathcal{P}|K}(\boldsymbol{x}_n) & \text{if } \boldsymbol{x}_n\notin\Gamma_D, \\ u_D(\boldsymbol{x}_n) & \text{if } \boldsymbol{x}_n\in\Gamma_D, \end{cases} \tag{56}$$

where Ω_n is the set of elements that have a vertex at the point $\boldsymbol{x}_n$, and $\#\Omega_n$ denotes its cardinality. Inserting the function into the above statement shows that the estimator

$$\eta^2_{\mathsf{NC},K} = (a\,\mathbf{grad}_{\mathcal{P}}(U^*_{\mathcal{P}} - U_{\mathcal{P}}), \mathbf{grad}_{\mathcal{P}}(U^*_{\mathcal{P}} - U_{\mathcal{P}}))_K \tag{57}$$

provides a computable upper bound on the nonconforming part of the error.

The next result asserts that the estimator provides two-sided bounds on the error.

LEMMA 9. *Let $\eta_{\mathsf{NC},K}$ denote the estimator defined in* (57). *Then*

$$(a^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi) \le \sum_{K\in\mathcal{P}} \eta^2_{\mathsf{NC},K} \tag{58}$$

and, for each element $K \in \mathcal{P}$,

$$c\eta_{\mathsf{NC},K} \le \|a^{1/2}\,\mathbf{grad}\,\chi\|_{K^*} + \|a^{-1/2}\,\mathbf{curl}\,\psi\|_{K^*} + \mathrm{osc}(g, \{\gamma \in \mathcal{E}_N(K^*)\}) + \mathrm{osc}(f, K^*), \tag{59}$$

where

$$K^* = \cup\{\widetilde{\gamma} : \mathcal{N}(K) \cap \mathcal{N}(\gamma) \text{ is nonempty}\} \tag{60}$$

and c is a positive constant independent of any mesh-size.

Proof. The upper bound follows from the foregoing arguments. A simple scaling argument shows that

$$\eta^2_{\mathsf{NC},K} \le Ca_K \sum_{n\in\mathcal{N}(K)} |U^*_{\mathcal{P}}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n)|^2.$$

We distinguish two cases.

Case (i). If $\boldsymbol{x}_n \notin \Gamma_D$, then

$$U^*_{\mathcal{P}}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n) = \frac{1}{\#\Omega_n} \sum_{K'\in\Omega_n} (U_{\mathcal{P}|K'}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n)).$$

If elements K and K' share a common edge γ, then

$$|U_{\mathcal{P}|K'}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n)| = |\,[U_{\mathcal{P}}]\,(\boldsymbol{x}_n)|$$

and, by writing $[U_{\mathcal{P}}]_\gamma$ in terms of its average value and (constant) gradient on the edge, we deduce that

$$|\,[U_{\mathcal{P}}]\,(\boldsymbol{x}_n)| \le \left|\frac{1}{|\gamma|}\int_\gamma [U_{\mathcal{P}}]\ \mathrm{d}s\right| + \frac{1}{2}|\gamma|^{1/2}\|[\partial U_{\mathcal{P}}/\partial s]\|_\gamma.$$

If K and K' are separated by intervening elements, we write the difference $U_{\mathcal{P}|K} - U_{\mathcal{P}|K'}$ as a telescoping sum of differences between neighboring elements and use the above estimate to deduce that

$$|U_{\mathcal{P}|K'}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n)| \le C \sum_{\gamma:\boldsymbol{x}_n\in\gamma} \left|\frac{1}{|\gamma|}\int_\gamma [U_{\mathcal{P}}]\ \mathrm{d}s\right| + \frac{1}{2}|\gamma|^{1/2}\|[\partial U_{\mathcal{P}}/\partial s]\|_\gamma$$

whenever K and $K' \in \Omega_n$.

Case (ii). If $\boldsymbol{x}_n \in \Gamma_D$, then $U^*_{\mathcal{P}}(\boldsymbol{x}_n) = u_D(\boldsymbol{x}_n)$. If K has an edge $\gamma \subset \Gamma_D$, then, arguing as in the previous case, we obtain

$$|U^*_{\mathcal{P}}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n)| \le \left| \frac{1}{|\gamma|} \int_\gamma (U_{\mathcal{P}} - u_D)\, \mathrm{d}s \right| + \frac{1}{2}|\gamma|^{1/2} \|\partial(U_{\mathcal{P}} - u_D)/\partial s\|_\gamma.$$

As before, if element K is separated from the Dirichlet boundary by intervening elements, then we write the difference as a telescoping sum of differences between neighboring elements and use the previous estimates.

Finally, making use of the above bounds and the estimates in Lemmas 2 and 3, we conclude that in both cases,

$$c|U^*_{\mathcal{P}}(\boldsymbol{x}_n) - U_{\mathcal{P}|K}(\boldsymbol{x}_n)|$$

$$\le \|a^{1/2}\, \mathbf{grad}\, \chi\|_{K^*} + \|a^{-1/2}\, \mathbf{curl}\, \psi\|_{K^*} + \mathrm{osc}(g, \{\gamma \in \mathcal{E}_N(K^*)\}) + \mathrm{osc}(f, K^*)$$

and the desired result is then a simple consequence of this estimate. ∎

Appendix. Some basic estimates. We collect some basic estimates needed in the upper bounds for the conforming part of the error in section 5. The following optimal Poincaré estimate is proved in [13].

THEOREM 4. *Let $K \in \mathcal{P}$ and $v \in H^1(K)$. Then*

$$\inf_{c \in \mathbb{R}} \|v - c\|_K \le \mathcal{C}_p(K) \|\mathbf{grad}\, v\|_K, \tag{61}$$

where $\mathcal{C}_p(K) = \frac{1}{\pi} \max_{\boldsymbol{x},\boldsymbol{y} \in K} |\boldsymbol{x} - \boldsymbol{y}|$.

The next estimate gives a bound for the L_2-norm of a function over an element edge in terms of the H^1-norm over the element. Again, the bound may be shown to be optimal.

LEMMA 10. *Let γ be any edge of a triangle $K \in \mathcal{P}$. Let $\boldsymbol{x}_\gamma$ denote the vertex of K opposite to edge γ and define $L_\gamma = \max_{\boldsymbol{x} \in \gamma} |\boldsymbol{x} - \boldsymbol{x}_\gamma|$ and $\ell_\gamma = \min_{\boldsymbol{x} \in \gamma} |\boldsymbol{x} - \boldsymbol{x}_\gamma|$. Then, for all $v \in H^1(K)$,*

$$\|v\|^2_\gamma \le \frac{2}{\ell_\gamma} \|v\|_K \left[\|v\|_K + L_\gamma \|\mathbf{grad}\, v\|_K \right]. \tag{62}$$

Proof. Let $\boldsymbol{\nu}$ denote the unit outward normal on ∂K, and observe that the quantity $\boldsymbol{\nu} \cdot (\boldsymbol{x} - \boldsymbol{x}_\gamma)$ equals ℓ_γ on γ and vanishes on the remaining portion of the boundary of K. Hence,

$$\ell_\gamma \|v\|^2_{\partial K} = \int_{\partial K} \boldsymbol{\nu} \cdot (\boldsymbol{x} - \boldsymbol{x}_\gamma) v^2\, \mathrm{d}s = \int_K \mathbf{div}[(\boldsymbol{x} - \boldsymbol{x}_\gamma) v^2]\, \mathrm{d}\boldsymbol{x}$$

$$= 2\|v\|^2_K + 2 \int_K v(\boldsymbol{x} - \boldsymbol{x}_\gamma) \cdot \mathbf{grad}\, v\, \mathrm{d}\boldsymbol{x}.$$

The result then follows by noting that the second term on the right-hand side is bounded by $2L_\gamma \|v\|_K \|\mathbf{grad}\, v\|_K$. In fact, we have proved a slightly stronger estimate whereby $\mathbf{grad}\, v$ could be replaced by the derivative of v in the direction of $\boldsymbol{x} - \boldsymbol{x}_\gamma$. A different argument used in [8] led to the same kind of estimate with a worse constant. ∎

The following simple corollary of the above results will be used in the main text.

LEMMA 11. *Let γ be any edge of a triangle $K \in \mathcal{P}$. Then*

$$\inf_{c\in\mathbb{R}} \|v - c\|_\gamma \leq \mathcal{C}_t(K,\gamma)\|\mathbf{grad}\, v\|_K, \tag{63}$$

where

$$\mathcal{C}_t(K,\gamma)^2 = \frac{2}{\ell_\gamma}\mathcal{C}_p(K)\left(\mathcal{C}_p(K) + L_\gamma\right) \tag{64}$$

and L_γ, ℓ_γ are given in Lemma 10, and $\mathcal{C}_p$ is given in Theorem 4.

REFERENCES

[1] M. AINSWORTH, *Robust a posteriori error estimation for nonconforming finite element approximation*, SIAM J. Numer. Anal., 42 (2005), pp. 2320–2341.

[2] M. AINSWORTH, *A synthesis of a posteriori error estimation techniques for conforming, nonconforming and discontinuous Galerkin finite element methods*, Contemp. Math., 383 (2005), pp. 1–14.

[3] M. AINSWORTH AND J. T. ODEN, *A posteriori error estimation in finite element analysis*, Pure Appl. Math., Wiley-Interscience, John Wiley & Sons, New York, 2000.

[4] D. N. ARNOLD, *An interior penalty finite element method with discontinuous elements*, SIAM J. Numer. Anal., 19 (1982), pp. 742–760.

[5] D. N. ARNOLD, F. BREZZI, B. COCKBURN, AND L. D. MARINI, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[6] G. A. BAKER, *Finite element methods for elliptic equations using nonconforming elements*, Math. Comp., 31 (1977), pp. 45–59.

[7] R. BECKER, P. HANSBO, AND M. G. LARSON, *Energy norm a posteriori error estimation for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 723–733.

[8] C. CARSTENSEN AND S. A. FUNKEN, *Constants in Clément-interpolation error and residual based a posteriori error estimates in finite element methods*, East-West J. Numer. Math., 8 (2000), pp. 153–175.

[9] B. COCKBURN, G. E. KARNIADAKIS, AND C.-W. SHU, *The development of discontinuous Galerkin methods*, in Discontinuous Galerkin Methods (Newport, RI, 1999), B. Cockburn and G. E. Karniadakis, eds., Lect. Notes Comput. Sci. Eng. 11, Springer, Berlin, 2000, pp. 3–50.

[10] E. DARI, R. DURAN, C. PADRA, AND V. VAMPA, *A posteriori error estimators for nonconforming finite element methods*, M2AN Math. Model. Numer. Anal., 30 (1996), pp. 385–400.

[11] A. ERN AND J. PROFT, *A posteriori discontinuous Galerkin error estimates for transient convection-diffusion equations*, Appl. Math. Lett., 18 (2005), pp. 833–841.

[12] O. A. KARAKASHIAN AND F. PASCAL, *A posteriori error estimates for a discontinuous Galerkin approximation of second-order elliptic problems*, SIAM J. Numer. Anal., 41 (2003), pp. 2374–2399.

[13] L. E. PAYNE AND H. F. WEINBERGER, *An optimal Poincaré inequality for convex domains*, Arch. Rational Mech. Anal., 5 (1960), pp. 286–292.

[14] B. RIVIÈRE AND M. F. WHEELER, *A posteriori error estimates for a discontinuous Galerkin method applied to elliptic problems*, Comput. Math. Appl., 46 (2003), pp. 141–163.

[15] A. ROMKES, S. PRUDHOMME, AND J. T. ODEN, *A posteriori error estimation for a new stabilized discontinuous Galerkin method*, Appl. Math. Lett., 16 (2003), pp. 447–452.

[16] S. SUN AND M. F. WHEELER, *$L^2(H^1)$-norm a posteriori error estimation for discontinuous Galerkin approximations of reactive transport problems*, J. Sci. Comput., 22/23 (2005), pp. 501–530.

[17] R. VERFÜRTH, *A Review of a Posteriori Error Estimation and Adaptive Mesh Refinement Techniques*, Wiley-Teubner, Stuttgart, 1996.

[18] M. F. WHEELER, *An elliptic collocation-finite element method with interior penalties*, SIAM J. Numer. Anal., 15 (1978), pp. 152–161.

[19] J. M. YANG AND Y. P. CHEN, *A unified a posteriori error analysis for discontinuous Galerkin approximations of reactive transport equations*, J. Comput. Math., 24 (2006), pp. 425–434.

SIAM J. NUMER. ANAL.
Vol. 45, No. 4, pp. 1799–1800

ERRATUM: ON A HERMITE INTERPOLATION BY POLYNOMIALS OF TWO VARIABLES*

BORISLAV BOJANOV† AND YUAN XU‡

Abstract. We correct several errors in [*SIAM J. Numer. Anal.,* 39 (2002), pp. 1780–1793] caused by a formula wrongly copied from the literature.

Key words. Hermite interpolation, polynomials, two variables

DOI. 10.1137/060676027

On p. 1789 of [1], the expression $h_{k,0}(r)$ of the basic polynomials in the classical Hermite interpolation was wrongly copied from the literature. It should be

$$h_{k,0}(r) = \left(\frac{1-r}{2}\right)^{s+1} \frac{2^k}{k!} \sum_{j=0}^{s-k} \binom{s+j}{j} \left(\frac{1+r}{2}\right)^{k+j}.$$

This mistake then manifested in some of the formulas in section 3. The corrected formulas are given below:

(1) The formula for $p(r)$ in the statement of Corollary 3.3 should be

$$p(r) = \sum_{k=0}^{[n/2]} \frac{(-1)^k 2^k F_k}{k!} \sum_{j=0}^{[n/2]-k} \binom{[n/2]+j}{j} \times \left[\left(\frac{1-r}{2}\right)^{[n/2]+1} \left(\frac{1+r}{2}\right)^{k+j} + \left(\frac{1+r}{2}\right)^{[n/2]+1} \left(\frac{1-r}{2}\right)^{k+j} \right].$$

(2) The formula for the coefficients Λ_k in Theorem 3.6 should be

$$\Lambda_k = \frac{2\pi}{2m+1} \frac{(-1)^k}{k!} \sum_{j=0}^{s-k} \binom{s+j}{j} \sum_{i=0}^{[(s+1-k-j)/2]} \binom{s+1-k-j}{2i} \times \frac{\Gamma(k+j+1)\Gamma(i+1)}{2^{s+j+1}\Gamma(k+j+i+2)}.$$

Also, on p. 1790, the expression following the product sign in the formula for $g(r)$ should be $r^2 - r_j^2$. We are grateful to M. Gachpazan and S. Serajzadeh for detecting these errors and giving the correct formulas.

We also take this opportunity to mention again that the proof of Theorem 2.5 covers only the case when all $\alpha_1, \ldots, \alpha_\lambda$ are equal. This was pointed out to us by H. Hakopian and was already acknowledged in the paper [2]. For further development in this direction, see [3, 4, 5].

*Received by the editors November 27, 2006; accepted for publication (in revised form) March 16, 2007; published electronically August 24, 2007.
http://www.siam.org/journals/sinum/45-4/67602.html
†Department of Mathematics, University of Sofia, Blvd. James Boucher 5, 1164 Sofia, Bulgaria (boris@fmi.uni-sofia.bg).
‡Department of Mathematics, University of Oregon, Eugene, OR 97403-1222 (yuan@uoregon.edu).

REFERENCES

[1] B. Bojanov and Y. Xu, *On a Hermite interpolation by polynomials of two variables*, SIAM J. Numer. Anal., 39 (2002), pp. 1780–1793.
[2] B. Bojanov and Y. Xu, *On polynomial interpolation of two variables*, J. Approx. Theory, 120 (2003), pp. 267–282.
[3] H. Hakopian and S. Ismail, *On Bojanov-Xu interpolation on conic sections*, East J. Approx. Theory, 9 (2003), pp. 251–267.
[4] M. Khalaf and H. Hakopian, *On the poisedness of Bojanov-Xu interpolation*, J. Approx. Theory, 115 (2005), pp. 11–28.
[5] M. Khalaf and H. Hakopian, *On the poisedness of Bojanov-Xu interpolation*, II, East J. Approx. Theory, 11 (2005), pp. 187–220.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1801–1813

A TWO-GRID METHOD OF A MIXED STOKES–DARCY MODEL FOR COUPLING FLUID FLOW WITH POROUS MEDIA FLOW*

MO MU† AND JINCHAO XU‡

Abstract. We study numerical methods for solving a coupled Stokes–Darcy problem in porous media flow applications. A two-grid method is proposed for decoupling the mixed model by a coarse grid approximation to the interface coupling conditions. Error estimates are derived for the proposed method. Both theoretical analysis and numerical experiments show the efficiency and effectiveness of the two-grid approach for solving multimodeling problems. Potential extensions and future directions are discussed.

Key words. porous media flow, Stokes equations, Darcy's law, multimodeling problems, two-grid method

AMS subject classifications. 65N15, 65N30, 76D07, 76S05

DOI. 10.1137/050637820

1. Introduction.

There are many multimodeling problems in real applications of complex systems. They consist of multiple models in different regions coupled through interface conditions. The local models may be very varied in type, scale, control variable, and many other physical and mathematical properties. The corresponding numerical treatments may, of course, also vary significantly in geometric and PDE discretization, algebraic solution, and so on, in order to cope with local properties. The mixture of coupled models also leads to various mathematical and numerical difficulties. For instance, interface coupling conditions involve different control variables from different local models and may have complex, or even nonlinear, forms. Coupling different models may lead to very singular and complex structures across the interface and strong stiffness due to different scales, which would present considerable numerical difficulties. Examples of coupled multimodel applications include viscous-inviscid flows [5], compressible-incompressible fluids [17], turbulent-laminar flows [9], viscous-porous media flows [11, 16, 21, 27], and inertial confinement fusion with high ratio of density and temperature [31].

In general, there are two types of approaches to solving multimodel problems. One is to solve coupled problems directly, and the other is to first decouple mixed models and then apply appropriate local solvers individually. There are many appealing reasons to use the decoupling approach. First, it allows one to tailor algorithm components flexibly and conveniently in terms of physical, mathematical, and numerical properties for each local model and solver. Second, it is suitable for today's grid computing environment because it can efficiently and effectively exploit the existing computing resources, including both hardware and software, that are distributed

*Received by the editors August 9, 2005; accepted for publication (in revised form) January 19, 2007; published electronically August 31, 2007.
http://www.siam.org/journals/sinum/45-5/63782.html

†Department of Mathematics, Hong Kong University of Science and Technology, Clearwater Bay, Kowloon, Hong Kong (mamu@ust.hk). This author's work was supported in part by Hong Kong RGC Competitive Earmarked Research grant HKUST6111/02P.

‡Department of Mathematics, Penn State University, University Park, PA 16802 (xu@math.psu.edu). This author's work was supported in part by NSF DMS-0209497 and NSF DMS-0215392. This author would also like to acknowledge the support of the Department of Mathematics, Hong Kong University of Science and Technology, during his visit to the department.

over the Internet and that have been developed by different experts for use in various application fields [26]. As a by-product, it naturally results in parallelism in the conventional sense.

There are various decoupling techniques. Many of them are in the spirit of domain decomposition in general. For instance, Quarteroni and Valli [29] have extensively investigated heterogeneous domain decomposition methods for various coupled models. The Lagrange multiplier approach is also widely used [15, 28] for decoupling multimodel problems. The interface relaxation approach [24, 25] has also been successfully applied in multimodel simulations. We note that two-grid methods were proposed in [34, 35] for discretizing nonsymmetric and indefinite PDEs. The approach was also used for linearizing nonlinear problems [23, 36, 37], for localization and parallelization [38, 39, 40], as well as for many other applications; see, for instance, Axelssson and coworkers [2, 3, 4], Girault and Lions [13], Layton and coworkers [18, 19, 20], and Utnes [32]. In this paper, we demonstrate that the two-grid approach can also be applied successfully to solve multimodel problems.

The rest of the paper is organized as follows. A coupled Stokes–Darcy model is described in the next section as our model problem. A two-grid algorithm is proposed in section 3 for decoupling the mixed model. The basic idea is to first solve a much smaller problem on a coarse grid. The coarse grid solution is then used to interpolate the interface condition, which leads to a decoupled problem on the fine grid. Section 4 contains the error analysis for the two-grid method and discusses its computational aspects as well as potential extensions and future directions. Both theoretical analysis and numerical experiments confirm that approximation accuracy does not deteriorate under the proposed two-grid decoupling technique so that the decoupled discrete problem is of the same accuracy as the couple discrete problem for approximating the mixed Stokes–Darcy model. Concluding remarks follow in section 5.

2. Coupled Stokes–Darcy model. Let us consider a mixed model of Stokes equations and Darcy equations for coupling a fluid flow with a porous media flow. There has been very active research done recently on its applications, mathematical analysis, finite element approximation, and numerical solution; see, e.g., [1, 11, 16, 21, 27] and references therein. In particular, a subdomain iterative method is proposed to decouple the Stokes–Darcy problem by applying the preconditioned Richardson–Franklin method to the interface equation with the Steklov–Poincaré pseudo-PDE operator [11].

We consider a fluid flow in Ω_f coupled with a porous media flow in Ω_p; see Figure 1, where Ω_f and Ω_p are two- or three-dimensional bounded domains, $\Omega_f \bigcap \Omega_p = \emptyset$, and $\overline{\Omega_f} \bigcap \overline{\Omega_p} = \Gamma$. Denote by $\Omega = \Omega_f \bigcup \Omega_p$, $\mathbf{n}_f$, and $\mathbf{n}_p$ as usual the unit outward normal directions on $\partial\Omega_f$ and $\partial\Omega_p$.

The fluid motion is governed by the Stokes equations for the velocity $\mathbf{V}_f$ and the pressure p_f: $\forall t > 0$,

$$(1) \qquad \begin{cases} \dfrac{\partial \mathbf{V}_f}{\partial t} - \mathrm{div}\mathbf{T}(\mathbf{V}_f, p_f) = \mathbf{g}_f & \forall \mathbf{x} \in \Omega_f \text{ (conservation of momentum)}, \\ \mathrm{div}\mathbf{V}_f = 0, & \forall \mathbf{x} \in \Omega_f \text{ (conservation of mass)}, \end{cases}$$

where

$$\mathbf{T}(\mathbf{V}_f, p_f) = -p_f \mathbf{I} + 2\mu \mathbf{D}(\mathbf{V}_f)$$

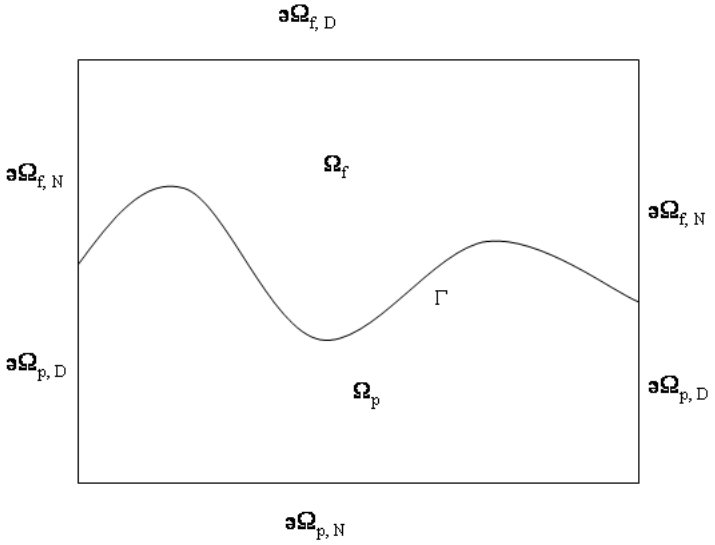


FIG. 1. *A global domain* Ω *consisting of a fluid region* Ω_f *and a porous media region* Ω_p *separated by an interface* Γ.

is the stress tensor, $\mu > 0$ is the kinematic viscosity, $\mathbf{g}_f$ is the external force, and

$$\mathbf{D}(\mathbf{V}_f) = \frac{1}{2}(\nabla \mathbf{V}_f + \nabla^T \mathbf{V}_f)$$

is the deformation rate tensor.

The porous media flow motion is governed by Darcy's law for the *piezometric head* ϕ and the *discharge vector* $\mathbf{q}$ that is proportional to the velocity $\mathbf{V}_p$, namely, $\mathbf{q} = n\mathbf{V}_p$ with n being the volumetric porosity: $\forall t > 0$,

$$\begin{cases} S_0 \dfrac{\partial \phi}{\partial t} + \operatorname{div} \mathbf{q} = g_p & \forall \mathbf{x} \in \Omega_p \text{ (conservation of mass)}, \\ \mathbf{q} = -\mathbf{K} \nabla \phi & \forall \mathbf{x} \in \Omega_p \text{ (Darcy's law)}, \end{cases} \tag{2}$$

where S_0 is the mass storativity, $\mathbf{K}$ is the hydraulic conductivity tensor of the porous medium, and the source g_p satisfies the solvability condition

$$\int_{\Omega_p} g_p = 0,$$

and

$$\phi = z + \frac{p_p}{\rho_f g},$$

where z is the elevation from a reference level, p_p is the pressure in Ω_p, ρ_f is the density, and g is the gravity acceleration.

We consider the following boundary conditions. Denote $\partial\Omega_f \setminus \Gamma = \partial\Omega_{f,D} \bigcup \partial\Omega_{f,N}$ and $\partial\Omega_p \setminus \Gamma = \partial\Omega_{p,D} \bigcup \partial\Omega_{p,N}$, as shown in Figure 1. For the fluid flow, we impose

$$\begin{cases} \mathbf{V}_f = 0 & \text{on } \partial\Omega_{f,D} \text{ with } meas(\partial\Omega_{f,D}) \neq 0, \\ -(\mathbf{T}(\mathbf{V}_f, p_f)) \cdot \mathbf{n}_f = \mathbf{h} & \text{on } \partial\Omega_{f,N}, \end{cases}$$

where $\mathbf{h}$ is a given vector. For the porous medium, we assume

$$\begin{cases} \phi = \phi_p & \text{on } \partial\Omega_{p,D}, \\ \mathbf{V}_p \cdot \mathbf{n}_p = v_p & \text{on } \partial\Omega_{p,N}. \end{cases}$$

A key part in a mixed model is the interface coupling conditions. The following interface conditions have been extensively used and studied in the literature [6, 21, 27]:

$$\begin{cases} \mathbf{V}_f \cdot \mathbf{n}_f + \mathbf{V}_p \cdot \mathbf{n}_p = 0, \\ -[(\mathbf{T}(\mathbf{V}_f, p_f)) \cdot \mathbf{n}_f] \cdot \mathbf{n}_f = \rho_f g \phi, \\ -[(\mathbf{T}(\mathbf{V}_f, p_f)) \cdot \mathbf{n}_f] \cdot \boldsymbol{\tau}_i = \dfrac{\alpha}{\sqrt{\boldsymbol{\tau}_i \cdot \mathbf{K} \cdot \boldsymbol{\tau}_i}} (\mathbf{V}_f - \mathbf{V}_p) \cdot \boldsymbol{\tau}_i, \quad i = 1, \ldots, d-1, \end{cases} \tag{3}$$

where $\{\boldsymbol{\tau}_i\}_{i=1}^{d-1}$ are linearly independent unit tangential vectors on Γ, d is the spacial dimension, and α is a positive parameter depending on the properties of the porous medium and must be experimentally determined. The first interface condition ensures mass conservation across Γ. The second one is a balance of normal forces across the interface. The third one states that the slip velocity along Γ is proportional to the shear stress along Γ. There have been many discussions in the literature on the slip condition along the interface. It is even unclear if the third condition in (3) leads to a well-posed problem. However, it has been observed that in practice the term $\mathbf{V}_p \cdot \boldsymbol{\tau}_i$ on the right-hand side from the porous media flow is much smaller than the other terms. The most accepted interface condition, known as the Beavers–Joseph–Saffman law, is then given by

$$-[2\mu \mathbf{D}(\mathbf{V}_f) \cdot \mathbf{n}_f] \cdot \boldsymbol{\tau}_i = \frac{\alpha}{\sqrt{\boldsymbol{\tau}_i \cdot \mathbf{K} \cdot \boldsymbol{\tau}_i}} \mathbf{V}_f \cdot \boldsymbol{\tau}_i, \ i = 1, \ldots, d-1, \tag{4}$$

which can be justified by a statistical approach and the Brinkman approximation [30]. We note that different interface conditions have been used in numerical studies. For instance, the Beavers–Joseph–Saffman condition is used in [1, 21], while the free-slip condition with $\alpha = 0$ is assumed in [10, 11, 12]. We will assume the Beavers–Joseph–Saffman condition (4) from now on.

For simplicity, let us assume n, ρ_f, and g are constants. We also assume the homogenous boundary condition on ϕ, $\phi_p = 0$, which can be easily handled by a lifting function in the nonhomogenous case.

Denote

$$\begin{aligned} H_f &= \{\mathbf{v} \in (H^1(\Omega_f))^d \mid \mathbf{v} = 0 \text{ on } \partial\Omega_{f,D}\}, \\ H_p &= \{\phi \in H^1(\Omega_p) \mid \phi = 0 \text{ on } \partial\Omega_{p,D}\}, \\ W &= H_f \times H_p, \\ Q &= L^2(\Omega_f). \end{aligned}$$

By integration by parts as in [21], the weak formulation for the above coupled (stationary) Stokes–Darcy problem reads as follows: For $f \in W'$, find $u = (\mathbf{u}, \phi) \in W$, $p \in Q$ such that

$$\begin{cases} a(u,v) + b(v,p) = f(v) & \forall v = (\mathbf{v}, \psi) \in W, \\ b(u,q) = 0 & \forall q \in Q, \end{cases} \tag{5}$$

where

$$a(u,v) = a_\Omega(u,v) + a_\Gamma(u,v),$$

with

$$\begin{aligned} a_\Omega(u,v) &= a_{\Omega_f}(\mathbf{u},\mathbf{v}) + a_{\Omega_p}(\phi,\psi), \\ a_{\Omega_f}(\mathbf{u},\mathbf{v}) &= \int_{\Omega_f} 2n\mu D(\mathbf{u}) \cdot D(\mathbf{v}) + \sum_{i=1}^{d-1} \frac{\alpha n}{\sqrt{\boldsymbol{\tau}_i \cdot \mathbf{K} \cdot \boldsymbol{\tau}_i}} \int_\Gamma (\mathbf{u} \cdot \boldsymbol{\tau}_i)(\mathbf{v} \cdot \boldsymbol{\tau}_i), \\ a_{\Omega_p}(\phi,\psi) &= \int_{\Omega_p} \rho_f g \nabla\psi \cdot \mathbf{K} \nabla\phi, \\ a_\Gamma(u,v) &= \int_\Gamma n\rho_f g[\phi\mathbf{v} - \psi\mathbf{u}] \cdot \mathbf{n}_f \end{aligned}$$

and with

$$b(v,p) \equiv b(\mathbf{v},p) = -\int_{\Omega_f} np \operatorname{div} \mathbf{v}.$$

Similarly to [10], it is easy to verify that (i) $a(\cdot,\cdot)$ is continuous and coercive on W, and that (ii) $b(\cdot,\cdot)$ is continuous on $W \times Q$ and satisfies the well-known Brezzi–Babuska condition as follows: There exists a positive constant $\beta > 0$ such that $\forall q \in Q, \exists w \in W$ such that

$$b(w,q) \geq \beta ||w||_W ||q||_Q. \tag{6}$$

The well-posedness of the model problem (5) then follows from Brezzi's theory for saddle-point problems [7]. The only difference from [10] is that the extension from the free-slip interface condition to the case of nonzero α results in the inclusion of an extra term $\sum_{i=1}^{d-1} \frac{\alpha n}{\sqrt{\boldsymbol{\tau}_i \cdot \mathbf{K} \cdot \boldsymbol{\tau}_i}} \int_\Gamma (\mathbf{u} \cdot \boldsymbol{\tau}_i)(\mathbf{v} \cdot \boldsymbol{\tau}_i)$ in the bilinear form $a_{\Omega_f}(\mathbf{u},\mathbf{v})$. Note that this extension does not affect property (i) for the bilinear form $a(\cdot,\cdot)$. The continuity is obvious, while the coercivity is still a consequence of the well-known Poincaré inequality and Korn inequality as in the free-slip case because α is positive and the corresponding term can thus be ignored in the estimation.

3. A two-grid algorithm. Let $W_h = H_{f,h} \times H_{p,h} \subset W$ and $Q_h \subset Q$ be two finite element spaces. The finite element discretization applied to the model problem (5) leads to a coupled discrete problem as follows: Find $u_h = (\mathbf{u_h}, \phi_h) \in W_h$, $p_h \in Q_h$ such that

$$\begin{cases} a(u_h, v_h) + b(v_h, p_h) = f(v_h) & \forall v_h = (\mathbf{v_h}, \psi_h) \in W_h, \\ b(u_h, q_h) = 0 & \forall q_h \in Q_h. \end{cases} \tag{7}$$

The construction of the finite element spaces W_h and Q_h is described more specifically as follows. Let the triangulation of the global domain be regular, as well as compatible and quasi-uniform on Γ as described in [11]. Furthermore, the finite element spaces $H_{f,h}$ and Q_h approximating the velocity and pressure fields in the fluid region are assumed to satisfy the discrete inf-sup condition as follows: There exists a positive constant $\beta^* > 0$, independent of h, such that $\forall \mathbf{v_h} \in H_{f,h}, q_h \in Q_h$,

$$b(\mathbf{v_h}, q_h) \geq \beta^* ||\mathbf{v_h}||_{H_f} ||q_h||_Q. \tag{8}$$

Several families of finite element spaces designed for the Stokes problem are provided in IV.2 and Chapter VI in [7]. They all satisfy the discrete inf-sup condition (8) and can thus be applied for $H_{f,h}$ and Q_h. Finally, standard finite element approximations of $H^m(\Omega_p)$, such as piecewise linear elements for $m = 1$, can be applied for $H_{p,h}$ in the porous media region. The well-posedness and error analysis of the coupled discrete model (7) can be found in [11].

We now propose a *two-grid algorithm* consisting of the following two steps.

ALGORITHM.

1. Solve a coarse grid problem (7) with spacing H as follows: Find $u_H = (\mathbf{u_H}, \phi_H) \in W_H \subset W_h$, $p_H \in Q_H \subset Q_h$ such that

$$\begin{cases} a(u_H, v_H) + b(v_H, p_H) = f(v_H) & \forall v_H = (\mathbf{v_H}, \psi_H) \in W_H, \\ b(u_H, q_H) = 0 & \forall q_H \in Q_H. \end{cases} \tag{9}$$

2. Solve a modified fine grid problem as follows: Find $u^h = (\mathbf{u^h}, \phi^h) \in W_h$, $p^h \in Q_h$ such that

$$\begin{cases} a_\Omega(u^h, v_h) + b(v_h, p^h) = f(v_h) - a_\Gamma(u_H, v_h) & \forall v_h \in W_h, \\ b(u^h, q_h) = 0 & \forall q_h \in Q_h. \end{cases} \tag{10}$$

It is easy to see that the modified fine grid problem (10) is also well-posed. More important, the discrete model (10) is in fact equivalent to two decoupled problems that correspond to the Stokes problem on Ω_f and the Darcy problem on Ω_p, respectively, with the boundary conditions defined by u_H on Γ. More specifically, the discrete Stokes problem on the fluid region reads as follows: Find $\mathbf{u^h} \in H_{f,h}$, $p^h \in Q_h$ such that

$$\begin{cases} a_{\Omega_f}(\mathbf{u}^h, \mathbf{v}_h) + b(\mathbf{v}_h, p^h) = (n\mathbf{g}_f, \mathbf{v}_h) - \displaystyle\int_\Gamma n\rho_f g \phi_H \mathbf{v}_h \cdot \mathbf{n}_f & \forall \mathbf{v}_h \in H_{f,h}, \\ b(\mathbf{u}^h, q_h) = 0 & \forall q_h \in Q_h. \end{cases} \tag{11}$$

Similarly, the discrete Darcy problem on the porous media region reads as follows: Find $\phi^h \in H_{p,h}$ such that

$$a_{\Omega_p}(\phi^h, \psi_h) = (\rho_f g g_p, \psi_h) + \int_\Gamma n\rho_f g \psi_h \mathbf{u}_H \cdot \mathbf{n}_f \quad \forall \psi_h \in H_{p,h}. \tag{12}$$

4. Error analysis. For convenience, from now on we will use $x \lesssim y$ to denote that there exists a constant C, such that $x \leq Cy$. Let W_h and Q_h be any finite element spaces as described in the previous section. In addition, for illustration assume the regularity $u \in (H^2(\Omega_f))^d \times H^2(\Omega_p)$ and $p \in H^1(\Omega_f)$, and thus finite element spaces as described above of first order approximation $O(h)$ are used for the fluid and

porous media regions. Then the error analysis for the coupled model in [11] yields the estimates

$$\begin{cases} ||u-u_h||_W \lesssim h, \\ ||p-p_h||_Q \lesssim h. \end{cases} \tag{13}$$

Note that estimates (13) apply to the coupled problem (7) but not to the decoupled problem (10). Furthermore, the extended framework of the Aubin–Nitsche duality technique [7] gives the following L^2-norm estimate.

LEMMA 1. *Let $W_- = (L^2(\Omega_f))^d \times L^2(\Omega_p)$. Then under the same assumptions as above, we have*

$$||u-u_h||_{W_-} \lesssim h^2. \tag{14}$$

Proof. As in the Aubin–Nitsche duality technique for the general framework of mixed problems in [7], consider the dual problem defined by the error pair $(u-u_h, p-p_h)$ to be (2.90) and (2.93) from [7]. For the solution (w,s) of the dual problem, from the regularity of the dual problem we have $(w,s) \in W_{++} \times Q_{++} = ((H^2(\Omega_f))^d \times H^2(\Omega_p)) \times H^1(\Omega_f)$ in the particular setting of our problem. Then, Theorem 2.2 (in particular the estimate of (2.100)) in [7] gives

$$||u-u_h||_{W_-} \lesssim m(h)(\|u-u_h\|_W + \|p-p_h\|_Q) + n(h)\|u-u_h\|_W,$$

where

$$\inf_{w_h \in W_h} \|w-w_h\|_W \le m(h)\|w\|_{W_{++}},$$

and

$$\inf_{q_h \in Q_h} \|s-q_h\|_Q \le n(h)\|s\|_{Q_{++}}.$$

Note that both $m(h)$ and $n(h)$ are of the order of $O(h)$ as shown in [7]. Estimate (14) then follows immediately from (13), which completes the proof. ∎

As a consequence, the following estimates, which will be used in the proof of the next theorem, follow immediately from (13) and (14):

$$\begin{cases} ||\mathbf{u}_h - \mathbf{u}_H||_{H_f} \lesssim H, \quad ||\mathbf{u}_h - \mathbf{u}_H||_{(L^2(\Omega_f))^d} \lesssim H^2, \\ \|\phi_h - \phi_H\|_{H_p} \lesssim H, \quad \|\phi_h - \phi_H\|_{L^2(\Omega_p)} \lesssim H^2. \end{cases} \tag{15}$$

THEOREM 2. *Let u_h, p_h and u^h, p^h be defined by the two discrete models* (7) *and* (10) *on the fine grid. The following error estimates hold:*

$$||\phi_h - \phi^h||_{H_p} \lesssim H^2, \tag{16}$$

$$||\mathbf{u}_h - \mathbf{u}^h||_{H_f} \lesssim H^{3/2}, \tag{17}$$

$$||p_h - p^h||_Q \lesssim H^{3/2}. \tag{18}$$

Proof. Note that by comparing the two discrete models (7) and (10) on the fine grid, we have

$$\begin{cases} a_\Omega(u_h - u^h, v_h) + a_\Gamma(u_h - u_H, v_h) + b(v_h, p_h - p^h) = 0 & \forall v_h \in W_h, \\ b(u_h - u^h, q_h) = 0 & \forall q_h \in Q_h. \end{cases} \tag{19}$$

First, taking $v_h = (\mathbf{0}, \psi_h) \in W_h$ in (19), we obtain

$$a_{\Omega_p}(\phi_h - \phi^h, \psi_h) + a_\Gamma(u_h - u_H, v_h) = 0.$$

In particular, when $\psi_h = \phi_h - \phi^h$, it is further reduced to

$$a_{\Omega_p}(\phi_h - \phi^h, \phi_h - \phi^h) = \int_\Gamma n\rho_f g(\phi_h - \phi^h)(\mathbf{u}_h - \mathbf{u}_H) \cdot \mathbf{n}_f.$$

Let $\theta \in H^1(\Omega_f)$ be a harmonic extension of $\phi_h - \phi^h$ to the fluid flow region, satisfying

$$\begin{cases} -\Delta\theta = 0 & \text{in } \Omega_f, \\ \theta = \phi_h - \phi^h & \text{on } \Gamma, \\ \theta = 0 & \text{on } \partial\Omega_f/\Gamma. \end{cases}$$

Let $H_{00}^{1/2}(\Gamma)$ denote the interpolation space [22]

$$H_{00}^{1/2}(\Gamma) = [L^2(\Gamma), H_0^1(\Gamma)]_{1/2}.$$

Apparently,

$$||\theta||_{H^1(\Omega_f)} \lesssim ||\phi_h - \phi^h||_{H_{00}^{1/2}(\Gamma)} \lesssim ||\phi_h - \phi^h||_{H_p}.$$

Note that $\forall q_H \in Q_H$,

$$\begin{aligned} &\int_\Gamma n\rho_f g(\phi_h - \phi^h)(\mathbf{u}_h - \mathbf{u}_H) \cdot \mathbf{n}_f \\ &\quad = \int_{\partial\Omega_f} n\rho_f g\theta(\mathbf{u}_h - \mathbf{u}_H) \cdot \mathbf{n}_f \\ &\quad = \int_{\Omega_f} \mathrm{div}(\mathbf{u}_h - \mathbf{u}_H)(n\rho_f g\theta) + \int_{\Omega_f} (\mathbf{u}_h - \mathbf{u}_H) \cdot \nabla(n\rho_f g\theta) \\ &\quad = n\rho_f g \left(\int_{\Omega_f} (\theta - q_H)\mathrm{div}(\mathbf{u}_h - \mathbf{u}_H) + \int_{\Omega_f} (\mathbf{u}_h - \mathbf{u}_H) \cdot \nabla\theta \right), \end{aligned}$$

where in the last equality we use the discrete divergence-free property for $\mathbf{u}_h$ and $\mathbf{u}_H$,

$$b(u_h - u_H, q_H) = \int_{\Omega_f} nq_H \mathrm{div}(\mathbf{u}_h - \mathbf{u}_H) = 0 \ \forall q_H \in Q_H.$$

Therefore, we have

$$\begin{aligned} &||\phi_h - \phi^h||_{H_p}^2 \\ &\quad \lesssim a_{\Omega_p}(\phi_h - \phi^h, \phi_h - \phi^h) \\ &\quad \lesssim \inf_{\forall q_H \in Q_H} \left| \int_{\Omega_f} (\theta - q_H)\mathrm{div}(\mathbf{u}_h - \mathbf{u}_H) \right| + \left| \int_{\Omega_f} (\mathbf{u}_h - \mathbf{u}_H) \cdot \nabla\theta \right| \\ &\quad \lesssim ||\mathbf{u}_h - \mathbf{u}_H||_{H_f} \inf_{\forall q_H \in Q_H} ||\theta - q_H||_{L^2(\Omega_f)} + ||\mathbf{u}_h - \mathbf{u}_H||_{(L^2(\Omega_f))^d} ||\theta||_{H^1(\Omega_f)} \\ &\quad \lesssim (H||\mathbf{u}_h - \mathbf{u}_H||_{H_f} + ||\mathbf{u}_h - \mathbf{u}_H||_{(L^2(\Omega_f))^d}) ||\theta||_{H^1(\Omega_f)} \\ &\quad \lesssim H^2 ||\phi_h - \phi^h||_{H_{00}^{1/2}(\Gamma)} \\ &\quad \lesssim H^2 ||\phi_h - \phi^h||_{H_p}, \end{aligned}$$

which leads to estimate (16).

To show (17), taking $v_h = (\mathbf{v}_h, 0) \in W_h$ in (19), we obtain

$$a_{\Omega_f}(\mathbf{u}_h - \mathbf{u}^h, \mathbf{v}_h) + a_\Gamma(u_h - u_H, v_h) + b(v_h, p_h - p^h) = 0.$$

In particular, when $\mathbf{v}_h = \mathbf{u}_h - \mathbf{u}^h$, due to the discrete divergence-free property of $\mathbf{u}_h$ and $\mathbf{u}^h$ so that $b(\mathbf{u}_h - \mathbf{u}^h, p_h - p^h) = 0$, we further have

$$a_{\Omega_f}(\mathbf{u}_h - \mathbf{u}^h, \mathbf{u}_h - \mathbf{u}^h) = \int_\Gamma n\rho_f g(\phi_h - \phi_H)(\mathbf{u}_h - \mathbf{u}^h) \cdot \mathbf{n}_f.$$

Hence,

$$\begin{aligned} ||\mathbf{u}_h - \mathbf{u}^h||^2_{H_f} &\lesssim a_{\Omega_f}(\mathbf{u}_h - \mathbf{u}^h, \mathbf{u}_h - \mathbf{u}^h) \\ &= \int_\Gamma n\rho_f g(\phi_h - \phi_H)(\mathbf{u}_h - \mathbf{u}^h) \cdot \mathbf{n}_f \\ &\lesssim \|\phi_h - \phi_H\|_{L^2(\Gamma)} \|\mathbf{u}_h - \mathbf{u}^h\|_{(L^2(\Gamma))^d} \\ &\lesssim \|\phi_h - \phi_H\|_{L^2(\Gamma)} \|\mathbf{u}_h - \mathbf{u}^h\|_{H_f}. \end{aligned} \tag{20}$$

Using a refined trace result (see [33, p. 27], with $\epsilon = H^{1/2}$), we get

$$\|\phi_h - \phi_H\|_{L^2(\Gamma)} \lesssim H^{-1/2}\|\phi_h - \phi_H\|_{L^2(\Omega_p)} + H^{1/2}\|\phi_h - \phi_H\|_{H^1(\Omega_p)} \lesssim H^{3/2}. \tag{21}$$

Applying (21) to (20) then yields estimate (17).

Finally, let us show (18). From the discrete Brezzi–Babuska condition on Ω_f, for $q_h = p_h - p^h \in Q_h, \exists \mathbf{v}_h \in H_{f,h}$ such that

$$||p_h - p^h||_{L^2(\Omega_f)} \lesssim \frac{-\int_{\Omega_f} n(p_h - p^h)\mathrm{div}\mathbf{v}_h}{||\mathbf{v}_h||_{H_f}}.$$

Recall that for $v_h = (\mathbf{v}_h, 0) \in W_h$ in (19), we have

$$a_{\Omega_f}(\mathbf{u}_h - \mathbf{u}^h, \mathbf{v}_h) + a_\Gamma(u_h - u_H, v_h) + b(v_h, p_h - p^h) = 0.$$

The first term above is easy to handle by

$$|a_{\Omega_f}(\mathbf{u}_h - \mathbf{u}^h, \mathbf{v}_h)| \lesssim ||\mathbf{u}_h - \mathbf{u}^h||_{H_f} ||\mathbf{v}_h||_{H_f}.$$

For the second term, we have

$$\begin{aligned} |a_\Gamma(u_h - u_H, v_h)| &= \left| \int_\Gamma n\rho_f g(\phi_h - \phi_H)\mathbf{v}_h \cdot \mathbf{n}_f \right| \\ &\lesssim \|\phi_h - \phi_H\|_{L^2(\Gamma)} \|\mathbf{v}_h\|_{(L^2(\Gamma))^d} \\ &\lesssim \|\phi_h - \phi_H\|_{L^2(\Gamma)} \|\mathbf{v}_h\|_{H_f}. \end{aligned}$$

Using (21) and (17), we have

$$\begin{aligned} ||p_h - p^h||_{L^2(\Omega_f)} &\lesssim \frac{|a_{\Omega_f}(\mathbf{u}_h - \mathbf{u}^h, \mathbf{v}_h)| + |a_\Gamma(u_h - u_H, v_h)|}{||\mathbf{v}_h||_{H_f}} \\ &\lesssim ||\mathbf{u}_h - \mathbf{u}^h||_{H_f} + \|\phi_h - \phi_H\|_{L^2(\Gamma)} \\ &\lesssim H^{3/2}, \end{aligned}$$

which leads to estimate (18). This completes the proof. ∎

COROLLARY 3. *Let $(u^h, p^h) \in W_h \times Q_h$ be the solution of the two-grid algorithm with $H = \sqrt{h}$. We have*

$$||\phi - \phi^h||_{H_p} \lesssim h \tag{22}$$

and

$$||\mathbf{u} - \mathbf{u}^h||_{H_f} + ||p - p^h||_Q \lesssim h^{3/4}. \tag{23}$$

If $H = h^{2/3}$, estimate (23) *is further improved to the optimal order as follows:*

$$||\mathbf{u} - \mathbf{u}^h||_{H_f} + ||p - p^h||_Q \lesssim h. \tag{24}$$

We remark that error estimates (17) and (18) for $\mathbf{u}_h - \mathbf{u}^h$ and $p_h - p^h$ may not be optimal due to technical reasons. These two estimates might be further improved to $O(H^2)$ by a finer analysis, as suggested by numerical experiments in [8], which could then lead to an improvement of (23) to an optimal estimate of the order of $O(h)$ for $\mathbf{u} - \mathbf{u}^h$ and $p - p^h$, yet still with $H = \sqrt{h}$. Furthermore, the error analysis may be extended to finite element spaces with higher order approximation $O(h^m)$, provided that the solution is locally smooth enough within each subdomain. Specifically, if $W_h \subset W$ and $Q_h \subset Q$ are finite element spaces with the approximation order $O(h^m)$, and the solution (u, p) is locally smooth enough within each subdomain, we expect the following estimates to hold:

$$||\phi_h - \phi^h||_{H_p} \lesssim H^{m+1} \tag{25}$$

and

$$||\mathbf{u_h} - \mathbf{u}^h||_{H_f} + ||p_h - p^h||_Q \lesssim H^{m+1}, \tag{26}$$

which implies the optimal error estimates if we take $H = h^{\frac{m}{m+1}}$:

$$||\phi - \phi^h||_{H_p} \lesssim h^m \tag{27}$$

and

$$||\mathbf{u} - \mathbf{u}^h||_{H_f} + ||p - p^h||_Q \lesssim h^m. \tag{28}$$

We refer readers to [8] for more details on this extension.

Comprehensive numerical experiments on various aspects of the proposed theoretical framework are under investigation and will be reported in [8]. For instance, if the well-known Taylor–Hood elements [7], also known as the P2-P1 elements, are applied to the Stokes model, and the P2 elements are applied to the Darcy model, and for convenience we simply take $H = \sqrt{h}$, the numerical approximations of the two-grid algorithm to a locally very smooth solution clearly demonstrate an optimal convergence rate of $O(h^2)$, which confirms our theoretical expectation. For more details, see [8].

Most important, the presented theory suggests that one can effectively and efficiently decouple a coupled multimodel problem by proper multigrid techniques. This allows for different submodel problems to be solved independently by applying the most appropriate numerical techniques individually. In addition, these decoupled local problems can be solved by different processors on a parallel multiprocessor or

by different computing nodes on a traditional cluster or even a remotely distributed computational grid. Furthermore, in a grid computing environment, powerful and efficient local solvers are usually available which were developed at different sites by different experts for various single models. Therefore, substantial coding tasks can also be reduced thanks to resource sharing in grid computing.

We also remark that the proposed two-grid algorithm still requires a coarse grid solver for the coupling purpose. The coarse grid problem usually has a much smaller size, say $H = \sqrt{h}$, and can thus be solved on a front end machine or a client machine. It is also numerically easier to solve than a fine grid problem in various aspects such as approximation accuracy, stiffness, and so on.

In addition, iterative strategies such as preconditioned error correction can be applied for the coarse grid solver by restricting the computed fine grid approximation to the coarse grid so that the coarse grid problem is also similarly decoupled. This then leads to a fully decoupled iterative two-grid algorithm. Finally, we remark that the same strategy can be applied recursively to the coarse grid problem, if necessary, which then leads to a multigrid algorithm.

5. Conclusions. We have proposed a two-grid method for solving the coupled Stokes–Darcy problem. Error estimates are obtained, which suggests that multigrid can provide a general framework for solving multimodeling problems. It is promising to extend this approach to more general settings, such as other boundary and interface conditions, Navier–Stokes/Darcy coupling, time-dependent problems, as well as other coupling applications. It is also possible to generalize the framework to other versions, including iterative two-grid methods and multilevel methods.

Acknowledgments. The authors would like to thank M. C. Cai for implementing the two-grid algorithm and conducting the numerical experiments. They also thank the referees very much for helpful comments and suggestions, which led to substantial improvements in the presentation.

REFERENCES

[1] T. Arbogast and D. S. Brunson, *A computational method for approximating a Darcy–Stokes system governing a vuggy porous medium*, Comput. Geosci., to appear.

[2] O. Axelsson and I. E. Kaporin, *Minimum residual adaptive multilevel finite element procedure for the solution of nonlinear stationary problems*, SIAM J. Numer. Anal., 35 (1998), pp. 1213–1229.

[3] O. Axelsson and W. Layton, *A two-level method for the discretization of nonlinear boundary value problems*, SIAM J. Numer. Anal., 33 (1996), pp. 2359–2374.

[4] O. Axelsson and A. Padiy, *On a two level Newton type procedure applied for solving nonlinear elasticity problems*, Internat. J. Numer. Methods Engrg., 49 (2000), pp. 1479–1493.

[5] D. Barberis and P. Molton, *Shock Wave/Turbulent Boundary Layer Interaction in a Three-Dimensional Flow*, AIAA paper 1995-227, American Institute of Aeronautics and Astronautics, Inc., Reston, VA, 1995.

[6] G. Beavers and D. Josephn, *Boundary conditions at a naturally permeable wall*, J. Fluid Mech., 30 (1967), pp. 197–207.

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

[8] M. C. Cai, M. Mu, and J. C. Xu, *Numerical Study on Two-Level and Multilevel Methods for Mixed Stokes/Darcy Model*, in preparation.

[9] B. Chanetz, R. Benay, J. Bousquet, R. Bur, T. Pot, F. Grasso, and J. Moss, *Experimental and numerical study of the laminar separation in hypersonic flow*, Aerospace Sci. Technol., 3 (1998), pp. 205–218.

[10] M. Discacciati and A. Quarteroni, *Analysis of a domain decomposition method for the coupling of Stokes and Darcy equations*, in Proceedings of the 3rd European Conference on Numerical Mathematics and Advanced Applications (ENUMATH 2001), F. Brezzi, A. Buffa, S. Corsaro, and A. Murli, eds., Springer, Milan, 2003, pp. 3–20.
[11] M. Discacciati and A. Quarteroni, *Convergence analysis of a subdomain iterative method for the finite element approximation of the coupling of Stokes and Darcy equations*, Comput. Vis. Sci., 6 (2005), pp. 1001–1026.
[12] M. Discacciati, E. Miglio, and A. Quarteroni, *Mathematical and numerical models for coupling surface and groundwater flows*, Appl. Numer. Math., 43 (2002), pp. 57–74.
[13] V. Girault and J.-L. Lions, *Two-grid finite-element schemes for the transient Navier-Stokes problem. Mathematical modelling and numerical analysis*, M2AN Math. Model. Numer. Anal., 35 (2001), pp. 945–980.
[14] V. Girault and P. Raviart, *Finite Element Methods for Navier-Stokes Equations: Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[15] R. Glowinski, T. Pan, and J. Periaux, *A Lagrange multiplier/fictitious domain method for the numerical simulation of incompressible viscous flow around moving grid bodies:* I. *Case where the rigid body motions are known a priori*, C. R. Acad. Sci. Paris Sér. I Math., 324 (1997), pp. 361–369.
[16] W. Jager and A. Mikelic, *On the boundary conditions at the contact interface between a porous medium and a free fluid*, Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4), 23 (1996), pp. 403–465.
[17] S. Klainerman and A. Majda, *Compressible and incompressible fluids*, Comm. Pure Appl. Math., 35 (1982), pp. 629–651.
[18] W. Layton and W. Lenferink, *Two-level Picard and modified Picard methods for the Navier-Stokes equations*, Appl. Math. Comput., 69 (1995), pp. 263–274.
[19] W. Layton, A. Meir, and P. Schmidt, *A two-level discretization method for the stationary MHD equations*, Electron. Trans. Numer. Anal., 6 (1997), pp. 198–210.
[20] W. Layton and L. Tobiska, *A two-level method with backtracking for the Navier–Stokes equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2035–2054.
[21] W. J. Layton, F. Schieweck, and I. Yotov, *Coupling fluid flow with porous media flow*, SIAM J. Numer. Anal., 40 (2003), pp. 2195–2218.
[22] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Vol. 1, Springer-Verlag, New York, Heidelberg, 1972.
[23] M. Marion and J. Xu, *Error estimates on a new nonlinear Galerkin method based on two-grid finite elements*, SIAM J. Numer. Anal., 32 (1995), pp. 1170–1184.
[24] S. Markus, E. Houstis, A. Catlin, J. Rice, P. Tsompanopoulou, E. Vavalis, D. Gottfried, K. Su, and G. Balakrishnan, *An agent-based netcentric framework for multi-disciplinary problem solving environments (MPSE)*, Internat. J. Comput. Engrg. Sci., 1 (2000), pp. 33–60.
[25] M. Mu, *Solving composite problems with interface relaxation*, SIAM J. Sci. Comput., 20 (1999), pp. 1394–1416.
[26] M. Mu, *PDE.Mart: A network-based problem-solving environment for PDEs*, ACM Trans. Math. Software, 31 (2005), pp. 508–531.
[27] L. Payne and B. Straughan, *Analysis of the boundary condition at the interface between a viscous fluid and a porous medium and related modelling questions*, J. Math. Pures Appl., 77 (1998), pp. 317–354.
[28] M. Peszynska, M. Wheeler, and I. Yotov, *Mortar upscaling for multiphase flow in porous media*, Comput. Geosci., 6 (2002), pp. 73–100.
[29] A. Quarteroni and A. Valli, *Domain Decomposition Methods for Partial Differential Equations*, Oxford University Press, Oxford, UK, 1999.
[30] P. Saffman, *On the boundary condition at the surface of a porous media*, Stud. Appl. Math., 50 (1971), pp. 93–101.
[31] A. Shestakov, M. Prasad, J. Milovich, N. Gentile, J. Painter, and G. Furnish, *The radiation-hydrodynamic ICF3D code*, Comput. Methods Appl. Mech. Engrg., 187 (2000), pp. 181–200.
[32] T. Utnes, *Two-grid finite element formulations of the incompressible Navier-Stokes equations*, Comm. Numer. Methods Engrg., 13 (1997), pp. 675–684.
[33] J. Xu, *Theory of Multilevel Methods*, Ph.D. dissertation, Cornell University, Ithaca, NY, 1989.
[34] J. Xu, *A new class of iterative methods for nonselfadjoint or indefinite problems*, SIAM J. Numer. Anal., 29 (1992), pp. 303–319.
[35] J. Xu, *Iterative methods by SPD and small subspace solvers for nonsymmetric or indefinite problems*, in Proceedings of the Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, SIAM, Philadelphia, 1992, pp. 106–118.

[36] J. Xu, *A novel two-grid method for semilinear elliptic equations*, SIAM J. Sci. Comput., 15 (1994), pp. 231–237.
[37] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs*, SIAM J. Numer. Anal., 33 (1996), pp. 1759–1777.
[38] J. Xu and A. Zhou, *Local and parallel finite element algorithms based on two-grid discretizations*, Math. Comp., 69 (2000), pp. 881–909.
[39] J. Xu and A. Zhou, *Local and parallel finite element algorithms based on two-grid discretizations for nonlinear problems*, Adv. Comput. Math., 14 (2001), pp. 293–327.
[40] J. Xu and A. Zhou, *Local and parallel finite element algorithms for eigenvalue problems*, Acta Math. Appl. Sin. Engl. Ser., 18 (2002), pp. 185–200.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1814–1842

SPECIALIZED PARTITIONED ADDITIVE RUNGE–KUTTA METHODS FOR SYSTEMS OF OVERDETERMINED DAES WITH HOLONOMIC CONSTRAINTS*

LAURENT O. JAY†

Abstract. We consider a general class of systems of overdetermined differential-algebraic equations (ODAEs). We are particularly interested in extending the application of the symplectic Gauss methods to Hamiltonian and Lagrangian systems with holonomic constraints. For the numerical approximation to the solution to these ODAEs, we present specialized partitioned additive Runge–Kutta (SPARK) methods, and in particular the new class of (s, s)-Gauss–Lobatto SPARK methods. These methods not only preserve the constraints, symmetry, symplecticness of the flow, and variational nature of the trajectories of holonomically constrained Hamiltonian and Lagrangian systems, but they also have an optimal order of convergence $2s$.

Key words. differential-algebraic equations, Gauss coefficients, Hamiltonian systems, holonomic constraints, Lagrangian systems, Lobatto coefficients, Runge–Kutta methods, symplecticness, variational integrators

AMS subject classifications. 65L05, 65L06, 65L80, 70F20, 70H03, 70H05, 70H45

DOI. 10.1137/060667475

1. Introduction.

Gauss methods for Hamiltonian systems are known to be symplectic [7, 8, 19, 25, 28]. For Lagrangian systems these methods are also known to be of a variational nature [21]. The main objective of this paper is to present extensions of Gauss methods to Hamiltonian and Lagrangian systems with holonomic constraints. For these systems we have found extensions of Gauss methods preserving symplecticness, the manifold of constraints, the variational nature of trajectories, and having an optimal order of convergence. When applied to nonstiff ordinary differential equations (ODEs), Gauss methods have a maximal order of convergence in the class of Runge–Kutta (RK) methods [3, 8]. However, for index 3 differential-algebraic equations (DAEs) such as Hamiltonian systems with holonomic constraints, standard Gauss methods either are divergent or have a very low order of convergence when the underlying differentiated constraints are not taken into account [5]. Gauss methods have thus not been considered of much practical interest for the numerical solution of high index DAEs. Recently, optimal methods based on Gauss coefficients have been obtained for index 2 DAEs [18] and have stirred renewed interest in Gauss methods for DAEs.

In this paper we consider a general class of systems of overdetermined differential-algebraic equations (ODAEs), including a unified formulation of Hamiltonian and Lagrangian systems with holonomic constraints. To approximate numerically the solution to these systems of ODAEs, we present the new class of specialized partitioned additive Runge–Kutta (SPARK) methods. We make great use of the structure of the ODAEs. The new class of (s, s)-Gauss–Lobatto SPARK methods extends to these ODAEs the application of Gauss methods to ODEs. These symmetric methods are

*Received by the editors August 14, 2006; accepted for publication (in revised form) January 22, 2007; published electronically August 31, 2007. This material is based upon work supported by the National Science Foundation under grant 9983708.
http://www.siam.org/journals/sinum/45-5/66747.html
†Department of Mathematics, The University of Iowa, 14 MacLean Hall, Iowa City, IA 52242-1419 (ljay@math.uiowa.edu, na.ljay@na-net.ornl.gov).

shown to be superconvergent of order $2s$ and constraint preserving. Moreover, for Hamiltonian and Lagrangian systems with holonomic constraints these methods are shown to be symplectic and to satisfy a discrete variational principle.

The paper is organized as follows. In section 2 we introduce the equations of Hamiltonian and Lagrangian systems with holonomic constraints. We state some of their relations and main properties. A unified formulation of Hamiltonian and Lagrangian systems is presented and generalized to a larger class of systems of ODAEs. In section 3 we introduce the new class of SPARK methods. Examples of SPARK methods are given. In section 4 we characterize symplectic SPARK methods and show their variational nature. In section 5 we give results about existence, uniqueness, local error, and global convergence of SPARK methods. Finally, in section 6 some numerical experiments are given to illustrate our theoretical results. A short conclusion is given in section 7.

Regarding notation, we denote by x' the total derivative of x with respect to the independent variable t. For a function $f(x, y)$, we denote by $f_x(x, y)$ its partial derivative with respect to x.

2. Hamiltonian and Lagrangian systems with holonomic constraints. In this section we introduce the equations of Hamiltonian and Lagrangian systems with holonomic constraints. For these systems some important relations and properties are stated [1, 4, 20]. A unified and generalized formulation of Hamiltonian and Lagrangian systems is presented.

2.1. Hamiltonian systems with holonomic constraints. The Hamiltonian system with Hamiltonian $H : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ and holonomic constraints $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ $(m < n)$ is given by

$$q' = H_p^T(q, p), \tag{2.1a}$$

$$p' = -H_q^T(q, p) - g_q^T(q)\lambda, \tag{2.1b}$$

$$0 = g(q). \tag{2.1c}$$

Differentiating (2.1c) once with respect to the independent variable t, we obtain $g_q(q)q' = 0$, and from (2.1a) this leads to

$$0 = g_q(q)H_p^T(q, p). \tag{2.1d}$$

We assume that $g_q(q)$ is of full row rank m and that the Hessian matrix

$$H_{pp}^T(q, p) \text{ is invertible.} \tag{2.2}$$

For example, $H_{pp}^T(q, p)$ is generally assumed to be (strictly) positive definite. Equations (2.1a,b,c) are DAEs of index 3 in Hessenberg form [2, 6, 9, 12, 15]. The whole system (2.1) can be considered as a system of index 2 ODAEs. For consistent initial values, i.e., for $(q_0, p_0) \in V$, where

$$V := \left\{(q, p) \in \mathbb{R}^n \times \mathbb{R}^n \mid 0 = g(q),\ 0 = g_q(q)H_p^T(q, p)\right\}, \tag{2.3}$$

we have existence and uniqueness of a solution.

2.2. Lagrangian systems with holonomic constraints. The Lagrangian system with Lagrangian $L : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ and holonomic constraints $g : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ $(m < n)$ is given by

$$q' = v, \tag{2.4a}$$

$$(L_v^T(q,v))' = L_q^T(q,v) - g_q^T(q)\lambda, \tag{2.4b}$$

$$0 = g(q). \tag{2.4c}$$

Differentiating (2.4c) once with respect to t, we obtain $g_q(q)q' = 0$, and from (2.4a) this leads to

$$0 = g_q(q)v. \tag{2.4d}$$

We assume that $g_q(q)$ is of full row rank m and that the Hessian matrix

$$L_{vv}^T(q,v) \text{ is invertible,} \tag{2.5}$$

for example, $L_{vv}^T(q,v)$ is generally assumed to be (strictly) positive definite. Equations (2.4a,b,c) are usually called *Euler–Lagrange equations* and are DAEs of index 3 [15]. The whole system (2.4) can be considered as a system of index 2 ODAEs. For consistent initial values, i.e., for $(q_0, v_0) \in W$, where

$$W := \{(q,v) \in \mathbb{R}^n \times \mathbb{R}^n \mid 0 = g(q),\ 0 = g_q(q)v\}, \tag{2.6}$$

we have existence and uniqueness of a solution. For Lagrangian systems with holonomic constraints (2.4), it is advantageous to consider directly the formulation (2.4b) instead of

$$L_{vv}^T(q,v)v' = -L_{vq}^T(q,v)v + L_q^T(q,v) - g_q^T(q)\lambda, \tag{2.7}$$

since this formulation (2.7) requires an extra term $L_{vq}^T(q,v)v$ which usually corresponds to Coriolis forces; see [14, 15]. Moreover, preserving the Lagrangian symplectic 2-form (2.10) for numerical methods is certainly more problematic with formulation (2.7) than with (2.4b); see also Corollary 4.2.

2.3. Relations and properties of Hamiltonian and Lagrangian systems with holonomic constraints. Lagrangian systems are closely related to Hamiltonian systems. The momenta p of a Lagrangian system are defined by

$$p := L_v^T(q,v). \tag{2.8}$$

From (2.5), the relation $p - L_v^T(q,v) = 0$ defines v as an implicit function $v(q,p)$. Under assumption (2.5) the Lagrangian system (2.4) is equivalent by the change of variables (2.8) to the Hamiltonian system (2.1) with Hamiltonian

$$H(q,p) := p^T v(q,p) - L(q, v(q,p)).$$

This is known as a Legendre transform. Assumption (2.5) is equivalent to (2.2). The velocities of a Hamiltonian system are defined by

$$v := H_p^T(q,p). \tag{2.9}$$

From (2.2), the relation $v - H_p^T(q,p) = 0$ defines p as an implicit function $p(q,v)$. Under assumption (2.2) Hamiltonian system (2.1) is equivalent by the change of variables (2.9) to Lagrangian system (2.4) with Lagrangian

$$L(q,v) := p^T(q,v)v - H(q,p(q,v)).$$

This is also a Legendre transform. Under the equivalent assumptions (2.2) and (2.5) we have the following symmetric relations between Lagrangian systems and their Hamiltonian counterparts:

$$p^T v = H(q,p) + L(q,v),$$

$$p = L_v^T(q,v),$$

$$v = H_p^T(q,p),$$

$$I_n = H_{pp}^T(q,p)L_{vv}^T(q,v).$$

Properties of Lagrangian systems can thus be transferred to Hamiltonian systems, and vice versa. Hence, here we state only five important properties of Lagrangian systems with holonomic constraints as follows:

1. Any solution to (2.4) must lie on the manifold of constraints W (2.6). In particular, any initial conditions (q_0, v_0) at t_0 must belong to W.

2. The energy function $E(q,v) := L_v(q,v)v - L(q,v)$ is invariant along a solution, i.e.,

$$E(q(t), v(t)) = \text{Const.}$$

3. The flow $\varphi_\tau : (q(t), v(t)) \mapsto (q(t+\tau), v(t+\tau))$ on the manifold of constraints W preserves the Lagrangian symplectic 2-form

$$\sum_{i=1}^{n} dq^i \wedge dL_{v^i}(q,v) = \sum_{i=1}^{n}\sum_{j=1}^{n}(L_{v^i q^j}(q,v)dq^i \wedge dq^j + L_{v^i v^j}(q,v)dq^i \wedge dv^j). \tag{2.10}$$

4. The *action* of the Lagrangian

$$\int_{t_a}^{t_b} L(q(t), v(t)) - g^T(q(t))\lambda(t)dt$$

is stationary. This is *Hamilton's variational principle*. The algebraic variables λ are Lagrange multipliers associated with the holonomic constraints (2.4c).

5. The flow may be γ-*reversible*, i.e., $\varphi_\tau = \gamma^{-1} \circ \varphi_\tau^{-1} \circ \gamma$ for some transformation γ of the variables (q,v). For example, for *conservative mechanical systems* in Lagrangian form, the Lagrangian is given by $L(q,v) = T(q,v) - U(q)$, where $T(q,v) = \frac{1}{2}v^T M(q)v$ is the *kinetic energy* with $M(q)$ being the (strictly) positive definite symmetric generalized mass matrix, and $U(q)$ is the *potential energy*. The flow is γ-reversible with respect to a reflection of the velocities $\gamma : (q,v) \mapsto (q,-v)$.

2.4. Unification and generalization of the formulation of Hamiltonian and Lagrangian systems with holonomic constraints. We present here a unified and generalized formulation of Hamiltonian and Lagrangian systems with holonomic constraints, consisting of a set of implicit ODAEs

$$y' = v(y,z), \tag{2.11a}$$

$$(p(y,z))' = f(y,z) + r(y,\lambda), \tag{2.11b}$$

$$0 = g(y), \tag{2.11c}$$

$$0 = g_y(y)v(y,z). \tag{2.11d}$$

These equations encompass the formulation of conservative mechanical systems with constraints of holonomic and scleronomic types [10, 22, 26, 27]. In mechanics the quantities y, v, p, f, r usually represent, respectively, generalized coordinates, generalized velocities, generalized momenta, generalized forces, and reaction forces due to the holonomic constraints (2.11c). These equations include Hamiltonian systems with holonomic constraints (2.1) and Lagrangian systems with holonomic constraints (2.4). For Hamiltonian systems (2.1) we have $q = y$, $p(y,z) = z$, $v(y,z) = H_z^T(y,z)$, $f(y,z) = -H_y^T(y,z)$, and $r(y,\lambda) = -g_y^T(y)\lambda$. For Lagrangian systems (2.4) we have $q = y$, $v(y,z) = z$, $p(y,z) = L_z^T(y,z)$, $f(y,z) = L_y^T(y,z)$, and $r(y,\lambda) = -g_y^T(y)\lambda$. Equation (2.11d) corresponds to $0 = (g(y))' = g_y(y)y'$. The variable $t \in \mathbb{R}$ is the independent variable and

$$\begin{aligned}
&y = (y^1, \ldots, y^{n_y})^T \in \mathbb{R}^{n_y},\\
&z = (z^1, \ldots, z^{n_z})^T \in \mathbb{R}^{n_z},\\
&\lambda = (\lambda^1, \ldots, \lambda^{n_\lambda})^T \in \mathbb{R}^{n_\lambda},\\
&p \,:\, \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \longrightarrow \mathbb{R}^{n_z},\\
&g \,:\, \mathbb{R} \times \mathbb{R}^{n_y} \longrightarrow \mathbb{R}^{n_\lambda},\\
&v \,:\, \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \longrightarrow \mathbb{R}^{n_y},\\
&f \,:\, \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_z} \longrightarrow \mathbb{R}^{n_z},\\
&r \,:\, \mathbb{R} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_\lambda} \longrightarrow \mathbb{R}^{n_z}.
\end{aligned}$$

The variables y, z are called the *differential* variables and the variables λ are called the *algebraic* variables. The latter correspond to Lagrange multipliers when the DAEs are derived from a constrained variational principle [10, 22]. The initial values y_0, z_0 at t_0 are assumed to be given and consistent, i.e., (2.11c) and (2.11d) must be satisfied. Some differentiability conditions on the above functions are also assumed to ensure existence and uniqueness of the solution. In a neighborhood of the solution the following conditions are assumed to be satisfied:

$$p_z \text{ is invertible,} \tag{2.12a}$$

$$\begin{pmatrix} p_z & -r_\lambda \\ g_y v_z & O \end{pmatrix} \text{ is invertible.} \tag{2.12b}$$

Differentiating the left-hand side of (2.11b), under the assumption (2.12a) we obtain the following expression:

$$z' = p_z(y,z)^{-1}\left(f(y,z) + r(y,\lambda) - p_y(y,z)v(y,z)\right). \tag{2.13}$$

Differentiating the constraints (2.11d) leads to

$$0 = g_{yy}(y)\left(v(y,z), v(y,z)\right) + g_y(y)(v_y(y,z)v(y,z) + v_z(y,z)z'). \tag{2.14}$$

Introducing the expression for z' from (2.13) into (2.14), we see that under assumption (2.12b) equations (2.14) form an implicit system of equations for λ whose solution exists and is locally unique by application of the implicit function theorem.

Introducing the new variables q, p and the relations

$$q = y, \qquad p = p(y,z), \tag{2.15}$$

under the assumption (2.12a) we can formally express the differential variables y and z as (implicit) functions of (q,p), i.e.,

$$y = q, \qquad z = z(q,p).$$

Defining

$$V(q,p) := v\left(q, z\left(q,p\right)\right), \; F(q,p) := f\left(q, z\left(q,p\right)\right), \; R(q,\lambda) := r\left(q,\lambda\right), \; G(q) := g\left(q\right),$$

the whole system (2.11) can be reformulated in an equivalent way as

$$q' = V(q,p), \tag{2.16a}$$

$$p' = F(q,p) + R(q,\lambda), \tag{2.16b}$$

$$0 = G(q), \tag{2.16c}$$

$$0 = G_q(q)V(q,p), \tag{2.16d}$$

and assumption (2.12b) is equivalent to

$$G_q V_p R_\lambda \text{ is invertible.} \tag{2.17}$$

There is no implicit derivative in (2.16b). Since the application of SPARK methods (3.2) below is invariant under the change of variables (2.15), for the analysis in section 5 we can simply consider $p(y,z) = z$ in (2.11b).

3. SPARK methods. After briefly considering the class of standard RK methods, we introduce the new class of SPARK methods for the ODAEs (2.11). Examples of SPARK methods are then given.

3.1. Standard RK methods. The standard application of RK methods to the system of index 3 DAEs (2.11a,b,c) with $p(y,z) = z$ is as follows [6]:

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij} v(Y_j, Z_j) \quad \text{for } i = 1,\dots,s,$$

$$Z_i = z_0 + h\sum_{j=1}^{s} a_{ij}\left(f(Y_j, Z_j) + r(Y_j, \Lambda_j)\right) \quad \text{for } i = 1,\dots,s,$$

$$0 = g(Y_i) \quad \text{for } i = 1, \dots, s,$$

$$y_1 = y_0 + h \sum_{j=1}^{s} b_j v(Y_j, Z_j),$$

$$z_1 = z_0 + h \sum_{j=1}^{s} b_j \left(f(Y_j, Z_j) + r(Y_j, \Lambda_j) \right).$$

For example, the standard $s = 1$-stage Gauss RK method for Hamiltonian systems with holonomic constraints (2.1a,b,c), based on the implicit midpoint rule for ODEs, reads as

$$Q_1 = q_0 + h\frac{1}{2} H_p^T(Q_1, P_1) = \frac{1}{2}(q_1 + q_0),$$

$$P_1 = p_0 - h\frac{1}{2} H_q^T(Q_1, P_1) - h\frac{1}{2} g_q^T(q_1)\Lambda_1 = \frac{1}{2}(p_1 + p_0),$$

$$0 = g(Q_1),$$

$$q_1 = q_0 + h H_p^T(Q_1, P_1),$$

$$p_1 = p_0 - h H_q^T(Q_1, P_1) - h g_q^T(q_1)\Lambda_1.$$

Unfortunately, this method is in general divergent. More generally, the standard definition of RK methods does not take advantage of the additive structure of (2.11b) and of the presence of the two sets of constraints (2.11c,d). A different extension of the implicit midpoint rule, convergent even for the more general system of ODAEs (2.11), can be found within the class of SPARK methods (3.2) to be discussed hereafter. When $p(y, z) = z$, this extension is given by

$$Y_1 = y_0 + h\frac{1}{2} v(Y_1, Z_1) = \frac{1}{2}(y_1 + y_0), \tag{3.1a}$$

$$Z_1 = z_0 + h\frac{1}{2} f(Y_1, Z_1) + h\frac{1}{2} r(y_0, \Lambda_0), \tag{3.1b}$$

$$y_1 = y_0 + h v(Y_1, Z_1), \tag{3.1c}$$

$$0 = g(y_1), \tag{3.1d}$$

$$z_1 = z_0 + h f(Y_1, Z_1) + h\frac{1}{2} r(y_0, \Lambda_0) + h\frac{1}{2} r(y_1, \Lambda_1), \tag{3.1e}$$

$$0 = g_y(y_1) v(y_1, z_1) \tag{3.1f}$$

and is named a $(1, 1)$-Gauss–Lobatto SPARK method; see subsection 3.3. Note that the quantity Λ_0 is local to the current step and does not come from the previous step. For Hamiltonian systems with holonomic constraints (2.1), we obtain

$$Q_1 = q_0 + h\frac{1}{2} H_p^T(Q_1, P_1) = \frac{1}{2}(q_1 + q_0),$$

$$P_1 = p_0 - h\frac{1}{2} H_q^T(Q_1, P_1) - h\frac{1}{2} g_q^T(q_0)\Lambda_0,$$

$$q_1 = q_0 + hH_p^T(Q_1, P_1),$$

$$0 = g(q_1),$$

$$p_1 = p_0 - hH_q^T(Q_1, P_1) - h\frac{1}{2}g_q^T(q_0)\Lambda_0 - h\frac{1}{2}g_q^T(q_1)\Lambda_1,$$

$$0 = g_q(q_1)H_p^T(q_1, p_1).$$

For separable Hamiltonian systems of the form $H(q,p) = \frac{1}{2}p^T M^{-1}p + U(q)$, this method is equivalent to a method proposed by Reich in [24].

3.2. Definition of SPARK methods. We propose here a class of methods based on RK coefficients taking advantage of the structure of (2.11), in particular of the additive and partitioned structure of (2.11a,b) and of the presence of the two sets of constraints (2.11c,d). The definition of SPARK methods is given below. A similar application of SPARK methods has been proposed for the numerical solution of mechanical systems in [15]; see also [16].

DEFINITION 3.1. *One step of an* $(s, \widetilde{s})$-SPARK method *applied to the system of implicit overdetermined partitioned DAEs* (2.11) *with consistent initial values* (y_0, z_0) *at* t_0 *and stepsize* h *is given as follows:*

$$Y_i = y_0 + h\sum_{j=1}^{s} a_{ij}v(Y_j, Z_j) \quad \text{for } i = 1, \ldots, s, \tag{3.2a}$$

$$p(Y_i, Z_i) = p_0 + h\sum_{j=1}^{s} \widehat{a}_{ij}f(Y_j, Z_j) + h\sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij}r(\widetilde{Y}_j, \Lambda_j) \quad \text{for } i = 1, \ldots, s, \tag{3.2b}$$

$$\widetilde{Y}_i = y_0 + h\sum_{j=1}^{s} \overline{a}_{ij}v(Y_j, Z_j) \quad \text{for } i = 0, 1, \ldots, \widetilde{s}, \tag{3.2c}$$

$$0 = g(\widetilde{Y}_i) \quad \text{for } i = 0, 1, \ldots, \widetilde{s}, \tag{3.2d}$$

$$y_1 = y_0 + h\sum_{j=1}^{s} b_j v(Y_j, Z_j), \tag{3.2e}$$

$$p(y_1, z_1) = p_0 + h\sum_{j=1}^{s} \widehat{b}_j f(Y_j, Z_j) + h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j r(\widetilde{Y}_j, \Lambda_j), \tag{3.2f}$$

$$0 = g(y_1), \tag{3.2g}$$

$$0 = g_y(y_1)v(y_1, z_1), \tag{3.2h}$$

where $p_0 := p(y_0, z_0)$. *We have four sets of coefficients* (b_j, a_{ij}, c_i), $(\widehat{b}_j, \widehat{a}_{ij})$, $(\widetilde{b}_j, \widetilde{a}_{ij})$, $(\overline{a}_{ij}, \widetilde{c}_i)$, *where we have defined*

$$c_i := \sum_{j=1}^{s} a_{ij} \quad \text{for } i = 1, \ldots, s, \qquad \widetilde{c}_i := \sum_{j=1}^{s} \overline{a}_{ij} \quad \text{for } i = 0, 1, \ldots, \widetilde{s}.$$

Notice that the coefficients $(b_j, c_j)_{j=1}^{s}$ *and* $(\widetilde{b}_j, \widetilde{c}_j)_{j=0}^{\widetilde{s}}$ *are generally two distinct quadrature formulas. The SPARK coefficients can be expressed concisely in four Butcher-style tableaux:*

$$\begin{array}{c|c} c_i & a_{ij} \\ \hline A & b_j \end{array} \qquad \begin{array}{c|c} & \widehat{a}_{ij} \\ \hline \widehat{A} & \widehat{b}_j \end{array} \qquad \begin{array}{c|c} & \widetilde{a}_{ij} \\ \hline \widetilde{A} & \widetilde{b}_j \end{array} \qquad \begin{array}{c|c} \widetilde{c}_i & \overline{a}_{ij} \\ \hline \overline{A} & \end{array}.$$

When the RK matrix $A = (a_{ij})_{i,j=1}^{s}$ is invertible we can express the values $\widetilde{Y}_i$ for $i = 0, 1, \ldots, \widetilde{s}$ and y_1 as linear combinations of y_0 and Y_j for $j = 1, \ldots, s$ as follows:

$$\widetilde{Y}_i = y_0 + \sum_{j=1}^{s} \eta_{ij}(Y_j - y_0), \qquad y_1 = y_0 + \sum_{j=1}^{s} \nu_j(Y_j - y_0),$$

where $\eta := \overline{A}A^{-1}$ and $\nu^T := b^T A^{-1}$. An $(s, \widetilde{s})$-SPARK method (3.2) can be seen as an extension of an s-stage standard (partitioned) RK method for partitioned ODEs

$$y' = v(y, z), \qquad z' = f(y, z).$$

To ensure existence and uniqueness of the SPARK solution (see Theorem 5.1), we assume the SPARK coefficients satisfy the following conditions:

(3.3a) $\overline{a}_{0j} = 0 \quad$ for $j = 1, \ldots, s$,

(3.3b) $\overline{a}_{\widetilde{s}j} = b_j \quad$ for $j = 1, \ldots, s$,

(3.3c) $$\sum_{j=1}^{s} \overline{a}_{ij} c_j = \sum_{j=1}^{s}\sum_{k=1}^{s} \overline{a}_{ij}\widehat{a}_{jk} = \sum_{j=1}^{s}\sum_{k=0}^{\widetilde{s}} \overline{a}_{ij}\widetilde{a}_{jk} = \frac{\widetilde{c}_i^2}{2} \quad \text{for } i = 0, 1, \ldots, \widetilde{s},$$

(3.3d) $$\overline{A}\widetilde{A} =: \begin{pmatrix} 0 & \cdots & 0 \\ & N & \end{pmatrix}, \qquad \begin{pmatrix} N \\ \widetilde{b}^T \end{pmatrix} \text{ is invertible.}$$

Condition (3.3a) implies that $\widetilde{c}_0 = 0$ and $\widetilde{Y}_0 = y_0$. Therefore $g(\widetilde{Y}_0) = 0$ is automatically satisfied since we assume $g(y_0) = 0$. Such SPARK methods generally do not require the evaluation of $v(y_0, z_0)$ and $f(y_0, z_0)$. However, $r(y_0, \Lambda_0)$ is required. Condition (3.3b) implies that $g(y_1) = 0$ is automatically satisfied since $g(\widetilde{Y}_{\widetilde{s}}) = 0$ from (3.2d) for $i = \widetilde{s}$ and $y_1 = \widetilde{Y}_{\widetilde{s}}$.

3.3. The (s, s)-Gauss–Lobatto SPARK methods. We are especially interested in extending Gauss RK methods for ODEs without constraints to corresponding (s, s)-SPARK methods (3.2) for the ODAEs (2.11) having an optimal order of convergence $2s$. The Gauss RK coefficients $\widehat{a}_{ij} = a_{ij}, \widehat{b}_j = b_j$ can be found, e.g., in [3, 7]. The Gauss RK coefficients satisfy

$$\sum_{i=1}^{s} b_i c_i^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s,$$

$$\sum_{j=1}^{s} a_{ij} c_j^{k-1} = \frac{c_i^k}{k}, \qquad i = 1, \ldots, s, \quad k = 1, \ldots, s.$$

For the coefficients $\widetilde{b}_i$ and $\widetilde{c}_i$, we take the coefficients of the $(s+1)$-stage Lobatto quadrature formula ($\widetilde{c}_0 = 0, \widetilde{c}_s = 1$) of order $2s$ which satisfy

$$\sum_{i=0}^{s} \widetilde{b}_i \widetilde{c}_i^{k-1} = \frac{1}{k}, \qquad k = 1, \ldots, 2s.$$

The coefficients $\overline{a}_{ij}$ can be taken according to

$$\sum_{j=1}^{s} \overline{a}_{ij} c_j^{k-1} = \frac{\widetilde{c}_i^k}{k}, \qquad i = 0, 1, \ldots, s, \quad k = 1, \ldots, s,$$

and the coefficients $\widetilde{a}_{ij}$ are then simply determined by

$$\widetilde{a}_{ij} = \widetilde{b}_j \left(1 - \frac{\overline{a}_{ji}}{b_i}\right), \qquad i = 1, \ldots, s, \quad j = 0, 1, \ldots, s.$$

These methods are called (s, s)-*Gauss–Lobatto SPARK methods*. They have order $2s$ of convergence; see Corollary 5.4. It can be shown that these methods satisfy conditions (3.3) and

$$\widetilde{a}_{i0} = \widetilde{b}_0, \quad \widetilde{a}_{is} = 0, \quad i = 1, \ldots, s.$$

The algebraic variable Λ_s appears only in (3.2f) and is thus determined by (3.2h).

The $(1, 1)$-Gauss–Lobatto SPARK method corresponds to the following Butcher-style tableaux of SPARK coefficients:

$$\begin{array}{c|c} 1/2 & 1/2 \\ \hline A & 1 \end{array} \qquad \begin{array}{c|c} & 1/2 \\ \hline \widehat{A} & 1 \end{array} \qquad \begin{array}{c|cc} & 1/2 & 0 \\ \hline \widetilde{A} & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|c} 0 & 0 \\ 1 & 1 \\ \hline \overline{A} & \end{array}.$$

We have $\widetilde{Y}_0 = y_0$ and $\widetilde{Y}_1 = y_1$. When $p(y, z) = z$ in (2.11b) the method simplifies to (3.1) for $Y_1, y_1, Z_1, z_1, \Lambda_0, \Lambda_1$.

The $(2, 2)$-Gauss–Lobatto SPARK method corresponds to the following Butcher-style tableaux of SPARK coefficients:

$$\begin{array}{c|cc} 1/2 - \sqrt{3}/6 & 1/4 & 1/4 - \sqrt{3}/6 \\ 1/2 + \sqrt{3}/6 & 1/4 + \sqrt{3}/6 & 1/4 \\ \hline A & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} & 1/4 & 1/4 - \sqrt{3}/6 \\ & 1/4 + \sqrt{3}/6 & 1/4 \\ \hline \widehat{A} & 1/2 & 1/2 \end{array}$$

$$\begin{array}{c|ccc} & 1/6 & 1/3 - \sqrt{3}/6 & 0 \\ & 1/6 & 1/3 + \sqrt{3}/6 & 0 \\ \hline \widetilde{A} & 1/6 & 2/3 & 1/6 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/4 + \sqrt{3}/8 & 1/4 - \sqrt{3}/8 \\ 1 & 1/2 & 1/2 \\ \hline \overline{A} & & \end{array}.$$

We have $\widetilde{Y}_0 = y_0$ and $\widetilde{Y}_2 = y_1$.

3.4. The Lobatto IIIA-B partitioned RK (PRK) methods. SPARK methods (3.2) include the Lobatto IIIA-B PRK methods of [12, 13]. For example, the

$(2,1)$-Lobatto IIIA-B SPARK method of order 2 (an extension of the Störmer/leap-frog/Verlet/RATTLE/ SHAKE methods) corresponds to the following Butcher-style tableaux of SPARK coefficients:

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline A & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} & 1/2 & 0 \\ & 1/2 & 0 \\ \hline \widehat{A} & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} & 1/2 & 0 \\ & 1/2 & 0 \\ \hline \widetilde{A} & 1/2 & 1/2 \end{array} \qquad \begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline \overline{A} & & \end{array}.$$

For this method we have $Y_1 = \widetilde{Y}_0 = y_0$ and $Y_2 = \widetilde{Y}_1 = y_1$. When $p(y,z) = z$ in (2.11b) the method simplifies to the following equations for $y_1, Z_1 = Z_2, z_1, \Lambda_0, \Lambda_1$:

$$\begin{aligned} y_1 &= y_0 + h\frac{1}{2}v(y_0, Z_1) + h\frac{1}{2}v(y_1, Z_2), \\ Z_1 &= z_0 + h\frac{1}{2}f(y_0, Z_1) + h\frac{1}{2}r(y_0, \Lambda_0), \\ 0 &= g(y_1), \\ z_1 &= z_0 + h\frac{1}{2}f(y_0, Z_1) + h\frac{1}{2}f(y_1, Z_2) + h\frac{1}{2}r(y_0, \Lambda_0) + h\frac{1}{2}r(y_1, \Lambda_1) \\ &= Z_1 + h\frac{1}{2}f(y_1, Z_2) + h\frac{1}{2}r(y_1, \Lambda_1), \\ 0 &= g_y(y_1)v(y_1, z_1). \end{aligned}$$

3.5. The symplectic Euler method. For Hamiltonian systems with holonomic constraints (2.1), the symplectic Euler method [7, 9, 23] is defined as follows:

$$P_1 = p_0 - hH_q^T(q_0, P_1) - hg_q^T(q_0)\Psi_0, \tag{3.4a}$$

$$q_1 = q_0 + hH_p^T(q_0, P_1), \tag{3.4b}$$

$$0 = g(q_1), \tag{3.4c}$$

$$p_1 = p_0 - hH_q^T(q_0, P_1) - hg_q^T(q_0)\Psi_0 - hg_q^T(q_1)\Psi_1 = P_1 - hg_q^T(q_1)\Psi_1, \tag{3.4d}$$

$$0 = g_q(q_1)H_p^T(q_1, p_1). \tag{3.4e}$$

It is a method of order 1 and the two quantities Ψ_0, Ψ_1 are locally determined by these equations. The symplectic Euler method can be interpreted as a SPARK method (3.2) with coefficients

$$\begin{array}{c|c} 0 & 0 \\ \hline A & 1 \end{array} \qquad \begin{array}{c|c} & 1 \\ \hline \widehat{A} & 1 \end{array} \qquad \begin{array}{c|cc} & \alpha & 0 \\ \hline \widetilde{A} & \alpha & 1-\alpha \end{array} \qquad \begin{array}{c|c} 0 & 0 \\ 1 & 1 \\ \hline \overline{A} & \end{array},$$

which we call the *"natural" symplectic Euler method*. The quantities Ψ_0 and Ψ_1 correspond to $\Psi_0 = \alpha\Lambda_0$ and $\Psi_1 = (1-\alpha)\Lambda_1$. Unfortunately, this method does not satisfy (3.3c), and when applied to the more general problem (2.11) this SPARK method is generally not convergent [17] when $r(y,\lambda)$ is nonlinear in λ. A convergent

extension of the symplectic Euler method to (2.11) (here given when $p(y,z)=z$ in (2.11b)) is as follows:

$$\begin{aligned}
Z_1 &= z_0 + hf(y_0, Z_1) + h\alpha r(y_0, \Lambda_0),\\
y_1 &= y_0 + hv(y_0, Z_1),\\
0 &= g(y_1),\\
z_1 &= z_0 + hf(y_0, Z_1) + h\alpha(r(y_0,\Lambda_0) - r(y_1,\Lambda_0)) + hr(y_1, \widetilde{\Lambda}_1)\\
&= Z_1 - h\alpha r(y_1,\Lambda_0) + hr(y_1,\widetilde{\Lambda}_1),\\
0 &= g_y(y_1)v(y_1, z_1),
\end{aligned}$$

with $\alpha \neq 0$. We call this method the *"true" symplectic Euler method*. It is convergent of order 1 [17]. It cannot be expressed in the format of a SPARK method (3.2) when $r(y,\lambda)$ is nonlinear in λ. When $r(y,\lambda)$ is affine in λ it is equivalent to the natural symplectic Euler method, which is symplectic for Hamiltonian systems with holonomic constraints (2.1) and for Lagrangian systems with holonomic constraints (2.4); see Theorems 4.1 and 4.2.

4. Symplecticness and variational properties of SPARK methods. The preservation of the symplecticness of the flow of Hamiltonian and Lagrangian systems with holonomic constraints by SPARK methods is considered in this section. The variational properties of the discrete trajectories of symplectic SPARK methods are also examined.

4.1. Symplectic SPARK methods. For Hamiltonian systems with holonomic constraints (2.1), SPARK methods whose numerical flow preserves (locally) the symplecticness property are characterized as follows.

THEOREM 4.1. *We consider Hamiltonian systems with holonomic constraints* (2.1) *satisfying the assumptions given in section* 2.1. *If the SPARK method* (3.2) *applied to* (2.1) *satisfies*

$$\widehat{b}_i = b_i \qquad \text{for } i = 1,\dots,s, \tag{4.1a}$$

$$\widehat{b}_i a_{ij} + b_j \widehat{a}_{ji} - \widehat{b}_i b_j = 0 \qquad \text{for } i,j = 1,\dots,s, \tag{4.1b}$$

$$\widetilde{b}_i \overline{a}_{ij} + b_j \widetilde{a}_{ji} - \widetilde{b}_i b_j = 0 \qquad \text{for } i = 0,1,\dots,\widetilde{s},\quad j = 1,\dots,s, \tag{4.1c}$$

then the numerical flow $(q_0,p_0)\mapsto(q_1,p_1)$ *preserves on* V (2.3) *the symplectic* 2*-form* $\sum_{i=1}^n dq^i \wedge dp^i$.

Proof. We denote

$$V_j := H_p^T(Q_j,P_j), \quad F_j := -H_q^T(Q_j,P_j), \quad R_j := -g_q^T(\widetilde{Q}_j)\Lambda_j.$$

We have

$$\begin{aligned}
dq_1^J \wedge dp_1^J - dq_0^J \wedge dp_0^J = {} & h\sum_{i=1}^{s} \widehat{b}_i dq_0^J \wedge dF_i^J + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i dq_0^J \wedge dR_i^J + h\sum_{j=1}^{s} b_j dV_j^J \wedge dp_0^J\\
& + h^2 \sum_{j=1}^{s} b_j dV_j^J \wedge \sum_{i=1}^{s} \widehat{b}_i dF_i^J + h^2 \sum_{j=1}^{s} b_j dV_j^J \wedge \sum_{i=0}^{\widetilde{s}} \widetilde{b}_i dR_i^J.
\end{aligned}$$

Introducing in the first three terms the following three relations for q_0 and p_0, respectively:

$$q_0 = Q_i - h\sum_{j=1}^{s} a_{ij}V_j, \quad q_0 = \widetilde{Q}_i - h\sum_{j=1}^{s} \overline{a}_{ij}V_j, \quad p_0 = P_j - h\sum_{i=1}^{s} \widehat{a}_{ji}F_i - h\sum_{i=0}^{\widetilde{s}} \widetilde{a}_{ji}R_i,$$

we obtain

$$\begin{aligned}\sum_{J=1}^{n} dq_1^J \wedge dp_1^J - \sum_{J=1}^{n} dq_0^J \wedge dp_0^J = {} & h\sum_{i=1}^{s}\left(\widehat{b}_i \sum_{J=1}^{n} dQ_i^J \wedge dF_i^J + b_i \sum_{J=1}^{n} dV_i^J \wedge dP_i^J\right) \\ & + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \left(\sum_{J=1}^{n} d\widetilde{Q}_i^J \wedge dR_i^J\right) \\ & + h^2 \sum_{J=1}^{n}\sum_{j=1}^{s}\sum_{i=1}^{s}\left(b_j\widehat{b}_i - b_j\widehat{a}_{ji} - \widehat{b}_i a_{ij}\right) dV_j^J \wedge dF_i^J \\ & + h^2 \sum_{J=1}^{n}\sum_{j=1}^{s}\sum_{i=0}^{\widetilde{s}}\left(b_j\widetilde{b}_i - b_j\widetilde{a}_{ji} - \widetilde{b}_i\overline{a}_{ij}\right) dV_j^J \wedge dR_i^J.\end{aligned}$$

The first term vanishes by assumption (4.1a) and

$$\sum_{J=1}^{n}(dQ_i^J \wedge dF_i^J + dV_i^J \wedge dP_i^J) = 0;$$

see [8, Formula (II.16.18)]. The last two terms also vanish by assumptions (4.1b,c). It remains to show that the second term also vanishes. We have

$$dR_i^J = -\sum_{L=1}^{m}\sum_{K=1}^{n} \frac{\partial^2 g^L}{\partial q^K \partial q^J}(\widetilde{Q}_i)\Lambda_i^L d\widetilde{Q}_i^K - \sum_{L=1}^{m} \frac{\partial g^L}{\partial q^J}(\widetilde{Q}_i) d\Lambda_i^L.$$

We thus get

$$\begin{aligned}\sum_{J=1}^{n} d\widetilde{Q}_i^J \wedge dR_i^J = {} & -\sum_{L=1}^{m} \Lambda_i^L \left(\sum_{J=1}^{n}\sum_{K=1}^{n} \frac{\partial^2 g^L}{\partial q^K \partial q^J}(\widetilde{Q}_i) d\widetilde{Q}_i^J \wedge d\widetilde{Q}_i^K\right) \\ & - \sum_{L=1}^{m}\left(\sum_{J=1}^{n} \frac{\partial g^L}{\partial q^J}(\widetilde{Q}_i) d\widetilde{Q}_i^J\right) \wedge d\Lambda_i^L.\end{aligned}$$

Since the second derivative of g^L is symmetric the expression in brackets in the first term vanishes. Moreover, since $g^L(\widetilde{Q}_i) = 0$ the expression in brackets in the second term also vanishes. This concludes the proof. ∎

Notice that by adding the terms $\overline{a}_{i0}v(y_0, z_0)$ in (3.2c) and $b_0 v(y_0, z_0)$ in (3.2e), we obtain in Theorem 4.1 the additional condition (4.1c) for $j = 0$. For $b_0 = 0$ this implies $\overline{a}_{i0} = 0$ if $\widetilde{b}_i \neq 0$ $(i = 0, 1, \ldots, \widetilde{s})$.

A consequence of Theorem 4.1 is the following.

COROLLARY 4.2. *We consider Lagrangian systems with holonomic constraints* (2.4) *satisfying the assumptions given in section* 2.2. *If the SPARK method* (3.2)

applied to (2.4) *satisfies* (4.1)*, then the numerical flow* $(q_0, v_0) \mapsto (q_1, v_1)$ *preserves on* W (2.6) *the Lagrangian symplectic* 2*-form* (2.10).

Proof. For systems without constraints this result was stated in [14]. The result follows from the equivalence between Hamiltonian and Lagrangian systems as described in section 2.3. Under assumption (2.5), Lagrangian system (2.4) with variables (q, v) can be reformulated in terms of an equivalent Hamiltonian system (2.1) with variables (q, p). A SPARK method (3.1) can be formally applied to this Hamiltonian system (2.1) and then rewritten in terms of the variables (q, v) of the Lagrangian form. This is in fact equivalent to applying a SPARK method (3.1) with the same coefficients directly to Lagrangian system (2.4). ∎

Assuming coefficients (b_i, a_{ij}) and $(\widehat{b}_i)$ are given, to satisfy the symplecticness conditions (4.1b) we must have

$$\widehat{a}_{ij} = \widehat{b}_j \left(1 - \frac{a_{ji}}{b_i}\right) \quad \text{for } i, j = 1, \dots, s, \quad \text{when } b_i \neq 0.$$

Assuming coefficients $(\widetilde{b}_i, \overline{a}_{ij})$ and (b_i) are given, to satisfy the symplecticness conditions (4.1c) we must have

$$\widetilde{a}_{ij} = \widetilde{b}_j \left(1 - \frac{\overline{a}_{ji}}{b_i}\right) \quad \text{for } i = 1, \dots, s, \quad j = 0, 1, \dots, \widetilde{s}, \quad \text{when } b_i \neq 0.$$

From the symplecticness condition (4.1c), the assumption $\overline{a}_{0j} = 0$ (3.3a) implies $b_j = 0$ or $\widetilde{a}_{j0} = \widetilde{b}_0$. We are thus particularly interested in SPARK methods satisfying

$$\widetilde{a}_{i0} = \widetilde{b}_0 \quad \text{for } i = 1, \dots, s. \tag{4.2}$$

From the symplecticness condition (4.1c), the assumption $\overline{a}_{\widetilde{s}j} = b_j$ implies $b_j = 0$ or $\widetilde{a}_{j\widetilde{s}} = 0$. We are thus particularly interested in SPARK methods satisfying

$$\widetilde{a}_{i\widetilde{s}} = 0 \quad \text{for } i = 1, \dots, s. \tag{4.3}$$

From this condition the algebraic variable $\Lambda_{\widetilde{s}}$ appears only in (3.2f) and is determined by (3.2h).

4.2. Symplectic SPARK methods are variational integrators. The application of a SPARK method to Lagrangian systems (2.4) with holonomic constraints and consistent initial values q_0, v_0 at t_0, i.e., $g(q_0) = 0$ and $g_q(q_0)v_0 = 0$, reads as

$$Q_i = q_0 + h \sum_{j=1}^{s} a_{ij} V_j \quad \text{for } i = 1, \dots, s, \tag{4.4a}$$

$$P_i = p_0 + h \sum_{j=1}^{s} \widehat{a}_{ij} F_j + h \sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij} R_j \quad \text{for } i = 1, \dots, s, \tag{4.4b}$$

$$\widetilde{Q}_i = q_0 + h \sum_{j=1}^{s} \overline{a}_{ij} V_j \quad \text{for } i = 0, 1, \dots, \widetilde{s}, \tag{4.4c}$$

$$0 = g(\widetilde{Q}_i) \quad \text{for } i = 0, 1, \dots, \widetilde{s}, \tag{4.4d}$$

$$q_1 = q_0 + h\sum_{j=1}^{s} b_j V_j, \tag{4.4e}$$

$$p_1 = p_0 + h\sum_{j=1}^{s} \widehat{b}_j F_j + h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j R_j, \tag{4.4f}$$

$$0 = g(q_1), \tag{4.4g}$$

$$0 = g_q(q_1)v_1, \tag{4.4h}$$

where

$$p_0 := L_v^T(q_0, v_0), \quad p_1 := L_v^T(q_1, v_1), \quad P_i := L_v^T(Q_i, V_i) \quad \text{for } i = 1, \dots, s,$$

$$F_i := L_q^T(Q_i, V_i) \quad \text{for } i = 1, \dots, s, \qquad R_i := -g_q^T(\widetilde{Q}_i)\Lambda_i \quad \text{for } i = 0, 1, \dots, \widetilde{s}.$$

When the SPARK coefficients satisfy symplecticness conditions (4.1), SPARK method (4.4) can also be derived from a variational point of view following the ideas introduced by Marsden and West [21]. Notice that the variational property in a backward analysis sense of symplectic PRK integrators was derived in [14]. The nonequivalent derivation of [7] would consider $V_1, \dots, V_s$ as independent variables and would remove the constraints (4.4b). This derivation would be difficult to apply in our context due to the presence of holonomic constraints.

Following Marsden and West [21], instead of considering the unknown quantities in (3.2) as implicit functions of q_0, v_0, and h, we consider them as implicit functions of q_0, q_1, and h. More precisely, assuming $g(q_0) = 0$ and $g(q_1) = 0$ we implicitly define as functions of q_0, q_1, and h the quantities $p_0, p_1, v_0, v_1, Q_i, P_i, V_i, F_i$ for $i = 1, \dots, s$ and $\widetilde{Q}_i, R_i, \Lambda_i$ for $i = 0, 1, \dots, \widetilde{s}$ by (4.4), except that we replace (4.4g) $g(q_1) = 0$ by $0 = g_q(q_0)v_0$. Formally speaking, we should make a distinction between the solution of (4.4) and the solution of (4.4) with the equation $g(q_1) = 0$ replaced by $0 = g_q(q_0)v_0$. In any case, the solution to one system is also the solution to the other under the assumptions $g(q_0) = 0$ and $g_q(q_0)v_0 = 0$ for the first system of equations and $g(q_0) = 0$ and $g(q_1) = 0$ for the second system of equations.

Considering the discrete action

$$A_d(q_0, q_1, h) := h\sum_{i=1}^{s} b_i L(Q_i, V_i) - h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \Lambda_i g(\widetilde{Q}_i),$$

we can show after some lengthy calculations (see the proof of Theorem 4.3) that when the SPARK coefficients satisfy the symplecticness assumptions (4.1), we have the relations

$$p_0 = -\nabla_1 A_d(q_0, q_1, h), \qquad p_1 = \nabla_2 A_d(q_0, q_1, h).$$

Therefore, the discrete Euler–Lagrange equations

$$\nabla_2 A_d(q_{n-1}, q_n, h) + \nabla_1 A_d(q_n, q_{n+1}, h) = 0$$

are satisfied for $n = 1, \dots, N-1$. This implies stationarity of the total discrete action

$$\sum_{n=1}^{N} A_d(q_{n-1}, q_n, h) \tag{4.5}$$

with respect to q_n for $n = 1, \ldots, N-1$. This is nothing else but a discrete version of Hamilton's principle applied to this sum (4.5). Therefore a SPARK symplectic integrator is also a variational integrator in this sense; more precisely, we have the following.

THEOREM 4.3. *For Lagrangian systems with holonomic constraints* (2.4) *and a corresponding SPARK method* (4.4)*, assume q_0 and q_N are fixed and consistent. Replace $0 = g(q_{n+1})$ for $n = 0, 1, \ldots, N-1$ by $0 = g_q(q_n)v_n$. If the SPARK coefficients satisfy symplecticness assumptions* (4.1)*, then we have a variational integrator in the sense of Marsden and West* [21]*; i.e., we have stationarity of the total discrete action* (4.5) *with respect to q_n for $n = 1, \ldots, N-1$.*

Proof. We show now the relation $-\nabla_1 A_d(q_0, q_1, h) = p_0$. We have

$$
\begin{aligned}
-\frac{\partial A_d}{\partial q_0}(q_0,q_1,h) &= -h\sum_{i=1}^{s} b_i L_q(Q_i,V_i)\frac{\partial Q_i}{\partial q_0} - h\sum_{i=1}^{s} b_i L_v(Q_i,V_i)\frac{\partial V_i}{\partial q_0} \\
&\quad + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \Lambda_i^T \left(g_q(\widetilde{Q}_i)\frac{\partial \widetilde{Q}_i}{\partial q_0}\right) + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g^T(\widetilde{Q}_i)\frac{\partial \Lambda_i}{\partial q_0} \\
&= -h\sum_{i=1}^{s} b_i F_i^T \left(I + h\sum_{j=1}^{s} a_{ij}\frac{\partial V_j}{\partial q_0}\right) - h\sum_{i=1}^{s} b_i P_i^T \frac{\partial V_i}{\partial q_0} \\
&\quad + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i \Lambda_i^T g_q(\widetilde{Q}_i) \left(I + h\sum_{j=1}^{s} \overline{a}_{ij}\frac{\partial V_j}{\partial q_0}\right) + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g^T(\widetilde{Q}_i)\frac{\partial \Lambda_i}{\partial q_0} \\
&= -h\sum_{i=1}^{s} b_i F_i^T I - h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} b_i a_{ij} F_i^T \frac{\partial V_j}{\partial q_0} \\
&\quad - h\sum_{i=1}^{s} b_i \left(p_0^T + h\sum_{j=1}^{s} \widehat{a}_{ij} F_j^T + h\sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij} R_j^T \right) \frac{\partial V_i}{\partial q_0} - h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i R_i^T I \\
&\quad - h^2 \sum_{i=0}^{\widetilde{s}}\sum_{j=1}^{s} \widetilde{b}_i \overline{a}_{ij} R_i^T \frac{\partial V_j}{\partial q_0} + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g^T(\widetilde{Q}_i)\frac{\partial \Lambda_i}{\partial q_0} \\
&= -h\sum_{j=1}^{s} b_j F_j^T I - h^2 \sum_{i=1}^{s}\sum_{j=1}^{s} (b_j a_{ji} + b_i \widehat{a}_{ij}) F_j^T \frac{\partial V_i}{\partial q_0} \\
&\quad - p_0^T h\sum_{i=1}^{s} b_i \frac{\partial V_i}{\partial q_0} - h^2 \sum_{i=1}^{s}\sum_{j=0}^{\widetilde{s}} b_i \widetilde{a}_{ij} R_j^T \frac{\partial V_i}{\partial q_0} - h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i R_i^T I \\
&\quad - h^2 \sum_{i=0}^{\widetilde{s}}\sum_{j=1}^{s} \widetilde{b}_i \overline{a}_{ij} R_i^T \frac{\partial V_j}{\partial q_0} + h\sum_{i=0}^{\widetilde{s}} \widetilde{b}_i g^T(\widetilde{Q}_i)\frac{\partial \Lambda_i}{\partial q_0}.
\end{aligned}
$$

From (4.4e) we have

$$
0 = I + h\sum_{i=1}^{s} b_i \frac{\partial V_i}{\partial q_0};
$$

hence

$$-\frac{\partial A_d}{\partial q_0}(q_0,q_1,h) = -h^2\sum_{i=1}^{s}\sum_{j=1}^{s}(b_j a_{ji} + b_i\widehat{a}_{ij} - b_j b_i)F_j^T\frac{\partial V_i}{\partial q_0} + p_0^T$$

$$-h^2\sum_{i=0}^{\widetilde{s}}\sum_{j=1}^{s}(b_j\widetilde{a}_{ji} + \widetilde{b}_i\overline{a}_{ij} - \widetilde{b}_i b_j)R_i^T\frac{\partial V_j}{\partial q_0} + h\sum_{i=0}^{\widetilde{s}}\widetilde{b}_i g^T(\widetilde{Q}_i)\frac{\partial \Lambda_i}{\partial q_0}.$$

From (4.4d) and symplecticness assumptions (4.1) we obtain the desired result of

$$-\frac{\partial A_d}{\partial q_0}(q_0,q_1,h) = p_0^T.$$

The relation $\nabla_2 A_d(q_0,q_1,h) = p_1$ can be shown in a similar way; thus we skip its proof. ∎

A consequence of Theorem 4.3 is the following.

COROLLARY 4.4. *For Hamiltonian systems with holonomic constraints* (2.1)*, assume q_0 and q_N are fixed and consistent. Replace the equations $0 = g(q_{n+1})$ for $n = 0,1,\ldots,N-1$ by $0 = g_q(q_n)H_p(q_n,p_n)$. If the SPARK coefficients satisfy symplecticness assumptions* (4.1)*, then we have a variational integrator in the sense of Marsden and West* [21]*; i.e., we have stationarity of the total discrete action*

$$\sum_{n=1}^{N} A_d(q_{n-1},q_n,h)$$

with respect to q_n for $n = 1,\ldots,N-1$, where

$$A_d(q_0,q_1,h) := h\sum_{i=1}^{s} b_i\ell(Q_i,P_i) - h\sum_{i=0}^{\widetilde{s}}\widetilde{b}_i\Lambda_i^T g(\widetilde{Q}_i)$$

and where $\ell(q,p) := p^T H_p^T(q,p) - H(q,p)$.

Proof. The result follows from the equivalence between Hamiltonian and Lagrangian systems described in section 2.3. Under assumption (2.2), Hamiltonian system (2.1) with variables (q,p) can be reformulated in terms of an equivalent Lagrangian system (2.4) with variables (q,v). A SPARK method (3.1) can be formally applied to this Lagrangian system (2.1) and then rewritten in terms of the variables (q,p) of the Hamiltonian form. This is in fact equivalent to applying a SPARK method (3.1) with the same coefficients directly to Hamiltonian system (2.1). ∎

5. Analysis of SPARK methods. In this section we give results about existence, uniqueness, local error, and global convergence of SPARK methods. Since SPARK methods are invariant under the change of variables (2.15) (see (2.16) and (2.17)), for the analysis we can simply consider $p(y,z) = z$ in (2.11b), under the assumption

$$g_y v_z r_\lambda \text{ is invertible.}$$

5.1. Existence and uniqueness. Generally there does not exist a solution to the nonlinear system of Definition 3.1 without any assumption on the coefficients of the SPARK method. For consistent SPARK methods satisfying (3.3), existence and

uniqueness for the nonlinear system can be shown under some additional assumptions (see Theorem 5.1). A very accurate value for λ_1 may be unnecessary. For a consistent SPARK method, by (3.3b) we have $\widetilde{c}_{\widetilde{s}} = 1$. Hence, a fairly good choice for λ_1 is given by $\lambda_1 := \Lambda_{\widetilde{s}}$ if one is not interested in enforcing constraints (2.14). The accuracy of the numerical λ-component does not influence the convergence of the (y, z)-components and the properties of the SPARK method. Existence and uniqueness for the system of nonlinear equations of SPARK methods (3.1) are shown in the following theorem.

THEOREM 5.1. *Suppose that* $y_0 = y_0(h), z_0 = z_0(h), \lambda_0 = \lambda_0(h)$ *satisfy*

$$0 = g(y_0), \tag{5.1a}$$

$$O(h^2) = g_y(y_0)v(y_0, z_0), \tag{5.1b}$$

$$\begin{aligned} O(h) = {} & g_{yy}(y_0)(v(y_0, z_0), v(y_0, z_0)) + g_y(y_0)v_y(y_0, z_0)v(y_0, z_0) \\ & + g_y(y_0)v_z(y_0, z_0)(f(y_0, z_0) + r(y_0, \lambda_0)), \end{aligned} \tag{5.1c}$$

where (2.12) *is satisfied in a neighborhood of* (y_0, z_0, λ_0). *Then for SPARK methods satisfying* (3.3) *and* $|h| \le h_0$ *there exists a locally unique SPARK solution of*

$$0 = Y_i - y_0 - h\sum_{j=1}^{s} a_{ij}v(Y_j, Z_j) \quad \text{for } i = 1, \dots, s, \tag{5.2a}$$

$$0 = Z_i - z_0 - h\sum_{j=1}^{s} \widehat{a}_{ij}f(Y_j, Z_j) - h\sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij}r(\widetilde{Y}_j, \Lambda_j) \quad \text{for } i = 1, \dots, s, \tag{5.2b}$$

$$0 = \widetilde{Y}_i - y_0 - h\sum_{j=1}^{s} \overline{a}_{ij}v(Y_j, Z_j) \quad \text{for } i = 0, 1, \dots, \widetilde{s}, \tag{5.2c}$$

$$0 = g(\widetilde{Y}_i) \quad \text{for } i = 0, 1, \dots, \widetilde{s}, \tag{5.2d}$$

$$0 = y_1 - y_0 - h\sum_{j=1}^{s} b_j v(Y_j, Z_j), \tag{5.2e}$$

$$0 = z_1 - z_0 - h\sum_{j=1}^{s} \widehat{b}_j f(Y_j, Z_j) - h\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j r(\widetilde{Y}_j, \Lambda_j), \tag{5.2f}$$

$$0 = g(y_1), \tag{5.2g}$$

$$0 = g_y(y_1)v(y_1, z_1), \tag{5.2h}$$

which satisfies

$$\begin{aligned} & & Y_i - y_0 = O(h) \quad & \text{for } i = 1, \dots, s, & & \\ \widetilde{Y}_0 = y_0, \quad & \widetilde{Y}_i - y_0 = O(h) \quad & \text{for } i = 1, \dots, \widetilde{s}, \quad & y_1 = \widetilde{Y}_{\widetilde{s}}, \\ & Z_i - z_0 = O(h) \quad & \text{for } i = 1, \dots, s, \quad & z_1 - z_0 = O(h), \\ & \Lambda_i - \lambda_0 = O(h) \quad & \text{for } i = 0, 1, \dots, \widetilde{s}. & \end{aligned}$$

Proof. The proof of this theorem can be done by application of the implicit function theorem, as in the proof of [12, Theorem V.4.1]. We have $\widetilde{Y}_0 = y_0$; hence $g(\widetilde{Y}_0) = 0$ is automatically satisfied by assumption. We have $\widetilde{Y}_{\widetilde{s}} = y_1$; hence (5.2g) can be removed since it is equivalent to (5.2d) for $i = \widetilde{s}$. We expand $g(\widetilde{Y}_i)$ for $i = 1, \dots, \widetilde{s}$ and $v(Y_i, Z_i)$ for $i = 1, \dots, s$ into Taylor series around y_0

$$
\begin{aligned}
g(\widetilde{Y}_i) &= g(y_0) + g_y(y_0)(\widetilde{Y}_i - y_0) \\
&\qquad + \int_0^1 (1-\tau) g_{yy}(y_0 + \tau(\widetilde{Y}_i - y_0)) d\tau (\widetilde{Y}_i - y_0, \widetilde{Y}_i - y_0), \\
v(Y_i, Z_i) &= v(y_0, z_0) + \int_0^1 v_y(y_0 + \tau(Y_i - y_0), z_0 + \tau(Z_i - z_0)) d\tau (Y_i - y_0) \\
&\qquad + \int_0^1 v_z(y_0 + \tau(Y_i - y_0), z_0 + \tau(Z_i - z_0)) d\tau (Z_i - z_0) \\
&= v(y_0, z_0) + h \int_0^1 v_y(y_0 + \tau(Y_i - y_0), z_0 + \tau(Z_i - z_0)) d\tau \sum_{j=1}^{s} a_{ij} v(Y_j, Z_j) \\
&\quad + h \int_0^1 v_z(y_0 + \tau(Y_i - y_0), z_0 + \tau(Z_i - z_0)) d\tau \left(\sum_{j=1}^{s} \widehat{a}_{ij} f(Y_j, Z_j) + \sum_{j=0}^{\widetilde{s}} \widetilde{a}_{ij} r(\widetilde{Y}_j, \Lambda_j) \right).
\end{aligned}
$$

Dividing $g(\widetilde{Y}_i)$ by h^2 and replacing the terms $\widetilde{Y}_i - y_0$, $Y_i - y_0$, and $Z_i - z_0$ by using (5.2a,b,c), we obtain

$$
\begin{aligned}
\frac{1}{h^2} g(\widetilde{Y}_i) &= \frac{1}{h^2} g(y_0) + \frac{1}{h} \sum_{j=1}^{s} \overline{a}_{ij} g_y(y_0) v(Y_j, Z_j) \\
&\qquad + \sum_{j=1}^{s} \sum_{k=1}^{s} \overline{a}_{ij} \overline{a}_{ik} \int_0^1 (1-\tau) g_{yy}(y_0 + \tau(\widetilde{Y}_i - y_0)) d\tau (v(Y_j, Z_j), v(Y_k, Z_k)) \\
&= \frac{1}{h^2} g(y_0) + \frac{1}{h} \sum_{j=1}^{s} \overline{a}_{ij} g_y(y_0) v(y_0, z_0) \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \overline{a}_{ij} a_{jk} g_y(y_0) \int_0^1 v_y(y_0 + \tau(Y_j - y_0), z_0 + \tau(Z_j - z_0)) d\tau v(Y_k, Z_k) \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \overline{a}_{ij} \widehat{a}_{jk} g_y(y_0) \int_0^1 v_z(y_0 + \tau(Y_j - y_0), z_0 + \tau(Z_j - z_0)) d\tau f(Y_k, Z_k) \\
&\quad + \sum_{j=1}^{s} \sum_{k=0}^{\widetilde{s}} \overline{a}_{ij} \widetilde{a}_{jk} g_y(y_0) \int_0^1 v_z(y_0 + \tau(Y_j - y_0), z_0 + \tau(Z_j - z_0)) d\tau r(\widetilde{Y}_k, \Lambda_k) \\
&\quad + \sum_{j=1}^{s} \sum_{k=1}^{s} \overline{a}_{ij} \overline{a}_{ik} \int_0^1 (1-\tau) g_{yy}(y_0 + \tau(\widetilde{Y}_i - y_0)) d\tau (v(Y_j, Z_j), v(Y_k, Z_k)).
\end{aligned}
$$

By (3.3c), for the values $Y_i := y_0$, $\widetilde{Y}_i := y_0$, $Z_i := z_0$, and $\Lambda_i = \lambda_0$ we obtain

$$\frac{1}{h^2}g(\widetilde{Y}_i) = \frac{\widetilde{c}_i^2}{2}\big(g_y(y_0)v_y(y_0,z_0)v(y_0,z_0) + g_y(y_0)v_z(y_0,z_0)f(y_0,z_0) + g_y(y_0)v_z(y_0,z_0)r(y_0,\lambda_0) + g_{yy}(y_0)(v(y_0,z_0),v(y_0,z_0))\big) = O(h).$$

Hence the values $Y_i(0) := y_0(0)$, $\widetilde{Y}_i(0) := y_0(0)$, $Z_i(0) := z_0(0)$, and $\Lambda_i(0) = \lambda_0(0)$ satisfy (5.2a,b,c) and

$$0 = \frac{1}{h^2}g(\widetilde{Y}_i) \tag{5.3}$$

$$= \sum_{j=1}^{s}\sum_{k=1}^{s}\overline{a}_{ij}a_{jk}g_y(y_0)\int_0^1 v_y(y_0+\tau(Y_j-y_0),z_0+\tau(Z_j-z_0))d\tau v(Y_k,Z_k) + \cdots.$$

Similarly we have

$$g_y(y_1) = g_y(y_0) + \int_0^1 g_{yy}(y_0+\tau(y_1-y_0))d\tau(y_1-y_0,\cdot)$$

$$= g_y(y_0) + h\sum_{j=1}^{s} b_j\int_0^1 g_{yy}(y_0+\tau(y_1-y_0))d\tau(v(Y_j,Z_j),\cdot),$$

$$v(y_1,z_1) = v(y_0,z_0) + \int_0^1 v_y(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau(y_1-y_0) + \int_0^1 v_z(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau(z_1-z_0)$$

$$= v(y_0,z_0) + h\int_0^1 v_y(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau\sum_{j=1}^{s} b_j v(Y_j,Z_j)$$

$$+ h\int_0^1 v_z(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau\left(\sum_{j=1}^{s}\widehat{b}_j f(Y_j,Z_j) + \sum_{j=0}^{\widetilde{s}}\widetilde{b}_j r(\widetilde{Y}_j,\Lambda_j)\right).$$

Hence, dividing $g_y(y_1)v(y_1,z_1)$ by h, we obtain

$$\frac{1}{h}g_y(y_1)v(y_1,z_1) = \frac{1}{h}g_y(y_0)v(y_0,z_0) + \sum_{j=1}^{s} b_j g_y(y_0)\int_0^1 v_y(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau v(Y_j,Z_j)$$

$$+ \sum_{j=1}^{s}\widehat{b}_j\int_0^1 v_z(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau f(Y_j,Z_j)$$

$$+ \sum_{j=0}^{\widetilde{s}}\widetilde{b}_j\int_0^1 v_z(y_0+\tau(y_1-y_0),z_0+\tau(z_1-z_0))d\tau r(\widetilde{Y}_j,\Lambda_j)$$

$$+ \sum_{j=1}^{s} b_j\int_0^1 g_{yy}(y_0+\tau(y_1-y_0))d\tau(v(Y_j,Z_j),v(y_1,z_1)).$$

By consistency $\sum_{j=1}^{s} b_j = 1$, $\sum_{j=1}^{s} \widehat{b}_j = 1$, $\sum_{j=0}^{\widetilde{s}} \widetilde{b}_j = 1$, for the values $Y_i := y_0$, $\widetilde{Y}_i := y_0$, $y_1 := y_0$, $Z_i := z_0$, $z_1 := z_0$, and $\Lambda_i = \lambda_0$, we obtain

$$\frac{1}{h} g_y(y_1) v(y_1, z_1) = g_y(y_0) v_y(y_0, z_0) v(y_0, z_0) + g_y(y_0) v_z(y_0, z_0) f(y_0, z_0) + g_y(y_0) v_z(y_0, z_0) r(y_0, \lambda_0) + g_{yy}(y_0)(v(y_0, z_0), v(y_0, z_0)) = O(h).$$

Hence the values $Y_i(0) := y_0(0)$, $\widetilde{Y}_i(0) := y_0(0)$, $y_1(0) := y_0(0)$, $Z_i(0) := z_0(0)$, $z_1(0) := z_0(0)$, and $\Lambda_i(0) = \lambda_0(0)$ satisfy

$$0 = \frac{1}{h} g_y(y_1) v(y_1, z_1) = \sum_{j=1}^{s} b_j g_y(y_0) \int_0^1 v_y(y_0 + \tau(y_1 - y_0), z_0 + \tau(z_1 - z_0)) d\tau v(Y_j, Z_j) + \cdots . \tag{5.4}$$

Replacing y_1 and z_1 in (5.3)–(5.4) by using (5.2e,f), and using tensor matrix product notations, we see that the Jacobian of (5.2a,b,c), (5.3), and (5.4) with respect to Y_i $(i = 1, \ldots, s)$, Z_i $(i = 1, \ldots, s)$, $\widetilde{Y}_i$ $(i = 1, \ldots, \widetilde{s})$, and Λ_i $(i = 0, 1, \ldots, \widetilde{s})$ is of the form

$$\begin{pmatrix} I_{sn_y} + O(h) & O & O(h) & O \\ O(h) & I_{sn_z} + O(h) & O(h) & O(h) \\ O(h) & O(h) & I_{\widetilde{s}n_y} & O \\ O(1) & O(1) & O(1) & \begin{pmatrix} N \\ \widetilde{b}^T \end{pmatrix} \otimes g_y(y_0) v_z(y_0, z_o) r_\lambda(y_0, \lambda_0) + O(h) \end{pmatrix}$$

with N as defined in (3.3d). This Jacobian matrix is invertible for $|h| \le h_0$ sufficiently small. Therefore, the implicit function theorem yields the existence of a locally unique solution to (5.2a,b,c), (5.3), and (5.4), and hence to the corresponding SPARK method (5.2). ∎

5.2. Local error of the (s, s)-Gauss–Lobatto SPARK methods. A thorough local error analysis of the whole class of SPARK methods (3.2) based on using simplifying assumptions is beyond the scope of this paper. SPARK methods include a class of PRK methods whose local error analysis based on trees is long and technical [12, 13]. For Lobatto IIIA-B methods, an alternative proof using the idea of discontinuous collocation can be found in [7, section VII.1]. Here we will analyze only the local error of the (s, s)-Gauss–Lobatto SPARK methods as defined in subsection 3.3.

THEOREM 5.2. *Consider the system of ODAEs* (2.11), *consistent initial values* y_0, z_0 *at* t_0, *where* (2.12) *is satisfied in a neighborhood of* (y_0, z_0, λ_0). *Then for* $|h| \le h_0$ *the local error of the* (s, s)*-Gauss–Lobatto SPARK methods satisfies*

$$y_1 - y(t_0 + h) = O(h^{2s+1}), \quad z_1 - z(t_0 + h) = O(h^{2s+1}). \tag{5.5}$$

Proof. For the proof we can consider $p(y, z) = z$ in (2.11b). To prove this theorem we use the same techniques of proof as used in [11] for collocation methods. We define the polynomials $Y(t)$, $\widetilde{Y}(t)$, $Z(t)$, and $\Lambda(t)$ of degree s by

$$Y(t) = \sum_{i=0}^{s} \ell_i \left(\frac{t - t_0}{h} \right) Y_i, \qquad \widetilde{Y}(t) = \sum_{i=0}^{s} \widetilde{\ell}_i \left(\frac{t - t_0}{h} \right) \widetilde{Y}_i,$$

$$Z(t)=\sum_{i=0}^{s}\ell_i\left(\frac{t-t_0}{h}\right)Z_i,\qquad \Lambda(t)=\sum_{i=0}^{s}\widetilde{\ell}_i\left(\frac{t-t_0}{h}\right)\Lambda_i,$$

where

$$\ell_i(\tau):=\prod_{\substack{j=0\\ j\neq i}}^{s}\left(\frac{\tau-c_j}{c_i-c_j}\right),\qquad \widetilde{\ell}_i(\tau):=\prod_{\substack{j=0\\ j\neq i}}^{s}\left(\frac{\tau-\widetilde{c}_j}{\widetilde{c}_i-\widetilde{c}_j}\right),$$

$c_0:=0$, $Y_0:=y_0$, and $Z_0:=z_0$. We have $Y(t_0)=\widetilde{Y}(t_0)=y_0$, $Z(t_0)=z_0$, $Y(t_0+h)=\widetilde{Y}(t_0+h)=y_1$, $Z(t_0+h)=z_1$, and

$$Y'(t)=v(Y(t),Z(t))+\delta(t), \tag{5.6a}$$

$$\widetilde{Y}'(t)=v(Y(t),Z(t))+\widetilde{\delta}(t), \tag{5.6b}$$

$$Z'(t)=f(Y(t),Z(t))+r(\widetilde{Y}(t),\Lambda(t))+\mu(t), \tag{5.6c}$$

$$0=g(\widetilde{Y}(t))+\widetilde{\theta}(t), \tag{5.6d}$$

$$0=g_y(\widetilde{Y}(t))(v(Y(t),Z(t))+\widetilde{\delta}(t))+\widetilde{\theta}'(t), \tag{5.6e}$$

with defects $\delta(t)$, $\widetilde{\delta}(t)$, $\mu(t)$, $\widetilde{\theta}(t)$ satisfying

$$\begin{aligned}
\delta(t_0+c_ih)&=0 \quad \text{for } i=1,\dots,s,\\
\widetilde{\delta}(t_0+\widetilde{c}_ih)&=0 \quad \text{for } i=0,1,\dots,s,\\
\mu(t_0+c_ih)&=0 \quad \text{for } i=1,\dots,s,\\
\widetilde{\theta}(t_0+\widetilde{c}_ih)&=0 \quad \text{for } i=0,1,\dots,s,\\
\widetilde{\theta}'(t_0)&=-g_y(\widetilde{Y}(t_0))\widetilde{\delta}(t_0)=0,\\
\widetilde{\theta}'(t_0+h)&=-g_y(\widetilde{Y}(t_0+h))\widetilde{\delta}(t_0+h)=0.
\end{aligned}$$

The exact solution $(y(t),y(t),z(t),\lambda(t))$ satisfies the same above relations (5.6) with $\delta(t)\equiv 0$, $\widetilde{\delta}(t)\equiv 0$, $\mu(t)\equiv 0$, and $\widetilde{\theta}(t)\equiv 0$. One more differentiation of (5.6e) yields

$$\begin{aligned}
0=&\,g_{yy}(\widetilde{Y}(t))(v(Y(t),Z(t))+\widetilde{\delta}(t),v(Y(t),Z(t))+\widetilde{\delta}(t))\\
&+g_y(\widetilde{Y}(t))v_y(Y(t),Z(t))(v(Y(t),Z(t))+\delta(t))\\
&+g_y(\widetilde{Y}(t))v_z(Y(t),Z(t))(f(Y(t),Z(t))+r(\widetilde{Y}(t),\Lambda(t))+\mu(t))\\
&+g_y(\widetilde{Y}(t))\widetilde{\delta}'(t)+\widetilde{\theta}''(t).
\end{aligned}$$

We can express $\Lambda(t)$ from this equation as an implicit function

$$\Lambda(t)=\Upsilon(Y(t),\widetilde{Y}(t),Z(t),\delta(t),\widetilde{\delta}(t),\widetilde{\delta}'(t),\mu(t),\widetilde{\theta}''(t)).$$

Inserting this relation into (5.6c), we obtain the system of ODEs

$$Y'(t) = v(Y(t), Z(t)) + \delta(t),$$

$$\widetilde{Y}'(t) = v(Y(t), Z(t)) + \widetilde{\delta}(t),$$

$$Z'(t) = f(Y(t), Z(t)) + r(\widetilde{Y}(t), \Upsilon(Y(t), \widetilde{Y}(t), Z(t), \delta(t), \widetilde{\delta}(t), \widetilde{\delta}'(t), \mu(t), \widetilde{\theta}''(t))) + \mu(t).$$

To apply the Gröbner–Alekseev formula [8, Theorem I.14.5] we need the defect $d(t) := (d_1(t), d_2(t), d_3(t))^T$:

$$d_1(t) := Y'(t) - v(Y(t), Z(t)) = \delta(t),$$

$$d_2(t) := \widetilde{Y}'(t) - v(Y(t), Z(t)) = \widetilde{\delta}(t),$$

$$d_3(t) := Z'(t) - f(Y(t), Z(t)) - r(\widetilde{Y}(t), \Upsilon(Y(t), \widetilde{Y}(t), Z(t), 0, 0, 0, 0, 0)).$$

We have

$$d_3(t) = \Phi_3(t, 1) - \Phi_3(t, 0) = \int_0^1 \frac{\partial \Phi_3}{\partial \tau}(t, \tau) d\tau,$$

where

$$\Phi_3(t, \tau) := r(\widetilde{Y}(t), \Upsilon(Y(t), \widetilde{Y}(t), Z(t), \tau\delta(t), \tau\widetilde{\delta}(t), \tau\widetilde{\delta}'(t), \tau\mu(t), \tau\widetilde{\theta}''(t))) + \tau\mu(t).$$

Hence, we get

$$d_3(t) = Q_1(t)\delta(t) + Q_2(t)\widetilde{\delta}(t) + Q_3(t)\widetilde{\delta}'(t) + (I + Q_4(t))\mu(t) + Q_5(t)\widetilde{\theta}''(t),$$

where we give only the expressions of $Q_3(t)$ and $Q_5(t)$:

$$Q_3(t) = -\int_0^1 (r_\lambda (g_y v_z r_\lambda)^{-1} g_y)(Y(t), \widetilde{Y}(t), Z(t), \Upsilon(Y(t), \widetilde{Y}(t), Z(t), \tau\delta(t), \tau\widetilde{\delta}(t), \tau\widetilde{\delta}'(t), \tau\mu(t), \tau\widetilde{\theta}''(t)) d\tau,$$

$$Q_5(t) = -\int_0^1 (r_\lambda (g_y v_z r_\lambda)^{-1}(Y(t), \widetilde{Y}(t), Z(t), \Upsilon(Y(t), \widetilde{Y}(t), Z(t), \tau\delta(t), \tau\widetilde{\delta}(t), \tau\widetilde{\delta}'(t), \tau\mu(t), \tau\widetilde{\theta}''(t)) d\tau.$$

We denote the resolvent of the exact solution

$$R(t, s) := R(t, s, y_s, \widetilde{Y}_s, z_s) = \frac{\partial(y, \widetilde{Y}, z)}{\partial(y_s, \widetilde{Y}_s, z_s)}(t, s, y_s, \widetilde{Y}_s, z_s).$$

From the Gröbner–Alekseev formula we have

$$\begin{pmatrix} Y(t) - y(t) \\ \widetilde{Y}(t) - y(t) \\ Z(t) - z(t) \end{pmatrix} = \int_{t_0}^t R(t, s) d(s) ds$$

$$= \int_{t_0}^t S_1(t, s)\delta(s) + S_2(t, s)\widetilde{\delta}(s) + S_3(t, s)\widetilde{\delta}'(s) + S_4(t, s)\mu(s) + S_5(t, s)\widetilde{\theta}''(s) ds,$$

where

$$S_1(t,s)=R(t,s)\begin{pmatrix} I \\ O \\ Q_1(s) \end{pmatrix}, \qquad S_2(t,s)=R(t,s)\begin{pmatrix} O \\ I \\ Q_2(s) \end{pmatrix},$$

$$S_3(t,s)=R(t,s)\begin{pmatrix} O \\ O \\ Q_3(s) \end{pmatrix}, \qquad S_4(t,s)=R(t,s)\begin{pmatrix} O \\ O \\ I+Q_4(s) \end{pmatrix},$$

$$S_5(t,s)=R(t,s)\begin{pmatrix} O \\ O \\ Q_5(s) \end{pmatrix}.$$

Hence, by integration by parts, we obtain

$$\begin{pmatrix} Y(t)-y(t) \\ \widetilde{Y}(t)-y(t) \\ Z(t)-z(t) \end{pmatrix} = S_3(t,s)\widetilde{\delta}(s) - \frac{\partial S_5}{\partial s}(t,s)\widetilde{\theta}(s) + S_5(t,s)\widetilde{\theta}'(s)\Big|_{s=t_0}^{t}$$

$$+\int_{t_0}^{t}\sigma(t,s)ds+\int_{t_0}^{t}\widetilde{\sigma}(t,s)ds,$$

where

$$\sigma(t,s):=S_1(t,s)\delta(s)+S_4(t,s)\mu(s),$$

$$\widetilde{\sigma}(t,s):=\left(S_2(t,s)-\frac{\partial S_3}{\partial s}(t,s)\right)\widetilde{\delta}(s)+\frac{\partial^2 S_5}{\partial s^2}(t,s)\widetilde{\theta}(s).$$

We have $\widetilde{\delta}(t_0)=0=\widetilde{\delta}(t_0+h)$, $\widetilde{\theta}(t_0)=0=\widetilde{\theta}(t_0+h)$, $\widetilde{\theta}'(t_0)=0=\widetilde{\theta}'(t_0+h)$; hence at $t=t_0+h$ we are left with

$$\begin{pmatrix} y_1-y(t_0+h) \\ y_1-y(t_0+h) \\ z_1-z(t_0+h) \end{pmatrix} = \int_{t_0}^{t_0+h}\sigma(t_0+h,s)ds+\int_{t_0}^{t_0+h}\widetilde{\sigma}(t_0+h,s)ds.$$

Applying the Gauss quadrature formula with s nodes of order $2s$ for the first integral, and the Lobatto quadrature formula with $s+1$ nodes of order $2s$ for the second integral, we obtain

$$\int_{t_0}^{t_0+h}\sigma(t_0+h,s)ds=h\sum_{i=1}^{s}\sigma(t_0+h,t_0+c_ih)+O(h^{2s+1}),$$

$$\int_{t_0}^{t_0+h}\widetilde{\sigma}(t_0+h,s)ds=h\sum_{i=0}^{s}\widetilde{\sigma}(t_0+h,t_0+\widetilde{c}_ih)+O(h^{2s+1}),$$

and since $\sigma(t,t_0+c_ih)=0$ for $i=1,\dots,s$ and $\widetilde{\sigma}(t,t_0+\widetilde{c}_ih)=0$ for $i=0,1,\dots,s$, this leads to the desired result (5.5). □

5.3. Global convergence of SPARK methods. Once local error estimates of SPARK methods are known, global convergence results can be obtained without too much difficulty.

THEOREM 5.3. *Consider the system of ODAEs* (2.11) *under assumptions* (2.12) *and a SPARK method* (3.2) *of local order p satisfying assumptions* (3.3). *Then it is globally convergent of order p, i.e.,*

$$y_n - y(t_n) = O(h^p), \quad z_n - z(t_n) = O(h^p)$$

for $t_n - t_0 = nh \leq \text{Const}$.

Proof. For the proof we can consider $p(y,z) = z$ in (2.11b). Replacing (5.2d,g,h), respectively, by

$$\begin{aligned} g(\widetilde{Y}_i) &= g(y_0) + h\widetilde{c}_i g_y(y_0)v(y_0,z_0) \quad \text{for } i = 0,1,\dots,\widetilde{s}, \\ g(y_1) &= g(y_0) + h g_y(y_0)v(y_0,z_0), \\ g_y(y_1)v(y_1,z_1) &= g_y(y_0)v(y_0,z_0) \end{aligned}$$

extends the definition of SPARK methods to a neighborhood of (y_0,z_0) in $\mathbb{R}^{n_y}\times\mathbb{R}^{n_z}$; i.e., SPARK methods are not restricted to just the manifold of constraints $\{(y,z)\in \mathbb{R}^{n_y}\times\mathbb{R}^{n_z} | 0 = g(y), 0 = g_y(y)v(y,z)\}$. Hence, SPARK methods can be locally expressed as a mapping

$$\begin{pmatrix} y_{n+1} \\ z_{n+1} \end{pmatrix} = \begin{pmatrix} y_n \\ z_n \end{pmatrix} + h_n \Phi(h_n, y_n, z_n)$$

from $\mathbb{R}^{n_y}\times\mathbb{R}^{n_z}$ to $\mathbb{R}^{n_y}\times\mathbb{R}^{n_z}$. Hence, classical convergence results, like those for RK methods applied to ODEs, can then be applied [8]. ∎

For the (s,s)-Gauss–Lobatto SPARK methods, as a consequence of Theorem 4.1, Corollary 4.2, Theorem 4.3, Corollary 4.4, and Theorems 5.2 and 5.3, we can now state a major result of this paper.

COROLLARY 5.4. *Consider the system of ODAEs* (2.11) *under assumptions* (2.12). *The* (s,s)*-Gauss–Lobatto SPARK method* (3.2) *is constraint-preserving, symmetric, and of maximal order* $2s$, *i.e.,*

$$y_n - y(t_n) = O(h^{2s}), \quad z_n - z(t_n) = O(h^{2s})$$

for $|t_n - t_0| \leq \text{Const}$ *and* $h := \max(|h_1|,\dots,|h_n|)$. *For holonomically constrained Hamiltonian systems* (2.1) *and Lagrangian systems* (2.4) *these methods are also symplectic and variational.*

6. Numerical experiments. Figure 6.3

To illustrate Corollary 5.4, we have applied (s,s)-Gauss–Lobatto SPARK methods with constant stepsize h to the following system of ODAEs:

$$\begin{pmatrix} y_1' \\ y_2' \end{pmatrix} = \begin{pmatrix} 2z_1 \\ -z_2 \end{pmatrix}, \tag{6.1a}$$

$$\begin{pmatrix} z_1' \\ z_2' \end{pmatrix} = \begin{pmatrix} 2y_1y_2z_1z_2 - y_1z_1z_2 \\ z_1 - y_1z_2^3 \end{pmatrix} + \begin{pmatrix} y_1y_2\lambda_1^2 \\ -\sqrt{y_1}\lambda_1 \end{pmatrix}, \tag{6.1b}$$

$$0 = y_1y_2^2 - 1, \tag{6.1c}$$

$$0 = 2y_2(z_1y_2 - y_1z_2). \tag{6.1d}$$

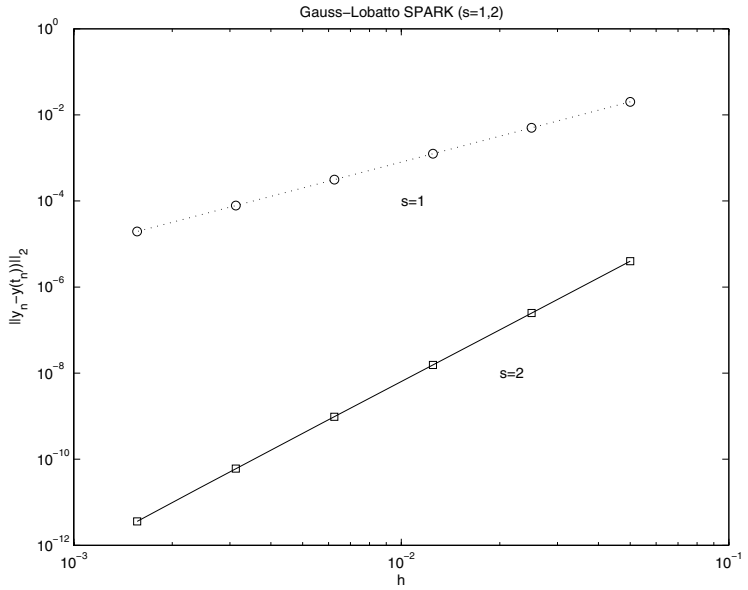


FIG. 6.1. *Global error in y at $t_n = 1$ of (s,s)-Gauss–Lobatto SPARK methods $(s = 1, 2)$ applied with various constant stepsizes h to the test problem* (6.1).

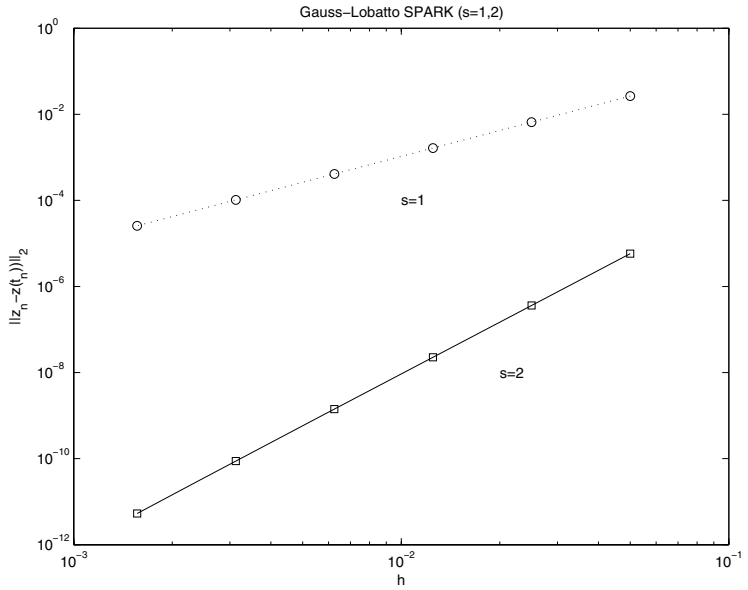


FIG. 6.2. *Global error in z at $t_n = 1$ of (s,s)-Gauss–Lobatto SPARK methods $(s = 1, 2)$ applied with various constant stepsizes h to the test problem* (6.1).

For the initial conditions $y_1(0) = y_2(0) = z_1(0) = z_2(0) = 1$ at $t_0 = 0$, the exact solution to this test problem is given by $y_1(t) = z_1(t) = e^{2t}$, $y_2(t) = z_2(t) = e^{-t}$, $\lambda_1(t) = e^t$. We have plotted in Figures 6.1 and 6.2 the global errors for the y- and z-components at $t_n = 1$ with respect to various constant stepsizes h. Logarithmic scales have been used so that a curve appears as a straight line of slope k whenever

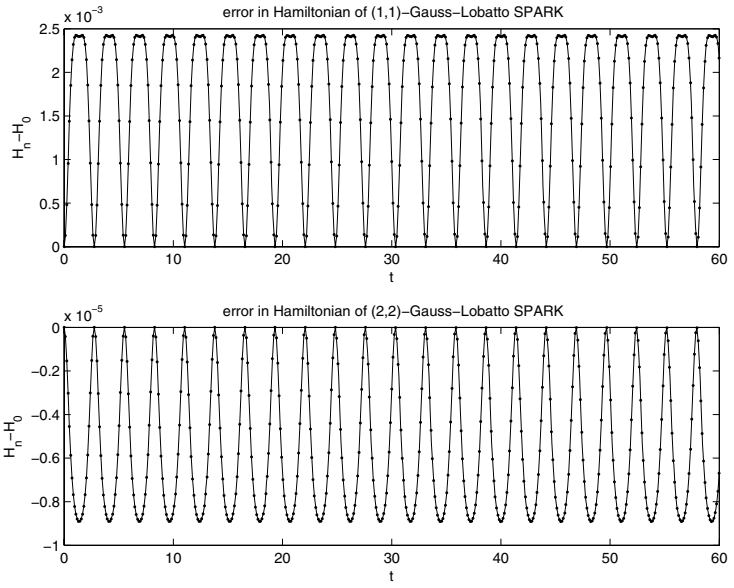


FIG. 6.3. *Error in Hamiltonian of* (s,s)*-Gauss–Lobatto SPARK methods* $(s=1,2)$ *applied with constant stepsize* $h=0.12$ *to the test problem* (6.2).

the leading term of the global error is of order k, i.e., when $\|y_n - y(t_n)\| = O(h^k)$. For the (s,s)-Gauss–Lobatto SPARK methods with $s=1,2$ of order $2s=2,4$ we observe straight lines of slope $2s=2,4$, thus confirming the orders of convergence predicted by Corollary 5.4.

As a second test problem, we consider the motion of a particle of mass m and electric charge e under the influence of an electric field $(0,0,E)^T$ and a magnetic field $(0,0,B)^T$ and restricted to a sphere of radius R [4, Problem 7.16]. This system can be described in term of Cartesian coordinates $(q_1,q_2,q_3)^T$ and generalized momenta $(p_1,p_2,p_3)^T$ with a nonseparable Hamiltonian

$$H = \frac{1}{2m}((p_1+m\omega q_2)^2 + (p_2 - m\omega q_1)^2 + p_3^2) - eEq_3 \tag{6.2a}$$

with $\omega := eB/(2mc)$ and holonomic constraint

$$\sqrt{q_1^2+q_2^2+q_3^2} - R = 0. \tag{6.2b}$$

We choose the parameters

$$m=1, \quad \omega=1, \quad R=1, \quad eE=1$$

and initial conditions

$$q_1(0)=0.2,\ q_2(0)=0.2,\ q_3(0)=\sqrt{0.92},\ p_1(0)=1,\ p_2(0)=-1,\ p_3(0)=0.$$

In Figure 6.3 we plot the Hamiltonian error of (s,s)-Gauss–Lobatto SPARK methods $(s=1,2)$ applied with constant stepsize $h=0.12$ to this system. As expected for a symplectic integrator, we observe that the Hamiltonian error remains bounded and small over long-time intervals.

7. Conclusion. We have considered a general class of ODAEs, and, more particularly, a unified formulation of Hamiltonian and Lagrangian systems with holonomic constraints. We have defined the application of SPARK methods for these systems, including in particular the new (s, s)-Gauss–Lobatto SPARK methods and also well-known schemes such as the Lobatto IIIA-B PRK methods. SPARK methods preserve the constraints. The (s, s)-Gauss–Lobatto SPARK methods have been proved to be of optimal order of convergence $2s$. For Hamiltonian and Lagrangian systems with holonomic constraints, these methods have also been shown to be symplectic and to preserve the variational nature of trajectories.

REFERENCES

[1] V. I. Arnold, *Mathematical Methods of Classical Mechanics*, 2nd ed., Grad. Texts in Math. 60, Springer-Verlag, New York, 1989.

[2] K. E. Brenan, S. L. Campbell, and L. R. Petzold, *Numerical Solution of Initial-Value Problems in Differential-Algebraic Equations*, Classics in Appl. Math., SIAM, Philadelphia, 1996.

[3] J. C. Butcher, *Numerical Methods for Ordinary Differential Equations*, 2nd ed., John Wiley & Sons, Chichester, 2003.

[4] P. Choquard, *Mécanique Analytique*, Vol. 1, of Cahiers Math. Ecole Polytech. Fédérale de Lausanne, Presses Polytechniques et Universitaires Romandes, Lausanne, Switzerland, 1992.

[5] E. Hairer and L. O. Jay, *Implicit Runge–Kutta Methods for Higher Index Differential-Algebraic Systems*, in Contributions in Numerical Mathematics, World Sci. Ser. Appl. Anal. 2, World Sci. Publ., River Edge, NJ, 1993, pp. 213–224.

[6] E. Hairer, Ch. Lubich, and M. Roche, *The Numerical Solution of Differential-Algebraic Systems by Runge–Kutta Methods*, Lecture Notes in Math. 1409, Springer, Berlin, 1989.

[7] E. Hairer, Ch. Lubich, and G. Wanner, *Geometric Numerical Integration*, 2nd ed., Comput. Math. 31, Springer, Berlin, 2006.

[8] E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations* I. *Nonstiff Problems*, 2nd revised ed., Comput. Math. 18, Springer, Berlin, 1993.

[9] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, 2nd revised ed., Comput. Math. 14, Springer, Berlin, 1996.

[10] E. J. Haug, *Computer Aided Kinematics and Dynamics of Mechanical Systems. Volume* I: *Basic Methods*, Allyn and Bacon, Boston, MA, 1989.

[11] L. O. Jay, *Collocation methods for differential-algebraic equations of index* 3, Numer. Math., 65 (1993), pp. 407–421.

[12] L. O. Jay, *Runge–Kutta Type Methods for Index Three Differential-Algebraic Equations with Applications to Hamiltonian Systems*, Ph.D. thesis, Department of Mathematics, University of Geneva, Geneva, Switzerland, 1994.

[13] L. Jay, *Symplectic partitioned Runge–Kutta methods for constrained Hamiltonian systems*, SIAM J. Numer. Anal., 33 (1996), pp. 368–387.

[14] L. O. Jay, *Lagrangian Integration with Symplectic Methods*, Tech. report 97-009, Army High Performance Computing Research Center, University of Minnesota, Minneapolis, MN, 1997.

[15] L. O. Jay, *Structure preservation for constrained dynamics with super partitioned additive Runge–Kutta methods*, SIAM J. Sci. Comput., 20 (1998), pp. 416–446.

[16] L. O. Jay, *Iterative solution of nonlinear equations for SPARK methods applied to DAEs*, Numer. Algorithms, 31 (2002), pp. 171–191.

[17] L. O. Jay, *A Note on the Symplectic Euler Method for DAEs with Holonomic Constraints*, Tech. report 165, Department of Mathematics, University of Iowa, Iowa City, IA, 2006.

[18] L. O. Jay, *Specialized Runge–Kutta methods for index* 2 *differential algebraic equations*, Math. Comput., 75 (2006), pp. 641–654.

[19] B. Leimkuhler and S. Reich, *Simulating Hamiltonian Dynamics*, Cambridge Monogr. Appl. Comput. Math. 14, Cambridge University Press, Cambridge, UK, 2004.

[20] J. E. Marsden and T. S. Ratiu, *Introduction to Mechanics and Symmetry*, Texts Applied Math. 17, Springer, New York, 1994.

[21] J. E. Marsden and M. West, *Discrete mechanics and variational integrators*, Acta Numer., 10 (2001), pp. 357–514.

[22] P. J. Rabier and W. C. Rheinboldt, *Nonholonomic Motion of Rigid Mechanical Systems from a DAE Viewpoint*, SIAM, Philadelphia, 2000.
[23] S. Reich, *Symplectic Integration of Constrained Hamiltonian Systems by Runge–Kutta Methods*, Tech. report 93-13, Department of Computer Science, University of British Columbia, Vancouver, BC, Canada, 1993.
[24] S. Reich, *Symplectic integration of constrained Hamiltonian systems by composition methods*, SIAM J. Numer. Anal., 33 (1996), pp. 475–491.
[25] J. M. Sanz-Serna and M. P. Calvo, *Numerical Hamiltonian Problems*, Appl. Math. Math. Comput. 7, Chapman & Hall, London, 1994.
[26] W. O. Schiehlen, ed., *Multibody Systems Handbook*, Springer-Verlag, Berlin, 1990.
[27] W. O. Schiehlen, ed., *Advanced Multibody System Dynamics, Simulation and Software Tools*, Kluwer Academic Publishers, London, 1993.
[28] A. M. Stuart and A. R. Humphries, *Dynamical Systems and Numerical Analysis*, Cambridge Monogr. Appl. Comput. Math. 2, Cambridge University Press, Cambridge, UK, 1998.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1843–1870

A SMALL EDDY CORRECTION METHOD FOR A 3D NAVIER–STOKES-TYPE SYSTEM OF EQUATIONS RELATED TO THE PRIMITIVE EQUATIONS OF THE OCEAN*

T. TACHIM MEDJO†

Abstract. Considering the interaction between the baroclinic and barotropic flows and using the idea of the Newton iteration, a small eddy correction method is proposed for approximating and numerically solving the primitive equations of the ocean. We assume that the barotropic approximation to the solution is known. Formally applying the Newton iterative procedure to the baroclinic flow equation, we then generate approximate systems. It is shown that the first step leads to the well-known quasi-geostrophic equations. The convergence analysis is presented and the results show that the small eddy correction method can greatly improve the accuracy of the quasi-geostrophic approximate solution. More precisely, we prove that the approximate system derived from the procedure converges to the original primitive equations, and we estimate the rate of convergence as a function of the aspect ratio of the domain. Some numerical simulations are presented to illustrate the method.

Key words. primitive equations, barotropic flow, baroclinic flow, small eddy

AMS subject classifications. 65N12, 49K35, 76D55

DOI. 10.1137/05063074X

1. Introduction. In dynamical systems theory the objective is to study the long-term behavior of solutions of an evolution equation. When the equation is dissipative all solutions converge as $t \longmapsto \infty$ to a complicated set $\mathcal{A}$, the global attractor, which may be a fractal set. This set embodies the large-time dynamics of the equations, corresponding to all sorts of regimes, including turbulent ones. Although this set may be fairly complicated, in general it has finite dimension [26]. Despite the considerable increase in available computing power during the past few years, the numerical approximation of the global attractor remains a difficult task especially for important systems such as the Navier–Stokes equations or the primitive equations (PEs) of the ocean. For the Navier–Stokes flows, there are some approaches to deriving simplified behavioral laws for the smallest structure set in motion with the aim of reducing the computational cost [9, 24]. In the nonlinear Galerkin (NLG) method introduced in [24], the small scales are given as a function of the large scales, and the nonlinear interaction between the large and the small scales is only approximately modeled. In [13], the authors presented a small eddy correction method for the two-dimensional (2D) Navier–Stokes equations. It is shown that the first step of this iterative method leads to the standard Galerkin method and the second step yields the NLG method.

Although the source of the extensive scale variability differs for the Navier–Stokes and the PE models (the scale variability in the Navier–Stokes system is mainly the result of the nonlinear term, while the sources are more varied for the PE model), there exists an energy cascade that is similar for the two models, and for which one can apply the main principle of description given by Charney [5].

*Received by the editors May 5, 2005; accepted for publication (in revised form) February 23, 2007; published electronically August 31, 2007. This work was supported in part by NSF grant DMS0305110, DOE grant DE-FG02-01ER63251:A000, and by the Research Fund of Indiana University.

http://www.siam.org/journals/sinum/45-5/63074.html

†Department of Mathematics, Florida International University, DM413B, University Park, Miami, FL 33199 (tachimt@fiu.edu).

Every mode undergoes constraints due to wind. Even for a fairly constant wind, there is still an infinite number of modes stimulated by the boundary conditions. These modes will exhibit different behaviors with respect to the stimulations based on their position in the spectrum. They can be grouped into three categories:

- At the largest scales, geophysical flows such as the ocean and the atmosphere are essentially 2D (barotropic component). These barotropic modes transmit their energy in the following two ways:
  - (i) at modes of greater dimension, through an inverse kinetic energy cascade; the surplus energy is then dissipated by the boundary conditions;
  - (ii) at modes of smaller dimension, through an entropy barotropic cascade.
- At the medium scales, we have the baroclinic modes. These modes will redistribute their energy as a baroclinic energy cascade, thus transporting the energy to the viscous dispersal area. This cascade is similar to the energy cascade predicted by Kolmogorov for the Navier–Stokes system [9, 8].
- At the very small scales, the energy provided by the surface forces is insufficient to oppose the viscous dispersion constraint.

Given the similarities with the Navier–Stokes system and inspired by the results obtained for the 2D Navier–Stokes equations with the small eddy correction method [13] and the NLG method [24, 12, 11], we present in this article a small eddy correction method for the PEs of the ocean. Considering the interaction between the baroclinic and barotropic flows and using the idea of the Newton iteration, a small eddy correction method is proposed for approximating and numerically solving the PEs of the ocean. We assume that the barotropic approximation to the solution is known. Formally applying the Newton iterative procedure to the baroclinic flow equation, we then generate approximate systems. It is shown that the first step leads to the well-known quasi-geostrophic equations. The convergence analysis is presented and the results show that the small eddy correction method can greatly improve the accuracy of the quasi-geostrophic approximate solution. More precisely, we prove that the approximate system derived from the procedure converges to the PEs of the ocean with a rate of convergence $O((\delta^{1/2})^{2^{l-1}})$, where δ is the shape ratio of the ocean and l is the number of small eddy iterations.

The article is organized as follows. In section 2, we recall the PEs of the ocean and their mathematical setting. The third section of this article presents the small eddy correction method. We prove the existence and uniqueness of solutions to the small eddy correction method when the aspect ratio is small enough. The fourth section is devoted to the convergence of small eddy correction models to the PEs as the aspect ratio goes to zero. We derive an estimate to the rate of convergence as a power of the aspect ratio of the ocean. Although the approach used here is similar to that of [13], there are several differences between the work of [13] and that presented here. First, our model is more complicated. In fact, the PEs of the ocean possess some specific difficulties to circumvent; for instance, the nonlocal constraint (incompressible condition) and the integral expression of the vertical velocity lead to a strong nonlinear term

$$\left(\int_z^0 \operatorname{div} v\, ds\right) \frac{\partial v}{\partial z}. \tag{1.1}$$

More importantly, in [13] the authors used the eigenvalues of the Stokes operator to split the solution between the large and small scales, while in this article the large scale is the depth average (barotropic mode) of the solution and the small scale is the

deviation (baroclinic mode), a decomposition commonly used in ocean modeling. To illustrate the method, some numerical simulations are presented in the last section of this article.

Let us recall that despite advances in the study of the PEs, the mathematical theory of the PEs is far from complete [15, 16, 19, 20, 21, 22, 23, 29, 4]. In [16], the authors prove the existence and uniqueness of global strong solutions of the PEs in thin domains for a broad class of data which include the most physically relevant ones. The most recent result appears in [3], in which the authors prove the existence and uniqueness of global strong solutions without any restriction on the size of the data and the domain occupied by the fluid. It is also worth mentioning that the existence and regularity of solutions to the primitive equations are also studied in [1, 2]. In all these works, the model is a 2D primitive equations, while this article deals with a three-dimensional (3D) problem. Another model of a fluid under the effect of a Coriolis force is studied in [6, 7]. In these articles, the model is a 3D Navier–Stokes equation with a Coriolis force. The authors proved in [6] the existence and uniqueness of strong solutions when the domain is the whole space $\Re^3$ and the Rossby number is small enough. In [7], the authors studied the convergence of the model when the Rossby number goes to zero.

2. The primitive equations and its mathematical setting.

2.1. Governing equations. We first recall the primitive equations. In the nondimensional form, the equations read

$$\begin{cases} \dfrac{\partial v}{\partial t} - \dfrac{1}{R_{e_1}}\Delta v - \dfrac{1}{R_{e_2}}\dfrac{\partial^2 v}{\partial z^2} + f k_0 \times v + (v\cdot\nabla)v + w\dfrac{\partial v}{\partial z} + \operatorname{grad} p = F_1, \\ \dfrac{\partial \rho}{\partial t} - \dfrac{1}{R_{t_1}}\Delta \rho - \dfrac{1}{R_{t_2}}\dfrac{\partial^2 \rho}{\partial z^2} + (v\cdot\nabla)\rho + w\dfrac{\partial \rho}{\partial z} = F_2, \\ \operatorname{div} v + \dfrac{\partial w}{\partial z} = 0, \\ \dfrac{\partial p}{\partial z} = -\rho. \end{cases} \tag{2.1}$$

The boundary conditions are given by

$$\begin{cases} \dfrac{\partial v}{\partial z} = \dfrac{\partial \rho}{\partial z} = 0,\ w = 0 \text{ at } z = -\delta, 0, \\ v, w, \rho \ \text{ periodic in the direction } x, y \text{ with period } \omega. \end{cases} \tag{2.2}$$

In (2.1)–(2.2), the unknown functions are the horizontal velocity v, the vertical velocity w, and the density ρ of the fluid. The constants $R_{e_1} = \frac{\mu}{L_1 U_1} > 0$, $R_{e_2} = \frac{\nu}{L_1 U_1} > 0$, $R_{t_1} = \frac{\mu_T}{L_1 U_1} > 0$, and $R_{t_2} = \frac{\nu_T}{L_1 U_1} > 0$ are the nondimensional Reynolds numbers, $\delta = \frac{H_1}{L_1}$ is the aspect ratio, p is the pressure of the fluid, F_1, F_2 are the volume forces, and k_0 is the unit vector in the vertical direction. Here U_1 is the reference value for the horizontal velocity, L_1 is the reference value for the horizontal length scale, H_1 is the reference value for the vertical length scales, ν and μ are the effective molecular dissipation in the horizontal and vertical directions, and μ_T and ν_T reflect the heat diffusion [19, 20, 21].

The Coriolis parameter f is defined by $f = f_0 + \beta y$, where $\beta > 0$, $f_0 > 0$ are positive constants.

Throughout this article, we use Δ, ∇, div to denote the 2D gradient, Laplacian, and divergence operators on the horizontal plane. The nondimensional domain $\mathcal{M}$ occupied by the fluid is given by

$$\mathcal{M} = \omega \times (-\delta, 0), \tag{2.3}$$

where $\omega \subset \Re^2$ is a smooth convex, bounded open set of $\Re^2$ with boundary $\partial\omega$, and $\delta > 0$ is a constant.

From $(2.1)_{3,4}$ we derive that

$$\begin{cases} W(v) = -\displaystyle\int_{-\delta}^{z} \text{div } v ds, \\ p = p_s + \displaystyle\int_{z}^{0} \rho ds, \\ \text{div} \displaystyle\int_{-\delta}^{0} u \, dz = 0. \end{cases} \tag{2.4}$$

Therefore, (2.1) becomes

$$\begin{cases} \dfrac{\partial v}{\partial t} - \dfrac{1}{R_{e_1}}\Delta v - \dfrac{1}{R_{e_2}}\dfrac{\partial^2 v}{\partial z^2} + f k_0 \times v + (u \cdot \nabla)v \\ \qquad + W(v)\dfrac{\partial v}{\partial z} + \text{grad } p_s + \text{grad} \displaystyle\int_{z}^{0} \rho ds = F_1, \\ \dfrac{\partial \rho}{\partial t} - \dfrac{1}{R_{t_1}}\Delta \rho - \dfrac{1}{R_{t_2}}\dfrac{\partial^2 \rho}{\partial z^2} + (v \cdot \nabla)\rho + W(v)\dfrac{\partial \rho}{\partial z} = F_2, \\ \text{div} \displaystyle\int_{-\delta}^{0} v \, dz = 0. \end{cases} \tag{2.5}$$

Remark 2.1. As in [14], on the surface $z = 0$ of the ocean, we consider a simplified boundary condition. A more physical boundary conditions should be

$$\frac{\partial v}{\partial z} + \alpha_v v = 0, \ \frac{\partial \rho}{\partial z} + \alpha_\rho \rho = 0 \ \text{ at } z = 0. \tag{2.6}$$

In [14], the author pointed out that his existence and uniqueness results cannot be easily extended to the boundary conditions (2.6). Since we extensively used some results given in [14], we will restrict ourselves to the boundary conditions (2.2).

2.2. Mathematical setting. In this section we first recall from [14] the functional spaces suitable for the mathematical setting of (2.5)–(2.2). Let $\mathcal{C}^\infty(\mathcal{M})$ be the usual infinitely differentiable function space in $\mathcal{M}$. Let $H^s(\mathcal{M})$, for $s \in \Re$, be the Sobolev spaces constructed on $L^2(\mathcal{M})$. Motivated by the boundary conditions (2.2), we define

$$\mathcal{C}^\infty_{l,per}(\mathcal{M}) = \{v \in \mathcal{C}^\infty(\mathcal{M}), \ v \text{ periodic in the } x, y \text{ direction with period } \omega\}, \tag{2.7}$$

where the subscript "l" represents the lateral boundary.

Let

$$\mathcal{V}_1 = \left\{ v \in \mathcal{C}^\infty_{l,per}(\mathcal{M}), \ \ \text{div} \int_{-\delta}^{0} v ds = 0 \right\}, \tag{2.8}$$

$$H_1 = \text{ closure of } \mathcal{V}_1 \text{ in } (L^2(\mathcal{M}))^2, \ H_2 = L^2(\mathcal{M}), H = H_1 \times H_2, \tag{2.9}$$

and

$$V_1 = \text{ closure of } \mathcal{V}_1 \text{ in the } (H^1(\mathcal{M}))^2 - \text{norm},\ V_2 = H^1(\mathcal{M}),\ V = V_1 \times V_2. \tag{2.10}$$

The scalar product in H is simply denoted by $(\cdot,\cdot)$, the one in V is denoted by $((\cdot,\cdot))$, and the associated norms are denoted by $|\cdot|_{L^2}$ and $\|\cdot\|$, respectively.

We define the function spaces X_1 and X_2 by

$$X_1 = \left\{ \bar{u} = (\bar{v}, \bar{q}),\ \text{for } u = (v,q) \in L^2(0,T;D(A)),\ \frac{du}{dt} \in L^2(0,T;H) \right\},$$

$$X_2 = \left\{ u^\flat = (v^\flat, q^\flat),\ \text{for } u = (v,q) \in L^2(0,T;D(A)),\ \frac{du}{dt} \in L^2(0,T;H) \right\},$$

where $\bar{u} = Mu$, $u^\flat = N$ are defined by (2.22)–(2.24).

The spaces X_1 and X_2 are endowed with the norms

$$\|\bar{u}\|_{X_1} = \left(\|\bar{u}\|^2_{L^2(0,T;D(A))} + \left| \frac{d\bar{u}}{dt} \right|^2_{L^2(0,T;H)} \right)^{\frac{1}{2}},$$

$$\|u^\flat\|_{X_2} = \left(\|u^\flat\|^2_{L^2(0,T;D(A))} + \left| \frac{dv^\flat}{dt} \right|^2_{L^2(0,T;H)} \right)^{\frac{1}{2}}.$$

We define the bilinear forms $a : V \times V \to \Re$, $a_i : V_i \times V_i \to \Re, i = 1,2$, and the corresponding linear operators $A : V \to V'$, $A_i : V_i \to V_i'$, $i = 1,2$, by

$$\begin{aligned} a_1(v,\tilde{v}) &= \langle A_1 v, \tilde{v} \rangle = \int_{\mathcal{M}} \left[\frac{1}{R_{e_1}} \nabla v \nabla \tilde{v} + \frac{1}{R_{e_2}} \frac{\partial v}{\partial z} \frac{\partial \tilde{v}}{\partial z} \right] dxdydz, \\ a_2(\rho,\tilde{\rho}) &= \langle A_2 \rho, \tilde{\rho} \rangle = \int_{\mathcal{M}} \left[\frac{1}{R_{t_1}} \nabla \rho \nabla \tilde{\rho} + \frac{1}{R_{t_2}} \frac{\partial \rho}{\partial z} \frac{\partial \tilde{\rho}}{\partial z} \right] dxdydz, \\ a(u,\tilde{u}) &= \langle Au, \tilde{u} \rangle = \langle A_1 v, \tilde{v} \rangle + \langle A_2 \rho, \tilde{\rho} \rangle \end{aligned} \tag{2.11}$$

for $u = (v,\rho), \tilde{u} = (\tilde{v},\tilde{\rho}) \in V$.

We have the following characterization of the operators A_i and their domains (see [14] for the details):

$$\begin{aligned} D(A_1) &= \left\{ v \in (H^2(\mathcal{M}))^2 \cap V_1;\ \frac{\partial v}{\partial z} = 0 \ \text{ at } z = -\delta, 0 \right\}, \\ D(A_2) &= \left\{ \rho \in H^2(\mathcal{M}) \cap V_2;\ \frac{\partial \rho}{\partial z} = 0 \ \text{ at } z = -\delta, 0 \right\}, \end{aligned} \tag{2.12}$$

and

$$\begin{aligned} A_1 v &= -\mathcal{P} \left[\frac{1}{R_{e_1}} \Delta v + \frac{1}{R_{e_2}} \frac{\partial^2 v}{\partial z^2} \right] \ \forall v \in D(A_1), \\ A_2 \rho &= -\left[\frac{1}{R_{t_1}} \Delta \rho + \frac{1}{R_{t_2}} \frac{\partial^2 \rho}{\partial z^2} \right] \ \forall \rho \in D(A_2), \end{aligned} \tag{2.13}$$

where $\mathcal{P}$ is the orthogonal projection from $(L^2(\mathcal{M}))^2$ onto H_1.

The linear operators A_i, $i = 1, 2$, which are isomorphisms from V_i onto V_i' are unbounded, self-adjoint linear operators on H_i; they are positive operators and admit compact inverses, so that the fractional power of A_i can also be defined; see [14] and also [28] in which the regularity of A_1 has been studied in a different context.

Note that there exist constants $m_1 > 0$ and $m_2 > 0$ such that

$$m_1\|w\|^2 \leq \langle Aw, w\rangle \leq m_2\|w\|^2 \ \forall w \in V. \tag{2.14}$$

For the nonlinear term, we define the following trilinear functionals and associated operators

$$\begin{aligned} b_1(v,\tilde{v},v') &= \langle B_1(v,\tilde{v}),v'\rangle = \int_{\mathcal{M}} \left[(v\cdot\nabla)\tilde{v} + W(v)\frac{\partial\tilde{v}}{\partial z}\right]\cdot v' dxdydz,\\ b_2(v,\tilde{\rho},\rho') &= \langle B_2(v,\tilde{\rho}),\rho'\rangle = \int_{\mathcal{M}} \left[(v\cdot\nabla)\tilde{\rho} + W(v)\frac{\partial\tilde{\rho}}{\partial z}\right]\cdot \rho' dxdydz,\\ b(u,\tilde{u},u') &= \langle B(u,\tilde{u}),u'\rangle = b_1(v,\tilde{v},v') + b_2(v,\tilde{\rho},\rho') \end{aligned} \tag{2.15}$$

for $u = (v, \rho)$, $\tilde{u} = (\tilde{v}, \tilde{\rho})$, $u' = (v', \rho')$ in V.

For the Coriolis term, we define the bilinear functional $e : H_1 \times H_1 \to \Re$ and the associated operator $E : H_1 \to H_1$ by

$$e(u,\tilde{u}) = \langle E(u),\tilde{u}\rangle = \int_{\mathcal{M}} (f_0 k\times v)\cdot\tilde{v}dxdydz \ \forall u = (v,\rho), \tilde{u} = (\tilde{v},\tilde{\rho}) \in H. \tag{2.16}$$

Finally we define γ and Λ by

$$\Lambda u = \int_z^0 \text{grad}\rho ds, \ \gamma(u,\tilde{u}) = \langle \Lambda u,\tilde{v}\rangle \ \forall u = (v,\rho), \tilde{u} = (\tilde{v},\tilde{\rho}) \in V. \tag{2.17}$$

With these notations, we have the following weak formulation of (2.5)–(2.2) (see page 431 of [14] for the details):

(2.18)

Find $u = (v, \rho) \in L^\infty(0, T; H) \cap L^2(0, T; V)$ $\forall T > 0$, such that

$$\frac{d}{dt}(u,u') + a(u,u') + b(u,u,u') + e(u,u') + \gamma(u,u') = (F,u') \ \forall u' = (v',\rho') \in V,$$

$$u(0) = a \in H.$$

The following result concerning the existence of weak solutions for the PEs is proved in [18, 19].

THEOREM 2.1. *For $T > 0$, there exists at least one solution $u = (v, \rho)$ for (2.18) defined on $(0, T)$ and such that*

$$u \in C([0,T]; H_w), \ \frac{du}{dt} \in L^2(0,T;(V\cap H^3(\mathcal{M}))'),$$

where H_w is the space H equipped with the weak topology and $(V \cap H^3(\mathcal{M}))'$ is the dual space of $V \cap H^3(\mathcal{M})$.

Let us recall that the existence and uniqueness of strong solutions to the PEs was recently proved in [3] without any restriction on the size of the data and the domain

occupied by the fluid. However, in this article, we restrict ourselves to the framework given in [14].

To state the results of [14], we introduce a monotone increasing function $R_0(\delta)$ which satisfies

$$\lim_{\delta\to 0} \delta^{1/2} R_0^2(\delta) = 0. \tag{2.19}$$

THEOREM 2.2. *Assume that the initial condition $a \in H$ and the forcing $F \in H$ satisfy*

$$\|a^\flat\|^2 + |F^\flat|_{L^2}^2 \le \delta^{1/4} R_0^2(\delta), \quad \|\bar a\|^2 + |\bar F|_{L^2}^2 \le \delta R_0^2(\delta). \tag{2.20}$$

Then there exist a constant δ_0 which depends on $R_{e_i}, R_{t_i}, i = 1, 2$, ω and a constant $\sigma > 1$ independent of δ such that, whenever $\delta \in (0, \delta_0)$, the strong solution $u = (v, \rho)$ of (2.18) exists and is unique for all times. More precisely, for all $T > 0$, $u \in L^\infty(0, \infty; V) \cap L^2(0, T; D(A))$ and the following estimates hold:

$$\|v^\flat\|^2 + \|\rho^\flat\|^2 \le \sigma\delta^{1/2} R_0^2(\delta), \quad \|\bar v\|^2 \le \sigma\delta R_0^2(\delta), \quad \|\bar\rho\|^2 \le \sigma R_0^2(\delta), \tag{2.21}$$

where $u^\flat = Nu$, $\bar u = Mu$. The operator M and N which are, respectively, the average operator in the thin direction and its complementary part are defined in (2.22).

Proof. See page 436 of [14]. □

Remark 2.2. The restriction (2.19) on the data is still physically relevant. In fact the baroclinic components $a^\flat$ and $F^\flat$ of the data can be of order $O(\delta^{-1/4})$, while the baroclinic components $\bar a$ and $\bar F$ are of order $O(1)$.

2.2.1. Functional inequalities in thin domains. We recall from [28] (see also [14]) some functional inequalities in thin domains. First we define average operator M in the thin direction and its complementary N as follows:

$$M\psi(x, y) = \frac{1}{\delta}\int_{-\delta}^{0} \psi(x, y, z)dz, \ N\psi(x, y, z) = \psi(x, y, z) - M\psi(x, y). \tag{2.22}$$

Clearly, M and N are orthogonal projections on $L^2(\mathcal{M})$. They commute with the partial derivatives $\frac{\partial}{\partial x}$, $\frac{\partial}{\partial y}$, and $\frac{\partial}{\partial z}$.

It follows that (see [14])

$$MPv = PMv \ \forall v \in (L^2(\mathcal{M}))^2, \qquad MA_1 v = A_1 Mv \ \forall v \in D(A_1). \tag{2.23}$$

Hereafter, we will also use the following notation:

$$\bar v = Mv, \qquad v^\flat = Nv. \tag{2.24}$$

We also have the Poincaré inequalities

$$\begin{aligned} &|v^\flat|_{L^2} \le \delta \left|\frac{\partial v^\flat}{\partial z}\right|_{L^2} \ \forall v \in V_1, \\ &|v^\flat|_{L^2} \le \delta^2 |Av^\flat|_{L^2} \ \forall v \in D(A_1). \end{aligned} \tag{2.25}$$

The following lemma is borrowed from [28] (see also [14]).

LEMMA 2.3. *Let $k \in (0,1/2)$. Then there exists $c_1 = c_1(k)$ independent of δ such that*

(2.26)
$$|b_1(\bar{u}, v^\flat, w)| \le c_1 \delta^k \|\bar{u}\| |A_1 v^\flat|_{L^2} |w|_{L^2} \ \forall u \in V_1, v \in D(A_1), w \in (L^2(\mathcal{M}))^2,$$

$$|b_1(u^\flat, \bar{v}, w)| \le c_1 \delta^{1/2} \|\bar{v}\| |A_1 u^\flat|_{L^2} |w|_{L^2} \ \forall u \in D(A_1), v \in V_1, w \in (L^2(\mathcal{M}))^2,$$

$$|b_1(u^\flat, v^\flat, w)| \le c_1 \delta^{1/2} \|v^\flat\| |A_1 u^\flat|_{L^2} |w|_{L^2} \ \forall u \in D(A_1), v \in V_1, w \in (L^2(\mathcal{M}))^2,$$

$$|b_2(\bar{v}, \rho^\flat, w)| \le c_1 \delta^k \|\bar{v}\| |A_2 \rho^\flat|_{L^2} |w|_{L^2} \ \forall v \in V_1, \rho \in D(A_2), w \in (L^2(\mathcal{M}))^2,$$

$$|b_2(v^\flat, \bar{\rho}, w)| \le c_1 \delta^{1/2} \|\bar{\rho}\| |A_1 v^\flat|_{L^2} |w|_{L^2} \ \forall v \in D(A_1), \rho \in V_2, w \in (L^2(\mathcal{M}))^2,$$

$$|b_2(v^\flat, \rho^\flat, w)| \le c_1 \delta^{1/2} \|\rho^\flat\| |A_1 v^\flat|_{L^2} |w|_{L^2} \ \forall v \in D(A_1), \rho \in V_2, w \in (L^2(\mathcal{M}))^2,$$

$$|\Lambda w|_{L^2}^2 \le c\delta^2 |Aw|_{L^2}^2 \ \forall w \in D(A),$$

$$|Ew|_{L^2}^2 \le c|w|_{L^2}^2 \ \forall w \in H.$$

We also recall the following well-known estimate on the trilinear form of the 2D Navier–Stokes equations (see [26]):

$$\begin{aligned}
&b(\bar{u}, \bar{v}, \bar{v}) = 0 \quad \forall u, v \in V,\\
&|b(\bar{u}, \bar{v}, \bar{w})| \le c_0 |\bar{u}|_{L^2}^{\frac{1}{2}} \|\bar{u}\|^{\frac{1}{2}} \|\bar{v}\|^{\frac{1}{2}} |A\bar{v}|_{L^2}^{\frac{1}{2}} |\bar{w}|_{L^2} \quad \forall u \in V, v \in D(A), w \in H,\\
&|b(\bar{u}, \bar{v}, \bar{w})| \le c_0 |\bar{u}|_{L^2}^{\frac{1}{2}} |A\bar{u}|_{L^2}^{\frac{1}{2}} \|\bar{v}\| |\bar{w}|_{L^2} \quad \forall u \in D(A), v \in V, w \in H,\\
&|b(\bar{u}, \bar{v}, \bar{w})| \le c_0 |\bar{u}|_{L^2}^{\frac{1}{2}} \|\bar{u}\|^{\frac{1}{2}} \|\bar{v}\| |\bar{w}|_{L^2}^{\frac{1}{2}} \|\bar{w}\|^{\frac{1}{2}} \quad \forall u, v, w \in V.
\end{aligned} \tag{2.27}$$

In [14, p. 436], the author derived the following weak formulation for the M and N components $\bar{u}$ and $u^\flat$ of u (note that $\gamma(u, \bar{\theta}) = 0$ for all $\theta \in V$ since for $\theta = (w, \phi) \in V$, div $\bar{w} = 0$):

$$\begin{aligned}
&\frac{d}{dt}(\bar{u}, \bar{\theta}) + a(\bar{u}, \bar{\theta}) + e(\bar{u}, \bar{\theta}) + b(\bar{u}, \bar{u}, \bar{\theta}) + b(u^\flat, u^\flat, \bar{\theta}) = (\bar{F}, \bar{\theta}) \ \forall \theta \in V,\\
&\bar{u}(0) = \bar{a} \in H,
\end{aligned} \tag{2.28}$$

$$\begin{aligned}
&\frac{d}{dt}(u^\flat, \theta^\flat) + a(u^\flat, \theta^\flat) + e(u^\flat, \theta^\flat) + b(\bar{u}, u^\flat, \theta^\flat) + b(u^\flat, \bar{u}, \theta^\flat)\\
&\qquad + b(u^\flat, u^\flat, \theta^\flat) + \gamma(\bar{u} + u^\flat, \theta^\flat) = (F^\flat, \theta^\flat) \ \forall \theta \in V,\\
&u^\flat(0) = a^\flat \in H.
\end{aligned} \tag{2.29}$$

Hereafter, $c_0 = c_0(\omega, T)$ will denote constants that depend only on the horizontal domain ω and T, $c_2 = c_2(\omega, k, T)$ will denote a constant that depends on ω, T, and k, and $c_1 = c_1(k, T)$ will denote a constant that depend on k and T. Finally c will denote a generic constant.

2.2.2. The 2D Navier–Stokes. We first recall the following 2D Navier–Stokes equations (with a Coriolis force) and an associated transport equation:

$$\begin{cases} \dfrac{\partial \bar{w}_1}{\partial t} - \dfrac{1}{R_{e_1}}\Delta \bar{w}_1 + fk_0 \times \bar{w}_1 + (\bar{U}_1 \cdot \nabla)\bar{w}_1 + (\bar{w}_1 \cdot \nabla)\bar{U}_1 \\ \qquad + (\bar{w}_1 \cdot \nabla)\bar{w}_1 + \text{grad } p_s = \bar{F}_1, \\ \dfrac{\partial \bar{q}_1}{\partial t} - \dfrac{1}{R_{t_1}}\Delta \bar{q}_1 + (\bar{U}_1 \cdot \nabla)\bar{q}_1 + (\bar{w}_1 \cdot \nabla)\bar{\psi} + (\bar{v}_1 \cdot \nabla)\bar{q} = \bar{F}_2, \\ \text{div } \bar{w}_1 = 0, \end{cases} \tag{2.30}$$

with the initial and boundary conditions

$$\begin{aligned} &\bar{v}_1 = (\bar{w}_1, \bar{q}) \text{ is periodic in the } x \text{ and } y \text{ direction with period } \omega, \\ &\bar{v}_1(0) = \bar{a}, \text{ at } t = 0. \end{aligned} \tag{2.31}$$

It is clear that $\bar{v}_1 = (\bar{w}_1, \bar{q}_1)$ satisfies

$$\begin{aligned} &\frac{d}{dt}(\bar{v}_1, \bar{\theta}) + a(\bar{v}_1, \bar{\theta}) + e(\bar{v}_1, \bar{\theta}) + b(\bar{U}, \bar{v}_1, \bar{\theta}) + b(\bar{v}_1, \bar{U}, \bar{\theta}) + b(\bar{v}_1, \bar{v}_1, \bar{\theta}) \\ &\qquad = (\bar{F}, \bar{\theta}) \ \forall \theta \in V, \ \bar{v}_1(0) = \bar{a}, \end{aligned} \tag{2.32}$$

where $\bar{U} = (\bar{U}_1, \bar{\psi}_1)$, $\bar{F} = (\bar{F}_1, \bar{F}_2)$, $\bar{a} = (\bar{a}_1, \bar{a}_2)$.

In (2.30), the unknown functions are the velocity $\bar{w}$, the temperature $\bar{q}_1$, and the surface pressure p_s. The volume force $\bar{F}$ and the initial condition $\bar{a}$ are given.

We assume the following regularity conditions:

$$\bar{F} \in L^2(0,T;H) \ \forall T > 0, \ \bar{a} \in V; \qquad \bar{U} \in L^\infty(0,T;V) \cap L^2(0,T;D(A)). \tag{2.33}$$

The domain ω occupied by the fluid is a smooth convex, bounded open set of $\Re^2$ with boundary $\partial\omega$. Finally the constant $R_{e_1} > 0$ is the nondimensional Reynolds number.

PROPOSITION 2.4. *The system* (2.30) *has a unique solution* $\bar{v}_1 \in L^2(0,T;D(A)) \cap L^\infty(0,T;V)$. *Moreover, we have the estimates*

$$\begin{aligned} &|\bar{v}_1(t)|_{L^2}^2 + \int_0^T \|\bar{v}_1\|^2 dt \le e^{M_0(T)} \left(|\bar{a}|_{L^2}^2 + \int_0^T |\bar{F}|_{L^2}^2 ds \right), \\ &\|\bar{v}_1(t)\|^2 + \int_0^T |A\bar{v}_1|_{L^2}^2 dt \le e^{M_1(T)} \left(\|\bar{a}\|^2 + \int_0^T |\bar{F}|_{L^2}^2 ds \right), \\ &\int_0^T \left| \frac{d\bar{v}_1}{dt} \right|_{L^2}^2 dt \le e^{M_1(T)} \left(\|\bar{a}\|^2 + \int_0^T |\bar{F}|_{L^2}^2 ds \right), \end{aligned} \tag{2.34}$$

where

$$\begin{aligned} M_0(t) = c_0 \int^t \|\bar{U}\|^2 ds, \ M_1(t) = c_0 \int_0^t |\bar{U}|_{L^2} |A_1 \bar{U}|_{L^2} ds \\ + e^{2M_0(t)} \left(|\bar{a}|_{L^2}^2 + \int_0^T |\bar{F}|_{L^2}^2 ds \right)^2. \end{aligned} \tag{2.35}$$

Proof. For the existence and uniqueness of solutions to (2.32), (2.30), see, for instance, [26]. The estimates (2.34) are also standard, but for the sake of clarity, we give a sketch of the proof.

For $(2.34)_1$, multiplying (2.30) by $\bar{v}_1$ and using (2.27) yield

$$\frac{d}{dt}|\bar{v}_1|_{L^2}^2 + c_0\|\bar{v}_1\|^2 \leq c_0|\bar{v}_1|_{L^2}^2\|\bar{U}\|^2 + c_0|\bar{F}|_{L^2}^2, \tag{2.36}$$

and $(2.34)_1$ follows from the Gronwall lemma [26].

For $(2.34)_2$, we first note that (see (2.27))

$$\begin{aligned}
|b(\bar{v}_1, \bar{U}, A\bar{v}_1)| &\leq c_0|\bar{v}_1|_{L^2}^{1/2}|A\bar{v}_1|_{L^2}^{3/2}\|\bar{U}\| \\
&\leq \frac{1}{8}|A\bar{v}_1|_{L^2}^2 + c_0|\bar{v}_1|_{L^2}^2\|\bar{U}\|^4, \\
|b(\bar{U}, \bar{v}_1, A\bar{v}_1)| &\leq \frac{1}{8}|A\bar{v}_1|_{L^2}^2 + c_0|\bar{U}|_{L^2}|A\bar{U}|_{L^2}, \\
|b(\bar{v}_1, \bar{v}_1, A\bar{v})| &\leq \frac{1}{8}|A\bar{v}_1|_{L^2}^2 + c_0|\bar{v}_1|_{L^2}^2\|\bar{v}_1\|^4, \\
|e(\bar{v}_1, A\bar{v}_1)| &\leq \frac{1}{8}|A\bar{v}_1|_{L^2}^2 + c_0|\bar{v}_1|_{L^2}^2.
\end{aligned} \tag{2.37}$$

Now multiplying (2.30) by $A\bar{v}$ and using (2.37) yield

$$\frac{d}{dt}\|\bar{v}_1\|^2 + |A\bar{v}_1|_{L^2}^2 \leq c_0|\bar{F}|_{L^2}^2 + c_0|\bar{v}_1|_{L^2}^2 + h(t)\|\bar{v}_1\|^2, \tag{2.38}$$

where

$$h(t) = c_0|\bar{U}|_{L^2}|A\bar{U}|_{L^2} + c_0|\bar{v}_1|_{L^2}^2\|\bar{v}_1\|^2 + c_0\|\bar{U}\|^4$$

and

$$\begin{aligned}
\int_0^t h(s)ds &\leq c_0\int_0^t |\bar{U}|_{L^2}|A\bar{U}|_{L^2}ds + \sup_s |\bar{v}_1(s)|_{L^2}^2\int_0^t \|\bar{v}_1\|^2 ds \\
&\quad + c_0\int_0^t \|\bar{U}\|^4 ds \equiv M_1(t),
\end{aligned} \tag{2.39}$$

and $(2.34)_2$ follows from the standard Gronwall lemma.

For $(2.34)_3$, we note that

$$\begin{aligned}
\left|\frac{d\bar{v}_1}{dt}\right|_{L^2}^2 &\leq c|A\bar{v}_1|_{L^2}^2 + c|E\bar{v}_1|_{L^2}^2 + c|B(\bar{U}, \bar{v}_1)|_{L^2}^2 \\
&\quad + c|B(\bar{v}_1, \bar{U})|_{L^2}^2 + c|B(\bar{v}_1, \bar{v}_1)|_{L^2}^2 \\
&\leq c|A\bar{v}_1|_{L^2}^2 + c_0|\bar{v}_1|_{L^2}^2 + c_0|\bar{U}|_{L^2}|A\bar{U}|_{L^2}\|\bar{v}_1\|^2 + c_0|\bar{v}_1|_{L^2}|A\bar{v}_1|_{L^2}\|\bar{U}\|^2 \\
&\quad + c_0|\bar{v}_1|_{L^2}|A\bar{v}_1|_{L^2}\|\bar{v}_1\|^2,
\end{aligned} \tag{2.40}$$

and $(2.34)_3$ follows from $(2.34)_2$. ∎

2.2.3. The 3D linear system. We also consider the following 3D heat-type equations:

$$\frac{d}{dt}(v^\flat,\theta^\flat)+a(v^\flat,\theta^\flat)+e(v^\flat,\theta^\flat)+\gamma(v^\flat,\theta^\flat)=(F^\flat,\theta^\flat)\ \forall\theta\in V,\ v^\flat(0)=a^\flat. \tag{2.41}$$

In (2.41), the unknown function $v^\flat=(v_1^\flat,q_1^\flat)$, the volume force $F^\flat=(F^\flat,F_2^\flat)$, and the initial condition $a^\flat=(a_1^\flat,a_2^\flat)$ are given. We assume the following regularity condition:

$$a^\flat\in V,\qquad F^\flat\in L^2(0,T;H). \tag{2.42}$$

PROPOSITION 2.5. *The heat-type equation* (2.41) *has a unique solution* $v^\flat\in L^2(0,T;D(A))\cap L^\infty(0,T;V)$. *Moreover, we have the estimates*

$$\begin{aligned}&\|v^\flat(t)\|^2\le c\left(\|a^\flat\|^2+\int_0^T|F^\flat|_{L^2}^2ds\right)\equiv c_5,\ \int_0^T|Av^\flat|_{L^2}^2dt\le c_5,\\&\int_0^T\left|\frac{dv^\flat}{dt}\right|_{L^2}^2dt\le c_5,\ |v^\flat(t)|_{L^2}^2\le\delta^2c_5.\end{aligned} \tag{2.43}$$

Proof. Multiplying (2.41) by $v^\flat$ gives

$$|v^\flat(t)|_{L^2}^2+\int_0^t\|v^\flat\|^2ds+\le c\left(|a^\flat|_{L^2}^2+\int_0^t|F^\flat|_{L^2}^2ds\right). \tag{2.44}$$

Now, multiplying (2.41) by $Av^\flat$ yields

$$\|v^\flat(t)\|^2+\int_0^t|Av^\flat|_{L^2}^2ds\le c\left(\|a^\flat\|^2+\int_0^t|F^\flat|_{L^2}^2ds\right) \tag{2.45}$$

and $(2.43)_{2,3}$ follow. Note that $(2.43)_4$ follows from (2.25).

3. A small eddy correction method. Hereafter we set

$$\begin{aligned}(\mathcal{F}(\bar u,u^\flat),\theta^\flat)=&\frac{d}{dt}(u^\flat,\theta^\flat)+a(u^\flat,\theta^\flat)+e(u^\flat,\theta^\flat)+\gamma(\bar u+u^\flat,\theta^\flat)+b(\bar u,u^\flat,\theta^\flat)\\&+b(u^\flat,\bar u,\theta^\flat)+b(u^\flat,u^\flat,\theta^\flat)-(F^\flat,\theta^\flat).\end{aligned} \tag{3.1}$$

Then $(2.29)_1$ is equivalent to

$$(\mathcal{F}(\bar u,u^\flat),\theta^\flat)=0\ \forall\theta\in V,\ u^\flat(0)=a^\flat. \tag{3.2}$$

Supposing that the barotropic flow $\bar u$ is known, and formally applying the Newton iteration to (3.2), we get the following iterative procedure: assuming that the initial guess for the baroclinic (or the small eddy) component $u_0^\flat=0$ and the $(j-1)$th approximation $u_{j-1}^\flat$ is known for some integer j, find the jth approximation $u_j^\flat$ such that

$$D_{u^\flat}\mathcal{F}(\bar u,u^\flat)(u_j^\flat-u_{j-1}^\flat)=-\mathcal{F}(\bar u,u_{j-1}^\flat). \tag{3.3}$$

Simple calculation shows that (3.3) reduces to

$$\begin{aligned}&\frac{d}{dt}(u_j^\flat,\theta^\flat)+a(u_j^\flat,\theta^\flat)+e(u_j^\flat,\theta^\flat)+\gamma(v+u_j^\flat,\theta^\flat)+b(\bar u,u_j^\flat,\theta^\flat)+b(u_j^\flat,\bar u,\theta^\flat)\\&\quad+b(u_{j-1}^\flat,u_j^\flat,\theta^\flat)+b(u_j^\flat,u_{j-1}^\flat,\theta^\flat)+b(u_{j-1}^\flat,u_{j-1}^\flat,\theta^\flat)\\&=(F^\flat,\theta^\flat)\ \forall\theta\in V,\ u_j^\flat(0)=a^\flat.\end{aligned} \tag{3.4}$$

Combining (3.4) with the barotropic equation (2.29) (with $\bar{u}$ replaced by v and $u_j^\flat$ replaced by w_j), we obtain the following small eddy correction method: let $w_0 = 0$ and let l be a fixed positive integer:

$$\frac{d}{dt}(v,\bar{\theta}) + a(v,\bar{\theta}) + e(v,\bar{\theta}) + b(v,v,\bar{\theta}) + b(w_l,w_l,\bar{\theta}) = (\bar{F},\bar{\theta})\ \forall \theta \in V,\ v(0) = \bar{a}, \tag{3.5}$$

$$\begin{aligned}&\frac{d}{dt}(w_j,\theta^\flat) + a(w_j,\theta^\flat) + e(w_j,\theta^\flat) + \gamma(v+w_j,\theta^\flat) + b(v,w_j,\theta^\flat) + b(w_j,v,\theta^\flat)\\ &\qquad + b(w_{j-1},u_j^\flat,\theta^\flat) + b(w_j,w_{j-1},\theta^\flat) + b(w_{j-1},w_{j-1},\theta^\flat)\\ &\quad = (F^\flat,\theta^\flat)\ \forall \theta \in V,\ w_j(0) = a^\flat\end{aligned} \tag{3.6}$$

for $j = 1, 2, \ldots, l$.

Remark 3.1. For $l = 0$, (3.5)–(3.6) reduce (since $w_0 = 0$) to

$$\frac{d}{dt}(v,\bar{\theta}) + a(v,\bar{\theta}) + e(v,\bar{\theta}) + b(v,v,\bar{\theta}) = (\bar{F},\bar{\theta})\ \forall \theta \in V,\ v(0) = \bar{a}, \tag{3.7}$$

which is the well-known quasi-geostrophic model.

For $l = 1$, (3.5)–(3.6) become

$$\frac{d}{dt}(v,\bar{\theta}) + a(v,\bar{\theta}) + e(v,\bar{\theta}) + b(v,v,\bar{\theta}) + b(w_1,w_1,\bar{\theta}) = (\bar{F},\bar{\theta})\ \forall \theta \in V,\ v(0) = \bar{a}, \tag{3.8}$$

$$\begin{aligned}&\frac{d}{dt}(w_1,\theta^\flat) + a(w_1,\theta^\flat) + e(w_1,\theta^\flat) + \gamma(v+w_1,\theta^\flat) + b(v,w_1,\theta^\flat) + b(w_1,v,\theta^\flat)\\ &\quad = (F^\flat,\theta^\flat)\ \forall \theta \in V,\ w_1(0) = a^\flat,\end{aligned} \tag{3.9}$$

which is similar to the NLG method studied in [27, 17, 24, 25] for the 2D Navier–Stokes equations.

3.1. Some a priori estimates. In this part, we prove the existence and uniqueness of a strong solution to (3.5)–(3.6) when δ is small enough.

Hereafter we set $\mathbf{X} = X_1 \times (X_2)^l$. For $v = (\bar{v}, w_1, w_2, \ldots, w_l) \in \mathbf{X}$, we set

$$\|v\|_{\mathbf{X}}^2 = \|\bar{v}\|_{X_1}^2 + \sup_i \|w_i\|_{X_2}^2. \tag{3.10}$$

3.1.1. Linear problems. To (3.5)–(3.6) we associate the following system: Find $(\bar{v}^0, w_1^0, w_2^0, \ldots, w_l^0) \in \mathbf{X}$ such that

$$\frac{d}{dt}(v^0,\bar{\theta}) + a(v^0,\bar{\theta}) + e(v^0,\bar{\theta}) = (\bar{F},\bar{\theta})\ \forall \theta \in V,\ v^0(0) = \bar{a}, \tag{3.11}$$

and for $j = 1, 2, \ldots, l$

$$\frac{d}{dt}(w_j^0,\theta^\flat) + a(w_j^0,\theta^\flat) + e(w_j^0,\theta^\flat) = (F^\flat,\theta^\flat)\ \forall \theta \in V,\ w_j^0(0) = a^\flat. \tag{3.12}$$

Following Propositions 2.4 and 2.5, the unique strong solution $(v^0, w_1^0, w_2^0, \ldots, w_l^0) \in \mathbf{X}$ to (3.11)–(3.12) satisfies

$$\|v^0\|_{X_1}^2 \le \alpha_1^2, \quad \sup_j \|w_j^0\|_{X_2}^2 \le \alpha_2^2, \quad \sup_j |w_j^0(t)|_{L^2}^2 \le \delta^2\alpha_2^2, \tag{3.13}$$

where

$$
\begin{aligned}
&\alpha_1^2 \equiv c\left(\|\bar{a}\|^2 + \int_0^T |\bar{F}|_{L^2}^2 ds\right) \le c\delta^{1/2} R_0^2(\delta),\\
&\alpha_2^2 \equiv c\left(\|a^\flat\|^2 + \int_0^T |F^\flat|_{L^2}^2 ds\right) \le c\delta^{1/4} R_0^2(\delta),\ \alpha_0^2 = \alpha_1^2 + \alpha_2^2.
\end{aligned} \tag{3.14}
$$

Note that from (2.19)–(2.20), α_1^2 goes to zero as δ goes to zero.

3.1.2. Nonlinear problems. Now let us set $\vartheta = v - v^0$, $\eta_j = w_j - w_j^0$. Then ϑ and η_j satisfy

$$
\begin{aligned}
&\frac{d}{dt}(\vartheta, \bar{\theta}) + a(\vartheta, \bar{\theta}) + e(\vartheta, \bar{\theta}) + b(\vartheta, v^0, \bar{\theta}) + b(v^0, \vartheta, \bar{\theta}) + b(\vartheta, \vartheta, \bar{\theta}) + (S_1, \bar{\theta})\\
&\qquad = 0\ \forall \theta \in V,\ \vartheta(0) = 0,
\end{aligned} \tag{3.15}
$$

$$
\frac{d}{dt}(\eta_j, \theta^\flat) + a(\eta_j, \theta^\flat) + e(\eta_j, \theta^\flat) + (S_2, \theta^\flat) = 0\ \forall \theta \in V,\ \eta_j(0) = 0, \tag{3.16}
$$

where

$$
\begin{aligned}
&S_1 = B(v^0, v^0) + B(\eta_l + w_l^0, \eta_l + w_l^0),\\
&S_2 = B(\vartheta + v^0, \eta_j + w_j^0) + B(\eta_j + w_j^0, \vartheta + v^0) + B(\eta_{j-1} + w_{j-1}^0, \eta_j + w_j^0)\\
&\qquad + B(\eta_j + w_j^0, \eta_{j-1} + w_{j-1}^0) - B(\eta_{j-1} + w_{j-1}^0, \eta_{j-1} + w_{j-1}^0) + \Lambda(\vartheta + \eta_j).
\end{aligned} \tag{3.17}
$$

To solve (3.15)–(3.16), we consider the following iterative process:

$$
\begin{aligned}
&\frac{d}{dt}(\vartheta^{n+1}, \bar{\theta}) + a(\vartheta^{n+1}, \bar{\theta}) + e(\vartheta^{n+1}, \bar{\theta}) + b(\vartheta^{n+1}, v^0, \bar{\theta}) + b(v^0, \vartheta^{n+1}, \bar{\theta})\\
&\qquad + b(\vartheta^{n+1}, \vartheta^{n+1}, \bar{\theta}) + (S_1^n, \bar{\theta}) = 0\ \forall \theta \in V,\ \vartheta^{n+1}(0) = 0,
\end{aligned} \tag{3.18}
$$

$$
\frac{d}{dt}(\eta_j^{n+1}, \theta^\flat) + a(\eta_j^{n+1}, \theta^\flat) + e(\eta_j^{n+1}, \theta^\flat) + (S_2^n, \theta^\flat) = 0\ \forall \theta \in V,\ \eta_j^{n+1}(0) = 0, \tag{3.19}
$$

where

$$
\begin{aligned}
&S_1^n = B(v^0, v^0) + B(\eta_l^n + w_l^0, \eta_l^n + w_l^0),\\
&S_2^n = B(\vartheta^n + v^0, \eta_j^n + w_j^0) + B(\eta_j^n + w_j^0, \vartheta^n + v^0) + B(\eta_{j-1}^n + w_{j-1}^0, \eta_j^n + w_j^0)\\
&\qquad + B(\eta_j^n + w_j^0, \eta_{j-1}^n + w_{j-1}^0) - B(\eta_{j-1}^n + w_{j-1}^0, \eta_{j-1}^n + w_{j-1}^0) + \Lambda(\vartheta^n + \eta_j^n)
\end{aligned} \tag{3.20}
$$

for $(v^0, \eta_1^0, \eta_2^0, \dots, \eta_l^0) \in \mathbf{X}$ such that

$$
\|\vartheta^0\|_{X_1}^2 + \sup_j \|\eta_j^0\|_{X_2}^2 \le R^2, \qquad \sup_j |\eta_j^0(t)|_{L^2}^2 \le \delta^2 R^2, \tag{3.21}
$$

where $R = R(a, g, T, R_{e_1}, R_{e_2}) > 0$ is given by (3.34).

The goal is to prove (using a fixed-point argument) that the sequence (ϑ^n, η_j^n) is convergent for δ small enough.

PROPOSITION 3.1. *Let $k \in (0, 1/2)$. We assume that*

$$\|\vartheta^n\|_{X_1}^2 + \sup_j \|\eta_j^n\|_{X_2}^2 \leq R^2, \tag{3.22}$$

where R will be made precise later. Then the following estimates hold true for S_1^n and S_2^n:

(3.23)
$$\int_0^T (|S_1^n|_{L^2}^2 + |S_2^n|_{L^2}^2)dt \leq c_2\delta^{2k}(R^2(\|\vartheta^n\|_{X_1}^2 + \sup_j \|\eta_j^n\|_{X_2}^2) + R^4 + \alpha_0^4 + R^2) + c_0\alpha_1^4.$$

Proof. The proof follows from the inequalities (see (2.26)–(2.27))

$$|S_1^n|_{L^2}^2 \leq c_0|v^0|_{L^2}|Av^0|_{L^2}\|v^0\|^2 + c_1\delta\|\eta_l^n + w_l^0\|^2|A(\eta_l^n + w_l^0)|_{L^2}^2, \tag{3.24}$$

which gives

$$\begin{aligned}\int_0^T |S_1^n|_{L^2}^2 dt &\leq c_0\alpha_1^4 + c_1\delta \sup_s \|\eta_l^n(s) + w_l^0(s)\|^2 \int_0^T |A(\eta_l^n + w_l^0)|_{L^2}^2 ds \\ &\leq c_0\alpha_1^4 + c_1\delta(R^2 \sup_j \|\eta_j^n\|_{X_2}^2 + \alpha_0^4 + R^4).\end{aligned} \tag{3.25}$$

We also have

(3.26)
$$\begin{aligned}|S_2^n|_{L^2}^2 \leq{}& c_1\delta^{2k}\|\vartheta^n + v^0\|^2|A(\eta_j^n + w_j^0)|_{L^2}^2 + c_1\delta\|\vartheta^n + v^0\|^2|A(\eta_j^n + w_j^0)|_{L^2}^2 \\ &+ c_1\delta\|\eta_j^n + w_j^0\|^2|A(\eta_{j-1}^n + w_{j-1}^0)|_{L^2}^2 + c_1\delta\|\eta_{j-1}^n + w_{j-1}^0\|^2|A(\eta_j^n + w_j^0)|_{L^2}^2 \\ &+ c_1\delta\|\eta_{j-1}^n + w_{j-1}^0\|^2|A(\eta_{j-1}^n + w_{j-1}^0)|_{L^2}^2 + c_1\delta^2\|\eta_j^n + \vartheta^n\|^2,\end{aligned}$$

which gives

$$\begin{aligned}\int_0^T |S_2^n|_{L^2}^2 dt \leq{}& c_1\delta^{2k} \sup_s \|\vartheta^n(s) + v^0(s)\|^2 \int_0^T |A(\eta_j^n + w_j^0)|_{L^2}^2 ds \\ &+ c_1\delta \sup_s \|\vartheta^n + v^0\|^2 + c_1\delta \sup_s \|\eta_j^n + w_j^0\|^2 \int_0^T |A(\eta_{j-1}^n + w_{j-1}^0)|_{L^2}^2 ds \\ &+ c_1\delta \sup_s \|\eta_{j-1}^n + w_{j-1}^0\|^2 \int_0^T |A(\eta_j^n + w_j^0)|_{L^2}^2 ds \\ &+ c_1\delta \sup_s \|\eta_{j-1}^n + w_{j-1}^0\|^2 \int_0^T |A(\eta_{j-1}^n + w_{j-1}^0)|_{L^2}^2 ds + c_1\delta^2 R^2\end{aligned} \tag{3.27}$$

and

$$\int_0^T |S_2^n|_{L^2}^2 dt \leq c_1\delta^{2k}(R^2(\|\vartheta^n\|^2 + \sup_j \|\eta_j^n\|^2) + \alpha_0^4 + R^4) + c_1\delta^2 R^2. \tag{3.28}$$

Therefore (3.23) follows from (3.25) and (3.28). ∎

PROPOSITION 3.2. *Let $k \in (1/4, 1/2)$. We assume that* (3.22) *holds true.* Then for δ small enough, we have

$$\|\vartheta^{n+1}\|^2_{X_1} + \sup_j \|\eta_j^{n+1}\|^2_{X_2} \le R^2, \qquad \sup_j |\eta_j^{n+1}(t)|^2_{L^2} \le \delta^2 R^2. \tag{3.29}$$

Proof. It clearly follows from Proposition 2.4 and the estimate (3.25) that

$$\begin{aligned}
&|\vartheta^{n+1}(t)|^2_{L^2} \le c_2 e^{N_0(T)}(\delta R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \delta\alpha_0^4 + \delta R^4 + \alpha_1^4),\\
&\|\vartheta^{n+1}(t)\|^2_{X_1} \le c_2 e^{N_1(T)}(\delta R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \delta\alpha_0^4 + \delta R^4 + \alpha_1^4),\\
&\int_0^T |A\vartheta^{n+1}|^2_{L^2} dt \le c_2 e^{N_1(T)}(\delta R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \delta\alpha_0^4 + \delta R^4 + \alpha_1^4),\\
&\int_0^T \left|\frac{d\vartheta^{n+1}}{dt}\right|^2_{L^2} dt \le c_2 e^{N_1(T)}(\delta R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \delta\alpha_0^4 + \delta R^4 + \alpha_1^4).
\end{aligned} \tag{3.30}$$

The exact form of $N_0(T)$ and $N_1(T)$ can be obtained, respectively, from (2.35) by replacing $\bar{U}$ by v^0; i.e., $N_0(T) = N_0(T, \alpha_1)$ and $N_1(T) = N_1(T, \alpha_1)$.

It also follows from Proposition 2.5 and the estimate (3.28) that

$$\begin{aligned}
&\sup_j |\eta_j^{n+1}(t)|^2_{L^2} \le c_1\delta^{2k}(R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \alpha_0^4 + R^4) + c_1\delta^2 R^2,\\
&\sup_j \|\eta_j^{n+1}(t)\|^2 \le c_1\delta^{2k}(R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \alpha_0^4 + R^4) + c_1\delta^2 R^2,\\
&\sup_j \int_0^T |A\eta_j^{n+1}|^2_{L^2} dt \le c_1\delta^{2k}(R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \alpha_0^4 + R^4) + c_1\delta^2 R^2,\\
&\sup_j \int_0^T \left|\frac{d\eta_j^{n+1}}{dt}\right|^2_{L^2} dt \le c_1\delta^{2k}(R^2(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + \alpha_0^4 + R^4) + c_1\delta^2 R^2.
\end{aligned} \tag{3.31}$$

From (3.30) and (3.31) we can write

$$\|\vartheta^{n+1}\|^2_{X_1} + \sup_j \|\eta_j^{n+1}\|^2_{X_2} \le \epsilon(\|\vartheta^n\|^2_{X_1} + \sup_j \|\eta_j^n\|^2_{X_2}) + L_0, \tag{3.32}$$

where

$$\epsilon = c_2\delta^{2k}R^2, \qquad L_0 = c_2\delta^{2k}(\alpha_0^4 + R^4) + c_2\delta^2 R^2 + c_0\alpha_1^4. \tag{3.33}$$

Let

$$R^2 = 8c_2\alpha_0^4, \tag{3.34}$$

where c_2 is the constant that appears in (3.33). Note that from (2.19) and (3.14), R does not go to zero as δ goes to zero.

Now we choose δ small enough such that

$$L_0 \le R^2/4, \qquad \epsilon \le \frac{1}{2}. \tag{3.35}$$

This is possible since from (2.19)–(2.20) and (3.14), all the quantities $\delta^2 R^2$, α_1^4, $\delta^{2k} R^2$, and

$$\frac{\delta^{2k}(\alpha_0^4 + R^4)}{R^2} = c_2 \delta^{2k}(1 + \alpha_0^4) \tag{3.36}$$

go to zero as δ goes to zero for $k \in (1/4, 1/2)$.

Therefore, using inequality (3.32) successively, we get

$$\begin{aligned} \|\vartheta^{n+1}\|_{X_1}^2 + \sup_j \|\eta_j^{n+1}\|_{X_2}^2 &\le \epsilon^n (\|\vartheta^0\|_{X_1}^2 + \sup_j \|\eta_j^0\|_{X_2}^2) + \frac{1-\epsilon^n}{1-\epsilon} L_0, \\ &\le R^2/2 + 2L_0 \le R^2, \end{aligned} \tag{3.37}$$

and $(3.29)_1$ follows. Note that $(3.29)_2$ follows from $(3.29)_1$ and (2.25). □

PROPOSITION 3.3. *Let $k \in (1/4, 1/2)$. Let R be given by* (3.34). *We assume that δ is small enough so that* (3.35) *is satisfied. Let $(\vartheta^0, \eta_1^0, \dots, \eta_l^0) \in \mathbf{X}$ such that* (3.21) *is satisfied. Then the sequence $(\vartheta^n, \eta_1^n, \eta_2^n, \dots, \eta_l^n) \in \mathbf{X}$ given by* (3.18)–(3.19) *satisfies the estimates*

$$\|\vartheta^n\|_{X_1}^2 + \sup_j \|\eta_j^n\|_{X_2}^2 \le R^2, \qquad \sup_j |\eta_j^n(t)|_{L^2}^2 \le \delta^2 R^2. \tag{3.38}$$

Proof. The proof follows by induction from (3.21) and Proposition 3.2. □

Now, let us set $\theta^{n+1} = \vartheta^{n+1} - \vartheta^n$, $q_j^{n+1} = \eta_j^{n+1} - \eta_j^n$. Then $(\theta^{n+1}, q_j^{n+1})$ satisfy

$$\begin{aligned} &\frac{d}{dt}(\theta^{n+1}, \bar\zeta) + a(\theta^{n+1}, \bar\zeta) + e(\theta^{n+1}, \bar\zeta) + b(\theta^{n+1}, \vartheta^{n+1} + v^0, \bar\zeta) \\ &+ b(\vartheta^n + v^0, \theta^{n+1}, \bar\zeta) + (K_1^n, \bar\zeta) = 0 \ \forall \zeta \in V, \ \theta^{n+1}(0) = 0, \end{aligned} \tag{3.39}$$

$$\frac{d}{dt}(q_j^{n+1}, \zeta^\flat) + a(q_j^{n+1}, \zeta^\flat) + e(q_j^{n+1}, \zeta^\flat) + (K_2^n, \zeta^\flat) = 0 \ \forall \zeta \in V, \quad q_j^{n+1}(0) = 0, \tag{3.40}$$

where

$$K_n^1 = B(q_l^n, \eta_l^{n-1} + w_l^0) + B(\eta_l^n + w_l^0, q_l^n) \tag{3.41}$$

and

$$\begin{aligned} K_n^2 &= B(\theta^n, \eta_j^n + w_j^0) + B(q_j^n, \vartheta^n + v^0) + B(q_j^n, \eta_j^n + w_j^0) \\ &\quad + B(\vartheta^{n-1} + v^0, q_j^n) + B(\eta_j^{n-1} + w_j^0, \theta^n) + B(\eta_j^{n-1} + w_j^0, q_j^n) + \Lambda(\theta^n + q_j^n). \end{aligned} \tag{3.42}$$

PROPOSITION 3.4. *Let $k \in (1/4, 1/2)$. We assume that δ is small enough so that* (3.35) *holds. Then the following estimate holds:*

$$\int_0^T (|K_n^1|_{L^2}^2 + |K_n^2|_{L^2}^2) dt \le c_3 \delta^{2k} (\|\theta^n\|_{X_1}^2 + \sup_j \|q_j^n\|_{X_2}^2). \tag{3.43}$$

Proof. The proof follows from the following estimates:

$$|K_n^1|_{L^2}^2 \le c_1 \delta \|\eta_l^n + w_l^0\|^2 |A q_l^n|_{L^2}^2 + c_1 \delta \|q_l^n\|^2 |A(\eta_l^n + w_l^0)|_{L^2}^2 \tag{3.44}$$

and

$$\begin{aligned}\int_0^T |K_n^1|_{L^2}^2 dt &\le c_1\delta \sup_s \|\eta_l^n(s)+w_l^0(s)\|^2 \int_0^T |Aq_l^n|_{L^2}^2 ds \\ &\quad + c_1\delta \sup_s \|q_l^n(s)\|^2 \int_0^T |A(\eta_l^n+w_l^0)|_{L^2}^2 ds \\ &\le c_3\delta \sup_j \|q_j^n\|_{X_2}^2.\end{aligned} \tag{3.45}$$

We also have

$$\begin{aligned}|K_n^2|_{L^2}^2 &\le c_1\delta^{2k}\|\theta^n\|^2|A(\eta_j^n+w_j^0)|_{L^2}^2 + c_1\delta\|\vartheta^n+v^0\|^2|Aq_j^n|_{L^2}^2 \\ &\quad + c_1\delta\|\eta_j^n+w_j^0\|^2|Aq_j^n|_{L^2}^2 + c_1\delta^{2k}\|\vartheta^{n-1}+v^0\|^2|Aq_j^n|_{L^2}^2 \\ &\quad + c_1\delta\|\theta^n\|^2|A(\eta_j^{n-1}+w_j^0)|_{L^2}^2 \\ &\quad + c_1\delta\|q_j^n\|^2|A(\eta_j^{n-1}+w_j^0)|_{L^2}^2 + c_1\delta^2|A(\theta^n+q_j^n)|_{L^2}^2\end{aligned} \tag{3.46}$$

and

$$\begin{aligned}\int_0^T |K_n^2|_{L^2}^2 dt &\le c_1\delta^{2k}\sup_s \|\theta^n(s)\|^2 \int_0^T |A(\eta_j^n+w_j^0)|_{L^2}^2 ds + c_1\delta \sup_s \|\vartheta^n(s) \\ &\quad + v^0(s)\|^2 \int_0^T |Aq_j^n|_{L^2}^2 dt + c_1\delta \sup_s \|\eta_j^n(s)+w_j^0(s)\|^2 \int_0^T |Aq_j^n|_{L^2}^2 dt \\ &\quad + c_1\delta^{2k}\sup_s \|\vartheta^{n-1}(s)+v^0(s)\|^2 \int_0^T |Aq_j^n|_{L^2}^2 dt \\ &\quad + c_1\delta \sup_s \|\theta^n(s)\|^2 \int_0^T |A(\eta_j^{n-1}+w_j^0)|_{L^2}^2 dt \\ &\quad + c_1\delta \sup_s \|q_j^n(s)\|^2 \int_0^T |A(\eta_j^{n-1}+w_j^0)|_{L^2}^2 dt \\ &\quad + c_1\delta^2 \int_0^T |A(\theta^n+q_j^n)|_{L^2}^2 dt \le c_3\delta^{2k}(\|\theta^n\|_{X_1}^2 + \sup_j \|q_j^n\|_{X_2}^2).\end{aligned} \tag{3.47}$$

Finally (3.43) follows from (3.45) and (3.47). ∎

Note that from the previous estimates, c_3 has the form

$$c_3 = c(\alpha_2^2 + R^2). \tag{3.48}$$

It follows from (2.20) and (3.14) that $\delta^{2k}c_3$ goes to zero as δ goes to zero. Hereafter, we choose δ small enough such that

$$\delta_1 = \delta^{2k}c_3 < 1. \tag{3.49}$$

PROPOSITION 3.5. *Let $k \in (1/4, 1/2)$. We assume that δ is small enough so that* (3.35), (3.49) *hold. Then the following estimate holds:*

$$\|\theta^{n+1}\|_{X_1}^2 + \sup_j \|q_j^{n+1}\|_{X_2}^2 \le c_3\delta^{2k}\left(\|\theta^n\|_{X_1}^2 + \sup_j \|q_j^n\|_{X_2}^2\right). \tag{3.50}$$

Proof. The proof, which is similar to those of Propositions 2.4 and 2.5, follows from (3.43). Moreover, the following result is proved.

PROPOSITION 3.6. *Let $k \in (1/4, 1/2)$. We assume that δ is small enough so that* (3.35) *and* (3.49) *are satisfied. Then the sequence $(\vartheta^n, \eta_1^n, \eta_2^n, \dots, \eta_l^n) \in \mathbf{X}$ defined by* (3.18), (3.19) *converges to a solution $(\vartheta, \eta_1, \eta_2, \dots, \eta_l)$ to* (3.18)–(3.19) *in* $\mathbf{X}$. *Moreover, $(\vartheta, \eta_1, \eta_2, \dots, \eta_l)$ is the unique solution to* (3.15)–(3.16) *in* $\mathbf{X}$ *that satisfies*

$$\|\vartheta^n - \vartheta\|_{X_1}^2 + \sup_j \|\eta_j^n - \eta_j\|_{X_2}^2 \le R^2.$$

Furthermore the following convergence rate holds true:

$$\|\vartheta^n - \vartheta\|_{X_1}^2 + \sup_j \|\eta_j^n - \eta_j\|_{X_2}^2 \le \frac{\delta_1^n}{1-\delta_1}, \tag{3.51}$$

where $\delta_1 < 1$ is given by (3.49).

4. Convergence of the method. In this part, we study the convergence of the small eddy correction method presented in the previous section. We prove that the method converges, and we estimate the rate of convergence with respect to the aspect ratio δ.

Hereafter, we set $u_l = v + w_l$, $\zeta = u - u_l$, $\varepsilon_j = w_j - w_{j-1}$. In particular, $\varepsilon_1 = w_1$.

Using (2.18) and (3.5)–(3.6), it is clear that u_l and ζ satisfy (see [13])

$$\begin{aligned}\frac{d}{dt}(u_l,\theta) + a(u_l,\theta) + e(u_l,\theta) + \gamma(u_l,\theta) + b(u_l,u_l,\theta) - (NB(\varepsilon_l,\varepsilon_l),\theta)\\ = (F,\theta)\ \forall\theta\in V,\ u_l(0) = a,\end{aligned} \tag{4.1}$$

$$\begin{aligned}\frac{d}{dt}(\zeta,\theta) + a(\zeta,\theta) + e(\zeta,\theta) + \gamma(\zeta,\theta) + b(\zeta,u,\theta) + b(u_l,\zeta,\theta)\\ + (NB(\varepsilon_l,\varepsilon_l),\theta) = 0\ \forall\theta\in V,\ \zeta(0) = 0.\end{aligned} \tag{4.2}$$

Taking the vertical average of (4.2), we derive that the barotropic and baroclinic flows $\bar\zeta$ and $\zeta^\flat$ satisfy

$$\begin{aligned}\frac{d}{dt}(\bar\zeta,\bar\theta) + a(\bar\zeta,\bar\theta) + e(\bar\zeta,\bar\theta) + b(\bar\zeta,\bar u,\bar\theta) + b(v,\bar\zeta,\bar\theta) + b(\zeta^\flat,u^\flat,\bar\theta)\\ + b(w_l,\zeta^\flat,\bar\theta) = 0\ \forall\theta\in V,\ \bar\zeta(0) = 0,\end{aligned} \tag{4.3}$$

$$\begin{aligned}&\frac{d}{dt}(\zeta^\flat,\theta^\flat) + a(\zeta^\flat,\theta^\flat) + e(\zeta^\flat,\theta^\flat) + \gamma(\bar\zeta+\zeta^\flat,\theta^\flat) + b(\bar\zeta,u^\flat,\theta^\flat) + b(\zeta^\flat,\bar u,\theta^\flat)\\ &\quad + b(v,\zeta^\flat,\theta^\flat) + b(w_l,\bar\zeta,\theta^\flat) + b(\zeta^\flat,u^\flat,\theta^\flat)\\ &\quad + b(w_l,\zeta^\flat,\theta^\flat) + b(\varepsilon_l,\varepsilon_l,\theta^\flat) = 0\ \forall\theta\in V,\ \zeta^\flat(0) = 0.\end{aligned} \tag{4.4}$$

Note that

$$|b(\bar\zeta,\bar u,A\bar\zeta)| \le c_0|\bar\zeta|_{L^2}^{1/2}\|\bar u\||A\bar\zeta|_{L^2}^{3/2} \le \frac{1}{8}|A\bar\zeta|_{L^2}^2 + c_0|\bar\zeta|_{L^2}^2\|\bar u\|^4, \tag{4.5}$$

$$|b(\zeta^\flat, u^\flat, A\bar{\zeta})| \le c_1\delta^{1/2}\|u^\flat\||A\zeta^\flat|_{L^2}|A\bar{\zeta}|_{L^2} \le \frac{1}{8}|A\bar{\zeta}|_{L^2}^2 + c_1\delta\|u^\flat\|^2|A\zeta^\flat|_{L^2}^2, \tag{4.6}$$

$$|b(w_l, \zeta^\flat, A\bar{\zeta})| \le c_1\delta^{1/2}\|\zeta^\flat\||Aw_l|_{L^2}|A\bar{\zeta}|_{L^2} \le \frac{1}{8}|A\bar{\zeta}|_{L^2}^2 + c_1\delta\|\zeta^\flat\|^2|Aw_l|_{L^2}^2, \tag{4.7}$$

$$|e(\bar{\zeta}, A\bar{\zeta})| \le c_0\|\bar{\zeta}\||A\bar{\zeta}|_{L^2} \le \frac{1}{8}|A\bar{\zeta}|_{L^2}^2 + c_0\|\bar{\zeta}\|^2. \tag{4.8}$$

We also have

$$|b(\bar{\zeta}, u^\flat, A\zeta^\flat)| \le c_1\delta^k\|\bar{\zeta}\||Au^\flat|_{L^2}|A\zeta^\flat|_{L^2} \le \frac{1}{8}|A\zeta^\flat|_{L^2}^2 + c_1\delta^{2k}\|\bar{\zeta}\|^2|Au^\flat|_{L^2}^2, \tag{4.9}$$

$$|b(\zeta^\flat, \bar{u}, A\zeta^\flat)| \le c_1\delta^{1/2}\|\bar{u}\||A\zeta^\flat|_{L^2}^2, \tag{4.10}$$

$$|b(v, \zeta^\flat, A\zeta^\flat)| \le c_1\delta^k\|v\||A\zeta^\flat|_{L^2}^2, \tag{4.11}$$

$$|b(w_l, \bar{\zeta}, A\zeta^\flat)| \le c_1\delta^{1/2}\|\bar{\zeta}\||Aw_l|_{L^2}|A\zeta^\flat|_{L^2} \le \frac{1}{8}|A\zeta^\flat|_{L^2}^2 + c_1\delta|Aw_l|_{L^2}^2\|\bar{\zeta}\|^2, \tag{4.12}$$

$$|b(\zeta^\flat, u^\flat, A\zeta^\flat)| \le c_1\delta^{1/2}\|u^\flat\||A\zeta^\flat|_{L^2}^2, \tag{4.13}$$

$$|b(w_l, \zeta^\flat, A\zeta^\flat)| \le c_1\delta^{1/2}\|\zeta^\flat\||Aw_l|_{L^2}|A\zeta^\flat|_{L^2} \le \frac{1}{8}|A\zeta^\flat|_{L^2}^2 + c_1\delta|Aw_l|_{L^2}^2\|\zeta^\flat\|^2, \tag{4.14}$$

$$|b(\varepsilon_l, \varepsilon_l, A\zeta^\flat)| \le c_1\delta^{1/2}\|\varepsilon_l\||A\varepsilon_l|_{L^2}|A\zeta^\flat|_{L^2} \le \frac{1}{8}|A\zeta^\flat|_{L^2}^2 + c_1\delta|A\varepsilon_l|_{L^2}^2\|\varepsilon_l\|^2, \tag{4.15}$$

$$|e(\zeta^\flat, A\zeta^\flat)| \le c_1\delta|A\zeta^\flat|_{L^2}^2, \tag{4.16}$$

$$|\gamma(\bar{\zeta} + \zeta^\flat, A\zeta^\flat)| \le c\delta|A(\bar{\zeta} + \zeta^\flat)|_{L^2}|A\zeta^\flat|_{L^2} \le c\delta(|A\bar{\zeta}|_{L^2}^2 + |A\zeta^\flat|_{L^2}^2). \tag{4.17}$$

Let

$$\beta_1 = 1 - c_1\delta\|u^\flat\|^2 - c_1\delta^{2k}\|\bar{\zeta}\|^2 - c_1\delta^{1/2}\|\bar{u}\| - c_1\delta^k\|v\| - c_1\delta^{1/2}\|u^\flat\| - c_1\delta - c\delta. \tag{4.18}$$

We choose $k \in (1/4, 1/2)$ and δ small enough such that

$$\beta_1 > \frac{1}{2}. \tag{4.19}$$

This choice is possible since

$$\begin{aligned}
&\|\bar{u}\|^2 \le 2\sigma(\delta R_0^2(\delta) + \sigma R_0^2(\delta)),\\
&\|v\|^2 \le 2(R^2 + \alpha_1^2),\\
&\|\bar{\zeta}\|^2 \le 2\|\bar{u}\|^2 + 2\|v\|^2 \le 4(R^2 + \alpha_1^2 + \sigma\delta R_0^2(\delta) + \sigma R_0^2(\delta)),\\
&\|u^\flat\|^2 \le \sigma\delta^{1/4}R_0^2(\delta),
\end{aligned} \tag{4.20}$$

and (2.19)–(2.20) are satisfied.

Therefore, multiplying (4.3) by $A\bar{\zeta}$, (4.4) by $A\zeta^\flat$, and using (4.5)–(4.19) yield

$$\begin{aligned}&\frac{d}{dt}(\|\bar{\zeta}\|^2+\|\zeta^\flat\|^2)+\frac{1}{2}(|A\bar{\zeta}|_{L^2}^2+|A\zeta^\flat|_{L^2}^2)\\&\quad\le c_1\delta^{2k}\|\bar{\zeta}\|^2|Au^\flat|_{L^2}^2+c_0\|\bar{\zeta}\|^2+c_0\|\bar{\zeta}\|^2\|\bar{u}\|^4\\&\qquad+c_1\delta(\|\zeta^\flat\|^2|Aw_l|_{L^2}^2+\|\bar{\zeta}\|^2|Au^\flat|_{L^2}^2+\|\varepsilon_l\|^2|A\varepsilon_l|_{L^2}^2+|Aw_l|_{L^2}^2\|\bar{\zeta}\|^2),\end{aligned}\tag{4.21}$$

which gives

$$\|\bar{\zeta}(t)\|^2+\|\zeta^\flat(t)\|^2+\frac{1}{2}\int_0^t(|A\bar{\zeta}|_{L^2}^2+|A\zeta^\flat|_{L^2}^2)ds\le c_1\delta\sup_s\|\varepsilon_l(s)\|^2\int_0^t|A\varepsilon_l|_{L^2}^2ds.\tag{4.22}$$

The next step is to derive some a priori estimates on ε_k.

Note that ε_j satisfies (for $2\le j\le l$)

$$\begin{aligned}&\frac{d}{dt}(\varepsilon_j,\theta^\flat)+a(\varepsilon_j,\theta^\flat)+e(\varepsilon_j,\theta^\flat)+\gamma(\varepsilon_j,\theta^\flat)+b(v,\varepsilon_j,\theta^\flat)+b(\varepsilon_j,v,\theta^\flat)\\&+b(w_{j-1},\varepsilon_j,\theta^\flat)+b(\varepsilon_j,w_{j-1},\theta^\flat)+b(\varepsilon_{j-1},\varepsilon_{j-1},\theta^\flat)=0\ \forall\theta\in V,\ \varepsilon_j(0)=0.\end{aligned}\tag{4.23}$$

We have

$$|b(v,\varepsilon_j,A\varepsilon_j)|\le c_1\delta^k\|v\||A\varepsilon_j|_{L^2}^2.\tag{4.24}$$

$$|b(\varepsilon_j,v,A\varepsilon_j)|\le c_1\delta^{1/2}\|v\||A\varepsilon_j|_{L^2}^2.\tag{4.25}$$

$$\begin{aligned}|b(w_{j-1},\varepsilon_j,A\varepsilon_j)|&\le c_1\delta^{1/2}\|\varepsilon_j\||Aw_{j-1}|_{L^2}|A\varepsilon_j|_{L^2}\\&\le\frac{1}{8}|A\varepsilon_j|_{L^2}^2+c_1\delta\|\varepsilon_j\|^2|Aw_{j-1}|_{L^2}^2,\end{aligned}\tag{4.26}$$

$$|b(\varepsilon_j,w_{j-1},A\varepsilon_j)|\le c_1\delta^{1/2}\|w_{j-1}\||A\varepsilon_j|_{L^2}^2,\tag{4.27}$$

$$\begin{aligned}|b(\varepsilon_{j-1},\varepsilon_{j-1},A\varepsilon_j)|&\le c_1\delta^{1/2}\|\varepsilon_{j-1}\||A\varepsilon_{j-1}|_{L^2}|A\varepsilon_j|_{L^2}\\&\le\frac{1}{8}|A\varepsilon_j|_{L^2}^2+c_1\delta\|\varepsilon_{j-1}\|^2|A\varepsilon_{j-1}|_{L^2}^2.\end{aligned}\tag{4.28}$$

Let

$$\beta_2=1-c_1\delta^{1/2}\|v\|-c_1\delta^k\|v\|.\tag{4.29}$$

We choose δ small enough such that

$$\beta_2>\frac{1}{2}.\tag{4.30}$$

This choice is possible since $\|v\|^2\le 2R^2+2\alpha_1^2$, which shows that $\delta^k\|v\|$ and $\delta^{1/2}\|v\|$ go to zero as δ goes to zero for $k\in(1/4,1/2)$.

Therefore, multiplying (4.23) by $A\varepsilon_j$ and using (4.25)–(4.28) yield

$$\|\varepsilon_j(t)\|^2 + \int_0^t |A\varepsilon_j|_{L^2}^2 ds \le c_1\delta \sup_s \|\varepsilon_{j-1}(s)\|^2 \int_0^t |A\varepsilon_{j-1}|_{L^2}^2 ds. \tag{4.31}$$

Therefore

$$\sup_s \|\varepsilon_j(s)\|^2 \int_0^T |A\varepsilon_j|_{L^2}^2 ds \le c_1\delta^2 \left(\sup_s \|\varepsilon_{j-1}(s)\|^2\right)^2 \left(\int_0^T |A\varepsilon_{j-1}|_{L^2}^2 ds\right)^2. \tag{4.32}$$

THEOREM 4.1. *We assume that the data satisfy* (2.20). *Then for δ small enough, the error $\zeta(t) = u(t) - u_l(t)$ satisfies*

$$\|\bar{\zeta}(t)\|^2 + \|\zeta^\flat(t)\|^2 + \int_0^t (|A\bar{\zeta}(s)|_{L^2}^2 + |A\zeta^\flat(s)|_{L^2}^2) ds \le c_4(\delta^{1/2})^{2^{l-1}}, \tag{4.33}$$

where c_4 is a constant that is independent of δ for δ small enough.

Proof. Let us set

$$\begin{aligned} E(t) &= \|\bar{\zeta}(t)\|^2 + \|\zeta^\flat(t)\|^2 + \int_0^t (|A\bar{\zeta}(s)|_{L^2}^2 + |A\zeta^\flat(s)|_{L^2}^2) ds, \\ x_j &= \sup_s \|\varepsilon_j(s)\|^2 \int_0^T |A\varepsilon_j|_{L^2}^2 ds. \end{aligned} \tag{4.34}$$

It follows from (4.22) and (4.31) that

$$E(t) \le c\delta x_l, \quad x_j \le c\delta^2 x_{j-1}^2 = \delta_2^2 x_{j-1}^2, \quad \delta_2 = \sqrt{c}\delta. \tag{4.35}$$

By iteration we derive that

$$x_j \le \delta_2^{(2^{j-1}-1)2} x_1^{2^{j-1}}. \tag{4.36}$$

Let us now estimate $\varepsilon_1 = w_1$ and x_1. Note that (see (2.26))

$$\begin{aligned} |e(w_1, Aw_1)| &\le c\delta^2 |Aw_1|_{L^2}^2, \\ |\gamma(v + w_1, Aw_1)| &\le c_1\delta |A(v + w_1)|_{L^2} |Aw_1|_{L^2} \\ &\le \frac{1}{4}|Aw_1|_{L^2}^2 + c_1\delta^2 |Av|_{L^2}^2 + c_1\delta |Aw_1|_{L^2}^2, \\ |b(v, w_1, Aw_1)| &\le c_1\delta^k \|v\| |Aw_1|_{L^2}^2, \\ |b(w_1, v, Aw_1)| &\le c_1\delta^{1/2} \|v\| |Aw_1|_{L^2}^2. \end{aligned} \tag{4.37}$$

Let us set

$$\beta_3 = 1 - c_1\delta^2 - c_1\delta - c_1\delta^k \|v\| - c_1\delta^{1/2} \|v\|. \tag{4.38}$$

Now we assume that δ is small enough (using an argument similar to that of β_2) such that

$$\beta_3 > \frac{1}{4}. \tag{4.39}$$

Then using (3.9) and (4.37)–(4.39) we derive that

$$\frac{d}{dt}\|w_1\|^2 + |Aw_1|_{L^2}^2 \leq c_1\delta^2|Av|_{L^2}^2 + c_1|F^\flat|_{L^2}^2, \tag{4.40}$$

which gives

$$\begin{aligned}\|w_1(t)\|^2 + \int_0^t |Aw_1|_{L^2}^2 &\leq \|a^\flat\|^2 + c\delta^2\int_0^t |Av|_{L^2}^2 ds + c_1\int_0^t |F^\flat|_{L^2}^2 ds \\ &\leq c_1(\delta^{1/4}R_0^2(\delta) + \delta^2R^2 + \delta^2\alpha_1^2)\end{aligned} \tag{4.41}$$

and

$$x_1 \leq c(\delta^{1/4}R_0^2(\delta) + \delta^2R^2 + \delta^2\alpha_1^2)^2. \tag{4.42}$$

From (2.19)–(2.20), it is clear that $\sqrt{c}\delta^{1/2}x_1$ goes to zero as δ goes to zero, where c is the constant that appears in (4.22). Finally from (4.35)–(4.36), we have

$$E(t) \leq c_4(\delta^{1/2})^{2^{l-1}}, \tag{4.43}$$

where $c_4 > 0$ is a constant independent of δ for δ small enough.

5. Numerical results. In this section, we present some numerical simulations obtained with the models (2.18) and (3.5)–(3.6). The goal is to compare the solution obtained with (2.18) to that of the small eddy correction method. Hereafter we restrict ourselves to $l = 1$, and (3.8)–(3.9) will be referred to hereafter as the reduced model. More simulations for larger values of l will be presented elsewhere. To avoid dealing with the divergence-free condition, we first rewrote the barotropic equations in the vorticity streamfunction formulation.

In our experiments, the basin configuration is the (nondimensional) cube $[0,1] \times [0,1] \times [-1,0]$. Let us simply recall that this is a two-gyre, wind-driven ocean-type problem with steady sinusoidal wind stress (maximum $\tau_0 = 1$ dyne cm^{-4}) in a basin that is $L_1 \times L_1 \times H_1$ km (east-west $\times$ north-south $\times$ bottom-surface extent). The Coriolis parameter is given by $f = f_0 + \beta y$, $f_0 = 9.3\times10^{-5}\text{s}^{-1}$, $\beta = 2.\times10^{-11}$ m^{-1}s^{-1}. The model does not include bottom topography. The ocean is forced by a steady wind stress $\tau_0 = (\tau_0^x, \tau_0^y) = (-10^{-4}\cos(2\pi y/L_1), 0)$. Other dimensional quantities are given by $U_1 = 10^{-1}$m s^{-1}, $g = 9.8$m s^{-2}, $L_1 = 2.10^6$ m, and $H_1 = 4000$m, which gives $\delta = H_1/L_1 = 2.10^{-3}$. The initial condition is given by $u = 0$ at $t = 0$. All the operators in (2.18) and (3.8)–(3.9) are discretized using a second-order central difference scheme. The Jacobian operator is approximated using Arakawa's method [30]. For the time integration, we use a fourth-order Adams–Bashforth method. For the space discretization, we take 100×100 points in the x-y plane and 10 points on the vertical direction. For the boundary conditions, we replace (2.2) by the following physically more acceptable boundary conditions:

$$\begin{cases} \dfrac{\partial v}{\partial z} = 0,\ \dfrac{\partial \rho}{\partial z} = -\alpha_T\rho \text{ at } z = 0, \\ \dfrac{\partial v}{\partial z} = 0,\ \dfrac{\partial \rho}{\partial z} = 0 \text{ at } z = -\delta, \\ v = 0,\ \dfrac{\partial \rho}{\partial n} = 0 \text{ on } \partial\omega \times (-\delta, 0). \end{cases} \tag{5.1}$$

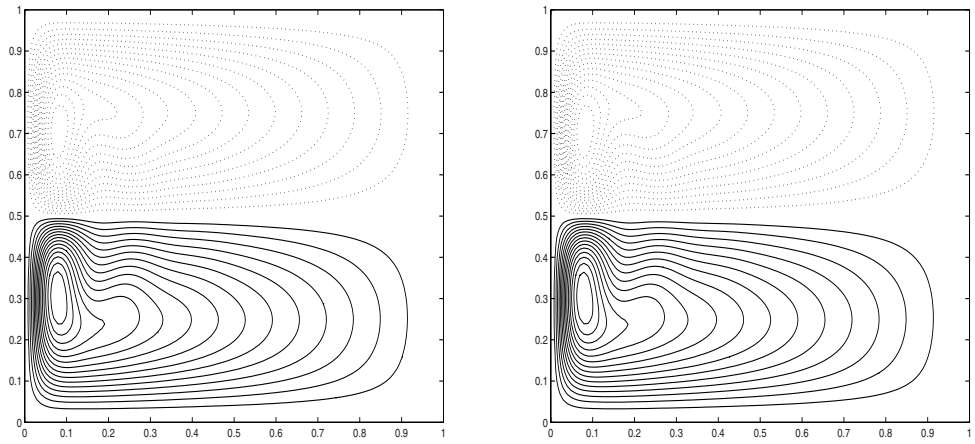


FIG. 1. *Barotropic streamfunction at the steady state. Original model* (2.18) *on the left and reduced model* (3.8)–(3.9) *on the right.*

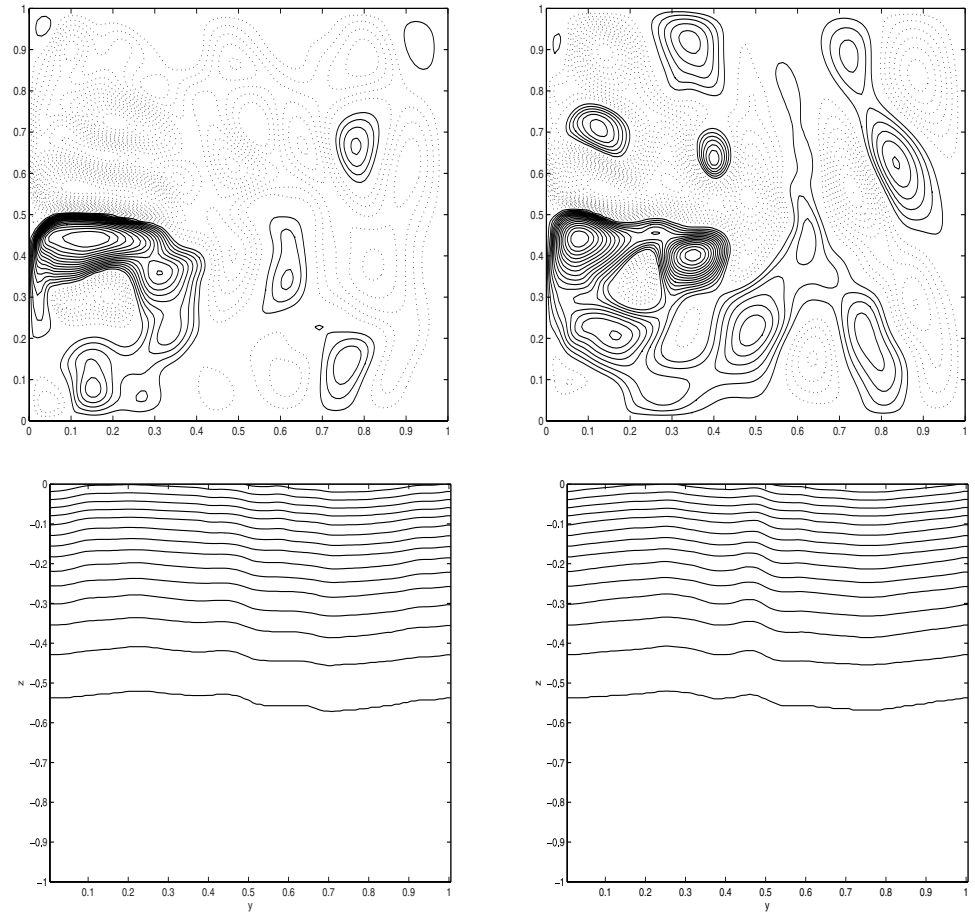


FIG. 2. *Snapshot at the (nondimensional) time* $t = 5$ *of the barotropic streamfunction and the total density at* $x = 0.25$. *PEs on the left and reduced model* (3.8)–(3.9) *on the right.*

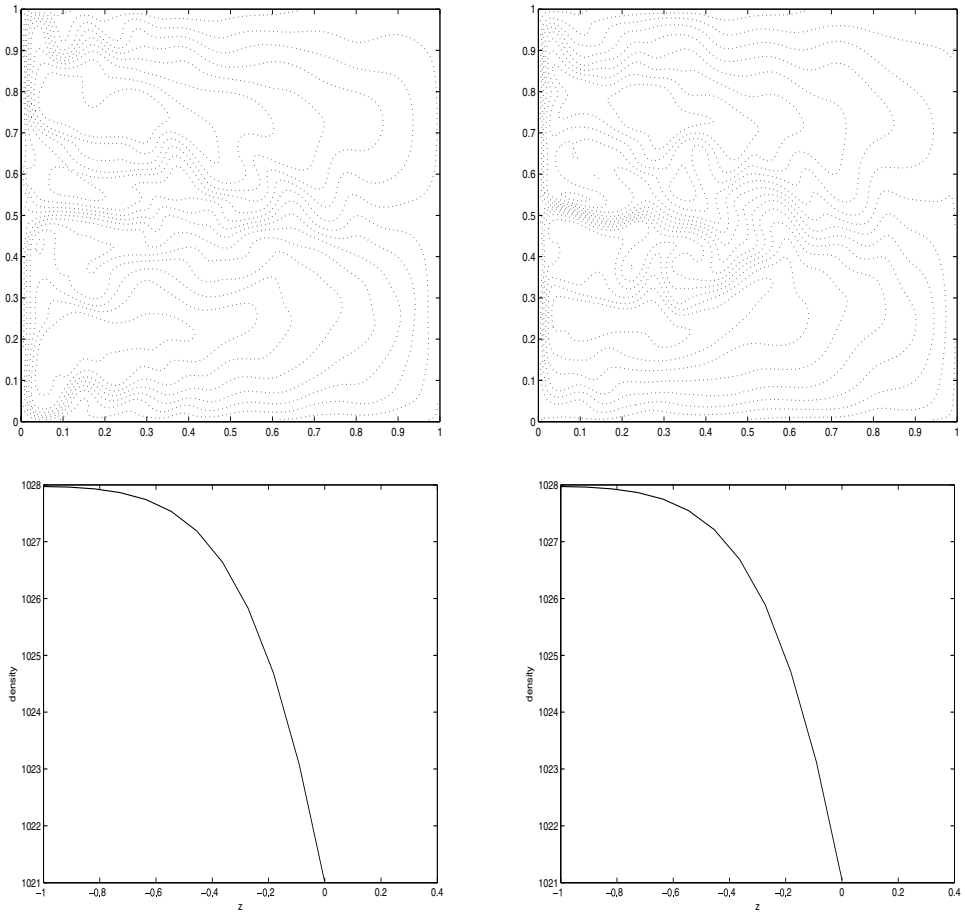


FIG. 3. *Snapshot at the (nondimensional) time $t = 5$ of the surface density deviation and the total density at $(x, y) = (0.5, 0.5)$. PEs on the left and reduced model* (3.8)–(3.9) *on the right.*

The forcing term $F = (F_1, F_2)$ in (2.5) is defined by

$$F_1 = c(g(z)\tau_0, 0), \qquad F_2(z) = c\frac{\partial^2 \rho_s}{\partial z^2}/\rho_0, \tag{5.2}$$

where $g(z)$ and ρ_s are given by

$$g(z) = 0.5(1 + \tanh((z/H_1 + z_1)/\epsilon_1)), \qquad \rho_s(z) = 1028 - 3\exp(10z/H_1). \tag{5.3}$$

In (5.2)–(5.3), z_1 and ϵ_1 are very small constants chosen such that the forcing F_1 is nonzero only on the first couple layers from the surface of the ocean and c is a constant.

Simulation 1: Steady state solutions.

In this simulation, we compare the two models (2.18) and (3.8)–(3.9) when the solution converges to a steady state. For $R_{e_1} = R_{e_2} = 100$, the solutions obtained with the two models converge to a steady state. Figure 1 shows the barotropic streamfunctions obtained with the two models. As one can see, model (3.8)–(3.9) approximates very well the original model (2.18).

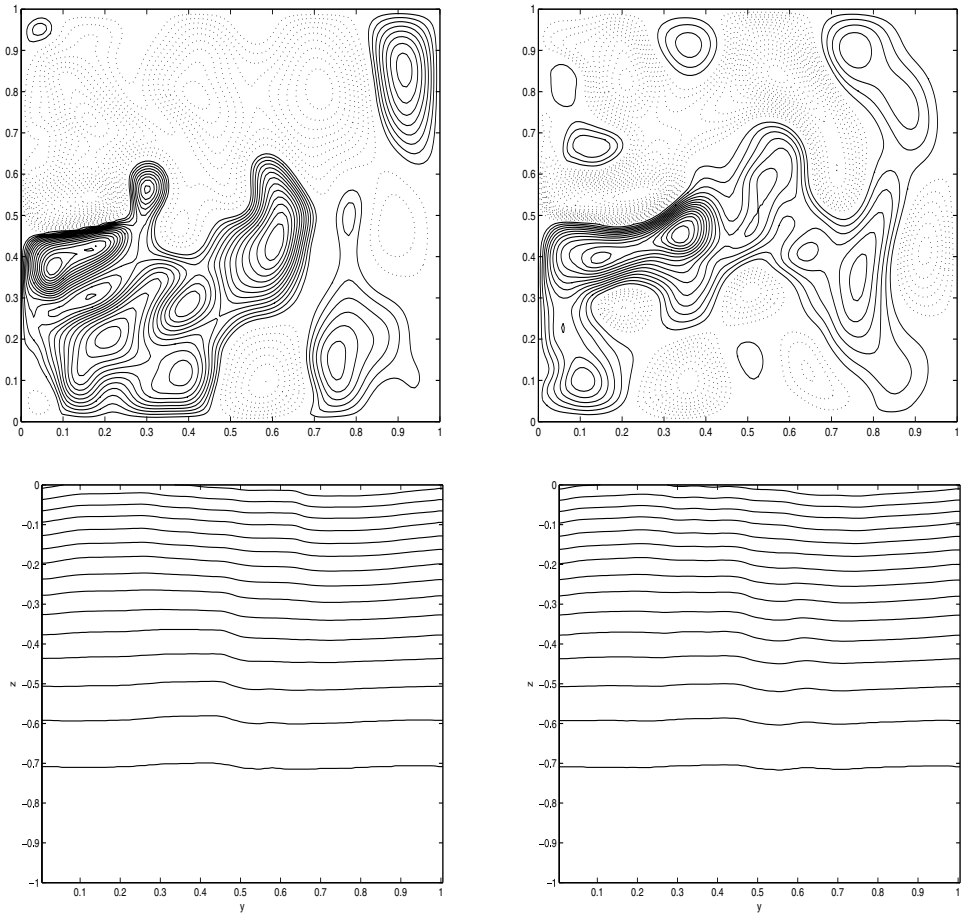


FIG. 4. *Snapshot at the (nondimensional) time* $t = 10$ *of the barotropic streamfunction and the total density at* $x = 0.25$. *PEs on the left and reduced model* (3.8)–(3.9) *on the right.*

Simulation 2: Time-dependent solutions.

For the two models (2.18) and (3.8)–(3.9), Figure 2 (resp., Figure 4) shows the barotropic streamfunction and the total density $\rho + \rho_s$ at $x = 0.25$ for the Reynolds number $R_{e_1} = 2.10^3, R_{e_2} = 10^2$ and at the time $t = 5$ (resp., $t = 10$). For the same values of the Reynolds number, Figure 3 (resp., Figure 5) shows the surface density deviation ρ and the total density $\rho + \rho_s$ at the point $(x, y) = (0.5, 0.5)$ and at the time $t = 5$ (resp., $t = 10$). For these values of the Reynolds number, the flow remains time-dependent. From these figures, we observe that the solutions obtained with the two models present some differences. However, Figure 6 shows that the time-averages of the two flows are very similar. This seems to confirm what is already believed in oceanography: from the climate point of view (where the main focus is on the time-average of the flow), the interactions between the baroclinic and the barotropic modes do not need to be accurately represented [10]. From an efficiency point of view, it was noticed that the reduced model (3.8)–(3.9) allows a larger time step than the original model (2.18). With $100 \times 100 \times 10$ grid points, for instance, the maximum time step allowed for (2.18) was around 10^{-4}, while the reduced model was still stable with $\triangle t = 3 \times 10^{-4}$. This result is not surprising. In fact, since the equation for the

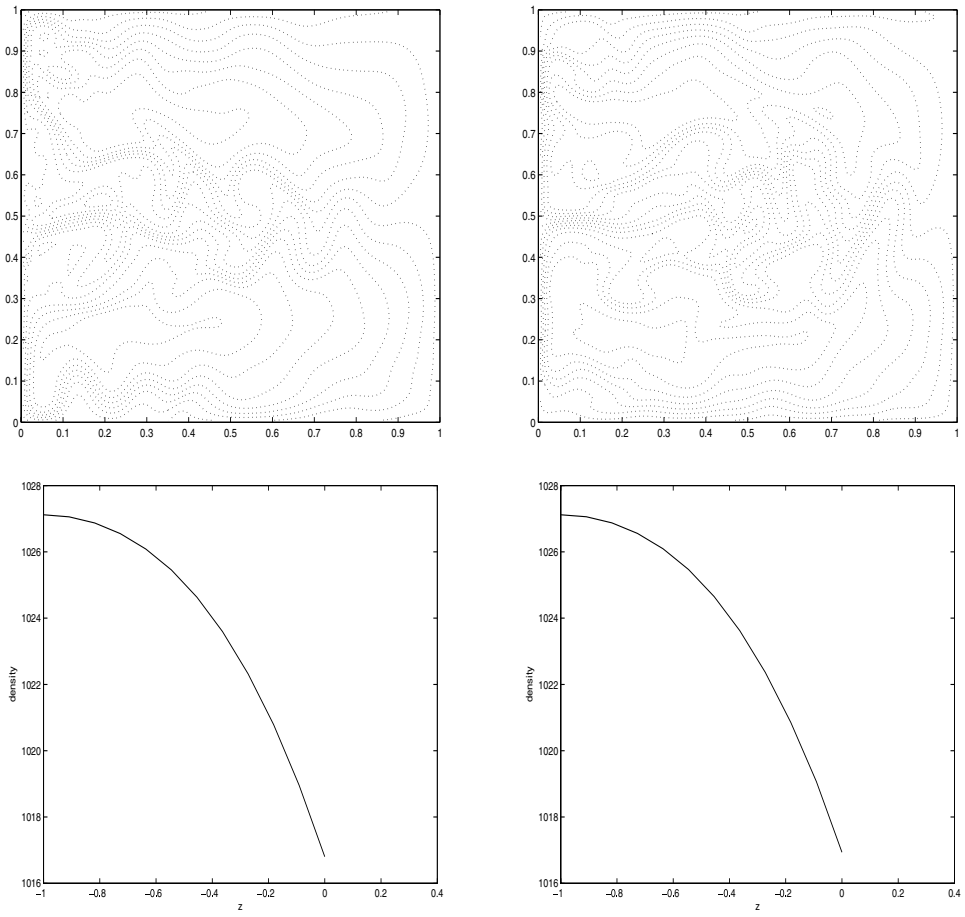


FIG. 5. *Snapshot at the (nondimensional) time* $t = 10$ *of the surface density deviation and the total density at* $(x, y) = (0.5, 0.5)$. *PEs on the left and reduced model* (3.8)–(3.9) *on the right.*

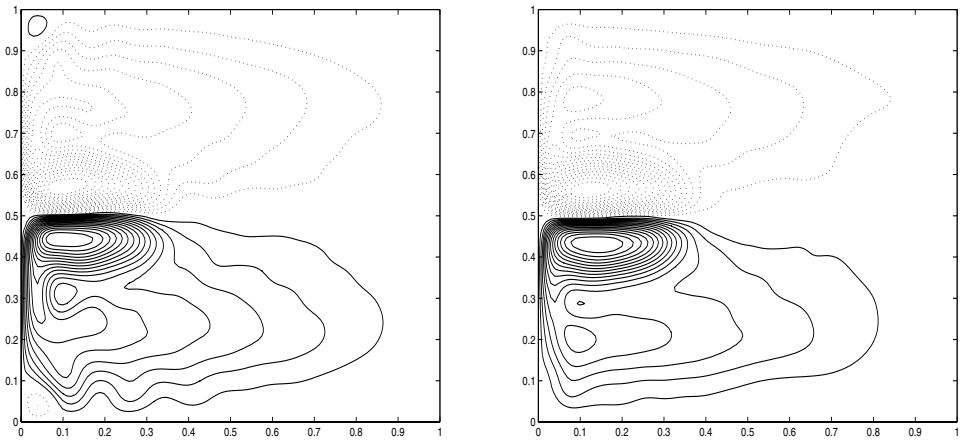


FIG. 6. *Time-average (over* $[0, 10]$*) of the barotropic streamfunction. PEs on the left and reduced model* (3.8)–(3.9) *on the right.*

small scales is (highly) nonlinear for the model (2.18) and linearized in the model (3.8)–(3.9), it is expected that the small scales will greatly influence the time step in (2.18) more than in (3.8)–(3.9), especially for turbulent flows.

6. Conclusion. The purpose of this article was to present a small eddy correction method for the numerical solution to (2.5). Considering the interaction between the baroclinic and barotropic flows and using the idea of the Newton iteration, a small eddy correction method was proposed for (2.5). We assume that the barotropic approximation to the solution is known. Formally applying the Newton iterative procedure to the baroclinic flow equation, we then generate approximate systems. It was shown that the initial step ($l = 0$) leads to the well-known quasi-geostrophic equations and the next step ($l = 1$) yields an NLG-type approximation. Some numerical simulations for $l = 1$ show that the method can accurately approximate the PEs system of (2.18). For the simulation presented in this article, it was observed that with the same time discretization, the reduced model allows a larger time step (about three times larger) than the original model. An efficient time discretization of the model presented in this article may lead to even more considerable savings in calculation costs. In fact, such method should take advantage of the time scale differences between the barotropic and baroclinic modes. For instance,

- one can use different time steps for the small and large scales;
- one can use different schemes for the small and large scales;
- one can freeze the small scales over an interval of time.

These approaches, already used with success (considerable reduction in CPU cost) in the context of the NLG method and the Navier–Stokes equations (see [9, 17]), are currently under development by the author. Let us also recall that in climate research, it would be extremely useful to be able to parameterize the nonlinear effects of the small scales on big scales, using simplified models that do not require costly solutions of the whole ocean model at high resolution. We believe that the research presented in this article is a step in this direction.

Acknowledgment. The author would like to thank the anonymous referees whose comments helped to greatly improve the content of this article.

REFERENCES

[1] D. Bresch, F. Guill'en-Gonzalez, N. Masmoudi, and M. A. Rodriguez-Bellido, *On the uniqueness of weak solutions of the two-dimensional primitive equations*, Differential Integral Equations, 16 (2003), pp. 77–94.

[2] D. Bresch, A. Kazhikhov, and J. Lemoine, *On the two-dimensional hydrostatic Navier–Stokes equation*, SIAM J. Math. Anal., 36 (2004), pp. 796–814.

[3] C. Cao and E. S. Titi, *Global well-posedness of the three-dimensional viscous primitive equations of large scale ocean and atmosphere dynamics*, Ann. Math., to appear.

[4] T. Chacón Rebollo, R. Lewandowski, and E. Chacón Vera, *Analysis of the hydrostatic approximation in oceanography with compression term*, Math. Model. Numer. Anal., 34 (2000), pp. 525–537.

[5] J. G. Charney, *Geostrophic turbulence*, J. Atmos. Sci., 28 (1971), pp. 1087–1095.

[6] J. Y. Chemin, B. Desjardins, I. Gallagher, and E. Grennier, *Anisotropie et dispersion dans les fluides tournants*, C. R. Acad. Sci. Paris Sér. I. Math., 329 (1999), pp. 1055–1058.

[7] J. Y. Chemin, B. Desjardins, I. Gallagher, and E. Grennier, *Ekman boundary layers in rotating fluids. A tribute to J. L. Lions*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 441–466.

[8] B. Di Martino and P. Orenga, *Resolution to a three-dimensional physical oceanographic problem using the non-linear Galerkin method*, Internat. J. Numer. Methods Fluids, 30 (1999), pp. 577–606.

[9] T. Dubois, F. Jauberteau, and R. Temam, *Dynamic Multilevel Methods and the Numerical Simulation of Turbulence*, Cambridge University Press, Cambridge, UK, 1999.
[10] S. M. Griffies, C. Boening, F. O. Bryan, E. P. Chassignet, R. Gerdes, H. Hasumi, A. Hirst, A. M. Treguier, and D. Webb, *Developments in ocean climate modelling*, Ocean Modelling, 2 (2000), pp. 123–192.
[11] Y. He, Y. Hou, and K. Li, *Stability and convergence of optimum spectral non-linear Galerkin methods*, Math. Methods Appl. Sci., 24 (2001), pp. 289–317.
[12] Y. He, K. M. Liu, and W. W. Sun, *A multi-level finite element method in space-time for the Navier-Stokes equations*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 1052–1078.
[13] Y. Hou and K. Li, *A small eddy correction method for nonlinear dissipative evolutionary equations*, SIAM J. Numer. Anal., 41 (2003), pp. 1101–1130.
[14] C. Hu, *Asymptotic analysis of the primitive equations under the small depth assumption*, Nonlinear Anal., 61 (2005), pp. 425–460.
[15] C. Hu, R. Temam, and M. Ziane, *Regularity results for linear elliptic problems related to the primitive equations*, Chinese Ann. Math. Ser. B, 23 (2002), pp. 1–16.
[16] C. Hu, R. Temam, and M. Ziane, *The primitive equations of the large scale ocean under the small depth hypothesis*, Discrete Contin. Dyn. Syst., 9 (2003), pp. 97–131.
[17] F. Jauberteau, C. Rosier, and R. Temam, *The nonlinear Galerkin method in computational fluid dynamic*, Appl. Numer. Math., 6 (1990), pp. 361–370.
[18] J. L. Lions, R. Temam, and S. Wang, *New formulations of the primitive equations of the atmosphere and applications*, Nonlinearity, 5 (1992), pp. 237–288.
[19] J. L. Lions, R. Temam, and S. Wang, *On the equations of large-scale ocean*, Nonlinearity, 5 (1992), pp. 1007–1053.
[20] J. L. Lions, R. Temam, and S. Wang, *Models of the coupled atmosphere and ocean (CAO* I*)*, Comput. Mech. Adv., 1 (1993), pp. 3–54.
[21] J. L. Lions, R. Temam, and S. Wang, *Numerical analysis of the coupled atmosphere and ocean models (CAO* II*)*, Comput. Mech. Adv., 1 (1993), pp. 55–120.
[22] J. L. Lions, R. Temam, and S. Wang, *Mathematical study of the coupled models of atmosphere and ocean (CAO* III*)*, Math. Pures Appl., 73 (1995), pp. 105–163.
[23] J. L. Lions, R. Temam, and S. Wang, *On mathematical problems for the primitive equations of the ocean: The mesoscale midlatitude case. Lakshmikantham's legacy: A tribute on his 75th birthday*, Nonlinear Anal., 40 (2000), pp. 439–482.
[24] M. Marion and R. Temam, *Nonlinear Galerkin methods*, SIAM J. Numer. Anal., 26 (1989), pp. 1139–1157.
[25] M. Marion and R. Temam, *Nonlinear Galerkin methods: The finite elements case*, Numer. Math., 57 (1990), pp. 205–226.
[26] R. Temam, *Infinite dimensional dynamical systems in mechanics and physics*, 2nd ed., Appl. Math. Sci. 68, Springer-Verlag, New York, 1988.
[27] R. Temam, *Stability analysis of the nonlinear Galerkin method*, Math. Comp., 57 (1991), pp. 477–505.
[28] R. Temam and M. Ziane, *Navier-Stokes equations in three-dimensional thin domains with various boundary conditions*, Adv. Differential Equations, 1 (1996), pp. 499–546.
[29] R. Temam and M. Ziane, *Some mathematical problems in geophysical fluid dynamics*, in Handbook of Mathematical Fluid Dynamics, Vol. III, S. Friedlander and D. Serre, eds., North-Holland, Amsterdam, 2004, pp. 535–658.
[30] W. M. Washington and C. L. Parkinson, *An Introduction to Three-Dimensional Climate Modeling*, Oxford University Press, Oxford, 1986.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1871–1901

LINEAR INSTABILITY OF THE FIFTH-ORDER WENO METHOD*

RONG WANG† AND RAYMOND J. SPITERI†

Abstract. The weighted essentially nonoscillatory (WENO) methods are popular spatial discretization methods for hyperbolic partial differential equations. In this paper we show that the combination of the widely used fifth-order WENO spatial discretization (WENO5) and the forward Euler time integration method is linearly unstable when numerically integrating hyperbolic conservation laws. Consequently it is not convergent. Furthermore we show that all two-stage, second-order explicit Runge–Kutta (ERK) methods are linearly unstable (and hence do not converge) when coupled with WENO5. We also show that all optimal first- and second-order strong-stability-preserving (SSP) ERK methods are linearly unstable when coupled with WENO5. Moreover the popular three-stage, third-order SSP(3,3) ERK method offers no linear stability advantage over non-SSP ERK methods, including ones with negative coefficients, when coupled with WENO5. We give new linear stability criteria for combinations of WENO5 with general ERK methods *of any order*. We find that a sufficient condition for the combination of an ERK method and WENO5 to be linearly stable is that the linear stability region of the ERK method should include the part of the imaginary axis of the form $[-\iota\mu, \iota\mu]$ for some $\mu > 0$. The linear stability analysis also provides insight into the behavior of ERK methods applied to nonlinear problems and problems with discontinuous solutions. We confirm the assertions of our analysis by means of numerical tests.

Key words. stability analysis, Runge–Kutta methods, WENO method, strong-stability-preserving

AMS subject classifications. 65L06, 65M20

DOI. 10.1137/050637868

1. Introduction.

The method of lines (MOL) is a general approach for the treatment of time-dependent partial differential equations (PDEs) [24]. The standard MOL involves two steps. The first step is to discretize the spatial variables of the PDE to obtain a large set of initial-value ordinary differential equations (ODEs). The second step is to integrate the ODEs using a time integration method such as a linear multistep or Runge–Kutta (RK) method [4, 5].

The essentially nonoscillatory (ENO) methods [6, 7] and the weighted essentially nonoscillatory (WENO) methods [15, 11] are popular and effective nonlinear spatial discretizations for hyperbolic PDEs. These methods are adept at handling the nonsmooth features that arise in the solutions to hyperbolic PDEs. For example, although these methods are formally first-order accurate once a shock is present, they still have uniform high-order accuracy right up to the location of the shock [11]. Specifically, the fifth-order WENO spatial discretization (WENO5) [11], which uses a convex combination of three third-order ENO stencils, is a widely used and robust spatial discretization for numerical solution of hyperbolic conservation laws.

The three-stage, third-order strong-stability-preserving (SSP) explicit RK (ERK) method, which has most recently been referred to as SSP(3,3) [21], is generally viewed as the time integration method of choice to couple with WENO5; see, e.g., [16] and references therein. Numerical results are generically stable and satisfactory [11, 20].

*Received by the editors August 9, 2005; accepted for publication (in revised form) April 23, 2007; published electronically August 31, 2007.

http://www.siam.org/journals/sinum/45-5/63786.html

†Department of Computer Science, University of Saskatchewan, Saskatoon, SK, S7N 5C9, Canada (rong@cs.usask.ca, spiteri@cs.usask.ca). The work of the first author was partially supported by MITACS and Martec, Inc. The work of the second author was partially supported by NSERC Canada, MITACS, and Martec, Inc.

Because of this, it is generally believed that, when WENO5 is used as the spatial discretization for hyperbolic PDEs, the SSP property is a necessary (or advantageous) property for the time integration method to possess [16, 20]. However, for ERK methods applied to hyperbolic conservation laws spatially discretized by WENO5, our work shows that there is a different property, i.e., linear stability, that must be considered.

In this paper we analyze the linear stability of some ERK methods when coupled with WENO5 to solve hyperbolic conservation laws. First we show that the forward Euler method is linearly unstable when coupled with WENO5; i.e., the corresponding CFL number is 0. Consequently the numerical solution does not converge to the true solution for any time step. This means that the stability of WENO5 in general is a product of its nonlinear nature and the particular time integration method with which it is coupled. Furthermore SSP ERK methods [18, 3] derived based on the SSP property of forward Euler cannot be SSP when coupled with WENO5. In fact, there is no guarantee for stability of any kind. In section 3 we show that any optimal s-stage, first- or second-order SSP ERK method [21] is linearly unstable when coupled with WENO5. This result is surprising and contrary to expectation based on existing literature; see, e.g., [16, 20, 25] and references therein. Moreover, we show that, in our analysis, the success of the SSP(3,3) method for the time integration of hyperbolic PDEs spatially discretized by WENO5 is not due to its SSP property; indeed any three-stage, third-order ERK method (even with negative coefficients or that is provably non-SSP) possesses the same linear stability properties. We demonstrate all of these results by means of numerical experiments.

For linear spatial discretizations of linear, constant-coefficient PDE problems posed on unbounded or periodic domains, linear instability guarantees *global* instability; the instability typically leads to spurious oscillations in the numerical solution that ultimately become unbounded [23]. However, WENO methods are not linear spatial discretizations; see, e.g., [9, p. 59] for a definition of linear spatial discretizations for periodic advection problems. Therefore, the behavior of a numerical solution computed from linearly unstable combinations of WENO methods and certain time integration methods for solving hyperbolic conservation laws is more subtle. Such a combination arises, for example, when using the combination of WENO5 and the forward Euler method; see Theorem 3.1 below. In such cases, the linear instability again manifests itself as spurious oscillations in the numerical solution. However, WENO methods attempt to adapt to the instability by changing the associated ENO stencil weights; see section 2 below. The spurious oscillations may still grow to large magnitudes; however, they do not necessarily become unbounded. In other words, the nonlinear nature of WENO methods may be successful in controlling potential instabilities, and this process may take a significant length of time to assert itself. It is important to note that, because they are linearly unstable, these combined methods are not convergent. Hence, although the discussions in this paper are phrased mainly in terms of linear stability analysis, an immediate corollary of every result presented regarding linear instability is the nonconvergence of the combined method. We offer further discussion and illustration of this in section 4; see also Example 1.

The remainder of this paper proceeds as follows. In section 2 we give a brief review of WENO5. In section 3 we prove that, when coupled with WENO5, the forward Euler method and all two-stage, second-order ERK methods are linearly unstable. We then provide criteria for the linear stability of general ERK methods of orders 1 and 2. Immediate consequences of these criteria are that the optimal s-stage ($s \geq 2$), first-order or s-stage ($s \geq 3$), second-order SSP ERK methods are linearly

unstable when coupled with WENO5. It is also easy to show that all three-stage, third-order ERK methods (including SSP(3,3)) and all four-stage, fourth-order ERK methods (including the classical four-stage, fourth-order ERK method) are linearly stable when coupled with WENO5. Finally we find that a sufficient condition for the combination of any ERK time integration method with WENO5 to be linearly stable is that the linear stability region of the ERK method must include the part of the imaginary axis in the form $[-\iota\mu, \iota\mu]$ for some $\mu > 0$. In section 4 we confirm our theoretical results by means of numerical experiments. We also derive and test four new non-SSP ERK methods (a two-stage, first-order method; a three-stage, second-order method; a three-stage, third-order method with negative coefficients; and a low-storage five-stage, third-order method) that are stable according to our linear stability analysis. Numerical results for both linear and nonlinear problems, as well as problems with continuous and discontinuous solutions, demonstrate the relevance of our analysis.

2. The WENO5 method. WENO methods are widely used for the spatial discretization of hyperbolic conservation laws. They were first introduced in [15] as an improvement to ENO methods. ENO methods are based on polynomial interpolation of solution data to define numerical fluxes. They were originally designed to suppress instabilities that lead to spurious oscillations in other commonly used spatial discretizations. To achieve this, ENO methods choose stencils that are adapted to the directions where the solution has an increased order of smoothness. WENO methods take a convex combination of r candidate ENO stencils of order r to produce a method of order $2r-1$ in regions where the solution is smooth while retaining the ENO property in regions where the solution exhibits discontinuous behavior. Specifically, for a given cell, the WENO5 method consists of a convex combination of the three possible third-order ENO stencils containing that cell [11]. Although in practice ENO or WENO methods are very robust, there are very few theoretical results for them [20]. We note that no proof of the stability of either family of methods has yet been given. The WENO5 method is perhaps the most commonly used of the WENO family of methods. We now give a brief summary of some theoretical aspects of WENO methods, with specific implementational details given for the WENO5 method.

Consider the one-dimensional scalar hyperbolic conservation law

$$u_t = -f_x(u), \qquad 0 < x < 1,\ t > 0. \tag{2.1}$$

Assume a uniform spatial mesh, i.e., $x_j = j\Delta x$, $j = 0, 1, \ldots, N$, where $\Delta x = \frac{1}{N}$, and define cells by $I_j = [x_{j-1}, x_j]$, $i = 1, 2, \ldots, N$. We use a conservative finite difference scheme in a MOL approach to write

$$\frac{du_j}{dt} = -\frac{1}{\Delta x}\left(\hat{f}_{j+\frac{1}{2}} - \hat{f}_{j-\frac{1}{2}}\right),$$

where $u_j(t) \approx u(x_j, t)$, $j = 0, 1, \ldots, N$. The term $\hat{f}_{j+\frac{1}{2}} = \hat{f}\left(u_{j-R}, \ldots, u_{j+S}\right)$ is the numerical flux. The numerical flux must be consistent with $f(u)$; i.e., $\hat{f}(u, \ldots, u) = f(u)$. The specification of $\hat{f}(u)$ determines the particular numerical method and its properties. We now derive the specific form of the numerical flux $\hat{f}(u)$ for WENO5.

We first split the flux into positive and negative parts

$$f(u) = f^+(u) + f^-(u). \tag{2.2}$$

This can be accomplished in different ways. In this paper we consider only the Lax–Friedrichs flux splitting [20]

$$f^+(u) = \frac{1}{2}\left(f(u) + mu\right), \qquad f^-(u) = \frac{1}{2}\left(f(u) - mu\right),$$

where $m = \max|f'(u)|$. It is easy to show that $\frac{df^+}{du} \geq 0$ and $\frac{df^-}{du} \leq 0$. However, we note that the same analysis and conclusions apply to other flux-splitting methods.

As in [11], we now calculate the indicators of smoothness IS_i, $i = 0, 1, \ldots, r-1$, associated with the ith stencil. For IS_i^+ we use

$$IS_i^+ = \sum_{m=1}^{r-1} \int_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \Delta x^{2m-1} \left(\frac{\partial^m p_i(x)}{\partial x^m}\right)^2 dx, \quad i = 0, 1, \ldots, r-1,$$

where $p_i(x)$ is the interpolating polynomial of order r to solution data on r cells; i.e., $I_{j+i-r+1}$, $\ldots$, I_{j+i}. In the case of WENO5, $r = 3$, this leads to

$$IS_0^+ = \frac{13}{12}\left(f_{j-2}^+ - 2f_{j-1}^+ + f_j^+\right)^2 + \frac{1}{4}\left(f_{j-2}^+ - 4f_{j-1}^+ + 3f_j^+\right)^2, \tag{2.3}$$

$$IS_1^+ = \frac{13}{12}\left(f_{j-1}^+ - 2f_j^+ + f_{j+1}^+\right)^2 + \frac{1}{4}\left(f_{j-1}^+ - f_{j+1}^+\right)^2, \tag{2.4}$$

$$IS_2^+ = \frac{13}{12}\left(f_j^+ - 2f_{j+1}^+ + f_{j+2}^+\right)^2 + \frac{1}{4}\left(3f_j^+ - 4f_{j+1}^+ + f_{j+2}^+\right)^2, \tag{2.5}$$

and, using an analagous formula for IS_i^-, we have

$$\begin{aligned}
IS_0^- &= \frac{13}{12}\left(f_{j+1}^- - 2f_{j+2}^- + f_{j+3}^-\right)^2 + \frac{1}{4}\left(3f_{j+1}^- - 4f_{j+2}^- + f_{j+3}^-\right)^2,\\
IS_1^- &= \frac{13}{12}\left(f_j^- - 2f_{j+1}^- + f_{j+2}^-\right)^2 + \frac{1}{4}\left(f_j^- - f_{j+2}^-\right)^2,\\
IS_2^- &= \frac{13}{12}\left(f_{j-1}^- - 2f_j^- + f_{j+1}^-\right)^2 + \frac{1}{4}\left(f_{j-1}^- - 4f_j^- + 3f_{j+1}^-\right)^2.
\end{aligned}$$

Next the (nonnormalized) stencil weights take the form

$$\alpha_i^{\pm} = \frac{d_i}{(\epsilon + IS_i)^2}, \quad i = 0, 1, \ldots, r-1,$$

where ϵ is a small positive number that is introduced to avoid the denominator becoming zero. In the numerical experiments of this paper, we choose $\epsilon = 10^{-6}$, which is the value recommended in [11]. In the case of the WENO5 method, we have $d_0 = \frac{1}{10}$, $d_1 = \frac{6}{10}$, $d_2 = \frac{3}{10}$ (see, e.g., [11] for a derivation of the d_i), and

$$\alpha_0^{\pm} = \frac{1}{10}\left(\frac{1}{\epsilon + IS_0^{\pm}}\right)^2, \quad \alpha_1^{\pm} = \frac{6}{10}\left(\frac{1}{\epsilon + IS_1^{\pm}}\right)^2, \quad \alpha_2^{\pm} = \frac{3}{10}\left(\frac{1}{\epsilon + IS_2^{\pm}}\right)^2. \tag{2.6}$$

In order to a achieve a convex combination of ENO stencils, the WENO stencil weights are normalized according to

$$w_i^{\pm} = \frac{\alpha_i}{\sum_{m=0}^{r-1} \alpha_m}, \quad i = 0, 1, \ldots, r-1,$$

to give

$$w_0^{\pm} = \frac{\alpha_0^{\pm}}{\alpha_0^{\pm} + \alpha_1^{\pm} + \alpha_2^{\pm}}, \quad w_1^{\pm} = \frac{\alpha_1^{\pm}}{\alpha_0^{\pm} + \alpha_1^{\pm} + \alpha_2^{\pm}}, \quad w_2^{\pm} = \frac{\alpha_2^{\pm}}{\alpha_0^{\pm} + \alpha_1^{\pm} + \alpha_2^{\pm}}. \tag{2.7}$$

We note that $w_j^{\pm} \in (0,1)$, $j = 0,1,2$, and $w_0^{\pm} + w_1^{\pm} + w_2^{\pm} = 1$, as required.

The numerical fluxes for WENO5 are given by

$$\begin{aligned} \hat{f}^+_{j+\frac{1}{2}} &= w_0^+ \left(\frac{2}{6} f^+_{j-2} - \frac{7}{6} f^+_{j-1} + \frac{11}{6} f^+_j \right) + w_1^+ \left(-\frac{1}{6} f^+_{j-1} + \frac{5}{6} f^+_j + \frac{2}{6} f^+_{j+1} \right) \\ &\quad + w_2^+ \left(\frac{2}{6} f^+_j + \frac{5}{6} f^+_{j+1} - \frac{1}{6} f^+_{j+2} \right) \end{aligned} \tag{2.8}$$

and

$$\begin{aligned} \hat{f}^-_{j+\frac{1}{2}} &= w_2^- \left(-\frac{1}{6} f^-_{j-1} + \frac{5}{6} f^-_j + \frac{2}{6} f^-_{j+1} \right) + w_1^- \left(\frac{2}{6} f^-_j + \frac{5}{6} f^-_{j+1} - \frac{1}{6} f^-_{j+2} \right) \\ &\quad + w_0^- \left(\frac{11}{6} f^-_{j+1} - \frac{7}{6} f^-_{j+2} + \frac{2}{6} f^-_{j+3} \right). \end{aligned}$$

Noting (2.2), the WENO5 method takes the final form

$$\frac{du_j}{dt} = -\frac{1}{\Delta x} \left[\left(\hat{f}^+_{j+\frac{1}{2}} - \hat{f}^+_{j-\frac{1}{2}} \right) + \left(\hat{f}^-_{j+\frac{1}{2}} - \hat{f}^-_{j-\frac{1}{2}} \right) \right]. \tag{2.9}$$

We refer the interested reader to [20] and references therein for further details and discussion of WENO methods.

3. Linear stability analysis. We consider the linear stability properties of various ERK methods when coupled with WENO5 to solve hyperbolic conservation laws (2.1). The CFL number associated with a uniform discretization in both space and time of (2.1) is defined as $\sigma = (\max \frac{\partial f}{\partial u}) \frac{\Delta t}{\Delta x}$. As is usual when performing linear stability analysis, we linearize and freeze coefficients to write

$$u_t = -\lambda u_x.$$

The corresponding CFL number σ is then $\sigma = \frac{\lambda \Delta t}{\Delta x}$. For the purposes of our analysis, it is sufficient to consider the one-dimensional scalar advection equation

$$u_t = -u_x;$$

i.e., $f(u) = u$; any CFL number appearing in our analysis can then be scaled appropriately for more general interpretations.

We begin this section by showing that the combination of WENO5 and the forward Euler ERK method is linearly unstable.

The Lax–Friedrichs flux splitting yields

$$f^+(u) = u, \quad f^-(u) = 0;$$

i.e., the negative part of the flux is zero. Thus, (2.8) takes the form

$$\begin{aligned} \hat{f}^+_{j+\frac{1}{2}} &= w_0^+ \left(\frac{2}{6} u_{j-2} - \frac{7}{6} u_{j-1} + \frac{11}{6} u_j \right) + w_1^+ \left(-\frac{1}{6} u_{j-1} + \frac{5}{6} u_j + \frac{2}{6} u_{j+1} \right) \\ &\quad + w_2^+ \left(\frac{2}{6} u_j + \frac{5}{6} u_{j+1} - \frac{1}{6} u_{j+2} \right). \end{aligned} \tag{3.1}$$

When the exact solution is smooth, it is well known that

$$w_0^+ = \frac{1}{10} + \epsilon_1, \quad w_1^+ = \frac{6}{10} + \epsilon_2, \quad w_2^+ = \frac{3}{10} + \epsilon_3, \tag{3.2}$$

where ϵ_1, ϵ_2, ϵ_3 are all $\mathcal{O}((\Delta x)^2)$, and $\epsilon_1+\epsilon_2+\epsilon_3 = 0$ [11]. We now prove the following theorem using von Neumann analysis.

THEOREM 3.1. *The combination of WENO5 and forward Euler is linearly unstable.*

Proof. Assuming periodic boundary conditions, we can expand the approximate solution $\{u_j\}_{j=0}^N$ as a finite Fourier series

$$u_j = \sum_{k=-\lfloor N/2 \rfloor}^{\lceil N/2 \rceil} \hat{u}_k e^{\iota j \xi_k \, \Delta x},$$

where ι is the imaginary unit, i.e., $\iota^2 = -1$, and ξ_k is the spatial frequency associated with $\hat{u}_k$. Because the wave equation is linear and has constant coefficients, it is sufficient to consider only one individual Fourier mode; i.e.,

$$u_j = \hat{u} e^{\iota j \xi \, \Delta x}.$$

Defining $\phi = \xi \, \Delta x$, we thus have

$$u_j = \hat{u} e^{\iota j \phi}. \tag{3.3}$$

By using (3.2), (3.1), and (3.3), we now obtain

$$\begin{aligned} \hat{f}^+_{j+\frac{1}{2}} = u_j \Bigg[& w_0^+ \left(\frac{2}{6} e^{-2\iota\phi} - \frac{7}{6} e^{-\iota\phi} + \frac{11}{6} \right) + \; w_1^+ \left(-\frac{1}{6} e^{-\iota\phi} + \frac{5}{6} + \frac{2}{6} e^{\iota\phi} \right) \\ & + w_2^+ \left(\frac{2}{6} + \frac{5}{6} e^{\iota\phi} - \frac{1}{6} e^{2\iota\phi} \right) \Bigg]. \end{aligned} \tag{3.4}$$

When the forward Euler method is used, the WENO5 method becomes

$$u_j^{n+1} = u_j^n - \frac{\Delta t}{\Delta x} L\left(u_{j-3}^n, u_{j-2}^n, \ldots, u_{j+2}^n\right), \tag{3.5}$$

where

$$L\left(u_{j-3}^n, u_{j-2}^n, \ldots, u_{j+2}^n\right) = \hat{f}^{+,n}_{j+\frac{1}{2}} - \hat{f}^{+,n}_{j-\frac{1}{2}}.$$

From (3.4) we see that $\frac{\hat{f}^{+,n}_{j+\frac{1}{2}} - \hat{f}^{+,n}_{j-\frac{1}{2}}}{u_j^n}$ is a function of ϕ. Thus we define $z(\phi) = \frac{\hat{f}^{+,n}_{j+\frac{1}{2}} - \hat{f}^{+,n}_{j-\frac{1}{2}}}{u_j^n}$ and obtain

$$L\left(u_{j-3}^n, u_{j-2}^n, \ldots, u_{j+2}^n\right) = z(\phi) u_j^n.$$

Defining the CFL number $\sigma = \frac{\Delta t}{\Delta x}$, (3.5) becomes

$$\begin{aligned} u_j^{n+1} &= u_j^n - \sigma z(\phi) u_j^n \\ &= u_j^n \left(1 - \sigma z(\phi)\right). \end{aligned} \tag{3.6}$$

Define the *amplification factor* $g(\sigma z(\phi)) = 1 - \sigma z(\phi)$. In order to prove the theorem, we need to show that for any $\sigma > 0$, there exists a ϕ such that $|g| > 1$. Let ϕ be a small positive number. Using the Taylor expansion of $\sin\phi$ and $\cos\phi$ in (2.3)–(2.5), we obtain

$$
\begin{aligned}
IS_0^+ &= u_j^2\phi^2\left(-1 + \frac{5}{12}\phi^2 - \frac{89}{40}\phi^4 + \frac{4889}{4032}\phi^6 - \frac{5}{3}\,\iota\phi^3 + \frac{17}{9}\,\iota\phi^5 + \mathcal{O}(\phi^7)\right),\\
IS_1^+ &= u_j^2\phi^2\left(-1 + \frac{17}{12}\phi^2 - \frac{9}{40}\phi^4 + \frac{337}{20160}\phi^6 + \mathcal{O}(\phi^7)\right),\\
IS_2^+ &= u_j^2\phi^2\left(-1 + \frac{5}{12}\phi^2 - \frac{89}{40}\phi^4 + \frac{4889}{4032}\phi^6 + \frac{5}{3}\,\iota\phi^3 - \frac{17}{9}\,\iota\phi^5 + \mathcal{O}(\phi^7)\right).
\end{aligned}
$$

For $0 < \epsilon \ll 1$, we can choose $\phi = \phi(\epsilon) = \mathcal{O}(\epsilon^{1/9})$ sufficiently small such that the α_i^+, $i = 0, 1, 2$, in (2.6) can be estimated as follows:

$$
\begin{aligned}
\alpha_0^+ &= \frac{1}{u_j^4\phi^4}\left(\frac{1}{10} + \frac{1}{12}\phi^2 - \frac{943}{2400}\phi^4 - \frac{7879}{4320}\phi^6 - \frac{1}{3}\,\iota\phi^3 - \frac{5}{9}\,\iota\phi^5 + \mathcal{O}(\phi^7)\right),\\
\alpha_1^+ &= \frac{1}{u_j^4\phi^4}\left(\frac{6}{10} + \frac{17}{10}\phi^2 + \frac{1337}{400}\phi^4 + \frac{19057}{3600}\phi^6 + \mathcal{O}(\phi^7)\right),\\
\alpha_2^+ &= \frac{1}{u_j^4\phi^4}\left(\frac{3}{10} + \frac{1}{4}\phi^2 - \frac{943}{800}\phi^4 - \frac{7879}{1440}\phi^6 + \,\iota\phi^3 + \frac{5}{3}\,\iota\phi^5 + \mathcal{O}(\phi^7)\right).
\end{aligned}
$$

Substituting the above expressions into (2.7), we obtain

$$
\begin{aligned}
w_0^+ &= \frac{1}{10} - \frac{3}{25}\phi^2 - \frac{163}{500}\phi^4 - \frac{30449}{30000}\phi^6 - \frac{2}{5}\,\iota\phi^3 + \frac{17}{75}\,\iota\phi^5 + \mathcal{O}(\phi^7),\\
w_1^+ &= \frac{6}{10} + \frac{12}{25}\phi^2 + \frac{163}{125}\phi^4 + \frac{20449}{7500}\phi^6 - \frac{2}{5}\,\iota\phi^3 - \frac{13}{75}\,\iota\phi^5 + \mathcal{O}(\phi^7),\\
w_2^+ &= \frac{3}{10} - \frac{9}{25}\phi^2 - \frac{489}{500}\phi^4 - \frac{51347}{30000}\phi^6 + \frac{4}{5}\,\iota\phi^3 - \frac{4}{75}\,\iota\phi^5 + \mathcal{O}(\phi^7).
\end{aligned}
$$

Thus the real and imaginary parts of $\hat{f}^{+,n}_{j+\frac{1}{2}}$ are

$$
\operatorname{Re}\hat{f}^{+,n}_{j+\frac{1}{2}} = u_j^n\left(1 - \frac{1}{12}\phi^2 - \frac{1}{720}\phi^4 + \frac{241}{21600}\phi^6 + \mathcal{O}(\phi^8)\right), \tag{3.7}
$$

$$
\operatorname{Im}\hat{f}^{+,n}_{j+\frac{1}{2}} = u_j^n\left(\frac{1}{2}\phi - \frac{7}{60}\phi^5 + \mathcal{O}(\phi^7)\right). \tag{3.8}
$$

Similarly it can be shown that

$$
\operatorname{Re}\hat{f}^{+,n}_{j-\frac{1}{2}} = u_j^n\left(1 - \frac{1}{12}\phi^2 - \frac{1}{720}\phi^4 - \frac{2279}{21600}\phi^6 + \mathcal{O}(\phi^8)\right), \tag{3.9}
$$

$$
\operatorname{Im}\hat{f}^{+,n}_{j-\frac{1}{2}} = u_j^n\left(-\frac{1}{2}\phi - \frac{7}{60}\phi^5 + \mathcal{O}(\phi^7)\right). \tag{3.10}
$$

Using (3.7)–(3.10), we obtain

$$
z(\phi) = \frac{7}{60}\phi^6 + \mathcal{O}(\phi^8) + \iota\left(\phi + \mathcal{O}(\phi^7)\right). \tag{3.11}
$$

The amplification factor becomes

$$g = 1 - \sigma\left(\mathcal{O}(\phi^6) + \iota\left(\phi + \mathcal{O}(\phi^7)\right)\right).$$

A simple calculation now shows that

$$\begin{aligned} |g|^2 &= \left(1 - \sigma\mathcal{O}(\phi^6)\right)^2 + \sigma^2\,\left(\phi + \mathcal{O}(\phi^7)\right)^2 \\ &= 1 + \sigma^2\,\phi^2 + \mathcal{O}(\phi^6) \\ &> 1 \qquad \forall\,\sigma > 0. \end{aligned}$$

This completes the proof. □

Remark 1. Equation (3.11) is valid only when ϕ is a small positive number. The general form of $z(\phi)$ for any ϕ is given later in (3.21).

Remark 2. This form of analysis applies to any linear finite difference method for $\frac{du_j}{dt} = -\frac{1}{\Delta x}L(u_{j-R}, \ldots, u_{j+S})$, where $L(u_{j-R}, \ldots, u_{j+S})$ can be written in the form

$$L(u_{j-R}, \ldots, u_{j+S}) = u_j z(\phi),$$

where $z(\phi)$ is uniquely determined by the spatial operator.

Remark 3. It should be noted that, although this instability argument applies to the WENO spatial discretization, it does not necessarily apply to the ENO spatial discretization. The reason is that this analysis needs to have a known stencil, and ENO methods may choose any of a number of candidate stencils even if the solution is smooth. Moreover, such a "randomly" chosen stencil may lead to an unstable method. This is why a biased choice for choosing ENO stencils is suggested in [19], and this strategy has been very successful for ENO methods in practice.

Remark 4. It follows immediately that, for the class of problems considered in this analysis, the combination of the forward Euler method and the WENO5 spatial discretization is not SSP for any step size $\Delta t > 0$. Because every ERK generates its second stage by a forward Euler step, this second stage cannot be SSP, and hence *in this framework, no ERK method can be SSP.*[1] Hence the SSP property offers no stability advantage.

Consider the general s-stage ERK method written in standard form:

$$\begin{aligned} y_n^{(1)} &= y_n, \\ y_n^{(k)} &= y_n + \Delta t \sum_{i=1}^{k-1} a_{k,i} f\left(t_n + c_i\Delta t, y_n^{(i)}\right), \quad k = 2, \ldots, s, \\ y_{n+1} &= y_n + \Delta t \sum_{i=1}^{s} b_i f\left(t_n + c_i\Delta t, y_n^{(i)}\right), \end{aligned}$$

where the c_k satisfy the conditions

$$c_k = a_{k1} + a_{k2} + \cdots + a_{k,k-1}$$

for $k = 1, \ldots, s$.

Its Butcher tableau is of the form given in Table 3.1.

[1] Recall that an ERK method written in Shu–Osher form is SSP if all of its stages i on step n of the numerical solution $Y_n^{(i)}$ satisfy $\|Y_n^{(i)}\| \le \|Y_n\|$, for all $i = 1, 2, \ldots, s$ and $n \ge 1$, for some suitable seminorm $\|\cdot\|$, where $Y_n^{(1)} = Y_n$ and $Y_n^{(s)} = Y_{n+1}$; see, e.g., [2, 8].

TABLE 3.1
Butcher tableau for s-stage ERK methods.

0						
c_2	a_{21}					
c_3	a_{31}	a_{32}				
c_4	a_{41}	a_{42}	a_{43}			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
c_s	a_{s1}	a_{s2}	a_{s3}	$\cdots$	$a_{s,s-1}$	
	b_1	b_2	b_3	$\cdots$	b_{s-1}	b_s

We have the following theorem and corollary [12].

THEOREM 3.2. *The amplification factor for an s-stage ERK method is*

$$g(\hat{z}) = 1 + \sum_{l=1}^{s} \left(\sum_{j=l}^{s} b_j \left(\sum_{j>m_1>\cdots>m_{l-1}\geq 1} a_{j,m_1} a_{m_1,m_2} \dots a_{m_{l-2},m_{l-1}} \right) \right) \hat{z}^l \tag{3.12}$$

$$= 1 + \hat{z} b^T (I - \hat{z}A)^{-1} e, \tag{3.13}$$

where I is the unit matrix, $A = (a_{ij})_{1\leq i,j\leq s}$ and $b = (b_1, b_2, \dots, b_s)$ are the coefficients of the Butcher tableau, $e = (1, 1, \dots, 1)$, $z(\phi)$ is determined by the spatial operator, and $\hat{z} = -\sigma z$.

Combining the order conditions for ERK methods with (3.12) or (3.13), we easily obtain the following corollary [14, 22].

COROLLARY 3.3. *The amplification factor of an s-stage, order-p ERK method is*

$$\begin{aligned} g(\hat{z}) &= 1 + \sum_{l=1}^{p} \frac{1}{l!} \hat{z}^l \\ &\quad + \sum_{l=p+1}^{s} \left(\sum_{j=l}^{s} b_j \left(\sum_{j>m_1>\cdots>m_{l-1}\geq 1} a_{j,m_1} a_{m_1,m_2} \dots a_{m_{l-2},m_{l-1}} \right) \right) \hat{z}^l \\ &= 1 + \sum_{l=1}^{p} \frac{\hat{z}^l}{l!} + \sum_{l=p+1}^{s} \hat{z}^l b^T A^{l-1} e. \end{aligned} \tag{3.14}$$

Therefore, a spatial discretization scheme combined with a given ERK method is linearly stable if and only if g in (3.12) satisfies $|g| \leq 1$ for all $\phi \in [0, 2\pi]$.

We can now prove the following theorem for any two-stage, second-order ERK method.

THEOREM 3.4. *The combination of WENO5 with any two-stage, second-order ERK method is linearly unstable.*

Proof. From (3.14), the amplification factor is given by

$$g(\hat{z}) = 1 - \sigma z + \frac{1}{2}(\sigma z)^2.$$

Choosing ϕ to be a small positive number and using (3.11), a simple calculation shows

$$\begin{aligned} |g|^2 &= \left(1 - \frac{1}{2}\sigma^2\phi^2 + \mathcal{O}(\phi^6)\right)^2 + \left(-\sigma\phi + \mathcal{O}(\phi^7)\right)^2 \\ &= 1 + \frac{1}{4}\sigma^4\phi^4 + \mathcal{O}(\phi^6) \\ &> 1. \end{aligned}$$

This finishes the proof. ∎

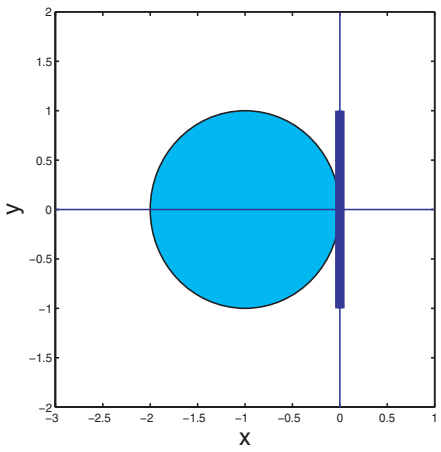


FIG. 1. *D_2 for the upwind (circle) and central (line segment) spatial discretizations.*

Before we derive the theorem for general ERK methods, we give a geometric interpretation for the CFL number σ [1, 10, 14, 22].

DEFINITION 3.5. *Let D_1 denote the classical (linear) stability domain of any ERK method, and let D_2 denote the region interior to the boundary $\{-z(\phi) : 0 \le \phi \le 2\pi\}$ in the complex domain. The CFL number σ is the largest nonnegative real number such that the scaled region σD_2 is contained in D_1.*

It is well known that (3.13) is the stability function for an ERK method; see, e.g., [5]. Thus $D_1 = \{\hat{z} : \ |g(\hat{z})| \le 1\}$. Note that the set $\{\hat{z} = -\sigma z(\phi) : 0 \le \phi \le 2\pi\}$ represents the boundary of the scaled region σD_2. It is clear that, in order to have $\hat{z} \in D_1$, the scaled region σD_2 must be contained in D_1. We now give two simple examples for the purposes of illustration. The first example is for the upwind spatial discretization; i.e., $du_j/dt = L(u_{j-1}, u_j) = -\frac{1}{\Delta x}(u_j - u_{j-1})$. Using von Neumann analysis, we obtain $z(\phi) = 1 - e^{-\iota\phi}$. It is easy to see that the set $\{-z(\phi) : 0 \le \phi \le 2\pi\}$ represents a circle in the complex plane with center $(-1, 0)$ and radius 1; i.e., D_2 is the shaded area shown in Figure 1.

If the forward Euler method is used for the time discretization, its classical linear stability domain D_1 is $\{z : |1 + z| \le 1\}$. In other words, D_1 is exactly the same as D_2 in this case. It is trivial to conclude therefore that $\sigma = 1$ is the largest number such that $\sigma D_2 \subseteq D_1$. Hence the CFL number is 1.

The second example is for the central finite difference spatial discretization; i.e., $du_j/dt = L(u_{j-1}, u_j, u_{j+1}) = -\frac{1}{\Delta x}\frac{u_{j+1} - u_{j-1}}{2}$. Using von Neumann analysis, we obtain $z(\phi) = \frac{1}{2}(e^{\iota\phi} - e^{-\iota\phi}) = \iota \sin\phi$. It is easy to see that the set $\{-z(\phi) : 0 \le \phi \le 2\pi\} = \{(0, y) : -1 \le y \le 1\}$; i.e., D_2 now represents a finite segment of the imaginary axis. If the forward Euler method is used for the time discretization, $\sigma D_2 \not\subseteq D_1$, no matter how small $\sigma > 0$ is chosen. Therefore, the central finite difference spatial discretization is linearly unstable when it is coupled with the forward Euler method.

We now derive the following lemma for any consistent ERK method. Lemma 3.6 is important for all of the theorems in this paper.

LEMMA 3.6. *The classical (linear) stability domain of any consistent ERK method contains a rectangle $[-\eta, 0] \times [-\iota\hat{\mu}, \iota\hat{\mu}]$ for some η, $\hat{\mu} > 0$ if and only if it has an intersection with the imaginary axis of the form $[-\iota\mu, \iota\mu]$ for some $\mu \ge \hat{\mu} > 0$.*

Proof. $\Longrightarrow$ If the rectangle $[-\eta, 0] \times [-\hat{\mu}, \hat{\mu}]$ is inside the classical (linear) stability domain of any consistent ERK method, by definition the part of the imaginary axis

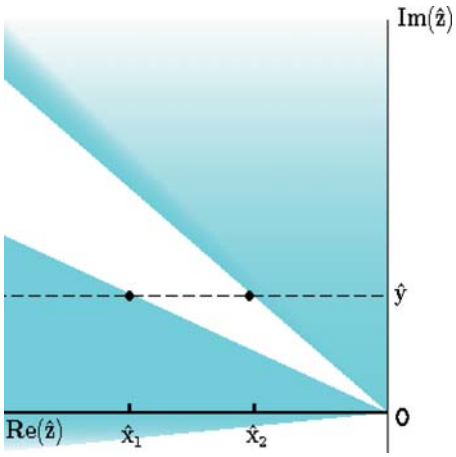


FIG. 2. *A schematic of the topology of the classical stability domain in a sufficiently small neighborhood of the origin if a rectangle* $[-\nu, 0] \times [-\hat{\mu}, \hat{\mu}]$ *is not contained in it. Shaded areas are inside in the stability domain.*

$[-\iota\hat{\mu}, \iota\hat{\mu}]$ is also inside the stability domain. In other words, the stability domain of the ERK method intersects the imaginary axis at $[-\iota\mu, \iota\mu]$ for some $\mu \geq \hat{\mu} > 0$.

$\Longleftarrow$ Assume that the stability domain of the ERK method intersects the imaginary axis at $[-\iota\mu, \iota\mu]$ for some $\mu > 0$. We first prove that the stability domain of the ERK method intersects the part of the (negative) real axis at $[-\hat{\eta}, 0]$ for some $\hat{\eta} > 0$. Recall that the stability function of any consistent ERK method is of the form

$$g(\hat{z}) = 1 + \hat{z} + \text{higher-order terms}.$$

Let $\hat{z} = -\gamma$, where γ is a small, positive real number. In other words, choose $\hat{z}$ to be close to the origin and on the negative real axis. It is easy to see $g(-\gamma) = 1-\gamma+\mathcal{O}(\gamma^2)$. It is obvious that $0 < |g(-\gamma)| < 1$ as $\gamma \to 0+$. That is, there is an intersection of the stability domain with the negative real axis. Assume the intersection is of the form $[-\hat{\eta}, 0]$ for some real $\hat{\eta} > 0$.

Now by contradiction assume that no rectangle $[-\eta, 0]\times[-\hat{\mu}, \hat{\mu}]$ is contained inside the stability domain. Using the facts that the stability domain intersects the negative real axis at $[-\hat{\eta}, 0]$ and that it intersects the imaginary axis at $[-\iota\mu, \iota\mu]$, we give a schematic representation of the topology of a sufficiently small neighborhood of the origin in Figure 2. In the figure, areas inside the stability domain are shaded. For simplicity we focus on the second quadrant in the complex plane, and we show only one region not contained in the stability domain. Thus, if we define $\hat{z} = \hat{x} + \iota\hat{y}$, where $\hat{x}$, $\hat{y}$ are sufficiently small real numbers, there are at least two numbers $\hat{x}_1$ and $\hat{x}_2$ for each $\hat{y}$ such that $|g(\hat{x}_1 + \iota\hat{y})| = |g(\hat{x}_2 + \iota\hat{y})| = 1$. Therefore, the equation $R(\hat{x}, \hat{y}) = |g(\hat{x} + \iota\hat{y})|^2 - 1 = 0$ must have more than one solution $\hat{x} = \hat{x}(\hat{y})$ in the neighborhood of $(0, 0)$.

We now obtain a contradiction by using the implicit function theorem to prove that, in fact, $\hat{x}(\hat{y})$ is a unique solution to $R(\hat{x}, \hat{y}) = 0$ in a small neighborhood of the origin. It is easy to see that

$$g(\hat{x} + \iota\hat{y}) = 1 + \hat{x} + \iota\hat{y} + Q_1(\hat{x}, \hat{y}) + \iota Q_2(\hat{x}, \hat{y}), \tag{3.15}$$

where $Q_1(\hat{x},\hat{y})$ and $Q_2(\hat{x},\hat{y})$ are two real polynomials of the form

$$Q_1(\hat{x},\hat{y}) = \sum_{\substack{l,k\geq 0,\\ l+k\geq 2}} \xi_{l,k}\hat{x}^l\hat{y}^k, \tag{3.16}$$

$$Q_2(\hat{x},\hat{y}) = \sum_{\substack{l,k\geq 0,\\ l+k\geq 2}} \zeta_{l,k}\hat{x}^l\hat{y}^k, \tag{3.17}$$

with real coefficients $\xi_{l,k}$, $\zeta_{l,k}$. Using (3.15)–(3.17), we obtain

$$R(\hat{x},\hat{y}) = (1+\hat{x}+Q_1(\hat{x},\hat{y}))^2 + (\hat{y}+Q_2(\hat{x},\hat{y}))^2 - 1.$$

A simple calculation shows that

$$\frac{\partial R}{\partial \hat{x}}(0,0) = 2 \neq 0.$$

From the implicit function theorem, we know that there is a unique solution $\hat{x} = \hat{x}(\hat{y})$ to $R(\hat{x},\hat{y}) = 0$ in a small neighborhood of the origin, yielding the desired contradiction. ∎

Remark 5. The result of Lemma 3.6 can be generalized to any consistent one-step method with very little modification of the proof.

Remark 6. Because of this result, the prospect that the spectra of the spatially discretized system may contain negative real components is not problematic.

Remark 7. This result is the equivalent of the result of [13] for *local stability*. In particular, the regions of stability described for ERK methods are equivalent.

Using Definition 3.5 and Lemma 3.6, we have the following theorem for first-order ERK methods.

THEOREM 3.7. *There exists a CFL number σ such that the combination of WENO5 with a first-order ERK method is linearly stable for $\Delta t/\Delta x \leq \sigma$ provided the first-order ERK method satisfies*

$$\sum_{1\leq i<j\leq s} b_j a_{ji} > \frac{1}{2}; \tag{3.18}$$

on the other hand, this combination is linearly unstable if

$$\sum_{1\leq i<j\leq s} b_j a_{ji} < \frac{1}{2}.$$

Note 1. When $\sum_{1\leq i<j\leq s} b_j a_{ji} = \frac{1}{2}$, the ERK method is second order. The corresponding results are given in Theorem 3.9.

Proof. We first prove the linearly unstable case. Let $\tau_2 = \sum_{1\leq i<j\leq s} b_j a_{ji}$; i.e., τ_2 is the coefficient of $\hat{z}^2$ in (3.12). For $\phi > 0$ sufficiently small, from (3.11) it is easy to show that

$$\begin{aligned} z^l &= \iota^l\phi^l + \mathcal{O}(\phi^6) + \iota\mathcal{O}(\phi^6) \qquad \text{if } 2\leq l<6,\\ z^l &= \mathcal{O}(\phi^6) \qquad \text{if } l\geq 6. \end{aligned}$$

From (3.12), we obtain

$$\begin{aligned} g(-\sigma z(\phi)) &= 1-\sigma z+\tau_2(\sigma z)^2+\mathcal{O}(z^3)\\ &= \left(1-\tau_2\sigma^2\phi^2+\mathcal{O}(\phi^4)\right) - \iota\left(\sigma\phi+\mathcal{O}(\phi^3)\right). \end{aligned}$$

A simple calculation shows that

$$\begin{aligned}|g|^2 &= \left(1-\tau_2\sigma^2\phi^2+\mathcal{O}(\phi^4)\right)^2+\left(\sigma\phi+\mathcal{O}(\phi^3)\right)^2\\ &= 1+(1-2\tau_2)\sigma^2\phi^2+\mathcal{O}(\phi^4)\\ &> 1,\end{aligned}$$

if $\tau_2 < \frac{1}{2}$. Therefore, the combined method is linearly unstable.

Now let us assume $\tau_2 > \frac{1}{2}$. We first show that the stability domain D_1 of the corresponding ERK intersects the imaginary axis. Later (in Theorem 3.14) we show that, in fact, this is a sufficient condition for linear stability of any ERK method when combined with WENO5. Recall that the stability function of the ERK method is of the form

$$g(\hat{z}) = 1+\hat{z}+\tau_2\,\hat{z}^2+\cdots.$$

Let $\hat{z}=\iota\gamma$, where γ is a small real number. In other words, choose $\hat{z}$ to be close to the origin and on the imaginary axis. It is easy to see

$$\begin{aligned}|g(\iota\gamma)|^2 &= \left(1-\tau_2\gamma^2+\mathcal{O}(\gamma^4)\right)^2+\left(\gamma+\mathcal{O}(\gamma^3)\right)^2\\ &= \left(1-2\tau_2\gamma^2+\mathcal{O}(\gamma^4)\right)+\left(\gamma^2+\mathcal{O}(\gamma^4)\right)\\ &= 1+(1-2\tau_2)\gamma^2+\mathcal{O}(\gamma^4).\end{aligned}$$

Using the condition $\tau_2 > \frac{1}{2}$, we obtain that $|g(\iota\gamma)| < 1$ as $\gamma \to 0+$. That is, there is an intersection of D_1 with the imaginary axis. Assume the intersection is the interval $[-\iota\mu, \iota\mu]$ for some real $\mu > 0$. Then from Lemma 3.6 there exists a rectangle $D_3 = [-\eta, 0]\times[-\hat{\mu},\hat{\mu}] \subseteq D_1$ for some $\eta > 0$, $0 < \hat{\mu} \le \mu$.

It is easy to derive the expression for $\hat{f}^+_{j+\frac{1}{2}}$ from (3.4) and (3.2):

$$\begin{aligned}\hat{f}^+_{j+\frac{1}{2}} = u_j\Bigg[&\left(\frac{2}{60}e^{-2\iota\phi}-\frac{13}{60}e^{-\iota\phi}+\frac{47}{60}+\frac{27}{60}e^{\iota\phi}-\frac{3}{60}e^{2\iota\phi}\right)\\ &+\epsilon_1\left(\frac{2}{6}e^{-2\iota\phi}-\frac{7}{6}e^{-\iota\phi}+\frac{11}{6}\right)\\ &+\epsilon_2\left(-\frac{1}{6}e^{-\iota\phi}+\frac{5}{6}+\frac{2}{6}e^{\iota\phi}\right)+\epsilon_3\left(\frac{2}{6}+\frac{5}{6}e^{\iota\phi}-\frac{1}{6}e^{2\iota\phi}\right)\Bigg].\end{aligned} \tag{3.19}$$

Similarly we obtain the expression for $\hat{f}^+_{j-\frac{1}{2}}$:

$$\begin{aligned}\hat{f}^+_{j-\frac{1}{2}} = u_j\Bigg[&\left(\frac{2}{60}e^{-3\iota\phi}-\frac{13}{60}e^{-2\iota\phi}+\frac{47}{60}e^{-\iota\phi}+\frac{27}{60}-\frac{3}{60}e^{\iota\phi}\right)\\ &+\epsilon_4\left(\frac{2}{6}e^{-3\iota\phi}-\frac{7}{6}e^{-2\iota\phi}+\frac{11}{6}e^{-\iota\phi}\right)\\ &+\epsilon_5\left(-\frac{1}{6}e^{-2\iota\phi}+\frac{5}{6}e^{-\iota\phi}+\frac{2}{6}\right)+\epsilon_6\left(\frac{2}{6}e^{-\iota\phi}+\frac{5}{6}-\frac{1}{6}e^{\iota\phi}\right)\Bigg],\end{aligned} \tag{3.20}$$

where ϵ_4, ϵ_5, and ϵ_6 are all $\mathcal{O}((\Delta x)^2)$, and $\epsilon_4+\epsilon_5+\epsilon_6 = 0$.

Using (3.19), (3.20), and the definition $z(\phi) = \frac{\hat{f}^{+,n}_{j+\frac{1}{2}}-\hat{f}^{+,n}_{j-\frac{1}{2}}}{u^n_j}$, we obtain

$$z(\phi) = \tilde{z} + M(\epsilon_1,\epsilon_2,\ldots,\epsilon_6,\phi), \tag{3.21}$$

where

$$\tilde{z} = -\frac{1}{30}e^{-3\iota\phi} + \frac{1}{4}e^{-2\iota\phi} - e^{-\iota\phi} + \frac{1}{3} + \frac{1}{2}e^{\iota\phi} - \frac{1}{20}e^{2\iota\phi}, \tag{3.22}$$

and

$$\begin{aligned} M(\epsilon_1,\epsilon_2,\ldots,\epsilon_6,\phi) &= \epsilon_1\left(\frac{2}{6}e^{-2\iota\phi} - \frac{7}{6}e^{-\iota\phi} + \frac{11}{6}\right) + \epsilon_2\left(-\frac{1}{6}e^{-\iota\phi} + \frac{5}{6} + \frac{2}{6}e^{\iota\phi}\right) \\ &\quad + \epsilon_3\left(\frac{2}{6} + \frac{5}{6}e^{\iota\phi} - \frac{1}{6}e^{2\iota\phi}\right) - \epsilon_4\left(\frac{2}{6}e^{-3\iota\phi} - \frac{7}{6}e^{-2\iota\phi} + \frac{11}{6}e^{-\iota\phi}\right) \\ &\quad - \epsilon_5\left(-\frac{1}{6}e^{-2\iota\phi} + \frac{5}{6}e^{-\iota\phi} + \frac{2}{6}\right) - \epsilon_6\left(\frac{2}{6}e^{-\iota\phi} + \frac{5}{6} - \frac{1}{6}e^{\iota\phi}\right). \end{aligned} \tag{3.23}$$

We note that M is made up of two pairs of three terms, corresponding to each of the ENO stencils associated with each of the flux terms. We now bound each of the terms that comprise M. Because $\operatorname{Re}\left(\frac{2}{6}e^{-2\iota\phi} - \frac{7}{6}e^{-\iota\phi} + \frac{11}{6}\right) = \frac{2}{6}\cos 2\phi - \frac{7}{6}\cos\phi + \frac{11}{6}$, and $-1 \le \cos\phi \le 1$ for all ϕ, we can write

$$\frac{1}{3} \le \operatorname{Re}\left(\frac{2}{6}e^{-2\iota\phi} - \frac{7}{6}e^{-\iota\phi} + \frac{11}{6}\right) \le \frac{10}{3}.$$

(In fact, the lower bound can be tightened to 95/96, but the proof is not sensitive to this value.)

Thus

$$\operatorname{Re}\left|\epsilon_1\left(\frac{2}{6}e^{-2\iota\phi} - \frac{7}{6}e^{-\iota\phi} + \frac{11}{6}\right)\right| \le \frac{10}{3}\left|\epsilon_1\right|.$$

Similarly we can bound the remaining terms of $M(\epsilon_1,\epsilon_2,\ldots,\epsilon_6,\phi)$. Finally we can obtain an expression of the form

$$\left|\operatorname{Re} M(\epsilon_1,\epsilon_2,\ldots,\epsilon_6,\phi)\right| \le \Gamma_1 \max_{1\le m\le 6}\left|\epsilon_m\right|, \tag{3.24}$$

where Γ_1 is a positive constant that is determined by the stencils. Applying the same analysis, we can write

$$\left|\operatorname{Im} M(\epsilon_1,\epsilon_2,\ldots,\epsilon_6,\phi)\right| \le \Gamma_2 \max_{1\le m\le 6}\left|\epsilon_m\right|, \tag{3.25}$$

where Γ_2 is a positive constant that is determined by the stencils.

We now examine the real and imaginary parts of $\tilde{z}$:

$$\begin{aligned} \operatorname{Re}\tilde{z} &= -\frac{1}{30}\cos 3\phi + \frac{1}{5}\cos 2\phi - \frac{1}{2}\cos\phi + \frac{1}{3} \\ &= -\frac{1}{30}\left(4\cos^3\phi - 3\cos\phi\right) + \frac{1}{5}\left(2\cos^2\phi - 1\right) - \frac{1}{2}\cos\phi + \frac{1}{3} \\ &= \frac{2}{15}(1-\cos\phi)^3; \end{aligned} \tag{3.26}$$

$$\operatorname{Im}\ \tilde{z} = \frac{1}{30}\sin 3\phi - \frac{3}{10}\sin 2\phi + \frac{3}{2}\sin\phi. \tag{3.27}$$

Let $D_4 = [-\frac{31}{15}, 0] \times [-\frac{17}{6}, \frac{17}{6}]$, and let $\sigma_0 > 0$ be such that the rectangle $\sigma_0 D_4 \subseteq D_3 = [-\eta, 0] \times [-\hat{\mu}, \hat{\mu}]$ defined previously. We now use D_3 and D_4 to prove that

$\sigma_0 D_2 \subseteq D_1$; i.e., the combination of an s-stage, first-order ERK method where $\tau_2 > \frac{1}{2}$ with WENO5 is linearly stable if $\frac{\Delta t}{\Delta x} \leq \sigma_0$. (Note that σ_0 may not be the same as the CFL number σ; in fact, we know only that $0 < \sigma_0 \leq \sigma$. However, this proves the existence of $\sigma > 0$.)

Using a similar analysis to the first part of this theorem, we conclude that, given σ_0, we can choose ϕ small enough such that $|g| < 1$. In other words, $\exists\, \alpha > 0$, such that the scaled domain $\{-\sigma_0 z(\phi) : 0 \leq \phi \leq \alpha \text{ or } 2\pi - \alpha \leq \phi \leq 2\pi\} \subseteq D_1$. We now complete the proof by showing that $D_5 = \{-\sigma_0 z(\phi) : \alpha \leq \phi \leq 2\pi - \alpha\} \subseteq D_1$.

Using (3.24) and (3.26), we obtain

$$\frac{2}{15}(1-\cos\phi)^3 - \Gamma_1 \max_{1\leq m\leq 6} |\epsilon_m| \leq \operatorname{Re} z(\phi) \leq \frac{2}{15}(1-\cos\phi)^3 + \Gamma_1 \max_{1\leq m\leq 6} |\epsilon_m|.$$

Because $\alpha \leq \phi \leq 2\pi - \alpha$, we see

$$\frac{2}{15}(1-\cos\alpha)^3 - \Gamma_1 \max_{1\leq m\leq 6} |\epsilon_m| \leq \operatorname{Re} z(\phi) \leq \frac{16}{15} + \Gamma_1 \max_{1\leq m\leq 6} |\epsilon_m|.$$

Note the ϵ_m, $m = 1, 2, \ldots, 6$, are $\mathcal{O}((\Delta x)^2)$, and Γ_1 is a constant. We can choose Δx small enough such that $\Gamma_1 \max_{1\leq m\leq 6} |\epsilon_m| \leq \min\left(\frac{2}{15}(1-\cos\alpha)^3, 1\right)$. Therefore,

$$0 \leq \operatorname{Re} z(\phi) \leq \frac{16}{15} + 1 = \frac{31}{15}. \tag{3.28}$$

From (3.27), using the fact that $-1 \leq \sin\phi \leq 1$ for all ϕ, we see that $-\frac{11}{6} \leq \operatorname{Im} \tilde{z} \leq \frac{11}{6}$. Again, we can choose Δx small enough such that $\Gamma_2 \max_{1\leq m\leq 6} |\epsilon_m| \leq 1$. Using (3.25) and (3.27), we have

$$-\frac{17}{6} = -\frac{11}{6} - 1 \leq \operatorname{Im} z(\phi) \leq \frac{11}{6} + 1 = \frac{17}{6}. \tag{3.29}$$

From (3.28) and (3.29), we conclude that $D_5 \subseteq D_4$. Because $\sigma_0 D_4 \subseteq D_3 = [-\eta, 0] \times [-\hat{\mu}, \hat{\mu}]$, and $D_3 \subseteq D_1$, we conclude that $\sigma_0 D_5 \subseteq D_1$.

This completes the proof. ∎

COROLLARY 3.8. *The combination of WENO5 and any optimal, s-stage, first-order SSP ERK method as in* [21] *is linearly unstable.*

Proof. The Butcher tableau of the optimal, s-stage, first-order SSP ERK method is of the form shown in Table 3.2.

The corresponding stability function is

$$\left(1 + \frac{1}{s}\hat{z}\right)^s = 1 + \hat{z} + \frac{s-1}{2s}\hat{z}^2 + \cdots.$$

Because $\frac{s-1}{2s} < \frac{1}{2}$ for all $s \geq 1$, linear instability follows from Theorem 3.7. ∎

TABLE 3.2
Butcher tableau for optimal s-stage, order-1 SSP ERK methods.

0						
$\frac{1}{s}$	$\frac{1}{s}$					
$\frac{2}{s}$	$\frac{1}{s}$	$\frac{1}{s}$				
$\frac{3}{s}$	$\frac{1}{s}$	$\frac{1}{s}$	$\frac{1}{s}$			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
$\frac{s-1}{s}$	$\frac{1}{s}$	$\frac{1}{s}$	$\frac{1}{s}$	$\cdots$	$\frac{1}{s}$	
	$\frac{1}{s}$	$\frac{1}{s}$	$\frac{1}{s}$	$\cdots$	$\frac{1}{s}$	$\frac{1}{s}$

Using the same analysis of Theorem 3.7, we obtain the following theorem for s-stage, second-order order ERK methods.

THEOREM 3.9. *Assume the ERK method is at least second order. Its stability function is of the form*

$$1 + \hat{z} + \frac{1}{2}\hat{z}^2 + \tau_3\hat{z}^3 + \tau_4\hat{z}^4 + \tau_5\hat{z}^5 + \tau_6\hat{z}^6 + \cdots . \tag{3.30}$$

Then there exists a CFL number σ such that the combination of this ERK method and WENO5 is linearly stable for $\Delta t/\Delta x \le \sigma$ if the ERK method satisfies

$$\tau_3 - \tau_4 > \frac{1}{8}; \tag{3.31}$$

on the other hand, the combination is linearly unstable if

$$\tau_3 - \tau_4 < \frac{1}{8}.$$

Proof. First we note that (3.30) is the amplification factor g if $\hat{z} = -\sigma z$. Now choosing ϕ to be a small positive number, we use (3.11) and have

$$\begin{aligned}
z &= \left(\frac{7}{60}\phi^6 + \mathcal{O}(\phi^8)\right) + \iota\ \left(\phi + \mathcal{O}(\phi^7)\right),\\
z^2 &= \left(-\phi^2 + \mathcal{O}(\phi^8)\right) + \iota\ \mathcal{O}(\phi^7),\\
z^3 &= \mathcal{O}(\phi^8) + \iota\ \left(-\phi^3 + \mathcal{O}(\phi^9)\right),\\
z^4 &= \left(\phi^4 + \mathcal{O}(\phi^{10})\right) + \iota\ \mathcal{O}(\phi^9),\\
z^5 &= \mathcal{O}(\phi^{10}) + \iota\ \left(\phi^5 + \mathcal{O}(\phi^{11})\right),\\
z^6 &= \left(-\phi^6 + \mathcal{O}(\phi^{12})\right) + \iota\ \mathcal{O}(\phi^{11}),
\end{aligned}$$

and $z^l = \mathcal{O}(\phi^7)$, $l \ge 7$. We now calculate $|g|^2$:

$$\begin{aligned}
|g|^2 &= \left(1 - \frac{1}{2}\sigma^2\phi^2 + \tau_4\sigma^4\phi^4 - \frac{7}{60}\sigma\phi^6 - \tau_6\sigma^6\phi^6 + \mathcal{O}(\phi^8)\right)^2\\
&\quad + \left(-\sigma\phi + \tau_3\sigma^3\phi^3 - \tau_5\sigma^5\phi^5 + \mathcal{O}(\phi^7)\right)^2\\
&= \left(1 - \sigma^2\phi^2 + \left(\frac{1}{4} + 2\tau_4\right)\sigma^4\phi^4 - \frac{7}{30}\sigma\phi^6 - \left(\tau_4 + 2\tau_6\right)\sigma^6\phi^6 + \mathcal{O}(\phi^8)\right)\\
&\quad + \left(\sigma^2\phi^2 - 2\tau_3\sigma^4\phi^4 + \left(\tau_3^2 + 2\tau_5\right)\sigma^6\phi^6 + \mathcal{O}(\phi^8)\right)\\
&= 1 + \left(\frac{1}{4} + 2\tau_4 - 2\tau_3\right)\sigma^4\phi^4 + \left(-\frac{7}{30}\sigma + \left(\tau_3^2 + 2\tau_5 - \tau_4 - 2\tau_6\right)\sigma^6\right)\phi^6 + \mathcal{O}(\phi^8).
\end{aligned} \tag{3.32}$$

If $(\frac{1}{4} + 2\tau_4 - 2\tau_3) > 0$, i.e., $\tau_3 - \tau_4 < \frac{1}{8}$, we have $|g| > 1$ as $\phi \to 0+$. Therefore, in this case the combination is linearly unstable.

However, if $(\frac{1}{4} + 2\tau_4 - 2\tau_3) < 0$, i.e., $\tau_3 - \tau_4 > \frac{1}{8}$, we have $|g| < 1$ as $\phi \to 0^+$. Moreover, there is an intersection between the stability domain of the ERK method and the imaginary axis. The rest of the proof of stability is similar to that of Theorem 3.7. ∎

COROLLARY 3.10. *The combination of WENO5 with any optimal s-stage, second-order SSP ERK method as in* [21] *is linearly unstable.*

TABLE 3.3
Butcher tableau for optimal s-stage, second-order SSP ERK methods.

0						
$\frac{1}{s-1}$	$\frac{1}{s-1}$					
$\frac{2}{s-1}$	$\frac{1}{s-1}$	$\frac{1}{s-1}$				
$\frac{3}{s-1}$	$\frac{1}{s-1}$	$\frac{1}{s-1}$	$\frac{1}{s-1}$			
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$		
$\frac{s-1}{s-1}$	$\frac{1}{s-1}$	$\frac{1}{s-1}$	$\frac{1}{s-1}$	$\cdots$	$\frac{1}{s-1}$	
	$\frac{1}{s}$	$\frac{1}{s}$	$\frac{1}{s}$	$\cdots$	$\frac{1}{s}$	$\frac{1}{s}$

Proof. The Butcher tableau of the optimal s-stage, second-order SSP ERK method is of the form given in Table 3.3.

The corresponding stability function is

$$\frac{1}{s}+\frac{s-1}{s}\left(1+\frac{1}{s-1}\hat{z}\right)^s=1+\hat{z}+\frac{1}{2}\hat{z}^2+\frac{s-2}{6(s-1)}\hat{z}^3+\frac{(s-2)(s-3)}{24(s-1)^2}\hat{z}^4+\cdots.$$

Because $\tau_3-\tau_4=\frac{s(s-2)}{(s-1)^2}(\frac{1}{8}-\frac{1}{24s})<\frac{1}{8}$, the linear instability follows from Theorem 3.9. ∎

Note that the stability function of any three-stage, third-order ERK method is $1+\hat{z}+\frac{1}{2}\hat{z}^2+\frac{1}{6}\hat{z}^3$. Because $\tau_3-\tau_4=\frac{1}{6}>\frac{1}{8}$, *the combination of WENO5 and any three-stage, third-order ERK method is linearly stable.* Furthermore, we have the following theorem for s-stage, third-order ERK methods.

THEOREM 3.11. *Assume the ERK method is third order. Its stability function is of the form*

$$1+\hat{z}+\frac{1}{2}\hat{z}^2+\frac{1}{6}\hat{z}^3+\tau_4\hat{z}^4+\cdots.$$

Then there exists a CFL number σ such that the combination of WENO5 and the ERK method is linearly stable for $\Delta t/\Delta x\le\sigma$ if the ERK method satisfies

$$\tau_4<\frac{1}{24}; \tag{3.33}$$

on the other hand, the combination is linearly unstable if

$$\tau_4>\frac{1}{24}.$$

When $\tau_4=\frac{1}{24}$, the ERK method has the same linear stability as the fourth-order ERK methods. The corresponding results are given in Theorem 3.12.

Proof. The proof is similar to that of Theorem 3.9. ∎

Note that, if the ERK method is at least fourth order, we have $\tau_3-\tau_4=\frac{1}{8}$. From (3.32), we obtain the following.

THEOREM 3.12. *Assume that the ERK method is at least order* 4. *Its stability function is of the form*

$$1+\hat{z}+\frac{1}{2}\hat{z}^2+\frac{1}{6}\hat{z}^3+\frac{1}{24}\hat{z}^4+\tau_5\hat{z}^5+\tau_6\hat{z}^6+\cdots.$$

Then there exists a CFL number σ such that the combination of the ERK method and WENO5 is linearly stable for $\Delta t/\Delta x\le\sigma$ if the ERK method satisfies

$$\tau_5-\tau_6<\frac{1}{144}.$$

Proof. From (3.32) we conclude that $-\frac{7}{30}\sigma + \left(\tau_3^2 + 2\tau_5 - \tau_4 - 2\tau_6\right)\sigma^6 < 0$ for σ sufficiently small. Therefore, $|g|^2 < 1$ for ϕ sufficiently small and positive. In other words, $\exists\, \sigma_0, \alpha > 0$, such that the scaled domain $\{-\sigma_0 z(\phi) : 0 \le \phi \le \alpha \text{ or } 2\pi - \alpha \le \phi \le 2\pi\}$ is inside the stability domain D_1 of the ERK method.

The stability function of the ERK method is now of the form

$$g(\hat{z}) = 1 + \hat{z} + \frac{1}{2}\hat{z}^2 + \frac{1}{6}\hat{z}^3 + \frac{1}{24}\hat{z}^4 + \tau_5\hat{z}^5 + \tau_6\hat{z}^6 + \cdots.$$

Let $\hat{z} = \iota\gamma$, where γ is a small real number. In other words, choose $\hat{z}$ to be close to the origin and on the imaginary axis. It is easy to see

$$|g|^2 = 1 + 2\left(\tau_5 - \tau_6 - \frac{1}{144}\right)\gamma^6 + \mathcal{O}(\gamma^8).$$

Using the condition $\tau_5 - \tau_6 < \frac{1}{144}$, we obtain that $|g(\iota\gamma)| < 1$ as $\gamma \to 0+$. In other words, there is an intersection between the stability domain D_1 of the ERK method and the imaginary axis if $\tau_5 - \tau_6 < \frac{1}{144}$. The rest of the proof is similar to the proof of linear stability for second-order ERK methods from Theorem 3.9. ∎

From Theorem 3.12, we see that $\tau_5 = \tau_6 = 0$ for the classical four-stage, fourth-order ERK method. Therefore, its combination with WENO5 is linearly stable.

Remark 8. There is no corresponding result for linear instability as in Theorem 3.9 if the ERK method is fourth order. This is because, regardless of the values of τ_5 and τ_6, $|g|^2 < 1$ whenever ϕ is a small positive number, and σ is sufficiently small. On the other hand, if the stability domain D_1 of the ERK method does not intersect the imaginary axis (i.e., $\tau_5 - \tau_6 > \frac{1}{144}$), linear instability cannot be proved as in Theorem 3.9.

Following immediately from Theorem 3.12 with $\tau_5 = \frac{1}{120}$, we have the following theorem for s-stage, fifth-order ERK methods.

THEOREM 3.13. *Assume that the ERK method is at least order* 5. *Its stability function is of the form*

$$1 + \hat{z} + \frac{1}{2}\hat{z}^2 + \frac{1}{6}\hat{z}^3 + \frac{1}{24}\hat{z}^4 + \frac{1}{120}\hat{z}^5 + \tau_6\hat{z}^6 + \cdots.$$

*Then there exists a CFL number σ such that the combination of the ERK method and WENO*5 *is linearly stable for $\Delta t/\Delta x \le \sigma$ if the ERK method satisfies*

$$\tau_6 > \frac{1}{720}.$$

Finally we have the more general result for s-stage, order-$p \ge 4$ ERK methods as follows.

THEOREM 3.14. *Assume that the ERK method is at least order $p \ge 4$. Its stability function is of the form*

$$1 + \hat{z} + \frac{1}{2}\hat{z}^2 + \frac{1}{6}\hat{z}^3 + \frac{1}{24}\hat{z}^4 + \cdots.$$

*Then there exists a CFL number σ such that the combination of the ERK method and WENO*5 *is linearly stable for $\Delta t/\Delta x \le \sigma$ if the stability domain of the ERK method includes the part of the imaginary axis of the form $[-\iota\mu, \iota\mu]$ for some $\mu > 0$.*

The proof is similar that of Theorem 3.12.

Remark 9. We note that the presence of an intersection of the stability domain of an ERK method with the imaginary axis of the form $[-\iota\mu, \iota\mu]$ for some $\mu > 0$ has been used to prove stability of s-stage methods of orders 1, 2, and 3 already. Combining this with the result in Theorem 3.14 allows us to conclude that the intersection of the stability domain with the imaginary axis in the form $[-\iota\mu, \iota\mu]$ for some $\mu > 0$ is a sufficient condition for linear stability of any ERK method when coupled with WENO5.

4. Numerical results. In this section, we study two classical scalar hyperbolic conservation laws: the (linear) advection equation and the (nonlinear) inviscid Burgers equation. In both cases, the problems are posed in one dimension, and WENO5 is employed as the spatial discretization. We use a uniform mesh with N spatial subintervals. Both SSP and non-SSP ERK time integration methods are considered. We illustrate the linear instability of some well-known first- and second-order SSP ERK methods by plotting the solution at a given time T_{out} with a specified Courant number $\sigma = \frac{\Delta t}{\Delta x}$. Extensive numerical tests have shown that smaller values of σ require larger values of T_{out} for the effect of the instability to clearly manifest itself.

4.1. ERK methods. In order to illustrate our theory, we consider the following four well-known ERK methods. Under appropriate assumptions, these methods can be SSP. The first three are linearly unstable when coupled with WENO5 and are used to solve hyperbolic conservation laws; the fourth is arguably the most widely used time integration method used with WENO spatial discretizations.

(1) The forward Euler (FE) method.

(2) The optimal two-stage, second-order SSP ERK method (which we call SSP(2,2)) with Butcher tableau (cf. Table 3.1)

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \end{array}.$$

(3) The optimal three-stage, second-order SSP ERK method [21] (which we call SSP(3,2)) with Butcher tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}.$$

(4) The (optimal) three-stage, third-order SSP ERK method, SSP(3,3) [3, 21], with Butcher tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \end{array}.$$

We also consider four ERK methods that are linearly stable according to our analysis when coupled with WENO5. These methods are provably not SSP for any time step $\Delta t > 0$. Stated differently, these methods have a radius of contractivity of 0 (see, e.g., [12]); this fact is obvious because each method has a 0 in its $\mathbf{b}$ vector (or *quadrature weights*) of the Butcher tableau.

(5) A two-stage, order-1 non-SSP ERK method (which we call NSSP(2,1)) with Butcher tableau

$$\begin{array}{c|cc} 0 & 0 & 0 \\ \frac{3}{4} & \frac{3}{4} & 0 \\ \hline & 0 & 1 \end{array}.$$

Because $b_2 a_{21} = \frac{3}{4}$, it is a stable ERK method for WENO5 according to Theorem 3.7. Its CFL number can be directly estimated to be $\sigma = 0.80$; see, e.g., [9, p. 150].

(6) A three-stage, second-order non-SSP ERK method (which we call NSSP(3,2)) with Butcher tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 \\ 1 & 0 & 1 & 0 \\ \hline & \frac{1}{2} & 0 & \frac{1}{2} \end{array}.$$

It is easy to show that the linear stability function (and hence the amplification factor) of this ERK method is the same as SSP(3,3) and indeed all three-stage, third-order ERK methods. According to our analysis, it has the same linear stability properties, and in particular CFL number $\sigma = 1.43$, as SSP(3,3); see [11].

(7) A three-stage, third-order non-SSP ERK method (which we call NSSP(3,3)) with Butcher tableau

$$\begin{array}{c|ccc} 0 & 0 & 0 & 0 \\ -\frac{4}{9} & -\frac{4}{9} & 0 & 0 \\ \frac{2}{3} & \frac{7}{6} & -\frac{1}{2} & 0 \\ \hline & \frac{1}{4} & 0 & \frac{3}{4} \end{array}.$$

This method has negative coefficients. It is sometimes necessary to specially treat right-hand side function evaluations $\hat{f}(u)$ that correspond to negative coefficients, e.g., by downwinding [18, 17]. However, according to our analysis, the linear stability properties of NSSP(3,3) are identical to SSP(3,3) (and all other three-stage, third-order ERK methods). We show that, even for a nonlinear problem (Example 2), it has the same stability performance as SSP(3,3).

(8) A five-stage, third-order non-SSP ERK method (which we call NSSP(5,3)) with Butcher tableau

$$\begin{array}{c|ccccc} 0 & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{7} & \frac{1}{7} & 0 & 0 & 0 & 0 \\ \frac{3}{16} & 0 & \frac{3}{16} & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & \frac{1}{3} & 0 & 0 \\ \frac{2}{3} & 0 & 0 & 0 & \frac{2}{3} & 0 \\ \hline & \frac{1}{4} & 0 & 0 & 0 & \frac{3}{4} \end{array}.$$

This is a new, low-storage ERK method, whose stability function is $1 + \hat{z} + \frac{1}{2}\hat{z}^2 + \frac{1}{6}\hat{z}^3 + \frac{1}{32}\hat{z}^4 + \frac{1}{224}\hat{z}^5$. It is a linearly stable ERK method for WENO5 according to Theorem 3.11. Its CFL number can be directly estimated to be

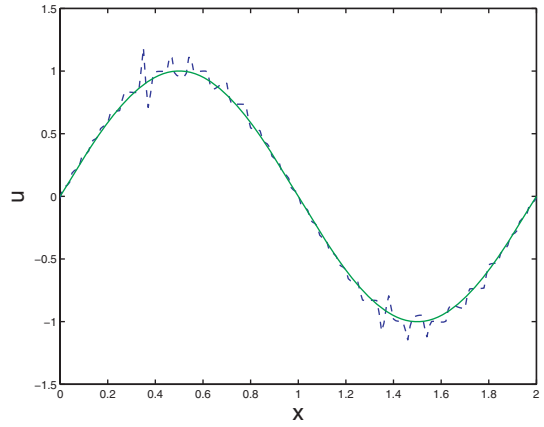


FIG. 3. *FE for* $u(x,0) = \sin(\pi x)$.

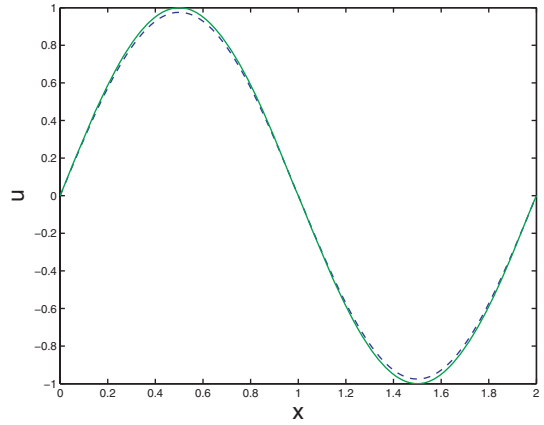


FIG. 4. *NSSP*(2,1) *for* $u(x,0) = \sin(\pi x)$.

$\sigma = 2.56$. Defining the effective CFL number to be $\frac{\sigma}{s}$, it is easy to see that a larger effective CFL number leads to more efficient time integration. The effective CFL number is 0.512 for NSSP(5,3), which is larger than 0.477 for SSP(3,3). We choose this scheme to illustrate that the theoretical principles described in this paper give us the ability to develop more efficient schemes than the popular SSP(3,3). We report on the results for the optimal ERK schemes for WENO5 elsewhere.

Example 1. The first example is the linear advection equation

$$u_t + u_x = 0, \qquad 0 < x < 2,\ t > 0,$$

with periodic boundary conditions. We consider three different initial conditions: (a) the smooth initial condition $u(x,0) = \sin(\pi x)$, (b) the smooth but more spatially varying initial condition $u(x,0) = \sin^9(\pi x)$, and (c) the discontinuous initial condition

$$u(x,0) = \begin{cases} 1 & \text{if } 0 < x < 0.5 \text{ or } 1.5 < x < 2, \\ 0 & \text{if } 0.5 \le x \le 1.5. \end{cases} \tag{4.1}$$

(a) Figures 3 and 4 show the performance of FE and NSSP(2,1) for the problem with the smooth initial condition $u(x,0) = \sin(\pi x)$. The solid lines in the figures are the

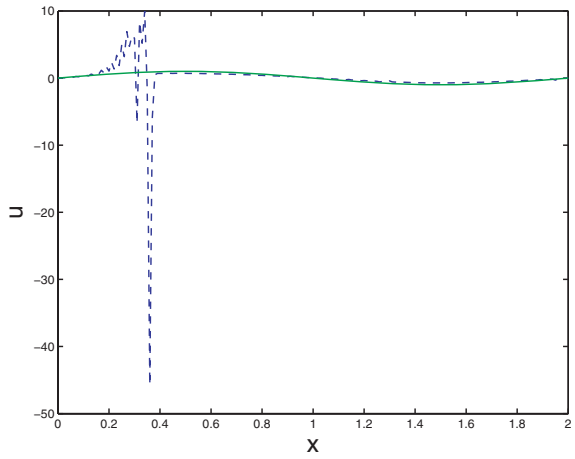


FIG. 5. *SSP*(2,2) *for* $u(x,0) = \sin(\pi x)$ *at* $T_{out} = 16$.

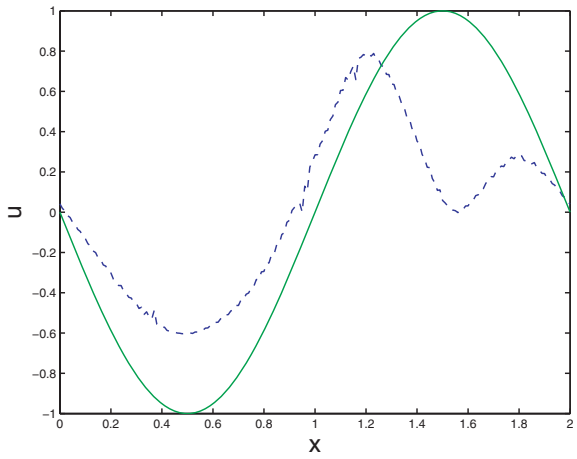


FIG. 6. *SSP*(2,2) *for* $u(x,0) = \sin(\pi x)$ *at* $T_{out} = 25$.

exact solutions, and the dashed lines are the computed solutions. Both solutions are computed with $N = 200$, $\sigma = 0.5$, and they are plotted at $T_{out} = 2$.

As expected, spurious oscillations due to linear instability are present when the FE method is used, whereas there is no instability exhibited for NSSP(2,1). Although we show only the numerical result for $\sigma = 0.5$, we emphasize that *the linear instability of the FE method appears for every* $\sigma > 0$*, no matter how small*. We observe the linear instability of the FE method for any of the later problems, whereas NSSP(2,1) is stable when $\sigma \leq 0.8$, which agrees with our expectation. However, the dissipation of NSSP(2,1) is very strong; this makes it unsuitable for computation in practice.

We now show an example of the nonconvergence effect of linear instability. Figures 5 and 6 show the result of the numerical integration using WENO5 coupled with SSP(2,2) at two different output times. The solutions are computed with $N = 200$ and $\sigma = 1.32$. Figure 5 gives the solution at $T_{out} = 16$, while Figure 6 gives it at $T_{out} = 25$. We note that oscillations are generated almost immediately at the start of the integration. By $T_{out} = 16$, Figure 5 shows significant oscillation. However, the

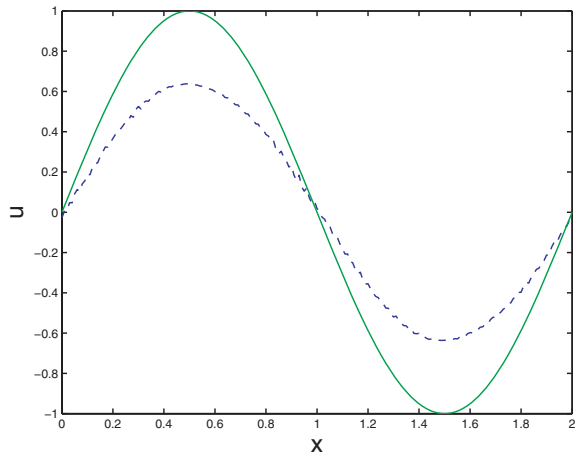


FIG. 7. *SSP*(3,3) *with* $\sigma = 1.5$ *for* $u(x,0) = \sin(\pi x)$.

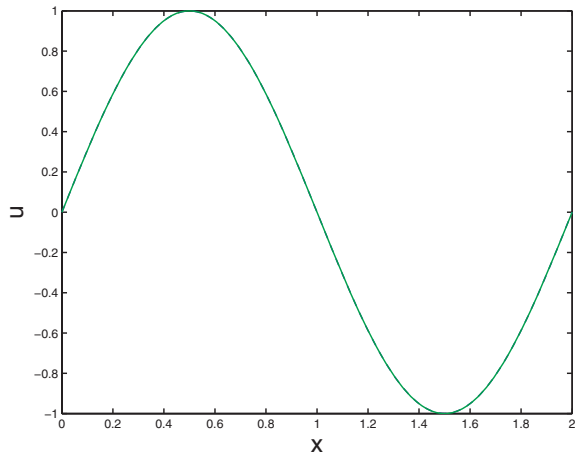


FIG. 8. *NSSP*(5,3) *with* $\sigma = 2.5$ *for* $u(x,0) = \sin(\pi x)$.

solution does not become unbounded, as can be seen in Figure 6 for $T_{out} = 25$; i.e., WENO5 has successfully adapted to the oscillations, not allowing them to become unbounded. However, with $N = 1000$ and the same σ, the solution quickly becomes unbounded. In other words, the spurious oscillations confirm the linear instability, but due to the nonlinear nature of WENO5, the numerical solution does not necessarily become unbounded. However, we point out that the linear instability of the combined method precludes convergence to the true solution; *i.e., irrespective of the long-term boundedness, the error of the numerical solution in such cases can at best be expected to be* $\mathcal{O}(1)$.

We now compare the performance of SSP(3,3) and NSSP(5,3). Figure 7 shows the performance of SSP(3,3) with $\sigma = 1.5$ for the problem just described, while Figure 8 shows the performance of NSSP(5,3) with $\sigma = 2.5$. The solid lines in the figures are the exact solutions, and the dashed lines are the computed solutions. Both solutions are computed with $N = 200$ and plotted at $T_{out} = 30$. We choose the two CFL numbers to make the computational costs equal for both experiments. Note that

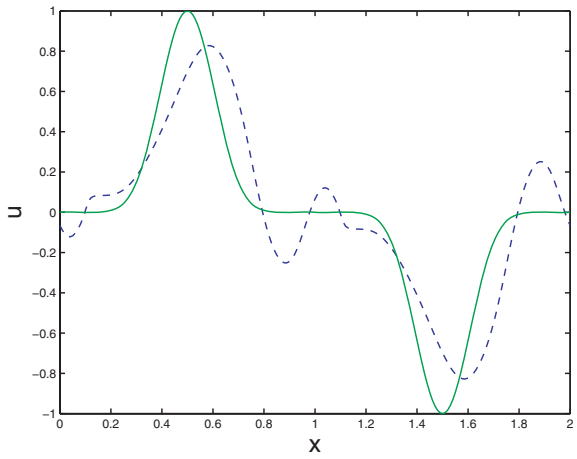


FIG. 9. *SSP*(2,2) *for* $u(x,0) = \sin^9(\pi x)$.

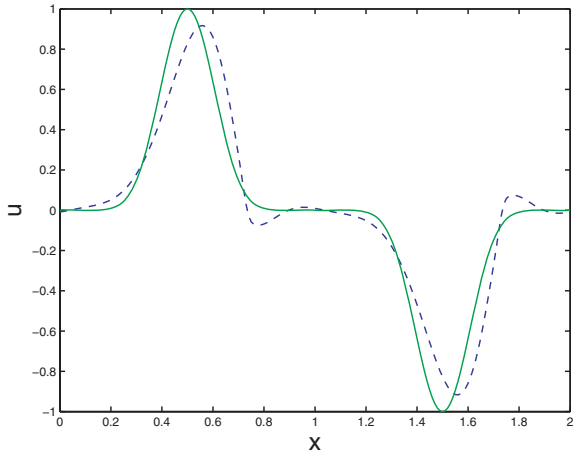


FIG. 10. *SSP*(3,2) *for* $u(x,0) = \sin^9(\pi x)$.

the difference between the computed solution by NSSP(5,3) and the exact solution is negligible. On the other hand, spurious oscillations appear for SSP(3,3) because it is linearly stable only for CFL numbers less than about 1.43. Once again we note that the solution does not become unbounded. Another interesting observation is that, when we choose $\sigma = 1.4$ for SSP(3,3), the difference between the computed solution and the exact solution is also negligible. This means that the solution by SSP(3,3) with $\sigma = 1.5$ is inaccurate due only to linear instability. The experiment clearly favors NSSP(5,3) for its larger effective CFL number.

Our experiments show that when SSP(2,2) and SSP(3,2) are used for the above problem with a small value for σ, the instability requires a long time to develop. In many cases the oscillations are not conspicuous in a relatively short time.

(b) We can introduce more spatial difficulty by using $u(x,0) = \sin^9(\pi x)$. Figures 9–12 show the performance of SSP(2,2), SSP(3,2), NSSP(3,2), and NSSP(3,3), respectively. The solid lines in the figures are the exact solutions, and the dashed lines are the computed solutions. All solutions are computed with $N = 200$, $\sigma = 0.5$, and they are plotted at $T_{out} = 150$.

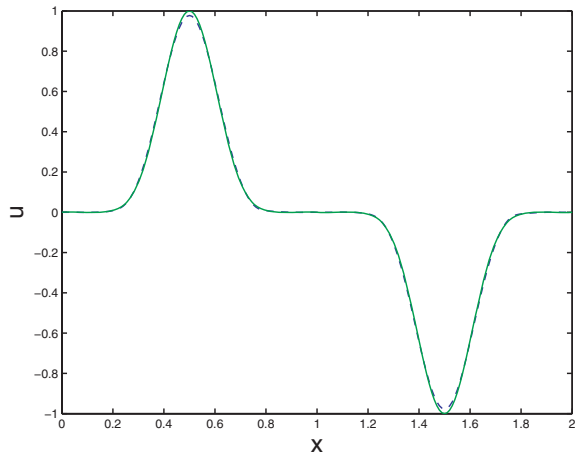


FIG. 11. *NSSP*(3,2) *for* $u(x,0) = \sin^9(\pi x)$.

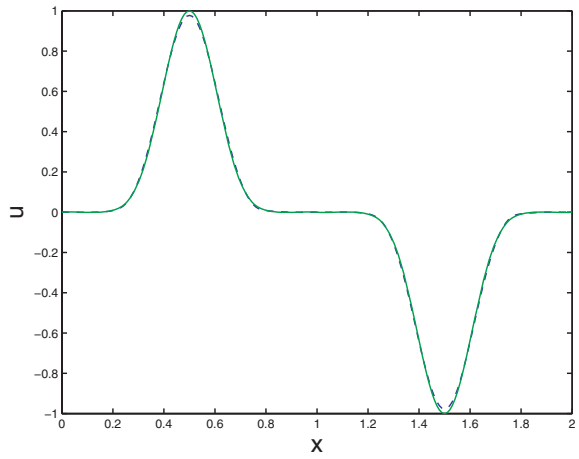


FIG. 12. *NSSP*(3,3) *for* $u(x,0) = \sin^9(\pi x)$.

The linear instability of WENO5 coupled with SSP(2,2) or SSP(3,2) takes a long time to become conspicuous for this problem, whereas NSSP(3,2) and NSSP(3,3) are stable. Moreover, smaller values of σ tend to delay the manifestation of the instability even further. We also observe that the linear instability of the combination of WENO5 and SSP(3,2) develops more slowly than that of WENO5 and SSP(2,2). This can be explained by the fact that the classical (linear) stability domain of SSP(3,2) includes the classical stability domain of SSP(2,2).

(c) The final initial condition is the step function (4.1). As we know, WENO5 is widely used in the numerical simulations of discontinuous solutions of hyperbolic PDEs. If the solution has only a few discontinuities and the wave speed is not zero, i.e., there is no stationary shock, the WENO5 stencil weights at any given point are the same as for the continuous case (3.2) for the majority of the time. That is, only the points close to the discontinuity use different discretizations because some stencil weights approach zero. However, the discretization at these points returns to those of the continuous case after the discontinuity passes through. Therefore, we

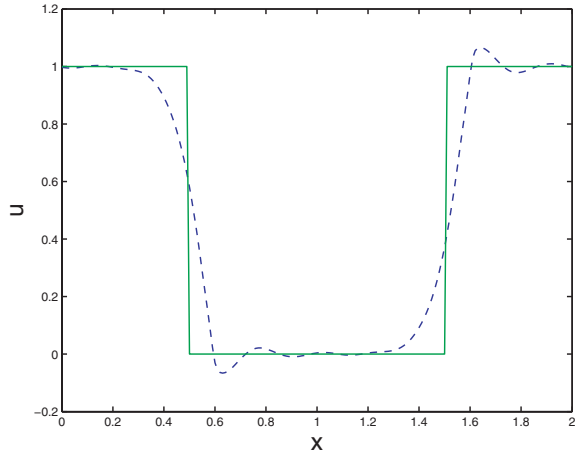


FIG. 13. *SSP*(2,2) *for* (4.1).

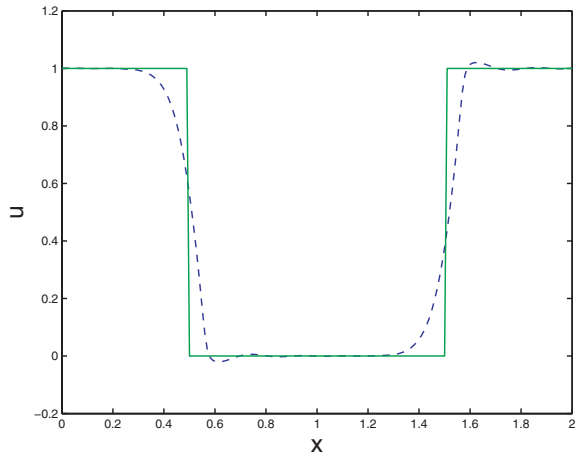


FIG. 14. *SSP*(3,2) *for* (4.1).

expect that our analysis for the continuous case is also relevant to such discontinuous problems. Figures 13–16 show the performance of SSP(2,2), SSP(3,2), NSSP(3,2), and NSSP(3,3), respectively, for the discontinuous initial condition (4.1). The solid lines in the figures are the exact solutions, and the dashed lines are the computed solutions. All solutions are computed with $N = 200$, $\sigma = 0.5$, and they are plotted at $T_{out} = 50$.

We again make the observation that NSSP(3,2) and NSSP(3,3) are stable, whereas SSP(2,2) and SSP(3,2) exhibit oscillations.

We now compare SSP(3,3) with NSSP(5,3) for the discontinuous initial condition (4.1). Figures 17 and 18 show the performance of SSP(3,3) with $\sigma = 1.11$ and NSSP(5,3) with $\sigma = 1.85$. We choose the two CFL numbers to make the computational costs equal for both experiments. The solid lines in the figures are the exact solutions, and the dashed lines are the computed solutions. All solutions are computed with $N = 200$, and they are plotted at $T_{out} = 10$. Spurious oscillations appear when we use SSP(3,3) with $\sigma = 1.11$, whereas there is no problem with NSSP(5,3) with

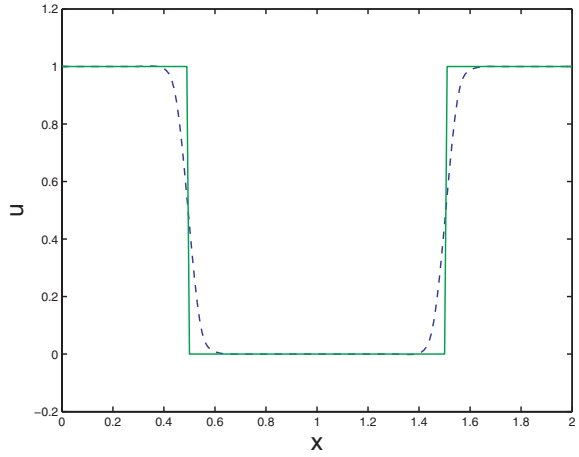


FIG. 15. *NSSP*(3,2) *for* (4.1).

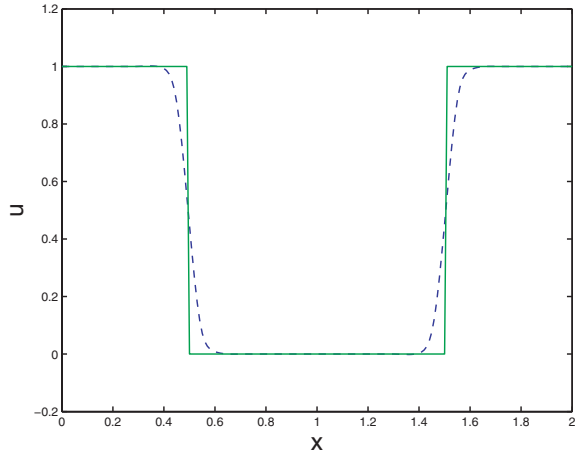


FIG. 16. *NSSP*(3,3) *for* (4.1).

$\sigma = 1.85$. These numerical experiments show that WENO5 coupled with SSP(3,3) is stable when $\sigma \leq 1$ for the discontinuous initial condition (4.1), whereas NSSP(5,3) is stable when $\sigma \leq 1.9$. This example shows that larger time steps can be used with NSSP(5,3) than with SSP(3,3) even when the solutions have discontinuities.

Example 2. The second example is the inviscid Burgers equation

$$u_t + \left(\frac{u^2}{2}\right)_x = 0, \qquad 0 < x < 2,\ t > 0,$$

with periodic boundary conditions and the initial condition $u(x,0) = 2 + \sin^9(\pi x)$. This is a nonlinear problem with $f(u) = u^2/2$. Thus $\partial f/\partial u = u$. For a given σ, Δt is chosen as $\Delta t = \sigma \Delta x(\max_j u_j)$. Figures 19–22 show the performance of SSP(2,2), SSP(3,2), NSSP(3,2), and NSSP(3,3), respectively. The solid lines in the figures represent the reference solution, which is generated using WENO5 with SSP(3,3), $N = 1000$, and $\sigma = 0.5$. The dashed lines are the computed solutions, all of which have $N = 100$, $\sigma = 0.5$, and they are plotted at $T_{out} = 40$. Again we see that NSSP(3,2)

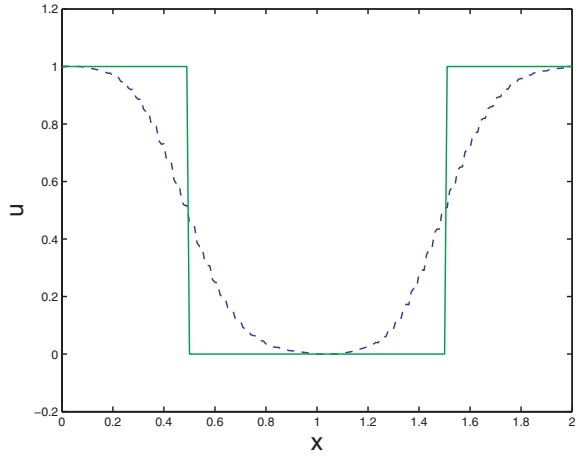


FIG. 17. *SSP*(3,3) *with* $\sigma = 1.11$ *for* (4.1).

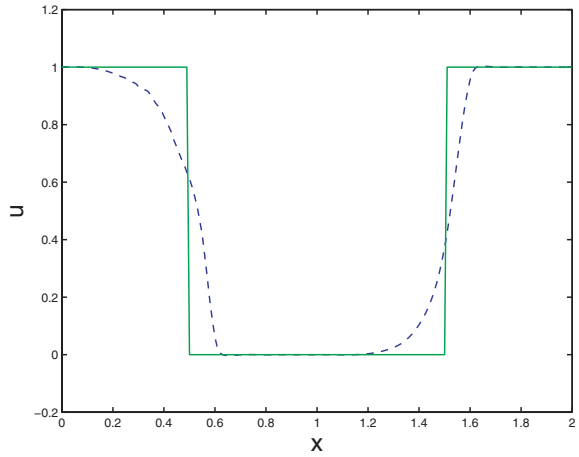


FIG. 18. *NSSP*(5,3) *with* $\sigma = 1.85$ *for* (4.1).

and NSSP(3,3) are stable, whereas SSP(2,2) and SSP(3,3) exhibit oscillations. This is a compelling illustration that our analysis accurately predicts the linear stability of ERK methods *even with negative coefficients and even when applied to a nonlinear problem whose solution develops a discontinuity.*

5. Conclusions. In this paper we employ a linear stability analysis for ERK time integration methods coupled with the WENO5 spatial discretization. We prove that the forward Euler method, all two-stage, second-order ERK methods, and all optimal SSP ERK methods of up to second order are linearly unstable when coupled with WENO5 and used for solving hyperbolic conservation laws. Hence all of these combined methods are also not convergent. Moreover, we show that, in our analysis, the success of the popular SSP(3,3) method is not due to the SSP property; indeed all three-stage, third-order ERK methods, including those with negative coefficients or those that are provably non-SSP, have precisely the same linear stability performance according to our analysis, and this has translated to very similar performance in

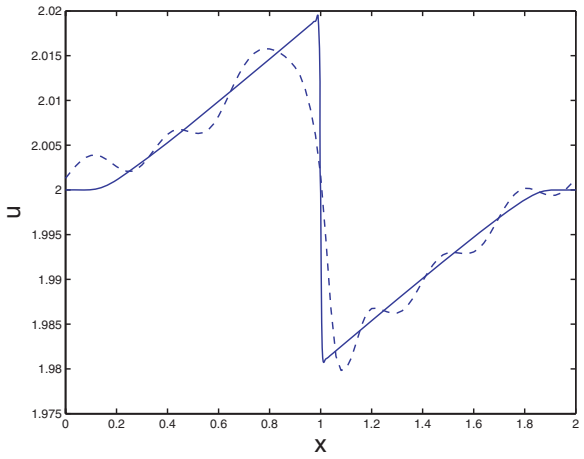


FIG. 19. *SSP*(2,2) *for the Burgers equation.*

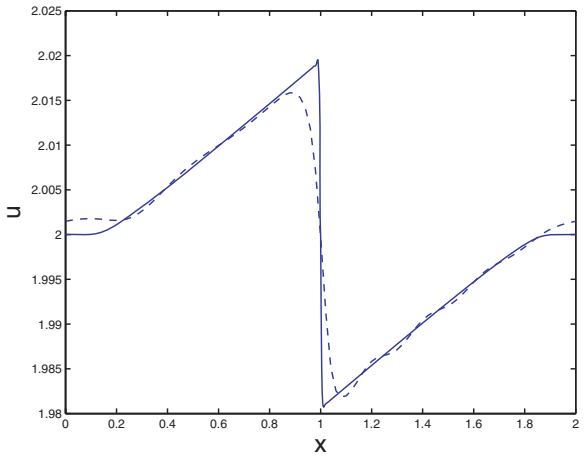


FIG. 20. *SSP*(3,2) *for the Burgers equation.*

the examples we have presented. New stability criteria are derived for general ERK methods *of any order*. Numerical experiments demonstrate that, although the analysis is strictly valid only for linear constant-coefficient problems with continuous initial conditions and periodic boundary conditions, it is relevant to both linear and nonlinear problems with continuous and discontinuous solutions. It is also relevant to ERK methods with negative coefficients without a special treatment (downwinding) of the spatial operator. For linear stability of an ERK time integration method coupled with WENO5, we show that it is sufficient that the classical linear stability region of the ERK method include a piece of the imaginary axis. The analysis techniques described in this paper apply to other WENO methods such as the seventh- or higher-order WENO methods. From this analysis it is also possible to derive optimal ERK methods in terms of the CFL number for WENO5. In particular, it is possible to derive methods such as NSSP(5,3) that are more efficient (i.e., have larger effective CFL numbers) than the benchmark method SSP(3,3). We report on these results elsewhere.

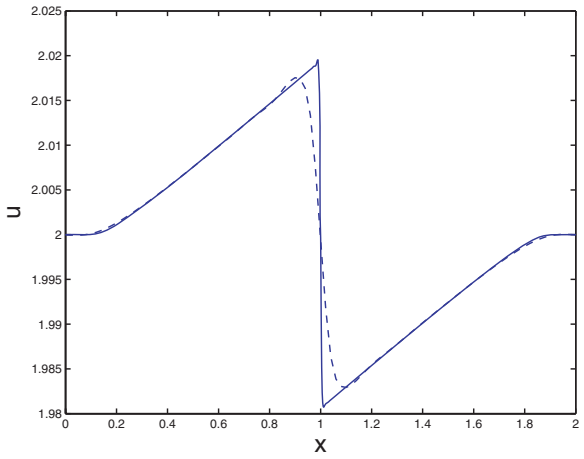


FIG. 21. *NSSP*(3,2) *for the Burgers equation.*

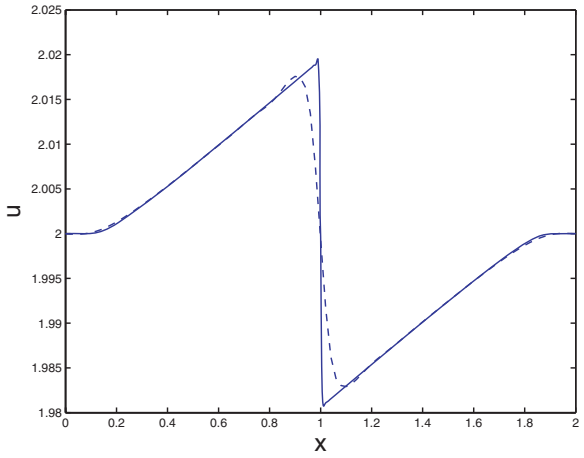


FIG. 22. *NSSP*(3,3) *for the Burgers equation.*

Acknowledgments. The authors express their thanks to the editor and the referees for their comments. R.J.S. also thanks G. Puppo, I. Higueras, and L. Ferracina for useful discussions.

REFERENCES

[1] J. BLAZEK, *Computational Fluid Dynamics: Principles and Applications*, Elsevier, Oxford, 2001.

[2] L. FERRACINA AND M. N. SPIJKER, *Stepsize restrictions for the total-variation-diminishing property in general Runge-Kutta methods*, SIAM J. Numer. Anal., 42 (2004), pp. 1073–1093.

[3] S. GOTTLIEB, C.-W. SHU, AND E. TADMOR, *Strong stability-preserving time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[4] E. HAIRER, S. P. NØRSETT, AND G. WANNER, *Solving Ordinary Differential Equations.* I. *Nonstiff Problems*, Springer Ser. Comput. Math. 8, Springer-Verlag, Berlin, 1987.

[5] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations.* II. *Stiff and Differential-Algebraic Problems*, Springer Ser. Comput. Math. 14, Springer-Verlag, Berlin, 1991.

[6] A. Harten and S. Osher, *Uniformly high-order accurate nonoscillatory schemes*. I, SIAM J. Numer. Anal., 24 (1987), pp. 279–309.
[7] A. Harten, B. Engquist, S. Osher, and S. Chakravarthy, *Uniformly high order essentially non-oscillatory schemes* III, J. Comput. Phys., 71 (1987), pp. 231–303.
[8] I. Higueras, *On strong stability preserving time discretization methods*, J. Sci. Comput., 21 (2004), pp. 193–223.
[9] W. Hundsdorfer and J. Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*, Springer Ser. Comput. Math. 33, Springer-Verlag, Berlin, 2003.
[10] A. Jameson, *Solution of the Euler equations for two dimensional transonic flow by a multigrid method*, Appl. Math. Comput., 13 (1983), pp. 327–355.
[11] G.-S. Jiang and C.-W. Shu, *Efficient implementation of weighted ENO schemes*, J. Comput. Phys., 126 (1996), pp. 202–228.
[12] J. F. B. M. Kraaijevanger, *Contractivity of Runge-Kutta methods*, BIT, 31 (1991), pp. 482–528.
[13] H.-O. Kreiss and G. Scherer, *Method of lines for hyperbolic differential equations*, SIAM J. Numer. Anal., 29 (1992), pp. 640–646.
[14] D. Levy and E. Tadmor, *From semidiscrete to fully discrete: Stability of Runge-Kutta schemes by the energy method*, SIAM Rev., 40 (1998), pp. 40–73.
[15] X.-D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 200–212.
[16] S. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, in Appl. Math. Sci., Springer-Verlag, New York, 2003.
[17] S. J. Ruuth and R. J. Spiteri, *High-order strong-stability-preserving Runge-Kutta methods with downwind-biased spatial discretizations*, SIAM J. Numer. Anal., 42 (2004), pp. 974–996.
[18] C.-W. Shu, *Total-variation-diminishing time discretizations*, SIAM J. Sci. Comput., 9 (1988), pp. 1073–1084.
[19] C.-W. Shu, *Numerical experiments on the accuracy of ENO and modified ENO schemes*, J. Sci. Comput., 5 (1990), pp. 127–149.
[20] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, Lecture Notes in Math. 1697, Springer-Verlag, Berlin, 1998, pp. 325–432.
[21] R. J. Spiteri and S. J. Ruuth, *A new class of optimal high-order strong-stability-preserving time discretization methods*, SIAM J. Numer. Anal., 40 (2002), pp. 469–491.
[22] E. Tadmor, *Stability analysis of finite difference, pseudospectral and Fourier-Galerkin approximations for time-dependent problems*, SIAM Rev., 29 (1987), pp. 525–555.
[23] J. W. Thomas, *Numerical partial differential equations: Finite difference methods*, in Texts Appl. Math., Springer-Verlag, New York, 1995.
[24] A. Vande Wouwer, P. Saucez, and W. E. Schiesser, eds., *Adaptive Method of Lines*, Chapman and Hall/CRC Press, Boca Raton, FL, 2001.
[25] Z. Xu and C.-W. Shu, *Anti-diffusive flux corrections for high order finite difference WENO schemes*, J. Comput. Phys., 205 (2005), pp. 458–485.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1902–1936

STABLE DIFFERENCE APPROXIMATIONS FOR THE ELASTIC WAVE EQUATION IN SECOND ORDER FORMULATION*

STEFAN NILSSON†, N. ANDERS PETERSSON†, BJÖRN SJÖGREEN†, AND HEINZ-OTTO KREISS‡

Abstract. We consider the three-dimensional elastic wave equation for an isotropic heterogeneous material subject to a stress-free boundary condition. Building on our recently developed theory for difference methods for second order hyperbolic systems [H.-O. Kreiss, N. A. Petersson, J. Yström, *SIAM J. Numer. Anal.*, 40 (2002), pp. 1940–1967], we develop an explicit, second order accurate technique which is stable for all ratios of longitudinal over transverse phase velocities. The spatial discretization is self-adjoint, and the stability is obtained through an energy estimate. Seismic events are often modeled using singular source terms, and we devise a technique to place sources independently of the grid while retaining second order accuracy away from the source. Several numerical examples are given.

Key words. elastic wave equation, finite differences, stability, energy estimate, seismic wave propagation

AMS subject classifications. 65M06, 74B05, 86A15

DOI. 10.1137/060663520

1. Introduction.

As a model for seismic wave propagation, we consider the elastic wave equation for an isotropic heterogeneous material in a three-dimensional domain Ω:

$$\rho \frac{\partial^2 \mathbf{u}}{\partial t^2} = \nabla \cdot \mathfrak{T} + \mathbf{f}, \quad \mathbf{x} \in \Omega,\ t \geq 0, \tag{1}$$
$$\mathfrak{T} = \lambda (\nabla \cdot \mathbf{u}) \mathbf{I} + \mu \left(\nabla \mathbf{u} + \nabla \mathbf{u}^T \right),$$

subject to initial data

$$\mathbf{u}(\mathbf{x}, 0) = \mathbf{U}_0(\mathbf{x}), \quad \mathbf{u}_t(\mathbf{x}, 0) = \mathbf{U}_1(\mathbf{x}), \quad \mathbf{x} \in \Omega.$$

Here $\mathfrak{T}$ is the stress tensor, $\mathbf{u} = \mathbf{u}(\mathbf{x}, t)$ is the displacement vector with Cartesian components $\mathbf{u} = (u, v, w)^T$, where $\mathbf{x} = (x, y, z)^T$ is the location, and t is time. $\mathbf{f}$ is the external (volume) forcing, and the material properties are characterized by the density $\rho(\mathbf{x}) > 0$ and the Lamé parameters $\lambda(\mathbf{x}) > 0$ and $\mu(\mathbf{x}) \geq 0$. The degenerate case $\mu = 0$ corresponds to acoustic wave propagation and will not be discussed here. We henceforth assume $\mu(\mathbf{x}) > 0$.

Common boundary conditions include a Dirichlet condition for $\mathbf{u}$ or a normal stress condition

$$\mathfrak{T} \cdot \hat{\mathbf{n}} = \lambda (\nabla \cdot \mathbf{u}) \hat{\mathbf{n}} + \mu (\nabla \mathbf{u} + \nabla \mathbf{u}^T) \cdot \hat{\mathbf{n}} = \mathbf{g}, \tag{2}$$

*Received by the editors June 21, 2006; accepted for publication (in revised form) April 26, 2007; published electronically August 31, 2007. This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contractW-7405-Eng-48.

http://www.siam.org/journals/sinum/45-5/66352.html

†Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Livermore, CA 94551 (nilsson2@llnl.gov, andersp@llnl.gov, sjogreen2@llnl.gov).

‡Träskö-Storö Institute of Mathematics, Stockholm, Sweden (hokreiss@nada.kth.se).

which prescribes the stresses on a boundary with unit normal $\hat{\mathbf{n}}$. When $\mathbf{g} = \mathbf{0}$, this boundary condition is often called a free surface or stress-free condition. The system (1) admits longitudinal (P, or primary) and transverse (S, or secondary) waves which propagate at phase velocities

$$c_p = \sqrt{(2\mu + \lambda)/\rho} \quad \text{and} \quad c_s = \sqrt{\mu/\rho},$$

respectively. There can also be surface waves, which travel along a free surface, as well as waves which travel along internal material discontinuities.

Finite difference approximations of the elastodynamic equations in second order formulation have been around for a long time [2, 3]. Early methods, based on explicit centered difference approximations, were initially very successful but suffered from instability problems when a free surface boundary condition was imposed, and the ratio between the P- and S-wave velocities

$$\nu = \frac{c_p}{c_s}$$

became too large [13] (note that $\nu > \sqrt{2}$). Ilan [14] proposed a remedy which applied only to materials with constant properties normal to the boundary, and an implicit boundary update technique was suggested by Vidale and Clayton [25]. However, no generally applicable, stable, explicit discretization was found for the second order formulation which worked for high values of ν. Due to the instability problems, alternative formulations were explored where the elastic wave equation was rewritten as a larger first order system for the three velocity and six stress components and discretized on a staggered grid [21]. Most current finite difference methods for seismic wave propagation are based on the staggered grid technique. It is, however, difficult to handle complex geometry (e.g., topography) with these staggered grid methods, so there has been recent interest in more expensive methods based on unstructured meshes, such as the spectral element technique described by Komatitsch and Tromp [15].

In this paper we revisit the problem of devising an explicit finite difference method for the elastic wave equation in second order formulation, subject to a free surface boundary condition. Building on our recently developed theory for difference methods for second order hyperbolic systems [18], we develop a technique which is stable for all ratios c_p/c_s. We focus on the long-wave approximation where topography is neglected, and the stress-free boundary condition is enforced on a flat surface which is aligned with a grid surface. However, our longer term goal is to extend the embedded boundary technique [19, 17, 16] to the elastic wave equation for handling general domains. In seismic applications, the material parameters ρ, μ, and λ often vary on a length scale which is significantly smaller than the wavelength of the elastic waves. Hence the material parameters can vary rapidly on the computational grid, and to guarantee stability it is desirable to develop a numerical method which satisfies an energy estimate. For a hyperbolic system in second order formulation, the key to an energy estimate is a spatial discretization which is self-adjoint, i.e., corresponds to a symmetric or symmetrizable matrix. In this paper, we present a discretization which makes the spatial approximation second order accurate, self-adjoint, and explicit. The self-adjoint property also implies that the method is conservative.

In section 1.1 we introduce the basic ideas behind our spatial discretization by studying the scalar wave equation with a cross term in two space dimensions. The discretization technique is generalized to the elastic wave equation in section 2, where

we present a theory proving that the method is second order accurate and stable for all values of c_p/c_s. The stability and accuracy of the new method are also illustrated with computational experiments. Seismic events (for example, earthquakes) are often modeled using singular source terms applied at points, along lines, or over surfaces in the three-dimensional domain. In section 3 we devise a technique to place sources independently of the grid while retaining second order accuracy away from the source. We also study how the temporal smoothness of a point source affects the spatial smoothness of the solution. In section 4 we first study how the phase velocity of surface waves depends on the number of grid points per wavelength. Thereafter, we solve a benchmark problem for a simplified earthquake where the sources are distributed along a plane. Some comments on our implementation of nonreflecting boundary conditions for truncating unbounded domains are also given.

1.1. A model problem. We introduce our discretization technique on the half-plane problem for the scalar wave equation with a cross term in two dimensions:

$$\frac{\partial^2 u}{\partial t^2} = \nabla \cdot \mathbf{F}, \quad x \geq 0,\ 0 \leq y \leq 2\pi,\ t \geq 0, \tag{3}$$
$$\mathbf{F} = \begin{pmatrix} u_x + \alpha u_y \\ u_y + \alpha u_x \end{pmatrix},$$

with 2π-periodic solutions in the y-direction, subject to the boundary condition

$$\mathbf{F} \cdot \hat{\mathbf{n}} = u_x + \alpha u_y = 0, \quad x = 0,\ 0 \leq y \leq 2\pi,\ t \geq 0, \quad \text{when} \quad \hat{\mathbf{n}} = (1,0)^T. \tag{4}$$

Here α is a real constant. Similar to the elastic wave equation, the problem (3)–(4) conserves an energy:

$$\|u_t\|^2 + \|u_x\|^2 + \|u_y\|^2 + 2\alpha(u_x, u_y) = \text{const},$$

where (u, v) is the L_2 scalar product and $\|u\|^2 = (u, u)$. We have

$$\|u_x\|^2 + \|u_y\|^2 + 2\alpha(u_x, u_y) \geq (1 - |\alpha|)\left(\|u_x\|^2 + \|u_y\|^2\right) > 0, \quad |\alpha| < 1.$$

Hence the conserved quantity is a norm, and the problem (3)–(4) is well-posed, when $|\alpha| < 1$. Conversely, it can be shown that the problem becomes ill-posed for $|\alpha| > 1$.

We introduce a grid with points $x_i = (i-1)h$, $y_j = (j-1)h$, $i = 0, 1, 2, \ldots$, $j = 1, 2, \ldots, N_y$, where $h = 2\pi/(N_y - 1)$ is the grid size. We denote a two-dimensional grid function by $u_{i,j}(t) = u(x_i, y_j, t)$. The time dependence will be suppressed when the meaning is obvious. We use the usual definitions of divided difference operators

$$D_+^x v_{i,j} = \frac{1}{h}(v_{i+1,j} - v_{i,j}), \quad D_-^x v_{i,j} = D_+^x v_{i-1,j}, \quad D_0^x = \frac{1}{2}\left(D_+^x + D_-^x\right)$$

and corresponding expressions in the y-direction.

A second order accurate centered spatial discretization of (3) is given by

$$\frac{d^2 u_{i,j}}{dt^2} = \left(D_-^x D_+^x + D_-^y D_+^y + 2\alpha D_0^x D_0^y\right) u_{i,j}, \quad i \geq 1, \quad 1 \leq j \leq N_y - 1. \tag{5}$$

There are several ways to discretize the boundary condition (4) to second order accuracy. As we shall see, a good choice is

$$D_0^x u_{1,j} + \alpha D_0^y \left(\frac{u_{2,j} + u_{0,j}}{2} \right) = 0, \quad 1 \le j \le N_y - 1. \tag{6}$$

After Fourier transforming in the y-direction (with dual variable ω), using the boundary condition (6) to eliminate the ghost point values at $i = 0$, and introducing the vector notation $\hat{\mathbf{u}} = (\hat{u}_1, \hat{u}_2, \dots)^T$, we can write the Fourier-transformed semidiscrete problem in matrix form

$$h^2 \frac{d^2 \hat{\mathbf{u}}}{dt^2} = (A + B)\hat{\mathbf{u}}, \tag{7}$$

where

$$A = \begin{pmatrix} -\left(2 + 4\sin^2 \frac{\omega h}{2}\right) & 2 & & \\ 1 & -\left(2 + 4\sin^2 \frac{\omega h}{2}\right) & 1 & \\ & \ddots & \ddots & \ddots \end{pmatrix},$$

$$B = \imath\alpha \sin(\omega h) \begin{pmatrix} 0 & 2 & & \\ -1 & 0 & 1 & \\ & \ddots & \ddots & \ddots \end{pmatrix},$$

and $\imath = \sqrt{-1}$. We can symmetrize (7) by the diagonal scaling

$$S = \begin{pmatrix} 1/\sqrt{2} & 0 & & \\ 0 & 1 & 0 & \\ & \ddots & \ddots & \ddots \end{pmatrix}, \quad \hat{\mathbf{w}} = S\hat{\mathbf{u}},$$

$$h^2 \frac{d^2 \hat{\mathbf{w}}}{dt^2} = (\tilde{A} + \tilde{B})\hat{\mathbf{w}}, \quad \tilde{A} + \tilde{B} = S(A + B)S^{-1},$$

where $\tilde{A} + \tilde{B}$ is self-adjoint. As we shall see in section 2, the semidiscrete problem is stable if $\tilde{A} + \tilde{B}$ also is negative definite. Furthermore, when $\tilde{A} + \tilde{B}$ is self-adjoint, it is straightforward to discretize time such that the fully discrete problem becomes stable and conserves a discrete energy which is a second order accurate approximation of the conserved energy in the continuous case.

Note that it is not necessary to solve a linear system to update the ghost points. Instead of (6), we can change the boundary condition to be

$$D_0^x u_{1j} + \alpha D_0^y u_{1j} = 0, \quad 1 \le j \le N_y - 1, \tag{8}$$

if we also modify the difference approximation on the boundary by taking the cross term one-sided in the direction normal to the boundary

$$\frac{d^2 u_{1,j}}{dt^2} = \left(D_-^x D_+^x + D_-^y D_+^y + 2\alpha D_+^x D_0^y \right) u_{1,j}, \quad 1 \le j \le N_y - 1. \tag{9}$$

After Fourier transforming (9) in the y-direction and eliminating the ghost point by use of (8), we obtain the same matrix representation as before, showing that the two formulations are equivalent.

2. The elastic wave equation. In Cartesian component form, the system (1) is

$$\rho u_{tt} = \frac{\partial}{\partial x}\left((2\mu+\lambda)u_x + \lambda v_y + \lambda w_z\right) + \frac{\partial}{\partial y}\left(\mu v_x + \mu u_y\right) + \frac{\partial}{\partial z}\left(\mu u_z + \mu w_x\right) + f^{(x)}, \tag{10}$$

$$\rho v_{tt} = \frac{\partial}{\partial x}\left(\mu v_x + \mu u_y\right) + \frac{\partial}{\partial y}\left((2\mu+\lambda)v_y + \lambda u_x + \lambda w_z\right) + \frac{\partial}{\partial z}\left(\mu v_z + \mu w_y\right) + f^{(y)}, \tag{11}$$

$$\rho w_{tt} = \frac{\partial}{\partial x}\left(\mu u_z + \mu w_x\right) + \frac{\partial}{\partial y}\left(\mu v_z + \mu w_y\right) + \frac{\partial}{\partial z}\left((2\mu+\lambda)w_z + \lambda u_x + \lambda v_y\right) + f^{(z)}. \tag{12}$$

In this paper, we consider box-shaped domains $0 \le x \le a$, $0 \le y \le b$, $0 \le z \le c$ and impose a normal stress boundary condition at $z = 0$. In component form, the boundary condition (2) is

$$\mu u_z + \mu w_x = g^{(x)}, \tag{13}$$

$$\mu v_z + \mu w_y = g^{(y)}, \quad z = 0,\ 0 \le x \le a,\ 0 \le y \le b,\ t \ge 0, \tag{14}$$

$$(2\mu+\lambda)w_z + \lambda u_x + \lambda v_y = g^{(z)}. \tag{15}$$

For the purpose of discussing the stability properties of our method, we impose homogeneous Dirichlet conditions at $z = c$

$$\mathbf{u}(x, y, c, t) = 0, \quad 0 \le x \le a,\ 0 \le y \le b,\ t \ge 0, \tag{16}$$

and periodic boundary conditions in the x- and y-directions. Note that the stability results can be extended to the case of Dirichlet conditions in the x- and y-directions.

To simplify our notation, we assume zero volume and boundary forcings ($\mathbf{f} = \mathbf{0}$ and $\mathbf{g} = \mathbf{0}$) throughout sections 2.1–2.3.

2.1. Spatial discretization. The conclusion from the model problem in section 1.1 is that a stable second order accurate discretization of (3)–(4) can be obtained by discretizing the differential equation with centered differences, except for the cross terms on the boundary, which should be taken one-sided in the direction normal to the boundary. The resulting approximation will be second order accurate, and the ghost points can be updated explicitly if the tangential derivatives in the boundary conditions are discretized by centered differences along the boundary. We shall use these principles to define the difference scheme for the three-dimensional elastic wave equation and proceed by verifying that the resulting approximation is stable and second order accurate. The underlying ideas are the same as for the model problem, even though the algebra gets more complicated.

We define a three-dimensional grid with points $x_i = (i-1)h$, $y_j = (j-1)h$, $z_k = (k-1)h$, $0 \le i \le N_x$, $0 \le j \le N_y$, $0 \le k \le N_z$, where $h > 0$ is the grid size, $x_{N_x} = a$, $y_{N_y} = b$, and $z_{N_z} = c$. Time is discretized with step size $\delta_t > 0$ on a grid $t_n = n\delta_t$, $n = 0, 1, \ldots$, and we denote a grid function by $u^n_{i,j,k} = u(x_i, y_j, z_k, t_n)$. The superscript for time will be supressed when the meaning is obvious. Apart from the difference operators already defined, we also introduce

$$\widetilde{D^z_0} v_{i,j,k} = \begin{cases} D^z_+ v_{i,j,1}, & k = 1, \\ D^z_0 v_{i,j,k}, & k \ge 2, \end{cases}$$

and the averaging operators

$$E^x_{1/2}(\gamma_{i,j,k}) = \gamma_{i+1/2,j,k} := \frac{\gamma_{i+1,j,k} + \gamma_{i,j,k}}{2},$$
$$E^y_{1/2}(\gamma_{i,j,k}) = \gamma_{i,j+1/2,k} := \frac{\gamma_{i,j+1,k} + \gamma_{i,j,k}}{2},$$
$$E^z_{1/2}(\gamma_{i,j,k}) = \gamma_{i,j,k+1/2} := \frac{\gamma_{i,j,k+1} + \gamma_{i,j,k}}{2}.$$

We form the spatially discrete equations at the grid points $1 \le i \le N_x - 1$, $1 \le j \le N_y - 1$, $1 \le k \le N_z - 1$,

$$\rho \frac{d^2u}{dt^2} = D^x_- \left(E^x_{1/2}(2\mu+\lambda) D^x_+ u \right) + D^y_- \left(E^y_{1/2}(\mu) D^y_+ u \right) + D^z_- \left(E^z_{1/2}(\mu) D^z_+ u \right)$$
$$+ D^x_0 \left(\lambda D^y_0 v + \lambda \widetilde{D^z_0} w \right) + D^y_0 \left(\mu D^x_0 v \right) + \widetilde{D^z_0} \left(\mu D^x_0 w \right) =: L^{(u)}(u,v,w), \tag{17}$$

$$\rho \frac{d^2v}{dt^2} = D^x_- \left(E^x_{1/2}(\mu) D^x_+ v \right) + D^y_- \left(E^y_{1/2}(2\mu+\lambda) D^y_+ v \right) + D^z_- \left(E^z_{1/2}(\mu) D^z_+ v \right)$$
$$+ D^x_0 \left(\mu D^y_0 u \right) + D^y_0 \left(\lambda D^x_0 u + \lambda \widetilde{D^z_0} w \right) + \widetilde{D^z_0} \left(\mu D^y_0 w \right) =: L^{(v)}(u,v,w), \tag{18}$$

$$\rho \frac{d^2w}{dt^2} = D^x_- \left(E^x_{1/2}(\mu) D^x_+ w \right) + D^y_- \left(E^y_{1/2}(\mu) D^y_+ w \right) + D^z_- \left(E^z_{1/2}(2\mu+\lambda) D^z_+ w \right)$$
$$+ D^x_0 \left(\mu \widetilde{D^z_0} u \right) + D^y_0 \left(\mu \widetilde{D^z_0} v \right) + \widetilde{D^z_0} \left(\lambda D^x_0 u + \lambda D^y_0 v \right) =: L^{(w)}(u,v,w), \tag{19}$$

where grid point indices have been suppressed to improve readability. The free surface boundary conditions (13)–(15) are discretized by

$$\frac{1}{2} \left(\mu_{i,j,3/2} D^z_+ u_{i,j,1} + \mu_{i,j,1/2} D^z_+ u_{i,j,0} \right) + \mu_{i,j,1} D^x_0 w_{i,j,1} = 0, \tag{20}$$

$$\frac{1}{2} \left(\mu_{i,j,3/2} D^z_+ v_{i,j,1} + \mu_{i,j,1/2} D^z_+ v_{i,j,0} \right) + \mu_{i,j,1} D^y_0 w_{i,j,1} = 0, \tag{21}$$

$$\frac{1}{2} \left((2\mu+\lambda)_{i,j,3/2} D^z_+ w_{i,j,1} + (2\mu+\lambda)_{i,j,1/2} D^z_+ w_{i,j,0} \right) + \lambda_{i,j,1} \left(D^x_0 u_{i,j,1} + D^y_0 v_{i,j,1} \right) = 0 \tag{22}$$

for $1 \le i \le N_x - 1$, $1 \le j \le N_y - 1$. The Dirichlet boundary condition (16) is discretized by

$$\mathbf{u}_{i,j,N_z} = \mathbf{0}, \quad 1 \le i \le N_x,\ 1 \le j \le N_y. \tag{23}$$

The discrete counterparts of the periodic boundary conditions are

$$\mathbf{u}_{N_x,j,k} = \mathbf{u}_{1,j,k}, \quad \mathbf{u}_{0,j,k} = \mathbf{u}_{N_x-1,j,k}, \tag{24}$$

$$\mathbf{u}_{i,N_y,k} = \mathbf{u}_{i,1,k}, \quad \mathbf{u}_{i,0,k} = \mathbf{u}_{i,N_y-1,k} \tag{25}$$

for $1 \le i \le N_x$, $1 \le j \le N_y$, $1 \le k \le N_z$.

In (17)–(19), z-derivatives in the cross terms are made one-sided at the grid line $k = 1$. Nevertheless, the semidiscrete approximation is a second order accurate approximation as demonstrated in the following theorem.

THEOREM 1. *The semidiscrete scheme* (17)–(19) *subject to the boundary conditions* (20)–(25) *is a second order accurate approximation of the continuous equation* (10)–(12) *subject to the boundary conditions* (13)–(16).

Proof. See Appendix A. ∎

We will show that the above scheme satisfies an energy estimate. The energy estimate relies on the spatial discretization being self-adjoint and negative definite (elliptic). These properties are stated in three lemmas below. The main stability estimate is stated after the lemmas.

The diagonal scaling S which was used to symmetrize the spatial discretization for the model problem in section 1.1 is related to a weighted scalar product for the unscaled problem. For the three-dimensional elastic wave equation, the appropriate scalar product and norm are

$$(w,v)_h = h^2 \sum_{i=1}^{N_x-1} \sum_{j=1}^{N_y-1} \left(\frac{h}{2} w_{i,j,1} v_{i,j,1} + h \sum_{k=2}^{N_z-1} w_{i,j,k} v_{i,j,k} \right), \quad \|v\|_h^2 = (v,v)_h.$$

The self-adjoint property is expressed in the following lemma.

LEMMA 1. *For all real-valued grid functions* (u^0, v^0, w^0), (u^1, v^1, w^1) *which satisfy the discrete boundary conditions* (20)–(25), *the spatial operator* $(L^{(u)}, L^{(v)}, L^{(w)})$ *is self-adjoint; i.e.,*

$$\begin{aligned} &\left(u^0, L^{(u)}(u^1,v^1,w^1)\right)_h + \left(v^0, L^{(v)}(u^1,v^1,w^1)\right)_h + \left(w^0, L^{(w)}(u^1,v^1,w^1)\right)_h \\ &\quad = \left(u^1, L^{(u)}(u^0,v^0,w^0)\right)_h + \left(v^1, L^{(v)}(u^0,v^0,w^0)\right)_h + \left(w^1, L^{(w)}(u^0,v^0,w^0)\right)_h. \end{aligned} \tag{26}$$

Proof. See Appendix B. ∎

From the self-adjoint property it follows that there exists a conserved quantity.

LEMMA 2. *All real-valued solutions* (u,v,w) *of the semidiscrete scheme* (17)–(19) *subject to the boundary conditions* (20)–(25) *satisfy*

$$\begin{aligned} &\|\rho^{1/2} u_t\|_h^2 + \|\rho^{1/2} v_t\|_h^2 + \|\rho^{1/2} w_t\|_h^2 - (u, L^{(u)}(u,v,w))_h - (v, L^{(v)}(u,v,w))_h \\ &\quad -(w, L^{(w)}(u,v,w))_h = C, \end{aligned} \tag{27}$$

where C *is a constant which depends on the initial data.*

Proof. Lemma 1 gives

$$\begin{aligned} \frac{1}{2}\frac{d}{dt} &\left(\|\rho^{1/2} u_t\|_h^2 + \|\rho^{1/2} v_t\|_h^2 + \|\rho^{1/2} w_t\|_h^2 \right) \\ &= (u_t, L^{(u)}(u,v,w))_h + (v_t, L^{(v)}(u,v,w))_h + (w_t, L^{(w)}(u,v,w))_h \\ &= \frac{1}{2}\left((u_t, L^{(u)}(u,v,w))_h + (v_t, L^{(v)}(u,v,w))_h + (w_t, L^{(w)}(u,v,w))_h \right) \\ &+ \frac{1}{2}\left((u, L^{(u)}(u_t,v_t,w_t))_h + (v, L^{(v)}(u_t,v_t,w_t))_h + (w, L^{(w)}(u_t,v_t,w_t))_h \right) \\ &= \frac{1}{2}\frac{d}{dt}\left((u, L^{(u)}(u,v,w))_h + (v, L^{(v)}(u,v,w))_h + (w, L^{(w)}(u,v,w))_h \right). \end{aligned}$$

Integrating the above relation in time starting at $t = 0$ gives (27) and shows that the constant C depends on the initial data. ∎

To prove that the semidiscrete scheme is stable, we need to show that the conserved quantity in (27) is a norm; i.e., we need to show that the spatial operator is negative definite. In particular, we need to show that the sum of the mixed terms in $(u, L^{(u)})_h$, $(v, L^{(v)})_h$, and $(w, L^{(w)})_h$ (such as $\left(D_0^x w, \mu \widetilde{D_0^z} u\right)_h$) is dominated by the sum of the strictly positive terms (such as $\left(D_+^x w, E_{1/2}^x(\mu) D_+^x w\right)_h$). This is straightforward in the corresponding continuous case and leads to the well-known formula for

the elastic energy. What makes the discrete case more challenging is that all derivatives in the strictly positive terms are discretized by operators such as $D_+^x D_-^x$, while they are discretized by centered differences (such as $D_0^x D_0^y$) in all mixed terms. We have the following.

LEMMA 3. *For all real-valued grid functions (u, v, w) which satisfy the boundary conditions* (20)–(25)*, we have*

(28)
$$\begin{aligned}(u, L^{(u)}(u,v,w))_h + (v, L^{(v)}(u,v,w))_h + (w, L^{(w)}(u,v,w))_h &= -2\|(E^x_{1/2}(\mu))^{1/2} D_+^x u\|_h^2 \\ - 2\|(E^y_{1/2}(\mu))^{1/2} D_+^y v\|_h^2 - 2\|(E^z_{1/2}(\mu))^{1/2} D_+^z w\|_h^2 &- \|\lambda^{1/2}(D_0^x u + D_0^y v + \widetilde{D_0^z} w)\|_h^2 \\ - \|\mu^{1/2}(D_0^x u + D_0^y v)\|_h^2 - \|\mu^{1/2}(\widetilde{D_0^z} v + D_0^y w)\|_h^2 &- \|\mu^{1/2}(\widetilde{D_0^z} u + D_0^x w)\|_h^2 - \frac{h^2}{4} R - B.\end{aligned}$$

The operator $\left(L^{(u)}, L^{(v)}, L^{(w)}\right)$ is negative definite when $\mu > 0$ and $\lambda > 0$. It is semidefinite when $\mu = 0$ and $\lambda > 0$. The remainder term R and the boundary term B are both positive. They are given by

$$\begin{aligned}R = \|\lambda^{1/2} D_+^x D_-^x u\|_h^2 + \|\mu^{1/2} D_+^y D_-^y u\|_h^2 + \|\mu^{1/2} D_+^z D_-^z u\|_{hr}^2 \\ + \|\mu^{1/2} D_+^x D_-^x v\|_h^2 + \|\lambda^{1/2} D_+^y D_-^y v\|_h^2 + \|\mu^{1/2} D_+^z D_-^z v\|_{hr}^2 \\ + \|\mu^{1/2} D_+^x D_-^x w\|_h^2 + \|\mu^{1/2} D_+^y D_-^y w\|_h^2 + \|\lambda^{1/2} D_+^z D_-^z w\|_{hr}^2\end{aligned} \tag{29}$$

and

(30)
$$B = h \sum_{j=1}^{N_y} \sum_{i=1}^{N_x} \Big(\frac{\lambda_{i,j,N_z}}{2} w^2_{i,j,N_z-1} + \frac{\mu_{i,j,N_z}}{2} (u^2_{i,j,N_z-1} + v^2_{i,j,N_z-1}) + h^2 \mu_{i,j,3/2} (D_+^z w_{i,j,1})^2 \Big),$$

respectively.

Note: The reduced scalar product $(u, v)_{hr}$ is similar to the standard scalar product, except that it starts the summation from $k = 2$:

$$(w, v)_{hr} = h^3 \sum_{i=1}^{N_x - 1} \sum_{j=1}^{N_y - 1} \sum_{k=2}^{N_z - 1} w_{i,j,k} v_{i,j,k}, \quad \|v\|_{hr}^2 = (v, v)_{hr}.$$

Proof. The identity (28) is derived in Appendix C. All terms on the right-hand side of (28) are nonpositive when the functions μ and λ are nonnegative. Therefore the operator is at least negative semidefinite. Negative definiteness is proved by showing that

$$(u, L^{(u)}(u,v,w))_h + (v, L^{(v)}(u,v,w))_h + (w, L^{(w)}(u,v,w))_h = 0 \tag{31}$$

implies $u_{i,j,k} = 0$, $v_{i,j,k} = 0$, and $w_{i,j,k} = 0$ at all grid points.

Assume that $\mu_{i,j,k} > 0$ and $\lambda_{i,j,k} > 0$ for all i, j, k and that (31) holds. The right-hand side of (28) is a sum of nonpositive terms. Therefore, each term must be zero to make the sum zero. Hence the third scalar product term on the right-hand side of (28) gives

$$D_+^z w_{i,j,k} = 0, \quad 1 \le i \le N_x - 1, \ \ 1 \le j \le N_y - 1, \ \ 1 \le k \le N_z - 1.$$

Together with the boundary condition $w_{i,j,N_z} = 0$, this gives

$$0 = w_{i,j,N_z} = w_{i,j,N_z-1} = \cdots = w_{i,j,1}.$$

Thus $w_{i,j,k} = 0$ everywhere, except possibly at $k = 0$. Next we show that $u_{i,j,k} = 0$ for all i, j, k except possibly for $k = 0$. The seventh scalar product term on the right-hand side of (28) gives

$$\widetilde{D_0^z} u_{i,j,k} + D_0^x w_{i,j,k} = 0, \quad 1 \le i \le N_x - 1, \ \ 1 \le j \le N_y - 1, \ \ 1 \le k \le N_z - 1. \tag{32}$$

Because $w_{i,j,k} = 0$, (32) gives

$$u_{i,j,N_z} = u_{i,j,N_z-2} = u_{i,j,N_z-4} = \ldots, \tag{33}$$

$$u_{i,j,N_z-1} = u_{i,j,N_z-3} = u_{i,j,N_z-5} = \ldots . \tag{34}$$

The boundary term B contains u^2_{i,j,N_z-1}, which therefore must be zero. Hence, (33) and (34) together with the boundary condition $u_{i,j,N_z} = 0$ give

$$0 = u_{i,j,N_z} = u_{i,j,N_z-1} = \ldots = u_{i,j,1}.$$

We have shown that $u_{i,j,k} = 0$ for all i, j, k except possibly for $k = 0$. The property $v_{i,j,k} = 0$, except possibly for $k = 0$, follows in exactly the same way as for $u_{i,j,k}$ by studying the sixth term on the right-hand side of (28). The possibilities $u_{i,j,0} \neq 0$, $v_{i,j,0} \neq 0$, or $w_{i,j,0} \neq 0$ remain. However, when (u, v, w) is zero for $1 \le k \le N_z$, the boundary conditions (20)–(22) give $u_{i,j,0} = v_{i,j,0} = w_{i,j,0} = 0$. We have now proved that the operator $(L^{(u)}, L^{(v)}, L^{(w)})$ is negative definite when μ and λ are positive functions.

If $\mu = 0$ and $\lambda > 0$, the operator has a nontrivial null space. Take, for example, $u_{i,j,k} = f_{j,k}$, $v_{i,j,k} = g_{i,k}$, and $w_{i,j,k} = 0$, with $f_{j,k}, g_{i,k}$ satisfying $f_{j,N_z} = g_{i,N_z} = 0$ and periodic in the j- and i-directions, respectively, but otherwise arbitrary. Because $\mu = 0$, these functions satisfy the free surface boundary conditions (20)–(22). It is an easy exercise to show that these functions make (28) equal to zero when $\mu = 0$ everywhere. Hence the operator $(L^{(u)}, L^{(v)}, L^{(w)})$ is negative semidefinite when $\mu = 0$ and $\lambda > 0$. ∎

The findings in Lemmas 1–3 are summarized in the following main theorem, showing that the semidiscrete problem is well-posed.

THEOREM 2. *The solution of the semidiscrete scheme* (17)–(19) *subject to the boundary conditions* (20)–(25) *satisfies*

$$\|\rho^{1/2} u_t\|_h^2 + \|\rho^{1/2} v_t\|_h^2 + \|\rho^{1/2} w_t\|_h^2 - (u, L^{(u)}(u,v,w))_h - (v, L^{(v)}(u,v,w))_h - (w, L^{(w)}(u,v,w))_h = C,$$

where C is a constant that depends on the initial data. The quantity

$$-(u, L^{(u)}(u,v,w))_h - (v, L^{(v)}(u,v,w))_h - (w, L^{(w)}(u,v,w))_h$$

is positive definite when $\mu > 0$ and $\lambda > 0$ and is therefore a norm.

2.2. Fully discrete equations. Following the theory in [18], we discretize (17)–(19) in time according to

$$\rho\left(\frac{u^{n+1}-2u^n+u^{n-1}}{\delta_t^2}\right)=L^{(u)}(u^n,v^n,w^n), \tag{35}$$

$$\rho\left(\frac{v^{n+1}-2v^n+v^{n-1}}{\delta_t^2}\right)=L^{(v)}(u^n,v^n,w^n), \tag{36}$$

$$\rho\left(\frac{w^{n+1}-2w^n+w^{n-1}}{\delta_t^2}\right)=L^{(w)}(u^n,v^n,w^n). \tag{37}$$

To simplify the notation, we introduce the weighted ρ-norm

$$(w,v)_\rho = h^2\sum_{i=1}^{N_x-1}\sum_{j=1}^{N_y-1}\left(\frac{h}{2}\rho_{i,j,1}w_{i,j,1}v_{i,j,1}+h\sum_{k=2}^{N_z-1}\rho_{i,j,k}w_{i,j,k}v_{i,j,k}\right),\quad \|v\|_\rho^2=(v,v)_\rho\,.$$

Trivial calculations give

$$\left(w,\rho^{-1}v\right)_\rho=(w,v)_h\,. \tag{38}$$

To show that the fully discrete scheme is energy conserving, we consider the quantity

$$\begin{aligned}C_e(t_{n+1}) &= \left\|D_+^t u^n\right\|_\rho^2+\left\|D_+^t v^n\right\|_\rho^2+\left\|D_+^t w^n\right\|_\rho^2-\left(u^{n+1},\rho^{-1}L^{(u)}(u^n,v^n,w^n)\right)_\rho\\ &\quad-\left(v^{n+1},\rho^{-1}L^{(v)}(u^n,v^n,w^n)\right)_\rho-\left(w^{n+1},\rho^{-1}L^{(w)}(u^n,v^n,w^n)\right)_\rho\\ &= \left\|D_+^t u^n\right\|_\rho^2+\left\|D_+^t v^n\right\|_\rho^2+\left\|D_+^t w^n\right\|_\rho^2\\ &\quad-\left(u^{n+1},D_+^tD_-^t u^n\right)_\rho-\left(v^{n+1},D_+^tD_-^t v^n\right)_\rho-\left(w^{n+1},D_+^tD_-^t w^n\right)_\rho\,.\end{aligned} \tag{39}$$

We have the following energy conservation result for the difference scheme.

THEOREM 3. *The solution computed by the difference scheme* (35)–(37) *together with the boundary conditions* (20)–(25) *satisfies*

$$C_e(t_{n+1})=C_e(t_n);$$

i.e., $C_e(t_n)$ is a conserved quantity for the fully discrete scheme.

Proof. Expanding the square in the term $||D_+^t u^n||_\rho^2$ (and similarly for v and w) gives the identity

$$\begin{aligned}\delta_t^2 C_e(t_{n+1}) &= \|u^{n+1}\|_\rho^2+\|u^n\|_\rho^2-\left(u^{n+1},2u^n+\delta_t^2\rho^{-1}L^{(u)}(u^n,v^n,w^n)\right)_\rho\\ &\quad+\|v^{n+1}\|_\rho^2+\|v^n\|_\rho^2-\left(v^{n+1},2v^n+\delta_t^2\rho^{-1}L^{(v)}(u^n,v^n,w^n)\right)_\rho\\ &\quad+\|w^{n+1}\|_\rho^2+\|w^n\|_\rho^2-\left(w^{n+1},2w^n+\delta_t^2\rho^{-1}L^{(w)}(u^n,v^n,w^n)\right)_\rho\,.\end{aligned} \tag{40}$$

We have

$$u^{n+1}+u^{n-1}=2u^n+\delta_t^2\rho^{-1}L^{(u)}(u^n,v^n,w^n)$$

and corresponding expressions for v and w. Hence,

$$\begin{aligned}\delta_t^2 C_e(t_{n+1}) &= \|u^{n+1}\|_\rho^2 + \|u^n\|_\rho^2 - \left(u^{n+1}, u^{n+1} + u^{n-1}\right)_\rho + \|v^{n+1}\|_\rho^2 + \|v^n\|_\rho^2 \\ &\quad - \left(v^{n+1}, v^{n+1} + v^{n-1}\right)_\rho + \|w^{n+1}\|_\rho^2 + \|w^n\|_\rho^2 - \left(w^{n+1}, w^{n+1} + w^{n-1}\right)_\rho \\ &= \|u^n\|_\rho^2 + \|u^{n-1}\|_\rho^2 - \left(u^{n-1}, 2u^n + \delta_t^2 \rho^{-1} L^{(u)}(u^n, v^n, w^n)\right)_\rho \\ &\quad + \|v^n\|_\rho^2 + \|v^{n-1}\|_\rho^2 - \left(v^{n-1}, 2v^n + \delta_t^2 \rho^{-1} L^{(v)}(u^n, v^n, w^n)\right)_\rho \\ &\quad + \|w^n\|_\rho^2 + \|w^{n-1}\|_\rho^2 - \left(w^{n-1}, 2w^n + \delta_t^2 \rho^{-1} L^{(w)}(u^n, v^n, w^n)\right)_\rho .\end{aligned}$$

The relation (38) gives

$$\left(u^{n-1}, \delta_t^2 \rho^{-1} L^{(u)}(u^n, v^n, w^n)\right)_\rho = \left(u^{n-1}, \delta_t^2 L^{(u)}(u^n, v^n, w^n)\right)_h ,$$

so Lemma 1 yields

$$\begin{aligned}&\left(u^{n-1}, \delta_t^2 \rho^{-1} L^{(u)}(u^n, v^n, w^n)\right)_\rho + \left(v^{n-1}, \delta_t^2 \rho^{-1} L^{(v)}(u^n, v^n, w^n)\right)_\rho \\ &\qquad + \left(w^{n-1}, \delta_t^2 \rho^{-1} L^{(w)}(u^n, v^n, w^n)\right)_\rho \\ &= \left(u^n, \delta_t^2 \rho^{-1} L^{(u)}(u^{n-1}, v^{n-1}, w^{n-1})\right)_\rho + \left(v^n, \delta_t^2 \rho^{-1} L^{(v)}(u^{n-1}, v^{n-1}, w^{n-1})\right)_\rho \\ &\qquad + \left(w^n, \delta_t^2 \rho^{-1} L^{(w)}(u^{n-1}, v^{n-1}, w^{n-1})\right)_\rho .\end{aligned} \tag{41}$$

We conclude that

$$C_e(t_{n+1}) = C_e(t_n);$$

i.e., $C_e(t_n)$ is a conserved quantity for the fully discrete scheme. ∎

To obtain an energy estimate we need to show that $C_e > 0$. This was done in [18] for approximations of the scalar wave equation. We here perform a similar analysis for the scheme (35)–(37). To make the presentation more compact, we introduce the vector notation

$$\begin{aligned}(\mathbf{u}^{n+1}, \mathbf{L}(\mathbf{u}^n))_h =: &(u^{n+1}, L^{(u)}(u^n, v^n, w^n))_h + (v^{n+1}, L^{(v)}(u^n, v^n, w^n))_h \\ &+ (w^{n+1}, L^{(w)}(u^n, v^n, w^n))_h.\end{aligned} \tag{42}$$

As we shall see below, it is natural to study the scaled eigenvalue problem

$$\rho^{-1}\mathbf{L}(\mathbf{w}) = \zeta \mathbf{w}, \tag{43}$$

where $\mathbf{w}$ satisfies the boundary conditions (20)–(25). We know from Lemma 1 that $\mathbf{L}$ is self-adjoint with respect to $(\cdot,\cdot)_h$. Therefore, $\rho^{-1}\mathbf{L}$ is self-adjoint with respect to $(\cdot,\cdot)_\rho$ because

$$(\mathbf{v}, \rho^{-1}\mathbf{L}(\mathbf{w}))_\rho = (\mathbf{v}, \mathbf{L}(\mathbf{w}))_h = (\mathbf{L}(\mathbf{v}), \mathbf{w})_h = (\rho^{-1}\mathbf{L}(\mathbf{v}), \mathbf{w})_\rho.$$

Hence, the eigenvalues of (43) are real and Lemma 3 implies that they are negative, i.e.,

$$-\max_m |\zeta_m| \|\mathbf{w}\|_\rho^2 \le (\mathbf{w}, \rho^{-1}\mathbf{L}(\mathbf{w}))_\rho \le -\min_m |\zeta_m| \|\mathbf{w}\|_\rho^2. \tag{44}$$

We have the following stability result.

THEOREM 4. *If the eigenvalues ζ_m of (43) satisfy the CFL condition*

$$\frac{\delta_t^2}{4}\max_m|\zeta_m| < 1, \tag{45}$$

then the conserved quantity $C_e(t_{n+1})$ is a norm which is bounded from below by

$$C_e(t_{n+1}) \geq \left(1 - \frac{\delta_t^2}{4}\max_m|\zeta_m|\right)\|D_+^t\mathbf{u}^n\|_\rho^2 + \frac{\min_m|\zeta_m|}{4}\|\mathbf{u}^{n+1}+\mathbf{u}^n\|_\rho^2. \tag{46}$$

Proof. Using the vector notation (42), we can write the conserved quantity (39) as follows:

$$C_e(t_{n+1}) = \left\|D_+^t\mathbf{u}^n\right\|_\rho^2 - \left(\mathbf{u}^{n+1},\mathbf{L}(\mathbf{u}^n)\right)_h.$$

Because the operator $\mathbf{L}$ is self-adjoint (Lemma 1),

$$(\mathbf{u}^{n+1},\mathbf{L}(\mathbf{u}^n))_h = \frac{1}{2}(\mathbf{u}^{n+1},\mathbf{L}(\mathbf{u}^n))_h + \frac{1}{2}(\mathbf{u}^n,\mathbf{L}(\mathbf{u}^{n+1}))_h.$$

Furthermore,

$$\begin{aligned}&(\mathbf{u}^{n+1}+\mathbf{u}^n,\mathbf{L}(\mathbf{u}^{n+1}+\mathbf{u}^n))_h - (\mathbf{u}^{n+1}-\mathbf{u}^n,\mathbf{L}(\mathbf{u}^{n+1}-\mathbf{u}^n))_h\\ &\quad = 2(\mathbf{u}^n,\mathbf{L}(\mathbf{u}^{n+1}))_h + 2(\mathbf{u}^{n+1},\mathbf{L}(\mathbf{u}^n))_h,\end{aligned}$$

and $(\mathbf{w},\mathbf{L}(\mathbf{w}))_h = (\mathbf{w},\rho^{-1}\mathbf{L}(\mathbf{w}))_\rho$. Hence,

$$\begin{aligned}\delta_t^2 C_e(t_{n+1}) = ||\mathbf{u}^{n+1}-\mathbf{u}^n||_\rho^2 &- \frac{\delta_t^2}{4}(\mathbf{u}^{n+1}+\mathbf{u}^n,\rho^{-1}\mathbf{L}(\mathbf{u}^{n+1}+\mathbf{u}^n))_\rho\\ &+ \frac{\delta_t^2}{4}(\mathbf{u}^{n+1}-\mathbf{u}^n,\rho^{-1}\mathbf{L}(\mathbf{u}^{n+1}-\mathbf{u}^n))_\rho.\end{aligned} \tag{47}$$

The eigenvalue bound (44) gives

$$\delta_t^2 C_e(t_{n+1}) \geq \left(1 - \frac{\delta_t^2}{4}\max_m|\zeta_m|\right)\|\mathbf{u}^{n+1}-\mathbf{u}^n\|_\rho^2 + \frac{\delta_t^2}{4}\min_m|\zeta_m|\|\mathbf{u}^{n+1}+\mathbf{u}^n\|_\rho^2. \tag{48}$$

Hence, $C_e(t_{n+1})$ is a norm when

$$1 - \frac{\delta_t^2}{4}\max_m|\zeta_m| > 0,$$

i.e., when the CFL condition (45) is satisfied. ∎

2.3. Time step restrictions. In the case of constant ρ, μ, λ, and periodic boundary conditions in all three directions, a von Neumann analysis gives the maximum eigenvalue

$$\zeta_{vN} = \begin{cases} -\dfrac{4}{h^2}\dfrac{4\mu+\lambda}{\rho}, & \lambda < 2\mu,\\ -\dfrac{9}{2h^2}\dfrac{(2\mu+\lambda)^2}{\rho(\mu+\lambda)}, & \lambda \geq 2\mu.\end{cases} \tag{49}$$

(We mention in passing that the largest eigenvalue occurs for the highest wave number on the grid ($\omega h = \pi$) when $\lambda < 2\mu$, while it arises for $\omega h = 2\pi/3$ when $\lambda \geq 2\mu$.

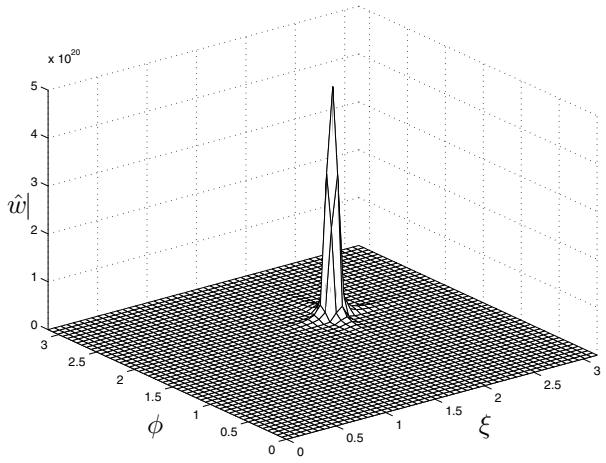


FIG. 1. *Magnitude of the two-dimensional discrete Fourier transform of w at $t = 1.78$, along the $z = 0$ (stress-free) surface calculated with a time step allowed by the von Neumann analysis, which underestimates the largest eigenvalue of the spatial operator. In this calculation, $\rho = 1$, $\mu = 1$, $\lambda = 79$, $h = 0.04$, $\delta_t = 0.95\delta_{tvN}$, and the initial data were given by* (66). *Note that all energy is concentrated around the wave numbers $\omega_x h \approx \omega_y h \approx 2\pi/3$.*

This behavior is different from the corresponding two-dimensional problem, where the largest eigenvalue always happens when $\omega h = \pi$.) If ζ_{vN} is used to estimate the largest eigenvalue $\max_m |\zeta_m|$, we get the time step restriction $\delta_t < \delta_{tvN}$, where

$$\delta_{tvN} = \begin{cases} h\sqrt{\dfrac{\rho}{4\mu+\lambda}} = \dfrac{h}{\sqrt{c_p^2+2c_s^2}}, & c_p < 2c_s, \\ \dfrac{\sqrt{8}h}{3}\dfrac{\sqrt{\rho(\mu+\lambda)}}{2\mu+\lambda} = \dfrac{\sqrt{8}h}{3}\dfrac{\sqrt{c_p^2-c_s^2}}{c_p^2}, & c_p \geq 2c_s. \end{cases} \tag{50}$$

Unfortunately, numerical simulations using a time step smaller but close to the limit (50) become unstable when a stress-free boundary is imposed and the ratio $\nu = c_p/c_s$ is large; see Figure 1.

To estimate how the free-surface boundary condition modifies the time step restriction, we study the stability of the discrete half-plane problem with constant values of ρ, μ, λ. In this approximation, we assume a 2π-periodic solution in the x- and y-directions, expand the grid in the z-direction by taking $N_z \to \infty$, and replace the Dirichlet boundary condition (23) by

$$\lim_{k\to\infty} |\mathbf{u}_{i,j,k}^n| = 0. \tag{51}$$

Several stability definitions for difference approximations are possible, and we refer to [11] for a discussion. Here we use a normal-mode approach and define the half-plane problem to be stable if there are no solutions of the form

$$\mathbf{u}(x_i, y_j, z_k, t_n) = \chi^n e^{\imath(\omega_x x_i + \omega_y y_j)}\hat{\mathbf{u}}_k, \quad \sum_{k=1}^{\infty} |\hat{\mathbf{u}}_k|^2 < \infty, \quad |\chi| > 1, \tag{52}$$

where $\imath = \sqrt{-1}$. For simplicity we assume that $N_x = N_y$ is odd. Then $\omega_x, \omega_y = 0, \pm 1, \pm 2, \ldots, \pm(N_x - 1)/2$.

It is straightforward to perform the stability analysis if we first rewrite our scheme (17)–(19) into an equivalent form, where the one-sided discretization of the cross-derivatives at $k = 1$ are replaced by the centered discretization used for $k \geq 2$, i.e., replace $\widetilde{D_0^z}$ by D_0^z in (17)–(19). We arrive at an equivalent problem by introducing compensating terms in the boundary conditions; see Appendix A. In the case of constant coefficients, the compensated stress-free boundary conditions are

$$D_0^z u_{i,j,1} + D_0^x \left(w_{i,j,1} + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z w_{i,j,1} \right) = 0, \tag{53}$$

$$D_0^z v_{i,j,1} + D_0^y \left(w_{i,j,1} + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z w_{i,j,1} \right) = 0, \tag{54}$$

$$\begin{aligned} \nu^2 D_0^z w_{i,j,1} &+ D_0^x \left((\nu^2 - 2) u_{i,j,1} + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z u_{i,j,1} \right) \\ &+ D_0^y \left((\nu^2 - 2) v_{i,j,1} + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z v_{i,j,1} \right) = 0. \end{aligned} \tag{55}$$

After inserting the ansatz (52) into the modified version of (17)–(19), we arrive at the eigenvalue problem

$$\begin{aligned} \frac{\zeta_{hp}}{c_s^2} \hat{\mathbf{u}}_k := \frac{\chi - 2 + \chi^{-1}}{\delta_t^2 c_s^2} \hat{\mathbf{u}}_k &= -\frac{4}{h^2} \left(\sin^2 \frac{\xi}{2} + \sin^2 \frac{\phi}{2} \right) \hat{\mathbf{u}}_k + D_+^z D_-^z \hat{\mathbf{u}}_k \\ &+ (\nu^2 - 1) \begin{pmatrix} -\frac{4}{h^2} \sin^2 \frac{\xi}{2} & -\frac{1}{h^2} \sin\xi \sin\phi & \frac{\imath}{h} \sin\xi D_0^z \\ -\frac{1}{h^2} \sin\xi \sin\phi & -\frac{4}{h^2} \sin^2 \frac{\phi}{2} & \frac{\imath}{h} \sin\phi D_0^z \\ \frac{\imath}{h} \sin\xi D_0^z & \frac{\imath}{h} \sin\phi D_0^z & D_+^z D_-^z \end{pmatrix} \hat{\mathbf{u}}_k, \end{aligned} \tag{56}$$

where $\xi = \omega_x h$ and $\phi = \omega_y h$ satisfy $-\pi \leq \xi \leq \pi$, $-\pi \leq \phi \leq \pi$. Inserting the ansatz (52) into the boundary conditions (53)–(55) gives

$$D_0^z \hat{u}_1 + \frac{\imath}{h} \sin\xi \left(\hat{w}_1 + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z \hat{w}_1 \right) = 0, \tag{57}$$

$$D_0^z \hat{v}_1 + \frac{\imath}{h} \sin\phi \left(\hat{w}_1 + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z \hat{w}_1 \right) = 0, \tag{58}$$

$$\begin{aligned} \nu^2 D_0^z \hat{w}_1 &+ \frac{\imath}{h} \sin\xi \left((\nu^2 - 2)\hat{u}_1 + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z \hat{u}_1 \right) \\ &+ \frac{\imath}{h} \sin\phi \left((\nu^2 - 2)\hat{v}_1 + (\nu^2 - 1)\frac{h^2}{4} D_+^z D_-^z \hat{v}_1 \right) = 0. \end{aligned} \tag{59}$$

The eigenvalue problem (56) can be solved using the ansatz

$$\hat{\mathbf{u}}_k = \mathbf{U}\kappa^k, \quad \text{where } |\kappa| < 1. \tag{60}$$

Lemma 1 is straightforward to generalize to the half-plane problem, so the spatial operator is self-adjoint, and the generalization of Lemma 3 shows that the spatial operator is negative semidefinite. All eigenvalues ζ_{hp} are therefore real and nonpositive.

Next we study the relation between ζ_{hp} and χ in (56). The roots of the quadratic equation $\chi^2 - (2 - |\zeta_{hp}|\delta_t^2)\chi + 1 = 0$ are given by

$$\chi_{1,2} = 1 - \frac{|\zeta_{hp}|\delta_t^2}{2} \pm \sqrt{\Delta}, \quad \Delta = -|\zeta_{hp}|\delta_t^2 \left(1 - \frac{|\zeta_{hp}|\delta_t^2}{4} \right).$$

If $\Delta < 0$, the roots are complex conjugates. Since the product of the roots equals one, both roots satisfy $|\chi_{1,2}| = 1$. If $\Delta = 0$, $\chi_{1,2} = -1$ is a double root. Finally, if $\Delta > 0$, both roots are real. One root will have magnitude greater than one and one less than one. Hence, the condition $|\chi| > 1$ in the normal-mode ansatz (52) is equivalent to $\Delta > 0$. Conversely, there are no solutions of the form (52) if all eigenvalues ζ_{hp} satisfy

$$-|\zeta_{hp}|\delta_t^2\left(1 - \frac{|\zeta_{hp}|\delta_t^2}{4}\right) \leq 0, \quad \text{i.e.,} \quad \frac{\delta_t^2}{4}|\zeta_{hp}| \leq 1.$$

Hence the normal-mode stability definition leads to the same type of time step restriction as in the energy method (Theorem 4), and we can use the most negative eigenvalue ζ_{hp} to approximate the eigenvalue in (45). This approximation will lead to a more restrictive time step limitation than in the von Neumann analysis if there are any eigenvalues ζ_{hp} such that

$$|\zeta_{hp}| > |\zeta_{vN}|.$$

Inserting (60) into (56) gives

$$Q\mathbf{U} = 0, \tag{61}$$

$$Q = \begin{pmatrix} \begin{matrix}-4(\nu^2\sin^2\frac{\xi}{2} + \sin^2\frac{\phi}{2}) + \\ \kappa - 2 + \kappa^{-1} - \tilde{\zeta}\end{matrix} & -(\nu^2-1)\sin\xi\sin\phi & (\nu^2-1)\imath\sin\xi(\kappa-\kappa^{-1}) \\ -(\nu^2-1)\sin\xi\sin\phi & \begin{matrix}-4(\sin^2\frac{\xi}{2} + \nu^2\sin^2\frac{\phi}{2}) + \\ \kappa - 2 + \kappa^{-1} - \tilde{\zeta}\end{matrix} & (\nu^2-1)\imath\sin\phi(\kappa-\kappa^{-1}) \\ (\nu^2-1)\imath\sin\xi(\kappa-\kappa^{-1}) & (\nu^2-1)\imath\sin\phi(\kappa-\kappa^{-1}) & \begin{matrix}-4(\sin^2\frac{\xi}{2} + \sin^2\frac{\phi}{2}) + \\ \nu^2(\kappa - 2 + \kappa^{-1}) - \tilde{\zeta}\end{matrix} \end{pmatrix},$$

where $\tilde{\zeta} = \zeta_{hp}h^2/c_s^2$. Multiply (61) by κ, and let

$$P(\tilde{\zeta}, \kappa, \xi, \phi, \nu) = 0 \tag{62}$$

be the corresponding characteristic equation. Here P is a cubic polynomial in $\tilde{\zeta}$ and a polynomial of degree six in κ. For fixed ν, ξ, and ϕ there are six roots κ for each $\tilde{\zeta}$. The following lemma is a standard result (see, e.g., [12]), which we here formulate for our discretization of the elastic wave equation.

LEMMA 4. *The characteristic equation $P = 0$ has six roots κ_l. For $\tilde{\zeta} < -|\tilde{\zeta}_{vN}| = -|\zeta_{vN}|h^2/c_s^2$, three of these roots have $|\kappa| < 1$ and three have $|\kappa| > 1$.*

Proof. A polynomial of degree six has six roots (counting multiplicity). If any κ is such that $|\kappa| = 1$, then $\kappa = e^{i\alpha}$ for some real α, and (62) becomes identical to the relation obtained in the von Neumann analysis of the fully periodic problem. We know that there are no eigenvalues with magnitude greater than $|\tilde{\zeta}_{vN}|$ in this case. Therefore there can be no κ on the unit circle when $\tilde{\zeta} < -|\tilde{\zeta}_{vN}|$. Second, take $\phi = \xi = 0$. It is not hard to see that the characteristic equation $P = 0$ becomes

$$\left[\kappa^2 - (2+\tilde{\zeta})\kappa + 1\right]\left[\kappa^2 - (2+\tilde{\zeta})\kappa + 1\right]\left[\kappa^2 - \left(2 + \frac{\tilde{\zeta}}{\nu^2}\right)\kappa + 1\right] = 0.$$

Therefore the six roots κ satisfy the pairwise relations $\kappa_1\kappa_4 = 1$, $\kappa_2\kappa_5 = 1$, and $\kappa_3\kappa_6 = 1$. Since no root can be on the unit circle when $\tilde{\zeta} < -|\tilde{\zeta}_{vN}|$, there must be three roots inside the unit circle and three roots outside of it. Furthermore, the roots

κ are smooth functions of ϕ, ξ, ν, and $\tilde{\zeta}$. Because they cannot move across the unit circle when $\tilde{\zeta} < -|\tilde{\zeta}_{vN}|$, the roots are always divided into these two groups for all values of ϕ, ξ, ν, and for any $\tilde{\zeta}$ such that $\tilde{\zeta} < -|\tilde{\zeta}_{vN}|$. □

It follows from Lemma 4 that the general solution of (56) subject to the boundary condition (51) is

$$\hat{\mathbf{u}}_k = C_1\mathbf{U}_1\kappa_1^k + C_2\mathbf{U}_2\kappa_2^k + C_3\mathbf{U}_3\kappa_3^k, \quad |\kappa_l| < 1, \quad l = 1, 2, 3, \tag{63}$$

where $\mathbf{U}_l$ are the eigenvectors corresponding to κ_l, $l = 1, 2, 3$, respectively. For each $\tilde{\zeta} < -|\tilde{\zeta}_{vN}|$, $\mathbf{U}_l$ is the null vector of the linear system (61) when the root κ_l is substituted for κ.

Inserting the general solution (63) into the stress-free boundary conditions (57)–(59) leads to a homogeneous linear system for the coefficients C_1, C_2, and C_3:

$$A \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} = \mathbf{0}, \tag{64}$$

where $A = A(\tilde{\zeta}, \xi, \phi, \nu)$ is a three by three matrix. There are nontrivial solutions of (64) if and only if $\det A = 0$. If (64) has a nontrivial solution $(C_1, C_2, C_3)^T$ for some $\tilde{\zeta}$, then the corresponding ζ_{hp} is an eigenvalue of (56).

Since the algebra involved in forming $\det A$ is rather complicated, we have resolved to calculate the roots of $\det A = 0$ numerically. The determinant depends on four parameters, where $\nu = c_p/c_s$ is a material constant and the angles ξ, ϕ satisfy $-\pi \leq \xi, \phi \leq \pi$. For each fixed ν, we need to find the angles ξ, ϕ that give the most negative solution $\tilde{\zeta}$ of $\det A = 0$. A straightforward approach is to discretize ξ, ϕ on a fine mesh:

$$\xi_p = -\pi + p\frac{2\pi}{N_\xi}, \quad p = 0, 1, 2, \ldots, N_\xi,$$
$$\phi_q = -\pi + q\frac{2\pi}{N_\phi}, \quad q = 0, 1, 2, \ldots, N_\phi.$$

At each mesh point $\det A$ is a complex-valued function of the real variable $\tilde{\zeta}$, and we need to consider only $\tilde{\zeta} < -|\tilde{\zeta}_{vN}|$, since only such eigenvalues can restrict the time step beyond the von Neumann limit. At each point (ξ_p, η_q), we apply a numerical root-finding routine to locate the most negative solution $\tilde{\zeta}_{p,q}$ of $\det A = 0$. We then use $\min_{p,q} \tilde{\zeta}_{p,q}$ as an approximation of the most negative solution $\tilde{\zeta}$ corresponding to ν. The fundamental operation when applying a numerical root-finding routine is to evaluate $\det A$ at a given value of $\tilde{\zeta}$, which can be broken down into the following steps:

1. Solve the characteristic equation (62) for κ. Select the three roots with $|\kappa_l| < 1$;
2. find the three eigenvectors $\mathbf{U}_l$ by solving (61) for each κ_l, $l = 1, 2, 3$;
3. form the matrix A by inserting (63) into (57)–(59);
4. compute the determinant of A.

Using the numerical root-finding procedure outlined above, we calculated the ratio between the largest stable time step for the half-plane problem with a stress-free boundary and the largest stable time step for the fully periodic case; see Figure 2. The numerical root-finding procedure located the largest eigenvalue $|\zeta_{hp}|$ at $\phi = \xi = 2\pi/3$. Hence, the spatial frequencies $\omega_x h = \omega_y h = 2\pi/3$ should grow the fastest if the time

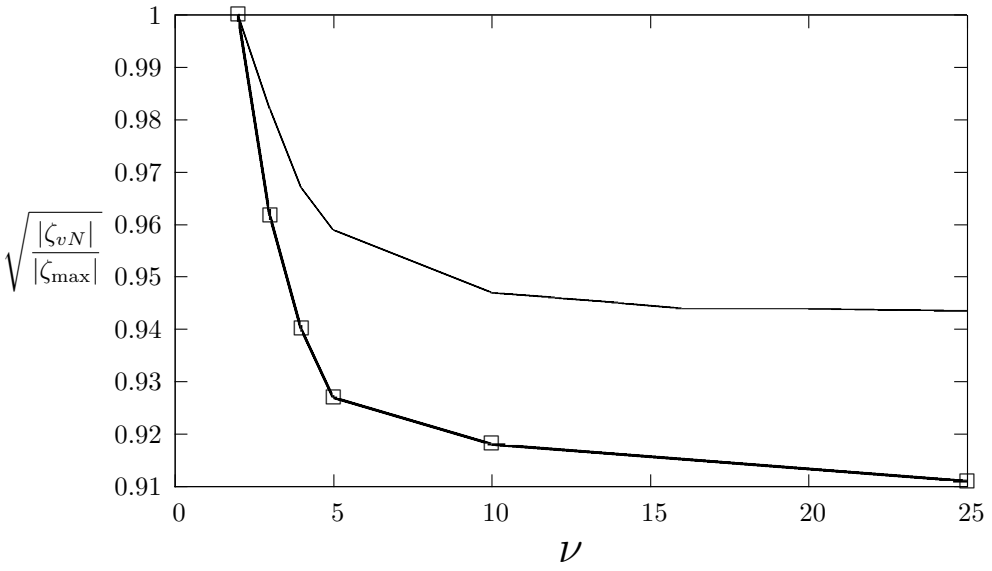


FIG. 2. *Ratio between the maximum stable time step for the half-plane problem with a free surface and that of the fully periodic problem. The line with boxes corresponds to the three-dimensional problem, and the two-dimensional case is shown with a solid line.*

step exceeds the stability limit and ν is large. This prediction was confirmed by spatially Fourier transforming an unstable numerical solution; see Figure 1.

For large ν, the solutions of $\det A = 0$ corresponding to the largest $|\tilde{\zeta}|$ were numerically found to occur when κ_1 is real with $-1 < \kappa_1 < 0$ and $\kappa_2 = \bar{\kappa_3}$ with $-1 < \mathrm{Re}(\kappa_{2,3}) < 0$. Thus, the eigenfunction corresponding to the largest eigenvalue oscillates in the z-direction with two different frequencies: the fastest frequency on the mesh $|\kappa_1|^k(-1)^k$ and more slowly $|\kappa_2|^k(\exp(\pm\imath \arg\kappa_2))^k$, where $\arg(\kappa_2) \approx 2\pi/3$ for large ν. This boundary layer behavior has been observed in numerical solutions when the time step exceeds the stability limit.

Note that the limitations imposed on the time step by the stress-free boundary are very moderate even for extreme ν values (most solid materials occuring in nature have $c_p/c_s \leq 3$). As ν gets large, the largest stable time step for the half-plane problem tends to a factor exceeding 0.91 of that for the fully periodic problem. Our practical experience with the time-stepping algorithm on bounded domains with variable coefficients and a free surface boundary condition on one side indicates that it is stable when the half-plane problem with constant coefficients is stable, using the smallest time step obtained by evaluating c_p and c_s at all grid points. Hence, we can handle all values of c_p/c_s by reducing the time step by less than 9% compared to the von Neumann value. This makes our method practically useful for all isotropic materials.

The additional time-step restriction due to the free surface boundary condition indicates that there are numerical surface waves which travel faster than any volume waves on the grid. In the continuous problem, Rayleigh (surface) waves always have a phase velocity which is smaller than c_s. Hence, it is likely that the numerical phase velocity for Rayleigh waves will depend on the grid resolution in terms of the number of grid points per wavelength. Numerical experiments along these lines are presented in section 4.1.

We also analyzed the two-dimensional version of the scheme by assuming that the solution does not depend on y. Here a von Neumann analysis of the doubly periodic case (ρ, μ, and λ constant) gives a time step restriction

$$\delta_t < \frac{h\sqrt{\rho}}{\sqrt{3\mu+\lambda}} = \frac{h}{\sqrt{c_p^2+c_s^2}}. \tag{65}$$

The stability restriction on the time step with the free surface boundary condition can be obtained using the above root-finding procedure with $\phi = 0$. The results are given in Figure 2 together with the three-dimensional case. When ν becomes large, the largest stable time step for the half-plane problem tends to a factor exceeding 0.94 of that for the fully periodic case (i.e., 6% smaller). As in the three-dimensional problem, the largest eigenvalue occurs for the spatial frequency $\omega_x h = 2\pi/3$.

2.4. Numerical tests of the scheme. In order to test the implementation of our method we first ran a number of computations without forcing with decreasing grid size h to evaluate the discrete energy C_e as a function of time. We took $\mu = 0.16$, $\lambda = 0.49$, $\rho = 1$, and started the computations with the initial data in spherical coordinates:

$$\begin{aligned} \mathbf{U}_0(r) = \nabla\left(\frac{P_{10}(r)}{r}\right), \quad \mathbf{U}_1(r) = -c_p\nabla\left(\frac{P'_{10}(r)}{r}\right), \\ r = \sqrt{(x-2)^2+(y-1.5)^2+(z-1.5)^2}, \end{aligned} \tag{66}$$

where $P_{10}(\xi)$ is the four times continously differentiable function

$$P_{10}(\xi) = \begin{cases} 0, & \xi <= 0, \\ 1024\xi^5\left(1-5\xi+10\xi^2-10\xi^3+5\xi^4-\xi^5\right), & 0<\xi<1, \\ 0, & \xi \geq 1. \end{cases} \tag{67}$$

(We note in passing that $\mathbf{u}(r,t) = \nabla(P_{10}(r-c_pt)/r)$ is an analytic solution of the free space problem.) We impose a stress-free boundary condition at $z = 0$ and enforce zero displacement conditions on all other boundaries. The size of the computational domain was $a = 4$, $b = 3$, and $c = 3$. Since there is no forcing, the discrete energy $C_e(t_n)$ should remain constant. The energy in the continuous problem is often decomposed into its kinematic and potential components

$$E(t) = K(t) + U(t),$$

where

$$\begin{aligned} K(t) =& \frac{1}{2}\int_\Omega \rho(u_t^2+v_t^2+w_t^2)\,d\Omega, \\ U(t) =& \frac{1}{2}\int_\Omega \lambda(u_x+v_y+w_z)^2 + 2\mu(u_x^2+v_y^2+w_z^2) \\ &+ \mu\left((u_y+v_x)^2+(u_z+w_x)^2+(v_z+w_y)^2\right)\,d\Omega. \end{aligned}$$

In the absence of forcing, $E(t) = \text{const}$. By dividing (47) by δ_t^2 it is straightforward to see that

$$C_e(t_{n+1}) = 2E(t_{n+1/2}) + \mathcal{O}(h^2).$$

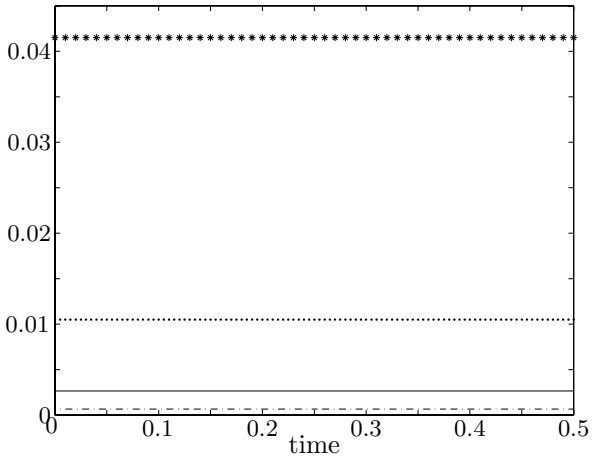


FIG. 3. *Time evolution of the relative error in the discrete energy* $(C_e(t) - 2E(t))/2E(t)$ *for different grid sizes. The discrete energy is conserved to within roundoff errors for all cases. As can be seen, the discrete energy converges towards the continuous value at the expected* $\mathcal{O}(h^2)$ *rate. Here* $h = 0.04\,(*)$, $0.02\,(\cdot)$, $0.01\,(-)$, *and* $0.005\,(-\cdot)$.

Hence, the discrete energy C_e should not only be conserved in time, but its value should also converge to $2E(t)$ as the grid is refined. Both of these properties are confirmed by our calculations; see Figure 3.

As a second test of our implementation, we check the order of accuracy of the scheme using the *method of analytical solutions* (also known as *twilight-zone* forcing [5]). The idea is to construct forcing functions $\mathbf{f}$ and $\mathbf{g}$ so that the solution of the test problem becomes a known function $\mathbf{u}^{\text{true}}(\mathbf{x}, t)$. We then solved the test problem using our implementation of the method and compared our numerical results to the known solution on a succession of finer grids in order to check the convergence properties. Our constructed solution was

$$\begin{aligned} u^{\text{true}}(\mathbf{x}, t) &= \sin(\omega(x - ct)) \sin(\omega y) \sin(\omega z), \\ v^{\text{true}}(\mathbf{x}, t) &= \sin(\omega x) \sin(\omega(y - ct)) \sin(\omega z), \\ w^{\text{true}}(\mathbf{x}, t) &= \sin(\omega x) \sin(\omega y) \sin(\omega(z - ct)), \end{aligned}$$

where ω and c are constants. The material properties were chosen to vary smoothly according to

$$\begin{aligned} \mu(\mathbf{x}) &= 1 + \cos^2(\pi x) \cos^2(\pi y) \cos^2(\pi z), \\ \lambda(\mathbf{x}) &= 1 + \sin^2(\pi x) \sin^2(\pi y) \sin^2(\pi z), \\ \rho(\mathbf{x}) &= 1. \end{aligned}$$

A normal stress condition was imposed on the $z = 0$ surface, and inhomogeneous Dirichlet conditions were imposed on all other boundaries. The computational domain had sizes $a = 2$, $b = 2$, and $c = 2$. A number of calculations with increasingly fine grid spacing were run, and the errors were evaluated in the discrete max-norm. (The discrete max-norm of a vector grid function $\mathbf{v}_h = (u_h, v_h, w_h)$ is defined as $||\mathbf{v}_h||_\infty = \max(\max_{i,j,k} |u_h|, \max_{i,j,k} |v_h|, \max_{i,j,k} |w_h|)$.) As expected we obtained second order convergence when both the forcing and the solution are smooth; see Table 1. Nonsmooth forcings and solutions will be discussed in section 3.

TABLE 1
Errors in max-norm for decreasing h and smooth analytical solution $\mathbf{u}^{\text{true}}$. Convergence rate indicates second order convergence. Here $c = 1$ and $\omega = 2\pi$.

	$t = 1$	
h	$\|\|\mathbf{v}_h - \mathbf{u}^{\text{true}}\|\|_\infty$	Rate
0.04	0.04331	
0.02	0.01062	4.079
0.01	0.002654	4.00
0.005	0.0006627	4.00

3. Singular source terms. In seismic wave propagation the source term is often applied at a point, along a line, or over a surface in three-dimensional space. Sources along lines or surfaces are commonly decomposed into a number of point sources distributed along the corresponding line or surface:

$$\mathbf{f}(\mathbf{x},t) = \sum_r \mathbf{f}_r^{(F)}(\mathbf{x},t) + \sum_r \mathbf{f}_r^{(M)}(\mathbf{x},t). \tag{68}$$

Two types of point sources occur in seismic applications. Point forces ($\mathbf{f}_r^{(F)}$) are, for example, used to model internal forcings due to volcanic eruptions or external forcings applied to the free surface

$$\mathbf{f}_r^{(F)}(\mathbf{x},t) = g_r(t)\mathbf{F}_r\delta(\mathbf{x}-\mathbf{x}_r), \tag{69}$$

where $\delta(\mathbf{x})$ is the Dirac distribution and $\mathbf{F}_r$ is a constant vector. The second type of point source is the point moment (or double couple), denoted by $\mathbf{f}_r^{(M)}$ in (68). Point moments are often used to model earthquakes and explosions [4] and are of the form

$$\mathbf{f}_r^{(M)}(\mathbf{x},t) = g_r(t)\,\mathfrak{M}_r \cdot \nabla\delta(\mathbf{x}-\mathbf{x}_r), \tag{70}$$

where $\nabla\delta(\mathbf{x})$ is the gradient of the Dirac distribution, and $\mathfrak{M}_r$ is a constant symmetric tensor.

Each term in (68) is applied at a location (x_r, y_r, z_r), and it is desirable to make this location independent of the grid so that the numerical modeling can be made as accurate as possible and no artifacts are generated by "stair stepping" the point sources along a smooth line or surface in three-dimensional space. Due to the singular nature of point sources, we can only expect the numerical solution to converge away from the location of the sources. Furthermore, we can expect that different numerical techniques are necessary for handling the two types of sources, since the point force depends on the Dirac distribution while the point moment depends on its gradient, which is a more singular function.

The analyses of Waldén [26] and Tornberg and Engquist [24] demonstrate that it is possible to derive regularized approximations of the Dirac distribution and its gradient, which result in pointwise convergence of the solution away from the sources. Based on these analyses, we define a hat function

$$\delta_{\text{hat}}(x) = \frac{1}{h}\begin{cases} 1-|x|/h, & |x|<h, \\ 0, & \text{elsewhere}, \end{cases} \tag{71}$$

and use $\delta_{\text{hat}}(x-x_r)\delta_{\text{hat}}(y-y_r)\delta_{\text{hat}}(z-z_r)$ to approximate $\delta(\mathbf{x})$ in (69). To approximate the gradient of a Dirac distribution, we start from the piecewise cubic function

$$\delta_{\text{cube}}(x) = \frac{1}{h}\begin{cases} 1-|x/h|/2-|x/h|^2+|x/h|^3/2, & |x|<h, \\ 1-11|x/h|/6+|x/h|^2-|x/h|^3/6, & h\le|x|<2h, \\ 0, & \text{elsewhere}. \end{cases} \tag{72}$$

We then use

$$\begin{pmatrix} \delta'_{\rm cube}(x-x_r)\delta_{\rm hat}(y-y_r)\delta_{\rm hat}(z-z_r) \\ \delta_{\rm hat}(x-x_r)\delta'_{\rm cube}(y-y_r)\delta_{\rm hat}(z-z_r) \\ \delta_{\rm hat}(x-x_r)\delta_{\rm hat}(y-y_r)\delta'_{\rm cube}(z-z_r) \end{pmatrix}$$

to approximate the Cartesian components of $\nabla\delta(\mathbf{x}-\mathbf{x}_r)$ in (70). Note that neither (71) nor (72) need to be aligned with the grid.

3.1. Spatial regularity. To study the relation between smoothness of the time function $g(t)$ in the source term and smoothness in the space of the solution, we analyze the related problem of the scalar wave equation with a singular source term. In particular, we study the problem on an infinite domain with the forcing term applied at the point $(0,0,0)$ with homogeneous initial data:

$$\begin{aligned} p_{tt} &= \nabla^2 p + g(t)\delta(\mathbf{x}), \quad \mathbf{x}\in R^3, \quad t\geq 0, \\ p(\mathbf{x},0) &= p_t(\mathbf{x},0) = 0. \end{aligned}$$

The Fourier transform of this equation is

$$\frac{d^2\hat{p}}{dt^2} = -(k_x^2+k_y^2+k_z^2)\hat{p} + g(t), \quad t\geq 0, \tag{73}$$

$$\hat{p}(k_x,k_y,k_z,0) = \hat{p}_t(k_x,k_y,k_z,0) = 0, \tag{74}$$

where the Fourier transform is given by

$$\hat{p}(k_x,k_y,k_z,t) = \int\int\int p(x,y,z,t)e^{-i(xk_x+yk_y+zk_z)}\,dx\,dy\,dz.$$

Equations (73)–(74) are solved by

$$\hat{p}(k_x,k_y,k_z,t) = \begin{cases} \int_0^t \int_0^\tau g(\tau')\,d\tau'\,d\tau, & k=0, \\ \frac{1}{k}\left(\sin(kt)\int_0^t \cos(k\tau)g(\tau)\,d\tau - \cos(kt)\int_0^t \sin(k\tau)g(\tau)\,d\tau\right), & k>0, \end{cases} \tag{75}$$

where $k=\sqrt{k_x^2+k_y^2+k_z^2}$. If $g(t)$ is continuously differentiable, we can integrate (75) by parts:

$$\begin{aligned} \hat{p}(k_x,k_y,k_z,t) = \frac{1}{k^2}\Bigg(g(t) - \cos(kt)g(0) - \sin(kt)\int_0^t \sin(k\tau)g'(\tau)d\tau \\ - \cos(kt)\int_0^t \cos(k\tau)g'(\tau)d\tau\Bigg). \end{aligned}$$

By assuming that $g(t)$ has compact support, i.e., $g(t)\equiv 0$ for $t\leq 0$ and $t\geq T$, we get

$$\begin{aligned} \hat{p}(k_x,k_y,k_z,t) = \frac{1}{k^2}\Bigg(-\sin(kt)\int_0^t \sin(k\tau)g'(\tau)\,d\tau \\ -\ \cos(kt)\int_0^t \cos(k\tau)g'(\tau)\,d\tau\Bigg), \quad t\geq T. \end{aligned}$$

The Fourier transform decays as $1/k^2$. We can continue integrating by parts as long as $g(t)$ is sufficiently differentiable, gaining one order of k for each integration. This

shows that the solution $p(\mathbf{x},t)$ has a Fourier transform that decays as $1/k^q$ for $t > T$ if $g(t)$ has compact support and is $q-1$ times differentiable in time. Furthermore, $\hat{p}$ is bounded because the singularity at $k = 0$ is removable:

$$\lim_{k\to 0} \hat{p}(k_x, k_y, k_z, t) = \int_0^t (t-\tau) g(\tau)\, d\tau = \int_0^t \int_0^\tau g(\tau')\, d\tau'\, d\tau.$$

Therefore,

$$\int\int\int (1 + k^{2q'}) |\hat{p}|^2\, dk_x\, dk_y\, dk_z < \infty$$

for $q' < q - 3/2$. By the Sobolev lemma [10], p can be identified with a function that has m continous derivatives for $m < q' - 3/2 < q - 3$. We conclude that for $t > T$ the solution $p(\mathbf{x},t)$ will have m continuous derivatives if g is compactly supported and smooth. Here m can be made arbitrarily large by choosing $g(t)$ sufficiently smooth.

If $g(t)$ does not tend to zero for large t, the solution will remain singular at the location of the point source but will be smooth away from it.

3.2. Free space solutions. Let the free space Green's (dyadic) function for the elastic wave equation in a homogeneous material be $\mathfrak{G}(\mathbf{x},t)$; see [4]. Assuming homogeneous initial data, the analytical solution of the elastic wave equation due to a source function $\mathbf{f}(\mathbf{x},\mathbf{t})$ follows as the space and time convolution between the Green's function and the source term

$$\mathbf{u}(\mathbf{x},t) = \int_t \int_\Omega \mathbf{f}(\mathbf{x}',t') \cdot \mathfrak{G}(\mathbf{x}-\mathbf{x}', t-t')\, d\mathbf{x}'\, dt'.$$

In the special case when the source is a point force, the spatial convolution becomes trivial due to the Dirac distributions in $\mathbf{f}_r^{(F)}$, and the expression reduces to a time integral over t'. Near the source, the solution behaves like $1/|\mathbf{x}-\mathbf{x}_r|$. A closed form solution can be obtained when the time integration can be performed analytically, for instance, when $g(t)$ is a polynomial function.

For a point moment source term $\mathbf{f}_r^{(M)}$, the analytical solution can be written

$$\begin{aligned}\mathbf{u}(\mathbf{x},t) &= \int_0^t \int_\Omega g_r(t') \left(\mathfrak{M}_r \cdot \nabla\delta(\mathbf{x}'-\mathbf{x}_r)\right) \cdot \mathfrak{G}(\mathbf{x}-\mathbf{x}', t-t')\, d\mathbf{x}'\, dt' \\ &= \int_0^t g_r(t')\, \mathfrak{M}_r : \nabla\mathfrak{G}(\mathbf{x}-\mathbf{x}_r, t-t')\, dt',\end{aligned}$$

where the colon represents the tensor contraction over two indices. Near the point moment, the solution behaves like $1/|\mathbf{x}-\mathbf{x}_r|^2$, so it is more singular than in the point force case.

To investigate how the numerical solution converges when the source function is singular, we ran a number of tests with point forces and point moments using the time function $g(t) = P_{10}(t)$ defined in (67). This function has compact support in $0 \le t \le 1$ and is four times continuously differentiable. We took a computational domain with $a = 2$, $b = 2$, $c = 2$, and used the material parameters $\rho = 1$, $\lambda = 0.32$, $\mu = 0.16$. Dirichlet boundary conditions were enforced on all boundaries, but the boundaries have no influence on the solution until $t > 1.25$ since $c_p = 0.8$ and the point sources were centered at $\mathbf{x}_r = (1,1,1)$. The errors were measured at two different times in discrete max-, 2-, and 1-norms. Since the analytical solution is singular

TABLE 2

Relative error in the numerical solution of the free space problem at time $t = 0.5$ (singular solution) due to a point force (top) and a point moment (bottom), measured in max-, 2-, and 1-norms. Here $\mathbf{v}_h$ and $\mathbf{u}$ denote the numerical and analytical solutions, respectively.

	Point force					
h	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_\infty}{\|\|\mathbf{v}_h\|\|_\infty}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_2}{\|\|\mathbf{v}_h\|\|_2}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_1}{\|\|\mathbf{v}_h\|\|_1}$	Rate^∞	Rate^2	Rate^1
0.04	0.04833	0.08293	0.1011			
0.02	0.04108	0.05174	0.03248	1.176	1.602	3.113
0.01	0.03936	0.03525	0.009970	1.043	1.467	3.257
0.005	0.03894	0.02470	0.002955	1.010	1.427	3.373
	Point moment					
h	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_\infty}{\|\|\mathbf{v}_h\|\|_\infty}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_2}{\|\|\mathbf{v}_h\|\|_2}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_1}{\|\|\mathbf{v}_h\|\|_1}$	Rate^∞	Rate^2	Rate^1
0.04	0.3051	0.2805	0.2272			
0.02	0.3208	0.2760	0.1154	0.9509	1.016	1.969
0.01	0.3253	0.2769	0.05759	0.9871	0.9967	2.003
0.005	0.3264	0.2782	0.02872	0.9970	0.9953	2.005

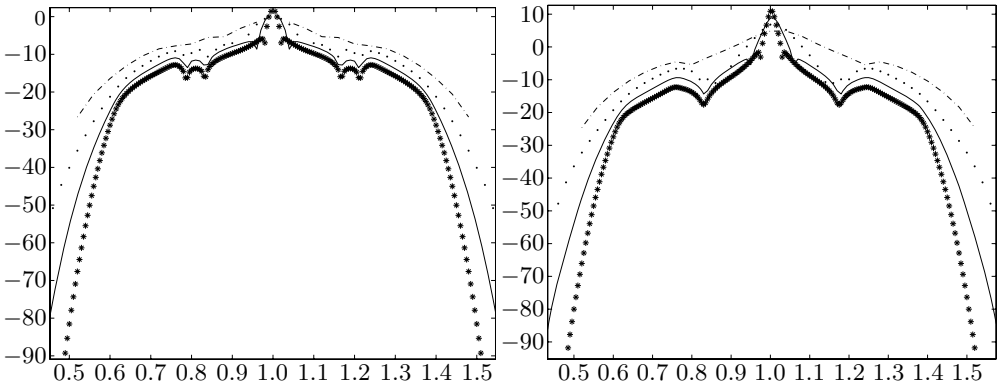


FIG. 4. *The 2-logarithm of the error along a line going through the source point for a point force (left) and a point moment source (right), both located at $x = 1$. Note that the error decays as $\mathcal{O}(h^2)$ away from the source but not near it. Near the source, the error is about $2^{11} \approx 2000$ times larger for the point moment than for the point force. The grid sizes were $h = 0.04\,(-\cdot)$, $0.02\,(\cdot)$, $0.01\,(-)$, and $0.005\,(*)$.*

at the point where the source applies, that point was excluded from the calculation of the norms. (The 2- and 1-norms for a vector grid function $\mathbf{u}$ are defined as $||\mathbf{u}||_2^2 = h^3 \sum_{i,j,k}(|u_{i,j,k}|^2+|v_{i,j,k}|^2+|w_{i,j,k}|^2)$ and $||\mathbf{u}||_1 = h^3 \sum_{i,j,k}(|u_{i,j,k}|+|v_{i,j,k}|+|w_{i,j,k}|)$.) First we evaluated the errors at $t = 0.5$ when $g(t) > 0$; see Table 2. As expected we did not achieve second order convergence because the solution of the continuous problem is singular. Also note that the convergence rate is slower for the point moment source than in the less singular point force case. In Figure 4, we show the errors as a function of the distance from the singularity. Away from the singularity, the errors are smooth in space and decay like $\mathcal{O}(h^2)$ as the grid size tends to zero. However, near the source the errors do not decay as the grid is refined, and this explains the convergence numbers in Table 2. Second, we evaluated the errors at $t = 1.2$, when $g(t) = 0$; see Table 3. After the source term has vanished the solution becomes smooth everywhere, and our results show the proper second order convergence rate in accordance with theory.

We remark that in the point moment source case it is important to use the δ'_{cube} approximation in the gradient of the Dirac distribution, as opposed to δ'_{hat}. Otherwise

TABLE 3

Relative error in the numerical solution of the free space problem at time $t = 1.2$ (smooth solution) due to a point force (top) and a point moment (bottom), measured in max-, 2-, and 1-norms. Here $\mathbf{v}_h$ and $\mathbf{u}$ denote the numerical and analytical solutions, respectively.

	Point force					
h	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_\infty}{\|\|\mathbf{v}_h\|\|_\infty}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_2}{\|\|\mathbf{v}_h\|\|_2}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_1}{\|\|\mathbf{v}_h\|\|_1}$	Rate$^\infty$	Rate^2	Rate^1
0.04	0.04516	0.03984	0.04122			
0.02	0.01180	0.01001	0.01025	3.831	3.984	4.021
0.01	0.003023	0.002512	0.002560	3.907	3.988	4.004
0.005	0.0007592	0.0006287	0.0006400	3.983	4.000	4.00
	Point moment					
h	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_\infty}{\|\|\mathbf{v}_h\|\|_\infty}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_2}{\|\|\mathbf{v}_h\|\|_2}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_1}{\|\|\mathbf{v}_h\|\|_1}$	Rate$^\infty$	Rate^2	Rate^1
0.04	0.1170	0.1016	0.09981			
0.02	0.03400	0.02762	0.02681	3.440	3.678	3.724
0.01	0.008872	0.007109	0.006855	3.833	3.885	3.908
0.005	0.002244	0.001793	0.001724	3.961	3.972	3.985

TABLE 4

Relative error in the numerical solution of Lamb's problem at $t = 0.5$ (top) (when the solution is singular) and at $t = 1.1$ (bottom) (when the solution is smooth), measured in max-, 2-, and 1-norms. Here $\mathbf{v}_h$ and $\mathbf{u}$ denote the numerical and analytical solutions, respectively.

	$t = 0.5$					
h	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_\infty}{\|\|\mathbf{v}_h\|\|_\infty}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_2}{\|\|\mathbf{v}_h\|\|_2}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_1}{\|\|\mathbf{v}_h\|\|_1}$	Rate$^\infty$	Rate^2	Rate^1
0.04	0.02797	0.08631	0.2007			
0.02	0.01758	0.05312	0.1102	1.591	1.625	1.821
0.01	0.01547	0.04002	0.05028	1.136	1.327	2.192
0.005	0.01696	0.03696	0.02305	0.9121	1.083	2.181
	$t = 1.1$					
h	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_\infty}{\|\|\mathbf{v}_h\|\|_\infty}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_2}{\|\|\mathbf{v}_h\|\|_2}$	$\frac{\|\|\mathbf{v}_h-\mathbf{u}\|\|_1}{\|\|\mathbf{v}_h\|\|_1}$	Rate$^\infty$	Rate^2	Rate^1
0.04	0.2892	0.3081	0.3686			
0.02	0.1082	0.1186	0.1408	2.673	2.598	2.618
0.01	0.03138	0.03496	0.04175	3.448	3.392	3.372
0.005	0.008189	0.009194	0.01100	3.832	3.802	3.795

the convergence rate will be slower than second order in the grid size (example not shown to conserve space).

3.3. Half spaces and Lamb's problem. Point forcing on the boundary of a half space is referred to as Lamb's problem [20]. Analytical solutions for the three-dimensional problem have been presented by a number of authors with different degrees of applicability. For the case of a point force directed normal to the free surface $z = 0$, the general solution can be found in [22] or [9]. To test the accuracy of the numerical solutions, we performed a grid refinement study on a computational domain with sizes $a = 4$, $b = 4$, $c = 2$, enforcing a free surface boundary condition along $z = 0$ and Dirichlet conditions on all other boundaries. We assumed a Poisson material with $\rho = 1$, $\mu = 1$, and $\lambda = 1$, i.e., $c_p/c_s = \sqrt{3}$, and used the same time function $g(t)$ as in the free space case. In this experiment, the point force was applied at $\mathbf{x}_r = (2, 2, 0)$, so the Dirichlet boundaries should not affect the solution until $t > 1.15$. The error in the numerical solution was evaluated both at $t = 0.5$, when the solution of the continuous problem is singular, and at $t = 1.1$, when the solution is smooth. We report only the error along the free surface, because the analytical solution is difficult to evaluate in the interior of the domain. As in the free space problem, we observe second order convergence only when the solution is smooth in space; see Table 4.

4. Applications and extensions of the method.

4.1. Surface waves. The elastodynamic equations together with the stress-free boundary condition admit solutions in the form of surface waves, i.e., waves propagating along the surface with amplitude decaying exponentially away from the surface. For the homogeneous two-dimensional half-plane problem in $z \geq 0$, these solutions are commonly referred to as Rayleigh waves and have the form

$$u(x,z,t) = A\left(e^{-\eta_p \omega z} - \left(1 - \frac{c_r^2}{2c_s^2}\right) e^{-\eta_s \omega z}\right) \sin(\omega(c_r t - x)), \tag{76}$$

$$w(x,z,t) = A\left(1 - \frac{c_r^2}{c_p^2}\right)^{1/2} \left(-e^{-\eta_p \omega z} + \left(1 - \frac{c_r^2}{2c_s^2}\right)^{-1} e^{-\eta_s \omega z}\right) \cos(\omega(c_r t - x)), \tag{77}$$

where

$$\eta_p = \left(1 - \frac{c_r^2}{c_p^2}\right)^{1/2}, \quad \eta_s = \left(1 - \frac{c_r^2}{c_s^2}\right)^{1/2}.$$

Here c_r is the phase velocity of the wave, which is the real root of the Rayleigh equation

$$\left(2 - \frac{c_r^2}{c_s^2}\right)^2 - 4\left(1 - \frac{c_r^2}{c_p^2}\right)^{1/2} \left(1 - \frac{c_r^2}{c_s^2}\right)^{1/2} = 0, \quad 0 < c_r < c_s.$$

The waves described by (76)–(77) are nondispersive; i.e., c_r is independent of ω. However, the discretization introduces errors that can be interpreted as a numerical dispersion relation where the phase velocity depends on the resolution on the grid. The numerical dispersion relation for our interior difference stencil coincides with previous central difference schemes which were analyzed by Cohen [7]. For surface waves, the numerical dispersion relation provides the numerical phase velocity c_r^* as a function of the resolution ωh, which often is expressed in terms of the number of grid points per wavelength

$$\mathrm{PPW} = \frac{2\pi}{\omega h}.$$

Since it is very complicated to analytically derive the numerical dispersion relation for surface waves, we instead investigate the relation by numerical experiments using a two-dimensional version of our method. A free surface condition was imposed at $z = 0$, and periodic boundary conditions were used in the x-direction. We enforced (76)–(77) as initial data, which contains only a single spatial frequency ω. Hence, the numerical solution should essentially advect the initial data with a modified phase velocity c_r^*. We determined c_r^* by visually inspecting the solution along the surface at time $t = 1/c_r$ and comparing the positions of the numerical and analytical solutions; see Table 5. Note that the visual inspection is not very precise when the solution is poorly resolved on the grid (PPW < 5), so these results should be interpreted accordingly. Despite this uncertainty, it is clear that the numerical phase velocity increases rapidly as ωh approaches $2\pi/3$ and $\nu \geq 3$. It is interesting to note that this value of ωh coincides with the spatial frequency of the fast surface waves which determine the stability limit of the time step; cf. section 2.3.

TABLE 5

Numerical dispersion relation for the finite difference scheme applied to Rayleigh waves. The table shows the ratio between the estimated phase velocity in the numerical solution and its continuous value, using different number of grid points per wavelength (PPW) and ν.

	c_r^*/c_r			
PPW	$\nu = 2$	$\nu = 3$	$\nu = 5$	$\nu = 10$
40	1.0028	1.0065	1.017	1.055
20	1.011	1.022	1.052	1.12
10	1.031	1.06	1.11	1.2
8	1.043	1.083	1.13	1.5
6	1.049	1.095	1.35	1.63
5	1.07	1.11	1.4	1.72
4	1.095	1.14	1.65	1.78
3.5	1.16	1.4	2.76	2.9

4.2. Nonreflecting boundary conditions. When modeling seismic events such as the simplified earthquake in section 4.3, it is desirable to truncate the computational domain without causing significant amounts of artificial reflections. Many different methods, including absorbing, nonreflecting, and perfectly matching techniques have been proposed in the literature. Here we will use the first order nonreflecting boundary conditions developed by Clayton and Engquist [6]. The well-known idea behind these boundary conditions is to impose a differential equation on the boundary which allows wave propagation only in the outward direction. For boundaries with $x = \text{const}$, the boundary conditions are

$$u_t = \pm c_p u_x,\ v_t = \pm c_s v_x,\ w_t = \pm c_s w_x, \tag{78}$$

where the positive signs are taken for the lower boundary $x = 0$ and the negative signs for the upper boundary $x = a$. Similar advection equations are imposed at boundaries with $y = \text{const}$ or $z = \text{const}$.

Away from edges in the computational domain, we have found that the box scheme discretization [6] of the boundary condition (78) works well. At the edges of the domain, i.e., where two nonreflecting boundaries meet, Clayton and Engquist suggested applying the nonreflecting boundary condition in a diagonal direction. However, we have found that imposing compatibility conditions along the edges results in a more rubust method which also is easier to implement. We examplify the compatibility conditions on the edge where $x = 0$ and $y = 0$. Along the boundary $x = 0$, we impose (78) (with the positive sign). The corresponding boundary conditions along $y = 0$ are

$$u_t = c_s u_y,\ v_t = c_p v_y,\ w_t = c_s w_y, \quad y = 0,\ 0 \le x \le a,\ 0 \le z \le c,\ t \ge 0.$$

Equating the time derivatives along the edge gives

$$\begin{aligned} c_p u_x &= c_s u_y, \\ c_s v_x &= c_p v_y, \quad y = 0,\ x = 0,\ 0 \le z \le c,\ t \ge 0, \\ c_s w_x &= c_s w_y. \end{aligned}$$

Similar relations can easily be derived for the other edges.

4.3. A simplified earthquake. The Pacific Earthquake Engineering Center and the Southern California Earthquake Center have defined a set of seismic model problems in an effort to evaluate and validate wave propagation software [8]. We have computed solutions to several of these problems, but in order to save space we report

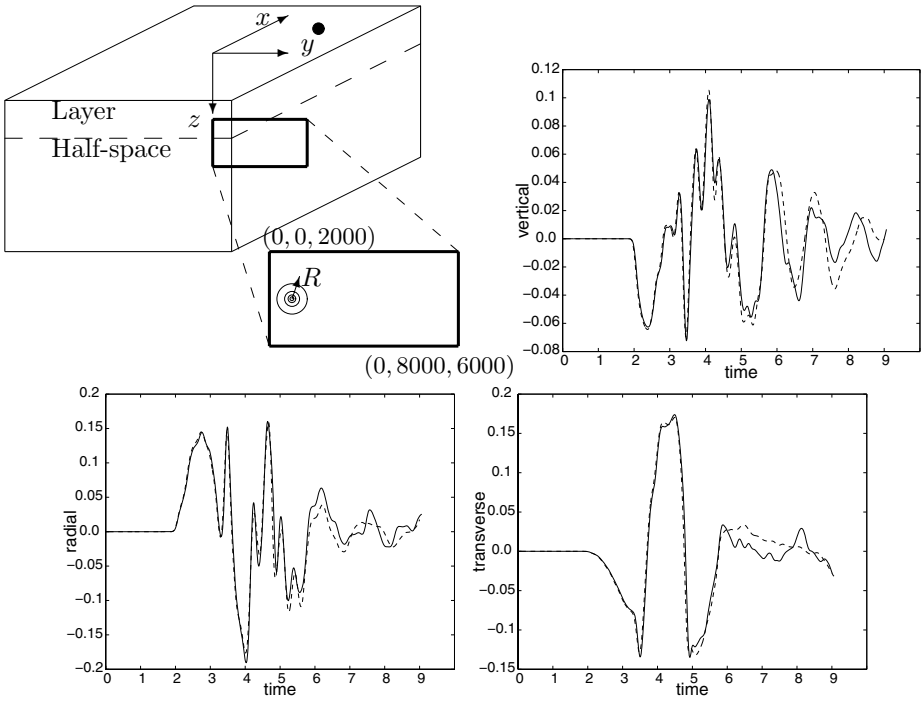


FIG. 5. *The computational domain and fault surface for the simplified earthquake problem LOH.2 (upper left). The • indicates the measurement station, and the magnified plane shows the fault surface, where the slip starts at the hypocenter indicated by concentric circles. Our results are shown with solid lines for the vertical (top right), radial (bottom left), and transverse (bottom right) velocity components, and the dashed lines are the results from the UCSB code; see* [8].

only our results for problem LOH.2, which models a simplified earthquake with slip on an extended fault surface; see Figure 5. The material in this model consists of a layer over a half-space, where the layer extends from depth $z = 0$ to $z = 1000$. The velocities and density in the layer are $c_p = 4000$, $c_s = 2000$, $\rho = 2600$. The half-space $z \geq 1000$ has the material properties $c_p = 6000$, $c_s = 3464$, $\rho = 2700$.

The slip on the extended fault is modeled by distributing point moment sources on a regular grid with size δ_s (which is independent of the grid size h) over the fault surface $x = 0$, $0 \leq y \leq 8000$, $2000 \leq z \leq 6000$. In this case, the fault slips by a constant amount in the y-direction, which means that the Cartesian components of the moment tensor $\mathfrak{M}_r$ in each source term (70) equal

$$\mathfrak{M}_r = \delta_s^2 \mu S_0 \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad S_0 = 1.$$

The modeled earthquake starts at the hypocenter $\mathbf{x}_H = (0, 1000, 4000)$, and the rupture propagates along the fault surface with a uniform rupture velocity of 3000. The propagation of the rupture is modeled by letting the source time function $g_r(t)$ depend on the distance between the hypocenter and the location of each source

$$R_r = |\mathbf{x}_r - \mathbf{x}_H|.$$

The time dependence of source number r is

$$g_r(t) = \begin{cases} 0, & t < R_r/3000, \\ 1 - \left(1 + \dfrac{\tau_r(t)}{T}\right) e^{-\tau_r(t)/T}, & t \geq R_r/3000, \end{cases} \qquad \tau_r(t) = t - R_r/3000,$$

where $T = 0.1$ is related to the rise time of the slip, i.e., how quickly the fault slips at each fixed point along the fault surface.

In our calculation, the extent of the computational domain was $-15000 \leq x \leq 15000$, $-15000 \leq y \leq 15000$, $0 \leq z \leq 17000$, and nonreflecting boundary conditions were imposed on all boundaries except at $z = 0$, where a free surface condition was enforced. The grid size was $h = 50$, corresponding to about 1.23×10^8 grid points, and 1742 time steps were taken to reach time $t = 9$. We discretized the fault surface with $\delta_s = 100$, giving 3200 point moment sources. Results for this problem are available from a number of finite difference and finite element codes [8]. To compare our results, we recorded the time evolution of the velocity (i.e., the time derivative of the displacement) at a number of stations along a line on the free surface. Since all codes predicted similar results, we show only the comparison with the UCSB code (using notation from [8]). This code solves the elastic wave equation as a first order system in velocity-stress formulation using a staggered grid finite difference method. Since the source time functions $g_r(t)$ trigger high frequency motions which are not resolvable on the mesh, the results from both our code and the UCSB code were low-pass filtered in time using a Gaussian with filter width $\sigma = 0.05$. In Figure 5 we compare solutions at a station located at $\mathbf{x} = (6000, 8000, 0)$. Velocities are given in a cylindrical coordinate system (radial, transverse, up) with the origin at $(0, 0, 0)$. Note that the nonreflecting boundary conditions affects the solution only after $t \approx 5$ and that our results compare especially well with the other code before that time. One way of determining the accuracy of the solution after $t \approx 5$ would be to repeat the simulation on a larger domain, but the computational cost was too great to perform that experiment.

5. Conclusions. We have described a stable, second order accurate finite difference method for the elastic wave equation in second order formulation subject to a stress-free boundary condition on a flat surface. We have proven that the method is stable even when the coefficients are discontinuous in space, as long as $\mu > 0$, $\lambda > 0$, and $\rho > 0$ at all grid points. The stability limit on the time step has been studied in detail, and we have shown that all values of $c_p/c_s > \sqrt{2}$ can be handled if the time step is reduced by 9% compared to the von Neumann value. We have also described a way to discretize point forces and moments on the mesh so that the solution becomes second order accurate away from the singularity in the solution.

In seismic applications it is common to have water (e.g., a lake or an ocean) in parts of the domain. Only compressional (P) waves can travel through water, and the acoustic wave propagation can be modeled by setting $\mu = 0$ in the elastic wave equation. We have generalized our scheme to handle the mixed elastic/acoustic case, and this scheme was used as part of a simulation effort coordinated by the U.S. Geological Survey to model ground motions during the great 1906 San Francisco earthquake [23]. Our results showed good agreement with other codes and measured Mercalli intensities. More details will be described in a forthcoming paper [1].

Future plans include generalizing our embedded boundary technique for the scalar wave equation [19, 17, 16] to the elastic wave equation. In the seismic application, embedded boundaries will allow us to include effects of topography and more accurately treat internal material discontinuities. We are also exploring generalizations to fourth order accuracy and curvilinear coordinates.

Appendix A. Accuracy (Theorem 1). We will prove the accuracy of the semidiscrete equations by showing that they are equivalent to another approximation which clearly is second order accurate. In particular, we want to analyze the accuracy of the spatial discretization (17)–(19) at the $z = 0$ boundary, where the free surface boundary condition is applied. At this boundary, the operator $\widetilde{D_0^z}$ simplifies to D_+^z, which would appear to give only a first order accurate difference formula. However, we proceed to show that this difference formula, in combination with the discrete free surface boundary condition, indeed results in a second order approximation.

We start by eliminating the ghost points above the free surface from the semidiscrete system (17)–(19), subject to the boundary conditions (20)–(22). To save space, we go through only the details for (17) subject to (20). The terms in $L^{(u)}$ that contain z-differences on the $z = 0$ grid line are

$$T_{i,j} =: D_-^z \left(\mu_{i,j,3/2} D_+^z u_{i,j,1}\right) + D_0^x \left(\lambda_{i,j,1} D_+^z w_{i,j,1}\right) + D_+^z \left(\mu_{i,j,1} D_0^x w_{i,j,1}\right).$$

The free surface boundary condition (20) gives

$$\mu_{i,j,1/2} D_+^z u_{i,j,0} = -\mu_{i,j,3/2} D_+^z u_{i,j,1} - 2\mu_{i,j,1} D_0^x w_{i,j,1}.$$

Hence,

(79)
$$T_{i,j} = \frac{2}{h}\left[\mu_{i,j,3/2} D_+^z u_{i,j,1} + \mu_{i,j,1} D_0^x w_{i,j,1}\right] + D_0^x \left(\lambda_{i,j,1} D_+^z w_{i,j,1}\right) + D_+^z \left(\mu_{i,j,1} D_0^x w_{i,j,1}\right).$$

We compare the spatial discretization to a fully centered scheme where the terms in $L^{(u)}$ that contain z-differences on the $z = 0$ grid line:

$$\widetilde{T}_{i,j} =: D_-^z \left(\mu_{i,j,3/2} D_+^z u_{i,j,1}\right) + D_0^x \left(\lambda_{i,j,1} D_0^z w_{i,j,1}\right) + D_0^z \left(\mu_{i,j,1} D_0^x w_{i,j,1}\right). \tag{80}$$

We can perturb the free surface boundary condition (20) by a second order term

$$\frac{1}{2}\left(\mu_{i,j,3/2} D_+^z u_{i,j,1} + \mu_{i,j,1/2} D_+^z u_{i,j,0}\right) + \mu_{i,j,1} D_0^x w_{i,j,1} = h^2 R_{i,j}. \tag{81}$$

The resulting spatial discretization will be second order accurate as long as R is a difference operator which is bounded independently of h for smooth functions. We will determine R such that (80) subject to (81) is equivalent to (79). The boundary condition (81) gives

$$\mu_{i,j,1/2} D_+^z u_{i,j,0} = -\mu_{i,j,3/2} D_+^z u_{i,j,1} - 2\mu_{i,j,1} D_0^x w_{i,j,1} + 2h^2 R_{i,j}. \tag{82}$$

Using (82), (80) can be written

$$\begin{aligned}\widetilde{T}_{i,j} = \frac{2}{h}\left[\mu_{i,j,3/2} D_+^z u_{i,j,1} + \mu_{i,j,1} D_0^x w_{i,j,1}\right] + D_0^x \left(\lambda_{i,j,1} D_0^z w_{i,j,1}\right) \\ + D_0^z \left(\mu_{i,j,1} D_0^x w_{i,j,1}\right) + 2hR_{i,j}.\end{aligned}$$

Hence, $T = \widetilde{T}$ if

$$\begin{aligned}&D_0^x \left(\lambda_{i,j,1} D_+^z w_{i,j,1}\right) + D_+^z \left(\mu_{i,j,1} D_0^x w_{i,j,1}\right) \\ &\quad = D_0^x \left(\lambda_{i,j,1} D_0^z w_{i,j,1}\right) + D_0^z \left(\mu_{i,j,1} D_0^x w_{i,j,1}\right) + 2hR_{i,j}.\end{aligned}$$

We have

$$D_0^z w = D_+^z w - \frac{h}{2} D_+^z D_-^z w,$$

which gives

$$R_{i,j} = \frac{1}{4} D_0^x (\lambda_{i,j,1} D_+^z D_-^z w_{i,j,1}) + \frac{1}{4} D_+^z D_-^z (\mu_{i,j,1} D_0^x w_{i,j,1}).$$

Similar calculations show that the boundary conditions (21) and (22) can be perturbed by second order terms to account for the difference between a fully centered and a one-sided spatial discretization in $L^{(v)}$ and $L^{(w)}$, respectively.

This proves that the semidiscrete approximation (17)–(19) subject to the boundary conditions (20)–(22) is second order accurate. ∎

Note. Inserting the expression for $R_{i,j}$ into (81) shows that the fully centered approximation couples all ghost points ($k = 0$) along the free surface. Hence, using this formulation would require a linear system to be solved to obtain the ghost point values at each time step. As we have demonstrated, the same solution can be obtained without solving a linear system by using our one-sided formula on the boundary.

Appendix B. Self-adjointness of the spatial operator (Lemma 1). It is straightforward to show the following summation by parts identities:

(83)
$$(w, D_-^z v)_h = -(D_+^z w, v)_h - \frac{h^2}{2} \sum_{i,j} (w_{i,j,2} v_{i,j,1} + w_{i,j,1} v_{i,j,0}) + h^2 \sum_{i,j} w_{i,j,N_z} v_{i,j,N_z-1},$$

(84)
$$(w, \widetilde{D_0^z} v)_h = -(\widetilde{D_0^z} w, v)_h - h^2 \sum_{i,j} w_{i,j,1} v_{i,j,1} + \frac{h^2}{2} \sum_{i,j} \left(w_{i,j,N_z-1} v_{i,j,N_z} + w_{i,j,N_z} v_{i,j,N_z-1} \right),$$

where $\sum_{i,j} = \sum_{i=1}^{N_x-1} \sum_{j=1}^{N_y-1}$. Since the solution satisfies periodic boundary conditions in the x- and y-directions, we have

$$(w, D_-^x v)_h = -(D_+^x w, v)_h, \quad (w, D_0^x v)_h = -(D_0^x w, v)_h, \tag{85}$$

$$(w, D_-^y v)_h = -(D_+^y w, v)_h, \quad (w, D_0^y v)_h = -(D_0^y w, v)_h. \tag{86}$$

Consider the three terms in the left-hand side of (26): LHS := I + II + III,

$$\mathrm{I} = \left(u^0, L^{(u)}(u^1, v^1, w^1)\right)_h, \quad \mathrm{II} = \left(v^0, L^{(v)}(u^1, v^1, w^1)\right)_h,$$
$$\mathrm{III} = \left(w^0, L^{(w)}(u^1, v^1, w^1)\right)_h.$$

Applying the summation by parts identities (83)–(86) on the first term gives

$$\begin{aligned} \mathrm{I} = &- \left(D_+^x u^0, E_{1/2}^x (2\mu + \lambda) D_+^x u^1\right)_h - \left(D_+^y u^0, E_{1/2}^y (\mu) D_+^y u^1\right)_h \\ &- \left(D_+^z u^0, E_{1/2}^z (\mu) D_+^z u^1\right)_h - \left(D_0^x u^0, \lambda D_0^y v^1 + \lambda \widetilde{D_0^z} w^1\right)_h \\ &- \left(D_0^y u^0, \mu D_0^x v^1\right)_h - \left(\widetilde{D_0^z} u^0, \mu D_0^x w^1\right)_h + B^{(u)}, \end{aligned} \tag{87}$$

where the boundary terms are

$$\begin{aligned} B^{(u)} = &- \frac{h^2}{2} \sum_{i,j} \left(u^0_{i,j,2} \mu_{i,j,3/2} D^z_+ u^1_{i,j,1} + u^0_{i,j,1} \mu_{i,j,1/2} D^z_+ u^1_{i,j,0} \right) \\ &- h^2 \sum_{i,j} u^0_{i,j,1} \mu_{i,j,1} D^x_0 w^1_{i,j,1} + h^2 \sum_{i,j} u^0_{i,j,N_z} \mu_{i,j,N_z-1/2} D^z_+ u^1_{i,j,N_z-1} \\ &+ \frac{h^2}{2} \sum_{i,j} \left(u^0_{i,j,N_z-1} \mu_{i,j,N_z} D^x_0 w^1_{i,j,N_z} + u^0_{i,j,N_z} \mu_{i,j,N_z-1} D^x_0 w^1_{i,j,N_z-1} \right). \end{aligned}$$

The homogeneous Dirichlet boundary condition (23) gives

$$u^0_{i,j,N_z} = 0, \quad D^x_0 w^1_{i,j,N_z} = 0.$$

Hence, the third and fourth terms in $B^{(u)}$ vanish. To analyze the first term, we note that

$$u^0_{i,j,2} = u^0_{i,j,1} + h D^z_+ u^0_{i,j,1}.$$

Therefore,

$$\begin{aligned} B^{(u)} = -\frac{h^2}{2} \sum_{i,j} u^0_{i,j,1} \left(\mu_{i,j,3/2} D^z_+ u^1_{i,j,1} + \mu_{i,j,1/2} D^z_+ u^1_{i,j,0} + 2\mu_{i,j,1} D^x_0 w^1_{i,j,1} \right) \\ - \frac{h^3}{2} \sum_{i,j} \mu_{i,j,3/2} D^z_+ u^0_{i,j,1} D^z_+ u^1_{i,j,1}. \end{aligned} \tag{88}$$

The first term in (88) vanishes because of the free surface boundary condition (20), and we arrive at

$$B^{(u)} = -\frac{h^3}{2} \sum_{i,j} \mu_{i,j,3/2} D^z_+ u^0_{i,j,1} D^z_+ u^1_{i,j,1}.$$

The second term in LHS can be analyzed in the same way, giving

$$\begin{aligned} \mathrm{II} = &- \left(D^x_+ v^0, E^x_{1/2}(\mu) D^x_+ v^1 \right)_h - \left(D^y_+ v^0, E^y_{1/2}(2\mu+\lambda) D^y_+ v^1 \right)_h \\ &- \left(D^z_+ v^0, E^z_{1/2}(\mu) D^z_+ v^1 \right)_h - \left(D^x_0 v^0, \mu D^y_0 u^1 \right)_h \\ &- \left(D^y_0 v^0, \lambda D^x_0 u^1 + \lambda \widetilde{D^z_0} w^1 \right)_h - \left(\widetilde{D^z_0} v^0, \mu D^y_0 w^1 \right)_h + B^{(v)}, \end{aligned} \tag{89}$$

where

$$B^{(v)} = -\frac{h^3}{2} \sum_{i,j} \mu_{i,j,3/2} D^z_+ v^0_{i,j,1} D^z_+ v^1_{i,j,1}.$$

For the third term in LHS, we get

$$\begin{aligned} \mathrm{III} = &- \left(D^x_+ w^0, E^x_{1/2}(\mu) D^x_+ w^1 \right)_h - \left(D^y_+ w^0, E^y_{1/2}(\mu) D^y_+ w^1 \right)_h \\ &- \left(D^z_+ w^0, E^z_{1/2}(2\mu+\lambda) D^z_+ w^1 \right)_h - \left(D^x_0 w^0, \mu \widetilde{D^z_0} u^1 \right)_h \\ &- \left(D^y_0 w^0, \mu \widetilde{D^z_0} v^1 \right)_h - \left(\widetilde{D^z_0} w^0, \lambda D^x_0 u^1 + \lambda D^y_0 v^1 \right)_h + B^{(w)}, \end{aligned} \tag{90}$$

where

$$B^{(w)} = -\frac{h^3}{2}\sum_{i,j}(2\mu_{i,j,3/2}+\lambda_{i,j,3/2})D_+^z w^0_{i,j,1} D_+^z w^1_{i,j,1}.$$

After applying the same summation by parts rules to the right-hand side of (26), it is straightforward to verify that the right-hand side equals the left-hand side. ∎

Appendix C. Ellipticity of the spatial operator (Lemma 3). We will mimic the construction of the energy in the continuous case by exploring the identity

$$D_-^x E_{1/2}^x(\mu) D_+^x u = D_0^x\left(\mu D_0^x u\right) - \frac{h^2}{4} D_+^x D_-^x\left(\mu D_+^x D_-^x u\right) \tag{91}$$

in the periodic x- and y-directions. The problem is not periodic in the z-direction. We will use the following summation-by-parts form of the above identity instead ($N = N_z$ in this appendix)

$$\begin{aligned}\left(u, D_-^z E_{1/2}^z(\mu) D_+^z u\right)_h &= -\left(\widetilde{D_0^z} u, \mu \widetilde{D_0^z} u\right)_h - \frac{h^2}{4}\left(D_+^z D_-^z u, \mu D_+^z D_-^z u\right)_{hr} \\ &+ h^2 \sum_{i,j}\left(-\frac{1}{2}\mu_{i,j,1/2} u_{i,j,1} D_+^z u_{i,j,0} - \frac{1}{2}\mu_{i,j,3/2} u_{i,j,1} D_+^z u_{i,j,1}\right. \\ &\left. + \frac{\mu_{i,j,N}}{2} u_{i,j,N-1} D_+^z u_{i,j,N-1} + \frac{\mu_{i,j,N-1}}{2} u_{i,j,N} D_+^z u_{i,j,N-1}\right).\end{aligned} \tag{92}$$

We obtain, by use of (91) in the periodic directions,

$$\begin{aligned}L^{(u)}(u,v,w) &= 2D_-^x\left(E_{1/2}^x(\mu) D_+^x u\right) + D_-^z\left(E_{1/2}^z(\mu) D_+^z u\right) \\ &+ D_0^x\left(\lambda(D_0^x u + D_0^y v + \widetilde{D_0^z} w)\right) + D_0^y\left(\mu(D_0^y u + D_0^x v)\right) + \widetilde{D_0^z}\left(\mu D_0^x w\right) \\ &- \frac{h^2}{4}\left(D_+^x D_-^x(\lambda D_+^x D_-^x u) + D_+^y D_-^y(\mu D_+^y D_-^y u)\right), \\ L^{(v)}(u,v,w) &= 2D_-^y\left(E_{1/2}^y(\mu) D_+^y v\right) + D_-^z\left(E_{1/2}^z(\mu) D_+^z v\right) \\ &+ D_0^y\left(\lambda(D_0^x u + D_0^y v + \widetilde{D_0^z} w)\right) + D_0^x\left(\mu(D_0^y u + D_0^x v)\right) + \widetilde{D_0^z}\left(\mu D_0^y w\right) \\ &- \frac{h^2}{4}\left(D_+^x D_-^x(\mu D_+^x D_-^x v) + D_+^y D_-^y(\lambda D_+^y D_-^y v)\right), \\ L^{(w)}(u,v,w) &= 2D_-^z\left(E_{1/2}^z(\mu) D_+^z w\right) + D_-^z\left(E_{1/2}^z(\lambda) D_+^z w\right) \\ &+ \widetilde{D_0^z}\left(\lambda(D_0^x u + D_0^y v)\right) + D_0^x\left(\mu(\widetilde{D_0^z} u + D_0^x w)\right) + D_0^y\left(\mu(\widetilde{D_0^z} v + D_0^y w)\right) \\ &- \frac{h^2}{4}\left(D_+^x D_-^x(\mu D_+^x D_-^x w) + D_+^y D_-^y(\mu D_+^y D_-^y w)\right).\end{aligned}$$

Identities (92) and (84) give

$$\begin{aligned}(u, L^{(u)})_h = &-2(D_+^x u, E_{1/2}^x(\mu) D_+^x u)_h - (\widetilde{D_0^z} u, \mu \widetilde{D_0^z} u)_h \\ &- \left(D_0^x u, \lambda(D_0^x u + D_0^y v + \widetilde{D_0^z} w)\right)_h - \left(D_0^y u, \mu(D_0^y u + D_0^x v)\right)_h \\ &- \left(\widetilde{D_0^z} u, \mu D_0^x w\right)_h - \frac{h^2}{4}\left[(D_+^x D_-^x u, \lambda D_+^x D_-^x u)_h\right. \\ &\left. + (D_+^y D_-^y u, \mu D_+^y D_-^y u)_h + (D_+^z D_-^z u, \mu D_+^z D_-^z u)_{hr}\right] + T_1^{(u)} + T_N^{(u)},\end{aligned}$$

where $T_1^{(u)}$ and $T_N^{(u)}$ are the boundary terms that correspond to the boundary at $k = 1$ and at $k = N$, respectively. The periodic directions do not contribute with any boundary terms as seen from (85) and (86). We have

$$\begin{aligned} T_1^{(u)} &= h^2 \sum_{i,j} \left(-\frac{1}{2}\mu_{i,j,1/2} u_{i,j,1} D_+^z u_{i,j,0} - \frac{1}{2}\mu_{i,j,3/2} u_{i,j,1} D_+^z u_{i,j,1} - u_{i,j,1}\mu_{i,j,1} D_0^x w_{i,j,1} \right) \\ &= h^2 \sum_{i,j} u_{i,j,1} \left(-\frac{1}{2}\mu_{i,j,1/2} D_+^z u_{i,j,0} - \frac{1}{2}\mu_{i,j,3/2} D_+^z u_{i,j,1} - \mu_{i,j,1} D_0^x w_{i,j,1} \right). \end{aligned}$$

It follows directly from the free surface boundary condition (20) that $T_1^{(u)} = 0$. The boundary terms at $k = N$ are given by

$$\begin{aligned} T_N^{(u)} = h^2 \sum_{i,j} \Big(&\frac{\mu_{i,j,N-1}}{2} u_{i,j,N} D_0^x w_{i,j,N-1} + \frac{\mu_{i,j,N} u_{i,j,N-1}}{2} D_0^x w_{i,j,N} \\ &+ \frac{\mu_{i,j,N}}{2} u_{i,j,N-1} D_-^z u_{i,j,N} + \frac{\mu_{i,j,N-1}}{2} u_{i,j,N} D_-^z u_{i,j,N} \Big). \end{aligned}$$

The Dirichlet boundary condition at $k = N$ gives

$$T_N^{(u)} = -h \sum_{i,j} \frac{\mu_{i,j,N}}{2} u_{i,j,N-1}^2.$$

Similarly, we obtain

$$\begin{aligned} (v, L^{(v)})_h = &- 2(D_+^y v, E_{1/2}^y(\mu) D_+^y v)_h - (\widetilde{D_0^z} v, \mu \widetilde{D_0^z} v)_h \\ &- \left(D_0^y v, \lambda (D_0^x u + D_0^y v + \widetilde{D_0^z} w) \right)_h - (D_0^x v, \mu (D_0^y u + D_0^x w))_h \\ &- \left(\widetilde{D_0^z} v, \mu D_0^y w \right)_h - \frac{h^2}{4} \left[(D_+^x D_-^x v, \mu D_+^x D_-^x v)_h \right. \\ &\left. + (D_+^y D_-^y v, \lambda D_+^y D_-^y v)_h + (D_+^z D_-^z v, \mu D_+^z D_-^z v)_{hr} \right] \\ &- h \sum_{i,j} \frac{\mu_{i,j,N}}{2} v_{i,j,N-1}^2. \end{aligned}$$

In the z-direction, we make use of (84) and (92) as well as

$$\begin{aligned} &\left(w, D_+^z \left(E_{1/2}^z(\mu) D_-^z w \right) \right)_h = - \left(D_+^z w, E_{1/2}^z(\mu) D_+^z w \right)_h \\ &+ h^2 \sum_{i,j} -\frac{1}{2}\mu_{i,j,1/2} w_{i,j,1} D_+^z w_{i,j,0} - \frac{1}{2}\mu_{i,j,3/2} w_{i,j,1} D_+^z w_{i,j,1} - \frac{h}{2}\mu_{i,j,3/2} (D_+^z w_{i,j,1})^2 \\ &\qquad + \mu_{i,j,N-1/2} w_{i,j,N} D_+^z w_{i,j,N-1}. \end{aligned}$$

We have

$$\begin{aligned} (w, L^{(w)})_h = &-2(D_+^z w, E_{1/2}^z(\mu) D_+^z w)_h - (\widetilde{D_0^z} w, \lambda \widetilde{D_0^z} w)_h - (\widetilde{D_0^z} w, \lambda (D_0^x u + D_0^y v))_h \\ &- (D_0^x w, \mu (D_0^x w + \widetilde{D_0^z} u))_h - (D_0^y w, \mu (D_0^y w + \widetilde{D_0^z} v))_h \\ &- \frac{h^2}{4} \left[(D_+^x D_-^x w, \mu D_+^x D_-^x w)_h + (D_+^y D_-^y w, \mu D_+^y D_-^y w)_h + (D_+^z D_-^z w, \lambda D_+^z D_-^z w)_{hr} \right] \\ &+ T_1^{(w)} + T_N^{(w)}, \end{aligned}$$

where $T_1^{(w)}$ are the boundary terms that belong to the free surface boundary, and $T_N^{(w)}$ are the boundary terms that belong to the Dirichlet boundary. We have

$$\begin{aligned}
T_1^{(w)} &= h^2 \sum_{i,j} -\frac{1}{2}\lambda_{i,j,1/2} w_{i,j,1} D_+^z w_{i,j,0} - \frac{1}{2}\lambda_{i,j,3/2} w_{i,j,1} D_+^z w_{i,j,1} \\
&\quad - w_{i,j,1}\lambda_{i,j,1}(D_0^x u_{i,j,1} + D_0^y v_{i,j,1}) - \mu_{i,j,1/2} w_{i,j,1} D_+^z w_{i,j,0} \\
&\quad - \mu_{i,j,3/2} w_{i,j,1} D_+^z w_{i,j,1} - h\mu_{i,j,3/2}(D_+^z w_{i,j,1})^2 \\
&= h^2 \sum_{i,j} w_{i,j,1} \left(-\frac{1}{2}\lambda_{i,j,1/2} D_+^z w_{i,j,0} - \frac{1}{2}\lambda_{i,j,3/2} D_+^z w_{i,j,1} \right. \\
&\quad - \lambda_{i,j,1}(D_0^x u_{i,j,1} + D_0^y v_{i,j,1}) - \mu_{i,j,1/2} D_+^z w_{i,j,0} \\
&\quad \left. - \mu_{i,j,3/2} D_+^z w_{i,j,1} \right) - h\mu_{i,j,3/2}(D_+^z w_{i,j,1})^2.
\end{aligned}$$

The free surface boundary condition (22) gives

$$T_1^{(w)} = -h^3 \sum_{i,j} \mu_{i,j,3/2}(D_+^z w_{i,j,1})^2.$$

At the Dirichlet boundary we have

$$\begin{aligned}
T_N^{(w)} &= h^2 \sum_{i,j} 2\mu_{i,j,N-1/2} w_{i,j,N} D_+^z w_{i,j,N-1} + \frac{1}{2} w_{i,j,N-1}\lambda_{i,j,N}(D_0^x u_{i,j,N} + D_0^y v_{i,j,N}) \\
&\quad + \frac{1}{2} w_{i,j,N}\lambda_{i,j,N-1}(D_0^x u_{i,j,N-1} + D_0^y v_{i,j,N-1}) + \frac{\lambda_{i,j,N}}{2} w_{i,j,N-1} D_+^z w_{i,j,N-1} \\
&\quad + \frac{\lambda_{i,j,N-1}}{2} w_{i,j,N} D_+^z w_{i,j,N-1} = -h \sum_{i,j} \frac{\lambda_{i,j,N}}{2} w_{i,j,N-1}^2.
\end{aligned}$$

Adding the expressions for $(u, L^{(u)})$, $(v, L^{(v)})$, and $(w, L^{(w)})$ results in (28)–(30).

Acknowledgments. The authors thank Dr. Arthur Rodgers for sharing his expertise in seismology and extensive experience with wave propagation codes, Mrs. Kathleen McCandless for implementing the method on massively parallel machines at LLNL, and Dr. Steven Blair and Dr. Hrvoje Tkalcic for testing and validating the new code.

REFERENCES

[1] B. T. Aagaard, T. M. Brocher, D. Dolenc, D. Dreger, A. Frankel, R. W. Graves, S. Harmsen, S. Hartzell, S. Larsen, K. McCandless, S. Nilsson, N. A. Petersson, A. Rodgers, B. Sjogreen, and M. L. Zoback, *Ground-Motion Modeling of the* 1906 *San Francisco Earthquake* II: *Ground-Motion Estimates for the* 1906 *Earthquake and Scenario Events*, Bull. Seismol. Soc. Amer., to appear.

[2] Z. S. Alterman and A. Rotenberg, *Seismic waves in a quarter plane*, Bull. Seismol. Soc. Amer., 59 (1969), pp. 347–368.

[3] A. Bamberger, G. Chavent, and P. Lailly, *Étude de schémas numériques de l'élastodynamique linéaire*, Technical report RR-0041, INRIA Rocquencourt, 1980, http://hal.inria.fr/inria-00076520.

[4] A. Ben-Menahem and S. J. Singh, *Seismic Waves and Sources*, Dover, New York, 2000.

[5] G. Chesshire and W. Henshaw, *Composite overlapping meshes for the solution of partial differential equations*, J. Comput. Phys., 90 (1990), pp. 1–64.

[6] R. Clayton and B. Engquist, *Absorbing boundary conditions for acoustic and elastic wave equations*, Bull. Seismol. Soc. Amer., 67 (1977), pp. 1529–1540.

[7] G. C. Cohen, *Higher-Order Numerical Methods for Transient Wave Equations*, Springer, New York, 2002.
[8] S. M. Day, J. Bielak, D. Dreger, S. Larsen, R. Graves, A. Pitarka, and K. B. Olsen, *Tests of* 3*D Elastodynamic Codes: Lifelines Program Task* 1*A*01, Technical report, Pacific Earthquake Engineering Center, 2001.
[9] A. C. Eringen and E. S. Şuhubi, *Elastodynamics, Volume* II, Elsevier, New York, 1975.
[10] G. B. Folland, *Introduction to Partial Differential Equations*, Princeton University Press, Princeton, NJ, 1976.
[11] B. Gustafsson, H.-O. Kreiss, and J. Oliger, *Time Dependent Problems and Difference Methods*, Wiley-Interscience, New York, 1995.
[12] B. Gustafsson, H.-O. Kreiss, and A. Sundström, *Stability theory of difference approximations for mixed initial boundary value problems*, II, Math. Comp., 26 (1972), pp. 649–686.
[13] A. Ilan and D. Loewenthal, *Instability of finite difference schemes due to boundary conditions in elastic media*, Geophys. Prospecting, 24 (1976), pp. 431–453.
[14] A. Ilan, *Stability of finite difference schemes for the problem of elastic wave propagation in a quarter plane*, J. Comput. Phys., 29 (1978), pp. 389–403.
[15] D. Komatitsch and J. Tromp, *Introduction to the spectral element method for three-dimensional seismic wave propagation*, Geophys. J. Int., 139 (1999), pp. 806–822.
[16] H.-O. Kreiss and N. A. Petersson, *An embedded boundary method for the wave equation with discontinuous coefficients*, SIAM J. Sci. Comput., 28 (2006), pp. 2054–2074.
[17] H.-O. Kreiss and N. A. Petersson, *A second order accurate embedded boundary method for the wave equation with Dirichlet data*, SIAM J. Sci. Comput., 27 (2006), pp. 1141–1167.
[18] H.-O. Kreiss, N. A. Petersson, and J. Yström, *Difference approximations for the second order wave equation*, SIAM J. Numer. Anal., 40 (2002), pp. 1940–1967.
[19] H.-O. Kreiss, N. A. Petersson, and J. Yström, *Difference approximations of the Neumann problem for the second order wave equation*, SIAM J. Numer. Anal., 42 (2004), pp. 1292–1323.
[20] H. Lamb, *On the propagation of tremors over the surface of an elastic solid*, Philos. Trans. R. Soc. Lond. Ser. A, 203 (1904), pp. 1–42.
[21] R. Madariaga, *Dynamics of an expanding circular fault*, Bull. Seismol. Soc. Amer., 66 (1976), pp. 639–666.
[22] H. M. Mooney, *Some numerical solutions for Lamb's problem*, Bull. Seismol. Soc. Amer., 64 (1974), pp. 473–492.
[23] N. A. Petersson, A. Rodgers, M. Duchaineau, S. Nilsson, B. Sjögreen, and K. McCandless, *Large scale seismic modeling and visualization of the* 1906 *San Francisco earthquake*, Seismol. Res. Lett., 77 (2006). Abstract for the Seismological Society of America meeting, San Francisco, CA.
[24] A.-K. Tornberg and B. Engquist, *Numerical approximation of singular source terms in differential equations*, J. Comput. Phys., 200 (2004), pp. 462–488.
[25] J. E. Vidale and R. W. Clayton, *A stable free-surface boundary condition for two-dimensional elastic finite-difference wave simulation*, Geophysics, 51 (1986), pp. 2247–2249.
[26] J. Waldén, *On the approximation of singular source terms in differential equations*, Numer. Methods Partial Differential Equations, 15 (1999), pp. 503–520.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1937–1953

CONVERGENCE OF A FINITE ELEMENT APPROXIMATION TO A STATE-CONSTRAINED ELLIPTIC CONTROL PROBLEM*

KLAUS DECKELNICK† AND MICHAEL HINZE‡

Abstract. We consider an elliptic optimal control problem with pointwise state constraints. The cost functional is approximated by a sequence of functionals which are obtained by discretizing the state equation with the help of linear finite elements and enforcing the state constraints in the nodes of the triangulation. The corresponding minima are shown to converge in L^2 to the exact control as the discretization parameter tends to zero. Furthermore, error bounds for the control and the state are obtained in both two and three space dimensions. Finally, we present numerical examples which confirm our analytical findings.

Key words. elliptic optimal control problem, state constraints, error estimates

AMS subject classifications. 49J20, 49K20, 35B37

DOI. 10.1137/060652361

1. Introduction.

The aim of this paper is to analyze a finite element discretization of a control problem with pointwise state constraints. Let $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ be a bounded domain with a smooth boundary. For a given function $u \in L^2(\Omega)$ we denote by $y = \mathcal{G}(u)$ the solution of the Neumann problem

$$\begin{aligned} -\Delta y + y &= u \quad \text{in } \Omega, \\ \partial_\nu y &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Here ν denotes the outward pointing unit normal to $\partial\Omega$. It is well known that $y \in H^2(\Omega)$ and

$$\|y\|_{H^2} \le C\|u\|_{L^2}. \tag{1.1}$$

We now consider the following control problem:

$$\min_{u \in L^2(\Omega)} J(u) = \frac{1}{2}\int_\Omega |y - y_0|^2 + \frac{\alpha}{2}\int_\Omega |u - u_0|^2$$

$$\text{subject to } y = \mathcal{G}(u) \text{ and } y(x) \le b(x) \text{ in } \Omega. \tag{1.2}$$

Here $\alpha > 0$ and $y_0, u_0 \in H^1(\Omega)$ as well as $b \in W^{2,\infty}(\Omega)$ are given functions. We denote by $\mathcal{M}(\bar\Omega)$ the space of Radon measures, which is defined as the dual space of $C^0(\bar\Omega)$ and endowed with the norm

$$\|\mu\|_{\mathcal{M}(\bar\Omega)} = \sup_{f \in C^0(\bar\Omega), |f| \le 1} \int_{\bar\Omega} f\, d\mu.$$

The analysis of (1.2) is well understood and sketched in [16, section 6.2.1] for the problem under consideration. Since the state constraints form a convex set and the

*Received by the editors February 17, 2006; accepted for publication (in revised form) May 4, 2007; published electronically August 31, 2007.
http://www.siam.org/journals/sinum/45-5/65236.html
†Institut für Analysis und Numerik, Otto–von–Guericke–Universität Magdeburg, Universitätsplatz 2, D-39106 Magdeburg, Germany (Klaus.Deckelnick@mathematik.uni-magdeburg.de)
‡Department Mathematik, Schwerpunkt Optimierung und Approximation, Universität Hamburg, Bundesstraße 55, D-20146 Hamburg, Germany (michael.hinze@uni-hamburg.de).

cost functional is quadratic, it is not difficult to establish the existence of a unique solution $u \in L^2(\Omega)$ to this problem. Moreover, from [3, Theorem 5.2] we infer (compare also [2, Theorem 2]) the following.

THEOREM 1.1. *A function $u \in L^2(\Omega)$ is a solution of (1.2) if and only if there exist $\mu \in \mathcal{M}(\bar\Omega)$ and $p \in L^2(\Omega)$ such that with $y = \mathcal{G}(u)$ there holds,*

$$\int_\Omega p\big(-\Delta v + v\big) = \int_\Omega (y - y_0)v + \int_{\bar\Omega} v\,d\mu \qquad \forall v \in H^2(\Omega) \text{ with } \partial_\nu v = 0 \text{ on } \partial\Omega, \tag{1.3}$$

$$p + \alpha(u - u_0) = 0 \qquad \text{a.e. in } \Omega, \tag{1.4}$$

$$\mu \ge 0,\ y(x) \le b(x) \text{ a.e. in } \Omega \text{ and } \int_{\bar\Omega} (b - y)\,d\mu = 0. \tag{1.5}$$

The study of (1.2) is complicated by the presence of the measure μ on the right-hand side of (1.3). As a consequence, the solution p of this problem is no longer in $H^1(\Omega)$ but only in $W^{1,s}(\Omega)$ for all $1 \le s < \frac{d}{d-1}$. This fact also accounts for the form of the weak formulation (1.3).

The aim of the present paper is to develop a finite element approximation of problem (1.2). The underlying idea consists in approximating the cost functional J by a sequence of functionals J_h, where h is a mesh parameter related to a sequence of triangulations. The definition of J_h involves the approximation of the state equation by linear finite elements and enforces constraints on the state in the nodes of the triangulation. We shall prove that the minima of J_h converge in L^2 to the minimum of J as $h \to 0$ and that the states convergence strongly in H^1 as well as uniformly and derive corresponding error bounds.

To our knowledge only a few attempts have been made to develop a finite element analysis for state-constrained elliptic control problems. In [4] Casas proves convergence of finite element approximations to optimal control problems for semilinear elliptic equations with finitely many state constraints. Casas and Mateos extend these results in [5] to a less regular setting for the states and prove convergence of finite element approximations to semilinear distributed and boundary control problems.

Let us comment on further approaches that tackle optimization problems for PDEs with state constraints. A *Lavrentiev-type regularization* of problem (1.2) is investigated in [11]. In this approach the state constraint $y \le b$ in (1.2) is replaced by the mixed constraint $\epsilon u + y \le b$, with $\epsilon > 0$ denoting a regularization parameter. It turns out that the associated Lagrange multiplier μ_ϵ belongs to $L^2(\Omega)$. The resulting optimization problems are solved by either interior-point methods or primal-dual active set strategies; compare [10]. The development of numerical approaches to tackle (1.2) is ongoing. An excellent overview can be found in [8, 9], where further references are also given.

The paper is organized as follows: In section 2 we describe our discretization and establish convergence of controls and states to their continuous counterparts for two- and three-dimensional domains. An error analysis is carried out in section 3. We obtain

$$\|u - u_h\|_{L^2},\ \|y - y_h\|_{H^1} = \begin{cases} O(h^{1-\epsilon}) & \text{if } d = 2, \\ O(h^{\frac{1}{2}-\epsilon}) & \text{if } d = 3 \end{cases}$$

($\epsilon > 0$ arbitrary), where u_h and y_h are the discrete control and state, respectively. Roughly speaking, the idea is to insert the discrete solution into the continuous functional and vice versa. An important tool in the analysis is the use of L^∞-error esti-

mates for finite element approximations of the Neumann problem developed in [13]. The need for uniform estimates is due to the presence of the measure μ in (1.3).

2. Finite element discretization. Let $\mathcal{T}_h$ be a triangulation of Ω with maximum mesh size $h := \max_{T\in\mathcal{T}_h} \operatorname{diam}(T)$ and vertices $x_1, \dots, x_m$. We suppose that $\bar{\Omega}$ is the union of the elements of $\mathcal{T}_h$; boundary elements are allowed to have one curved face. In addition, we assume that the triangulation is quasi-uniform in the sense that there exists a constant $\kappa > 0$ (independent of h) such that each $T \in \mathcal{T}_h$ is contained in a ball of radius $\kappa^{-1}h$ and contains a ball of radius κh. Let us define the space of linear finite elements:

$$X_h := \{v_h \in C^0(\bar{\Omega}) \,|\, v_h \text{ is a linear polynomial on each } T \in \mathcal{T}_h\}.$$

We have the following approximation and inverse properties:

(a) There exists an interpolation operator $\Pi_h : W^{l,p}(\Omega) \to X_h$ $(l = 1, 2; 1 \le p \le \infty)$ such that for $T \in \mathcal{T}_h$

$$\|v - \Pi_h v\|_{W^{m,p}(T)} \le Ch^{l-m}\|v\|_{W^{l,p}(S_T)}, \quad 0 \le m \le l, \tag{2.1}$$

where $S_T = \cup\{\tilde{T} \in \mathcal{T}_h \,|\, \tilde{T} \cap T \neq \emptyset\}$.

(b) If $p > d$, the usual nodal interpolate satisfies

$$\|v - I_h v\|_{W^{1,\infty}(T)} \le Ch^{1-\frac{d}{p}}\|v\|_{W^{2,p}(T)}, \quad T \in \mathcal{T}_h. \tag{2.2}$$

(c) For $v_h \in X_h$, $m = 0, 1, l \ge 0$, and $1 \le q \le p \le \infty$ we have

$$\|v_h\|_{W^{m,p}(T)} \le Ch^{-(\frac{d}{q}-\frac{d}{p})-m-l}\|v_h\|_{W^{-l,q}(T)}, \quad T \in \mathcal{T}_h, \tag{2.3}$$

where, for $l \ge 1$, $W^{-l,q}(T)$ is the dual of $W_0^{l,q'}(T)$, $\frac{1}{q} + \frac{1}{q'} = 1$.

The interpolation operator Π_h in (a) can be defined as in [15] using averages over $(d-1)$-simplices σ_i. Since boundary elements are allowed to have only one curved face we can associate with a boundary node x_i a straight $(d-1)$-simplex σ_i such that $x_i \in \sigma_i$. It turns out that the arguments presented in [15] for a polydedral domain can be used with small changes in order to derive (2.1) for boundary elements T. Note that functions $v_h \in X_h$ do not have to satisfy boundary conditions. The estimates (2.2) and (2.3) can be proved similarly as in [14, pp. 685–686].

In what follows it is convenient to introduce a discrete approximation of the solution operator $\mathcal{G}$. For a given function $v \in L^2(\Omega)$ we denote by $z_h = \mathcal{G}_h(v) \in X_h$ the solution of the discrete Neumann problem

$$\int_\Omega \big(\nabla z_h \cdot \nabla v_h + z_h v_h\big) \;=\; \int_\Omega v v_h \quad \text{for all } v_h \in X_h.$$

It is well known that for all $v \in L^2(\Omega)$

$$\|\mathcal{G}(v) - \mathcal{G}_h(v)\| \le Ch^2\|v\|, \tag{2.4}$$

$$\|\mathcal{G}(v) - \mathcal{G}_h(v)\|_{L^\infty} \le Ch^{2-\frac{d}{2}}\|v\|. \tag{2.5}$$

Here $\|\cdot\|$ denotes the L^2-norm. We propose the following approximation of the control problem (1.2):

$$\min_{u\in L^2(\Omega)} J_h(u) := \frac{1}{2}\int_\Omega |y_h - P_h y_0|^2 + \frac{\alpha}{2}\int_\Omega |u - P_h u_0|^2$$

$$\text{subject to } \; y_h = \mathcal{G}_h(u) \text{ and } y_h(x_j) \le b(x_j) \text{ for } j = 1, \dots, m. \tag{2.6}$$

Here P_h denotes the L^2-projection, i.e.,

$$\int_\Omega P_h z\, v_h = \int_\Omega z\, v_h \qquad \forall v_h \in X_h. \tag{2.7}$$

It is well known that

$$\|z - P_h z\| \le Ch\|z\|_{H^1} \qquad \forall z \in H^1(\Omega). \tag{2.8}$$

Problem (2.6) represents a convex infinite-dimensional optimization problem of similar structure as problem (1.2) but with only finitely many equality and inequality constraints which form a convex admissible set. Again we can apply [3, Theorem 5.2] which together with [2, Corollary 1] yields (compare also the analysis of problem (P) in [4]) the following.

LEMMA 2.1. *Problem* (2.6) *has a unique solution* $u_h \in L^2(\Omega)$. *There exist* $\mu_1, \dots, \mu_m \in \mathbb{R}$ *and* $p_h \in X_h$ *such that with* $y_h = \mathcal{G}_h(u_h)$ *and* $\mu_h = \sum_{j=1}^m \mu_j \delta_{x_j}$ *we have*

$$\int_\Omega \big(\nabla p_h \cdot \nabla v_h + p_h v_h\big) = \int_\Omega (y_h - P_h y_0) v_h + \int_{\bar\Omega} v_h d\mu_h \quad \text{for all } v_h \in X_h, \tag{2.9}$$

$$p_h + \alpha(u_h - P_h u_0) = 0 \text{ in } \Omega, \tag{2.10}$$

$$\mu_j \ge 0,\ y_h(x_j) \le b(x_j), j = 1, \dots, m, \text{ and } \int_{\bar\Omega} \big(I_h b - y_h\big) d\mu_h = 0. \tag{2.11}$$

Here δ_x denotes the Dirac measure concentrated at x, and I_h is the usual Lagrange interpolation operator.

Remark 2.2. From (2.10) we deduce that in problem (2.6) it is sufficient to minimize over controls $u \in X_h$ instead of $u \in L^2(\Omega)$ in order to obtain the same unique solution u_h. For the resulting finite-dimensional optimization problem the result of Lemma 2.1 then follows from, e.g., [12, Theorem 12.1].

We have the following convergence result.

THEOREM 2.3. *Let* $u_h \in L^2(\Omega)$ *be the optimal solution of* (2.6) *with corresponding state* $y_h \in X_h$ *and adjoint variables* $p_h \in X_h$ *and* $\mu_h \in \mathcal{M}(\bar\Omega)$. *Then as* $h \to 0$ *we have*

$$u_h \to u \text{ in } L^2(\Omega), \quad y_h \to y \text{ in } H^1(\Omega) \text{ and in } C^0(\bar\Omega),$$

where u *is the solution of* (1.2) *with corresponding state* y.

Proof. Let $\underline{b} := \min_{x \in \bar\Omega} b(x)$. Since $\underline{b} = \mathcal{G}_h(\underline{b})$ and $\underline{b} \le b(x_j)$ for $j = 1, \dots, m$ we have

$$\frac{1}{2}\int_\Omega |y_h - P_h y_0|^2 + \frac{\alpha}{2}\int_\Omega |u_h - P_h u_0|^2 = J_h(u_h) \le J_h(\underline{b}) \le C(y_0, u_0, \underline{b}).$$

This implies that there exists a constant C which is independent of h such that

$$\|y_h\|,\ \|u_h\|,\ \|p_h\| \le C \qquad \text{for all } 0 < h \le 1. \tag{2.12}$$

Note that the bound on p_h follows from (2.10). In order to estimate μ_h we use $v_h \equiv 1$ in (2.9) and obtain for every $f \in C^0(\bar\Omega), |f| \le 1$,

$$\int_{\bar\Omega} f d\mu_h \le \sum_{j=1}^m \mu_j |f(x_j)| \le \sum_{j=1}^m \mu_j = \int_{\bar\Omega} 1 d\mu_h = \int_\Omega \big(p_h + P_h y_0 - y_h\big) \le C$$

by (2.12). This yields

$$\|\mu_h\|_{\mathcal{M}(\bar\Omega)} \le C \qquad \text{for all } 0 < h \le 1. \tag{2.13}$$

In view of (2.12) and (2.13) there exists a sequence $h \to 0$ and $\hat u, \hat p \in L^2(\Omega)$ as well as $\hat\mu \in \mathcal{M}(\bar\Omega)$ such that

$$u_h \rightharpoonup \hat u, \quad p_h \rightharpoonup \hat p \text{ in } L^2(\Omega), \quad \text{and } \mu_h \rightharpoonup \hat\mu \text{ in } \mathcal{M}(\bar\Omega). \tag{2.14}$$

Since $\mathcal{G}$ is compact as an operator from $L^2(\Omega)$ into $C^0(\bar\Omega)$ we have, after passing to a further subsequence if necessary,

$$\mathcal{G}(u_h) \to \mathcal{G}(\hat u) \qquad \text{in } C^0(\bar\Omega) \tag{2.15}$$

and hence

$$\begin{aligned}\|y_h - \mathcal{G}(\hat u)\|_{L^\infty} &\le \|\mathcal{G}_h(u_h) - \mathcal{G}(u_h)\|_{L^\infty} + \|\mathcal{G}(u_h) - \mathcal{G}(\hat u)\|_{L^\infty}\\ &\le Ch^{2-\frac{d}{2}}\|u_h\| + \|\mathcal{G}(u_h) - \mathcal{G}(\hat u)\|_{L^\infty}\end{aligned}$$

so that $y_h \to \mathcal{G}(\hat u) =: \hat y$ in $C^0(\bar\Omega)$ as $h \to 0$ by (2.12) and (2.15). A similar argument shows that $y_h \to \hat y$ in $H^1(\Omega)$.

Let us now pass to the limit in (2.9)–(2.11). To begin, let $v \in H^2(\Omega)$ with $\partial_\nu v = 0$ on $\partial\Omega$ and denote by $R_h v$ the Ritz projection of v. Recalling (2.14) and (2.9) and the fact that $R_h v \to v$ in $C^0(\bar\Omega)$ we obtain

$$\begin{aligned}\int_\Omega \hat p\big(-\Delta v + v\big) &\leftarrow \int_\Omega p_h\big(-\Delta v + v\big) = \int_\Omega \big(\nabla p_h \cdot \nabla v + p_h v\big)\\ &= \int_\Omega \big(\nabla p_h \cdot \nabla R_h v + p_h R_h v\big) = \int_\Omega (y_h - P_h y_0) R_h v + \int_{\bar\Omega} R_h v\, d\mu_h\\ &\to \int_\Omega (\hat y - y_0) v + \int_{\bar\Omega} v\, d\hat\mu.\end{aligned}$$

Using (2.14) we may pass to the limit in (2.10) and deduce $\hat p + \alpha(\hat u - u_0) = 0$ a.e. in Ω. Clearly, $\hat\mu \ge 0$; since $y_h \le I_h b$ in $\bar\Omega$ and $y_h \to \hat y$ in $C^0(\bar\Omega)$ we have $\hat y \le b$ in $\bar\Omega$. Furthermore, recalling that $\int_{\bar\Omega}(I_h b - y_h)\, d\mu_h = 0$ we obtain in the limit

$$\int_{\bar\Omega} (b - \hat y)\, d\hat\mu = 0.$$

Theorem 1.1 now implies that $\hat u$ is a solution of (1.2); as the solution of this problem is unique we must have $u = \hat u$ and hence $y = \hat y$, and the whole sequence is convergent.

Let us finally prove that $u_h \to u$ in $L^2(\Omega)$. To begin, note that by (2.5)

$$\mathcal{G}_h\big(u - \gamma h^{2-\frac{d}{2}}\big) = \mathcal{G}_h(u) - \mathcal{G}(u) + \mathcal{G}(u) - \gamma h^{2-\frac{d}{2}} \le Ch^{2-\frac{d}{2}}\|u\| + b - \gamma h^{2-\frac{d}{2}} \le b$$

in $\bar\Omega$, provided that γ is large enough. Evaluating the above inequality at the nodes $x_1, \dots, x_m$ we see that $\mathcal{G}_h(u - \gamma h^{2-\frac{d}{2}})$ is admissible for the discrete problem, and hence $J_h(u_h) \le J_h(u - \gamma h^{2-\frac{d}{2}})$ or

$$\frac{\alpha}{2}\|u_h - P_h u_0\|^2 \le \frac{\alpha}{2}\|u - \gamma h^{2-\frac{d}{2}} - P_h u_0\|^2 + \frac{1}{2}\|\mathcal{G}_h(u) - \gamma h^{2-\frac{d}{2}} - P_h y_0\|^2 - \frac{1}{2}\|y_h - P_h y_0\|^2.$$

Since $y_h \to y$, $\mathcal{G}_h(u) \to \mathcal{G}(u) = y$ in $L^2(\Omega)$ we infer that

$$\limsup_{h\to 0} \|u_h - P_h u_0\|^2 \le \|u - u_0\|^2 \le \liminf_{h\to 0} \|u_h - P_h u_0\|^2,$$

where the second inequality is a consequence of the weak convergence $u_h - P_h u_0 \rightharpoonup u - u_0$. Thus, $\|u_h - P_h u_0\|^2 \to \|u - u_0\|^2$, which implies $u_h - P_h u_0 \to u - u_0$ in L^2 and hence $u_h \to u_0$ in L^2. ∎

Remark 2.4. The above convergence result also holds if Ω is assumed to be a bounded, convex, and polyhedral domain in $\mathbb{R}^d$. To see this we note that (1.1) still holds in this case, and hence both Theorem 1.1 (see [2]) and the estimates (2.4), (2.5) remain true. An inspection of the proof of Theorem 2.3 then shows that this is sufficient in order to carry out the analysis.

3. Error analysis. Let us now turn to the error analysis and start with a couple of auxiliary results.

LEMMA 3.1. *Suppose that $u, u_h \in L^2(\Omega)$ are the optimal solutions of* (1.2) *and* (2.6)*, respectively, with corresponding states $y \in H^2(\Omega), y_h \in X_h$. Let $v \in L^2(\Omega)$ and $z = \mathcal{G}(v), z_h = \mathcal{G}_h(v)$. Then*

$$J(u) + \frac{1}{2}\int_\Omega |z - y|^2 + \frac{\alpha}{2}\int_\Omega |v - u|^2 + \int_{\bar\Omega} (b - z)\, d\mu = J(v), \tag{3.1}$$

$$J_h(u_h) + \frac{1}{2}\int_\Omega |z_h - y_h|^2 + \frac{\alpha}{2}\int_\Omega |v - u_h|^2 + \int_{\bar\Omega} \big(I_h b - z_h\big)\, d\mu_h = J_h(v). \tag{3.2}$$

Proof. An elementary calculation using (1.3) shows

$$\begin{aligned} J(v) - J(u) &= \frac{1}{2}\int_\Omega |z - y|^2 + \frac{\alpha}{2}\int_\Omega |v - u|^2 + \int_\Omega (z - y)(y - y_0) + \alpha \int_\Omega (u - u_0)(v - u) \\ &= \frac{1}{2}\int_\Omega |z - y|^2 + \frac{\alpha}{2}\int_\Omega |v - u|^2 + \int_\Omega p\big(-\Delta(z - y) + (z - y)\big) \\ &\quad - \int_{\bar\Omega} (z - y)\, d\mu + \alpha \int_\Omega (u - u_0)(v - u). \end{aligned}$$

Since $z = \mathcal{G}(v), y = \mathcal{G}(u)$ we have

$$\int_\Omega p\big(-\Delta(z - y) + (z - y)\big) = \int_\Omega p(v - u),$$

so that (1.4) and (1.5) finally imply

$$J(v) - J(u) = \frac{1}{2}\int_\Omega |z - y|^2 + \frac{\alpha}{2}\int_\Omega |v - u|^2 + \int_{\bar\Omega} (b - z)\, d\mu.$$

The second claim follows in a similar way. ∎

Remark 3.2. Note that in the above $z = \mathcal{G}(v), z_h = \mathcal{G}_h(v)$ do not necessarily have to be admissible for the minimization problems.

The next lemma examines in more detail the approximation of J by J_h.

LEMMA 3.3. *Suppose that $v \in W^{1,s}(\Omega)$ for some $\frac{2d}{d+2} \le s \le 2$. Then*

$$|J(v) - J_h(v)| \le C h^{2 + \frac{d}{2} - \frac{d}{s}} \big(\|u_0\|_{H^1}\|v\|_{W^{1,s}} + \|v\|^2 + \|y_0\|_{H^1}^2 + \|u_0\|_{H^1}^2\big).$$

Proof. Let $z = \mathcal{G}(v), z_h = \mathcal{G}_h(v)$. Then

$$J(v) - J_h(v) = \frac{1}{2}\int_\Omega \big(|z - y_0|^2 - |z_h - P_h y_0|^2\big) + \frac{\alpha}{2}\int_\Omega \big(|v - u_0|^2 - |v - P_h u_0|^2\big).$$

Using (2.7), (2.8), (2.4), and (1.1) we obtain

$$\begin{aligned}
&\left|\int_\Omega \big(|z-y_0|^2-|z_h-P_hy_0|^2\big)\right| = \left|\int_\Omega (z-y_0-z_h+P_hy_0)(z-y_0+z_h-P_hy_0)\right| \\
&= \left|\int_\Omega \big((z-z_h)(z-y_0+z_h-P_hy_0)-(y_0-P_hy_0)(z-y_0-P_h(z-y_0)\big)\right| \\
&\le C\|z-z_h\|\big(\|z\|+\|z_h\|+\|y_0\|\big)+Ch^2\|y_0\|_{H^1}\big(\|z\|_{H^1}+\|y_0\|_{H^1}\big) \\
&\le Ch^2\big(\|v\|^2+\|y_0\|_{H^1}^2\big).
\end{aligned}$$

For the second term we obtain in a similar way

$$\int_\Omega \big(|v-u_0|^2-|v-P_hu_0|^2\big)=\int_\Omega (u_0-P_hu_0)w=\int_\Omega (u_0-P_hu_0)(w-P_hw),$$

where $w=u_0+P_hu_0-2v$ and where we have used (2.7). Applying Lemma 5.1 from the appendix we infer

$$\begin{aligned}
\left|\int_\Omega \big(|v-u_0|^2-|v-P_hu_0|^2\big)\right| &\le Ch^{2+\frac{d}{2}-\frac{d}{s}}\|u_0\|_{H^1}\|w\|_{W^{1,s}} \\
&\le Ch^{2+\frac{d}{2}-\frac{d}{s}}\|u_0\|_{H^1}\big(\|u_0\|_{H^1}+\|v\|_{W^{1,s}}\big).
\end{aligned}$$

This proves the lemma. ∎

LEMMA 3.4. *Suppose that $v\in W^{1,s}(\Omega)$ for some $1<s<\frac{d}{d-1}$. Then*

$$\|\mathcal{G}(v)-\mathcal{G}_h(v)\|_{L^\infty}\le Ch^{3-\frac{d}{s}}|\log h|\ \|v\|_{W^{1,s}}.$$

Proof Let $z=\mathcal{G}(v)$, $z_h=\mathcal{G}_h(v)$. From a well-known embedding theorem we infer that $v\in L^q(\Omega)$, with $q=\frac{ds}{d-s}$. Hence, elliptic regularity theory implies that $z\in W^{2,q}(\Omega)$ and

$$\|z\|_{W^{2,q}}\le C\|v\|_{L^q}\le C\|v\|_{W^{1,s}}. \tag{3.3}$$

Using Theorem 2.2 and the ensuing Remark in [13] we have

$$\|z-z_h\|_{L^\infty}\le C|\log h|\inf_{\chi\in X_h}\|z-\chi\|_{L^\infty}. \tag{3.4}$$

The result in [13] holds under certain approximation and inverse properties (see A.1–A.4 on pp. 883–884) on the finite element space X_h which follow from (2.1)–(2.3). Combining (3.4) with a well-known interpolation estimate yields

$$\|z-z_h\|_{L^\infty}\le Ch^{2-\frac{d}{q}}|\log h|\|z\|_{W^{2,q}}\le Ch^{3-\frac{d}{s}}|\log h|\|v\|_{W^{1,s}}$$

in view of (3.3) and the relation between s and q. ∎

Our next aim is to derive a uniform bound on $\|u_h\|_{W^{1,s}}$ for $s<\frac{d}{d-1}$.

LEMMA 3.5. *Let $1<s<\frac{d}{d-1}$. Then there exists a constant c, which is independent of h, such that*

$$\|u_h\|_{W^{1,s}}\le c \qquad \text{for all } 0<h\le 1.$$

Proof. In view of (2.10) we have

$$\|u_h\|_{W^{1,s}}\le\frac{1}{\alpha}\|p_h\|_{W^{1,s}}+\|P_hu_0\|_{H^1}\le\frac{1}{\alpha}\|p_h\|_{W^{1,s}}+c,$$

so that it is sufficient to bound $\|p_h\|_{W^{1,s}}$.

Let s' be such that $\frac{1}{s}+\frac{1}{s'}=1$, and suppose that $\phi\in L^{s'}(\Omega)$. Let us denote by $\psi\in W^{2,s'}(\Omega)$ the unique solution of the Neumann problem

$$\begin{aligned} -\Delta\psi+\psi &= \phi \quad \text{in } \Omega,\\ \partial_\nu\psi &= 0 \quad \text{on } \partial\Omega. \end{aligned}$$

Integration by parts and (2.9) yield

$$\begin{aligned} \int_\Omega p_h\,\phi &= \int_\Omega \big(\nabla p_h\cdot\nabla\psi + p_h\psi\big) = \int_\Omega \big(\nabla p_h\cdot\nabla R_h\psi + p_h R_h\psi\big)\\ &= \int_\Omega (y_h-P_hy_0)R_h\psi + \int_{\bar\Omega} R_h\psi\, d\mu_h, \end{aligned} \tag{3.5}$$

where $R_h\psi$ is the Ritz projection of ψ. Arguing similarly as in Theorem 1 of [1] one shows that there exists a unique solution $p^h\in W^{1,s}(\Omega)$ of the problem

$$\int_\Omega p^h\big(-\Delta v+v\big) = \int_\Omega (y_h-P_hy_0)v + \int_{\bar\Omega} v\,d\mu_h \quad \forall v\in H^2(\Omega) \text{ with } \partial_\nu v=0 \text{ on } \partial\Omega. \tag{3.6}$$

Furthermore, there exists a constant $c=c(s)>0$ such that

$$\|p^h\|_{W^{1,s}} \le c\big(\|y_h-P_hy_0\|+\|\mu_h\|_{\mathcal{M}(\bar\Omega)}\big)\le C \tag{3.7}$$

uniformly in h in view of (2.12) and (2.13). If we use $v=\psi$ in (3.6) and combine it with (3.5), we obtain

$$\begin{aligned} \int_\Omega (p^h-p_h)\phi &= \int_\Omega (y_h-P_hy_0)(\psi-R_h\psi) + \int_{\bar\Omega}(\psi-R_h\psi)d\mu_h\\ &\le Ch^2\|\psi\|_{H^2}\big(\|y_h\|+\|P_hy_0\|\big) + \|\psi-R_h\psi\|_{L^\infty}\|\mu_h\|_{\mathcal{M}(\bar\Omega)}\\ &\le Ch^2\|\psi\|_{H^2} + ch^{2-\frac{d}{s'}}|\log h|\|\psi\|_{W^{2,s'}}\\ &\le Ch^{2-\frac{d}{s'}}|\log h|\|\phi\|_{L^{s'}}. \end{aligned}$$

Note that we have again applied (3.4) in order to control $\|\psi-R_h\psi\|_{L^\infty}$. Since $\phi\in L^{s'}(\Omega)$ is arbitrary we infer

$$\|p^h-p_h\|_{L^s}\le Ch^{2-\frac{d}{s'}}|\log h|.$$

Interpolation and inverse estimates then give

$$\|\nabla p_h\|_{L^s}\le C\|\nabla p^h\|_{L^s} + ch^{1-\frac{d}{s'}}|\log h|\le C$$

by (3.7) and since $1-\frac{d}{s'}=\frac{d-1}{s}\big(\frac{d}{d-1}-s\big)>0$. ∎

Let us finally turn to an error estimate for the optimal controls and the optimal states.

THEOREM 3.6. *Let u and u_h be the solutions of* (1.2) *and* (2.6)*, respectively. For every $\epsilon>0$ there exists $C_\epsilon>0$ such that*

$$\|u-u_h\|+\|y-y_h\|_{H^1}\le C_\epsilon h^{2-\frac{d}{2}-\epsilon}.$$

Proof. Let us define $\tilde y^h:=\mathcal{G}(u_h)\in H^2(\Omega)$ and $\tilde y_h:=\mathcal{G}_h(u)\in X_h$. Then Lemma 3.1 implies

$$\begin{aligned} J(u)+\frac12\int_\Omega|\tilde y^h-y|^2+\frac{\alpha}{2}\int_\Omega|u_h-u|^2+\int_{\bar\Omega}(b-\tilde y^h)d\mu = J(u_h),\\ J_h(u_h)+\frac12\int_\Omega|\tilde y_h-y_h|^2+\frac{\alpha}{2}\int_\Omega|u-u_h|^2+\int_{\bar\Omega}\big(I_hb-\tilde y_h\big)d\mu_h = J_h(u). \end{aligned}$$

Since $u = u_0 - \frac{1}{\alpha}p \in W^{1,s}(\Omega)$ for all $\frac{2d}{d+2} \le s < \frac{d}{d-1}$ we obtain with the help of Lemma 3.3

$$\begin{aligned}
&\frac{1}{2}\int_\Omega |\tilde{y}^h - y|^2 + \frac{1}{2}\int_\Omega |\tilde{y}_h - y_h|^2 + \alpha\int_\Omega |u_h - u|^2 \\
&= J(u_h) - J(u) + J_h(u) - J_h(u_h) - \int_{\bar\Omega}(b - \tilde{y}^h)d\mu - \int_{\bar\Omega}\big(I_h b - \tilde{y}_h\big)d\mu_h \\
&\le Ch^{2+\frac{d}{2}-\frac{d}{s}}\Big(\|u_0\|_{H^1}\big(\|u\|_{W^{1,s}} + \|u_h\|_{W^{1,s}}\big) + \|u\|^2 + \|u_h\|^2 + \|y_0\|^2_{H^1} + \|u_0\|^2_{H^1}\Big) \\
&\qquad + \int_{\bar\Omega}(\tilde{y}^h - b)d\mu + \int_{\bar\Omega}(\tilde{y}_h - I_h b)d\mu_h. \qquad (3.8)
\end{aligned}$$

Let us first consider the last two integrals. We have for $x \in \bar\Omega$

$$\begin{aligned}
\tilde{y}^h(x) - b(x) &= (\tilde{y}^h(x) - y_h(x)) + (y_h(x) - (I_h b)(x)) + ((I_h b)(x) - b(x)) \\
&\le \|\mathcal{G}(u_h) - \mathcal{G}_h(u_h)\|_{L^\infty} + \|I_h b - b\|_{L^\infty},
\end{aligned}$$

since $y_h(x_j) \le b(x_j), j = 1, \dots, m$, implies that $y_h \le I_h b$ in $\bar\Omega$. If we combine Lemma 3.4 with Lemma 3.5 we infer

$$\int_{\bar\Omega}(\tilde{y}^h - b)d\mu \le Ch^{3-\frac{d}{s}}|\log h|\,\|u_h\|_{W^{1,s}} + Ch^2|b|_{W^{2,\infty}} \le Ch^{3-\frac{d}{s}}|\log h|.$$

Similarly we have from (1.5)

$$\begin{aligned}
\tilde{y}_h(x) - (I_h b)(x) &= (\tilde{y}_h(x) - y(x)) + (y(x) - b(x)) + (b(x) - (I_h b)(x)) \\
&\le \|\mathcal{G}_h(u) - \mathcal{G}(u)\|_{L^\infty} + \|b - I_h b\|_{L^\infty},
\end{aligned}$$

so that (2.13) and Lemma 3.4 give

$$\int_{\bar\Omega}\big(y_h - I_h b\big)d\mu_h \le Ch^{3-\frac{d}{s}}|\log h|\,\|u\|_{W^{1,s}} + Ch^2|b|_{W^{2,\infty}} \le Ch^{3-\frac{d}{s}}|\log h|.$$

Inserting these estimates into (3.8) and applying again Lemma 3.5 we derive

$$\|u - u_h\|^2 + \|y - y_h\|^2 \le Ch^{3-\frac{d}{s}}|\log h|.$$

If we now choose s sufficiently close to $\frac{d}{d-1}$ we obtain

$$\|u - u_h\|^2 + \|y - y_h\|^2 \le C_\epsilon h^{4-d-2\epsilon}.$$

Finally, in order to obtain the error bound for y in H^1 we note that

$$\int_\Omega\big(\nabla(y - y_h)\cdot\nabla v_h + (y - y_h)v_h\big) = \int_\Omega (u - u_h)v_h$$

for all $v_h \in X_h$, from which one derives the desired estimate using standard finite element techniques and the bound on $\|u - u_h\|$. ∎

In general we expect only weak convergence of μ_h to μ. Nevertheless, we have the following partial result.

COROLLARY 3.7. *Let $K \subset \bar\Omega$ be compact with $K \cap \operatorname{supp}\mu = \emptyset$. For every $\epsilon > 0$ there exists a constant C_ϵ such that*

$$\mu_h(K) \le C_\epsilon h^{2-\frac{d}{2}-\epsilon}.$$

Proof. By Lemma 5.2 in the appendix there exists a nonnegative function $\phi \in C^2(\bar\Omega)$ which satisfies

$$\phi \geq 1 \text{ on } K,\ \phi = 0 \text{ on } \operatorname{supp}\mu,\ \partial_\nu \phi = 0 \text{ on } \partial\Omega.$$

Since $\mu_h \geq 0$ we obtain from (2.9)

$$\begin{aligned}
\mu_h(K) &\leq \int_{\bar\Omega} \phi \, d\mu_h = \int_{\bar\Omega} (\phi - R_h\phi) d\mu_h + \int_{\bar\Omega} R_h\phi \, d\mu_h \\
&= \int_{\bar\Omega} (\phi - R_h\phi) d\mu_h + \int_\Omega \big(\nabla p_h \cdot \nabla R_h\phi + p_h R_h\phi\big) - \int_\Omega (y_h - P_h y_0) R_h\phi \\
&= \int_{\bar\Omega} (\phi - R_h\phi) d\mu_h + \int_\Omega \big(\nabla p_h \cdot \nabla \phi + p_h \phi\big) - \int_\Omega (y_h - P_h y_0) R_h\phi \\
&= \int_{\bar\Omega} (\phi - R_h\phi) d\mu_h + \int_\Omega p_h(-\Delta\phi + \phi) - \int_\Omega (y_h - P_h y_0) R_h\phi,
\end{aligned}$$

where R_h is again the Ritz projection. On the other hand, (1.3) and the fact that $\phi = 0$ on $\operatorname{supp}\mu$ imply

$$\int_\Omega (y - y_0)\phi - \int_\Omega p(-\Delta\phi + \phi) = 0.$$

Combining this relation with the first estimate we derive

$$\begin{aligned}
\mu_h(K) &\leq \int_{\bar\Omega} (\phi - R_h\phi) d\mu_h + \int_\Omega (p_h - p)(-\Delta\phi + \phi) + \int_\Omega (y_h - P_h y_0)(\phi - R_h\phi) \\
&\qquad + \int_\Omega (y - y_h - y_0 + P_h y_0)\phi \\
&\leq \|\phi - R_h\phi\|_{L^\infty} \|\mu_h\|_{\mathcal{M}(\bar\Omega)} + \|p - p_h\| \|\phi\|_{H^2} + \big(\|y_h\| + \|P_h y_0\|\big)\|\phi - R_h\phi\| \\
&\qquad + \big(\|y - y_h\| + \|y_0 - P_h y_0\|\big)\|\phi\| \\
&\leq C\|\phi - R_h\phi\|_{L^\infty} + C_\epsilon h^{2-\frac{d}{2}-\epsilon} \leq C_\epsilon h^{2-\frac{d}{2}-\epsilon}
\end{aligned}$$

in view of (1.4), (2.10), and Theorem 3.6. ∎

Remark 3.8. We mention here a second approach that differs from the one discussed above in the way in which the inequality constraints are realized. Denote by $D_1, \dots, D_m$ the cells of the dual mesh. Each cell D_i is associated with a vertex x_i of $\mathcal{T}_h$, and we have

$$\bar\Omega = \cup_{i=1}^m D_i, \quad \operatorname{int}(D_i) \cap \operatorname{int}(D_j) = \emptyset, i \neq j.$$

In (2.6), we now impose the constraints

$$\fint_{D_j} (y_h - I_h b) \leq 0 \text{ for } j = 1, \dots, m \tag{3.9}$$

on the discrete solution $y_h = \mathcal{G}_h(u)$. Here we have abbreviated $\fint_{D_j} f = \frac{1}{|D_j|}\int_{D_j} f$. The measure μ_h that appears in Lemma 2.1 now has the form $\mu_h = \sum_{j=1}^m \mu_j \fint_{D_j} \cdot\, dx$, and the pointwise constraints in (2.11) are replaced by those of (3.9). The error analysis for the resulting numerical method can be carried out in the same way as

shown above with the exception of Theorem 3.6, where the bounds on $\tilde{y} - b$ and $\tilde{y}_h - I_h b$ require a different argument. In this case, additional terms of the form

$$\left\| f - \fint_{D_j} f \right\|_{L^\infty(D_j)}$$

have to be estimated. Since these will, in general, be only of order $O(h)$, this analysis would only give $\|u-u_h\|, \|y-y_h\|_{H^1} = O(\sqrt{h})$. The numerical test example in section 4 suggests that at least $\|u - u_h\| = O(h)$, but we are presently unable to prove such an estimate.

4. Numerical examples.

Example 4.1. The following test problem is taken—in a slightly modified form—from [10, Example 6.2]. Let $\Omega := B_1(0)$, $\alpha > 0$,

$$y_0(x) := 4 + \frac{1}{\pi} - \frac{1}{4\pi}|x|^2 + \frac{1}{2\pi}\log|x|, \quad u_0(x) := 4 + \frac{1}{4\alpha\pi}|x|^2 - \frac{1}{2\alpha\pi}\log|x|,$$

and $b(x) := |x|^2 + 4$. We consider the cost functional

$$J(u) := \frac{1}{2}\int_\Omega |y - y_0|^2 + \frac{\alpha}{2}\int_\Omega |u - u_0|^2,$$

where $y = \mathcal{G}(u)$. By checking the optimality conditions of first order, one verifies that $u \equiv 4$ is the unique solution of (1.2) with corresponding state $y \equiv 4$ and adjoint states

$$p(x) = \frac{1}{4\pi}|x|^2 - \frac{1}{2\pi}\log|x| \quad \text{and} \quad \mu = \delta_0.$$

The finite element counterparts of y, u, p, and μ are denoted by y_h, u_h, p_h, and μ_h, respectively.

For an error functional $E(h)$ we define the experimental order of convergence as

$$\text{EOC} \ = \frac{\ln E(h_1) - \ln E(h_2)}{\ln h_1 - \ln h_2}.$$

To investigate EOCs for our model problem we choose a sequence of uniform partitions of Ω containing five refinement levels, starting with eight triangles forming a uniform octagon as the initial triangulation of the unit disc. The corresponding grid sizes are $h_i = 2^{-i}$ for $i = 1, \ldots, 5$. As error functionals we take $E(h) = \|(u, y) - (u_h, y_h)\|$ and $E(h) = \|(u, y) - (u_h, y_h)\|_{H^1}$ and note that the error $p - p_h$ is related to $u - u_h$ via (2.10). We solve problems (2.6) using the QUADPROG routine of the MATLAB OPTIMIZATION TOOLBOX. The required finite element matrices for the discrete state and adjoint systems are generated with the help of the MATLAB PDE TOOLBOX. Furthermore, for discontinuous functions f we use the quadrature rule

$$\int_\Omega f(x)dx \approx \sum_{T \in \mathcal{T}_h} f\left(x_{s(T)}\right)|T|,$$

where $x_{s(T)}$ denotes the barycenter of T. In all computations we set $\alpha = 1$.

In Table 1, we present EOCs for problem (2.6) (case $S = D$) and the approach sketched in Remark 3.8 (case $S = M$). As one can see, the error $\|u - u_h\|$ behaves in the case $S = D$ as predicted by Theorem 3.6, whereas the errors $\|y - y_h\|$ and

TABLE 1
Experimental order of convergence.

Level	$(S=D)$ $\|u-u_h\|$	$(S=M)$ $\|u-u_h\|$	$(S=D)$ $\|y-y_h\|$	$(S=M)$ $\|y-y_h\|$	$(S=D)$ $\|y-y_h\|_{H^1}$	$(S=M)$ $\|y-y_h\|_{H^1}$
1	0.97615	0.65403	0.53646	0.69030	0.86051	0.68853
2	0.82724	1.97278	1.14786	2.01783	1.27240	2.01560
3	0.95585	1.96219	1.38937	2.00438	1.45709	2.00428
4	0.98378	1.85668	1.51838	1.98972	1.56420	1.99056
5	0.99544	1.58872	1.59842	1.97908	1.63277	1.97994

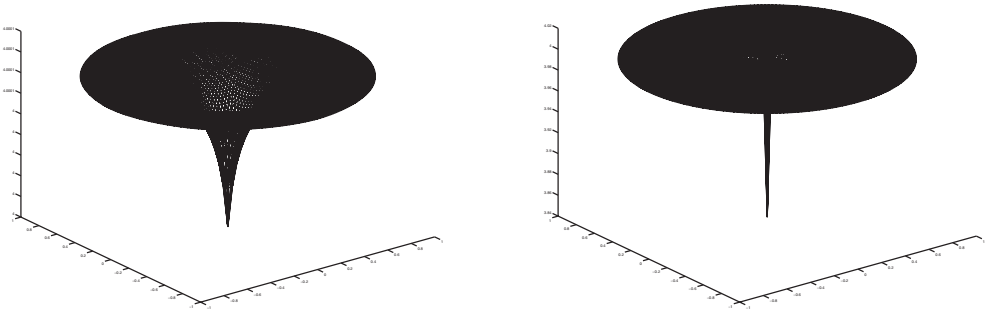


FIG. 1. *Numerically computed state y_h (left) and control u_h (right) for $h = 2^{-5}$ in the case $S = D$.*

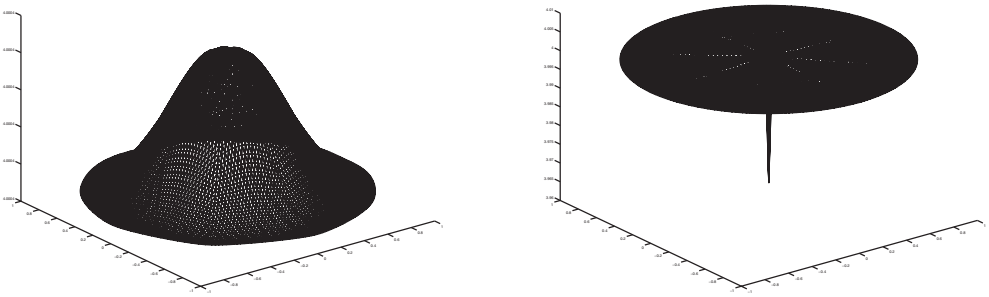


FIG. 2. *Numerically computed state y_h (left) and control u_h (right) for $h = 2^{-5}$ in the case $S = M$.*

$\|y-y_h\|_{H^1}$ show a better convergence behavior. On the finest level we have $\|u-u_h\| = 0.003117033$, $\|y-y_h\| = 0.000123186$, and $|y-y_h|_{H^1} = 0.000083757$. Furthermore, all coefficients of μ_h are equal to zero, except the one in front of δ_0, whose value is 0.99946494. The errors $\|u-u_h\|$, $\|y-y_h\|$, and $\|y-y_h\|_{H^1}$ in the case $S = M$ show a better EOC than in the case $S = D$. This can be explained by the facts that the exact solutions y and u are very smooth and that the relaxed form of the state constraints introduce a smearing effect on the numerical solutions at the origin. On the finest level we have $\|u-u_h\| = 0.001020918$, $\|y-y_h\| = 0.000652006$, and $|y-y_h|_{H^1} = 0.000037656$. Furthermore, the coefficient of μ_h corresponding to the patch containing the origin has the value 0.66505911271141.

Figures 1 and 2 present the numerical solutions y_h and u_h for $h = 2^{-5}$ in the case $S = D$ and $S = M$, respectively. We note that using equal scales on all axes would give completely flat graphs in all four figures.

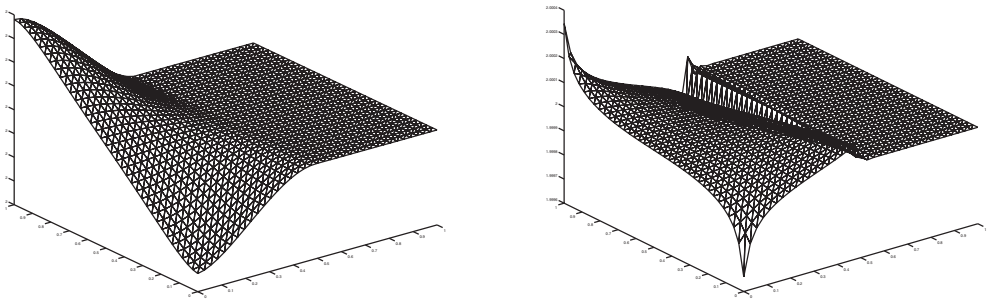


FIG. 3. *Numerically computed state y_h (left) and control u_h (right) for $h = \frac{\sqrt{2}}{36}$ in the case $S = D$.*

Example 4.2. The second test problem is taken from [11, Example 2]. It reads

$$\min_{u \in L^2(\Omega)} J(u) = \frac{1}{2}\int_\Omega |y - y_0|^2 + \frac{1}{2}\int_\Omega |u - u_0|^2$$
$$\text{subject to } y = \mathcal{G}(u) \text{ and } y(x) \geq b(x) \text{ in } \Omega.$$

Here Ω denotes the unit square,

$$b(x) = \begin{cases} 2x_1 + 1, & x_1 < \frac{1}{2}, \\ 2, & x_1 \geq \frac{1}{2}, \end{cases} \quad y_0(x) = \begin{cases} x_1^2 - \frac{1}{2}, & x_1 < \frac{1}{2}, \\ \frac{1}{4}, & x_1 = \frac{1}{2}, \\ \frac{3}{4}, & x_1 > \frac{1}{2}, \end{cases}$$

and

$$u_0(x) = \begin{cases} \frac{5}{2} - x_1^2, & x_1 < \frac{1}{2}, \\ \frac{9}{4}, & x_1 \geq \frac{1}{2}. \end{cases}$$

We remark that the inequality sign in the state constraint has been reversed in order to construct the example. The exact solution is given by $y \equiv 2$ and $u \equiv 2$ in Ω. The corresponding Lagrange multiplier $p \in H^1(\Omega)$ is given by

$$p(x) = \begin{cases} \frac{1}{2} - x_1^2, & x_1 < \frac{1}{2}, \\ \frac{1}{4}, & x_1 \geq \frac{1}{2}. \end{cases}$$

The multiplier μ has the form

$$\int_{\bar\Omega} f d\mu = \int_{\{x_1 = \frac{1}{2}\}} f ds + \int_{\{x_1 > \frac{1}{2}\}} f dx, \qquad f \in C^0(\bar\Omega). \tag{4.1}$$

In our numerical computations we use uniform grids generated with the POIMESH function of the MATLAB PDE TOOLBOX. Integrals containing y_0, u_0 are numerically evaluated by substituting y_0, u_0 by their piecewise linear, continuous finite element interpolations $I_h y_0, I_h u_0$. The grid size of a grid containing l horizontal and l vertical lines is given by $h_l = \frac{\sqrt{2}}{l+1}$. Figure 3 presents the numerical results for a grid with $h = \frac{\sqrt{2}}{36}$ in the case $(S = D)$. The corresponding values of μ_h on the same grid are presented in Figure 4. They reflect the fact that the measure consists of a lower dimensional part which is concentrated on the line $\{x \in \Omega \,|\, x_1 = \frac{1}{2}\}$ and a regular

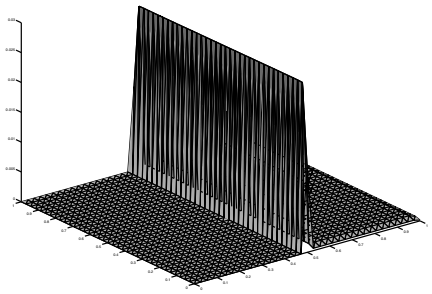


FIG. 4. *Numerically computed multiplier* μ_h *for* $h = \frac{\sqrt{2}}{36}$ *in the case* $S = D$.

TABLE 2
Experimental order of convergence, $x_1 = \frac{1}{2}$ *grid line.*

Level	$(S=D)$ $\|u-u_h\|$	$(S=M)$ $\|u-u_h\|$	$(S=D)$ $\|y-y_h\|$	$(S=M)$ $\|y-y_h\|$	$(S=D)$ $\|y-y_h\|_{H^1}$	$(S=M)$ $\|y-y_h\|_{H^1}$
1	1.669586	0.448124	1.417368	0.544284	1.594104	0.384950
2	1.922925	1.184104	1.990906	1.473143	1.992097	1.239771
3	2.000250	1.456908	2.101633	1.871948	2.080739	1.745422
4	2.029556	1.530303	2.125168	2.427634	2.108241	2.348036
5	2.041913	1.260744	2.124773	2.743918	2.116684	2.563363
6	2.047106	1.142668	2.117184	1.430239	2.117739	1.318617
7	2.048926	1.177724	2.107828	1.503463	2.115633	1.409563
8	2.049055	1.194893	2.098597	1.578342	2.112152	1.497715
9	2.048312	1.194802	2.090123	1.622459	2.108124	1.549495

TABLE 3
Experimental order of convergence, $x_1 = \frac{1}{2}$ *not a grid line.*

Level	$(S=D)$ $\|u-u_h\|$	$(S=M)$ $\|u-u_h\|$	$(S=D)$ $\|y-y_h\|$	$(S=M)$ $\|y-y_h\|$	$(S=D)$ $\|y-y_h\|_{H^1}$	$(S=M)$ $\|y-y_h\|_{H^1}$
1	0.812598	0.460528	1.160789	2.154570	0.885731	1.473561
2	1.361946	0.406917	2.042731	0.597846	1.918942	0.405390
3	1.228268	1.031763	1.832573	1.392796	1.700124	1.088595
4	1.245030	1.262257	1.678233	1.621110	1.570580	1.392408
5	1.252221	1.416990	1.646124	1.844165	1.554434	1.686808
6	1.256861	1.505759	1.696309	2.128776	1.620231	2.021210
7	1.264456	1.489061	1.627539	2.507863	1.559065	2.415552
8	1.260157	1.316627	1.640964	2.989867	1.580113	2.818148
9	1.265599	1.169109	1.686579	1.601263	1.635084	1.460153

part with a density $\chi_{|\{x_1>\frac{1}{2}\}}$. We again note that using equal scales on all axes would give completely flat graphs for y_h as well as for u_h.

We compute EOCs for the two different sequences of grid sizes $s_o = \{h_1, h_3, \dots, h_{19}\}$ and $s_e = \{h_0, h_2, \dots, h_{18}\}$. We note that the grids corresponding to s_o contain the line $x_1 = \frac{1}{2}$. Table 2 presents EOCs for s_o, and Table 3 presents EOCs for s_e. For the sequence s_o we observe superconvergence in the case $(S = D)$, although the discontinuous function y_0 for the quadrature is replaced by its piecewise linear, continuous finite element interpolant $I_h y_0$. Let us note that further numerical experiments show that the use of the quadrature rule (4.1) for integrals containing the function y_0 decreases the EOC for $\|u-u_h\|$ to $\frac{3}{2}$, whereas EOCs remain close to 2 for the other two errors $\|y-y_h\|$ and $\|y-y_h\|_{H^1}$. For this sequence also the case $(S = M)$ behaves twice as good as expected by our arguments in Remark 3.8. For the sequence s_e the

TABLE 4
Approximation of the multiplier in the case ($S = D$), $x_1 = \frac{1}{2}$ grid line.

Level	$\sum_{x_i \in \{x_1 = 1/2\}} \mu_i$	$\sum_{x_i \in \{x_1 > 1/2\}} \mu_i$
1	1.13331662624081	0.36552954225441
2	1.06315278164899	0.43644163287114
3	1.03989323182608	0.45990635060758
4	1.02893022155910	0.47095098878247
5	1.02265064139378	0.47727091447291
6	1.01855129775903	0.48139306499280
7	1.01569011772403	0.48426838085822
8	1.01359012331610	0.48637773715316
9	1.01198410389649	0.48799027450619

error $\|u - u_h\|$ in the case ($S = D$) approximately behaves as predicted by our theory; in the case ($S = M$) it behaves as for the sequence s_o. The errors $\|y - y_h\|$ and $\|y - y_h\|_{H^1}$ behave that well, since the the exact solutions y and u are very smooth. For h_{19} we have in the case ($S = D$) $\|u - u_h\| = 0.000103428$, $\|y - y_h\| = 0.000003233$, and $|y - y_h|_{H^1} = 0.000015155$, and in the case ($S = M$) $\|u - u_h\| = 0.011177577$, $\|y - y_h\| = 0.000504815$, and $|y - y_h|_{H^1} = 0.001547907$. We observe that the errors in the case $S = M$ are two magnitudes larger than in the case ($S = D$). This can be explained by the fact that an ansatz for the multiplier μ with a linear combination of Dirac measures is better suited to approximate measures concentrated on singular sets than a piecewise constant ansatz as in the case ($S = M$). Finally, Table 4 presents $\sum_{x_i \in \{x_1 = 1/2\}} \mu_i$ and $\sum_{x_i \in \{x_1 > 1/2\}} \mu_i$ for s_o in the case ($S = D$). As one can see $\sum_{x_i \in \{x_1 = 1/2\}} \mu_i$ tends to 1, the length of $\{x_1 = 1/2\}$, and $\sum_{x_i \in \{x_1 > 1/2\}} \mu_i$ tends to $1/2$, the area of $\{x_1 > 1/2\}$. These numerical findings indicate that $\mu_h = \sum_{i=1}^m \mu_i \delta_{x_i}$ well approximates μ, since $\int_{\bar{\Omega}} d\mu_h = \sum_{i=1}^m \mu_i$, and that μ_h also well resolves the structure of μ; see (4.1). For all numerical computations of this example we have $\mu_i = 0$ for $x_i \in \{x_1 < 1/2\}$.

5. Appendix.

LEMMA 5.1. *Let $\frac{2d}{d+2} \le s \le 2$ and $v \in W^{1,s}(\Omega)$. Then*

$$\|v - P_h v\| \le C h^{1 + \frac{d}{2} - \frac{d}{s}} \|v\|_{W^{1,s}}.$$

Proof. The assertion is clear if $s = \frac{2d}{d+2}$ or if $s = 2$ so that we may assume $\frac{2d}{d+2} < s < 2$. Let us write

$$\int_\Omega |v - P_h v|^2 = \int_\Omega |v - P_h v|^{\frac{sd - 2d + 2s}{s}} |v - P_h v|^{\frac{d(2-s)}{s}}$$

and apply Hölder's inequality with $p = \frac{s^2}{sd - 2d + 2s}$, $q = \frac{s^2}{(d-s)(2-s)}$, which implies

$$\begin{aligned} \|v - P_h v\|^2 &\le \|v - P_h v\|_{L^s}^{\frac{sd-2d+2s}{s}} \|v - P_h v\|_{L^{\frac{ds}{d-s}}}^{\frac{d(2-s)}{s}} \\ &\le \|v - P_h v\|_{L^s}^{\frac{sd-2d+2s}{s}} \left(\|v\|_{L^{\frac{ds}{d-s}}} + \|P_h v\|_{L^{\frac{ds}{d-s}}} \right)^{\frac{d(2-s)}{s}}. \end{aligned}$$

We infer from [6] that

$$\|v - P_h v\|_{L^s} \le C h \|v\|_{W^{1,s}}, \quad \|P_h v\|_{L^{\frac{ds}{d-s}}} \le C \|v\|_{L^{\frac{ds}{d-s}}},$$

which, together with the continuous embedding $W^{1,s}(\Omega) \hookrightarrow L^{\frac{ds}{d-s}}(\Omega)$, gives

$$\|v - P_h v\|^2 \le C h^{\frac{sd-2d+2s}{s}} \|v\|^2_{W^{1,s}}$$

so that the assertion follows. ∎

LEMMA 5.2. *Suppose that K and $\tilde{K}$ are two disjoint compact subsets of $\bar{\Omega}$. Then there exists a nonnegative function $\phi \in C^2(\bar{\Omega})$ which satisfies*

$$\partial_\nu \phi = 0 \text{ on } \partial\Omega, \quad \phi \ge 1 \text{ on } K, \quad \phi = 0 \text{ on } \tilde{K}.$$

Proof. For $r > 0$ let us define $\Omega_r := \{x \in \bar{\Omega} \,|\, \operatorname{dist}(x, \partial\Omega) < r\}$. In view of the smoothness of $\partial\Omega$ there exists $\delta > 0$ such that for each $x \in \Omega_\delta$ there exists a unique point $y = y(x) \in \partial\Omega$, with

$$x = y - \operatorname{dist}(x, \partial\Omega)\nu(y)$$

(see [7, Section 14.6]). Since $K \cap \tilde{K} = \emptyset$ we may assume that $\operatorname{dist}(K, \tilde{K}) > \delta$. Let us define

$$\Gamma_K := \{y(x) \,|\, x \in K \cap \overline{\Omega_{\frac{\delta}{2}}}\}, \quad \Gamma_{\tilde{K}} := \{y(x) \,|\, x \in \tilde{K} \cap \overline{\Omega_{\frac{\delta}{2}}}\}.$$

Γ_K and $\Gamma_{\tilde{K}}$ are disjoint, compact subsets of $\partial\Omega$, since $\operatorname{dist}(K, \tilde{K}) > \delta$ and $x \mapsto y(x)$ is continuous. Let $\phi_1 \in C^2(\partial\Omega)$ be a nonnegative function satisfying $\phi_1 \ge 1$ on Γ_K, $\phi_1 = 0$ on $\Gamma_{\tilde{K}}$. By setting $\phi_1(x) = \phi_1(y(x))$ we extend ϕ_1 as a C^2 function to Ω_δ. Clearly, $\partial_\nu \phi_1 = 0$ on $\partial\Omega$. Let $\psi \in C^2(\bar{\Omega})$ be a nonnegative cutoff function, with $\psi = 1$ in $\Omega_{\frac{\delta}{4}}$ and $\psi = 0$ in $\bar{\Omega} \setminus \Omega_{\frac{\delta}{2}}$. Then $\phi_2 := \psi\phi_1$ satisfies

$$\partial_\nu \phi_2 = 0 \text{ on } \partial\Omega, \quad \phi_2 \ge 1 \text{ on } K \cap \Omega_{\frac{\delta}{4}}, \quad \phi_2 = 0 \text{ on } \tilde{K}.$$

Finally, choose a nonnegative function $\phi_3 \in C^2(\bar{\Omega})$, with

$$\phi_3 \ge 1 \text{ on } K \cap (\bar{\Omega} \setminus \Omega_{\frac{\delta}{4}}), \quad \phi_3(x) = 0 \text{ if } \operatorname{dist}(x, K \cap (\bar{\Omega} \setminus \Omega_{\frac{\delta}{4}})) \ge \frac{\delta}{8}.$$

Then $\partial_\nu \phi_3 = 0$ on $\partial\Omega$, $\phi_3 = 0$ on $\tilde{K}$, and $\phi := \phi_2 + \phi_3$ has the required properties. ∎

Acknowledgments. We thank Ulrich Matthes (TU Dresden) for coding the numerical examples and Alan Demlow (University of Kentucky) for pointing out reference [13] to us.

REFERENCES

[1] E. CASAS, *L^2 estimates for the finite element method for the Dirichlet problem with singular data*, Numer. Math., 47 (1985), pp. 627–632.

[2] E. CASAS, *Control of an elliptic problem with pointwise state constraints*, SIAM J. Control Optim., 24 (1986), pp. 1309–1318.

[3] E. CASAS, *Boundary control of semilinear elliptic equations with pointwise state constraints*, SIAM J. Control Optim., 31 (1993), pp. 993–1006.

[4] E. CASAS, *Error estimates for the numerical approximation of semilinear elliptic control problems with finitely many state constraints*, ESAIM Control Optim. Calc. Var., 8 (2002), pp. 345–374.

[5] E. CASAS AND M. MATEOS, *Uniform convergence of the FEM. Applications to state constrained control problems*, Comput. Appl. Math. 21 (2002), pp. 67–100.

[6] J. Douglas, T. Dupont, and L. Wahlbin, *The stability in L^q of the L^2-projection into finite element function spaces*, Numer. Math., 23 (1975), pp. 193–197.

[7] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Springer, New York, 1983.

[8] M. Hintermüller and K. Kunisch, *Path following methods for a class of constrained minimization methods in function spaces*, SIAM J. Optim., 17 (2006), pp. 159–187.

[9] M. Hintermüller and K. Kunisch, *Feasible and non-interior path following in constrained minimization with low multiplier regularity*, SIAM J. Control Optim., to appear.

[10] C. Meyer, U. Prüfert, and F. Tröltzsch, *On two numerical methods for state-constrained elliptic control problems*, Optimization Methods and Software, to appear.

[11] C. Meyer, A. Rösch, and F. Tröltzsch, *Optimal control problems of PDEs with regularized pointwise state constraints*, Comput. Optim. Appl., 33 (2006), pp. 209–228.

[12] J. Nocedal and S. J. Wright, *Nonlinear Optimization*. Springer Ser. Oper. Res., Springer, New York, 1999.

[13] A. H. Schatz, *Pointwise error estimates and asymptotic error expansion inequalities for the finite element method on irregular grids.* I: *Global estimates*, Math. Comp., 67 (1998), pp. 877–899.

[14] R. Scott, *Optimal L^∞ estimates for the finite element method on irregular meshes*, Math. Comp., 30 (1976), pp. 681–897.

[15] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

[16] F. Tröltzsch, *Optimale Steuerung mit partiellen Differentialgleichungen*, Wiesbaden, Vieweg, 2005.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1954–1978

ANALYSIS OF A SPECTRAL-GALERKIN APPROXIMATION TO THE HELMHOLTZ EQUATION IN EXTERIOR DOMAINS*

JIE SHEN† AND LI-LIAN WANG‡

Abstract. An error analysis is presented for the spectral-Galerkin method to the Helmholtz equation in 2- and 3-dimensional exterior domains. The problem in unbounded domains is first reduced to a problem on a bounded domain via the Dirichlet-to-Neumann operator, and then a spectral-Galerkin method is employed to approximate the reduced problem. The error analysis is based on exploring delicate asymptotic behaviors of the Hankel functions and on deriving a priori estimates with explicit dependence on the wave number for both the continuous and the discrete problems. Explicit error bounds with respect to the wave number are derived, and some illustrative numerical examples are also presented.

Key words. Helmholtz equation, wave scattering, error analysis, spectral-Galerkin, unbounded domain

AMS subject classifications. 65N35, 65N22, 65F05, 35J05

DOI. 10.1137/060665737

1. Introduction.

We consider in this paper the acoustic wave scattering from a bounded obstacle $D \subset \mathbb{R}^d$, $d = 2, 3$. In this case, the scattered wave satisfies the Helmholtz equation

$$-\Delta U - k^2 U = F \quad \text{in } \mathbb{R}^d \backslash \bar{D}, \tag{1.1}$$

along with the Sommerfeld radiation condition at infinity

$$\partial_r U - \mathrm{i}kU = o\big(r^{\frac{1-d}{2}}\big) \quad \text{as } r \to \infty,\ d = 2, 3, \tag{1.2}$$

which ensures that waves do not reflect from the far field. On the surface of the scatterer D, a Dirichlet (sound soft) or Neumann (sound hard) condition is assumed.

Although the Helmholtz equation with (1.2) is linear, its numerical approximation and associated analysis are notoriously difficult due to the following: (i) the domain is unbounded; (ii) the system is not positive definite; and (iii) when the wave number $k \gg 1$, the solution is highly oscillatory. In particular, it remains a challenge to design numerical algorithms which are robust and efficient for moderate to high wave numbers.

There has been extensive research work devoted to overcoming these difficulties (see, for instance, [16, 23, 22] and the references therein). In particular, it has been shown, at least for some simple cases, that errors of pth order numerical methods for the Helmholtz equation behave like $O(k^{p+1}h^p)$ (see, for instance, [18, 4, 30]). Hence, high-order methods are particularly preferable for this type of problem over low-order methods. We note also that some very detailed analyses were carried out in [2, 3]

*Received by the editors July 21, 2006; accepted for publication (in revised form) May 11, 2007; published electronically August 31, 2007.
http://www.siam.org/journals/sinum/45-5/66573.html

†Department of Mathematics, Purdue University, West Lafayette, IN 47907 (shen@math.purdue.edu). The work of this author was partially supported by NFS grants DMS-0311915 and DMS-0610646.

‡Division of Mathematics, SPMS, Nanyang Technological University, 637616, Singapore (lilian@ntu.edu.sg). The work of this author was partially supported by a Start-Up grant of NTU.

on the discrete dispersive relation by the hp version of finite element method (FEM) and by the high-order discontinuous Galerkin method. These results indicate, once again, that high-order methods are preferable, if not necessary, for highly oscillatory problems.

On the other hand, the linear system from a discretization of the Helmholtz equation with moderate to high wave numbers is usually highly indefinite and difficult to solve. It is with these considerations in mind that we choose to use the transformed field expansion (TFE) method (cf. [26]), which improves over the classical field expansion method [27, 5, 6], coupled with a fast spectral-Galerkin solution (cf. [28, 29, 24]).

There are a few recent works on wave number independent boundary element methods and on error estimates with explicit dependence on wave numbers for acoustic scattering problems. In [19, 8], the authors introduced a novel Galerkin boundary element method using a graded mesh and special basis functions and derived a quasi-optimal error estimate which is independent of wave number for the Helmholtz equation in a half-plane and exterior of a convex polygon.

We now briefly describe the TFE method for a 2-dimensional (2-D) obstacle enclosed by $\{r = a + g(\theta) : 0 \le \theta < 2\pi\}$. The TFE algorithm consists of the following steps:

- Assuming F is compactly supported and choosing b such that $b > a + \max_{0\le\theta<2\pi} |g(\theta)|$ and $\mathrm{supp}F \subset \Omega_g := \{(r,\theta) : a + g(\theta) < r < b\}$, we then use the Dirichlet-to-Neumann operator T (see [15, 13] and the next section) to reduce the problem in the unbounded domain to

$$\begin{aligned} &-\Delta U - k^2 U = F \quad \text{in } \Omega_g, \\ &U|_{r=a+g(\theta)} = \xi, \qquad (\partial_r U + T(U))|_{r=b} = 0. \end{aligned} \tag{1.3}$$

- Make a change of variables

$$r' = \frac{(b-a)r - bg(\theta)}{(b-a) - g(\theta)}, \qquad \theta' = \theta, \tag{1.4}$$

which maps Ω_g to an annulus Ω_0. To simplify the notation, we still use (r,θ) to denote (r',θ') and U, F, ξ to denote the functions U, F, ξ after the change of variables. Then the problem (1.3) becomes

$$\begin{aligned} &-\Delta U - k^2 U = F + J(g,U) \quad \text{in } \Omega_0, \\ &U(a,\theta) = \xi(\theta), \ \ (\partial_r U + TU)|_{r=b} = \eta(g,U), \end{aligned} \tag{1.5}$$

where $J(g,U)$ and $\eta(g,U)$ contain differential operators with nonconstant coefficients for which a fast direct/iterative solution is not available.

- Consider the obstacle $\{(r,\theta) : r < a + g(\theta)\}$ as a perturbation of the disk $\{r < a\}$; i.e., write $g = \varepsilon h$ and expand u as

$$U(r,\theta;\varepsilon) = \sum_{n=0}^{\infty} U_n(r,\theta)\ \varepsilon^n.$$

Plugging the above expansion into (1.5) and collecting terms with ε^n, we find that [24]

$$\begin{aligned} &-\Delta U_n - k^2 U_n = \delta_{n,0} F + \tilde{J}(g, U_{n-4}, \ldots, U_{n-1}) \quad \text{in } \Omega_0, \\ &U_n(a,\theta) = \delta_{n,0}\xi(\theta), \ \ (\partial_r U_n + TU_n)|_{r=b} = \tilde{\eta}(g, U_{n-1}). \end{aligned} \tag{1.6}$$

- Solve (1.6) for $n = 0, 1, 2, \ldots$, and sum up the series by using a Padé approximation.

It is shown in [25, 26] that this TFE method is stable and robust at high order, and it is demonstrated in [24] that this method, coupled with a spectral-Galerkin solution for (1.6), is very efficient and capable of providing very accurate results for bounded obstacle scattering with moderate to high wave numbers.

Notice that the whole algorithm boils down to solving a sequence of the following nonhomogeneous Helmholtz equation in an annulus (2-D) or a spherical shell (3-D):

$$\begin{aligned} &-\Delta U - k^2 U = F \quad \text{in } \Omega_0, \\ &U(a,\theta) = \xi(\theta), \ \ (\partial_r U + TU)|_{r=b} = \eta(\theta). \end{aligned} \tag{1.7}$$

The purpose of this paper is to present a detailed error analysis of the spectral-Galerkin method for (1.7). The main difficulty here is to obtain error estimates with explicit dependence on the wave number. Among the very few results available in this regard are those in [18, 30], where the Helmholtz equation in bounded domains with a first-order approximation to the radiation boundary condition was considered and error estimates with explicit dependence on the wave numbers were derived. To the authors' best knowledge, there seems to be no rigorous error estimate available with explicit dependence on the wave number for a numerical scheme to bounded obstacle scattering.

We now introduce some notations to be used throughout this paper. Let ϖ be a given positive weight function in $I := (a, b)$. We denote by $L^2_\varpi(I)$ a Hilbert space of real or complex functions with inner product and norm

$$(u, v)_\varpi = \int_I u(r)\overline{v(r)}\varpi(r)dr, \qquad \|u\|_\varpi = \sqrt{(u, u)_\varpi},$$

respectively, where $\bar{v}$ is the complex conjugate of v. Then the weighted Sobolev spaces $H^s_\varpi(I)$ $(s = 0, 1, 2, \ldots)$ can be defined as usual with inner products, norms, and seminorms denoted by $(\cdot,\cdot)_{s,\varpi}$, $\|\cdot\|_{s,\varpi}$, and $|\cdot|_{s,\varpi}$, respectively. For real $s > 0$, $H^s_\varpi(I)$ is defined by space interpolation (cf. [20]). The subscript ϖ will be omitted from the notations in the case of $\varpi \equiv 1$. For simplicity, we denote $\partial_r^l v = \frac{d^l v}{dr^l}$, $l \geq 1$. We shall also use $(\cdot,\cdot)_\omega$ and $\|\cdot\|_\omega$ to denote the weighted inner product and the weighted L^2-norm, respectively, in two and three dimensions.

Let S be the unit circle in 2-D and the unit sphere in 3-D; we also use the nonisotropic periodic-type Sobolev space on $\Omega = S \times I$: $H^{s'}_p\big(S; H^s_\varpi(I)\big)$, $s' \geq 0$ (subscript p stands for periodicity in the θ-direction) with the norm

$$\|U\|_{H^{s'}_p(S;H^s_\varpi(I))} = \begin{cases} \Big(\sum_{|m|=0}^\infty (1+m^2)^{s'}\|\hat{u}_m\|^2_{s,\varpi}\Big)^{1/2} & \text{if } \ d = 2, \\ \Big(\sum_{m=0}^\infty \sum_{l=-m}^m (1+m)^{2s'}\|\hat{u}_{lm}\|^2_{s,\varpi}\Big)^{1/2} & \text{if } \ d = 3, \end{cases} \tag{1.8}$$

where $\{\hat{u}_m\}$ (resp., $\{\hat{u}_{lm}\}$) are the expansion coefficients of U in terms of Fourier (resp., spherical harmonic) basis, i.e.,

$$U = \sum_{|m|=0}^\infty \hat{u}_m e^{im\theta} \quad \text{or} \quad U = \sum_{m=0}^\infty \sum_{l=-m}^m \hat{u}_{lm} Y^l_m(\theta, \phi). \tag{1.9}$$

The norm of the Sobolev space $H^{s'}_p(S)$ on S can be defined in the same fashion by replacing the norm $\|\cdot\|_{s,\varpi}$ by the absolute value $|\cdot|$. In particular, we have

$$L^2_p(\Omega) = H^0_p(S, H^0_\varpi(I)) \ \text{ with } \ \varpi = r^{d-1}, \ \ d = 2, 3.$$

We assume that $a > 0$ is a fixed parameter representing the radius of the scatterer. Throughout this paper, we denote by c a generic positive constant which depends only on a and possibly on a fixed $k_0 > 0$. We use the expression $A \lesssim B$ to mean that $A \leq cB$.

2. Dirichlet-to-Neumann (DtN) map. The error analysis relies heavily on the properties of the DtN map which we investigate below.

2.1. Formulation of the DtN operator. We start with the 3-D case and consider an "auxiliary" exterior problem

$$\begin{cases} -\Delta U - k^2 U = 0 & \text{in } \Omega_{\text{ext}} := \mathbb{R}^3 \setminus \bar{B}, \\ U = \Psi & \text{on } \partial B, \end{cases} \tag{2.1}$$

where B is a ball of radius b. This problem can be solved analytically via separation of variables; namely, we can express its solution as

$$U(r,\theta,\phi) = \sum_{m=0}^{\infty} h_m^{(1)}(kr) \sum_{l=-m}^{m} \hat{u}_{lm} Y_m^l(\theta,\phi), \tag{2.2}$$

where $(r,\theta,\phi) \in [b,\infty) \times [0,2\pi) \times [0,\pi)$, $h_m^{(1)}(z)$ is the spherical Hankel function of the first kind of order m, and $\{Y_m^l\}$ are the spherical harmonic functions. To determine the coefficients $\{\hat{u}_{lm}\}$, we expand the Dirichlet boundary value Ψ on the sphere ∂B as

$$U(b,\theta,\phi) = \Psi(\theta,\phi) = \sum_{m=0}^{\infty} \sum_{l=-m}^{m} \hat{\psi}_{lm} Y_m^l(\theta,\phi). \tag{2.3}$$

Letting $r = b$ in (2.2) and comparing the coefficients of the two expansions yield that

$$\hat{u}_{lm} = \frac{\hat{\psi}_{lm}}{h_m^{(1)}(kb)} \quad \text{for} \quad m \geq |l| \geq 0. \tag{2.4}$$

Plugging it into (2.2) leads to the exact solution of (2.1):

$$U(r,\theta,\phi) = \sum_{m=0}^{\infty} \frac{h_m^{(1)}(kr)}{h_m^{(1)}(kb)} \sum_{l=-m}^{m} \hat{\psi}_{lm} Y_m^l(\theta,\phi). \tag{2.5}$$

Differentiating (2.5) with respect to r and setting $r = b$, we find

$$\partial_r U(b,\theta,\phi) = \sum_{m=0}^{\infty} k \frac{h_m^{(1)'}(kb)}{h_m^{(1)}(kb)} \sum_{l=-m}^{m} \hat{\psi}_{lm} Y_m^l(\theta,\phi). \tag{2.6}$$

Hence, the DtN map is defined explicitly as

$$T(U) = \frac{\partial U}{\partial \mathbf{n}}\Big|_{\partial B} = -\frac{\partial U}{\partial r}\Big|_{r=b} = -\sum_{m=0}^{\infty} k \frac{h_m^{(1)'}(kb)}{h_m^{(1)}(kb)} \sum_{l=-m}^{m} \hat{\psi}_{lm} Y_m^l(\theta,\phi), \tag{2.7}$$

where $\mathbf{n}$ is the outward normal of Ω_{ext}.

The counterpart of (2.1) in 2-D is

$$\begin{cases} -\Delta U - k^2 U = 0 & \text{in } \Omega_{\text{ext}} := \mathbb{R}^2 \setminus \bar{B}, \\ U = \Phi & \text{on } \partial B, \end{cases} \tag{2.8}$$

where B is a circle of radius b which can be solved analytically with the exact solution

$$U(r,\theta) = \sum_{|m|=0}^{\infty} \hat{u}_m H_m^{(1)}(kr) e^{\mathrm{i}m\theta} \quad \forall (r,\theta) \in [b,\infty) \times [0,2\pi). \tag{2.9}$$

Here $H_m^{(1)}(z)$ is the Hankel function of the first kind of order m. The coefficients $\{\hat{u}_m\}$ are determined by the boundary value $\Phi(\theta)$ with the expansion

$$U(b,\theta) = \Phi(\theta) = \sum_{|m|=0}^{\infty} \hat{\phi}_m e^{\mathrm{i}m\theta}. \tag{2.10}$$

Hence, letting $r = b$ in (2.9) and comparing the coefficients of the above two expansions lead to $\hat{u}_m = \hat{\phi}_m / H_m^{(1)}(kb)$. As a consequence, the exact solution of (2.8) is

$$U(r,\theta) = \sum_{|m|=0}^{\infty} \frac{H_m^{(1)}(kr)}{H_m^{(1)}(kb)} \hat{\phi}_m e^{\mathrm{i}m\theta} \quad \forall (r,\theta) \in [b,\infty) \times [0,2\pi). \tag{2.11}$$

The 2-D DtN map is given by

$$T(U) = \frac{\partial U}{\partial \mathbf{n}}\Big|_{\partial B} = -\frac{\partial U}{\partial r}\Big|_{r=b} = -\sum_{|m|=0}^{\infty} k \frac{H_m^{(1)'}(kb)}{H_m^{(1)}(kb)} \hat{\phi}_m e^{\mathrm{i}m\theta}. \tag{2.12}$$

By using the DtN map T and choosing b sufficiently large so that B contains both D and supp$F$, the original problem (1.1)–(1.2) with a Dirichlet boundary condition is reduced to:

$$\begin{cases} -\Delta U - k^2 U = F & \text{in } \Omega := B \cap \mathbb{R}^d \setminus \bar{D},\ d = 2,3, \\ U = \xi & \text{on } \partial D, \\ \partial_r U + TU = 0 & \text{on } \partial B. \end{cases} \tag{2.13}$$

To fix the idea, we prescribed a Dirichlet boundary condition on the scatterer D; other types of boundary conditions can be used as well.

2.2. Properties of the DtN kernel. In order to carry out a rigorous mathematical analysis for the problem (2.13), we need to study carefully the properties of the DtN kernel associated with (2.7) and (2.12), i.e., the properties of the coefficients:

$$\mathcal{T}_{m,\kappa} = \begin{cases} \dfrac{H_m^{(1)'}(\kappa)}{H_m^{(1)}(\kappa)} & \text{if } d = 2, \\ \dfrac{h_m^{(1)'}(\kappa)}{h_m^{(1)}(\kappa)} & \text{if } d = 3. \end{cases} \tag{2.14}$$

2.2.1. Behavior of the 3-D kernel. In this case, we have $\kappa > 0$ and $m \geq 0$. We recall that

$$h_m^{(1)}(\kappa) = j_m(\kappa) + \mathrm{i}y_m(\kappa) = \sqrt{\frac{\pi}{2\kappa}}J_{m+1/2}(\kappa) + \mathrm{i}\sqrt{\frac{\pi}{2\kappa}}Y_{m+1/2}(\kappa), \tag{2.15}$$

where J_ν and Y_ν (resp., j_ν and y_ν) are the Bessel (resp., spherical Bessel) functions of the first and second kinds, respectively, of order ν. Using the relevant properties of the Bessel functions (cf. [31]), one verifies that

$$\begin{aligned}\mathrm{Re}(\mathcal{T}_{m,\kappa}) &= \frac{m}{\kappa} - \frac{j_m(\kappa)j_{m+1}(\kappa) + y_m(\kappa)y_{m+1}(\kappa)}{\big|h_m^{(1)}(\kappa)\big|^2} \\ &= \frac{m}{\kappa} - \frac{J_{m+1/2}(\kappa)J_{m+3/2}(\kappa) + Y_{m+1/2}(\kappa)Y_{m+3/2}(\kappa)}{J_{m+1/2}^2(\kappa) + Y_{m+1/2}^2(\kappa)};\end{aligned} \tag{2.16a}$$

$$\mathrm{Im}(\mathcal{T}_{m,\kappa}) = \frac{1}{\kappa^2\big|h_m^{(1)}(\kappa)\big|^2} = \frac{2}{\pi\kappa}\frac{1}{J_{m+1/2}^2(\kappa) + Y_{m+1/2}^2(\kappa)}. \tag{2.16b}$$

An explicit expression of $\mathcal{T}_{m,\kappa}$ is given by Theorem 2.6.1 of [23]:

$$\mathcal{T}_{m,\kappa} = \mathrm{Re}(\mathcal{T}_{m,\kappa}) + \mathrm{i}\,\mathrm{Im}(\mathcal{T}_{m,\kappa}) = -\frac{P_m(\kappa)}{\kappa Q_m(\kappa)} + \frac{\mathrm{i}}{Q_m(\kappa)}, \tag{2.17}$$

where

$$\begin{aligned}P_m(\kappa) &= 1 + 2a_1^m\frac{1}{\kappa^2} + \cdots + (m+1)a_m^m\frac{1}{\kappa^{2m}}, \\ Q_m(\kappa) &= 1 + a_1^m\frac{1}{\kappa^2} + \cdots + a_m^m\frac{1}{\kappa^{2m}},\end{aligned} \tag{2.18}$$

with

$$a_j^m = \frac{(m+j)!(2j)!}{4^j(j!)^2(m-j)!}. \tag{2.19}$$

We now study the monotonic property of $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ with respect to m and κ. We observe from (2.17)–(2.19) that, for a fixed $m \geq 0$, $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is an *increasing* function of κ, as illustrated by Figure 2.1(b). However, for a given $\kappa > 0$, $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is a *decreasing* function of m, which follows from Nicholson's formula (see p. 444 of [31])

$$J_{m+1/2}^2(\kappa) + Y_{m+1/2}^2(\kappa) = \frac{8}{\pi^2}\int_0^{+\infty}\mathcal{K}_0(2\kappa\sinh t)\cosh\big((2m+1)t\big)dt, \tag{2.20}$$

where $\mathcal{K}_0(\xi) > 0$ is Kelvin's function defined by (A.2) in the appendix.

We next consider the bounds and asymptotic behavior of $\mathcal{T}_{m,\kappa}$. An immediate consequence of (2.17)–(2.19) is that

$$\mathrm{Re}(\mathcal{T}_{m,\kappa}) < 0, \qquad \mathrm{Im}(\mathcal{T}_{m,\kappa}) > 0, \tag{2.21}$$

which ensures the well-posedness of the problem (2.13) (cf. [11]). Moreover, we have the following bounds (see, e.g., p. 87 of [23]):

$$-\frac{m+1}{\kappa} \leq \mathrm{Re}(\mathcal{T}_{m,\kappa}) \leq -\frac{1}{\kappa}, \qquad 0 < \mathrm{Im}(\mathcal{T}_{m,\kappa}) \leq 1, \tag{2.22}$$

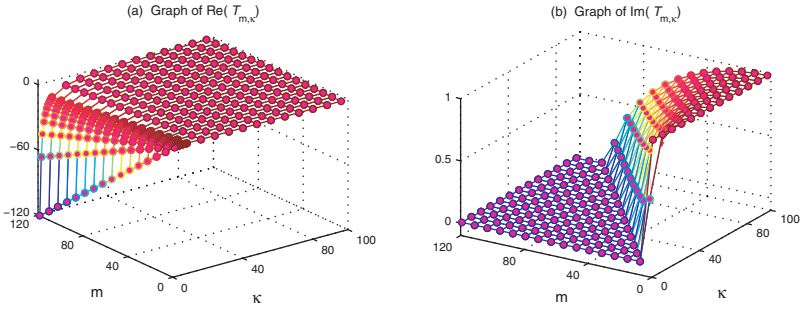


FIG. 2.1. *Graphs of* $\mathrm{Re}(\mathcal{T}_{m,\kappa})$ *and* $\mathrm{Im}(\mathcal{T}_{m,\kappa})$, *with* $(\kappa, m) \in [1{,}100] \times [0{,}120]$, *in the* 3-*D case.*

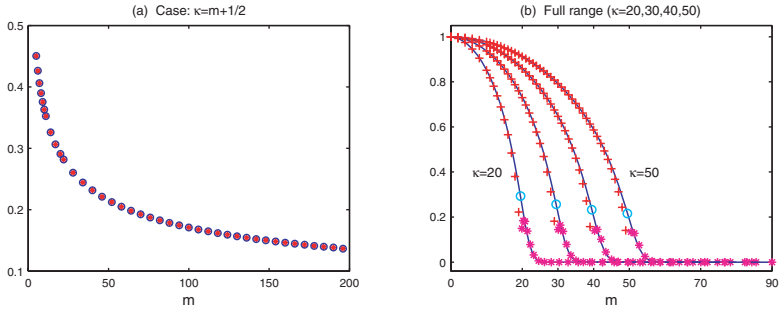


FIG. 2.2. (a) $\mathrm{Im}(\mathcal{T}_{m,m+1/2})$ ($\star$) *against* $E_{m,m+1/2}$ ($\circ$) *with* $m \in [1, 200]$; (b) $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ *(solid line) against* $E_{m,\kappa}$ ($+$ *for* $\kappa > m + 1/2$, ($\circ$) *for* $k = m + 1/2$, *and* ($\star$) *for* $\kappa < m + 1/2$), *with* $\kappa = 20, 30, 40, 50$.

in particular, by (2.17)–(2.19),

$$\mathrm{Re}(\mathcal{T}_{0,\kappa}) = -\frac{1}{\kappa}, \qquad \mathrm{Im}(\mathcal{T}_{0,\kappa}) = 1. \tag{2.23}$$

We now seek more precise estimates of $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ and proceed separately with three cases:

(i) $\kappa > m + 1/2$. We first recall the estimate (see p. 447 of [31])

$$\frac{2}{\pi\kappa} < J_\nu^2(\kappa) + Y_\nu^2(\kappa) < \frac{2}{\pi\sqrt{\kappa^2 - \nu^2}} \quad \text{if} \quad \frac{1}{2} \le \nu < \kappa, \tag{2.24}$$

which, together with (2.16b), implies that

$$E_{m,\kappa} := \frac{\sqrt{\kappa^2 - (m + 1/2)^2}}{\kappa} < \mathrm{Im}(\mathcal{T}_{m,\kappa}) < 1 \quad \text{if} \quad \kappa > m + \frac{1}{2}. \tag{2.25}$$

We observe from Figure 2.2 that the lower bound $E_{m,\kappa}$ provides an acceptable approximation to $\mathrm{Im}(\mathcal{T}_{m,\kappa})$.

(ii) $\kappa = m + 1/2$. Using the formulas (see p. 232 of [31])

$$J_\nu(\nu) = C_1\nu^{-1/3} + O(\nu^{-5/3}), \qquad Y_\nu(\nu) = -C_2\nu^{-1/3} + O(\nu^{-5/3}), \tag{2.26}$$

with

$$C_1 = \frac{\Gamma(1/3)}{2^{2/3}3^{1/6}\pi} \approx 0.4473, \quad C_2 = \frac{3^{1/3}\Gamma(1/3)}{2^{2/3}\pi} \approx 0.7748,$$

we obtain from (2.16b) that

$$\mathrm{Im}\big(\mathcal{T}_{m,m+1/2}\big) \sim C_\circ \big(m+1/2\big)^{-1/3} := E_{m,m+1/2}, \tag{2.27}$$

with $C_\circ = 2/\big(\pi(C_1^2+C_2^2)\big) \approx 0.7954$.

In Figure 2.2(a), we plot $\mathrm{Im}(\mathcal{T}_{m,m+1/2})$ against $E_{m,m+1/2}$ for $1 \le m \le 200$, which shows that, even for small m, the asymptotic estimate $C_\circ\big(m+1/2\big)^{-1/3}$ provides a very good approximation to $\mathrm{Im}\big(\mathcal{T}_{m,m+1/2}\big)$.

(iii) $\kappa < m+1/2$. By the asymptotic formulas (see p. 243 of [31])

$$J_\nu(\nu\mathrm{sech}\alpha) \sim \frac{e^{\nu(\tanh\alpha-\alpha)}}{\sqrt{2\pi\nu\tanh\alpha}}, \qquad Y_\nu(\nu\mathrm{sech}\alpha) \sim -\frac{e^{\nu(\alpha-\tanh\alpha)}}{\sqrt{\frac{1}{2}\pi\nu\tanh\alpha}}, \tag{2.28}$$

one verifies that for $m+1/2 = \kappa\cosh\alpha$, with $\alpha > 0$,

$$\mathrm{Im}(\mathcal{T}_{m,\kappa}) \sim \frac{2(2m+1)\tanh\alpha}{\kappa\big[e^{(2m+1)(\tanh\alpha-\alpha)} + 4e^{(2m+1)(\alpha-\tanh\alpha)}\big]} := E_{m,\kappa}. \tag{2.29}$$

Hence, $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ becomes exponentially small for large m. The exponential decay of $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is shown more clearly from the asymptotic estimate

$$\mathrm{Im}(\mathcal{T}_{m,\kappa}) \sim \Big(\frac{e\kappa}{2m+1}\Big)^{2m}, \quad m \gg \kappa, \tag{2.30}$$

which follows from formula 9.3.1 of [1]:

$$J_\nu(\kappa) \sim \frac{1}{\sqrt{2\pi\nu}}\Big(\frac{e\kappa}{2\nu}\Big)^{\nu}, \quad Y_\nu(\kappa) \sim -\frac{2}{\sqrt{\pi\nu}}\Big(\frac{e\kappa}{2\nu}\Big)^{-\nu}, \ \ \nu \gg \kappa. \tag{2.31}$$

We plot in Figure 2.2(b) the estimate $E_{m,\kappa}$ (defined in (2.25), (2.27), and (2.29)) versus $\mathrm{Im}(\mathcal{T}_{m,\kappa})$, with $\kappa = 20, 30, 40, 50$ and various m, which indicates that $E_{m,\kappa}$ provides an accurate picture of $\mathrm{Im}(\mathcal{T}_{m,\kappa})$.

2.2.2. Behavior of the 2-D kernel. The identity $H_{-\nu}^{(1)}(z) = e^{\nu\pi\mathrm{i}}H_\nu^{(1)}(z)$ and the definition (2.14) imply that

$$\mathcal{T}_{-m,\kappa} = \frac{H_{-m}^{(1)}{}'(\kappa)}{H_{-m}^{(1)}(\kappa)} = \frac{(-1)^m H_m^{(1)}{}'(\kappa)}{(-1)^m H_m^{(1)}(\kappa)} = \mathcal{T}_{m,\kappa}. \tag{2.32}$$

Hence, it suffices to consider $\mathcal{T}_{m,\kappa}$ with $m \ge 0$. Using the recursion formulas of the Bessel functions, one verifies that

$$\mathrm{Re}(\mathcal{T}_{m,\kappa}) = \frac{m}{\kappa} - \frac{J_m(\kappa)J_{m+1}(\kappa) + Y_m(\kappa)Y_{m+1}(\kappa)}{J_m^2(\kappa)+Y_m^2(\kappa)}; \tag{2.33a}$$

$$\mathrm{Im}(\mathcal{T}_{m,\kappa}) = \frac{2}{\pi\kappa}\frac{1}{|H_m^{(1)}(\kappa)|^2} = \frac{2}{\pi\kappa}\frac{1}{J_m^2(\kappa)+Y_m^2(\kappa)}. \tag{2.33b}$$

We observe that the 2-D kernel has an expression similar to that of the 3-D kernel (cf. (2.16)). In fact, they share similar properties and asymptotic behaviors except for $m = 0$ (comparison: Figure 2.1(a) versus Figure 2.3(a) and Figure 2.2(b) versus Figure 2.3(b)).

Indeed, we notice that the same monotonic property holds for the 2-D $\mathrm{Im}(\mathcal{T}_{m,\kappa})$:

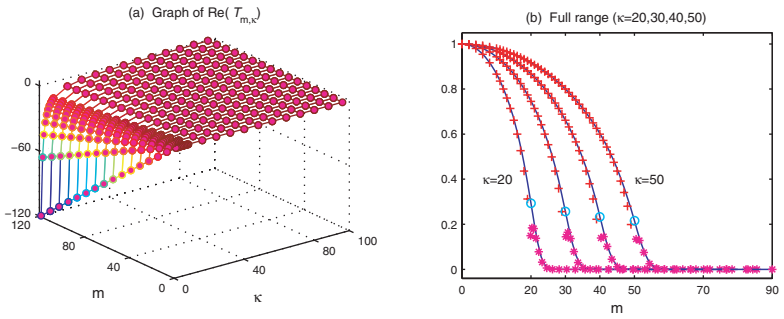


FIG. 2.3. (a) *Graph of* 2-*D* $\mathrm{Re}(\mathcal{T}_{m,\kappa})$, *with* $(\kappa, m) \in [1{,}100] \times [1{,}120]$; (b) 2-*D* $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ *(solid line) against* $E_{m,\kappa}$ *defined in* (2.35) (+ *for* $\kappa > m+1/2$, ○ *for* $\kappa = m+1/2$, *and* (⋆) *for* $\kappa < m+1/2$), *with* $\kappa = 20, 30, 40, 50$.

(i) For a given $m \geq 1$, $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is a strictly *increasing* function of κ, which follows from (2.33b) and the fact that $\kappa|H_m^{(1)}(\kappa)|^2$ is a strictly decreasing function of κ (cf. p. 446 of [31]);

(ii) for a fixed $\kappa > 0$, $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is a strictly *decreasing* function of m, which is a direct consequence of Nicholson's formula (A.3a).

As in (2.22), we have the following bound for the 2-D kernel (see the appendix for the proof):

$$0 < \mathrm{Im}(\mathcal{T}_{m,\kappa}) < 1, \ m \geq 1; \tag{2.34a}$$

$$-\frac{m}{\kappa} \leq \mathrm{Re}(\mathcal{T}_{m,\kappa}) \leq -\frac{1}{2\kappa}, \quad m \geq 1; \qquad -\frac{1}{2\kappa} \leq \mathrm{Re}(\mathcal{T}_{0,\kappa}) < 0; \tag{2.34b}$$

$$\mathrm{Im}(\mathcal{T}_{0,\kappa}) > 1 \ \ \forall \kappa > 0. \tag{2.34c}$$

As in the 3-D case, applying the general formulas (2.24), (2.26), and (2.28) to the 2-D $\mathrm{Im}(\mathcal{T}_{m,\kappa})$, we find that an accurate approximation for $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is

$$E_{m,\kappa} := \begin{cases} \sqrt{1 - m^2/\kappa^2} & \text{if } \ \kappa > m \geq 1, \\ C_\circ m^{-1/3} & \text{if } \ \kappa = m, \\ \dfrac{4m \tanh\alpha}{\kappa\left[e^{2m(\tanh\alpha - \alpha)} + 4e^{2m(\alpha - \tanh\alpha)}\right]} & \text{if } \ \kappa = m\,\mathrm{sech}\alpha, \ \alpha > 0, \end{cases} \tag{2.35}$$

where the constant $C_\circ$ is defined by (2.27).

In Figure 2.3(b), we plot $E_{m,\kappa}$ against $\mathrm{Im}(\mathcal{T}_{m,\kappa})$, which indicates that the estimate $E_{m,\kappa}$ gives an accurate picture of the behavior of $\mathrm{Im}(\mathcal{T}_{m,\kappa})$.

3. A priori estimates. In order to carry out error analysis for the spectral-Galerkin approximation to (1.7), we need to establish some a priori estimates for the solution of (1.7). Without loss of generality, we shall set $\xi = 0$ since the nonhomogeneous boundary condition at $r = a$ can be simply converted to a homogeneous one by subtracting a suitable function from the solution.

3.1. Dimension reduction. We now rewrite (1.7) with $\xi = 0$ in polar coordinates (r, θ) or spherical coordinates (r, θ, ϕ):

$$\begin{cases} -\Big(\dfrac{1}{r^{d-1}}\partial_r\big(r^{d-1}\partial_r U\big) + \dfrac{1}{r^2}\Delta_S U\Big) - k^2 U = F \quad \text{in } \Omega = (a,b) \times S, \\ U\big|_{r=a} = 0, \quad \big[\partial_r U + T(U)\big]\big|_{r=b} = \eta, \end{cases} \tag{3.1}$$

where

$$\Delta_S U = \begin{cases} \partial_\theta^2 U & \text{if } d=2, \\ \frac{1}{\sin^2\phi}\partial_\theta^2 U + \frac{1}{\sin\phi}\partial_\phi\big(\sin\phi\partial_\phi U\big) & \text{if } d=3, \end{cases} \tag{3.2}$$

and its eigenfunctions are the Fourier basis $\{e^{im\theta}\}$ (in 2-D) or the spherical harmonic functions $\{Y_m^l(\theta,\phi)\}$ (in 3-D), i.e.,

$$-\Delta_S e^{im\theta} = m^2 e^{im\theta}\ (d=2);\quad -\Delta_S Y_m^l(\theta,\phi) = m(m+1)Y_m^l(\theta,\phi)\ (d=3). \tag{3.3}$$

We shall denote

$$\beta_m = \begin{cases} m^2, & m=0,\pm1,\pm2,\ldots, \quad \text{if } d=2, \\ m(m+1), & m=0,1,2,\ldots, \quad \text{if } d=3. \end{cases} \tag{3.4}$$

Expanding the solution and given data in terms of the eigenfunctions of Δ_S :

$$\big(U,\ F,\ \eta\big) = \begin{cases} \sum_{|m|=0}^{\infty}\big(\hat{u}_m(r),\ \hat{f}_m(r),\ \hat{h}_m\big)e^{im\theta} & \text{if } d=2, \\ \sum_{m=0}^{\infty}\sum_{l=-m}^{m}\big(\hat{u}_{lm}(r),\ \hat{f}_{lm}(r),\ \hat{h}_{lm}\big)Y_m^l(\theta,\phi) & \text{if } d=3, \end{cases} \tag{3.5}$$

we find from (3.3) that the problem (3.1)–(3.2) is reduced to the following sequence of 1-dimensional equations (for brevity, we use u to denote $\hat{u}_m$ or $\hat{u}_{lm}$ and likewise for f and h below):

$$\begin{cases} -\dfrac{1}{r^{d-1}}\dfrac{d}{dr}\Big[r^{d-1}\dfrac{du}{dr}\Big] + \beta_m\dfrac{u}{r^2} - k^2 u = f, \quad r\in(a,b),\ d=2,3, \\ u(a)=0, \qquad u'(b) - kT_{m,k}u(b) = h, \end{cases} \tag{3.6}$$

where $T_{m,k}$ is derived from (2.7) and (2.12):

$$T_{m,k} = \begin{cases} \dfrac{H_m^{(1)\prime}(kb)}{H_m^{(1)}(kb)} & \text{if } d=2, \\ \dfrac{h_m^{(1)\prime}(kb)}{h_m^{(1)}(kb)} & \text{if } d=3. \end{cases} \tag{3.7}$$

Notice that $T_{m,k} = \mathcal{T}_{m,kb}$ (defined by (2.14)).

3.2. Variational formulation and a priori estimates. We denote the weight functions $\omega^\alpha(r) = r^\alpha$ and $\omega(r) = r$. Define the 1-D weighted space

$$X := X(d) = \big\{u\in H^1_{\omega^{d-1}}(I)\cap L^2_{\omega^{d-3}}(I):\ \ u(a)=0\big\}. \tag{3.8}$$

We define a bilinear form on $H_p^1(S;X)\times H_p^1(S;X)$:

$$\begin{aligned} \mathcal{B}(U,V) =& \big(\partial_r U,\partial_r V\big)_{\omega^{d-1}} + \big(\nabla_S U,\nabla_S V\big)_{\omega^{d-3}} - k^2\big(U,V\big)_{\omega^{d-1}} \\ &+ b^{d-1}\big\langle T(U)(\cdot),V(b,\cdot)\big\rangle_S, \end{aligned} \tag{3.9}$$

where $\langle\cdot,\cdot\rangle_S$ is the $L^2(S)$-inner product (cf. the appendix), and the gradient operator ∇_S is defined by

$$\nabla_S U = \begin{cases} \partial_\theta U & \text{if } d=2, \\ \Big(\frac{1}{\sin\theta}\partial_\theta U\Big)\vec{e}_\theta + \big(\partial_\phi U\big)\vec{e}_\phi & \text{if } d=3. \end{cases} \tag{3.10}$$

The variational formulation of (3.1) is as follows: Given $F \in L^2_{\omega^{d-1}}(\Omega)$ and $\eta \in L^2(S)$, find $U \in H^1_p(S;X)$ such that

$$\mathcal{B}(U,V) = (F,V)_{\omega^{d-1}} + b^{d-1}\big\langle \eta, V(b,\cdot)\big\rangle_S \quad \forall\, V \in H^1_p(S;X),\ \ d=2,3, \tag{3.11}$$

which admits a unique solution (see, e.g., [23]). The first main result of this paper is the following a priori estimates.

THEOREM 3.1. *Let U be the solution of* (3.11). *If $F \in L^2(\Omega)$ and $\eta \in L^2(S)$, then we have*

$$\|\nabla U\| + k\|U\|_\Omega \lesssim \Big(\sqrt{b^d} + \sqrt{b|I|}(kb)^{1/3}\Big)\|\eta\|_{L^2(S)} + (kb)^{1/3}|I|\|F\|, \tag{3.12}$$

where $|I| = b - a$.

The rest of this section is devoted to the proof of this estimate. Observe that, for each mode m or (l,m), the expansion coefficient $u = \hat{u}_m$ or $\hat{u}_{lm}$ (cf. (3.5)) satisfies the following reduced problem (i.e., the variational formulation of (3.6)–(3.7)):

$$\begin{aligned} &\text{Given } f \in L^2_{\omega^{d-1}}(I) \text{ and } h \in \mathbb{C}, \text{ find } u \in X \text{ such that}\\ &\mathcal{B}_m(u,v) = (f,v)_{\omega^{d-1}} + b^{d-1}h\overline{v(b)} \quad \forall v \in X,\ \ d=2,3, \end{aligned} \tag{3.13}$$

where $f = \hat{f}_m$ or $\hat{f}_{lm}$, $h = \hat{h}_m$ or $\hat{h}_{lm}$, and the sesquilinear form

$$\mathcal{B}_m(u,v) := (\partial_r u, \partial_r v)_{\omega^{d-1}} + \beta_m(u,v)_{\omega^{d-3}} - k^2(u,v)_{\omega^{d-1}} - kb^{d-1}T_{m,k}u(b)\overline{v(b)}, \tag{3.14}$$

where β_m is defined in (3.4).

An essential step is to derive a priori estimates for each $u = \hat{u}_m$ or $\hat{u}_{lm}$ and then combine these estimates to get the desired result for the original problem (3.11).

We have the following a priori estimate for the solution of (3.13)–(3.14).

LEMMA 3.1. *Let $|I| = b-a$ be the length of the interval $I = (a,b)$. If $f \in L^2_{\omega^{d-1}}(I)$, then given $k_0 > 0$, we have that, for $k \ge k_0$ and $d = 2,3$,*

$$\begin{aligned} &\|\partial_r u\|_{\omega^{d-1}} + \sqrt{\beta_m}\|u\|_{\omega^{d-3}} + k\|u\|_{\omega^{d-1}}\\ &\qquad \lesssim \big(\sqrt{b^d} + \sqrt{b|I|}C_{m,k}\big)|h| + C_{m,k}|I|\|f\|_{\omega^{d-1}}, \end{aligned} \tag{3.15}$$

where

$$C_{m,k} = \begin{cases} (kb)^{\frac{1}{3}} & \text{if} \quad |m| \le kb,\\ 1 & \text{if} \quad |m| > kb. \end{cases} \tag{3.16}$$

Proof. Some early work (cf. [12, 17, 18]) in this direction relies on the explicit form of Green's function which is very difficult, if not possible, to extend to more general cases. Our proof is based on an argument in [21, 10] (see also [30, 9]). More precisely, we take two test functions $v = u,\ (r-a)\partial_r u \in X$ in (3.13) successively to obtain a priori estimates without using Green's functions. In the following, $\varepsilon_j > 0$ $(j = 1,\dots,5)$ are some suitable real numbers.

We first take $v = u$ in (3.13). The imaginary and real parts are, respectively,

$$-kb^{d-1}\mathrm{Im}(T_{m,k})|u(b)|^2 = b^{d-1}\mathrm{Im}(h\overline{u(b)}) + \mathrm{Im}(f,u)_{\omega^{d-1}}, \tag{3.17a}$$

$$\begin{aligned} &\|\partial_r u\|^2_{\omega^{d-1}} + \beta_m\|u\|^2_{\omega^{d-3}} - k^2\|u\|^2_{\omega^{d-1}} - kb^{d-1}\mathrm{Re}(T_{m,k})|u(b)|^2\\ &\qquad = b^{d-1}\mathrm{Re}(h\overline{u(b)}) + \mathrm{Re}(f,u)_{\omega^{d-1}}. \end{aligned} \tag{3.17b}$$

In order to derive an upper bound for $\|\partial_r u\|^2_{\omega^{d-1}} + \beta_m \|u\|^2_{\omega^{d-3}}$, we proceed separately with two cases: (i) $d = 2,\ |m| > 0$ or $d = 3,\ m \geq 0$ and (ii) $d = 2,\ m = 0$. In the first case, we have from (2.22) and (2.34b) with $|m| \geq 1$ (note that $\kappa = kb$) that

$$\frac{1}{k|\mathrm{Re}(T_{m,k})|} \leq b \quad \text{for } d = 2,\ |m| > 0 \text{ or } d = 3,\ m \geq 0. \tag{3.18}$$

In what follows, we shall repeatedly use the inequality $2AB \leq \varepsilon A^2 + \frac{B^2}{\varepsilon}$ for all $A, B, \varepsilon > 0$.

Applying the Cauchy–Schwarz inequality to (3.17b) leads to

$$\begin{aligned} &\|\partial_r u\|^2_{\omega^{d-1}} + \beta_m \|u\|^2_{\omega^{d-3}} - kb^{d-1}\mathrm{Re}(T_{m,k})|u(b)|^2 \\ &\qquad \leq k^2 \|u\|^2_{\omega^{d-1}} + \frac{kb^{d-1}|\mathrm{Re}(T_{m,k})|}{2}|u(b)|^2 + \frac{b^{d-1}}{2k|\mathrm{Re}(T_{m,k})|}|h|^2 \\ &\qquad\quad + \varepsilon_1 k^2 \|u\|^2_{\omega^{d-1}} + \frac{1}{4\varepsilon_1 k^2}\|f\|^2_{\omega^{d-1}}. \end{aligned} \tag{3.19}$$

Thus, by (3.18), the estimate (3.19) becomes (for $d = 2,\ |m| > 0$ or $d = 3,\ m \geq 0$)

$$\begin{aligned} &\|\partial_r u\|^2_{\omega^{d-1}} + \beta_m \|u\|^2_{\omega^{d-3}} - \frac{kb^{d-1}\mathrm{Re}(T_{m,k})}{2}|u(b)|^2 \\ &\qquad \leq (1+\varepsilon_1) k^2 \|u\|^2_{\omega^{d-1}} + \frac{b^d}{2}|h|^2 + \frac{1}{4\varepsilon_1 k^2}\|f\|^2_{\omega^{d-1}}. \end{aligned} \tag{3.20}$$

To treat the only remaining case, (ii) $d = 2$ and $m = 0$, we apply the Cauchy–Schwarz inequality to (3.17a) and get that

$$\begin{aligned} kb\mathrm{Im}(T_{0,k})|u(b)|^2 \leq{}& \frac{kb\mathrm{Im}(T_{0,k})}{2}|u(b)|^2 + \frac{b}{2k\mathrm{Im}(T_{0,k})}|h|^2 \\ &+ \frac{\varepsilon_2 k\mathrm{Im}(T_{0,k})}{2}\|u\|^2_\omega + \frac{1}{2\varepsilon_2 k\mathrm{Im}(T_{0,k})}\|f\|^2_\omega, \end{aligned} \tag{3.21}$$

which implies that

$$k^2 b|u(b)|^2 \leq \varepsilon_2 k^2 \|u\|^2_\omega + \frac{b}{|\mathrm{Im}(T_{0,k})|^2}|h|^2 + \frac{1}{\varepsilon_2 |\mathrm{Im}(T_{0,k})|^2}\|f\|^2_\omega. \tag{3.22}$$

Thanks to (2.34c), we can rewrite the inequality (3.22) as

$$k^2 b|u(b)|^2 \leq \varepsilon_2 k^2 \|u\|^2_\omega + b|h|^2 + \frac{1}{\varepsilon_2}\|f\|^2_\omega. \tag{3.23}$$

We now apply the Cauchy–Schwarz inequality to (3.17b) (with $d = 2$ and $m = 0$) and use (3.23) to bound the term involving $|u(b)|^2$ to get

$$\begin{aligned} &\|\partial_r u\|^2_\omega + \beta_0 \|u\|^2_{\omega^{-1}} - kb\mathrm{Re}(T_{0,k})|u(b)|^2 \\ &\qquad \leq k^2 \|u\|^2_\omega + k^2 b|u(b)|^2 + \frac{b}{4k^2}|h|^2 + \frac{\varepsilon_1 k^2}{2}\|u\|^2_\omega + \frac{1}{2\varepsilon_1 k^2}\|f\|^2_\omega \\ &\qquad \leq (1+\varepsilon_1) k^2 \|u\|^2_\omega + cb|h|^2 + \frac{1}{\varepsilon_1}\|f\|^2_\omega, \end{aligned} \tag{3.24}$$

where we took $\varepsilon_2 = \varepsilon_1/2$ in (3.23). In view of (3.20) and (3.24), we have the following estimate which is valid for all cases:

$$\|\partial_r u\|^2_{\omega^{d-1}} + \beta_m \|u\|^2_{\omega^{d-3}} \le (1+\varepsilon_1)k^2\|u\|^2_{\omega^{d-1}} + cb^d|h|^2 + c\|f\|^2_{\omega^{d-1}}. \tag{3.25}$$

Now the main difficulty is how to bound the term $k^2\|u\|^2_{\omega^{d-1}}$. To do this, we need to derive further estimates by testing (3.13) with another function. Using a standard regularity argument, one can easily verify that for $f \in L^2_{\omega^{d-1}}(I)$ the weak solution of (3.13) satisfies $(r-a)\partial_r u \in X$. Hence, we can take $v = 2(r-a)\partial_r u$ in (3.13), and after integration by parts and thanks to the identity

$$(u,v)_\omega + (v,u)_\omega = 2\mathrm{Re}(u,v)_\omega, \tag{3.26}$$

we find that the first three terms of the real part of (3.13) with $v = 2(r-a)\partial_r u$ are

$$\begin{aligned} 2\mathrm{Re}\big(\partial_r u, \partial_r((r-a)\partial_r u)\big)_{\omega^{d-1}} &= b^{d-1}|I||\partial_r u(b)|^2 \\ &\quad + \int_a^b \Big[(2-d) + (d-1)\frac{a}{r}\Big]|\partial_r u|^2 r^{d-1}dr; \end{aligned} \tag{3.27a}$$

$$\begin{aligned} 2\beta_m \mathrm{Re}\big(u, (r-a)\partial_r u\big)_{\omega^{d-3}} &= \beta_m b^{d-3}|I||u(b)|^2 \\ &\quad - \beta_m \int_a^b \Big[(d-2) - (d-3)\frac{a}{r}\Big]|u|^2 r^{d-3}dr; \end{aligned} \tag{3.27b}$$

$$\begin{aligned} -2k^2\mathrm{Re}(u, (r-a)\partial_r u)_{\omega^{d-1}} &= -k^2 b^{d-1}|I||u(b)|^2 \\ &\quad + k^2 \int_a^b \Big[d - (d-1)\frac{a}{r}\Big]|u|^2 r^{d-1}dr. \end{aligned} \tag{3.27c}$$

Accordingly, we find that the real part of (3.13) with $v = 2(r-a)\partial_r u$ becomes

$$\begin{aligned} &b^{d-1}|I|\Big(|\partial_r u(b)|^2 + \beta_m b^{-2}|u(b)|^2\Big) + a(d-1)\|\partial_r u\|^2_{\omega^{d-2}} \\ &\qquad + k^2\int_a^b \Big[d - (d-1)\frac{a}{r}\Big]|u|^2 r^{d-1}dr \\ &\le k^2 b^{d-1}|I||u(b)|^2 + (d-2)\|\partial_r u\|^2_{\omega^{d-1}} + \beta_m \int_a^b \Big[(d-2) + (3-d)\frac{a}{r}\Big]|u|^2 r^{d-3}dr \\ &\qquad + 2b^{d-1}|I|\big|\mathrm{Re}(h\overline{\partial_r u(b)})\big| + 2\big|\mathrm{Re}(f, (r-a)\partial_r u)_{\omega^{d-1}}\big|. \end{aligned} \tag{3.28}$$

Note that in the third term the factor $d - (d-1)\frac{a}{r} > 1$ for all $r \in (a,b)$, so we can use this term to bound $k^2\|u\|^2_{\omega^{d-1}}$ in (3.25).

By the Cauchy–Schwarz inequality, we can treat the last two terms at the right-hand side of (3.28) as, respectively,

$$2b^{d-1}|I|\big|\mathrm{Re}\big(h\overline{\partial_r u(b)}\big)\big| \le \frac{b^{d-1}|I|}{2}|\partial_r u(b)|^2 + 2b^{d-1}|I||h|^2, \tag{3.29}$$

and

$$2\big|\mathrm{Re}(f, (r-a)\partial_r u)_{\omega^{d-1}}\big| \le \varepsilon_3\|\partial_r u\|^2_{\omega^{d-1}} + \frac{|I|^2}{\varepsilon_3}\|f\|^2_{\omega^{d-1}}. \tag{3.30}$$

We now proceed separately for $d = 2$ and $d = 3$.

Case I: $d = 2$. In this case, a combination of (3.28)–(3.30) leads to

$$\begin{aligned} b|I|\Big(|\partial_r u(b)|^2 + \beta_m b^{-2}|u(b)|^2\Big) + a\|\partial_r u\|^2 + k^2\int_a^b \Big[2 - \frac{a}{r}\Big]|u|^2 r dr \\ \le k^2 b|I||u(b)|^2 + \Big(\varepsilon_3\|\partial_r u\|_\omega^2 + a\beta_m\|u\|_{\omega^{-2}}^2\Big) + \frac{b|I|}{2}|\partial_r u(b)|^2 \\ + 2b|I||h|^2 + \frac{|I|^2}{\varepsilon_3}\|f\|_\omega^2. \end{aligned} \tag{3.31}$$

Using (3.25) with $d = 2$, we have that, for $\varepsilon_3 < 1$ and for certain $\xi_1 \in (a, b)$,

$$\begin{aligned} \varepsilon_3\|\partial_r u\|_\omega^2 + a\beta_m\|u\|_{\omega^{-2}}^2 &\le \max\Big\{\varepsilon_3, \frac{a}{\xi_1}\Big\}\Big(\|\partial_r u\|_\omega^2 + \beta_m\|u\|_{\omega^{-1}}^2\Big) \\ &\le (1+\varepsilon_1)k^2\|u\|_\omega^2 + cb^2|h|^2 + c\|f\|_\omega^2. \end{aligned} \tag{3.32}$$

Hence, it remains to bound the term $k^2 b|I||u(b)|^2$ in (3.31).

(i) $|m| > kb$. In this case, the term $b|I|k^2|u(b)|^2$ can be absorbed by $b^{-1}|I|\beta_m|u(b)|^2$ at the left-hand side of (3.31). Hence, a combination of (3.31)–(3.32) leads to the desired result:

$$\begin{aligned} b|I|\Big(\frac{1}{2}|\partial_r u(b)|^2 + (\beta_m b^{-2} - k^2)|u(b)|^2\Big) + a\|\partial_r u\|^2 + Ck^2\|u\|_\omega^2 \\ \lesssim b^2|h|^2 + (1+|I|^2)\|f\|_\omega^2, \end{aligned} \tag{3.33}$$

where, with a suitable choice of ε_1, the constant

$$C = 1 - \frac{a}{\xi_2} - \varepsilon_1 > 0 \quad \text{for certain } \xi_2 \in (a, b). \tag{3.34}$$

(ii) $|m| \le kb$. Similar to the derivation of (3.22), we apply the Cauchy–Schwarz inequality to (3.17a):

$$k^2 b|I||u(b)|^2 \le \varepsilon_3|I|k^2\|u\|_\omega^2 + \frac{b|I|}{|\mathrm{Im}(T_{m,k})|^2}|h|^2 + \frac{|I|}{\varepsilon_3|\mathrm{Im}(T_{m,k})|^2}\|f\|_\omega^2. \tag{3.35}$$

Then a combination of the estimates (3.31), (3.32), and (3.35) leads to

$$\begin{aligned} b|I|\Big(\frac{1}{2}|\partial_r u(b)|^2 + \beta_m b^{-2}|u(b)|^2\Big) + a\|\partial_r u\|^2 + \widetilde{C}k^2\|u\|_\omega^2 \\ \lesssim C_{m,k}^{(1)}|h|^2 + C_{m,k}^{(2)}\|f\|_\omega^2, \end{aligned} \tag{3.36}$$

where, with a suitable choice of ε_1 and ε_3 and using the fact that $\mathrm{Im}(T_{m,k}) < 1$, the constants are

$$\begin{aligned} &\widetilde{C} = 1 - \frac{a}{\xi_3} - \varepsilon_1 - \varepsilon_3|I| > 0 \quad \text{for certain } \xi_3 \in (a, b), \\ &C_{m,k}^{(1)} = b^2 + \frac{b|I|}{|\mathrm{Im}(T_{m,k})|^2}, \\ &C_{m,k}^{(2)} = |I|^2 + \frac{|I|}{\varepsilon_3|\mathrm{Im}(T_{m,k})|^2} \lesssim |I|^2\Big(1 + \frac{1}{|\mathrm{Im}(T_{m,k})|^2}\Big). \end{aligned} \tag{3.37}$$

Notice that we have

$$\mathrm{Im}(T_{m,k}) \ge c(kb)^{-\frac{1}{3}} \quad \text{for } |m| \le kb, \tag{3.38}$$

since $\mathrm{Im}(T_{m,k})$ is a decreasing function of m and the estimate (2.35).

Therefore, the desired result (3.15) with $d = 2$ follows from (3.25), (3.33), and (3.36).

Case II: $d = 3$. In this case, a combination of (3.28)–(3.30) leads to

$$\begin{aligned} b^2|I|\big(|\partial_r u(b)|^2 + \beta_m b^{-2}|u(b)|^2\big) + 2a\|\partial_r u\|_\omega^2 + k^2\int_a^b \Big[3 - \frac{2a}{r}\Big]|u|^2 r dr \\ \le k^2 b^2|I||u(b)|^2 + \frac{b^2|I|}{2}|\partial_r u(b)|^2 + \Big(\|\partial_r u\|_{\omega^2}^2 + \beta_m\|u\|^2\Big) \\ + 2b^2|I||h|^2 + \frac{|I|^2}{\varepsilon_3}\|f\|_{\omega^2}^2. \end{aligned} \tag{3.39}$$

By (3.25),

$$\|\partial_r u\|_{\omega^2}^2 + \beta_m\|u\|^2 \le (1+\varepsilon_1)k^2\|u\|_{\omega^2}^2 + cb^3|h|^2 + c\|f\|_{\omega^2}^2. \tag{3.40}$$

The rest of the proof is essentially the same as that in the 2-D case. More precisely, we can derive the 3-D version of inequalities (3.33)–(3.38) with slightly different constants

$$\begin{aligned} C &= 2 - \frac{2a}{\xi_3} - \varepsilon_1 > 0, \quad \xi_3 \in (a,b), \quad \text{if } \ m \ge kb, \\ \tilde{C} &= 2 - \frac{2a}{\xi_3} - \varepsilon_1 - 2\varepsilon_3|I|, \quad \xi_3 \in (a,b), \quad \text{if } \ m < kb. \end{aligned} \tag{3.41}$$

Finally, since $\mathrm{Im}(T_{m,k})$ is a decreasing function of m and $\mathrm{Im}(T_{m,k}) = \mathrm{Im}(\mathcal{T}_{m,kb})$ (cf. (2.14) and (3.7)), the desired bound follows from (2.27) and (2.35). ∎

The proof of Theorem 3.1. Since the proof of the 2-D and 3-D cases is essentially the same, we prove only (3.12) with $d = 3$. Thanks to the orthogonality of the spherical harmonic functions, we deduce from Lemma 3.1 that

$$\begin{aligned} \|\nabla U\|^2 + k^2\|U\|^2 &\lesssim \|\partial_r U\|_{\omega^2}^2 + \|\nabla_S U\|^2 + k^2\|U\|_{\omega^2}^2 \\ &\lesssim \sum_{m=0}^{\infty}\sum_{l=-m}^{m}\Big(\|\partial_r \hat{u}_{lm}\|_{\omega^2}^2 + \beta_m\|\hat{u}_{lm}\|^2 + k^2\|\hat{u}_{lm}\|_{\omega^2}^2\Big) \\ &\lesssim \sum_{m=0}^{\infty}\sum_{l=-m}^{m}\Big(\big(\sqrt{b^3} + \sqrt{b|I|}C_{m,k}\big)^2|\hat{h}_{lm}|^2 + C_{m,k}^2|I|^2\|\hat{f}_{lm}\|_{\omega^2}^2\Big) \\ &\lesssim \sum_{m=0}^{\infty}\sum_{l=-m}^{m}\Big(\big(\sqrt{b^3} + \sqrt{b|I|}(kb)^{1/3}\big)^2\hat{h}_{lm}^2 + (kb)^{2/3}|I|^2\|\hat{f}_{lm}\|_{\omega^2}^2\Big) \\ &\lesssim \big(\sqrt{b^3} + \sqrt{b|I|}(kb)^{1/3}\big)^2\|\eta\|_{L^2(S)} + (kb)^{2/3}|I|^2\|F\|^2. \end{aligned}$$

This ends the proof.

4. Spectral-Galerkin approximation.

4.1. The spectral-Galerkin method and its well-posedness. Let P_N be the space of all complex polynomials of degree at most N on $\bar{I}$. Define $X_N := \big\{u \in P_N \ : \ u(a) = 0\big\}$ and

$$Y_M := \begin{cases} \mathrm{span}\big\{e^{\mathrm{i}m\theta} \ : \ -M \le m \le M\big\} & \text{if } \ d = 2, \\ \mathrm{span}\big\{Y_m^l(\theta,\phi) \ : \ 0 \le |l| \le m \le M\big\} & \text{if } \ d = 3, \end{cases} \tag{4.1}$$

where $\mathcal{B}(\cdot,\cdot)$ is defined in (3.9).

The spectral-Galerkin approximation to (3.11) is as follows:

$$\begin{aligned}&\text{Find } U_{MN} \in \mathcal{V}_{MN} := X_N \times Y_M \text{ such that}\\ &\mathcal{B}(U_{MN}, V_{MN}) = (F, V_{MN})_{\omega^{d-1}} + b^{d-1}\big\langle \eta, V_{MN}(b,\cdot)\big\rangle_S \quad \forall\, V_{MN} \in \mathcal{V}_{MN}.\end{aligned} \tag{4.2}$$

Since the sesquilinear form $\mathcal{B}(\cdot,\cdot)$ is not coercive in $\mathcal{V}_{MN} \times \mathcal{V}_{MN}$ even for small wave number k, an important issue is to prove the well-posedness of the discrete scheme (4.4).

Expanding the numerical solution and test function as

$$\big(U_{MN},\ V_{MN}\big) = \begin{cases} \sum_{|m|=0}^{M} \big(\hat{u}_m^N(r),\ \hat{v}_m^N(r)\big)e^{\mathrm{i}m\theta} & \text{if } d=2,\\ \sum_{m=0}^{M}\sum_{l=-m}^{m} \big(\hat{u}_{lm}^N(r),\ \hat{v}_{lm}^N(r)\big)Y_m^l(\theta,\phi) & \text{if } d=3,\end{cases} \tag{4.3}$$

one verifies that $u_N := \hat{u}_m^N$ or $\hat{u}_{lm}^N$ satisfies the reduced problem

$$\begin{cases} \text{Find } u_N \in X_N \text{ such that}\\ \mathcal{B}_m(u_N, v_N) = (f, v_N)_{\omega^{d-1}} + b^{d-1}h\overline{v_N(b)} \quad \forall v_N \in X_N,\ \ d=2,3,\end{cases} \tag{4.4}$$

where $\mathcal{B}_m(\cdot,\cdot)$ is defined in (3.14); for brevity, we denote $v_N := \hat{v}_m^N$ or v_{lm}^N, and f and h are the same as those in (3.13).

It is important to note that, unlike in the Galerkin finite-element method, the spectral-Galerkin approximation space X_N has the following property: *For $u_N \in X_N$, we have $(r-a)\partial_r u_N \in X_N$. Hence, the proof of Lemma* 3.1 *is also valid for the discrete system* (4.4). In particular, Theorem 3.1 holds with u_N in the place of u. As a consequence, the problem (4.4) has at most one solution. Since (4.4) is finite-dimensional, we then derive from a simple fact in linear algebra that the problem (4.4) admits a unique solution.

Therefore, following the same procedure as in the proof of Theorem 3.1 leads to the following result.

THEOREM 4.1. *If $F \in L^2(\Omega)$ and $\eta \in L^2(S)$, the problem* (4.2) *admits a unique solution satisfying*

$$\|\nabla U_{MN}\| + k\|U_{MN}\| \lesssim \Big(\sqrt{b^d} + \sqrt{b|I|}(kb)^{1/3}\Big)\|\eta\|_{L^2(S)} + (kb)^{1/3}|I|\|F\|. \tag{4.5}$$

4.2. Error estimates. In this part, we shall estimate the error between U (solution of (3.11)) and U_{MN} (solution of (4.2)). Our starting point is to analyze the error of 1-dimensional approximation (4.4).

4.2.1. Analysis of the 1-D scheme. In order to carry out the error analysis, we define the orthogonal projection ${}_0\pi_N^1 : X \to X_N$ by

$$\big(\partial_r(u - {}_0\pi_N^1 u),\ \partial_r v_N\big) = 0 \quad \forall v_N \in X_N. \tag{4.6}$$

For $s \geq 1$ and $s \in \mathbb{N}$, we introduce the weighted Sobolev space

$$B^s(I) := \big\{u \in L^2(I)\ :\ [(r-a)(b-r)]^{\frac{l-1}{2}}\partial_r^l u \in L^2(I),\ 1 \leq l \leq s\big\},$$

with the norm and seminorm

$$\begin{aligned}\|u\|_{B^s} &= \Big(\|u\|^2 + \sum_{l=1}^{s}\big\|[(r-a)(b-r)]^{\frac{l-1}{2}}\partial_r^l u\big\|^2\Big)^{\frac{1}{2}},\\ |u|_{B^s} &= \big\|[(r-a)(b-r)]^{\frac{s-1}{2}}\partial_r^s u\big\|.\end{aligned}$$

LEMMA 4.1. *For any $u \in X \cap B^s(I)$, with $s \geq 1$ and $s \in \mathbb{N}$,*

$$\|\partial_r({}_0\pi_N^1 u - u)\| + N|I|^{-1}\|{}_0\pi_N^1 u - u\| \lesssim N^{1-s}|u|_{B^s}. \tag{4.7}$$

Proof. This result is a direct consequence of the Legendre polynomial approximation (with a scaling and a direct extension to complex functions), which can be found, for instance, in [7], with an improvement of the weighted seminorm in the upper bound given by [14]. □

With the aid of Lemmas 3.1 and 4.1, we are able to obtain the following error estimates.

THEOREM 4.2. *Let u and u_N be, respectively, the solutions of* (3.13) *and* (4.4). *If $u \in X \cap B^s(I)$, with integer $s \geq 1$, then for $d = 2, 3$*

$$\begin{aligned}\|\partial_r(u - u_N)\|_{\omega^{d-1}} + \sqrt{\beta_m}\|u - u_N\|_{\omega^{d-3}} + k\|u - u_N\|_{\omega^{d-1}} \\ \lesssim C_\star(m, N, k; a, b, d)N^{1-s}|u|_{B^s},\end{aligned} \tag{4.8}$$

where

$$\begin{aligned}C_\star(m, N, k; a, b, d) :&= (1 + \sqrt{\beta_m})b^{(d-1)/2} + \sqrt{\beta_m}a^{\frac{d-3}{2}}|I|N^{-1} \\ &+ k^{1/3}(\sqrt{\beta_m}b^{3d/2-2}\sqrt{|I|}N^{-1/2} + |I|^2 b^{d/2}k^2 N^{-1}).\end{aligned} \tag{4.9}$$

Proof. Let $e_N = u_N - {}_0\pi_N^1 u$ and $\tilde{e}_N = u - {}_0\pi_N^1 u$. By (3.13) and (4.4),

$$\mathcal{B}_m(u - u_N, v_N) = 0 \quad \forall v_N \in X_N. \tag{4.10}$$

Then we derive from (3.14), (4.6), and (4.10) that for any $v_N \in X_N$

$$\begin{aligned}\mathcal{B}_m(e_N, v_N) =& \mathcal{B}_m(\tilde{e}_N, v_N) = \beta_m(\tilde{e}_N, v_N)_{\omega^{d-3}} \\ &- k^2(\tilde{e}_N, v_N)_{\omega^{d-1}} - kb^{d-1}T_{m,k}\tilde{e}_N(b)\overline{v_N(b)}.\end{aligned} \tag{4.11}$$

Hence, we can view (4.11) in the form of (3.13) with $u = e_N$, $h = -kb^{d-1}T_{m,k}\tilde{e}_N(b)$, $f = -k^2\tilde{e}_N$, and an additional term $\beta_m(\tilde{e}_N, v_N)_{\omega^{d-3}}$. As with the proof of Theorem 3.1, we take two different test functions $v_N = e_N, 2(r-a)\partial_r e_N \in X_N$ and treat the extra term as

$$\beta_m|(\tilde{e}_N, e_N)_{\omega^{d-3}}| \leq \varepsilon_6\beta_m\|e_N\|^2_{\omega^{d-3}} + \frac{\beta_m}{4\varepsilon_6}\|\tilde{e}_N\|^2_{\omega^{d-3}} \tag{4.12}$$

and

$$\begin{aligned}&2\beta_m|\big(\tilde{e}_N, (r-a)\partial_r e_N\big)_{\omega^{d-3}}| \leq 2\beta_m\Big\{b^{d-3}|I|\big|\tilde{e}_N(b)\overline{e_N(b)}\big| \\ &\quad + \big|\big(\partial_r\tilde{e}_N, (1 - ar^{-1})e_N\big)_{\omega^{d-2}}\big| + \big|\big(\tilde{e}_N, ((d-2) - a(d-3)r^{-1})e_N\big)_{\omega^{d-3}}\big|\Big\} \\ &\leq \varepsilon_7\beta_m b^{d-3}|I||e_N(b)|^2 + \frac{\beta_m b^{d-3}|I|}{\varepsilon_7}|\tilde{e}_N(b)|^2 + \varepsilon_8\beta_m\|e_N\|^2_{\omega^{d-3}} \\ &\quad + \frac{c\beta_m}{\varepsilon_8}\Big(\|\partial_r\tilde{e}_N\|^2_{\omega^{d-1}} + \|\tilde{e}_N\|^2_{\omega^{d-3}}\Big).\end{aligned} \tag{4.13}$$

Thus, choosing suitable constants $\{\varepsilon_j\}_{j=6}^8$ and following the same lines as for the proof of Theorem 3.1 (with $u = e_N$, $h = -kb^{d-1}T_{m,k}\tilde{e}_N(b)$, and $f = -k^2\tilde{e}_N$), we can

derive that

$$
\begin{aligned}
&\|\partial_r e_N\|^2_{\omega^{d-1}} + \beta_m \|e_N\|^2_{\omega^{d-3}} + k^2 \|e_N\|^2_{\omega^{d-1}} \\
&\quad \lesssim \beta_m \big(\|\partial_r \tilde{e}_N\|^2_{\omega^{d-1}} + \|\tilde{e}_N\|^2_{\omega^{d-3}}\big) + \beta_m b^{d-3} |I| |\tilde{e}_N(b)|^2 \\
&\qquad + k^2 b^{2(d-1)} |T_{m,k}|^2 \big(\sqrt{b^d} + \sqrt{b|I|} C_{m,k}\big)^2 |\tilde{e}_N(b)|^2 \\
&\qquad + k^4 |I|^2 C^2_{m,k} \|\tilde{e}_N\|^2_{\omega^{d-1}}.
\end{aligned} \tag{4.14}
$$

To estimate the term $|\tilde{e}_N(b)|$, we use the Sobolev inequality and Lemma 4.1 to obtain that

$$
|\tilde{e}_N(b)|^2 \lesssim \big(2 + |I|^{-1}\big) \|\tilde{e}_N\| \|\tilde{e}_N\|_1 \lesssim N^{1-2s} |I| |u|^2_{B^s}. \tag{4.15}
$$

Next, using the inequality $\|v\|^2_{\omega^\alpha} \le \max\{b^\alpha, a^\alpha\} \|v\|^2$ and Lemma 4.1 leads to

$$
\begin{aligned}
&\|\partial_r^\mu \tilde{e}_N\|^2_{\omega^{d-1}} \le b^{d-1} \|\partial_r^\mu \tilde{e}_N\|^2 \lesssim b^{d-1} |I|^{2-2\mu} N^{2\mu-2s} |u|^2_{B^s}, \quad \mu = 0, 1, \\
&\|\tilde{e}_N\|^2_{\omega^{d-3}} \le a^{d-3} \|\tilde{e}_N\|^2 \lesssim a^{d-3} |I|^2 N^{-2s} |u|^2_{B^s}.
\end{aligned} \tag{4.16}
$$

Hence, by the triangle inequality, (4.14)–(4.16), and Lemma 4.1, we have that

$$
\begin{aligned}
&\|\partial_r (u - u_N)\|^2_{\omega^{d-1}} + \beta_m \|u - u_N\|^2_{\omega^{d-3}} + k^2 \|u - u_N\|^2_{\omega^{d-1}} \\
&\quad \le \Big(\|\partial_r e_N\|^2_{\omega^{d-1}} + \beta_m \|e_N\|^2_{\omega^{d-3}} + k^2 \|e_N\|^2_{\omega^{d-1}}\Big) \\
&\qquad + \Big(\|\partial_r \tilde{e}_N\|^2_{\omega^{d-1}} + \beta_m \|\tilde{e}_N\|^2_{\omega^{d-3}} + k^2 \|\tilde{e}_N\|^2_{\omega^{d-1}}\Big) \\
&\quad \lesssim (1 + \beta_m) \|\partial_r \tilde{e}_N\|^2_{\omega^{d-1}} + \beta_m \|\tilde{e}_N\|^2_{\omega^{d-3}} + \beta_m b^{d-3} |I| |\tilde{e}_N(b)|^2 \\
&\qquad + k^2 b^{2(d-1)} |T_{m,k}|^2 \big(\sqrt{b^d} + \sqrt{b|I|} C_{m,k}\big)^2 |\tilde{e}_N(b)|^2 \\
&\qquad + k^4 |I|^2 C^2_{m,k} \|\tilde{e}_N\|^2_{\omega^{d-1}} \\
&\quad \lesssim C^*(m, N, k; a, b, d) N^{2-2s} |u|^2_{B^s},
\end{aligned} \tag{4.17}
$$

where

$$
\begin{aligned}
C^*(m, N, k; a, b, d) &:= (1 + \beta_m) b^{d-1} + \beta_m a^{d-3} |I|^2 N^{-2} \\
&\quad + \beta_m b^{d-3} |I|^2 N^{-1} + k^2 b^{2(d-1)} |T_{m,k}|^2 \big(\sqrt{b^d} + \sqrt{b|I|} C_{m,k}\big)^2 |I| N^{-1} \\
&\quad + k^4 |I|^4 b^{d-1} C^2_{m,k} N^{-2}.
\end{aligned} \tag{4.18}
$$

We now derive an upper bound for $C^*(m, N, k; a, b, d)$. Since by (2.22) and (2.34)

$$
|T_{m,k}|^2 \le 1 + \frac{(m+1)^2}{(kb)^2}
$$

and by (3.16) $C_{m,k} \lesssim (kb)^{1/3}$, we deduce that for $N \gg 1$

$$
\begin{aligned}
C^*(m, N, k; a, b, d) &\lesssim (1 + \beta_m) b^{d-1} + \beta_m a^{d-3} |I|^2 N^{-2} \\
&\quad + \beta_m b^{3d-4} |I| k^{2/3} N^{-1} + |I|^4 b^d k^{4+2/3} N^{-2}.
\end{aligned}
$$

This implies the desired result. ∎

Remark 4.1. Note that, in the error estimate (4.8), $N^{1-s}|u|_{B^s}$ is the best approximation error, and $k^2 N^{-1}$ in C_* is the so-called "pollution error" which is typical for the numerical approximations to the Helmholtz equation (cf. [4]). The extra term

$k^{1/3}$ in C_* is due to the asymptotic behavior of the DtN kernel (see section 2), and it is unlikely that this extra term can be removed.

Remark 4.2. To illustrate how the error behaves with respect to N, k, and b with $a > 0$ being fixed, we consider a typical oscillatory function $u(r) = e^{ikr} - e^{ika}$. Then, for any $s > 0$, we have

$$\begin{aligned} |u|^2_{B^s} &= \int_a^b |\partial_r^s u|^2 \big((r-a)(b-r)\big)^{s-1} dr \\ &\le k^{2s} \int_a^b \big((r-a)(b-r)\big)^{s-1} dr \lesssim k\Big(k\frac{b-a}{2}\Big)^{2s-1}. \end{aligned} \tag{4.19}$$

Plugging this into (4.8), we find that for this particular but typical solution we have that for any $s \ge 1$

$$\begin{aligned} &\|\partial_r(u-u_N)\|_{\omega^{d-1}} + \sqrt{\beta_m}\|u-u_N\|_{\omega^{d-3}} + k\|u-u_N\|_{\omega^{d-1}} \\ &\qquad \lesssim C_\star(m,N,k;a,b,d)k\sqrt{\frac{b-a}{2}}\Big(\frac{k(b-a)}{2N}\Big)^{1-s}. \end{aligned} \tag{4.20}$$

Hence, the error will decay exponentially as soon as $\frac{k(b-a)}{2N} < 1$, as opposed to the usual condition $\frac{kb}{2N} < 1$. Hence, we can significantly reduce the computational cost by choosing b as close to a as we wish (note, however, that, for scattering from a general obstacle $D = r > a + g(\theta)$ in 2-D or $D = r > a + g(\theta,\phi)$ in 3-D, we have to make sure that $b > a + \|g\|_{L^\infty}$).

With the above preparations, we are ready to perform the error analysis of the full scheme (4.2).

4.2.2. Multidimensional cases. To describe the error, we introduce the following nonisotropic Sobolev space:

$$\mathcal{H}^{s,s'}_{p,\omega^{d-1}}(\Omega) = L^2_p\big(S; B^s(I)\big) \cap H^{s'-1}_p\big(S; H^1_{\omega^{d-1}}(I)\big) \cap H^{s'}_p\big(S; L^2_{\omega^{d-3}}(I) \cap L^2_{\omega^{d-1}}(I)\big), \tag{4.21}$$

with $d = 2, 3$, $s, s' \ge 1$, and the norm

$$\begin{aligned} \big\|U\big\|_{\mathcal{H}^{s,s'}_{p,\omega}(\Omega)} =& \Big(\sum_{|m|=0}^{\infty} \Big[|\hat{u}_m|^2_{B^s} + (1+m^2)^{s'-1}\|\partial_r\hat{u}_m\|^2_\omega \\ &+ (1+m^2)^{s'}\big(\|\hat{u}_m\|^2_{\omega^{-1}} + \|\hat{u}_m\|^2_\omega\big)\Big]\Big)^{\frac{1}{2}}; \\ \big\|U\big\|_{\mathcal{H}^{s,s'}_{p,\omega^2}(\Omega)} =& \Big(\sum_{m=0}^{\infty}\sum_{l=-m}^{m} \Big[|\hat{u}_{lm}|^2_{B^s} + (1+m)^{2s'-s}\|\partial_r\hat{u}_{lm}\|^2_{\omega^2} \\ &+ (1+m)^{2s'}\big(\|\hat{u}_{lm}\|^2 + \|\hat{u}_{lm}\|^2_{\omega^2}\big)\Big]\Big)^{\frac{1}{2}}. \end{aligned} \tag{4.22}$$

THEOREM 4.3. *Let U and U_{MN} be the solutions of* (3.11) *and* (4.2)*, respectively. If $U \in L^2_p(S;X) \cap \mathcal{H}^{s,s'}_{p,\omega^{d-1}}(\Omega)$, with $d = 2,3$ and $s, s' \ge 1$, then we have*

$$\begin{aligned} &\|\nabla(U-U_{MN})\| + k\|U-U_{MN}\| \\ &\qquad \lesssim \Big(C_*(M,N,k;a,b,d)N^{1-s} + (1+kM^{-1})M^{1-s'}\Big)\big\|U\big\|_{\mathcal{H}^{s,s'}_{p,\omega^{d-1}}(\Omega)}, \end{aligned} \tag{4.23}$$

where

$$
\begin{aligned}
C_\star(M,N,k;a,b,d) :&= (1+M)b^{(d-1)/2} + Ma^{\frac{d-3}{2}}|I|N^{-1} \\
&\quad + k^{1/3}(Mb^{3d/2-2}\sqrt{|I|}N^{-1/2} + |I|^2 b^{d/2}k^2N^{-1}).
\end{aligned} \tag{4.24}
$$

Proof. Since the proof of $d = 2, 3$ is quite similar, we shall prove only the case $d = 2$. For notational convenience, let $E_{MN} = U - U_{MN}$ and $\hat{e}_m = \hat{u}_m - \hat{u}_m^N$. Thanks to the orthogonality of the Fourier series, we have that

$$
\begin{aligned}
\|\nabla E_{MN}\|^2 + k^2\|E_{MN}\|^2 &\lesssim \sum_{|m|=0}^{M}\Big(\|\partial_r\hat{e}_m\|_\omega^2 + m^2\|\hat{e}_m\|_{\omega^{-1}}^2 \\
&\quad + k^2\|\hat{e}_m\|_\omega^2\Big) + \sum_{|m|>M}\Big(\|\partial_r\hat{u}_m\|_\omega^2 + m^2\|\hat{u}_m\|_{\omega^{-1}}^2 + k^2\|\hat{u}_m\|_\omega^2\Big) := S_1^2 + S_2^2.
\end{aligned} \tag{4.25}
$$

Using Theorem 4.2 leads to

$$
\begin{aligned}
S_1 &\lesssim \Big(\max_{0\le|m|\le M}\{C_\star(m,\dots)\}\Big)N^{1-s}\Big(\sum_{|m|=0}^{M}|\hat{u}_m|_{B^s}^2\Big)^{\frac{1}{2}} \\
&\lesssim C_\star(M,N,k;a,b,d)N^{1-s}\big\|U\big\|_{\mathcal{H}^{s,s'}_{p,\omega}(\Omega)}.
\end{aligned} \tag{4.26}
$$

We treat S_2 as

$$
\begin{aligned}
S_2 &\lesssim M^{1-s'}\Big(\sum_{|m|>M} m^{2s'-2}\big(\|\partial_r\hat{u}_m\|_\omega^2 + m^2\|\hat{u}_m\|_{\omega^{-1}}^2\big)\Big)^{\frac{1}{2}} \\
&\quad + kM^{-s'}\Big(\sum_{|m|>M} m^{2s'}\|\hat{u}_m\|_\omega^2\Big)^{\frac{1}{2}} \\
&\lesssim \big(1 + kM^{-1}\big)M^{1-s'}\big\|U\big\|_{\mathcal{H}^{s,s'}_{p,\omega}(\Omega)}.
\end{aligned} \tag{4.27}
$$

Hence, a combination of (4.25)–(4.27) yields the desired result. ∎

5. Numerical results and discussions. We now present some numerical results to complement our error estimates for the spectral-Galerkin scheme (4.2). We consider the problem (3.1) in 2-D and take

$$
F(r,\theta) = 0, \quad \eta(\theta) = 0, \quad \xi(\theta) = H_m^{(1)}(ka)e^{im\theta}. \tag{5.1}
$$

In this case the exact solution is $U(r,\theta) = H_m^{(1)}(kr)e^{im\theta}$. Since for a given m, $e^{im\theta}$ can be exactly determined with the number of mode $M = N_\theta \ge 2m$, we will concentrate on the approximation behavior of our scheme with respect to the frequency k and the thickness of the annulus $b - a$.

In the first set of tests, we take $a = 1$ and $b = 2$. In Figure 5.1, we present the relative L^2-error versus the number of mode $N = N_r$ for a wide range of wave numbers. We note that, as soon as $N_r > k(b-a)/2$, the errors start to decay, for moderate to large wave numbers, the errors decay slowly until about $N_r \sim k(b-a)$, and finally, for $N_r > k(b-a)$, all errors converge to zero at an exponential rate.

In the second set of tests, we take $a = 1$ and $b = 1.25$. The results are plotted in Figure 5.2. We observe similar behaviors as in the first set except that now we have

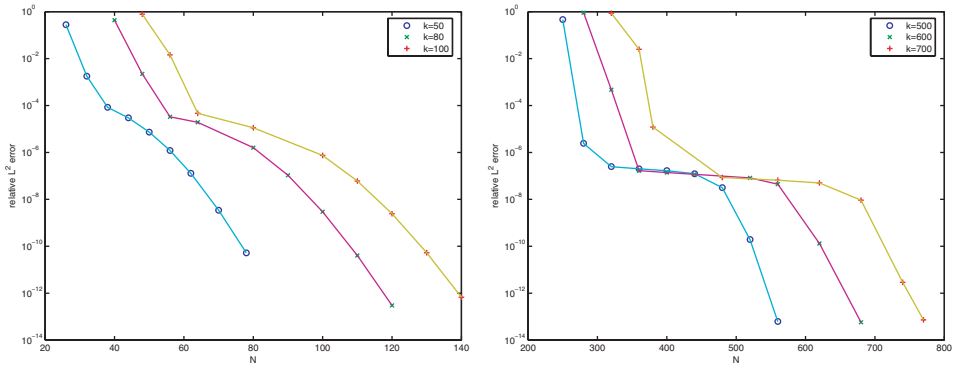


FIG. 5.1. *Relative L^2-error versus N_r as compared to an exact solution: $a = 1, b = 2$.*

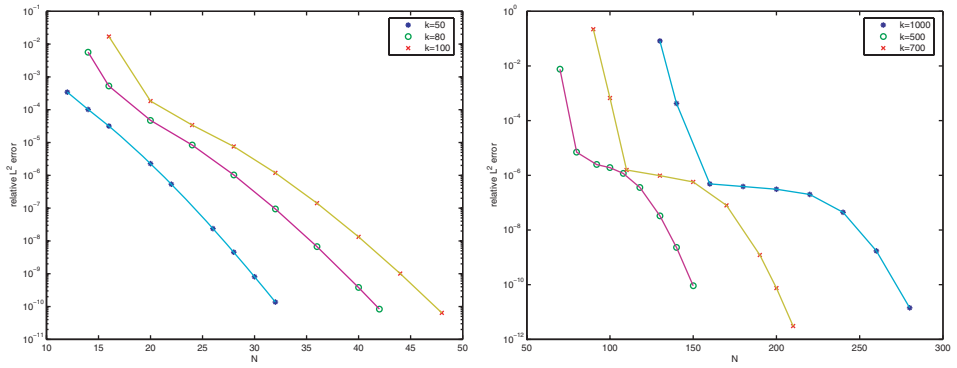


FIG. 5.2. *Relative L^2-error versus N_r as compared to an exact solution: $a = 1, b = 1.25$.*

$b - a = \frac{1}{4}$ and only about 1/4 of the modes are needed to achieve a similar accuracy. These behaviors are consistent with our error estimates (cf. Remark 4.2).

These results indicate that (i) the approximate solution U_{N_r,N_θ} will converge to the exact solution $U(r,\theta)$ exponentially fast as $N_r, N_\theta \to +\infty$ provided that all $F(r,\theta)$, $\xi(\theta)$, and $\eta(\theta)$ are analytic in Ω, and (ii) our numerical scheme is stable for large N_r and capable of providing accurate results for moderate to large wave numbers.

To summarize, we have presented a complete analysis for the spectral-Galerkin method to the Helmholtz equation in exterior domains. We first studied asymptotic behaviors of the Hankel functions which play essential roles for our error analysis. Using these asymptotic estimates, we then derived a priori estimates with explicit dependence on the wave number for both the continuous and the discrete problems. Finally, we performed an error analysis and derived error bounds with explicit dependence on the wave number. To the authors' best knowledge, our error estimates seem to be the first of their kind, i.e., with explicit dependence on the wave number for a numerical method on bounded obstacle scattering via the DtN map. A particular advantage of this approach, verified by our error estimates and numerical results, is that we can choose the artificial boundary very close to the scatterer while still maintaining the spectral accuracy.

Appendix A. The proof of (2.34). We first prove (2.34a). It is clear that, by (2.33b), $\mathrm{Im}(\mathcal{T}_{m,\kappa}) > 0$. On the other hand, since $\mathrm{Im}(\mathcal{T}_{m,\kappa})$ is a strictly increasing

(resp., decreasing) function of κ (resp., m), we have

$$\mathrm{Im}(\mathcal{T}_{m,\kappa}) \le \mathrm{Im}(\mathcal{T}_{1,\kappa}) < \mathrm{Im}(\mathcal{T}_{1,\infty}) = 1,$$

due to the asymptotic formula $|H_1^{(1)}(\kappa)|^2 \sim \frac{2}{\pi\kappa}$ for $\kappa \gg 1$ (see Formula 9.2.3 of [1]).

We now turn to the proof of (2.34b). Recall that the modified Bessel function of the second kind of order ν is defined by

$$\mathcal{K}_\nu(z) = \int_0^\infty e^{-z\cosh t}\cosh(\nu t)dt. \tag{A.1}$$

In particular, we have

$$\mathcal{K}_0(z) = \int_0^\infty e^{-z\cosh t}dt, \qquad \mathcal{K}_1(z) = -\mathcal{K}_0'(z). \tag{A.2}$$

By Formula (4) on p. 445 of [31],

$$J_m^2(\kappa) + Y_m^2(\kappa) = \frac{8}{\pi^2}\int_0^\infty \mathcal{K}_0(2\kappa\sinh t)\cosh(2mt)dt, \tag{A.3a}$$

$$\big[J_mJ_{m+1} + Y_mY_{m+1}\big](\kappa) = \frac{8}{\pi^2}\int_0^\infty \mathcal{K}_1(2\kappa\sinh t)\sinh\big((2m+1)t\big)dt. \tag{A.3b}$$

Using the identity $\mathcal{K}_1(z) = -\mathcal{K}_0'(z)$ and integration by parts leads to

$$\begin{aligned}\big[J_mJ_{m+1} + Y_mY_{m+1}\big](\kappa) &= -\frac{8}{\pi^2}\int_0^\infty \frac{\sinh\big((2m+1)t\big)}{2\kappa(\sinh t)'}d\Big(\mathcal{K}_0(2\kappa\sinh t)\Big)\\ &= -\frac{4}{\kappa\pi^2}\frac{\sinh\big((2m+1)t\big)}{\cosh t}\mathcal{K}_0(2\kappa\sinh t)\Big|_0^\infty\\ &\quad + \frac{4}{\kappa\pi^2}\int_0^\infty \mathcal{K}_0(2\kappa\sinh t)\Big(\frac{\sinh\big((2m+1)t\big)}{\cosh t}\Big)'dt\\ &= \frac{4}{\kappa\pi^2}\int_0^\infty \mathcal{K}_0(2\kappa\sinh t)\cosh(2mt)W_m(t)dt,\end{aligned}$$

where

$$W_m(t) = \frac{1}{\cosh(2mt)}\Big(\frac{\sinh\big((2m+1)t\big)}{\cosh t}\Big)'.$$

Note that in the last step we used the asymptotic formula (see Formula 9.7. 2 of [1])

$$\mathcal{K}_0(2\kappa\sinh t) \sim \sqrt{\frac{\pi}{2\kappa\sinh t}}e^{-2\kappa\sinh t} \sim e^{-\kappa e^t - t/2}, \qquad t \gg 1, \tag{A.4}$$

to claim that

$$\frac{\sinh\big((2m+1)t\big)}{\cosh t}\mathcal{K}_0(2\kappa\sinh t) \to 0 \quad \text{as } t \to \infty.$$

Using the identities of the hyperbolic functions, $W_m(t)$ can be written as

$$W_m(t) = 2m\big(1 + (\tanh t)\tanh(2mt)\big) + \mathrm{sech}^2 t, \quad 0 \le t < \infty. \tag{A.5}$$

We now seek the maximum and minimum values of $W_m(t)$. Taking the derivative of $W_m(t)$ yields

$$\begin{aligned} W_m'(t) &= 2m\big((\mathrm{sech}^2 t)\tanh(2mt) + 2m(\tanh t)\mathrm{sech}^2(2mt)\big) - 2(\tanh t)\mathrm{sech}^2 t \\ &= 2\big(m\tanh(2mt) - \tanh t\big)\mathrm{sech}^2 t + 4m^2(\tanh t)\mathrm{sech}^2(2mt). \end{aligned}$$

It is obvious that

$$m\tanh(2mt) - \tanh t > 0 \quad \forall t > 0,\ \forall m \geq 1.$$

Hence, $W_m(t)$ is an increasing function of t, and consequently,

$$2m + 1 = W_m(0) \leq W_m(t) \leq W_m(\infty) = 4m \quad \forall t \geq 0,\ \forall m \geq 1. \tag{A.6}$$

Therefore, for $m \geq 1$,

$$\frac{2m+1}{2\kappa} \leq \frac{J_m(\kappa)J_{m+1}(\kappa) + Y_m(\kappa)Y_{m+1}(\kappa)}{J_m^2(\kappa) + Y_m^2(\kappa)} \leq \frac{2m}{\kappa}, \tag{A.7}$$

which, together with (2.33a), yields the bounds

$$-\frac{m}{\kappa} \leq \mathrm{Re}(\mathcal{T}_{m,\kappa}) \leq -\frac{1}{2\kappa} \quad \text{for} \quad m \geq 1.$$

Finally, for $m = 0$, (A.5) implies that $0 < W_0(t) \leq 1$. Accordingly, we find that

$$-\frac{1}{2\kappa} \leq \mathrm{Re}(\mathcal{T}_{0,\kappa}) < 0.$$

It remains to prove (2.34c).

Let us first show that $\mathrm{Im}(\mathcal{T}_{0,\kappa})$ is a strictly decreasing function of κ. By (2.33b), it suffices to show that

$$f(x) := x|H_0^{(1)}(x)|^2 = x\big(J_0^2(x) + Y_0^2(x)\big), \quad x > 0,$$

is a strictly increasing function of x. Indeed, by (A.3a),

$$f(x) = \frac{8}{\pi^2}\int_0^\infty x\mathcal{K}_0(2x\sinh t)dt.$$

Differentiating it gives

$$f'(x) = \frac{8}{\pi^2}\int_0^\infty \big\{\mathcal{K}_0(2x\sinh t) + 2x\sinh t\ \mathcal{K}_0'(2x\sinh t)\big\}dt.$$

Integrating the second term by parts leads to

$$f'(x) = \frac{8}{\pi^2}\mathcal{K}_0(2x\sinh t)\tanh t\big|_0^\infty + \frac{8}{\pi^2}\int_0^\infty \mathcal{K}_0(2x\sinh t)\tanh^2 t dt.$$

Notice that the first term is zero due to the decay property of $\mathcal{K}_0$ (cf. (A.4)). Therefore, we have

$$f'(x) > 0 \quad \forall x > 0.$$

Finally, (2.34c) follows immediately from (2.33b) and the facts that $\mathrm{Im}(\mathcal{T}_{0,\kappa})$ is a strictly decreasing function of κ and $\kappa|H_0^{(1)}(\kappa)|^2 \to \frac{2}{\pi}$ as $\kappa \to \infty$.

This ends the proof of (2.34). ∎

REFERENCES

[1] M. Abramowitz and I. A. Stegun, eds., *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, New York, 1984. Reprint of the 1972 edition, Selected Government Publications.

[2] M. Ainsworth, *Discrete dispersion relation for hp-version finite element approximation at high wave number*, SIAM J. Numer. Anal., 42 (2004), pp. 553–575.

[3] M. Ainsworth, *Dispersive and dissipative behaviour of high order discontinuous Galerkin finite element methods*, J. Comput. Phys., 198 (2004), pp. 106–130.

[4] I. M. Babuška and S. A. Sauter, *Is the pollution effect of the FEM avoidable for the Helmholtz equation considering high wave numbers?*, SIAM J. Numer. Anal., 34 (1997), pp. 2392–2423.

[5] O. P. Bruno and F. Reitich, *Numerical solution of diffraction problems: A method of variation of boundaries*, J. Opt. Soc. Amer. A, 10 (1993), pp. 1168–1175.

[6] O. P. Bruno and F. Reitich, *Numerical solution of diffraction problems: A method of variation of boundaries*. II. *Finitely conducting gratings, Padé approximants, and singularities*, J. Opt. Soc. Amer. A, 10 (1993), pp. 2307–2316.

[7] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang, *Spectral Methods in Fluid Dynamics*, Springer-Verlag, Berlin, 1987.

[8] S. N. Chandler-Wilde and S. Langdon, *A Galerkin boundary element method for high frequency scattering by convex polygons*, SIAM J. Numer. Anal., 45 (2007), pp. 610–640.

[9] S. N. Chandler-Wilde and P. Monk, *Existence, uniqueness, and variational methods for scattering by unbounded rough surfaces*, SIAM J. Math. Anal., 37 (2005), pp. 598–618.

[10] P. Cummings and X. Feng, *Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations*, Math. Models Methods Appl. Sci., 16 (2006), pp. 139–160.

[11] L. Demkowicz and F. Ihlenburg, *Analysis of a coupled finite-infinite element method for exterior Helmholtz problems*, Numer. Math., 88 (2001), pp. 43–73.

[12] J. Douglas, J. E. Santos, D. Sheen, and L. S. Bennethum, *Frequency domain treatment of one-dimensional scalar waves*, Math. Models Methods Appl. Sci., 3 (1993), pp. 171–194.

[13] M. J. Grote and J. B. Keller, *On nonreflecting boundary conditions*, J. Comput. Phys., 122 (1995), pp. 231–243.

[14] B. Guo and L. Wang, *Jacobi approximations in non-uniformly Jacobi-weighted Sobolev spaces*, J. Approx. Theory, 128 (2004), pp. 1–41.

[15] I. Harari and T. J. R. Hughes, *Analysis of continuous formulations underlying the computation of time-harmonic acoustics in exterior domains*, Comput. Methods Appl. Mech. Engrg., 97 (1992), pp. 103–124.

[16] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Appl. Math. Sci. 132, Springer-Verlag, New York, 1998.

[17] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number, part* I: *The h-version of FEM*, Comput. Math. Appl., 30 (1995), pp. 9–37.

[18] F. Ihlenburg and I. Babuška, *Finite element solution of the Helmholtz equation with high wave number, part* II: *The h-p version of the FEM*, SIAM J. Numer. Anal., 34 (1997), pp. 315–358.

[19] S. Langdon and S. N. Chandler-Wilde, *A wavenumber independent boundary element method for an acoustic scattering problem*, SIAM J. Numer. Anal., 43 (2006), pp. 2450–2477.

[20] J.-L. Lions and E. Magenes, *Non-homogeneous Boundary Value Problems and Applications*, Vol. I, Springer-Verlag, New York, 1972 (translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181).

[21] J. M. Melenk, *On Generalized Finite Element Methods*, Ph.D. thesis, The University of Maryland, College Park, MD, 1995.

[22] P. Monk, *Finite Element Methods for Maxwell's Equations*, Numer. Math. Sci. Comput., Oxford University Press, New York, 2003.

[23] J.-C. Nédélec, *Acoustic and Electromagnetic Equations*, Appl. Math. Sci. vol. 144, Springer-Verlag, New York, 2001.

[24] D. P. Nicholls and J. Shen, *A stable, high–order method for two–dimensional bounded–obstacle scattering*, SIAM J. Sci. Comput., 28 (2006), pp. 1398–1419.

[25] D. P. Nicholls and F. Reitich, *Stability of high-order perturbative methods for the computation of Dirichlet-Neumann operators*, J. Comput. Phys., 170 (2001), pp. 276–298.

[26] D. P. Nicholls and F. Reitich, *Shape deformations in rough surface scattering: Improved algorithms*, J. Opt. Soc. Amer. A, 21 (2004), pp. 606–621.

[27] L. Rayleigh, *On the dynamical theory of gratings*, Proc. R. Soc. Lond. Ser. A, 79 (1907), pp. 399–416.
[28] J. Shen, *Efficient spectral-Galerkin method* I. *Direct solvers of second- and fourth-order equations using Legendre polynomials*, SIAM J. Sci. Comput., 15 (1994), pp. 1489–1505.
[29] J. Shen, *Efficient spectral-Galerkin methods* III: *Polar and cylindrical geometries*, SIAM J. Sci. Comput., 18 (1997), pp. 1583–1604.
[30] J. Shen and L.-L. Wang, *Spectral approximation of the Helmholtz equation with high wave numbers*, SIAM J. Numer. Anal., 43 (2005), pp. 623–644.
[31] G. N. Watson, *A Treatise on the Theory of Bessel Functions*, 2nd ed., Cambridge University Press, Cambridge, UK, 1966.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 1979–2011

FAST SEMI-LAGRANGIAN SCHEMES FOR THE EIKONAL EQUATION AND APPLICATIONS*

EMILIANO CRISTIANI† AND MAURIZIO FALCONE‡

Abstract. We introduce and analyze a fast version of the semi-Lagrangian algorithm for front propagation originally proposed in [M. Falcone, "The minimum time problem and its applications to front propagation," in *Motion by Mean Curvature and Related Topics*, A. Visintin and G. Buttazzo, eds., de Gruyter, Berlin, 1994, pp. 70–88]. The new algorithm is obtained using the local definition of the approximate solution typical of semi-Lagrangian schemes and redefining the set of "neighboring nodes" necessary for fast marching schemes. A new proof of convergence is needed since that definition produces a new *narrow band* centered at the interphase which is larger than the one used in fast marching methods based on finite differences. We show that the new algorithm converges to the viscosity solution of the problem and that its complexity is $O(N \log N_{nb})$, as it is for the fast marching method based on finite difference (N and N_{nb} being, respectively, the total number of nodes and the number of nodes in the *narrow band*). A new sufficient condition for the convergence of the standard finite difference fast marching method is also given. We present several tests comparing the two algorithms and other fast methods (e.g., fast sweeping) on a series of benchmarks which include the minimum time problem and the shape-from-shading problem.

Key words. front propagation, eikonal equation, semi-Lagrangian schemes, finite differences, fast marching methods

AMS subject classifications. Primary, 65N12; Secondary, 49L25, 49L20

DOI. 10.1137/050637625

1. Introduction.

The level set method is a clever and rather simple way to describe an interface separating two or more regions with different physical phases. As is well known, the method describes the evolution of the front by a continuous representation function $u(x,t)$ which is negative in the domain Ω_t corresponding to one of the phases, positive outside that domain, and changes sign across the interfaces. A comprehensive introduction to the level set method as well as to several applications and references can be found in [19] and [32].

The level set method leads to a nonlinear first order PDE whenever the interface evolution is simply driven by a normal velocity and (possibly) a given advection term. More complicated types of evolution consider the normal velocity as a function of the curvature and/or of other geometric parameters of the interface, and this leads to second order nonlinear PDEs (or integrodifferential equations).

The typical model problem for an interface which evolves in the normal direction driven by a given scalar velocity $c(x) : \mathbb{R}^n \to \mathbb{R}$ leads to the first order Hamilton–Jacobi equation

$$\begin{cases} u_t(x,t) + c(x)|\nabla u(x,t)| = 0 & \text{for } x \in \mathbb{R}^n\,, \quad t \in (0,+\infty), \\ u(x,0) = u_0(x) & \text{for } x \in \mathbb{R}^n, \end{cases} \tag{1}$$

*Received by the editors August 5, 2005; accepted for publication (in revised form) January 18, 2007; published electronically September 7, 2007. This research was partially supported by the MIUR Project 2003 "Modellistica Numerica per il Calcolo Scientifico ed Applicazioni Avanzate."
http://www.siam.org/journals/sinum/45-5/63762.html
†Dipartimento di "Metodi e Modelli Matematici per le Scienze Applicate," SAPIENZA - Università di Roma, Via A. Scarpa, 16 - 00161, Roma, Italy (cristiani@dmmm.uniroma1.it).
‡Dipartimento di Matematica, SAPIENZA - Università di Roma, P.le Aldo Moro, 5 - 00185, Roma, Italy (falcone@mat.uniroma1.it).

where the initial condition u_0 must be a representation function of the initial position of the front $\Gamma_0 = \partial\Omega_0$. Note that u_0 is unknown since Γ_0 is the only initial datum, so that the first step is usually to compute u_0. The above problem can be simplified when the evolution is monotone (increasing or decreasing), i.e., when either $\Omega_t \subset \Omega_{t+s}$ or the reverse inclusion are satisfied for any $t, s > 0$. For monotone types of evolution, it has been proved in [17] (see also [26]) that (1) can be replaced by the stationary equation

$$\begin{cases} c(x)|\nabla T(x)| = 1 & \text{for } x \in \mathbb{R}^n \backslash \Omega_0, \\ T(x) = 0 & \text{for } x \in \partial\Omega_0, \end{cases} \tag{2}$$

where we assume

$$c > 0 \tag{3}$$

and T represents the time needed to transfer a point $x \in \mathbb{R}^n \backslash \Omega_0$ to Ω_0 by appropriate dynamics (see below). In fact, the link between the two equations is simple: if T is the viscosity solution of (2), then $u(x,t) = T(x) - t$ is the viscosity solution of (1). It is worth noting that the second problem is easier to solve since it does not require the additional computation of u_0, which requires the solution of another Hamilton–Jacobi equation of type (2) to compute the (signed) distance function to Ω_0. Moreover, the knowledge of T gives a description of the interface for every time t using the fact that $\Gamma_t = \partial\Omega_t = \{x \in \mathbb{R}^n : T(x) = t\}$. On the other hand, (1) is preferable to its stationary version whenever it is needed to derive a high order scheme that has the same efficiency as the formally first order one (see, e.g., [1]). However, some results on high order methods for stationary first order Hamilton–Jacobi equations including (2) are available, e.g., in [15].

Note that the above stationary approach relies on the link between the propagations of fronts and the minimum time problem of control theory. In fact, as shown in [17], by the change of variable (Kružkov transform)

$$v(x) = 1 - e^{-T(x)} \tag{4}$$

we can transform (2) into the equation

$$\begin{cases} v(x) + \max\limits_{a \in B(0,1)} \{c(x)a \cdot \nabla v(x)\} = 1 & \text{for } x \in \mathbb{R}^n \backslash \Omega_0, \\ v(x) = 0 & \text{for } x \in \partial\Omega_0, \end{cases} \tag{5}$$

where $B(0,1)$ is the unit ball centered in 0. This is the Hamilton–Jacobi–Bellman equation of a minimum time problem for the dynamics

$$\begin{cases} \dot{y}(t) = -c(y)\alpha(t), & t \in (0, +\infty), \\ y(0) = x, \end{cases} \tag{6}$$

where $\alpha(\cdot) \in \mathcal{A} = \big\{\alpha(\cdot) : [0, +\infty) \to B(0,1) \subset \mathbb{R}^n, \text{ measurable}\big\}$. We will denote by $y(t; \alpha, x)$ the solution of the system corresponding to the control α and to the initial condition x. The usual requirement in order to have existence and uniqueness for the trajectories under the Carathéodory conditions is

$$c(x) \text{ Lipschitz continuous and bounded.} \tag{7}$$

Let us define the cost functional

$$J_x(\alpha(\cdot)) = \inf\left\{t : y(t;\alpha,x) \in \Omega_0\right\} \leq +\infty.$$

It is well known that the minimum time function

$$T(x) = \inf_{\alpha(\cdot)\in\mathcal{A}} J_x(\alpha(\cdot)) \tag{8}$$

is the unique viscosity solution of (2) (see, e.g., [2, 4]).

We will focus our attention on the numerical solution of (2). It should be noted that a fast marching method has a lower cost with respect to the corresponding classical iterative method (or fixed point method), which computes the solution on the whole grid at every iteration. The classical fast marching method based on finite differences (FM-FD) was proposed in [33] as an acceleration method for a monotone first order iterative finite difference scheme (see [7] for a second order version of the scheme and [8] for a general convergence result). Since semi-Lagrangian schemes have shown to be more accurate than the finite difference schemes corresponding to the same order, it is natural to extend the ideas behind the FM-FD method to this class of schemes. In the framework of semi-Lagrangian schemes several convergence results and a priori error estimates have been obtained via control arguments since these schemes correspond to a discrete version of the dynamic programming principle; see [2] and [13]. Moreover, these schemes do not require an explicit and restrictive CFL condition for stability (see [16]). It is interesting to note that the first tentative steps in this direction can be found in [38] using a different approximation scheme. More recently, a semi-Lagrangian scheme has been proposed by Sethian and Vladimirsky in [34] in a more general framework which includes anisotropic front propagation on unstructured grids.

Our main contribution here is to introduce and analyze a new fast marching version of a semi-Lagrangian scheme, for which a priori error estimates are available in [12], and to prove an upper bound on its computational cost. We will also review the basic features of the FM-FD method and give a complete proof of its convergence under an explicit CFL condition which guarantees that the scheme is always meaningful and there are no complex solutions. To our knowledge this condition appears for the first time in the literature; our proof is presented in the appendix. Lastly we will recall the fast sweeping method studied by Zhao [39] (see also [37, 20, 21, 27, 28] for other sweeping methods and extensions) and also provide a sweeping version of our algorithm. Further extensions and new applications of the scheme presented in this paper can be found in [9].

The paper is organized as follows. In section 2 we recall the basic features of the FM-FD method introduced in [33] to solve (2) when $c(x)$ has a constant sign in its domain of definition. An example which shows that the FM-FD scheme can produce complex solutions is given in the same section, and the proof of convergence under a new CFL condition which always guarantees real solutions is presented in the appendix. Section 3 is devoted to the presentation of the fast marching semi-Lagrangian method (FM-SL) for (5). Section 4 contains some properties of the FM-SL scheme that will be useful in establishing its convergence, which will be proved in section 5. In the same section we analyze the computational complexity, showing that the FM-SL scheme has a complexity of order $O(N\ln(N_{nb}))$, where N is the total number of nodes of our computational grid and N_{nb} is the number of nodes in the *narrow band* (bounded by N). In section 6 we also present other fast algorithms and

give the sweeping version of our method. Finally, section 7 is devoted to numerical tests and to comparisons between several FM schemes on a number of benchmarks.

2. The fast marching methods based on finite differences. The fast marching method has been introduced to reduce the computational effort needed to solve (2). The basic level set algorithm is based on a finite difference discretization and on an iterative procedure $T^{n+1} = F(T^n)$ which computes the approximate solution everywhere in $\mathbb{R}^n \setminus \Omega_0$ at every iteration. The FM-FD method instead follows the front concentrating the computational effort where it is needed, i.e., in a small neighborhood of the front, and it updates that neighborhood at every iteration to avoid useless computations. This is done by dividing the grid nodes into three subsets: *far* nodes, *accepted* nodes, and *narrow band* nodes. The narrow band nodes are the nodes where the computation actually takes place and their value can still change at the following iterations. The accepted nodes are those where the solution has been already computed and where the value cannot change in the following iterations. Finally, the far nodes are the remaining nodes where an approximate solution has never been computed. In physical terms, the far nodes are those in the space region which has never been touched by the front, the accepted nodes are those where the front has already passed through, and the narrow band nodes are, iteration by iteration, those lying in a neighborhood of the front.

The algorithm starts labeling as *accepted* only the nodes belonging to the initial front, i.e., belonging to $\Gamma_0 = \partial\Omega_0$, and ends only when all the nodes have been accepted. In this section, we will briefly sketch the FM-FD scheme for (2). In order to avoid cumbersome notation we will restrict the presentation to the case $n = 2$. In what follows, we will always consider the case of a positive normal velocity; i.e., we assume $c(x) > 0$ to guarantee a monotone (increasing) evolution of the front. The results in this section can be easily generalized to the n-dimensional case and to the case $c(x) < 0$.

We will take a square Q large enough to contain Ω_0; this is the domain where we want to compute T. Boundary conditions will be given on ∂Q and Γ_0 but, as a first step, we will consider the algorithm *without* boundary conditions on ∂Q. The implementation of boundary conditions in the scheme will be discussed in section 5.

We will assume that we are working on a structured grid of $M \times N$ nodes (x_i, y_j), $i = 1, \ldots, N$ and $j = 1, \ldots, M$. Δx and Δy will denote the (uniform) discretization steps, respectively, on the x and y axes. We will denote by $T_{i,j}$ and $c_{i,j}$, respectively, the values of T and c at (x_i, y_j).

Let us write (2) as

$$T_x^2 + T_y^2 = \frac{1}{c^2(x,y)}. \tag{9}$$

We replace the partial derivatives T_x and T_y by first order finite differences, and we choose for simplicity $M = N$ and $\Delta x = \Delta y$. It is well known that in order to obtain an approximation of the viscosity solution, an *up-wind* correction must be introduced. This leads to the equation

$$\begin{aligned} &\left(\max\left\{ \max\left\{ \frac{T_{i,j} - T_{i-1,j}}{\Delta x}, 0 \right\}, -\min\left\{ \frac{T_{i+1,j} - T_{i,j}}{\Delta x}, 0 \right\} \right\} \right)^2 \\ &+ \left(\max\left\{ \max\left\{ \frac{T_{i,j} - T_{i,j-1}}{\Delta x}, 0 \right\}, -\min\left\{ \frac{T_{i,j+1} - T_{i,j}}{\Delta x}, 0 \right\} \right\} \right)^2 = \frac{1}{c_{i,j}^2}. \end{aligned} \tag{10}$$

2.1. The FM-FD algorithm. Let us briefly recall the main definitions and steps of the FM-FD method.

DEFINITION 2.1 (neighboring nodes for the finite difference scheme). *Let* $X = (x_i, y_j)$ *be a node. We define the set of neighboring nodes to* X *as*

$$N_{FD}(X) = \Big\{(x_{i+1}, y_j), (x_{i-1}, y_j), (x_i, y_{j+1}), (x_i, y_{j-1})\Big\}.$$

These are the nodes appearing in the stencil of the first order finite difference discretization. The definition can be easily extended to the n-dimensional case.

Sketch of the FM-FD algorithm.

Initialization.

1. The nodes belonging to the initial front Γ_0 are located and labeled as *accepted*. Their value is set to $T = 0$ (they form the set $\widetilde{\Gamma}_0$).
2. The initial narrow band is defined taking the nodes belonging to $N_{FD}(\widetilde{\Gamma}_0)$, external to Γ_0. These nodes are labeled as *narrow band*, setting the value to $T = \frac{\Delta x}{c}$.
3. The remaining nodes are labeled as *far*, and their value is set to $T = +\infty$ (in practice, the maximum floating point number).

Main cycle.

1. Among all the nodes in the narrow band we search for the minimum value of T. Let us denote this node by A.
2. A is labeled as *accepted* and is removed from the *narrow band*.
3. The nodes in $N_{FD}(A)$ which are not accepted are labeled as *active*. If among these nodes there are nodes labeled as *far*, they are transferred to the narrow band.
4. The value of T in the active nodes is computed (or recomputed), solving the second order equation (10) and taking the largest root.
5. If the narrow band is not empty, go back to 1.

Note that the narrow band is a reasonable approximation of the level set of $T(x, y)$.

The main interest in the FM-FD method is that its computational cost is bounded. In fact, every node cannot be accepted more than one time and every node has just four neighbors, so the bound on the maximum number of times a single node can be recomputed is four. This corresponds to a computational cost of $O(N)$, where N is the total number of nodes. We should add to that cost the search for the minimum value of T among the nodes in the narrow band, which costs $O(\ln(N_{nb}))$, where N_{nb} is the number of nodes in the narrow band. In conclusion, the algorithm has a global cost of $O(N \ln(N_{nb}))$ operations (see [38, 32, 33] for further details on the computational cost). This is not the case for the usual iterative/fixed point algorithm since in that case the approximate solution is obtained in the limit and, in practice, no one knows when the stopping criterion will apply; i.e., the maximum number of iterations is virtually unbounded.

Let us observe that it is necessary to introduce some conditions or to modify the scheme in order to avoid inconsistencies due to the appearance of imaginary solutions. In fact, let us consider the discretization (10) and suppose that

$$T_{i,j} < T_{i+1,j}, \qquad T_{i,j} < T_{i,j-1}, \qquad T_{i,j} > T_{i-1,j}, \qquad T_{i,j} > T_{i,j+1}.$$

It is easy to check that (10) corresponds to

$$\left(\frac{T_{i,j} - T_{i-1,j}}{\Delta x}\right)^2 + \left(\frac{T_{i,j+1} - T_{i,j}}{\Delta x}\right)^2 = \frac{1}{c_{i,j}^2},$$

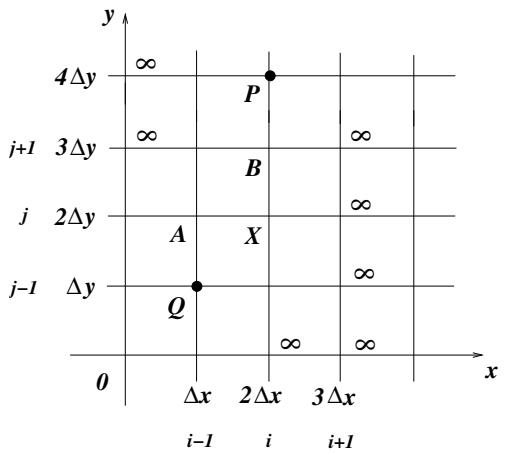


FIG. 1. *A configuration with complex roots.*

which gives

$$T_{i,j} = \frac{T_{i-1,j} + T_{i,j+1} \pm \sqrt{2\left(\frac{\Delta x}{c_{i,j}}\right)^2 - (T_{i-1,j} - T_{i,j+1})^2}}{2}. \tag{11}$$

We already noted that the term under the square root can be negative. Obviously this must be avoided since complex roots have no physical meaning. A situation where this occurs is the following example.

Consider the case where the initial front is the union of two points, i.e., $\Gamma_0 = P \cup Q$, $Q = (\Delta x, \Delta y)$ and $P = (2\Delta x, 4\Delta y)$ (see Figure 1). Let us consider the following velocity:

$$c(x,y) = \begin{cases} \varepsilon, & y \le \Delta y, \\ \varepsilon + \frac{1-\varepsilon}{\Delta y}(y - \Delta y), & \Delta y \le y \le 2\Delta y, \\ 1 - \frac{1-\varepsilon}{\Delta y}(y - 2\Delta y), & 2\Delta y \le y \le 3\Delta y, \\ \varepsilon, & y \ge 3\Delta y. \end{cases} \tag{12}$$

In this case the algorithm initializes the narrow band, computing a large value for B when ε is small and a small value for the node A which will be the first node accepted (after Γ_0). When the node X has to be computed, its value depends on $T(A)$ and $T(B)$. Since $c(X) = 1$ and $T(A) - T(B)$ is large (for ε small) the radicand in (11) will be negative (as numerical tests confirm).

This difficulty can be solved by either choosing the positive part of the radicand (as suggested in [22]) or changing discretization, as in [39]. However, both choices lead to a modification of the scheme, which can be difficult to handle when looking for theoretical results. We prefer to avoid changing the scheme, and we prove that under the CFL-like condition

$$\Delta x \le (\sqrt{2} - 1)\frac{c_{min}}{L_c} \tag{13}$$

the algorithm always computes *real* solutions at every node (here c_{min} is the minimum value of c, L_c is its Lipschitz constant, and again $\Delta x = \Delta y$). Condition (13) has a

clear meaning and allows us to give a proof of convergence to the viscosity solution. To our knowledge this is the first time this condition appears in the literature; a complete proof of the convergence result (Proposition 2.1) will be given in the appendix.

Let us denote by A the node in the narrow band where the minimum value of T is attained. The algorithm labels A as *accepted* and starts to compute the neighboring nodes which are not accepted.

PROPOSITION 2.1. *Let $X = (x_i, y_j) \in N_{FD}(A)$ be the node where the FM-FD method computes a solution. Let us assume that*

$$c_{min} = \min_{Q\setminus\Omega_0} c(x) > 0 \tag{14}$$

and that the following CFL-like condition holds true:

$$\Delta x \leq (\sqrt{2}-1)\frac{c_{min}}{L_c}, \tag{15}$$

where L_c denotes the Lipschitz constant of c. Then we have

$$T(A) \leq T(X) \leq T(A) + f_X, \tag{16}$$

where $f_X := \Delta x/c(X)$.

The above result is crucial in order to obtain convergence in a finite number of steps. In fact, it shows that the minimum value of the nodes in the narrow band (which is actually the only value accepted at every iteration) is exact within the consistency error of the scheme. An approximate value is considered to be exact if the algorithm cannot replace it with a strictly lower value at any of the following iterations.

3. The fast marching method based on the semi-Lagrangian scheme. We will study a fast marching version of the semi-Lagrangian scheme studied in [12] under the assumptions (3) and (7) that we will keep here. It was proved in [3] that the numerical scheme stems from a discrete version of the dynamic programming principle applied to (6); this leads to the equation

$$\begin{cases} w(x) = \min\limits_{a\in B(0,1)} \{\beta w(x - hc(x)a)\} + 1 - \beta & \text{for } x \in \mathbb{R}^n\setminus\Omega_0, \\ w(x) = 0 & \text{for } x \in \partial\Omega_0, \end{cases} \tag{17}$$

where $\beta = e^{-h}$, h is the time step for the (hidden) dynamics, and w is an approximation of v. We will consider for simplicity a structured grid G, denoting its nodes by x_i, $i = 1, \ldots, N$, i.e., $G = \{x_i,\ i = 1, \ldots, N\}$. Note that the same scheme can be implemented on an unstructured grid as in [31]. We write (17) at every node, obtaining

$$\begin{cases} w(x_i) = \min\limits_{a\in B(0,1)} \{\beta w(x_i - hc(x_i)a)\} + 1 - \beta & \text{for } x_i \in G\setminus\Omega_0, \\ w(x_i) = 0 & \text{for } x_i \in G \cap \Omega_0, \end{cases} \tag{18}$$

where we defined $w = 0$ also in the internal nodes of Γ_0. It has been shown in [12] that under our assumptions (3) and (7), equation (18) has a unique solution w in the class of piecewise linear functions ($\mathbb{P}_1$ in the finite element notation) defined on the grid. Let us note that by applying the Kružkov transform (4) to the equation, one can also treat the case when $c = 0$ since in that case the minimum time function to

the target (i.e., the initial configuration of the front in the front propagation problem) will have infinite value at some points, whereas v will always stay bounded by 1. This allows us to run the computations also for $c = 0$ and to treat problems with state constraints (as we will see in the last section).

We will always approximate the v variable and use the fact that the Kružkov transform is monotone. In fact, since $T_1 > T_2$ if and only if $v_1 > v_2$ we can work on the v variable without changing the rules for the update of the narrow band because the crucial point is to label as *accepted* the node in the narrow band, where T (or v) attains its minimum. The above rule guarantees that we will process the nodes in an ordering which corresponds to increasing values of v.

The idea which is behind the FM-SL method is rather simple: we follow the initialization and all the steps of the classical FM-FD method except the step where the value at the node x_i is actually computed. That step would require us to iterate until convergence the scheme

$$w(x_i) = \min_{a \in B(0,1)} \{\beta w(x_i - hc(x_i)a)\} + 1 - \beta, \tag{19}$$

so that the typical fixed point iteration is applied "locally" at every single node following the order indicated by the FM-FD method. We will prove that for a semi-Lagrangian scheme based on a piecewise linear space reconstruction, just a single iteration is needed to compute the exact (within the accuracy of the scheme) value at every node so that the computational effort is very limited and of the same order as the FM-FD method.

3.1. Fast minimum search in $B(0,1)$. We will start improving the minimum search which is typical of the semi-Lagrangian schemes. The search for a minimum in the unit ball $B(0,1)$ will be solved algebraically for a linear interpolation, which allows us to compute the values $w(x_i - hc(x_i)a)$ using the known values at the nodes. Clearly, a new algebraic solution must be obtained (if possible) for other high order interpolations. Let us just recall that for the standard semi-Lagrangian scheme the search for the minimum is usually restricted to a discretization of the unit ball $B(0,1)$ which takes into account r points (or *controls* in the minimum time terminology) $a_1, a_2, \ldots, a_k, \ldots, a_r \in B(0,1)$.

For example, one can construct a uniform grid on $\partial B(0,1)$ with step $\Delta\theta = 2\pi/r$. To find the minimum, for every a_k the value $w(x_i - hc(x_i)a_k)$ is actually computed by interpolation. Although the choice of the type and order of the interpolation is completely free, the most popular choices are *linear*, using the three values at the nodes which are closer to $x_i - hc(x_i)a_k$, and *bilinear*, using the four values of w at the vertices of the cell containing $x_i - hc(x_i)a_k$.

Once all the values for a_k, $k = 1, \ldots, r$, are computed the minimum is obtained by comparison. It is worth noting that this algorithm is quite slow and requires a high computational cost; however, it can be applied to every high order interpolation. Moreover, it should be noted that this minimization problem is quite difficult since we expect to have nondifferentiable or even discontinuous solutions (if state constraints/obstacles are present in the domain) and that the comparison algorithm is very simple to implement and reasonably fast in low dimension especially when the search for the minimum can be restricted to the boundary of $B(0,1)$ (as will be the case in many examples). However, other algorithms for the minimization of nonsmooth functions can be applied, and the interested reader can find in [6] and [14] recent improvements on the solution of this problem. These algorithms converge to

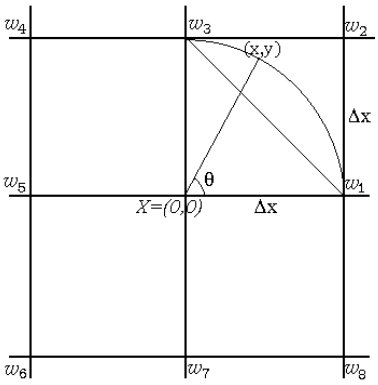


FIG. 2. *Search for optimal control.*

the minimum in the limit, so they cannot be applied here since we want to have an exact evaluation of the computational cost.

It is important to note that the time step h in (19) can vary at every node. We will denote by $h_i = h(x_i)$ the time step corresponding to the node x_i, by $c_i = c(x_i)$ the velocity at x_i, and by $\beta_i = e^{-h_i}$. When $c_i > 0$ it is always possible to choose

$$h_i = \frac{\Delta x}{c_i}. \tag{20}$$

In this way (19) can be written as

$$w(x_i) = \min_{a \in B(0,1)} \{\beta_i w(x_i - \Delta x\ a)\} + 1 - \beta_i. \tag{21}$$

In this situation, the nodes where $c_i = 0$ are actually treated apart from the other nodes: we just assign them the value $w = 1$ (which corresponds to $T = +\infty$) without any additional computation.

The method we propose here for the minimization problem has a low dimensional cost since for linear interpolation the search is restricted to the boundary of the unit ball. This is not a real restriction since, for our applications, the minimum in the unit ball is attained at the boundary. Later in this section we will show how this algorithm can be applied as a building block of our FM-SL scheme.

For simplicity, let us examine the situation in $\mathbb{R}^2$ considering a set of four cells each of side length Δx centered at the origin (see Figure 2). We want to compute the minimum of the function $w((0,0) - \Delta x\ a)$ for $a = (\cos\theta, \sin\theta)$ and $\theta \in [0, 2\pi)$. Let us introduce a vector $\mathbf{m} = (m_1, m_2, \ldots, m_8)$; the values of its components will be defined below. The minimum value for which we search will be given by $p = \min\{m_1, m_2, \ldots, m_8\}$.

Let us define the first four components of $\mathbf{m}$,

$$m_1 = w(\Delta x, 0)\,, \quad m_2 = w(0, \Delta x)\,, \quad m_3 = w(-\Delta x, 0)\,, \quad m_4 = w(0, -\Delta x),$$

and let us search for the minimum in every orthant.

Orthant I. Let w_1, w_2, and w_3 be the values of w corresponding, respectively, to the nodes $(\Delta x, 0)$, $(\Delta x, \Delta x)$, and $(0, \Delta x)$. The unique linear function $f(x, y)$ satisfying the conditions

$$f(\Delta x, 0) = w_1\,, \quad f(\Delta x, \Delta x) = w_2\,, \quad f(0, \Delta x) = w_3$$

is

$$f(x,y) = ax + by + c, \tag{22}$$

where

$$a = \left(\frac{w_2 - w_3}{\Delta x}\right), \qquad b = \left(\frac{w_2 - w_1}{\Delta x}\right), \qquad c = w_1 - w_2 + w_3.$$

Let us define the real function

$$F(\theta) = f(\Delta x \cos\theta, \Delta x \sin\theta) = a\Delta x \cos\theta + b\Delta x \sin\theta + c\,, \quad \theta \in [0, 2\pi) \tag{23}$$

and look for the minimum of $F(\theta)$ in the interval $(0, \pi/2)$. Note that the extreme values $\theta = 0$ and $\theta = \pi/2$ are not included since the values at the extrema of that interval have already been included in $\mathbf{m}$ (they are m_1 and m_2). By differentiating with respect to θ we obtain

$$F'(\theta) = 0 \Leftrightarrow \theta = \arctan(b/a).$$

The interesting case is when $w_2 < w_1$ and $w_2 < w_3$; otherwise the minimum is w_1 or w_3.

In this case, we get

$$a \neq 0\,, \quad b \neq 0\,, \quad b/a > 0\,, \quad \arctan(b/a) \in (0, \pi/2),$$

which means that the relative minimum is at $\theta_1^* = \arctan(b/a)$ and we set $m_5 = F(\theta_1^*)$.

If $w_2 \geq w_1$ or $w_2 \geq w_3$, we set $m_5 = +\infty$ (or the highest machine number).

Orthant II. Let w_3, w_4, and w_5 be the values of w, respectively, at the nodes $(0, \Delta x)$, $(-\Delta x, \Delta x)$, and $(-\Delta x, 0)$. The unique linear function $f(x, y)$ such that

$$f(0, \Delta x) = w_3\,, \quad f(-\Delta x, \Delta x) = w_4\,, \quad f(-\Delta x, 0) = w_5$$

is

$$f(x,y) = ax + by + c,$$

where

$$a = \left(\frac{w_3 - w_4}{\Delta x}\right), \qquad b = \left(\frac{w_4 - w_5}{\Delta x}\right), \qquad c = w_3 - w_4 + w_5.$$

Again we will consider the composite function $F(\theta)$ defined in (23), and we observe that it has a relative minimum in $(\pi/2, \pi)$ if and only if $w_4 < w_3$ and $w_4 < w_5$. In this case we have

$$a \neq 0\,, \quad b \neq 0\,, \quad b/a < 0\,, \quad \arctan(b/a) \in (-\pi/2, 0). \tag{24}$$

Since we are in the second orthant the value of θ where the minimum for F is attained is $\theta_2^* = \arctan(b/a) + \pi$. Proceeding as in the first orthant we set $m_6 = F(\theta_2^*)$.

If $w_4 \geq w_3$ or $w_4 \geq w_5$, we set $m_6 = +\infty$.

The analysis of the third and fourth orthants follows in the same way and will be skipped.

Once all the components of $\mathbf{m}$ have been set, we just compute $p = \min\{m_1, m_2, \ldots, m_8\}$ and substitute it in the expression

$$w(0,0) = \beta p + 1 - \beta. \tag{25}$$

This is done at every fixed point iteration until convergence. It is important to note that the above linear interpolation has a great advantage: the computation of the correct value of $w(0,0)$ does not require more than one iteration given the values at the neighboring nodes (along the axis directions and the diagonals) since $F(\theta)$ will not depend on $w(0,0)$. This property will *not* hold for other high-order interpolations, e.g., quadratic interpolation. Another advantage of linear interpolation with respect to the comparison of the values in a discrete unit ball is that it gives the exact value of the optimal direction at the cost corresponding to a discretization of $B(0,1)$ by just 8 directions.

3.2. The FM-SL scheme. This section is devoted to the presentation of the fast marching version of the SL-algorithm. For simplicity the presentation is given in $\mathbb{R}^2$, but the algorithm can be easily extended to $\mathbb{R}^n$. Let us start introducing the following definitions.

DEFINITION 3.1 (neighboring nodes for the SL scheme). *Let $X = (x_i, y_j)$ be a node of the grid. We define*

$$N_{FD}(X) = \Big\{(x_i, y_{j+1}), (x_i, y_{j-1}), (x_{i-1}, y_j), (x_{i+1}, y_j)\Big\},$$

$$D(X) = \Big\{(x_{i+1}, y_{j+1}), (x_{i+1}, y_{j-1}), (x_{i-1}, y_{j+1}), (x_{i-1}, y_{j-1})\Big\},$$

$$N_{SL}(X) = N_{FD}(X) \cup D(X).$$

The above definition is a natural extension of Definition 2.1 for the semi-Lagrangian scheme. According to the new definition, the nodes in the narrow band will also include the diagonal directions and not only the four directions N, S, E, W, as in the FM-FD method of section 2.

Sketch of the FM-SL algorithm.

Initialization (see Figure 3).

1. The nodes belonging to the initial front Γ_0 are located and labeled as *accepted*. Their value is set to $w = 0$. We will denote this set of nodes by $\widetilde{\Gamma}_0$.
2. The initial narrow band is defined according to the Definition 3.1, taking the nodes belonging to $N_{SL}(\widetilde{\Gamma}_0)$ external to Γ_0. These nodes are labeled as *narrow band*. Their value is set to $w = 1 - e^{-\frac{\Delta x}{c}}$ (which corresponds to $T = \Delta x/c$) if they belong to $N_{FD}(\widetilde{\Gamma}_0)$, or to $w = 1 - e^{-\frac{\sqrt{2}\Delta x}{c}}$ (which corresponds to $T = \sqrt{2}\Delta x/c$) if they belong to $D(\widetilde{\Gamma}_0)$.
3. We label as *far* all the remaining nodes of the grid; their value is set to $w = 1$ (which corresponds to the value $T = +\infty$).

Main cycle.

1. Among all the nodes in the narrow band we search for the minimum value of w. Let us denote this node by A.
2. The node A is labeled as *accepted* and is removed from the narrow band.
3. We label as *active* the nodes in $N_{SL}(A)$ which are not accepted. If there are far nodes, they are moved into the narrow band.

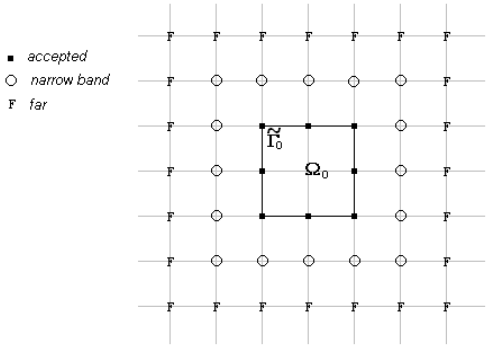


FIG. 3. *Initialization for FM-SL method, case* $c > 0$.

4. We compute (or recompute) the value w at the nodes belonging to $N_{FD}(A)$ which are active, iterating the fixed point operator

$$w(x_i) = \min_{a \in B(0,1)} \{\beta_i w(x_i - h_i c_i a)\} + 1 - \beta_i \,, \tag{26}$$

where $h_i c_i = \Delta x$. Note that just one iteration is needed, as we will see in the following sections. Then we compute by the same formula the value at the remaining active nodes in $N_{SL}(A) \setminus N_{FD}(A)$.
5. If the narrow band is empty, the algorithm stops; otherwise it goes back to step 1.

Although the algorithm advances the narrow band also in the diagonal directions, according to the new definition, it computes first the values at the neighboring nodes in the directions N, S, E, W (i.e., the finite difference directions) and then passes to the diagonal directions.

Some extensions: Obstacles, infinite velocity. We have seen that one can use our algorithm to deal with a front propagation with obstacles, i.e., regions where c vanishes. In [36, 18] the problem has been analyzed and several tests have been presented for a semi-Lagrangian method based on the linear interpolation, which treats the obstacle in a very simple way. The algorithm just assigns to the nodes belonging to an obstacle the value $w = 1$ in order to impose (indirectly and easily) a state constraints boundary conditions. In order to use the fast marching technique we just have to be careful and distinguish between nodes initialized to the value $w = 1$ because they are far and the ones to which was assigned the value $w = 1$ because they belong to an obstacle. In section 7 (Test 5) we will show a front propagating in the presence of obstacles.

Another interesting extension for applications to image processing is when the domain of computation contains points with infinite velocity. This is the case, for example, in the shape-from-shading problem when we have a point of maximal light intensity in the image (see, e.g., [29, 24]). Let us illustrate the idea which is behind our solution. Let x_{i_0} be a node such that

$$\lim_{x \to x_{i_0}} c(x) = +\infty.$$

Our equation $c(x)|\nabla T(x)| = 1$ can be written as

$$|\nabla T(x)| = g(x), \tag{27}$$

where $g(x) = 1/c(x)$. Clearly, (27) is a degenerate eikonal equation since g vanishes at x_{i_0}.

In order to compute $w(x_{i_0})$, we can set, according to (20), $h_{i_0} = 0$ and $\beta_{i_0} = 1$ and proceed as before, setting in (26)

$$h_{i_0} c_{i_0} = \Delta x. \tag{28}$$

Let us extend the function $h(x)$ outside the nodes in the domain $Q\backslash\Omega_0$. Our choice (28) can be justified by the fact that we would expect in our algorithm

$$\lim_{x\to x_{i_0}} c(x) = +\infty\,, \quad \lim_{x\to x_{i_0}} h(x) = 0, \quad \text{and} \quad \lim_{x\to x_{i_0}} c(x)h(x) = \Delta x.$$

Note that even if this argument is heuristic, it assigns to the node x_{i_0} the exact value for w. In fact, by (26), we get

$$w(x_{i_0}) = \min_{a\in B(0,1)} \{1w(x_{i_0} - \Delta x\ a)\} + 1 - 1 = w(x_{i_0} - \Delta x a^*),$$

where a^* is the *optimal control*. Since the front has an infinite velocity at x_{i_0} the minimum time of arrival on it coincides with the minimum time of arrival on the circle of radius Δx centered at x_{i_0}. In section 7 (Tests 6 and 7) we will show an application to a front propagation problem and to the shape-from-shading problem. It is interesting to note that theoretical results on discontinuous Hamiltonians can be found in [35] and [5].

4. Properties of the FM-SL scheme. We start with the following easy result on the semi-Lagrangian discretization.

PROPOSITION 4.1. *Let X be a node and assume that $w(X)$, defined by* (26), *is computed by interpolation using the three values $w^{(1)}$, $w^{(2)}$, $w^{(3)}$. Then*

$$w(X) \geq \min\left\{w^{(1)}, w^{(2)}, w^{(3)}\right\}. \tag{29}$$

Proof. Let $\beta = e^{-h}$, $h > 0$, and a^* be the optimal direction/control at X. The inequality

$$\beta w(X - h_i c_i a^*) + 1 - \beta \geq w(X - h_i c_i a^*)$$

is satisfied if and only if $w(X - h_i c_i a^*) \leq 1$. Since w is always less than or equal to 1 (due to the Kružkov transform) we have proved that

$$w(X) \geq w(X - h_i c_i a^*). \tag{30}$$

Since a simple property of linear interpolation guarantees that

$$\max\left\{w^{(1)}, w^{(2)}, w^{(3)}\right\} \geq w(X - h_i c_i a^*) \geq \min\left\{w^{(1)}, w^{(2)}, w^{(3)}\right\} \tag{31}$$

by (30) and (31) we end the proof. ∎

In order to prove that the fast marching version of our semi-Lagrangian scheme converges to the viscosity solution in a finite number of steps we have to prove first that the fast method for the minimum analyzed in section 3.1 matches the fast marching technique. This is necessary since the narrow band of the FM-SL method is larger than the narrow band of the FM-FD method as a consequence of the new definition of neighboring nodes. In particular we will show that the algorithm automatically

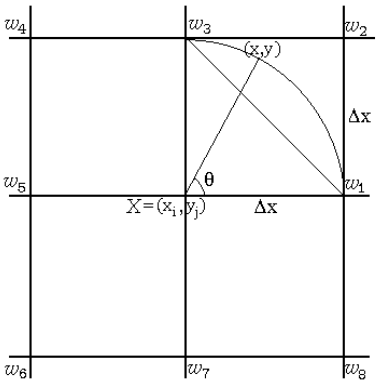


FIG. 4. *Analysis of the minimum in orthant I.*

rejects far nodes from the computation as in the standard up-wind finite difference discretization.

Let X be the node where we want to compute $w(X)$. Without loss of generality, we will assume that the optimal value is attained at a direction $\theta^* \in [0, \pi/2]$, i.e.,

$$a^* = (\cos\theta,\ \sin\theta)\,, \quad \theta \in [0, \pi/2]. \tag{32}$$

We will examine in detail all the possible configurations for this situation, which will be referred to in the following as the "minimum in orthant I" case (see Figure 4). For simplicity, let us assume $c > 0$ so that a node is labeled as *far* if and only if its value is $w = 1$.

PROPOSITION 4.2. *Let X be a node and let $w(X)$ be defined by* (26). *The value $w(X)$ will* not *be computed by interpolation using nodes labeled as* far.

Proof. Let us give the proof for the minimum in orthant I. The analysis for the other orthants is similar and can be easily obtained by symmetry arguments.

1. $w_1 = w_2 = w_3 = 1$: This configuration cannot occur. In fact, since the minimum is attained in orthant I we should have $w_4 = w_5 = w_6 = w_7 = w_8 = 1$. But this is not possible since we compute at X only when at least one of the nodes belonging to $N_{SL}(X)$ has been labeled as *accepted* in one of the previous iterations, and an accepted node must have a value lower than 1.
2. Among w_1, w_2, and w_3 there are two values equal to 1.
   (a) $w_1 = w_3 = 1$: this case cannot occur. In fact, since the minimum is attained in orthant I we must have $w_2 \leq w_1, w_3, w_4, \ldots, w_8$. The node that must be labeled as *accepted* is the one corresponding to the value w_2. This implies that the values w_1 and w_3 must be computed before X (see the sketch of the algorithm).
   (b) $w_1 = w_2 = 1$: the minimum value is w_3. A new iteration to compute $w(X)$ would not give a lower value, so the optimal value is obtained in just one iteration.
   (c) $w_2 = w_3 = 1$: the minimum value is w_1. Again, we will not get a lower value iterating, and the optimal value is obtained in just one iteration.
3. Among w_1, w_2, and w_3 only one value is equal to 1.
   (a) $w_2 = 1$: since f is linear the minimum will be attained by w_1 or w_3. The optimal value is obtained in just one iteration.
   (b) $w_1 = 1$, $w_3 \leq w_2$: the minimum is w_3.

(c) $w_1 = 1$, $w_3 > w_2$: this is the most delicate case since $w_2 < w_1, w_3$. The minimum for F will be attained at some $\theta^* \in (0, \pi/2)$. The value $w(X)$, obtained by linear interpolation, will not be correct since it depends on $w_1 = 1$, which is a conventional value. Moreover, note that a new iteration of the fixed point map at X will not make $w(X)$ decrease since w_1 is frozen and so is $w(X)$. If this case could occur, we would not get convergence to the correct value even in the limit on the number of iterations. Note that this difficulty can occur neither for the global semi-Lagrangian scheme where *all* the nodes are computed at the same iteration nor for the FM-FD method where the values corresponding to far nodes are not used in the stencil. The following argument shows that this case also cannot occur for the FM-SL scheme. Since $w_1 = 1$, the corresponding node is labeled as *far* at the current iteration. This implies that the nodes labeled as *accepted* at the previous iteration do not belong to $N_{SL}(w_1)$. As a consequence, w_2 belongs to the narrow band. By Proposition 4.1 we have $w(X) > w_2$. This implies that X cannot be labeled as *accepted* before the nodes corresponding to w_2. Once w_2 becomes accepted the algorithm computes w_1 and w_3 before computing $w(X)$ so that the values at nodes labeled as *far* will not contribute.

(d) $w_3 = 1$, $w_1 \leq w_2$: the minimum is w_1. The optimal value is obtained in just one iteration.

(e) $w_3 = 1$, $w_1 > w_2$: analogous to case (3c). □

5. Convergence of the FM-SL scheme in a finite number of steps. As for the FM-FD method we have to prove that the minimal value of the nodes of the narrow band cannot decrease if we iterate the fixed point operator; i.e., it coincides with the value obtained by the discrete operator working on all the nodes. As we have seen, the values at the nodes belonging to the narrow band are not accepted all together. Only the minimal value is accepted at every iteration (this is a very pessimistic choice which simplifies the theoretical result). The following proposition shows the bounds on the number of times that one node can be recomputed, and it is a building block for the convergence of the scheme.

PROPOSITION 5.1. *Let X be a node in the narrow band such that $w(X) = w_{old}(X)$. Let us assume that at the current iteration the algorithm needs to compute a new value $w_{new}(X)$ for X. Moreover, let us assume that at the current iteration the following property holds true:*

(33) *If A belongs to the narrow band and B is* accepted, *then $w(A) \geq w(B)$.*

The following properties hold:

1. *If the value $w_{old}(X)$ was computed at an iteration in which a grid point $A_1 \in N_{FD}(X)$ was labeled as* accepted, *then it is impossible that $w_{new}(X) < w_{old}(X)$.*
2. *If the value $w_{old}(X)$ was computed at an iteration in which a grid point $A_2 \in D(X)$ was labeled as* accepted, *then to the node X a new value $w_{new}(X) < w_{old}(X)$ can be assigned but it will always satisfy the inequality $w_{new}(X) \geq w(A_2)$.*

Proof. Let us start with the first statement.

1. Let us assume that when the value w_{old} was assigned to X the node A_1 was the (unique) node belonging to $N_{FD}(X)$, which had been labeled as *accepted*.

When the algorithm computed $w(X) = w_{old}(X)$ we certainly had

$$\min_{a \in \partial B(0,1)} w(X - \Delta x\ a) = w^* \leq w(A_1)$$

since there is a direction/control $\bar{a} \in \{(1,0),(0,1),(-1,0),(0,-1)\}$ such that $w(X - \Delta x\ \bar{a}) = w(A_1)$. The only possibility of having at X a value lower than $w_{old}(X)$ in the following iterations of the algorithm is that a value assigned to a node belonging to $N_{SL}(X)$ was lower than w^*. However, by Proposition 4.1 we know that this value cannot be computed using in the stencil the values at the nodes of the actual narrow band because they are all greater than $w(A_1) \geq w^*$, which has been accepted (as (33) assures). A lower value could be computed only using a stencil which contains nodes already accepted in one of the previous iterations since they all have values lower than $w(A_1)$. This is not possible since all the nodes which are neighbors of those accepted nodes have been computed already and they have a value greater than or equal to $w(A_1)$ since they have not been labeled as *accepted*.

2. Let us assume, for simplicity, that the node A_2 is the unique node belonging to $D(X)$ which has been labeled as *accepted* and let $w_{old}(X)$ be the value assigned at X at the same iteration. When a node $A_1 \in N_{FD}(X)$ has been labeled as *accepted* before A_2, the result holds true by the arguments of the above case 1.
   Let us assume that A_2 is the unique neighbor of X which has been labeled as *accepted*. Then we have

$$\min_{a \in \partial B(0,1)} w(X - \Delta x\ a) = w^* \geq w(A_2).$$

   It is always possible that using $w(A_2)$ one can obtain a new value $w_{new}(X)$ lower than $w_{old}(X)$. However, by (33) and Proposition 4.1 all the new values will be greater than or equal to $w(A_2)$; therefore $w_{new}(X) \geq w(A_2)$. □

Remark 5.1. Note that the previous proposition allows us to accelerate the algorithm. In fact, one can save CPU time by avoiding recomputing the values at the nodes corresponding to case 1. However, they cannot be labeled as *accepted* before their value is the minimum in the narrow band. An important consequence of Proposition 5.1 and the above observation is that every node can be computed at most 5 times; this is one of the reasons why the CPU time for FM-SL is slightly larger than that for the FM-FD method, where a node can be computed at most 4 times. We will see in the last section that the FM-SL method produces a more accurate approximation of the viscosity solution, which justifies a small increment in the CPU time.

The following result is an analogue of Proposition 2.1, and it is crucial to prove convergence in a finite number of steps.

PROPOSITION 5.2. *Let w be defined in* (26) *and let $w(X)$ be the value assigned at X at the same iteration when a node $Z \in N_{SL}(X)$ is labeled as* accepted. *Assume that*

$$c(x) \geq 0 \quad \textit{for any } \ x \in Q \backslash \Omega_0.$$

Then we have

$$w(X) \geq w(Z). \tag{34}$$

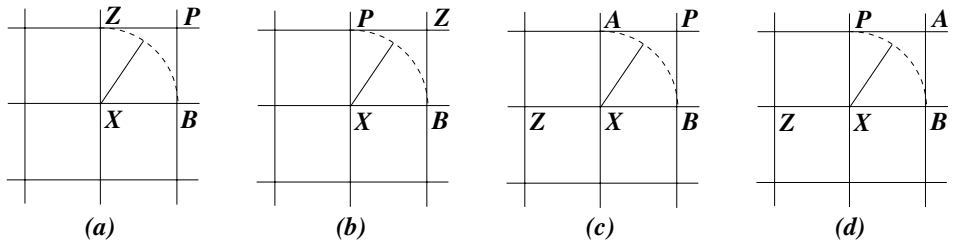


FIG. 5. *Four different configurations for Case* 2.

Proof. We examine all the cases corresponding to a minimum in orthant I (see Figure 4). The proof will be obtained by induction on the number of iterations of the algorithm.

At the first step the result holds true by our initialization.

Let us consider the nth step of the algorithm. The induction hypothesis implies that at the current iteration the values of nodes in the narrow band are greater than values of nodes labeled as accepted. Therefore (33) holds true, so we can apply Proposition 5.1. Our proof will be divided into three parts.

Case 1. $w_1, \ldots, w_8$ are narrow band or far (before Z is labeled as *accepted*).

If Z belongs to orthant I, we have seen by Proposition 4.1 that

$$w(X) \geq \min\{w_1, w_2, w_3\} = w(Z).$$

If Z does not belong to orthant I, we have

$$w(X) \geq \min\{w_1, w_2, w_3\} \geq w(Z)$$

since Z as been labeled as *accepted*.

Case 2. One node $w_1, \ldots, w_8$ is accepted (before Z is labeled as *accepted*).

Let us denote by P this node. When P was accepted the value at X was $w_{old}(X)$. Now the value at X has to be recomputed. We can only have one of the following situations:

1. P belongs to orthant I.
   (a) Z belongs to orthant I
      i. See Figure 5(a). By Proposition 5.1, Z and B cannot be assigned to a lower value after P became accepted, so $w_{new}(X) = w_{old}(X)$ and $w_{old}(X) \geq w(Z)$ since Z is the node chosen to be labeled as *accepted*.
      ii. See Figure 5(b). When Z is accepted the minimum is attained at P, and this implies again $w_{new}(X) = w_{old}(X)$.
   (b) Z does not belong to orthant I
      i. See Figure 5(c). In the iterations between the acceptance of P and that of Z the values $w(A)$ and $w(B)$ cannot be changed. Moreover, the minimum is attained in orthant I so we have $w_{new}(X) = w_{old}(X)$.
      ii. See Figure 5(d). We know that the value $w(A)$ has not been replaced, $w(B)$ cannot be lower than $w(P)$, and the minimum is attained in orthant I. Then the minimum is attained at P and $w_{new}(X) = w_{old}(X)$.

2. P does not belong to orthant I.
   Since the minimum is attained in orthant I this means that P has no effect on the computation at X and we are back to Case 1.

Case 3. More than one value among $w_1, \dots, w_8$ has been labeled as *accepted* (before Z is labeled as *accepted*).

This case can be solved by the arguments in Case 2. □

As for the FM-FD method (see [33]) we can now conclude that the value of the node which is labeled as *accepted* at every iteration cannot be decreased if we iterate the fixed point operator. In fact, let us denote this value w_{min}. Since all the nodes in the narrow band have values greater than w_{min}, the previous result implies that using those nodes we cannot assign to a node a value lower than w_{min}. In conclusion, the up-winding is respected and the value w_{min} can be considered exact since it cannot be improved on the same grid (of course it can be improved if we reduce the discretization steps).

Remark 5.2. The FM-SL scheme does not require a stability CFL-like condition, as required by the FM-FD scheme.

5.1. Convergence to the viscosity solution and conclusions. The semi-Lagrangian scheme is consistent, as has been proved, e.g., in [16]. Moreover, choosing $\Delta x = \Delta y$, we get that the local truncation error is $O(\Delta x)$.

We will prove that the solution computed by the FM-SL method is identical to the solution computed by the standard semi-Lagrangian scheme where the computation is repeated on every node of the grid until convergence. Naturally, if the two schemes compute the same values, convergence of the FM-SL method to the viscosity solution is just a consequence of that of the standard semi-Lagrangian scheme.

THEOREM 1. *Let* $(V_i)_{i=1,\dots,N}$ *be the matrix containing the final values on the* n*-dimensional grid and let*

$$V_i = F(V_{i-k}, \dots, V_{i+l}) \tag{35}$$

be the iteration corresponding to the numerical scheme. Let $\widehat{V}$ *be the matrix of the approximate solution corresponding to the fixed point iteration* (35) *and let* $\overline{V}$ *be the matrix containing the final values of the approximate solution corresponding to the fast marching technique applied to the* same *scheme (i.e., the result obtained when the narrow band is empty). Then* $\overline{V} = \widehat{V}$.

Proof. The two matrices coincide if and only if

$$\overline{V}_i = F(\overline{V}_{i-k}, \dots, \overline{V}_{i+l}) \quad \text{for any} \quad i = 1, \dots, N. \tag{36}$$

Assume the narrow band is empty and take $\overline{V}$ as initial guess for the fixed point technique; this will not change the solution since the value is computed by the same scheme. When all the nodes are accepted the equality (36) must hold for every i. In fact, if the equality is not true at one node, then its value can still be improved, implying that the list of narrow band or far nodes is not empty, which gives us a contradiction. □

The above results allow us to draw some conclusions about the order of complexity of the FM-SL scheme. The values $w(X)$ computed by (26) are an approximation of $v(X)$, which has been computed at most 5 times for every nodes. This means that the computational cost can be estimated as in the FM-FD scheme. One component is given by the cost of the heap-sort method to select the minimum value in the narrow band, and the other component is given by the computational cost at every node.

This globally gives a cost $O(N\log(N_{nb}))$, where N is the total number of nodes and N_{nb} the number of nodes in the narrow band (see [33]).

Since the values which have been labeled as *accepted* at every iteration cannot be improved by the global fixed point iteration, i.e., they coincide with the same values obtained by the global fixed point operator, the a priori error estimates in [12] are still valid for the solution obtained by the FM-SL method. In the last section we will present several tests which confirm these theoretical results.

Boundary conditions on ∂Q. We define outside Q a strip of ghost nodes where we set $w = 1$. If they enter the narrow band, at the end of the iteration, their value is set back to $w = 1$ to avoid their contributing to the computation of other internal nodes. When the minimal value on the nodes of the narrow band is 1, the ghost nodes will be the only nonaccepted nodes and we can stop the computation. In general, any constant larger than the maximum of the solution in Q can be used to assign the value at the ghost nodes (a typical choice is to set the solution to $+\infty$ if there is no a priori estimate on the solution).

Note that in our case, the normal velocity has always the same (positive) sign, so in the case of a constant velocity the front propagation starting from $\Gamma_0 \subset Q$ will hit the boundary of Q and both T and w are increasing approaching ∂Q. The values computed by the algorithm on the nodes of the boundary will always be lower than 1, and the choice of the above boundary condition is then well adapted to this situation. However, when c is variable or when there are obstacles in the domain we can also have a different situation: the front propagates more rapidly in some directions, and this could require enlarging the domain to get a correct solution to our problem. Finally, let us observe that the use of homogeneous Neumann boundary conditions is less appealing because it strongly affects the fronts near the boundary because all the level curves must be orthogonal to the boundary to satisfy $\nabla v(x)\cdot\eta(x) = 0$ for any $x \in \partial Q$ (here $\eta(x)$ denotes the exterior normal to Q).

6. Other fast schemes. As some authors have remarked, it is possible to improve the finite difference method. In the paper by Tsitsiklis [38] one can find an algorithm which can be parallelized directly with a complexity $O(N)$. There are at least two ways to accelerate convergence and/or reduce the CPU time:

1. Reduce the computational effort for the minimum search by accepting more than one node in the narrow band at every iteration (group marching method).
2. Avoid searching for the minimum value in the narrow band (fast sweeping method), obtaining convergence in more than one iteration.

We will briefly illustrate these two techniques.

Group marching. The group marching (GM) method has been introduced by Kim [22] to solve the eikonal equation on a structured grid by a discretization as that of FM-FD. Although we do not compare this algorithm with the others studied in the previous section, we will give a brief presentation of its main features for completeness. Let us denote by Γ the set of nodes belonging to the narrow band, and let us choose $\Delta x = \Delta y$. Define

$$T_{\Gamma,min} = \min\{T_{i,j} \mid (x_i, y_j) \in \Gamma\} \quad \text{and} \quad c_{\Gamma,\max} = \max\{c_{i,j} \mid (x_i, y_j) \in \Gamma\}.$$

The GM method labels as *accepted*, *all at once*, the nodes belonging to the set G defined by

$$G := \left\{(x_i, y_j) \in \Gamma \ : \ T_{i,j} \le T_{\Gamma,min} + \frac{\Delta x}{\sqrt{2}}\ \frac{1}{c_{\Gamma,\max}}\right\}. \tag{37}$$

At every iteration the update of the narrow band is obtained as in the FM-FD method, including the four neighbors of every node that have been labeled as *accepted*. It is clear that if the set G is large, the GM method can be much faster than the FM-FD method because more than one node at a time in accepted. On the other hand, it is rather difficult to give an estimate of the acceleration parameter since the cardinality of G depends on the values $\{T_{i,j} : (x_i, y_j) \in \Gamma\}$ and on the velocity of propagation. It could be that $G = \{T_{\Gamma,min}\}$, and this would imply a computational cost of $O(N \ln(N_{nb}))$ instead of getting $O(N)$, as one would expect by some tests in [22].

Fast sweeping. The fast sweeping (FS) method is based on an idea first introduced in [11] and was extensively analyzed in [39] and [37]. The crucial idea is that the algorithm sweeps the whole (two-dimensional) domain with four alternating orderings repeatedly,

$$(38) \qquad (1)\ i = 1, \dots, N\,,\ j = 1, \dots, M; \qquad (2)\ i = N, \dots, 1\,,\ j = 1, \dots, M;$$

$$(39) \qquad (3)\ i = N, \dots, 1\,,\ j = M, \dots, 1; \qquad (4)\ i = 1, \dots, N\,,\ j = M, \dots, 1$$

(where N and M are the number of nodes in each dimension), and it updates the value at a grid point only if the new value is smaller than the current one. This idea can be easily extended to n-dimensional domains.

Computing the values in this special ordering, the algorithm is able to follow simultaneously a family of characteristics in a certain direction. As proved in [39], the FS method converges in 2^n iterations, where n is the dimension of the problem if the initial front Γ_0 is just a point on the grid and the function c is constant. If those assumptions do not hold, the FS method has been shown to be of complexity $O(N)$ and to converge in a finite number of iterations although the bound for the number of iterations is not explicitly written out. See [27] for an extension on triangular meshes and an upper bound to the number of iterations needed by the FS method to reach convergence.

Let us note that the discretization used in [39] is the same as that used in the FM-FD method described in section 2, and that in any case the numerical evidence shows that the convergence is more rapid with respect to the classical iterative method.

The FS method has an easy extension to the semi-Lagrangian case. In fact, we can easily substitute the finite difference discretization by the semi-Lagrangian discretization maintaining the ordering in which nodes are visited. Obviously, we expect that at least in the case $c(x) \equiv$ const. the FS semi-Lagrangian scheme (FS-SL) can compute in four iterations exactly the same solution as FM-SL.

In the next section we run this algorithm in the case $c(x) \equiv 1$ with two different initial fronts and see that this intuition is actually true.

7. Numerical experiments. In this section we present some numerical experiments performed with MATLAB 7 on a PC equipped with a Pentium IV 2.80 GHz processor, 512 MB RAM.

The main goal is to compare the FM-FD method and the FM-SL method described in previous sections. We also compare these methods with the semi-Lagrangian iterative method and FS method based on a semi-Lagrangian discretization described in section 6. First, two tests are devoted to approximate the solution of model problems where we know the exact solution, so we can compute the L^∞ error and L^1 error. Other tests are devoted to solving more complicated problems and applications

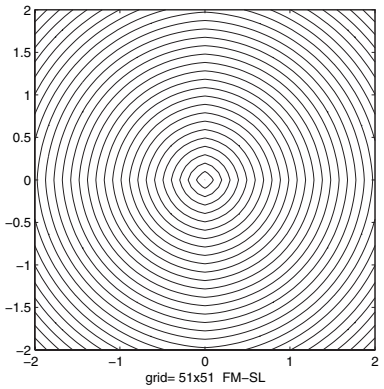


FIG. 6. *Level sets of $T(x)$ computed by the FM-SL method,* 51×51 *grid.*

in which the velocity function $c(x)$ does not satisfy standard assumptions such as Lipschitz continuity and boundedness.

If not specified otherwise, we choose $Q = [-2, 2]^2$ as our computational domain.

7.1. Tests on model problems. In the following tests we compare the exact solution T with the solution $\widehat{T}$ computed by the FM-FD method and the FM-SL method described above. Note that in the implementation of the FM-SL algorithm we have used the observation in Remark 5.1 to speed up the computation.

We compute

$$E_{\infty,\Delta x} = \max_{i,j} |T_{i,j} - \widehat{T}_{i,j}|\,, \qquad E_{1,\Delta x} = (\Delta x)^2 \sum_{i,j} |T_{i,j} - \widehat{T}_{i,j}| \tag{40}$$

and the rate of convergence r in some model problems in $\mathbb{R}^2$. We consider 51×51, 101×101, and 201×201 grids[1] corresponding, respectively, to $\Delta x = 0.08$, $\Delta x = 0.04$, and $\Delta x = 0.02$.

Since we know that there is a constant C such that

$$E_{p,\Delta x} \le C \Delta x^r \quad \text{and} \quad E_{p,\Delta x/2} \le C \left(\frac{\Delta x}{2}\right)^r, \qquad p = 1, \infty,$$

we obtain that the numerical rate of convergence is

$$r = \log_2 \left(\frac{E_{p,\Delta x}}{E_{p,\Delta x/2}}\right), \qquad p = 1, \infty.$$

Moreover, we compare these algorithms with the classical iterative semi-Lagrangian method in which we choose $\max_{i,j} |w_{i,j}^{(k)} - w_{i,j}^{(k-1)}| < \varepsilon$, $\varepsilon = 10^{-7}$, as the stopping criterion and with the FS-SL method performing just four iterations in different order.

Let us finally remark that in all cases condition (15) holds.

Test 1. $\Gamma_0 = (0, 0)$, $c(x, y) \equiv 1$. Exact solution: $T(x, y) = \sqrt{(x^2 + y^2)}$.

Results are summarized in Figure 6 and Table 1. As expected, in all cases errors reduce as Δx decreases. The numerical rate of convergence (Table 2) is in the interval $[0.5, 1]$ for both methods.

[1] In these grids there is a node corresponding to the point $(0, 0)$.

TABLE 1
Errors for Test 1.

Method	Δx	L^∞ error	L^1 error	CPU time (sec)
FM-FD	0.08	0.0875	0.7807	0.5
FM-SL	0.08	0.0329	0.3757	0.7
SL (46 it)	0.08	0.0329	0.3757	8.4
FS-SL	0.08	0.0329	0.3757	0.8
FM-FD	0.04	0.0526	0.4762	2.1
FM-SL	0.04	0.0204	0.2340	3.1
SL (86 it)	0.04	0.0204	0.2340	60
FS-SL	0.04	0.0204	0.2340	3.2
FM-FD	0.02	0.0309	0.2834	9.4
FM-SL	0.02	0.0122	0.1406	14
SL (162 it)	0.02	0.0122	0.1406	443.7
FS-SL	0.02	0.0122	0.1406	12.5

TABLE 2
Rate of convergence in L^∞ *and* L^1 *norms computed by errors in Table* 1.

Method	L^∞ $(0.08 \to 0.04)$	L^∞ $(0.04 \to 0.02)$	L^1 $(0.08 \to 0.04)$	L^1 $(0.04 \to 0.02)$
FM-FD	0.7342	0.7675	0.7132	0.7487
FM-SL	0.6895	0.7417	0.6831	0.7349

The FM-SL and semi-Lagrangian methods give exactly the same errors in accordance with Theorem 1, and they are also equal to the errors of FS-SL, as expected, since FS-SL converges in four iterations in the case c is constant. These errors number about half that of the FM-FD method, although both are first order methods. This is due to the fact that semi-Lagrangian discretization is able to follow every direction of the characteristic flow.

Both methods based on the fast marching technique are dramatically faster than the iterative semi-Lagrangian method. Nevertheless we want to note that only one iteration of the iterative scheme is less expensive with respect to the single iteration needed by fast marching–based algorithms. This is due to the fact that the narrow band technique requires that we (1) compute a minimum over nodes in the narrow band and (2) access the data in an almost random manner rather than in a systematic way along the loop indices (see [22]). Finally we note that the CPU time needed by the FM-SL method is slightly larger than the CPU time needed by the FM-FD method. This due to the fact that (1) the narrow band is bigger in the first method; and therefore the search for the minimum in the narrow band is more expensive; and (2) in the FM-SL method we need to compute the minimum over the unit ball $B(0,1)$.

Test 2. Γ_0 = unit square centered in $(-1,1)$ and rotated by $11.25^\circ \cup$ circle with radius $R = 0.5$ centered in $(0,-1) \cup$ square with side 0.4 centered in $(1.4, 1.4)$, $c(x,y) \equiv 1$. Exact solution: $T(x,y) =$ minimum between the distance function of the square rotated, the circle, and the square.

Results are summarized in Figure 7 and Table 3. In this test the shape of the initial front is much more complicated, but errors have the same behavior as in the previous simple Test 1, although the difference between errors is smaller.

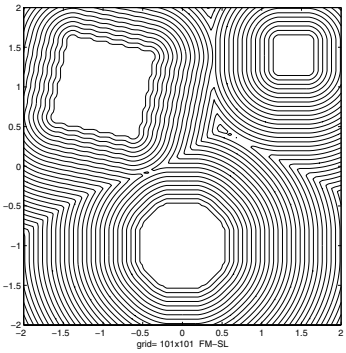


FIG. 7. *Level sets of $T(x)$ computed by the FM-SL method,* 101×101 *grid.*

TABLE 3
Errors for Test 2.

Method	Δx	L^∞ error	L^1 error	CPU time (sec)
FM-FD	0.08	0.0625	0.2154	0.5
FM-SL	0.08	0.0440	0.1849	0.7
SL (30 it)	0.08	0.0440	0.1849	4.9
FS-SL	0.08	0.0440	0.1849	0.7
FM-FD	0.04	0.0393	0.1120	2.2
FM-SL	0.04	0.0215	0.1044	3.1
SL (55 it)	0.04	0.0215	0.1044	34.1
FS-SL	0.04	0.0215	0.1044	2.9
FM-FD	0.02	0.0248	0.0669	10.2
FM-SL	0.02	0.0135	0.0633	14.5
SL (102 it)	0.02	0.0135	0.0633	246.6
FS-SL	0.02	0.0135	0.0633	11.4

TABLE 4
Rate of convergence in L^∞ and L^1 norms computed by errors in Table 3.

Method	L^∞ $(0.08 \to 0.04)$	L^∞ $(0.04 \to 0.02)$	L^1 $(0.08 \to 0.04)$	L^1 $(0.04 \to 0.02)$
FM-FD	0.6693	0.6642	0.9435	0.7434
FM-SL	1.0332	0.6714	0.8246	0.7218

FS-SL seems to be the best method. It has the smallest error and the CPU time is slightly larger than that of FM-FD. This is probably due to the fact that the structure of the narrow band is very complicated and is very large in terms of nodes.

Also in this case the rate of convergence (Table 4) is greater than 0.5.

7.2. Applications. In the following we try to use the FM-SL method in some classical applications of the eikonal equation such as the minimum time problem and shape-from-shading. We consider some cases not covered by the theory in which $c(x, y)$ is discontinuous, $c(x, y)$ vanishes in some regions (state constraints), and $c(x, y)$ has infinite values. We also consider the *anisotropic* case in which the velocity field c depends on (x, y) and on the control a. The results we obtained are very satisfactory even in these cases.

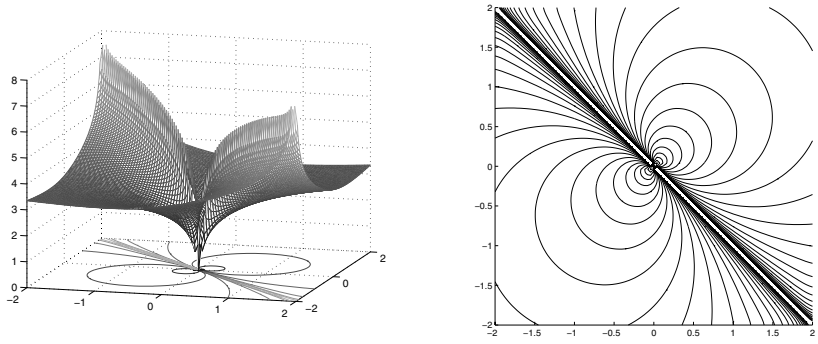


FIG. 8. *Value function T (left) and level sets of T (right).*

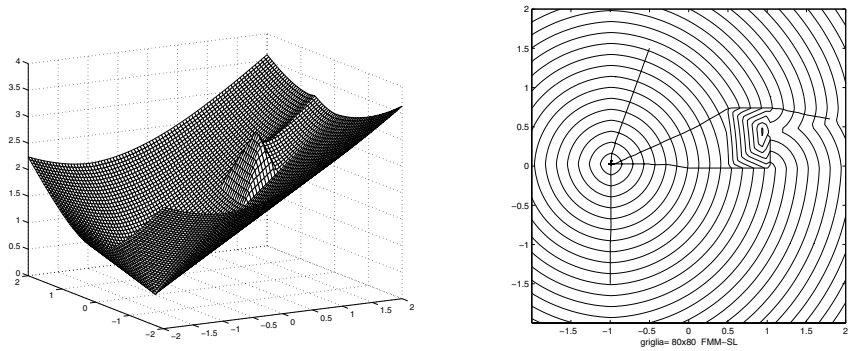


FIG. 9. *Value function T (left) and level sets of T with some optimal trajectories (right).*

Test 3: Nonconstant velocity. $\Gamma_0 = \partial B(0, \frac{\Delta x}{2})$, $c(x,y) = |x+y|$. In this case the velocity field is nonconstant. Figure 8 shows the value function $T(x,y)$ and level sets of T. On the line $x = -y$ the solution T is not defined since its correct value is $T = +\infty$. The FS-SL method needs 12 iterations to reach convergence and is more than three times slower than the FM-SL method on a 101×101 grid.

Test 4: Discontinuous vector field. $\Gamma_0 = (-1, 0)$.

$$c(x,y) = \begin{cases} 0.4, & (x,y) \in [0.5,1] \times [0,0.5], \\ 1 & \text{elsewhere.} \end{cases}$$

In this case the velocity field is discontinuous. Figure 9 shows the value function $T(x,y)$ and level sets of T. Figure 9 (right) also shows some optimal trajectories which start from four different points and reach the target Γ_0 in the minimum time with speed $c(x,y)$. The FS-SL method converges in 8 iterations.

Test 5: State constraint problem. $\Gamma_0 = (-1, -1)$.

$$c(x,y) = \begin{cases} 0, & (x,y) \in ([0,0.5] \times [-2,1.5]) \cup ([1,1.5] \times [-1.5,2]), \\ 1 & \text{elsewhere.} \end{cases}$$

In this test the velocity field vanishes in two different regions (the obstacles). Figure 10 shows the computational domain, the value function $T(x,y)$, and level sets of

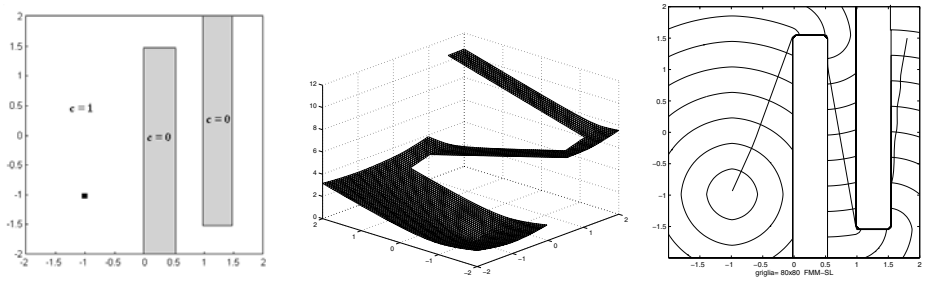


FIG. 10. *Domain of the equation (left), value function T (center), and level sets of T with one optimal trajectory (right).*

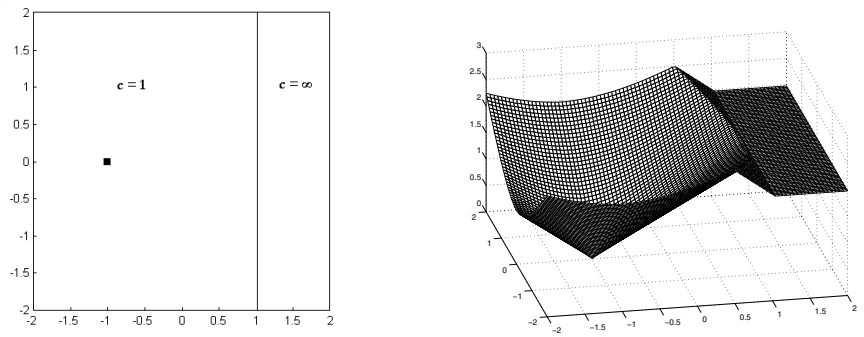


FIG. 11. *Domain of the equation (left) and value function T (right).*

T. Figure 10 (right) also shows one optimal trajectory which starts from the point $(1.8, 1.5)$ and reaches Γ_0 in the minimum time avoiding obstacles. We remark that since we use the Kružkov transform and compute v, we do not need to modify the numerical scheme to deal with state constraints. Also in this case the FS-SL method converges in 8 iterations.

Test 6: Infinite velocity. $\Gamma_0, = (-1, 0)$.

$$c(x,y) = \begin{cases} +\infty, & x \geq 1, \\ 1 & \text{elsewhere.} \end{cases}$$

In this case the front can propagate instantaneously in the region $R = \{x \geq 1\}$. It corresponds to the case of the following degenerate eikonal equation (see [30]):

$$|\nabla T(x,y)| = f(x,y) \quad \text{with } f = 0 \text{ in } R.$$

Figure 11 shows the computational domain and value function $T(x,y)$. In this test we used the technique described in section 3.2 in order to deal with this kind of vector field.

This technique allows us to reconstruct a perfect *flat* surface on R as the theory and the physical sense require. This technique can be very useful in shape-from-shading problems.

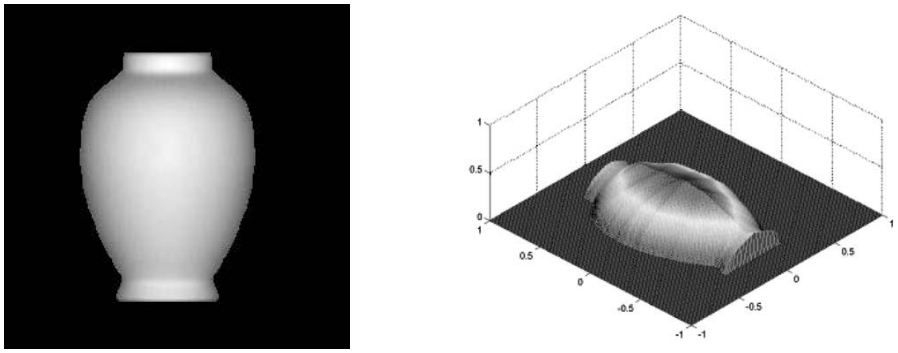


FIG. 12. *Initial image (left) and reconstructed surface (right).*

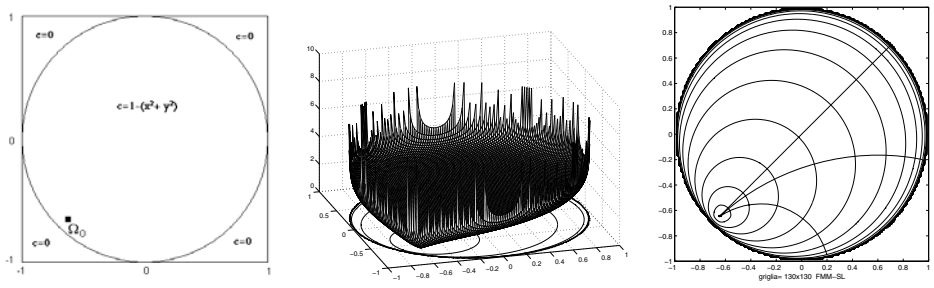


FIG. 13. *Domain of the equation (left), value function T (center), and level sets of T with some optimal trajectories (right).*

Test 7: Shape-from-shading. $Q = [-1,1]^2$, $\Gamma_0 =$ silhouette of a vase.

$$c(x,y) = \left(\sqrt{\frac{1}{I(x,y)^2} - 1}\right)^{-1}, \quad I(x,y) = \text{intensity light function.}$$

In this test we solve the shape-from-shading problem in the simple case of a vase. Figure 12 (left) shows the initial image and Figure 12 (right) shows the reconstructed surface. By the symmetry of the problem we guess that all characteristic curves start from the right and left sides of the image, so we can impose Dirichlet boundary condition just on the right and left sides of the domain and state constraints elsewhere as in [10] (see also [29, 24], where different boundary conditions are applied).

Test 8: Poincaré model. $Q = [-1,1]^2$, $\Gamma_0 = (-0.65, -0.65)$.

$$c(x,y) = \begin{cases} 1 - (x^2 + y^2), & x^2 + y^2 < 1, \\ 0 & \text{elsewhere.} \end{cases}$$

This example is an interesting application of the eikonal equation to the Poincaré model of the hyperbolic geometry. Figure 13 shows the computational domain, the value function $T(x,y)$, and level sets of T. The FS-SL method converges in 8 iterations.

As result of the particular choice of the velocity field (see [25]), the optimal trajectories of the associated minimum time problem correspond to the hyperbolic

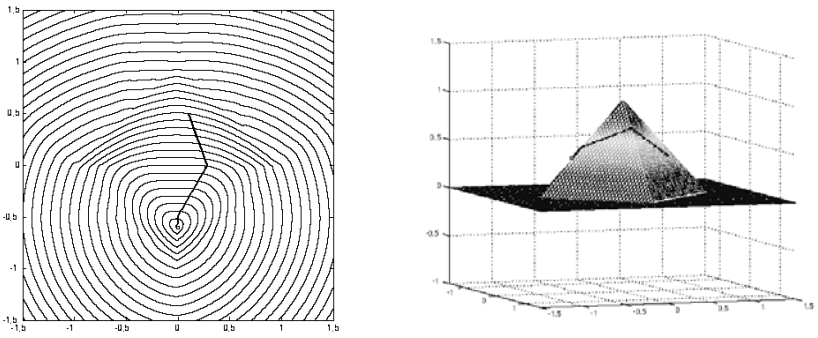


FIG. 14. *Level sets of T (left) and an optimal trajectory on the surface z (right).*

straight lines. Moreover, the level sets of T are hyperbolic circles with center Γ_0 (i.e., the sets of points which have the same hyperbolic distance from Γ_0).

Test 9: Geodesics on a nonsmooth surface. $Q = [-1.5, 1.5]^2, \Gamma_0 = (0, -0.6)$.

$$\text{Surface} : z(x,y) = \begin{cases} 1 - (|x| + |y|), & |x| + |y| < 1, \\ 0 & \text{elsewhere.} \end{cases}$$

In this case we want to solve a minimum time problem on a surface $z = z(x, y)$. The three-dimensional problem can be easily reduced to a two-dimensional problem modifying the velocity field according to the function z. In fact, if the intrinsic velocity on the surface in equal to 1, it can be shown (see [32, 23]) that the velocity of the corresponding two-dimensional problem becomes

$$c(x, y, a) = \frac{1}{\sqrt{1 + (\nabla z \cdot a)^2}}.$$

Figure 14 shows the level sets of T and the surface with an optimal trajectory on it. The starting point is $(0, 0.5)$.

We remark that the dependence of c on a changes the properties of the solution of the equation. In fact the equation for anisotropic front propagation is

$$\begin{cases} v(x) + \max\limits_{a \in B(0,1)} \{c(x,a)a \cdot \nabla v(x)\} = 1, & x \in \mathbb{R}^n \backslash \Omega_0, \\ v(x) = 0, & x \in \partial\Omega_0. \end{cases} \tag{41}$$

In this case the fast marching technique is no longer directly applicable (there is no guarantee that convergence is reached in just one iteration; see [34]). This is true for the FM-SL method too, but we stress that scheme (26) requires tiny modifications to deal with this kind of velocity field. Moreover, if we use the function w computed by the FM-SL method as a starting point of the iterative semi-Lagrangian scheme, we can reach convergence in very few iterations.

Appendix. Convergence of the FM-FD method in a finite number of steps.

Proof of Proposition 2.1. We will assume that B, C, and D are the neighbors of X which can have a label *accepted*, *narrow band,* or *far* (see Figure 15). We will

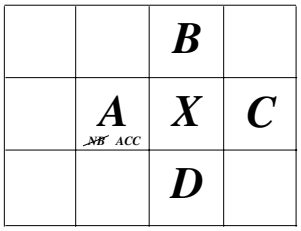


FIG. 15. *The neighboring nodes of* X.

prove the result by induction on the number of iterations of the algorithm. We will always assume

$$T(B) \leq T(D), \tag{42}$$

which is not restrictive since we can always switch the B and D.

In the first iteration we simply have $T(X) = 0 + f_X$ and (16) is satisfied. Let us consider the nth step of the algorithm. The induction hypothesis implies that at each iteration the values of nodes in the narrow band are greater than values of nodes labeled as *accepted* at the same iteration. Therefore, by construction we have that, given two nodes Y and Z,

$$\text{if } Y \text{ has become accepted before } Z, \text{ then } T(Y) \leq T(Z).$$

The proof will be divided into four cases.

Case 1. B is far. C and D are narrow band or far.

By assumption $T(B) = +\infty$, and since $T(B) \leq T(D)$ this implies that D must be far. Moreover, we have

$$T(C) \geq T(A)$$

since A has been chosen among all the nodes of the narrow band to become accepted. Also X must be far, since it has never been computed. Then by (10) we get

$$T(X) = T(A) + f_X. \tag{43}$$

Since $f_X > 0$, (43) implies

$$T(A) \leq T(X) \leq T(A) + f_X.$$

Case 2. B is narrow band. C and D are narrow band or far.

Also in this case X is far. We have

$$T(A) \leq T(B), \qquad T(A) \leq T(C)$$

since A is the minimal node in narrow band. Moreover, the assumption (42) implies that $T(X)$ will be computed by the values at A and B. From (10) we get

$$T(X) = \frac{T(A) + T(B) + \sqrt{2f_X^2 - (T(A) - T(B))^2}}{2} \tag{44}$$

and then

$$T(X) \geq \frac{T(A) + T(B)}{2} \geq \frac{T(A) + T(A)}{2} = T(A). \tag{45}$$

Since $T(X)$ solves

$$(T(X)-T(A))^2+(T(X)-T(B))^2=f_X^2$$

we have

$$(T(X)-T(A))^2\le f_X^2.$$

Since all the terms in the above equation are positive we conclude that

$$T(X)-T(A)\le f_X. \tag{46}$$

Case 3. B is accepted. C and D are narrow band or far.

This situation occurs when X has been already computed once (when B has been labeled as *accepted*). Let us denote its value by $T_{old}(X)$. The node X is then in the narrow band and has to be recomputed because A has just been labeled as *accepted*. Let us note that in the previous computation $T_{old}(X)$ has been computed according to the rules examined in Case 1 or 2. Then we have

$$T(B)\le T_{old}(X)\le T(B)+f_X.$$

Moreover $T(A)\le T_{old}(X)$ because A just became accepted and $T(B)\le T(A)$ since B became accepted before A (induction).

These inequalities imply

$$T(B)\le T(A)\le T_{old}(X)\le T(B)+f_X \tag{47}$$

and

$$0\le T(A)-T(B)\le f_X. \tag{48}$$

The value at X, which will be denoted by $T_{new}(X)$, will depend on $T(A)$ and $T(B)$.

By (48) and (47) we derive

$$\begin{aligned} T_{new}(X) &= \frac{T(A)+T(B)+\sqrt{2f_X^2-(T(A)-T(B))^2}}{2} \\ &\ge \frac{T(A)+(T(B)+f_X)}{2}\ge\frac{T(A)+T(A)}{2}=T(A) \end{aligned} \tag{49}$$

and

$$T_{new}(X)\le\frac{T(A)+T(B)+\sqrt{2}f_X}{2}\le T(A)+\frac{\sqrt{2}}{2}f_X\le T(A)+f_X.$$

Case 4. B is narrow band or far. C is accepted. D is narrow band or far.

In this case X has already been computed because it is a neighbor of C. It belongs to the narrow band and has a value $T_{old}(X)$. Besides

$$T(A)\le T_{old}(X) \tag{50}$$

since on the contrary X would have been chosen instead of A as the node to be accepted and

$$T(A)\le T(B) \tag{51}$$

for the same reason. Moreover we have $T(C) \leq T(A)$ by induction and $T(B) \leq T(D)$ by assumption. In conclusion, the nodes contributing to the computation of $T(X)$ are C and B or only C. The fact that A has been labeled as *accepted* has no effect on the computation so we are again in Case 1 or 2. This implies,

$$T(C) \leq T_{new}(X) \leq T(C) + f_X \leq T(A) + f_X.$$

Now we prove that $T_{new} \geq T(A)$. When $T_{old}(X)$ was computed the algorithm was in the Case 1 or 2, so

$$T(C) \leq T_{old}(X) \leq T(C) + f_X. \tag{52}$$

Moreover we have

$$T(C) \leq T(B) \tag{53}$$

by induction.

If $T(B) > T_{old}(X)$, then the node contributing to the computation of $T(X)$ is only C, so we have

$$T_{new}(X) = T(C) + f_X \geq T_{old}(X) \geq T(A).$$

Otherwise, if $T(B) \leq T_{old}(X)$, the nodes contributing to the computation of $T(X)$ are C and B.

Using this last assumption, (52), and (53) we have

$$\begin{aligned} T(C) \leq T(B) \leq T_{old}(X) \leq T(C) + f_X &\Rightarrow 0 \leq T(B) - T(C) \leq f_X \\ &\Rightarrow (T(B) - T(C))^2 \leq f_X^2. \end{aligned} \tag{54}$$

Moreover, by (50) and (52) we have

$$T(C) + f_X \geq T(A). \tag{55}$$

Computation of X leads to

$$\begin{aligned} T_{new}(X) &= \frac{T(C) + T(B) + \sqrt{2f_X^2 - (T(C) - T(B))^2}}{2} \\ &= \frac{(T(C) + f_X) - f_X + T(B) + \sqrt{2f_X^2 - (T(C) - T(B))^2}}{2}. \end{aligned} \tag{56}$$

Using (55), (54), and (51) we obtain

$$\begin{aligned} T_{new}(X) &\geq \frac{T(A) - f_X + T(B) + \sqrt{2f_X^2 - (T(C) - T(B))^2}}{2} \\ &\geq \frac{T(A) - f_X + T(B) + \sqrt{f_X^2}}{2} \geq \frac{T(A) + T(A)}{2} = T(A). \end{aligned} \tag{57}$$

Finally, let us remark that the cases when two or more nodes among B, C, and D are accepted can be treated as in the previous cases. Note that if D is accepted, then B must also be accepted since $T(B) \leq T(D)$.

To complete the proof, it is necessary to show that the expression appearing under the square root in the computation of $T(X)$ expressed as a function of its two

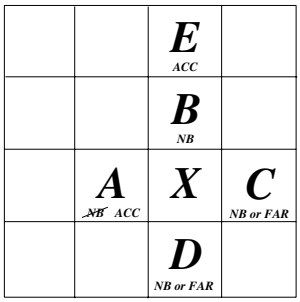


FIG. 16. *Proof that radicand is positive under the CFL-like condition* (15).

neighbors is nonnegative. Let us start by proving that the hypothesis (15) guarantees that

$$\frac{c(Z)}{c(Z')} \leq \sqrt{2} \tag{58}$$

for any couple of nodes Z and Z' such that

$$|Z - Z'| = \Delta x.$$

In fact, by assumption we have

$$|c(Z) - c(Z')| \leq L_c |Z - Z'|.$$

If $|Z - Z'| = \Delta x$, we have that

$$|c(Z) - c(Z')| \leq L_c \Delta x \leq (\sqrt{2} - 1) c_{min} \leq (\sqrt{2} - 1) c(Z'),$$

which implies

$$c(Z) - c(Z') \leq (\sqrt{2} - 1) c(Z')$$

and then

$$c(Z) \leq \sqrt{2}\, c(Z').$$

Let us examine the three cases where we need to show that the radicand is nonnegative.

Case 2. Since B is in the narrow band, there must be at least one neighbor belonging to accepted. Let E be this node (see Figures 16 and 1). Moreover, $T(A) \leq T(B)$ since A has been chosen to be labeled as *accepted* and $T(E) \leq T(A)$ because E became accepted before A. By the previous results, we get

$$T(E) \leq T(B) \leq T(E) + f_B,$$

which implies

$$T(A) \leq T(B) \leq T(E) + f_B \leq T(A) + f_B$$

and

$$0 \leq T(B) - T(A) \leq f_B. \tag{59}$$

Choosing $Z = X$ and $Z' = B$ in (58), we get

$$\frac{c(X)}{c(B)} \leq \sqrt{2}$$

and then

$$\sqrt{2} f_X \geq f_B. \tag{60}$$

Finally (59) and (60) imply

$$\sqrt{2} f_X \geq T(B) - T(A) \geq 0,$$

so we can conclude that

$$2 f_X^2 - (T(B) - T(A))^2 \geq 0.$$

Case 3 and 4. In these cases, (48) and (54) guarantee, respectively, that the expression appearing under the radicand is always positive. □

Let us show now that the value at the node which is labeled as *accepted* at every iteration is exact. Let us denote this value by T_{min}. Since all the nodes in the narrow band have values greater than T_{min}, the previous result implies that using those nodes we cannot assign to a node a value lower than T_{min}. In conclusion (see [33]), the up-winding is respected and the value T_{min} can be considered exact since it cannot be improved on the same grid (of course it can be improved if we reduce the discretization steps).

Note that Theorem 1 is valid also for the FM-FD method.

Acknowledgment. The authors wish to thank M. Sagona for helpful discussions on the FM-FD and FM-SL methods.

REFERENCES

[1] S. Augoula and R. Abgrall, *High order numerical discretization for Hamilton-Jacobi equations on triangular meshes*, J. Sci. Comput., 15 (2000), pp. 197–229.
[2] M. Bardi and I. Capuzzo Dolcetta, *Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations*, Birkhäuser, Boston, 1997.
[3] M. Bardi and M. Falcone, *An approximation scheme for the minimum time function*, SIAM J. Control Optim., 28 (1990), pp. 950–965.
[4] G. Barles, *Solutions de viscosité des equations de Hamilton-Jacobi*, Springer, Berlin, 1994.
[5] F. Camilli and A. Siconolfi, *Hamilton-Jacobi equations with measurable dependence on the state variable*, Adv. Differential Equations, 8 (2003), pp. 733–768.
[6] E. Carlini, M. Falcone, and R. Ferretti, *An efficient algorithm for Hamilton-Jacobi equations in high dimension*, Comput. Vis. Sci., 7 (2004), pp. 15–29.
[7] D. L. Chopp, *Some improvements of the fast marching method*, SIAM J. Sci. Comput., 23 (2001), pp. 230–244.
[8] M. G. Crandall and P.-L. Lions, *Two approximations of solutions of Hamilton-Jacobi equations*, Math. Comp., 43 (1984), pp. 1–19.
[9] E. Cristiani, *Fast Marching and Semi-Lagrangian Methods for Hamilton-Jacobi Equations with Applications*, Ph.D. thesis, Dipartimento di Metodi e Modelli Matematici per le Scienze Applicate, SAPIENZA - Università di Roma, Rome, Italy, 2007.
[10] E. Cristiani, M. Falcone, and A. Seghini, *Numerical solution of the perspective shape from shading problem*, in Proceedings of Control Systems: Theory, Numerics and Applications PoS (CSTNA2005) 008, http://pos.sissa.it/.
[11] P. Danielsson, *Euclidean distance mapping*, Comput. Graphics Image Process., 14 (1980), pp. 227–248.
[12] M. Falcone, *The minimum time problem and its applications to front propagation*, in Motion by Mean Curvature and Related Topics, A. Visintin and G. Buttazzo, eds., de Gruyter, Berlin, 1994, pp. 70–88.

[13] M. Falcone, *Numerical solution of dynamic programming equations*, in Optimal Control and Viscosity Solutions of Hamilton-Jacobi-Bellman Equations, M. Bardi and I. Capuzzo Dolcetta, eds., Birkhäuser, Boston, 1997, Appendix A, pp. 471–504.

[14] M. Falcone, *Some remarks on the synthesis of feedback controls via numerical methods*, in Optimal Control and Partial Differential Equations, J. L. Menaldi, E. Rofman, and A. Sulem, eds., IOS Press, The Netherlands, 2001, pp. 456–465.

[15] M. Falcone and R. Ferretti, *Discrete time high-order schemes for viscosity solutions of Hamilton-Jacobi-Bellman equations*, Numer. Math., 67 (1994), pp. 315–344.

[16] M. Falcone and R. Ferretti, *Semi-Lagrangian schemes for Hamilton-Jacobi equations, discrete representation formulae and Godunov methods*, J. Comput. Phys., 175 (2002), pp. 559–575.

[17] M. Falcone, T. Giorgi, and P. Loreti, *Level sets of viscosity solutions: Some applications to fronts and rendez-vous problems*, SIAM J. Appl. Math., 54 (1994), pp. 1335–1354.

[18] M. Falcone and C. Truini, *A Level-Set Algorithm for Front Propagation in the Presence of Obstacles*, Technical Report, Dipartimento di Matematica, Università di Roma "La Sapienza," Rome, Italy, 2003.

[19] R. Fedkiw and S. Osher, *Level set methods and dynamic implicit surfaces*, Appl. Math. Sci. 153, Springer, New York, 2003.

[20] C.-Y. Kao, S. Osher, and J. Qian, *Lax-Friedrichs sweeping scheme for static Hamilton–Jacobi equations*, J. Comput. Phys., 196 (2004), pp. 367–391.

[21] C.-Y. Kao, S. Osher, and Y.-H. Tsai, *Fast sweeping methods for static Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 42 (2005), pp. 2612–2632.

[22] S. Kim, *An $O(N)$ level set method for eikonal equations*, SIAM J. Sci. Comput., 22 (2001), pp. 2178–2193.

[23] R. Kimmel and J. A. Sethian, *Computing geodesic paths on manifold*, Proc. Natl. Acad. Sci. USA, 95 (1998), pp. 8431–8435.

[24] R. Kimmel and J. A. Sethian, *Optimal algorithm for shape from shading and path planning*, J. Math. Imaging Vision, 14 (2001), pp. 237–244.

[25] B. O'Neill, *Elementary Differential Geometry*, Academic Press, New York, London, 1966.

[26] S. Osher, *A level set formulation for the solution of the Dirichlet problem for Hamilton–Jacobi equations*, SIAM J. Math. Anal., 24 (1993), pp. 1145–1152.

[27] J. Qian, Y.-T. Zhang, and H. K. Zhao, *Fast sweeping methods for eikonal equations on triangular meshes*, SIAM J. Numer. Anal., 45 (2007),pp. 83–107.

[28] J. Qian, Y.-T. Zhang, and H. Zhao, *A fast sweeping method for static convex Hamilton-Jacobi equations*, J. Sci. Comput., 31 (2007), pp. 237–271.

[29] E. Rouy and A. Tourin, *A viscosity solutions approach to shape-from-shading*, SIAM J. Numer. Anal., 29 (1992), pp. 867–884.

[30] M. Sagona, *Numerical Methods for Degenerate Eikonal Type Equation and Applications*, Ph.D. thesis, Dipartimento di Matematica, Università di Napoli "Federico II," Naples, Italy, 2001.

[31] M. Sagona and A. Seghini, *An adaptive scheme for the shape from shading problem*, in Numerical Methods for Viscosity Solutions and Applications, M. Falcone and Ch. Makridakis, eds., World Scientific, Singapore, 2001, pp. 197–219.

[32] J. A. Sethian, *Level Set Methods and Fast Marching Methods. Evolving Interfaces in Computational Geometry, Fluid Mechanics, Computer Vision, and Materials Science*, Cambridge University Press, Cambridge, UK, 1999.

[33] J. A. Sethian, *A fast marching level set method for monotonically advancing fronts*, Proc. Natl. Acad. Sci. USA, 93 (1996), pp. 1591–1595.

[34] J. A. Sethian and A. Vladimirsky, *Ordered upwind methods for static Hamilton–Jacobi equations: Theory and algorithms*, SIAM J. Numer. Anal., 41 (2003), pp. 325–363.

[35] P. Soravia, *Boundary value problems for Hamilton-Jacobi equations with discontinuous Lagrangian*, Indiana Univ. Math. J., 51 (2002), pp. 451–477.

[36] C. Truini, *Approssimazione numerica del problema controllistico di tempo minimo e applicazione all'evoluzione dei fronti*, Master Thesis, Dipartimento di Matematica, Università di Roma "La Sapienza," Rome, Italy, 1995.

[37] Y.-H. R. Tsai, L.-T. Cheng, S. Osher, and H.-K. Zhao, *Fast sweeping algorithms for a class of Hamilton–Jacobi equations*, SIAM J. Numer. Anal., 41 (2003), pp. 673–694.

[38] J. N. Tsitsiklis, *Efficient algorithms for globally optimal trajectories*, IEEE Trans. Automat. Control, 40 (1995), pp. 1528–1538.

[39] H. Zhao, *A fast sweeping method for eikonal equations*, Math. Comp., 74 (2005), pp. 603–627.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2012–2031

AN ERROR ESTIMATE FOR THE SIGNORINI PROBLEM WITH COULOMB FRICTION APPROXIMATED BY FINITE ELEMENTS*

PATRICK HILD† AND YVES RENARD‡

Abstract. The present paper is concerned with the unilateral contact model and the Coulomb friction law in linear elastostatics. We consider a mixed formulation in which the unknowns are the displacement field and the normal and tangential constraints on the contact area. The chosen finite element method involves continuous elements of degree one and continuous piecewise affine multipliers on the contact zone. A convenient discrete contact and friction condition is introduced in order to perform a convergence study. We finally obtain a first a priori error estimate under the assumptions ensuring the uniqueness of the solution to the continuous problem.

Key words. unilateral contact, Coulomb friction, uniqueness of solution, finite elements, a priori error estimate

AMS subject classifications. 35J85, 65N30, 74M10, 74M15

DOI. 10.1137/050645439

Introduction. This study deals with the unilateral contact problem governed by the Coulomb friction law in linear elasticity. We consider a simplified model, the so-called static friction problem, which roughly corresponds to an incremental problem in the time discretized quasi-static model and whose solutions are also some particular equilibrium configurations of the dynamic problem.

From a mathematical point of view the early progress made on the static problem was accomplished in [15, 17]. These studies concerned the weak formulation of the problem. The first existence results were obtained in [39] for an infinite elastic strip. Thereafter, many existence results followed for general domains, in particular in [18] (see also the references quoted therein). These existence results hold for small friction coefficients, and uniqueness is not discussed. In fact, uniqueness does not hold in the general case, at least for large friction coefficients; see [26, 27]. More recently a first uniqueness result has been obtained in [41] with the assumption that a "regular" solution exists and that the friction coefficient is sufficiently small. Additionally, the so-called nonlocal Coulomb frictional models mollifying the normal stresses were introduced in [16] and developed in [14, 11, 31]. The smoothing map used in the nonlocal friction model allows one to obtain existence results for any friction coefficient. Moreover, uniqueness of a solution can also be established if the friction coefficient is small enough (see [16, 14, 11, 31]). The same type of result (existence for any friction coefficient and uniqueness for small friction coefficients) was obtained in [32, 33] for the normal compliance model, introduced in [40, 37].

From a numerical point of view, the finite element method is commonly used when approximating such frictional contact problems (see, e.g., [31, 24, 21, 35, 43]). It is well known that the finite element problem, associated with the continuous static

*Received by the editors November 16, 2005; accepted for publication (in revised from) March 9, 2007; published electronically September 12, 2007. This work is supported by "l'Agence Nationale de la Recherche," project ANR-05-JCJC-0182-01.

http://www.siam.org/journals/sinum/45-5/64543.html

†Laboratoire de Mathématiques de Besançon, CNRS UMR 6623, Université de Franche-Comté, 16 route de Gray, 25030 Besançon, France (patrick.hild@univ-fcomte.fr).

‡MIP, CNRS UMR 5640, INSAT, Complexe scientifique de Rangueil, 31077 Toulouse, France (yves.renard@insa-lyon.fr)

Coulomb friction model, always admits a solution and that the solution is unique if the friction coefficient is small enough. (Unfortunately the definition of small depends on the discretization parameter, and the bound ensuring uniqueness vanishes as the mesh is refined; see, e.g., [24].) The former result holds for any reasonable choice of the approximated contact and friction conditions (see [30]). Moreover, a first convergence study of the finite element problem towards the continuous model was accomplished in [22], where convergence was obtained under the assumptions ensuring the existence of a solution in [39] (i.e., small friction coefficient). This result proves the existence of a subsequence of discrete solutions converging towards a solution to the continuous problem. A similar result is obtained in [42] for the quasi-static model.

Our purpose is to carry out a convergence analysis and to obtain an a priori error estimate for a finite element discretization of the frictional contact conditions under the assumptions ensuring the uniqueness of a solution to the continuous problem obtained in [41]. As far as we know, this work presents the first error estimate with a convergence rate for this model.

Our paper is outlined as follows. Section 1 is concerned with the setting of the continuous problem, several equivalent weak formulations, and a presentation of the tools and techniques leading to the uniqueness result. In section 2 we consider a discretization of the problem with finite elements of degree one and continuous piecewise affine multipliers on the contact zone. We introduce a convenient discrete contact and friction condition which allows us to perform a convergence analysis and to obtain an a priori estimate of the discretization error with a quasi-optimal convergence rate of order $h^{1/2}$ in the energy norm under $H^{(3/2)+\varepsilon}$-regularity assumptions on the displacements.

1. The Signorini problem with Coulomb friction. Let $\Omega \subset R^d$ ($d = 2$ or 3) be a polygonal domain representing the reference configuration of a linearly elastic body whose boundary $\partial\Omega$ consists of three nonoverlapping open parts $\Gamma_{_N}$, $\Gamma_{_D}$, and $\Gamma_{_C}$ with $\overline{\Gamma_{_N}} \cup \overline{\Gamma_{_D}} \cup \overline{\Gamma_{_C}} = \partial\Omega$. We assume that the measures of $\Gamma_{_C}$ and $\Gamma_{_D}$ are positive and, in order to simplify that $\Gamma_{_C}$ is a straight line segment when $d = 2$ or a plane surface when $d = 3$. The body is submitted to a Neumann condition on $\Gamma_{_N}$ with a density of loads $F \in (L^2(\Gamma_{_N}))^d$, a Dirichlet condition on $\Gamma_{_D}$ (the body is assumed to be clamped on $\Gamma_{_D}$ to simplify), and to volume loads denoted $f \in (L^2(\Omega))^d$ in Ω. Finally, a unilateral contact condition with static Coulomb friction between the body and a flat rigid foundation holds on $\Gamma_{_C}$ (see Figure 1.1).

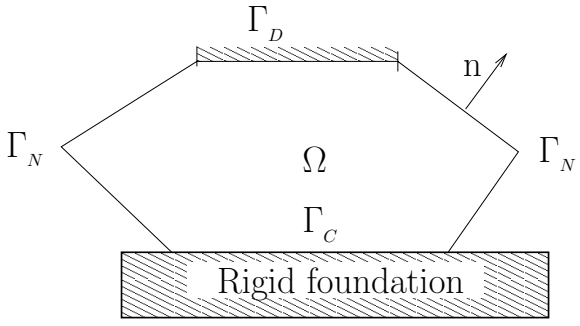


FIG. 1.1. *Elastic body* Ω *in frictional contact.*

The problem consists of finding the displacement field $u : \overline{\Omega} \to R^d$ satisfying

$$-\operatorname{div} \sigma(u) = f \quad \text{in } \Omega, \tag{1.1}$$

$$\sigma(u) = \mathcal{A}\varepsilon(u) \quad \text{in } \Omega, \tag{1.2}$$

$$\sigma(u)\mathrm{n} = F \quad \text{on } \Gamma_{_N}, \tag{1.3}$$

$$u = 0 \quad \text{on } \Gamma_{_D}, \tag{1.4}$$

where $\sigma(u)$ represents the stress tensor field, $\varepsilon(u) = (\nabla u + (\nabla u)^T)/2$ denotes the linearized strain tensor field, n stands for the outward unit normal to Ω on $\partial\Omega$, and $\mathcal{A}$ is the fourth order elastic coefficient tensor which satisfies the usual symmetry and ellipticity conditions and whose components are in $L^\infty(\Omega)$.

On $\Gamma_{_C}$, we decompose the displacement and the stress vector fields in normal and tangential components as follows:

$$u_{_N} = u.\mathrm{n}, \quad u_{_T} = u - u_{_N}\mathrm{n},$$

$$\sigma_{_N}(u) = (\sigma(u)\mathrm{n}).\mathrm{n}, \quad \sigma_{_T}(u) = \sigma(u)\mathrm{n} - \sigma_{_N}(u)\mathrm{n}.$$

The unilateral contact condition on $\Gamma_{_C}$ is expressed by the following complementary condition:

$$u_{_N} \le 0, \qquad \sigma_{_N}(u) \le 0, \qquad u_{_N}\sigma_{_N}(u) = 0, \tag{1.5}$$

where a vanishing gap between the elastic solid and the rigid foundation has been chosen in the reference configuration.

Denoting by $\mathcal{F} \ge 0$ the given friction coefficient on $\Gamma_{_C}$ (which is supposed constant for the sake of simplicity), the *static Coulomb friction* condition reads as

$$\text{if } \ u_{_T} = 0, \text{ then } |\sigma_{_T}(u)| \le -\mathcal{F}\sigma_{_N}(u), \tag{1.6}$$

$$\text{if } \ u_{_T} \ne 0, \text{ then } \sigma_{_T}(u) = \mathcal{F}\sigma_{_N}(u)\frac{u_{_T}}{|u_{_T}|}. \tag{1.7}$$

When $\mathcal{F} = 0$ the friction conditions (1.6)–(1.7) merely reduce to $\sigma_{_T}(u) = 0$ on $\Gamma_{_C}$.

1.1. Classical weak formulations. This section is devoted to the presentation of different and equivalent weak formulations of the Coulomb friction problem. Let us introduce the following Hilbert spaces:

$$V = \left\{ v \in (H^1(\Omega))^d, v = 0 \text{ on } \Gamma_{_D} \right\},$$

$$X = \left\{ v_{|_{\Gamma_{_C}}} : v \in V \right\} \subset (H^{1/2}(\Gamma_{_C}))^d,$$

$$X_{_N} = \left\{ v_{_N}|_{\Gamma_{_C}} : v \in V \right\}, \qquad X_{_T} = \left\{ v_{_T}|_{\Gamma_{_C}} : v \in V \right\},$$

and their topological dual spaces V', X', $X'_{_N}$, and $X'_{_T}$, endowed with their usual norms. Since $\Gamma_{_C}$ is a straight line segment ($d = 2$) or a plane surface ($d = 3$), we have $H^{1/2}_{00}(\Gamma_{_C}) \subset X_{_N} \subset H^{1/2}(\Gamma_{_C})$, $(H^{1/2}_{00}(\Gamma_{_C}))^{d-1}, \subset X_{_T} \subset (H^{1/2}(\Gamma_{_C}))^{d-1}$, which implies $X'_{_N} \subset H^{-1/2}(\Gamma_{_C})$ and $X'_{_T} \subset (H^{-1/2}(\Gamma_{_C}))^{d-1}$, where we denote by H^s the standard Sobolev spaces (see [1]). Classically, $H^{1/2}(\Gamma_{_C})$ is the space of the restrictions on $\Gamma_{_C}$

of traces on $\partial\Omega$ of functions in $H^1(\Omega)$, and $H^{-1/2}(\Gamma_C)$ is the dual space of $H^{1/2}_{00}(\Gamma_C)$, which is the space of the restrictions on Γ_C of functions in $H^{1/2}(\partial\Omega)$ vanishing outside Γ_C. We refer to [36, 1, 31] for a detailed presentation of trace operators and/or trace spaces.

The set of admissible displacements satisfying the noninterpenetration conditions on the contact zone is

$$K = \{v \in V, v_N \leq 0 \quad \text{a.e. on } \Gamma_C\}.$$

Take as given the following forms for any u and v in V:

$$a(u,v) = \int_\Omega \mathcal{A}\varepsilon(u) : \varepsilon(v)\, d\Omega,$$

$$l(v) = \int_\Omega f.v\, d\Omega + \int_{\Gamma_N} F.v\, d\Gamma,$$

which represent the virtual work of the elastic forces and of the external loads, respectively. If $\langle\cdot,\cdot\rangle_{X'_N,X_N}$ stands for the duality pairing between X'_N and X_N, then the "virtual work" of the friction forces is given by

$$j(\mathcal{F}\lambda_N, v_T) = -\langle \mathcal{F}\lambda_N, |v_T|\rangle_{X'_N,X_N}$$

for any $\lambda_N \in X'_N$ and $v_T \in X_T$. From the previous assumptions it follows that

$a(\cdot,\cdot)$ is a bilinear symmetric V-elliptic and continuous form on $V \times V$:
$\exists\ \alpha > 0, \exists\ M > 0, \quad a(v,v) \geq \alpha\|v\|^2_V, a(u,v) \leq M\|u\|_V\|v\|_V \quad \forall u, v \in V,$
$l(\cdot)$ linear continuous form on V, i.e., $\exists\ L > 0, \quad |l(v)| \leq L\|v\|_V \quad \forall v \in V.$

Moreover, $j(\mathcal{F}\lambda_N, v_T)$ is linear continuous with respect to λ_N and convex lower semicontinuous with regard to v_T if λ_N is a nonpositive element of X'_N (see, for instance, [2]).

Clearly $a(\cdot,\cdot)$ is an inner product on V, and the associated norm,

$$\|v\|_a = (a(v,v))^{1/2},$$

is equivalent to the usual norm of V:

$$\sqrt{\alpha}\|v\|_V \leq \|v\|_a \leq \sqrt{M}\|v\|_V \quad \forall v \in V.$$

The continuity constant of $l(\cdot)$ can also be given with respect to $\|\cdot\|_a$:

$$\exists\ L_a > 0, \quad |l(v)| \leq L_a\|v\|_a \quad \forall v \in V.$$

Constants L and L_a can be chosen such that

$$\sqrt{\alpha}L_a \leq L \leq \sqrt{M}L_a.$$

The weak formulation of problem (1.1)–(1.7) (written as an inequality), introduced in [15] (see also [17]), is

$$\text{(1.8)} \qquad \begin{cases} \text{Find } u \in K \text{ satisfying} \\ a(u, v-u) + j(\mathcal{F}\sigma_N(u), v_T) - j(\mathcal{F}\sigma_N(u), u_T) \geq l(v-u) \ \ \forall\ v \in K. \end{cases}$$

Introducing the stresses on the contact boundary as an unknown in the previous formulation, one obtains the following equivalent formulation (see [30]):

$$
\left\{
\begin{array}{l}
\text{Find } u \in V, \lambda_{_N} \in X_{_N}, \text{ and } \lambda_{_T} \in X_{_T} \text{ satisfying} \\
\\
a(u,v) = l(v) + \langle \lambda_{_N}, v_{_N} \rangle_{X'_{_N},X_{_N}} + \langle \lambda_{_T}, v_{_T} \rangle_{X'_{_T},X_{_T}} \quad \forall v \in V, \\
\\
u_{_N} \leq 0, \quad \langle \lambda_{_N}, v_{_N} - u_{_N} \rangle_{X'_{_N},X_{_N}} \geq 0 \quad \forall v_{_N} \in X_{_N}, v_{_N} \leq 0, \\
\\
\langle \lambda_{_T}, v_{_T} - u_{_T} \rangle_{X'_{_T},X_{_T}} + j(\mathcal{F}\lambda_{_N}, v_{_T}) - j(\mathcal{F}\lambda_{_N}, u_{_T}) \geq 0 \quad \forall v_{_T} \in X_{_T}.
\end{array}
\right. \tag{1.9}
$$

Inverting contact and friction relations, one also obtains the classical equivalent hybrid formulation (see [30]):

$$
\left\{
\begin{array}{l}
\text{Find } u \in V, \lambda_{_N} \in X_{_N}, \text{ and } \lambda_{_T} \in X_{_T} \text{ satisfying} \\
\\
a(u,v) = l(v) + \langle \lambda_{_N}, v_{_N} \rangle_{X'_{_N},X_{_N}} + \langle \lambda_{_T}, v_{_T} \rangle_{X'_{_T},X_{_T}} \quad \forall v \in V, \\
\\
\lambda_{_N} \in \Lambda_{_N}, \quad \langle \mu_{_N} - \lambda_{_N}, u_{_N} \rangle_{X'_{_N},X_{_N}} \geq 0 \quad \forall \mu_{_N} \in \Lambda_{_N}, \\
\\
\lambda_{_T} \in \Lambda_{_T}(\mathcal{F}\lambda_{_N}), \quad \langle \mu_{_T} - \lambda_{_T}, u_{_T} \rangle_{X'_{_T},X_{_T}} \geq 0 \quad \forall \mu_{_T} \in \Lambda_{_T}(\mathcal{F}\lambda_{_N}),
\end{array}
\right. \tag{1.10}
$$

where $\Lambda_{_N}$ and $\Lambda_{_T}(\mathcal{F}\lambda_{_N})$ denote the sets of admissible normal and tangential stresses:

$$
\Lambda_{_N} = \left\{ \lambda_{_N} \in X'_{_N} : \langle \lambda_{_N}, v_{_N} \rangle_{X'_{_N},X_{_N}} \geq 0 \ \ \forall v_{_N} \leq 0 \right\},
$$

$$
\Lambda_{_T}(\mathcal{F}\lambda_{_N}) = \left\{ \lambda_{_T} \in X'_{_T} : \langle \lambda_{_T}, v_{_T} \rangle_{X'_{_T},X_{_T}} + j(\mathcal{F}\lambda_{_N}, v_{_T}) \geq 0, \ \ \forall v_{_T} \in X_{_T} \right\}.
$$

It is easy to check that the multipliers $\lambda_{_N}$ and $\lambda_{_T}$ solving (1.9) and (1.10) satisfy $\lambda_{_N} = \sigma_{_N}(u)$ and $\lambda_{_T} = \sigma_{_T}(u)$ at least in a weak sense. The main difficulty in the existence and uniqueness analysis of (1.8), (1.9), or (1.10) comes from the coupling between the friction threshold $\mathcal{F}\sigma_{_N}(u)$ and the contact pressure $\sigma_{_N}(u)$.

Remark 1. The equivalence between problems (1.8) and (1.9) is easy to obtain here since the assumption $f \in L^2(\Omega)^d$ implies that a generalized Green formula holds (see [31], for instance). The proof can also be made directly as follows. A solution to problem (1.9) is obviously a solution to problem (1.8). Conversely, if u is solution to problem (1.8), then the map $X \ni v \longmapsto a(u, \Pi(v)) - l(\Pi(v))$ is linear continuous for any continuous lifting operator $\Pi : X \to V$. Thus there exists $\lambda \in X'$ such that $\langle \lambda, v \rangle_{X',X} = a(u, \Pi(v)) - l(\Pi(v))$ for all $v \in X$. It is easy to state that in fact $\langle \lambda, v \rangle_{X',X} = a(u,v) - l(v)$ for all $v \in V$, proving $a(u, \Pi(v|_{\Gamma_{_C}}) - v) - l(\Pi(v|_{\Gamma_{_C}}) - v) = 0$ for all $v \in V$. Indeed, $\Pi(v|_{\Gamma_{_C}}) - v$ has a vanishing trace on $\Gamma_{_C}$, and replacing successively $v - u$ by $(\Pi(v|_{\Gamma_{_C}}) - v + u) - u$ and by $(v - \Pi(v|_{\Gamma_{_C}}) + u) - u$ in the inequality of (1.8) leads to this result. The two inequalities of (1.9) result then from the replacement of $a(u, v-u) - l(v-u)$ by $\langle \lambda, v-u \rangle_{X',X}$ in the inequality of (1.8), separating normal and tangential components ($\Gamma_{_C}$ is straight here), and remarking that applying a Green formula, one has $\sigma_{_N}(u) = \lambda_{_N}$. The equivalence between (1.9) and (1.10) is developed in [30] by computing the Fenchel conjugate of $j(\mathcal{F}\sigma_{_N}, \cdot)$ and inverting the normal cone to K.

1.2. Neumann to Dirichlet operator. We introduce the Neumann to Dirichlet operator on Γ_C and its basic properties. This will allow us to restrict the contact and friction problem on Γ_C and obtain useful estimates.

Let $\lambda = (\lambda_N, \lambda_T) \in X'$. The solution u to

$$\left\{ \begin{array}{l} \text{Find } u \in V \text{ satisfying} \\ a(u,v) = l(v) + \langle \lambda, v \rangle_{X',X} \quad \forall\ v \in V \end{array} \right. \tag{1.11}$$

is unique (see [17]). So it is possible to define the operator

$$\begin{aligned} \mathbb{E} : X' &\longrightarrow X \\ \lambda &\longmapsto u_{|_{\Gamma_C}}. \end{aligned}$$

It is easy to check that the operator $\mathbb{E}$ is affine and continuous. We define the following norms on Γ_C relative to $a(\cdot,\cdot)$:

$$\|v\|_{a,\Gamma_C} = \inf_{w \in V,\ w_{|_{\Gamma_C}} = v} \|w\|_a,$$

$$\|v_N\|_{a,\Gamma_C} = \inf_{\substack{w \in V, \\ w_N = v_N \text{ on } \Gamma_C}} \|w\|_a = \inf_{\substack{w \in V, \\ w_N = v_N \text{ on } \Gamma_C}} \|w\|_{a,\Gamma_C},$$

$$\|v_T\|_{a,\Gamma_C} = \inf_{\substack{w \in V, \\ w_T = v_T \text{ on } \Gamma_C}} \|w\|_a = \inf_{\substack{w \in V, \\ w_T = v_T \text{ on } \Gamma_C}} \|w\|_{a,\Gamma_C},$$

$$\|\lambda\|_{-a,\Gamma_C} = \sup_{\substack{v \in X \\ v \neq 0}} \frac{\langle \lambda, v \rangle_{X',X}}{\|v\|_{a,\Gamma_C}} = \sup_{\substack{v \in V \\ v \neq 0}} \frac{\langle \lambda, v \rangle_{X',X}}{\|v\|_a},$$

$$\|\lambda_N\|_{-a,\Gamma_C} = \sup_{\substack{v_N \in X_N \\ v_N \neq 0}} \frac{\langle \lambda_N, v_N \rangle_{X'_N, X_N}}{\|v_N\|_{a,\Gamma_C}} = \|(\lambda_N, 0)\|_{-a,\Gamma_C},$$

$$\|\lambda_T\|_{-a,\Gamma_C} = \sup_{\substack{v_T \in X_T \\ v_T \neq 0}} \frac{\langle \lambda_T, v_T \rangle_{X'_T, X_T}}{\|v_T\|_{a,\Gamma_C}} = \|(0, \lambda_T)\|_{-a,\Gamma_C},$$

which are equivalent, respectively, to the norms in X and X':

$$\frac{\sqrt{\alpha}}{C_1} \|v\|_X \leq \|v\|_{a,\Gamma_C} \leq \sqrt{M}\gamma \|v\|_X,$$

$$\frac{\sqrt{\alpha}}{C_1} \|v_N\|_{X_N} \leq \|v_N\|_{a,\Gamma_C} \leq \sqrt{M}\gamma \|v_N\|_{X_N},$$

$$\frac{\sqrt{\alpha}}{C_1} \|v_T\|_{X_T} \leq \|v_T\|_{a,\Gamma_C} \leq \sqrt{M}\gamma \|v_T\|_{X_T},$$

$$\frac{1}{\sqrt{M}\gamma}\|\lambda\|_{X'} \leq \|\lambda\|_{-a,\Gamma_C} \leq \frac{C_1}{\sqrt{\alpha}}\|\lambda\|_{X'}.$$

One also has

$$\|v_N\|_{a,\Gamma_C} \leq \|v\|_{a,\Gamma_C}$$

and

$$\|\lambda_N\|_{-a,\Gamma_C} \leq C_\alpha \|\lambda\|_{-a,\Gamma_C}$$

with a constant $C_\alpha \leq C_1\gamma\sqrt{M/\alpha}$. (But a better estimate should be possible: following [19], C_α is close to 1 when the Poisson ratio is close to 0.)

With the previous norms, it is possible to state (see [41]) the following equalities, when $u = \mathbb{E}(\lambda)$ and $\overline{u} = \mathbb{E}(\overline{\lambda})$ are the solutions to problem (1.11):

$$\|u - \overline{u}\|_a = \|\mathbb{E}(\lambda) - \mathbb{E}(\overline{\lambda})\|_{a,\Gamma_C} = \|\lambda - \overline{\lambda}\|_{-a,\Gamma_C}. \tag{1.12}$$

1.3. Direct weak inclusion formulation. Let

$$K_N = \{v_N \in X_N : v_N \leq 0 \text{ a.e. on } \Gamma_C\}$$

be the set of admissible normal displacements on Γ_C. The normal cone in X'_N to K_N at $v_N \in X_N$ is defined as

$$N_{K_N}(v_N) = \left\{\mu_N \in X'_N : \langle \mu_N, w_N - v_N\rangle_{X'_N,X_N} \leq 0 \ \forall w_N \in K_N\right\} \qquad \text{if } v_N \in K_N,$$

and $N_{K_N}(v_N) = \emptyset$ if $v_N \notin K_N$. The subdifferential of $j(\mathcal{F}\lambda_N, .)$ (i.e., with respect to the second variable) at u_T is given by

$$\begin{aligned}&\partial_2 j(\mathcal{F}\lambda_N, u_T)\\ &= \left\{\mu_T \in X'_T : j(\mathcal{F}\lambda_N, v_T) \geq j(\mathcal{F}\lambda_N, u_T) + \langle \mu_T, v_T - u_T\rangle_{X'_T,X_T} \ \ \forall v_T \in X_T\right\}.\end{aligned}$$

With this notation, problem (1.8) can be written

$$\begin{cases} \text{Find } u \in V, \lambda_N \in X'_N, \text{ and } \lambda_T \in X'_T \text{ satisfying}\\ (u_N, u_T) = \mathbb{E}(\lambda_N, \lambda_T),\\ -\lambda_N \in N_{K_N}(u_N) \quad \text{in } X'_N,\\ -\lambda_T \in \partial_2 j(\mathcal{F}\lambda_N, u_T) \quad \text{in } X'_T. \end{cases} \tag{1.13}$$

More details resulting from this equivalence can be found in [34].

1.4. A uniqueness criterion. In [26, 27] some multisolutions of the problem (1.1)–(1.7) are exhibited for triangular or quadrangular domains. These multiple solutions involve either an infinite set of slipping solutions or two isolated (stick and separation) configurations. Note that these examples of nonuniqueness involve large friction coefficients (i.e., $\mathcal{F} > 1$) and tangential displacements with a constant sign on Γ_C. Actually, it seems that no multisolution has been detected for an arbitrary

small friction coefficient in the continuous case, although such a result exists for finite element approximations in [25], but for a variable geometry. The forthcoming partial uniqueness result is obtained in [41]: it defines some cases where it is possible to affirm that a solution to the Coulomb friction problem is in fact the unique solution. More precisely, if a regular solution to the Coulomb friction problem exists (here the term regular means, roughly speaking, that the transition is smooth when the slip direction changes) and if the friction coefficient is small enough, then this solution is the only one. We recall the main useful tools leading to that result.

LEMMA 1.1. *Let u and $\overline{u}$ be two solutions to problem* (1.8)*, and let λ and $\overline{\lambda}$ be the corresponding contact stresses on Γ_C. Then the following estimate holds:*

$$\|u-\overline{u}\|_a^2 = \|\lambda-\overline{\lambda}\|_{-a,\Gamma_C}^2 \le \langle \zeta-\lambda_T, \overline{u}_T-u_T\rangle_{X'_T,X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\overline{\lambda}_N, u_T).$$

Proof. From (1.12), the Green formula, and (1.5), we get

$$\begin{aligned}\|u-\overline{u}\|_a^2 = \|\lambda-\overline{\lambda}\|_{-a,\Gamma_C}^2 &= \left\langle \overline{\lambda}_N-\lambda_N, \overline{u}_N-u_N\right\rangle_{X'_N,X_N} + \left\langle \overline{\lambda}_T-\lambda_T, \overline{u}_T-u_T\right\rangle_{X'_T,X_T}\\ &\le \left\langle \overline{\lambda}_T-\lambda_T, \overline{u}_T-u_T\right\rangle_{X'_T,X_T}.\end{aligned}$$

Thus

$$\|u-\overline{u}\|_a^2 \le \left\langle (\overline{\lambda}_T-\zeta)+(\zeta-\lambda_T), \overline{u}_T-u_T\right\rangle_{X'_T,X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\overline{\lambda}_N, u_T).$$

The conclusion follows from (1.13) and the fact that $-\partial_2 j(\mathcal{F}\overline{\lambda}_N, .)$ is a monotone set-valued mapping. ∎

We now introduce the space of multipliers $M(X_T \to X_N)$ of the functions $\xi : \Gamma_C \to R^d$ satisfying $\xi.\mathrm{n} = 0$ a.e. on Γ_C and such that the following equivalent norms are finite:

$$\|\xi\|_{M(X_T\to X_N)} = \sup_{\substack{v_T\in X_T\\ v_T\neq 0}} \frac{\|\xi.v_T\|_{X_N}}{\|v_T\|_{X_T}}, \quad \text{and} \quad \|\xi\|_a = \sup_{\substack{v_T\in X_T\\ v_T\neq 0}} \frac{\|\xi.v_T\|_{a,\Gamma_C}}{\|v_T\|_{a,\Gamma_C}}.$$

Since Γ_C is assumed to be straight, $M(X_T \to X_N)$ contains for any $\varepsilon > 0$ the space $H^{1/2+\varepsilon}(\Gamma_C)$ when $d = 2$. When $d = 3$, $M(X_T \to X_N)$ contains $H^1(\Gamma_C) \cap L^\infty(\Gamma_C)$. (See [38] for a complete discussion on the theory of multipliers in a pair of Hilbert spaces.)

The partial uniqueness result is given assuming that $\lambda_T = \mathcal{F}\lambda_N\xi$, with $\xi \in M(X_T \to X_N)$. The product $\lambda_N\xi$ has to be understood in the sense that $\langle \lambda_N\xi, v_T\rangle_{X'_T,X_T} = \langle \lambda_N, \xi.v_T\rangle_{X'_N,X_N}$ for all $v_T \in X_T$. It is easy to see that this implies $|\xi| \le 1$ a.e. on the support of λ_N. More precisely, this implies that $\xi \in \mathrm{Dir}_T(u_T)$ a.e. on the support of λ_N, where $\mathrm{Dir}_T(.)$ is the subdifferential of the convex map $R^d \ni x \longmapsto |x_T|$. This means that it is possible to assume that $\xi \in \mathrm{Dir}_T(u_T)$ a.e. on Γ_C.

PROPOSITION 1.2. *Let u be a solution to problem* (1.8) *such that $\lambda_T = \mathcal{F}\lambda_N\xi$, with $\xi \in M(X_T \to X_N)$, $\xi \in Dir_T(u_T)$ a.e. on Γ_C, and $\mathcal{F} < (C_\alpha\|\xi\|_a)^{-1}$. Then u is the unique solution to problem* (1.8).

Proof. Let $\overline{u}$ be another solution to problem (1.8), where $\overline{\lambda}_N$ and $\overline{\lambda}_T$ denote the corresponding contact stresses on Γ_C. According to Lemma 1.1, we write

$$\|u-\overline{u}\|_a^2 \le \langle \zeta-\lambda_T, \overline{u}_T-u_T\rangle_{X'_T,X_T} \qquad \forall \zeta \in -\partial_2 j(\mathcal{F}\overline{\lambda}_N, u_T).$$

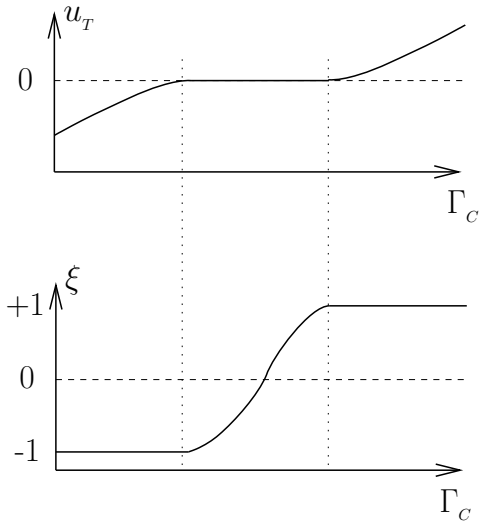


FIG. 1.2. *Example of a tangential displacement u_T and a possible corresponding multiplier ξ when $d = 2$.*

It is easy to see that a possible choice is $\zeta = \mathcal{F}\overline{\lambda}_N\xi$. Therefore

$$\begin{aligned}\|u-\overline{u}\|_a^2 \le \langle \mathcal{F}\xi(\overline{\lambda}_N-\lambda_N), \overline{u}_T-u_T\rangle_{X'_T,X_T} &\le \mathcal{F}\|\lambda_N-\overline{\lambda}_N\|_{-a,\Gamma_C}\|\xi.(u_T-\overline{u}_T)\|_{a,\Gamma_C}\\ &\le C_\alpha\mathcal{F}\|\xi\|_a\|\lambda-\overline{\lambda}\|_{-a,\Gamma_C}\|u-\overline{u}\|_a\\ &= C_\alpha\mathcal{F}\|\xi\|_a\|u-\overline{u}\|_a^2,\end{aligned}$$

which implies that $\overline{u} = u$ when $\mathcal{F} < (C_\alpha\|\xi\|_a)^{-1}$. □

In two space dimensions ($d = 2$), the case $\xi \equiv 1$ corresponds to a homogeneous sliding direction, and the previous result is complementary to the nonuniqueness results obtained in [26, 27].

As illustrated in Figure 1.2, when $d = 2$ the multiplier ξ has to vary from -1 to $+1$ each time the sign of the tangential displacement changes from negative to positive. The set $M(X_T \to X_N)$ does not contain any multiplier having a singularity of the first kind. Consequently, in order to satisfy the assumptions of Proposition 1.2, the tangential displacement of the solution u cannot pass from a negative value to a positive value and be zero at only a single point of Γ_C.

Remark 2. This remark deals with a more precise discussion concerning the assumption $\lambda_T = \mathcal{F}\lambda_N\xi$, $\xi \in M(X_T \to X_N)$, $\xi \in \mathrm{Dir}_T(u_T)$ and the cases where the assumption cannot be fulfilled independently of the regularity of the solution when $d = 2$. On the one hand, it is easy to show that the choice of ξ is unique at any point where $\lambda_N \neq 0$ or $u_T \neq 0$. In the first case $\xi = \lambda_T/(\mathcal{F}\lambda_N)$, in the second case $\xi = u_T/|u_T|$, and both expressions coincide when $\lambda_N \neq 0$ and $u_T \neq 0$. On the other hand, any $\xi \in [-1, 1]$ can be chosen at the points where $\lambda_N = u_T = 0$. So it remains to determine when ξ lies in $M(X_T \to X_N)$. If there are no points such that $\lambda_N = u_T = 0$ on Γ_C, then the condition $\xi \in M(X_T \to X_N)$ is linked to the regularity of u (in other words, if u is regular enough, then $\xi \in M(X_T \to X_N)$). If there are some points such that $\lambda_N = u_T = 0$, then it is easy to show that the continuity of ξ can be lost (whatever the regularity of u is) only if some of these points are isolated. A discussion

shows then that $\xi \notin M(X_T \to X_N)$ in three cases. The first one is when u_T passes from a negative to a positive value at such a point (note that this could also occur at a point which is separated from the foundation). The second case corresponds to a stick area surrounding such an isolated point and where the right and left limits of $\lambda_T/(\mathcal{F}\lambda_N)$ differ at this point (where $\lambda_T = \lambda_N = u_T = u_N = 0$). The third case is a combination of both previous cases: a side where $u_T \neq 0$, the other one with $u_T = 0$ and $\lambda_N \neq 0$, and a limit of $\lambda_T/(\mathcal{F}\lambda_N)$ which differs from $u_T/|u_T|$. If the solution is less regular, then other cases of nonfulfillment could appear, but we think that this assumption (which is needed to obtain the uniqueness of a solution to the continuous problem) takes into account many frictional contact configurations.

2. Finite element approximation. Let $V^h \subset V$ be a family of finite dimensional vector spaces indexed by h coming from a regular family $\mathcal{T}^h$ (see [9]) of triangulations of the domain Ω (h represents the largest diameter among all elements). We choose standard continuous and piecewise affine functions, i.e.,

$$V^h = \left\{ v^h \in (\mathcal{C}(\overline{\Omega}))^d, v^h{}_{|_T} \in P_1(T) \quad \forall T \in \mathcal{T}^h, v^h = 0 \text{ on } \Gamma_D \right\}. \tag{2.1}$$

Define

$$X_N^h = \left\{ v_N^h{}_{|_{\Gamma_C}} : v^h \in V^h \right\},$$

$$X_T^h = \left\{ v_T^h{}_{|_{\Gamma_C}} : v^h \in V^h \right\},$$

$$X^h = \left\{ v^h{}_{|_{\Gamma_C}} : v^h \in V^h \right\} = X_N^h \times X_T^h. \tag{2.2}$$

Identifying X_N^h and X_T^h with their dual spaces using the L^2 scalar product, we consider that X_N^h and X_T^h are also the finite-dimensional approximations of X_N' and X_T', respectively.

The finite element discretization of problem (1.10) becomes

$$\begin{cases} \text{Find } u^h \in V^h, \lambda_N^h \in X_N^h, \text{ and } \lambda_T^h \in X_T^h \text{ satisfying} \\ a(u^h, v^h) = l(v^h) + \displaystyle\int_{\Gamma_C} \lambda_N^h v_N^h \, d\Gamma + \int_{\Gamma_C} \lambda_T^h . v_T^h \, d\Gamma \quad \forall v^h \in V^h, \\ \lambda_N^h \in \Lambda_N^h, \quad \displaystyle\int_{\Gamma_C} (\mu_N^h - \lambda_N^h) u_N^h \, d\Gamma \geq 0 \quad \forall \mu_N^h \in \Lambda_N^h, \\ \lambda_T^h \in \Lambda_T^h(\mathcal{F}\lambda_N^h), \quad \displaystyle\int_{\Gamma_C} (\mu_T^h - \lambda_T^h) . u_T^h d\Gamma \geq 0 \quad \forall \mu_T^h \in \Lambda_T^h(\mathcal{F}\lambda_N^h), \end{cases} \tag{2.3}$$

where the approximations of Λ_N and $\Lambda_T(\mathcal{F}\lambda_N)$ have been chosen in the following way:

$$\Lambda_N^h = \Lambda_N \cap X_N^h, \tag{2.4}$$

$$\Lambda_T^h(\mathcal{F}\lambda_N^h) = \left\{ \lambda_T^h \in X_T^h : \int_{\Gamma_C} \lambda_T^h . v_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, v_T^h) \geq 0 \quad \forall v_T^h \in X_T^h \right\}. \tag{2.5}$$

To simplify our discussion, we assume afterwards that the mesh inherited on the contact zone is quasi-uniform (although there exist some less restrictive assumptions; see, e.g., [13]) of size h (to simplify). Another simplification is that we restrict ourselves to the two-dimensional case ($d = 2$), and we assume that the end points of Γ_C do not belong to $\overline{\Gamma_D}$ (in other words, $\overline{\Gamma_C} \cap \overline{\Gamma_D} = \emptyset$). More general cases will be discussed in some remarks at the end of the paper.

With this choice of discretization, the following discrete Babuška–Brezzi inf-sup condition holds (see, e.g., [12, 6]):

$$\inf_{\lambda^h \in X^h} \sup_{v^h \in V^h} \frac{\int_{\Gamma_C} \lambda^h . v^h d\Gamma}{\|v^h\|_a \|\lambda^h\|_{-a,\Gamma_C}} \geq c_{is} > 0, \tag{2.6}$$

where $c_{is} \leq 1$ is independent of h. As a consequence, problem (2.3) admits a solution for any friction coefficient, and the solution is unique for a sufficiently small friction coefficient (where the label "small" may depend on h) (see [30]).

The following lemma shows the relation between the hybrid formulation and the direct formulation of the friction condition in the discrete framework.

LEMMA 2.1. *For $\lambda_N^h \in \Lambda_N^h$, a pair $(\lambda_T^h, u_T^h) \in X_T^h \times X_T^h$ satisfies*

$$\lambda_T^h \in \Lambda_T^h(\mathcal{F}\lambda_N^h), \quad \int_{\Gamma_C} (\mu_T^h - \lambda_T^h).u_T^h d\Gamma \geq 0 \quad \forall \mu_T^h \in \Lambda_T^h(\mathcal{F}\lambda_N^h) \tag{2.7}$$

if and only if the pair satisfies

$$\int_{\Gamma_C} \lambda_T^h.(v_T^h - u_T^h)d\Gamma + j(\mathcal{F}\lambda_N^h, v_T^h) - j(\mathcal{F}\lambda_N^h, u_T^h) \geq 0 \quad \forall v_T^h \in X_T^h. \tag{2.8}$$

Proof. Let us first assume that (λ_T^h, u_T^h) satisfies (2.7). For an arbitrary choice $\xi \in \mathcal{F}\lambda_N^h \mathrm{Dir}_T(u_T^h)$ the map $v_T^h \longmapsto \int_{\Gamma_C} \xi.v_T^h d\Gamma$ is a linear form on X_T^h, and thus by the Riesz representation theorem there exists $\mu_T^h \in X_T^h$ such that $\int_{\Gamma_C} \mu_T^h.v_T^h d\Gamma = \int_{\Gamma_C} \xi.v_T^h d\Gamma$ for all $v_T^h \in X_T^h$. This μ_T^h satisfies $\int_{\Gamma_C} \mu_T^h.u_T^h d\Gamma = \int_{\Gamma_C} \mathcal{F}\lambda_N^h |u_T^h| d\Gamma = -j(\mathcal{F}\lambda_N^h, u_T^h)$, and μ_T^h is an element of $\Lambda_T^h(\mathcal{F}\lambda_N^h)$. Now considering this particular μ_T^h in (2.7) leads to

$$\int_{\Gamma_C} \lambda_T^h.u_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, u_T^h) \leq 0.$$

Together with the fact that λ_T^h is in $\Lambda_T^h(\mathcal{F}\lambda_N^h)$ this leads to the complementarity relation

$$\int_{\Gamma_C} \lambda_T^h.u_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, u_T^h) = 0,$$

which straightforwardly implies (2.8).

Conversely, let us assume that (λ_T^h, u_T^h) satisfies (2.8). Then choosing $v_T^h = 0$ in (2.8) gives $-\int_{\Gamma_C} \lambda_T^h.u_T^h d\Gamma - j(\mathcal{F}\lambda_N^h, u_T^h) \geq 0$, and choosing $v_T^h = 2u_T^h$ gives $\int_{\Gamma_C} \lambda_T^h.u_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, u_T^h) \geq 0$, which implies the complementarity relation

$$\int_{\Gamma_C} \lambda_T^h.u_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, u_T^h) = 0.$$

Taking this into account in (2.8) leads to

$$\int_{\Gamma_C} \lambda_T^h . v_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, v_T^h) \geq 0 \qquad \forall v_T^h \in X_T^h,$$

which implies $\lambda_T^h \in \Lambda_T^h(\mathcal{F}\lambda_N^h)$. Now, from the complementarity relation and for all $\mu_T^h \in \Lambda_T^h(\mathcal{F}\lambda_N^h)$ one has

$$\int_{\Gamma_C} (\mu_T^h - \lambda_T^h).u_T^h d\Gamma = \int_{\Gamma_C} \mu_T^h.u_T^h d\Gamma + j(\mathcal{F}\lambda_N^h, u_T^h) \geq 0,$$

which implies (2.7). □

Remark 3. The equivalence given by this lemma is a classical result when it deals with the continuous problem. With the particular finite element discretization considered in this section, the result is still valid in the finite-dimensional case. One of the reasons is that the space of the multipliers has been chosen in such a way that it can represent the dual space of the discrete trace space X_T^h. But this result does not remain valid when a smaller space for the multipliers is chosen.

3. The error estimate.

THEOREM 3.1. *Let (u, λ) be the solution to problem* (1.8) *(for $d = 2$) such that $\lambda_T = \mathcal{F}\lambda_N\xi$, with $\xi \in M(X_T \to X_N)$, $\xi \in Dir_T(u_T)$ a.e. on Γ_C, and $\mathcal{F} < c_{is}(C_\alpha\|\xi\|_a)^{-1}$. Assume that $u \in (H^{(3/2)+\varepsilon}(\Omega))^2$ with $\varepsilon > 0$, and let (u^h, λ^h) be a solution to the discrete problem* (2.3). *Then there exists a constant $C > 0$ independent of h and u such that*

$$\|u - u^h\|_a + \|\lambda - \lambda^h\|_{-a,\Gamma_C} \leq Ch^{1/2}\|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}. \tag{3.1}$$

Proof. Let $v^h \in V^h$. Then

$$\begin{aligned}
\|u - u^h\|_a^2 &= a(u - u^h, u - u^h) \\
&= a(u - u^h, u - v^h) + a(u, v^h - u^h) - a(u^h, v^h - u^h) \\
&= a(u - u^h, u - v^h) + \int_{\Gamma_C} \lambda_N (v_N^h - u_N^h)\, d\Gamma + \int_{\Gamma_C} \lambda_T.(v_T^h - u_T^h)\, d\Gamma \\
&\quad - \int_{\Gamma_C} \lambda_N^h (v_N^h - u_N^h)\, d\Gamma - \int_{\Gamma_C} \lambda_T^h.(v_T^h - u_T^h)\, d\Gamma \\
&= a(u - u^h, u - v^h) \\
&\quad + \int_{\Gamma_C} (\lambda_N - \lambda_N^h)(v_N^h - u_N)\, d\Gamma + \int_{\Gamma_C} (\lambda_T - \lambda_T^h).(v_T^h - u_T)\, d\Gamma \\
&\quad + \int_{\Gamma_C} (\lambda_N - \lambda_N^h)(u_N - u_N^h)\, d\Gamma + \int_{\Gamma_C} (\lambda_T - \lambda_T^h).(u_T - u_T^h)\, d\Gamma.
\end{aligned}$$

The continuous and discrete complementary conditions imply

$$\int_{\Gamma_C} \lambda_N u_N \, d\Gamma = \int_{\Gamma_C} \lambda_N^h u_N^h \, d\Gamma = 0.$$

Hence

$$\begin{aligned}\|u-u^h\|_a^2 = a(u-u^h,u-v^h)&\\ +\int_{\Gamma_C}(\lambda_N-\lambda_N^h)(v_N^h-u_N)\,d\Gamma+\int_{\Gamma_C}(\lambda_T-\lambda_T^h).(v_T^h-u_T)\,d\Gamma&\\ -\int_{\Gamma_C}\lambda_N u_N^h+\lambda_N^h u_N\,d\Gamma+\int_{\Gamma_C}(\lambda_T-\lambda_T^h).(u_T-u_T^h)\,d\Gamma.&\end{aligned}$$

Using the continuity of the bilinear form, we obtain

$$\begin{aligned}\|u-u^h\|_a^2 \le \|u-u^h\|_a\|u-v^h\|_a+\|\lambda-\lambda^h\|_{-a,\Gamma_C}\|u-v^h\|_{a,\Gamma_C}&\\ -\int_{\Gamma_C}\lambda_N u_N^h+\lambda_N^h u_N\,d\Gamma+\int_{\Gamma_C}(\lambda_T-\lambda_T^h).(u_T-u_T^h)\,d\Gamma.&\end{aligned}\tag{3.2}$$

Additionally, we consider the equilibrium equation. From $V^h\subset V$, we get

$$a(u,v^h)=l(v^h)+\int_{\Gamma_C}\lambda.v^h\,d\Gamma\qquad\forall\ v^h\in V^h.$$

Since

$$a(u^h,v^h)=l(v^h)+\int_{\Gamma_C}\lambda^h.v^h\,d\Gamma\qquad\forall\ v^h\in V^h,$$

we deduce by subtraction that

$$a(u-u^h,v^h)=\int_{\Gamma_C}(\lambda-\lambda^h).v^h\,d\Gamma\qquad\forall\ v^h\in V^h.$$

Consequently, for any $v^h\in V^h$ and any $\mu^h\in X^h$

$$\begin{aligned}\int_{\Gamma_C}(\lambda^h-\mu^h).v^h\,d\Gamma = a(u^h-u,v^h)+\int_{\Gamma_C}(\lambda-\mu^h).v^h\,d\Gamma&\\ \le(\|u-u^h\|_a+\|\lambda-\mu^h\|_{-a,\Gamma_C})\|v^h\|_a.&\end{aligned}$$

The mesh independent inf-sup condition (2.6) implies, for any $\mu^h\in X^h$,

$$c_{is}\|\lambda^h-\mu^h\|_{-a,\Gamma_C}\le\sup_{v^h\in V^h}\frac{\int_{\Gamma_C}(\lambda^h-\mu^h).v^h\,d\Gamma}{\|v^h\|_a}\le\|u-u^h\|_a+\|\lambda-\mu^h\|_{-a,\Gamma_C}.$$

By the triangular inequality we come to the conclusion that

$$\|\lambda-\lambda^h\|_{-a,\Gamma_C}\le\frac{1}{c_{is}}\|u-u^h\|_a+\left(1+\frac{1}{c_{is}}\right)\inf_{\mu^h\in X^h}\|\lambda-\mu^h\|_{-a,\Gamma_C}.\tag{3.3}$$

Keeping in mind that $u\in(H^{(3/2)+\varepsilon}(\Omega))^2$ with $\varepsilon>0$, so that $\lambda\in(L^2(\Gamma_C))^2$ according to the trace theorem, we choose $v^h=I^hu$, where I^h denotes the Lagrange interpolation operator mapping onto V^h, and $\mu^h=\pi^h\lambda$, where π^h represents the $(L^2(\Gamma_C))^2$-projection operator mapping onto X^h. As a consequence (see [7, 9, 13]) if $\varepsilon>0$ is small enough, we have

$$\inf_{v^h\in V^h}\|u-v^h\|_a\le Ch^{(1/2)+\varepsilon}\|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}\tag{3.4}$$

and

$$\inf_{\mu^h \in X^h} \|\lambda - \mu^h\|_{-a,\Gamma_C} \leq Ch^{(1/2)+\varepsilon} \|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}, \tag{3.5}$$

where C denotes here and afterwards a positive constant independent of h. We now estimate the terms in (3.2) coming from the contact approximation. Since $u_N \leq 0$ and $\lambda_N^h \leq 0$ on Γ_C, we deduce that the first term is nonpositive:

$$-\int_{\Gamma_C} \lambda_N^h u_N \, d\Gamma \leq 0. \tag{3.6}$$

In order to estimate the second term in (3.2) coming from the contact approximation we introduce a specific operator. Namely, $r^h : L^1(\Gamma_C) \mapsto X_N^h$ is the quasi-interpolation operator defined for any function v in $L^1(\Gamma_C)$ by

$$r^h v = \sum_{x \in N^h} \alpha_x(v) \psi_x,$$

where N^h represents the set of nodes of $\overline{\Gamma_C}$, ψ_x is the scalar basis function of X_N^h (defined on $\overline{\Gamma_C}$) at node x verifying $\psi_x(x') = \delta_{x,x'}$ for all $x' \in N^h$, and

$$\alpha_x(v) = \left(\int_{\Gamma_C} v \psi_x \, d\Gamma \right) \left(\int_{\Gamma_C} \psi_x \, d\Gamma \right)^{-1}.$$

Remark 4. It is straightforward to check that r^h is linear and that it preserves nonpositivity. It is also obvious that $r^h v^h \neq v^h$ when $v^h \in X_N^h$. This operator is different from Clément's (which consists of making local projections onto P_1 functions; see [10]), from Chen–Nochetto's (which uses local projections onto P_0 functions; see [8]), and from Ben Belgacem–Renard's (which consists of making local projections onto the convex cone of nonpositive P_1 functions; see [6]). The main particularity of the operator r^h, which directly follows from its definition, is that $r^h v^h \leq 0$ when $v^h \in X_N^h$ satisfies only "weak nonpositivity conditions"; i.e.,

$$\int_{\Gamma_C} \mu_N^h v^h d\Gamma \geq 0 \qquad \forall \mu_N^h \in \Lambda_N^h. \tag{3.7}$$

This property is not satisfied by the operators in [8] and [10]. Moreover, as we see hereafter, the approximation properties of r^h hold for any function without sign condition, contrary to the operator in [6].

The approximation properties of r^h are established in [28]. We recall them to render the proof of Theorem 3.1 self-contained. We first show the L^2-stability property of r^h.

LEMMA 3.2. *There is a positive constant C independent of h such that for any $v \in L^2(\Gamma_C)$ and any $E \in E_C^h$ (E_C^h denotes the set of closed edges lying in $\overline{\Gamma_C}$)*

$$\|r^h v\|_{L^2(E)} \leq C \|v\|_{L^2(\gamma_E)},$$

where $\gamma_E = \cup_{\{F \in E_C^h :\, F \cap E \neq \emptyset\}} F$.

Proof. Let γ_x be the support of the basis function ψ_x in Γ_C. Using the definition of $\alpha_x(v)$, the Cauchy–Schwarz inequality, and the uniform regularity of the mesh, we get

$$|\alpha_x(v)| \leq \|v\|_{L^2(\gamma_x)} \|\psi_x\|_{L^2(\gamma_x)} \|\psi_x\|_{L^1(\gamma_x)}^{-1} \leq Ch^{-\frac{1}{2}} \|v\|_{L^2(\gamma_x)}.$$

We obtain by a triangular inequality

$$\|r^h v\|_{L^2(E)} = \left\| \sum_{x \in N^h \cap E} \alpha_x(v) \psi_x \right\|_{L^2(E)} \leq C \|v\|_{L^2(\gamma_E)}. \qquad \square$$

The next lemma is concerned with the L^2-approximation properties of r^h.

LEMMA 3.3. *There is a positive constant C independent of h such that for any $v \in H^\eta(\Gamma_C)$, $0 \leq \eta \leq 1$, and any $E \in E_C^h$ (E_C^h denotes the set of closed edges lying in $\overline{\Gamma_C}$)*

$$\|v - r^h v\|_{L^2(E)} \leq C h^\eta \|v\|_{H^\eta(\gamma_E)}, \tag{3.8}$$

where $\gamma_E = \cup_{\{F \in E_C^h :\, F \cap E \neq \emptyset\}} F$.

Proof. When $\eta = 0$ the bound results from the previous lemma. Note that r^h preserves the constant functions on Γ_C. Let be given an arbitrary constant function $c(x) = c$, for all $x \in \Gamma_C$. From the definition of r^h, we may write, for any $v \in H^\eta(\Gamma_C)$,

$$v - r^h v = v - c - r^h(v - c).$$

Therefore by Lemma 3.2 we get

$$\|v - r^h v\|_{L^2(E)} \leq C \left(\|v - c\|_{L^2(E)} + \|v - c\|_{L^2(\gamma_E)} \right) \leq C \|v - c\|_{L^2(\gamma_E)} \quad \forall c \in R. \tag{3.9}$$

We then choose $c = \int_{\gamma_E} v(x)\, dx / |\gamma_E|$ in (3.9), where $|\gamma_E|$ denotes the length of γ_E. Then if $x \in \gamma_E$ and $0 < \eta < 1$, we have

$$\begin{aligned} v(x) - c &= |\gamma_E|^{-1} \int_{\gamma_E} v(x) - v(y)\, dy \\ &= |\gamma_E|^{-1} \int_{\gamma_E} \frac{v(x) - v(y)}{|x - y|^{\frac{1+2\eta}{2}}} |x - y|^{\frac{1+2\eta}{2}} dy. \end{aligned}$$

Using the Cauchy–Schwarz inequality, we deduce

$$\begin{aligned} \int_{\gamma_E} (v(x) - c)^2 dx &= |\gamma_E|^{-2} \int_{\gamma_E} \left(\int_{\gamma_E} \frac{v(x) - v(y)}{|x - y|^{\frac{1+2\eta}{2}}} |x - y|^{\frac{1+2\eta}{2}} dy \right)^2 dx \\ &\leq |\gamma_E|^{-2} \int_{\gamma_E} \left(\int_{\gamma_E} \frac{(v(x) - v(y))^2}{|x - y|^{1+2\eta}} dy \int_{\gamma_E} |x - y|^{1+2\eta} dy \right) dx \\ &\leq |\gamma_E|^{2\eta} \int_{\gamma_E} \int_{\gamma_E} \frac{(v(x) - v(y))^2}{|x - y|^{1+2\eta}}\, dy dx \\ &\leq C h^{2\eta} \|v\|^2_{H^\eta(\gamma_E)}. \end{aligned}$$

Hence the result.

If $x \in \gamma_E$ and $\eta = 1$, we have

$$v(x) - c = |\gamma_E|^{-1} \int_{\gamma_E} v(x) - v(y)\, dy = |\gamma_E|^{-1} \int_{\gamma_E} \int_y^x v'(t)\, dt dy.$$

Hence

$$|v(x) - c| \leq |\gamma_E|^{\frac{1}{2}} \|v'\|_{L^2(\gamma_E)}.$$

The result is then straightforward. ∎

End of the proof of Theorem 3.1. The second term coming from the contact approximation in (3.2) is handled as follows:

$$-\int_{\Gamma_C} \lambda_N u_N^h \, d\Gamma = -\int_{\Gamma_C} \lambda_N (u_N^h - r^h u_N^h) \, d\Gamma - \int_{\Gamma_C} \lambda_N \, r^h u_N^h \, d\Gamma.$$

According to (2.3), u_N^h satisfies a weak nonnegativity condition as in (3.7). From Remark 4 we deduce that $r^h u_N^h \leq 0$. Hence we have, for any small $\varepsilon > 0$,

$$\begin{aligned}
-\int_{\Gamma_C} \lambda_N u_N^h \, d\Gamma &\leq -\int_{\Gamma_C} \lambda_N (u_N^h - r^h u_N^h) \, d\Gamma \\
&\leq \|\lambda_N\|_{L^2(\Gamma_C)} \|u_N^h - r^h u_N^h\|_{L^2(\Gamma_C)} \\
&\leq \|\lambda_N\|_{L^2(\Gamma_C)} \|(u_N^h - u_N) - r^h(u_N^h - u_N)\|_{L^2(\Gamma_C)} \\
&\quad + \|\lambda_N\|_{L^2(\Gamma_C)} \|u_N - r^h u_N\|_{L^2(\Gamma_C)} \\
&\leq Ch^{1/2} \|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2} \|u_N^h - u_N\|_{H^{1/2}(\Gamma_C)} \\
&\quad + Ch \|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2} \|u_N\|_{H^1(\Gamma_C)}, \qquad (3.10)
\end{aligned}$$

where the trace theorem $\|\lambda_N\|_{L^2(\Gamma_C)} \leq C\|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}$ (see [20]) and the estimates in Lemma 3.3 have been used. Putting together estimates (3.6) and (3.10) yields for any small $\varepsilon > 0$

$$\begin{aligned}
&-\int_{\Gamma_C} \lambda_N^h u_N + \lambda_N u_N^h \, d\Gamma \leq Ch^{1/2} \|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2} \\
&\left(\|u - u^h\|_a + h^{1/2} \|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2} \right). \qquad (3.11)
\end{aligned}$$

We now estimate the terms corresponding to the friction approximation in (3.2). From the assumptions in the theorem we write

$$\begin{aligned}
\int_{\Gamma_C} (\lambda_T - \lambda_T^h).(u_T - u_T^h) \, d\Gamma &= \int_{\Gamma_C} (\lambda_T - \mathcal{F}\lambda_N^h \xi).(u_T - u_T^h) \, d\Gamma \\
&\quad + \int_{\Gamma_C} (\mathcal{F}\lambda_N^h \xi - \lambda_T^h).(u_T - u_T^h) \, d\Gamma \\
&= \int_{\Gamma_C} \mathcal{F}(\lambda_N - \lambda_N^h)\xi.(u_T - u_T^h) \, d\Gamma \\
&\quad + \int_{\Gamma_C} (\mathcal{F}\lambda_N^h \xi - \lambda_T^h).(u_T - u_T^h) \, d\Gamma. \qquad (3.12)
\end{aligned}$$

The estimate of the first integral term in (3.12) gives

$$\int_{\Gamma_C} \mathcal{F}(\lambda_N - \lambda_N^h)\xi.(u_T - u_T^h) \, d\Gamma \leq C_\alpha \mathcal{F} \|\xi\|_a \|\lambda - \lambda^h\|_{-a,\Gamma_C} \|u - u^h\|_a.$$

The second integral term in (3.12) is written as follows:

$$\begin{aligned}
\int_{\Gamma_C} (\mathcal{F}\lambda_N^h \xi - \lambda_T^h).(u_T - u_T^h) \, d\Gamma &= \int_{\Gamma_C} \mathcal{F}\lambda_N^h \xi.(u_T - u_T^h) \, d\Gamma - \int_{\Gamma_C} \lambda_T^h.u_T \, d\Gamma \\
&\quad + \int_{\Gamma_C} \lambda_T^h.u_T^h \, d\Gamma.
\end{aligned}$$

Using the equivalent discrete friction conditions in (2.8), we obtain for any $v^h \in V^h$

$$\int_{\Gamma_C} (\mathcal{F}\lambda_N^h \xi - \lambda_T^h).(u_T - u_T^h)\,d\Gamma \le \int_{\Gamma_C} \mathcal{F}\lambda_N^h \xi.(u_T - u_T^h)\,d\Gamma - \int_{\Gamma_C} \lambda_T^h.u_T\,d\Gamma + \int_{\Gamma_C} \lambda_T^h.v_T^h\,d\Gamma - \int_{\Gamma_C} \mathcal{F}\lambda_N^h |v_T^h|\,d\Gamma + \int_{\Gamma_C} \mathcal{F}\lambda_N^h |u_T^h|\,d\Gamma.$$

Choosing $v_T^h = I^h u_T$ and since $\xi.u_T = |u_T|$, we obtain

$$\int_{\Gamma_C} (\mathcal{F}\lambda_N^h \xi - \lambda_T^h).(u_T - u_T^h)\,d\Gamma \le \int_{\Gamma_C} \lambda_T^h.(I^h u_T - u_T)\,d\Gamma + \int_{\Gamma_C} \mathcal{F}\lambda_N^h (|u_T^h| - \xi.u_T^h)\,d\Gamma + \int_{\Gamma_C} \mathcal{F}\lambda_N^h (|u_T| - |I^h u_T|)\,d\Gamma. \tag{3.13}$$

The estimate of the first term in (3.13) is achieved as follows by using the error estimates in [9]:

$$\int_{\Gamma_C} \lambda_T^h.(I^h u_T - u_T)\,d\Gamma = \int_{\Gamma_C} (\lambda_T^h - \lambda_T).(I^h u_T - u_T)\,d\Gamma + \int_{\Gamma_C} \lambda_T.(I^h u_T - u_T)\,d\Gamma \le Ch^{1/2}\|u\|_{(H^{3/2}(\Omega))^2}\Big(\|\lambda - \lambda^h\|_{-a,\Gamma_C} + h^{1/2}\|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}\Big).$$

The estimate of the second term in (3.13) uses the fact that $\lambda_N^h \le 0$ and $|u_T^h| - \xi.u_T^h \ge 0$ so that

$$\int_{\Gamma_C} \mathcal{F}\lambda_N^h (|u_T^h| - \xi.u_T^h)\,d\Gamma \le 0.$$

Finally, the third term in (3.13) yields

$$\int_{\Gamma_C} \mathcal{F}\lambda_N^h (|u_T| - |I^h u_T|)\,d\Gamma \le \mathcal{F}\|\lambda_N^h\|_{L^2(\Gamma_C)}\||u_T| - |I^h u_T|\|_{L^2(\Gamma_C)} \le \mathcal{F}\|\lambda_N^h\|_{L^2(\Gamma_C)}\|u_T - I^h u_T\|_{L^2(\Gamma_C)} \le C\mathcal{F}\|\lambda_N^h\|_{L^2(\Gamma_C)} h\|u\|_{(H^{3/2}(\Omega))^2}.$$

Further, using the (global) $L^2(\Gamma_C)$-projection operator π_N^h onto X_N^h (the notation π^h stands for the $(L^2(\Gamma_C))^2$-projection operator onto X^h) and an inverse inequality (see, e.g., [7, 9, 13]), we write

$$\begin{aligned}\|\lambda_N^h\|_{L^2(\Gamma_C)} &\le \|\lambda_N^h - \pi_N^h\lambda_N\|_{L^2(\Gamma_C)} + \|\pi_N^h\lambda_N - \lambda_N\|_{L^2(\Gamma_C)} + \|\lambda_N\|_{L^2(\Gamma_C)} \\ &\le C\left(h^{-1/2}\|\lambda^h - \pi^h\lambda\|_{-a,\Gamma_C} + \|\lambda_N\|_{L^2(\Gamma_C)}\right) \\ &\le C\left(h^{-1/2}\|\lambda - \lambda^h\|_{-a,\Gamma_C} + \|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}\right).\end{aligned}$$

Therefore,

$$\int_{\Gamma_C} \mathcal{F}\lambda_N^h (|u_T| - |I^h u_T|)\,d\Gamma \le C\mathcal{F}h^{1/2}\|u\|_{(H^{3/2}(\Omega))^2} \left(\|\lambda - \lambda^h\|_{-a,\Gamma_C} + h^{1/2}\|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}\right).$$

We come to the conclusion that the term dealing with the friction approximation in (3.2) is bounded as follows:

$$\begin{aligned}\int_{\Gamma_C} (\lambda_T - \lambda_T^h).(u_T - u_T^h)\, d\Gamma \le{}& C_\alpha \mathcal{F}\|\xi\|_a \|\lambda - \lambda^h\|_{-a,\Gamma_C} \|u - u^h\|_a \\ &+ C(1+\mathcal{F})h\|u\|^2_{(H^{(3/2)+\varepsilon}(\Omega))^2} \\ &+ C(1+\mathcal{F})h^{1/2}\|u\|_{(H^{3/2}(\Omega))^2}\|\lambda - \lambda^h\|_{-a,\Gamma_C}. \end{aligned} \tag{3.14}$$

Finally, the result is obtained by using (3.4)–(3.5) and putting together (3.2), (3.3), (3.11), and (3.14). ∎

Remark 5. The quasi-optimal rate of convergence of order 1/2 in the theorem does not depend on $\varepsilon > 0$. Actually we are not able to obtain a better convergence rate even if ε increases. A bit more regularity than $H^{3/2}$ is needed to apply the trace theorem, when writing $\|\lambda_N\|_{L^2(\Gamma_C)} \le C\|u\|_{(H^{(3/2)+\varepsilon}(\Omega))^2}$ (see [20]). The choice of the regularity assumptions $u \in (H^{(3/2)+\varepsilon}(\Omega))^2$ in the numerical analysis of contact problems is discussed in [4, Remark 2.4(i)] and [5, Remark 4.4]. If u is less regular than $H^{3/2}$, then the normal and tangential constraints cannot be expressed pointwise, and the frictional contact conditions cannot be simply written as in (1.5)–(1.7). In the frictionless case, when $u \in (H^\nu(\Omega))^2$ with $1 < \nu < 3/2$, the error analysis of a finite element approximation is achieved in [4]. Actually we are not able to extend these results to the frictional case.

Remark 6. If one (or both) end points of $\overline{\Gamma_C} = [x_0, x_n]$ is subjected to Dirichlet conditions, then the previous study can be extended with some modifications. Suppose, for instance, that $\overline{\Gamma_C} \cap \overline{\Gamma_D} = \{x_0\}$ and that the definition of V^h in (2.1) remains unchanged. If we still keep the same definition of X^h as in (2.2), then the estimate (3.5), does not hold in the general case. Thus we use a mortar approach introduced in [7]: denoting by $x_i, 0 \le i \le n$, the nodes on $\overline{\Gamma_C}$, we set

$$X_N^h = \left\{ \mu_N^h \in \mathcal{C}(\overline{\Gamma_C}), \mu_N^h{}_{|_{[x_i,x_{i+1}]}} \in P_1([x_i, x_{i+1}])\ \forall 1 \le i \le n-1, \mu_N^h{}_{|_{[x_0,x_1]}} \in P_0([x_0, x_1]) \right\}.$$

The particularity of this space is that the functions are constant on the extreme segment $[x_0, x_1]$. We choose the same kind of approximation for X_T^h, and we set $X^h = X_N^h \times X_T^h$. In this case the discrete Babuška–Brezzi inf-sup condition (2.6) still holds (see [3]). Moreover, estimate (3.5) remains valid (see [7, Lemma 4.1]). Keeping the same definitions of Λ_N^h and $\Lambda_T^h(\mathcal{F}\lambda_N^h)$ as in (2.4) and (2.5), we note that the equivalence in Lemma 2.1 still holds in this case since the dimensions of the multiplier and tangential displacement spaces are the same (see also Remark 1) and the inf-sup condition is satisfied. According to [30], problem (2.3) admits a solution for any friction coefficient, and the solution is unique for a sufficiently small friction coefficient. The following result is then obtained: if (u^h, λ^h) is a solution to (2.3), then estimate (3.1) is recovered.

Remark 7. In the three-dimensional case, the convergence result should hold (at least when $\overline{\Gamma_C} \cap \overline{\Gamma_D} = \emptyset$), and the main task would be to generalize the estimate (3.8).

Remark 8. If $\mathcal{F} = 0$, then the continuous problem admits a unique solution. Choosing then the same approximation method as in (2.3) (therefore $\lambda_T^h = 0$ and the discrete solution is unique) and accomplishing the convergence analysis has led to an upper bound of the error of order $h^{1/2}$ under H^2-regularity hypotheses (see [6]). The estimate obtained in the present paper improves the bound in [6], since we obtain the

same convergence rate with fewer regularity assumptions ($H^{(3/2)+\varepsilon}$ with ε arbitrary small instead of H^2). Moreover, we observe that there is no loss of convergence when the friction terms are added. Nevertheless we mention that there exists in the frictionless case a standard finite element approximation, which leads to an upper bound of the error of order $h^{3/4}$ under H^2-regularity hypotheses (see [24, 23]) and of order h with some additional assumptions concerning the finiteness of transition points between contact and separation (see [29]). Actually we are not able to extend these results to the frictional case.

Remark 9. Note that we do not prove that the solution to the discrete problem is unique under the assumptions of Theorem 3.1. This seems to be an open question which is actually under investigation. Note also that this possible loss of uniqueness would not be embarrassing in the a priori error analysis of Theorem 3.1. As a matter of fact, even if there are multiple solutions to the discrete problem, any solution would converge towards the unique solution of the continuous model. Additionally, the bound ensuring uniqueness in Proposition 1.2 is $\mathcal{F} < (C_\alpha \|\xi\|_a)^{-1}$, and we establish the error estimate only for $\mathcal{F} < c_{is}(C_\alpha \|\xi\|_a)^{-1}$. It should be interesting to see whether or not it is possible to prove an error estimate for all the uniqueness cases of Proposition 1.2.

Conclusion. This work is a contribution to the numerical analysis of the unilateral contact problem governed by Coulomb's law of friction in elastostatics. As far as we know, this study establishes a first error estimate with a convergence rate for this model. From the previous remarks we can reasonably conclude that the present convergence analysis could be generalized in many directions.

REFERENCES

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

[2] L.-E. Andersson, *Existence results for quasistatic contact problems with Coulomb friction*, Appl. Math. Optim., 42 (2000), pp. 169–202.

[3] F. Ben Belgacem, *The mortar finite element method with Lagrange multipliers*, Numer. Math., 84 (1999), pp. 173–197.

[4] F. Ben Belgacem, *Numerical simulation of some variational inequalities arisen from unilateral contact problems by the finite element methods*, SIAM J. Numer. Anal., 37 (2000), pp. 1198–1216.

[5] F. Ben Belgacem, P. Hild, and P. Laborde, *Extension of the mortar finite element method to a variational inequality modeling unilateral contact*, Math. Models Methods Appl. Sci., 9 (1999), pp. 287–303.

[6] F. Ben Belgacem and Y. Renard, *Hybrid finite element methods for the Signorini problem*, Math. Comp., 72 (2003), pp. 1117–1145.

[7] C. Bernardi, Y. Maday, and A.T. Patera, *A new nonconforming approach to domain decomposition: The mortar element method*, in Collège de France seminar, H. Brezis and J.–L. Lions, eds., Pitman, Boston, 1994, pp. 13–51.

[8] Z. Chen and R.H. Nochetto, *Residual type a posteriori error estimates for elliptic obstacle problems*, Numer. Math., 84 (2000), pp. 527–548.

[9] P.G. Ciarlet, *The finite element method for elliptic problems*, in Handbook of Numerical Analysis, Volume II, Part 1, P.G. Ciarlet and J.L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–352.

[10] P. Clément, *Approximation by finite elements functions using local regularization*, RAIRO Anal. Numer., 9 (1975), pp. 77–84.

[11] M. Cocu, *Existence of solutions of Signorini problems with friction*, Internat. J. Engrg. Sci., 22 (1984), pp. 567–575.

[12] P. Coorevits, P. Hild, K. Lhalouani, and T. Sassi, *Mixed finite element methods for unilateral problems: Convergence analysis and numerical studies*, Math. Comp., 71 (2002), pp. 1–25.

[13] M. Crouzeix and V. Thomée, *The stability in L^p and $W^{1,p}$ of the L^2 projection on finite element function spaces*, Math. Comp., 48 (1987), pp. 521–532.

[14] L. Demkovicz and J.T. Oden, *On some existence and uniqueness results in contact problems with nonlocal friction*, Nonlinear Anal., 6 (1982), pp. 1075–1093.
[15] G. Duvaut, *Problèmes unilatéraux en mécanique des milieux continus*, in Actes du Congrès International des Mathématiciens (Nice 1970), Tome 3, Gauthier-Villars, Paris, 1971, pp. 71–77.
[16] G. Duvaut, *Equilibre d'un solide élastique avec contact unilatéral et frottement de Coulomb*, C. R. Acad. Sci. Sér. I Math., 290 (1980), pp. 263–265.
[17] G. Duvaut and J.L. Lions, *Les inéquations en mécanique et en physique*, Dunod, Paris, 1972.
[18] C. Eck and J. Jarušek, *Existence results for the static contact problem with Coulomb friction*, Math. Models Methods Appl. Sci., 8 (1998), pp. 445–468.
[19] C. Eck, J. Jarušek, and M. Krbec, *Unilateral Contact Problems: Variational Methods and Existence Theorems*, Pure Appl. Math. 270, CRC Press, Boca Raton, FL, 2005.
[20] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, 1985.
[21] W. Han and M. Sofonea, *Quasistatic Contact Problems in Viscoelasticity and Viscoplasticity*, American Mathematical Society, Providence, RI, 2002.
[22] J. Haslinger, *Approximation of the Signorini problem with friction, Obeying the Coulomb law*, Math. Methods Appl. Sci., 5 (1983), pp. 422-437.
[23] J. Haslinger and I. Hlaváček, *Contact between elastic bodies—2. Finite element analysis*, Aplikace Matematiky, 26 (1981), pp. 263–290.
[24] J. Haslinger, I. Hlaváček, and J. Nečas, *Numerical methods for unilateral problems in solid mechanics*, in Handbook of Numerical Analysis, Volume IV, Part 2, P.G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1996, pp. 313–485.
[25] R. Hassani, P. Hild, I. Ionescu, and N.-D. Sakki, *A mixed finite element method and solution multiplicity for Coulomb frictional contact*, Comput. Methods Appl. Mech. Engrg., 192 (2003), pp. 4517–4531.
[26] P. Hild, *Non-unique slipping in the Coulomb friction model in two-dimensional linear elasticity*, Quart. J. Mech. Appl. Math., 57 (2004), pp. 225–235.
[27] P. Hild, *Multiple solutions of stick and separation type in the Signorini model with Coulomb friction*, Z. Angew. Math. Mech., 85 (2005), pp. 673–680.
[28] P. Hild, *A Priori Error Analysis of a Sign Preserving Mixed Finite Element Method for Contact Problems*, Internal Report 2006-33 of the Laboratoire de Mathématiques de Besançon; Appl. Numer. Math., submitted.
[29] S. Hüeber and B. I., Wohlmuth, *An optimal a priori error estimate for nonlinear multibody contact problems*, SIAM J. Numer. Anal., 43 (2005), pp. 156–173.
[30] H. Khenous, J. Pommier, and Y. Renard, *Hybrid discretization of the Signorini problem with Coulomb friction, Theoretical aspects and comparison of some numerical solvers*, Appl. Numer. Math., 56 (2006), pp. 163–192.
[31] N. Kikuchi and J.T. Oden, *Contact Problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Stud. Appl. Math. 8, SIAM, Philadelphia, 1988.
[32] A. Klarbring, A. Mikelíc, and M. Shillor, *Frictional contact problems with normal compliance*, Internat. J. Engrg. Sci., 26 (1988), pp. 811–832.
[33] A. Klarbring, A. Mikelíc, and M. Shillor, *On friction problems with normal compliance*, Nonlinear Anal., 13 (1989), pp. 935–955.
[34] P. Laborde and Y. Renard, *Fixed points strategies for elastostatic frictional contact problems*, Math. Methods Appl. Sci., to appear.
[35] T. Laursen, *Computational Contact and Impact Mechanics*, Springer-Verlag, New York, 2002.
[36] J.–L. Lions and E. Magenes, *Problèmes aux limites non homogènes*, Dunod, Paris, 1968.
[37] J.A.C. Martins and J.T. Oden, *Existence and uniqueness results for dynamic contact problems with nonlinear normal and friction interface laws*, Nonlinear Anal., 11 (1987), pp. 407–428.
[38] V.G. Maz'ya and T.O. Shaposhnikova, *Theory of Multipliers in Spaces of Differentiable Functions*, Pitman, Boston, 1985.
[39] J. Nečas, J. Jarušek, and J. Haslinger, *On the solution of the variational inequality to the Signorini problem with small friction*, Boll. Unione Mat. Ital., 17 (1980), pp. 796–811.
[40] J.T. Oden and J.A.C. Martins, *Models and computational methods for dynamic friction phenomena*, Comput. Methods. Appl. Mech. Engrg., 52 (1985), pp. 527–634.
[41] Y. Renard, *A uniqueness criterion for the Signorini problem with Coulomb friction*, SIAM J. Math. Anal., 38 (2006), pp. 452–467.
[42] R. Rocca and M. Cocou, *Numerical analysis of quasi-static unilateral contact problems with local friction*, SIAM J. Numer. Anal., 39 (2001), pp. 1324–1342.
[43] P. Wriggers, *Computational Contact Mechanics*, Wiley, New York, 2002.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2032–2046

SUPERCONVERGENT DERIVATIVE RECOVERY FOR LAGRANGE TRIANGULAR ELEMENTS OF DEGREE p ON UNSTRUCTURED GRIDS*

RANDOLPH E. BANK†, JINCHAO XU‡, AND BIN ZHENG‡

Abstract. In this paper, we develop a postprocessing derivative recovery scheme for the finite element solution u_h on general unstructured but shape regular triangulations. In the case of continuous piecewise polynomials of degree $p \geq 1$, by applying the global L^2 projection (Q_h) and a smoothing operator (S_h), the recovered pth derivatives ($S_h^m Q_h \partial^p u_h$) superconverge to the exact derivatives ($\partial^p u$). Based on this technique we are able to derive a local error indicator depending only on the geometry of corresponding element and the $(p+1)$st derivatives approximated by $\partial S_h^m Q_h \partial^p u_h$. We provide several numerical examples illustrating the effectiveness of our schemes. We also observe that higher order elements are likely to require more conservative refinement strategies to create meshes corresponding to optimal orders of convergence.

Key words. superconvergence, derivative recovery, a posteriori error estimates

AMS subject classifications. 65N50, 65N30

DOI. 10.1137/060675174

1. Introduction.

In this work we introduce a derivative recovery scheme for Lagrange triangular elements of degree p. It is an extension of the gradient recovery scheme for linear elements proposed by Bank and Xu [3]. The recovered pth derivatives are shown to be superconvergent to the exact ones for general shape regular meshes. Due to the superconvergent property of this scheme, some a posteriori error estimates and local error indicators can be derived for mesh adaptation.

The recovery techniques for finite element analysis have been studied extensively in the literature [6, 7, 10, 12, 13, 16, 17]. The main goal of the recovery techniques is to construct better approximations of the solution function or derivative using certain postprocessing procedures. Typically these techniques involve some kind of local or global averaging, including local or global L^2 projection. Due to the superconvergence property, recovery techniques are often used to construct a posteriori error estimators (see, e.g., [5, 11, 14, 15, 18]) which are asymptotically exact. For the literature regarding superconvergence analysis of recovery techniques, we refer to [3] and the references therein.

Most recovery schemes are concerned only with the recovery of the gradient, the finite element solution itself, and the second order derivatives. There was also some work on the recovery of higher order derivatives on uniform grids (see, e.g., [4]). It is the purpose of the current work to recover $(p+1)$st derivatives for Lagrange elements of degree $p \geq 1$ on unstructured grids.

*Received by the editors November 16, 2006; accepted for publication (in revised form) April 16, 2007; published electronically September 19, 2007.

http://www.siam.org/journals/sinum/45-5/67517.html

†Department of Mathematics, University of California, San Diego, La Jolla, CA 92093-0112 (rbank@ucsd.edu). The work of this author was supported by the National Science Foundation under contracts DMS-0511766 and DMS-0208449.

‡Center for Computational Mathematics and Applications and Department of Mathematics, Pennsylvania State University, University Park, PA 16802, and LMAM and School of Mathematical Sciences, Peking University, Beijing 100871 China (xu@math.psu.edu, bzz102@psu.edu). The work of the second author was supported by the National Science Foundation under contracts DMS-0209497 and DMS-0609727, the National Natural Science Foundation of China under contract NSFC-10528102, and the Center for Computational Mathematics and Applications, Penn State University.

Our development has two major components. First, we develop a postprocessing derivative recovery scheme for the finite element solution u_h on general shape regular triangulations. In particular, in section 2 of this paper we compute $S_h^m Q_h \partial^p u_h$, where S_h is an appropriate smoothing operator, $m \in \{1, 2, \dots\}$ is the number of smoothing steps, and Q_h is the L^2 projection operator. The recovered pth derivatives superconverge to the exact ones. In the case of a small number of smoothing steps (the most interesting case), Theorem 2.5 shows that

$$\|\partial^p u - S_h^m Q_h \partial^p u_h\|_{0,\Omega} \lesssim h \left(m h^{1/2} + \left[\frac{\kappa - 1}{\kappa} \right]^m \right) \left(\|u\|_{p+2,\Omega} + |u|_{p+1,\infty,\Omega} \right).$$

Here $\kappa > 1$ is a constant independent of h and u.

The second major component, presented in section 3 of this paper, is the development of a posteriori error estimates based on the derivative recovery scheme. As an example, we discuss quadratic finite elements in detail. For the case of quadratic elements, we define our local error indicator as

$$\epsilon_\tau = \frac{1}{12} \prod_{k=1}^{3} \left(\ell_{k+1} \partial_{k+1} - \ell_{k-1} \partial_{k-1} \right) \bar{u}_3 \phi_0 + \frac{1}{12} \sum_{k=1}^{3} \ell_k^3 \partial_k^3 \bar{u}_3 \phi_k,$$

where $\bar{u}_3$ is any cubic polynomial with third derivatives equal to $\partial S_h^m Q_h \partial^2 u_h$, ℓ_k are the edge lengths of the triangular element, and ϕ_k's are hierarchical basis functions for the 4-dimensional space of cubic polynomials that are zero at the vertices and midpoints of the element. Note that the above local error indicator depends only on the geometry of the corresponding element and the gradients of the recovered second order derivatives.

The rest of this paper is organized as follows: In section 2, we describe our derivative scheme and give superconvergence estimates of the pth derivatives for shape regular meshes. In section 3, we develop and analyze our a posteriori error estimate. Finally, in section 4, we present several numerical examples, involving both uniform and adaptively refined (nonuniform) meshes, with some solutions that satisfy our smoothness assumptions and some that do not. In the latter case, we observe that high order elements require conservative refinement strategies to create meshes corresponding to optimal orders of convergence.

2. A derivative recovery scheme for shape regular triangulations. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with Lipschitz boundary $\partial\Omega$.[1] For simplicity of exposition, we assume that Ω is a polygon. Let $\mathcal{V}_h^{(p)}$ denote the finite element space consisting of C^0 piecewise polynomials of degree p associated with a shape regular triangulation $\mathcal{T}_h$, and let $u_h \in \mathcal{V}_h^{(p)}$ be the finite element approximation to a (possibly nonlinear) second order elliptic boundary value problem.

We analyze a superconvergent approximation to the pth order derivatives of u. This approximation is generated by applying the global L^2 projection operator Q_h and a multigrid smoothing operator S_h to the discrete pth order derivatives of the finite element solution u_h, and it can be represented as $S_h^m Q_h \partial_h^p u_h$.

We first recall that the L^2 projection $Q_h u \in \mathcal{V}_h^{(1)}$ of a given function $u \in L^2(\Omega)$ is defined by solving the variational problem

$$(Q_h u, v_h) = (u, v_h) \; \forall \; v_h \in \mathcal{V}_h^{(1)}. \tag{2.1}$$

Here $(\cdot, \cdot)$ denotes the inner product on $L^2(\Omega)$.

[1] It is easy to see that our theory in this paper is also valid for domains with cracks, such as the slit domain in the third example of section 4.

Consider the following bilinear form:

$$a(u,v) = (\nabla u, \nabla v) + (u,v). \tag{2.2}$$

By the Riesz representation theorem, $a(\cdot,\cdot)$ induces a bounded linear operator $A_h : \mathcal{V}_h^{(1)} \to \mathcal{V}_h^{(1)}$ uniquely determined by

$$(A_h u_h, v_h) = a(u_h, v_h) \; \forall \; u_h, v_h \in \mathcal{V}_h^{(1)},$$

and it follows that the operator A_h is symmetric with respect to the L^2-inner product. We further notice that the discrete operator A_h is symmetric positive definite on the finite dimensional space $\mathcal{V}_h^{(1)}$ and

$$\lambda \equiv \rho(A_h) \simeq \mathcal{O}(h^{-2}).$$

Using A_h, we introduce the smoothing operator S_h defined by

$$S_h = I - \lambda^{-1} A_h.$$

The usual multigrid convergence function

$$f(\alpha,\beta) = \frac{\alpha^\alpha \beta^\beta}{(\alpha+\beta)^{(\alpha+\beta)}} = \sup_{x\in[0,1]} x^\alpha (1-x)^\beta,$$

$\alpha, \beta > 0$, plays an important role [3].

For convenience in notation, we let $\partial^p u$ denote some pth order derivative of u and $\partial^p u_h$ denote some discrete pth order derivative of u_h. We also use the notation $\|\cdot\|'_{p,\Omega}$ to indicate the discrete norm $\sum_{K\in\mathcal{T}_h} \|\cdot\|_{p.K}$. We now state and prove some preliminary lemmas leading to the main Theorem 2.5 in this section.

LEMMA 2.1. *For any $z \in \mathcal{V}_h^{(1)}$, $u \in H^{p+2}(\Omega)$,*

$$\|(I - S_h^m)z\|_{0,\Omega} \lesssim mh(\|z - \partial^p u\|_{1,\Omega} + h\|u\|_{p+2,\Omega} + h^{1/2}|u|_{p+1,\partial\Omega}).$$

Proof. We note, from the definition of S_h, that

$$\begin{aligned}
\|(I - S_h^m)z\|_{0,\Omega} &= \lambda^{-1}\|(I - S_h^m)(I - S_h)^{-1} A_h z\|_{0,\Omega} \\
&\le \lambda^{-1} \max_{s\in[0,1]} [(1-s^m)(1-s)^{-1}] \|A_h z\|_{0,\Omega} \\
&\le \lambda^{-1} m \|A_h z\|_{0,\Omega} \\
&\lesssim mh^2 \|A_h z\|_{0,\Omega}.
\end{aligned}$$

Let $w = A_h z$. By definition,

$$(w,\varphi) = (\nabla z, \nabla\varphi) + (z,\varphi) \tag{2.3}$$

for all $\varphi \in \mathcal{V}_h^{(1)}$. We take $\varphi = w$ in (2.3),

$$\|w\|_{0,\Omega}^2 = (w,w) = (\nabla z, \nabla w) + (z,w).$$

We estimate the terms on the right-hand side

$$\begin{aligned}
(\nabla z, \nabla w) &= (\nabla(z - \partial^p u), \nabla w) + (\nabla \partial^p u, \nabla w) \\
&\lesssim \|\nabla(z - \partial^p u)\|_{0,\Omega} \|\nabla w\|_{0,\Omega} - (\triangle \partial^p u, w) + \int_{\partial\Omega} \nabla \partial^p u \cdot n w ds \\
&\lesssim h^{-1} \|\nabla(z - \partial^p u)\|_{0,\Omega} \|w\|_{0,\Omega} + \|u\|_{p+2,\Omega} \|w\|_{0,\Omega} + |u|_{p+1,\partial\Omega} \|w\|_{0,\partial\Omega} \\
&\lesssim (h^{-1}\|z - \partial^p u\|_{1,\Omega} + \|u\|_{p+2,\Omega} + h^{-1/2}|u|_{p+1,\partial\Omega}) \|w\|_{0,\Omega}.
\end{aligned}$$

Also

$$\begin{aligned}(z,w) &= (z-\partial^p u, w) + (\partial^p u, w)\\ &\lesssim (\|z-\partial^p u\|_{0,\Omega} + \|u\|_{p,\Omega})\|w\|_{0,\Omega}.\end{aligned}$$

Thus for $z \in \mathcal{V}_h^{(1)}$,

$$\|A_h z\|_{0,\Omega} = \|w\|_{0,\Omega} \lesssim h^{-1}\|z-\partial^p u\|_{1,\Omega} + \|u\|_{p+2,\Omega} + h^{-1/2}|u|_{p+1,\partial\Omega},$$

completing the proof. ∎

LEMMA 2.2 ([3]). *Suppose that for $v \in \mathcal{V}_h^{(1)}$ and some $0 < \alpha \le 1$ we have*

$$\|v\| \le \omega(h,v),$$
$$\|v\|_{-\alpha} \equiv \|A_h^{-\alpha/2} v\| \le (\mathcal{C}h)^\alpha \omega(h,v).$$

Then

$$\|S_h^m v\| \le \varepsilon_m\, \omega(h,v),$$

where

$$\varepsilon_m = \begin{cases} \kappa^{\alpha/2} f(m,\alpha/2) \lesssim m^{-\alpha/2} & \text{for } m > (\kappa-1)\alpha/2,\\ [(\kappa-1)/\kappa]^m & \text{for } m \le (\kappa-1)\alpha/2,\end{cases}$$

and $\kappa = (\mathcal{C}h)^2\lambda$.

Proof. See Lemma 2.3 of [3]. ∎

LEMMA 2.3. *Let $w|_K \in H^p(K) \cap W^{p-1,\infty}(K)$ for all $K \in \mathcal{T}_h$. Then, for $1/2 < \alpha \le 1$,*

$$\|S_h^m Q_h \partial^p w\|_{0,\Omega} \lesssim \varepsilon_m (h^{-1}\|w\|'_{p-1,\Omega} + \|w\|'_{p,\Omega} + h^{-\alpha}\|w\|'_{p-1,\infty,\Omega}),$$

with ε_m defined as in Lemma 2.2.

Proof. Our plan is to apply Lemma 2.2 to $v = Q_h\partial^p w$. Note that

$$\|v\|_{-\alpha} = \|Q_h\partial^p w\|_{-\alpha} = \sup_{\phi\in\mathcal{V}_h^{(1)}} \frac{(Q_h\partial^p w,\phi)}{\|\phi\|_\alpha} = \sup_{\phi\in\mathcal{V}_h^{(1)}} \frac{(\partial^p w,\phi)}{\|\phi\|_\alpha}.$$

Here $\|\phi\|_\alpha \equiv \|A_h^{\alpha/2}\phi\|$. Using integration by parts,

$$\begin{aligned}(\partial^p w,\phi) &= -(\partial^{p-1}w, \partial\phi) + \sum_{K\in\mathcal{T}_h}\int_{\partial K} \partial^{p-1} w\phi n_i ds\\ &\le \|w\|'_{p-1,\Omega}\|\phi\|_{1,\Omega} + \|w\|'_{p-1,\infty,\Omega}\|\phi\|_{\alpha,\Omega}\\ &\lesssim (h^{\alpha-1}\|w\|'_{p-1,\Omega} + \|w\|'_{p-1,\infty,\Omega})\|\phi\|_{\alpha,\Omega}.\end{aligned}$$

Thus

$$\|v\|_{-\alpha} \lesssim h^\alpha \omega(h,v),$$

with $\omega(h,v) = h^{-1}\|w\|'_{p-1,\Omega} + \|w\|'_{p,\Omega} + h^{-\alpha}\|w\|'_{p-1,\infty,\Omega}$.

Since

$$\|v\|_{0,\Omega} \le \|\partial^p w\|'_{0,\Omega} \le \omega(h,v),$$

the desired estimate now follows from Lemma 2.2. ∎

LEMMA 2.4. *Let $u \in H^{p+2}(\Omega) \cap W^{p+1,\infty}(\Omega)$. Then for any $v_h \in \mathcal{V}_h^{(p)}$ and $1/2 < \alpha \leq 1$ we have*

$$\|\partial^p u - S_h^m Q_h \partial^p v_h\|_{0,\Omega} \lesssim mh^{3/2}(h^{1/2}\|u\|_{p+2,\Omega} + |u|_{p+1,\partial\Omega}) + \varepsilon_m(h^{-1}\|u - v_h\|'_{p-1,\Omega} + h^{-\alpha}\|u - v_h\|'_{p-1,\infty,\Omega}),$$

with ε_m defined as in Lemma 2.2.

Proof. By the triangle inequality,

$$\|\partial^p u - S_h^m Q_h \partial^p v_h\|_{0,\Omega} \leq \|(I - Q_h)\partial^p u\|_{0,\Omega} + \|(I - S_h^m)Q_h\partial^p u\|_{0,\Omega} + \|S_h^m Q_h \partial^p (u - v_h)\|_{0,\Omega}.$$

By standard arguments, the first term

$$\|(I - Q_h)\partial^p u\|_{0,\Omega} \lesssim h^2\|u\|_{p+2,\Omega}.$$

The second term is estimated by Lemma 2.1. For the third term, we apply Lemma 2.3. ∎

In the case in which $v_h = u_h \in \mathcal{V}_h^{(p)} \cap H_0^1(\Omega)$ is the finite element approximation to $u \in H_0^1(\Omega)$, the boundary terms vanish and

$$\|\partial^p u - S_h^m Q_h \partial^p u_h\|_{0,\Omega} \lesssim h(mh + \varepsilon_m)\|u\|_{p+2,\Omega}.$$

In the more general case, we have the following theorem based on the results developed in this section.

THEOREM 2.5. *Let $u \in H^{p+2}(\Omega) \cap W^{p+1,\infty}(\Omega)$ and $u_h \in \mathcal{V}_h^{(p)}$ be an approximation of u satisfying*

$$\|u - u_h\|'_{p-1,\Omega} \lesssim h^2|u|_{p+1,\Omega},$$

$$\|u - u_h\|'_{p-1,\infty,\Omega} \lesssim h^2|\log h||u|_{p+1,\infty,\Omega}.$$

Then

$$\|\partial^p u - S_h^m Q_h \partial^p u_h\|_{0,\Omega} \lesssim h(mh^{1/2} + \varepsilon_m)(\|u\|_{p+2,\Omega} + |u|_{p+1,\infty,\Omega}),$$

where ε_m is defined as in Lemma 2.2 and $1/2 < \alpha < 1$.

We can easily derive the following estimate for $(p+1)$st order derivatives with the help of Theorem 2.5.

THEOREM 2.6. *Assume the hypotheses of Theorem 2.5 are satisfied. Then*

$$\|\partial(\partial^p u - S_h^m Q_h \partial^p u_h)\|_{0,\Omega} \lesssim (mh^{1/2} + \varepsilon_m)(\|u\|_{p+2,\Omega} + |u|_{p+1,\infty,\Omega}),$$

where ε_m is defined as in Lemma 2.2 and $1/2 < \alpha < 1$.

Proof. Let $z = I_h \partial^p u \in \mathcal{V}_h^{(1)}$. Then

$$\begin{aligned}\|\partial(\partial^p u - S_h^m Q_h \partial^p u_h)\|_{0,\Omega} &\leq \|\partial(\partial^p u - z)\|_{0,\Omega} + \|\partial(z - S_h^m Q_h \partial^p u_h\|)_{0,\Omega} \\ &\lesssim h\|u\|_{p+2,\Omega} + h^{-1}\|z - S_h^m Q_h \partial^p u_h\|_{0,\Omega} \\ &\lesssim h\|u\|_{p+2,\Omega} + h^{-1}(\|z - \partial^p u\|_{0,\Omega} + \|\partial^p u - S_h^m Q_h \partial^p u_h\|_{0,\Omega}) \\ &\lesssim (mh^{1/2} + \varepsilon_m)(\|u\|_{p+2,\Omega} + |u|_{p+1,\infty,\Omega}).\end{aligned}$$ ∎

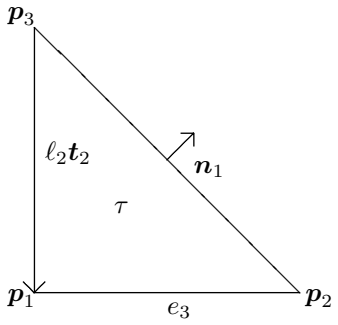


FIG. 3.1. *Parameters associated with the triangle* τ.

3. A posteriori error estimates. We begin with a description of our a posteriori error estimator. Our approach follows the development in [3] for the case of piecewise linear finite elements. For the general case of Lagrange elements of degree p, our goal is to find an expression for the error that involves only (approximate) derivatives of order $p+1$ of u and known parameters describing the geometry of a given element τ. Let a canonical element $\tau \in \mathcal{T}_h$ have vertices $\boldsymbol{p}_k^t = (x_k, y_k)$, $1 \le k \le 3$, oriented counterclockwise, and corresponding linear nodal basis functions (barycentric coordinates) $\{\psi_k\}_{k=1}^3$. Let $\{e_k\}_{k=1}^3$ denote the edges of element τ, $\{\boldsymbol{n}_k\}_{k=1}^3$ the unit outward normal vectors, $\{\boldsymbol{t}_k\}_{k=1}^3$ the unit tangent vectors with counterclockwise orientation, and $\{\ell_k\}_{k=1}^3$ the edge lengths (see Figure 3.1).

As an example, we now restrict attention to quadratic finite elements, since it is the quadratic space that is used in the numerical illustrations. We first seek an expression for $\hat{u}_3 - u_2$ on τ, where u_2 is the quadratic Lagrange interpolant and $\hat{u}_3$ is the cubic hierarchical extension. Thus $\hat{u}_3 - u_2$ is a cubic polynomial vanishing at vertices and edge midpoints of τ. A hierarchical basis for this 4-dimensional space is given by

$$\phi_0 = \psi_1\psi_2\psi_3,$$
$$\phi_k = \psi_{k-1}\psi_{k+1}(\psi_{k+1} - \psi_{k-1}),$$

for $1 \le k \le 3$, and $(k-1, k, k+1)$ is a cyclic permutation of $(1, 2, 3)$. Let $\partial_k u$ denote the directional derivative in the direction $\boldsymbol{t}_k$. Then

$$\hat{u}_3 - u_2 = \frac{1}{12}\prod_{k=1}^{3}\left(\ell_{k+1}\partial_{k+1} - \ell_{k-1}\partial_{k-1}\right)\hat{u}_3\phi_0 + \frac{1}{12}\sum_{k=1}^{3}\ell_k^3\partial_k^3\hat{u}_3\phi_k. \tag{3.1}$$

Equation (3.1) can be verified using the identities

$$\psi_1 + \psi_2 + \psi_3 = 1,$$
$$\nabla\psi_1 + \nabla\psi_2 + \nabla\psi_3 = 0,$$
$$\ell_1\boldsymbol{t}_1 + \ell_2\boldsymbol{t}_2 + \ell_3\boldsymbol{t}_3 = 0,$$
$$\begin{pmatrix}\ell_1\boldsymbol{t}_1^t\\ \ell_2\boldsymbol{t}_2^t\\ \ell_3\boldsymbol{t}_3^t\end{pmatrix}\begin{pmatrix}\nabla\psi_1 & \nabla\psi_2 & \nabla\psi_3\end{pmatrix} = \begin{pmatrix}0 & -1 & 1\\ 1 & 0 & -1\\ -1 & 1 & 0\end{pmatrix}.$$

In our local error indicator, we simply approximate the third derivatives needed to compute the directional derivatives appearing in (3.1) by

$$
\begin{aligned}
\partial_{xxx}\hat{u}_3 &\approx \alpha_\tau \partial_x S_h^m Q_h \partial_{xx} u_h, \\
\partial_{xxy}\hat{u}_3 &\approx \frac{\alpha_\tau}{2}(\partial_y S_h^m Q_h \partial_{xx} u_h + \partial_x S_h^m Q_h \partial_{xy} u_h), \\
\partial_{xyy}\hat{u}_3 &\approx \frac{\alpha_\tau}{2}(\partial_y S_h^m Q_h \partial_{xy} u_h + \partial_x S_h^m Q_h \partial_{yy} u_h), \\
\partial_{yyy}\hat{u}_3 &\approx \alpha_\tau \partial_y S_h^m Q_h \partial_{yy} u_h,
\end{aligned}
\tag{3.2}
$$

where $\alpha_\tau > 0$ is a constant described below. Let $\bar{u}_3$ be any cubic polynomial with third derivatives given by the right-hand sides of (3.2). Then our local error indicator is given by

$$
\epsilon_\tau = \frac{1}{12}\prod_{k=1}^{3}\left(\ell_{k+1}\partial_{k+1} - \ell_{k-1}\partial_{k-1}\right)\bar{u}_3\phi_0 + \frac{1}{12}\sum_{k=1}^{3}\ell_k^3\partial_k^3\bar{u}_3\phi_k. \tag{3.3}
$$

The normalization constant α_τ is chosen such that

$$
\begin{aligned}
|\epsilon_\tau|_{2,\tau}^2 = \|(I - S_h^m Q_h)\partial_{xx}^2 u_h\|_{0,\tau}^2 + 2\|(I - S_h^m Q_h)\partial_{xy}^2 u_h\|_{0,\tau}^2 \\
+ \|(I - S_h^m Q_h)\partial_{yy}^2 u_h\|_{0,\tau}^2 \equiv |u_h - R(u_h)|_{2,\tau}^2.
\end{aligned}
$$

Normally we expect that $\alpha_\tau \approx 1$, which is likely to be the case in regions where the third derivatives of the true solution are well defined. Near singularities, u is not smooth, and we anticipate difficulties in estimating the third derivatives. For elements near such singularities, α_τ provides a heuristic for partly compensating for poor approximation. Note that ϵ_τ is a cubic polynomial on each element depending only on the geometry of τ and the approximate third derivatives derived from our superconvergent approximations.

In the general case, $\hat{u}_{p+1} - u_p$ on element τ is a polynomial of degree $p+1$ that is zero at the degrees of freedom defining u_p. One can express this polynomial in terms of hierarchical basis functions depending on the geometry of τ and the derivatives of order $p+1$ of $\hat{u}_{p+1}$. The derivatives can be approximated by $\partial S_h^m Q_h \partial^p u_h$ as in the example above. This yields a polynomial ϵ_τ of degree $p+1$ for each element. The local error indicator η_τ is given by

$$
\eta_\tau = \|\nabla \epsilon_\tau\|_{0,\tau}. \tag{3.4}
$$

Since ϵ_τ is a discontinuous piecewise polynomial on all of Ω, we can also formally approximate errors in global norms and other functionals using ϵ_τ. For example,

$$
\begin{aligned}
\|u - u_h\|_{0,\Omega}^2 &\approx \sum_\tau \|\epsilon_\tau\|_{0,\tau}^2, \\
|u - u_h|_{1,\Omega}^2 &\approx \sum_\tau |\epsilon_\tau|_{1,\tau}^2 = \sum_\tau \eta_\tau^2.
\end{aligned}
$$

In the case of $|\cdot|_{1,\Omega}$ there is a bit of theory. In particular,

$$|u-u_h|_{1,\Omega} \le |u-\hat{u}_{p+1}|_{1,\Omega} + |u_p-u_h|_{1,\Omega} + |u_p-\hat{u}_{p+1}|_{1,\Omega}, \tag{3.5}$$

$$|u_p-\hat{u}_{p+1}|_{1,\Omega} \le |u-\hat{u}_{p+1}|_{1,\Omega} + |u_p-u_h|_{1,\Omega} + |u-u_h|_{1,\Omega}. \tag{3.6}$$

Suppose $|u-u_h|_{1,\Omega} \ge ch^p$, and $|u-\hat{u}_{p+1}|_{1,\Omega} \le Ch^{p+1}$. Then if $|u_p-u_h|_{1,\Omega}$ is also higher order, estimates (3.5)–(3.6) show $|u_p-\hat{u}_{p+1}|_{1,\Omega}$ to be an asymptotically exact estimate for $|u-u_h|_{1,\Omega}$. Such superapproximation results for $|u_p-u_h|_{1,\Omega}$ are known for $p=1,2$; see [2, 3, 8]. However, superapproximation estimates for $|u_p-u_h|_{1,\Omega}$ are not yet known to hold for $p \ge 3$; see [9]. For general p, estimate (3.5) can be replaced by

$$|u-u_h|_{1,\Omega} \le C\left(|u-\hat{u}_{p+1}|_{1,\Omega} + |\hat{u}_{p+1}-u_p|_{1,\Omega}\right) \tag{3.7}$$

due to the best approximation property for the energy norm and norm comparability of $|\!|\!|\cdot|\!|\!|_\Omega$ and $|\cdot|_{1,\Omega}$. Here we may lose asymptotic exactness but still have a useful upper bound for the error. Insofar as we know, the lower bound for $p>2$ is still an open question, as are general norms and functionals. Nonetheless, this informal analysis suggests that the error indicators η_τ will provide a useful and reliable basis for adaptive meshing algorithms.

4. Numerical experiments. We now present some numerical illustrations of our recovery scheme in the cases of uniform and adaptively refined (nonuniform) meshes. The gradient recovery scheme and a posteriori error estimate described above for the case of continuous piecewise quadratic elements were implemented in the PLTMG package [1], which was then used for our numerical experiments. The experiments were done on a dual Opteron Linux workstation, using the *g77* compiler and double precision arithmetic. We reprise some experiments given in [3] for the case of continuous piecewise linear finite elements.

In our first example, we consider the solution of the problem

$$\begin{aligned} -\Delta u &= f \qquad \text{in } \Omega=(0,1)\times(0,1),\\ u &= g \qquad \text{on } \partial\Omega, \end{aligned}$$

where f and g are chosen such that $u=\mathrm{e}^{x+y}$ is the exact solution. This is a very smooth solution that satisfies all of the assumptions of our theory. Here we will compare the recovery scheme with $m=2$ smoothing steps for the case of uniform and adaptive meshes. We begin with a uniform 3×3 mesh consisting of eight right triangles as shown in Figure 4.1. Elements in Figure 4.1 are colored according to size; this allows one to obtain some impression of the structure of highly refined meshes with many elements, even if individual elements can no longer be resolved.

In Tables 4.1–4.2, we record the results of the computation. We give the error as a function of the number of elements, choosing targets for the adaptive refinement procedure to produce adaptive meshes with similar numbers of elements to the uniform refinement case. Note that the dimension of the quadratic finite element space is approximately $2\,nt$, where nt is the number of elements reported in the tables. Other

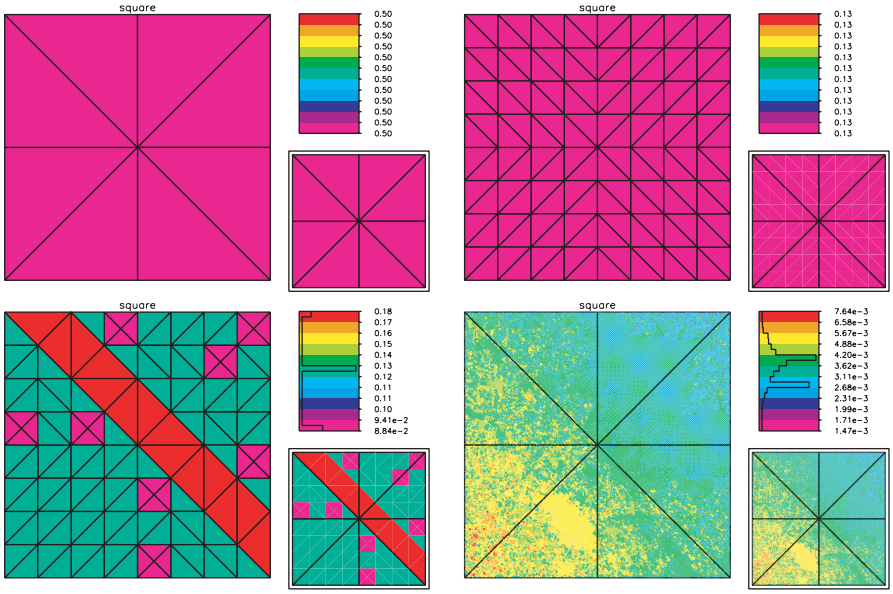


FIG. 4.1. *Top left:* 3×3 *initial mesh. Top right: Uniform refinement with* $nt = 128$. *Bottom left: Adaptive refinement with* $nt = 137$. *Bottom right: Adaptive refinement with* $nt = 131105$. *Elements are colored according to size.*

TABLE 4.1
Error estimates for uniform refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	8.8e-3	1.0e-2	1.1	0.1	0.2	1.6	1.3	2.1	1.7
32	1.0e-3	1.8e-3	1.8	3.0e-2	0.1	1.8	0.7	1.2	2.0
128	1.2e-4	2.0e-4	1.6	7.5e-3	1.2e-2	1.6	0.3	0.5	1.7
512	1.6e-5	2.4e-5	1.6	1.9e-3	2.9e-3	1.5	0.2	0.2	1.5
2048	1.9e-6	2.7e-6	1.4	4.7e-4	6.5e-4	1.4	0.1	0.1	1.4
8192	2.4e-7	3.1e-7	1.3	1.2e-4	1.5e-4	1.3	4.2e-2	3.4e-2	1.3
32768	3.0e-8	3.5e-8	1.2	3.0e-5	3.4e-5	1.2	2.1e-2	1.3e-2	1.2
131072	3.8e-9	4.1e-9	1.1	7.4e-6	8.0e-6	1.1	1.0e-2	4.7e-3	1.1
Order	3.04	3.15		2.02	2.13		1.01	1.43	

TABLE 4.2
Error estimates for adaptive refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	6.9e-4	3.7e-4	0.5	1.0e-2	5.6e-3	0.6	0.2	0.2	0.5
33	2.5e-4	1.8e-4	0.7	5.1e-3	4.8e-3	0.9	0.1	0.2	1.0
137	1.6e-5	2.2e-5	1.4	8.9e-4	1.5e-3	1.6	0.1	0.1	1.7
523	1.8e-6	2.2e-6	1.2	1.8e-4	2.6e-4	1.4	2.2e-2	3.1e-2	1.6
2063	2.0e-7	2.0e-7	1.0	3.7e-5	4.4e-5	1.2	1.0e-2	1.0e-2	1.3
8207	1.8e-8	1.6e-8	0.9	7.9e-6	7.9e-6	1.0	4.7e-3	2.6e-3	1.1
32775	2.2e-9	1.7e-9	0.8	1.9e-6	1.7e-6	0.9	2.3e-3	7.3e-4	1.0
131105	2.6e-1	2.0e-1	0.8	4.5e-7	4.1e-7	0.9	1.1e-3	2.1e-4	1.0
Order	3.15	3.24		2.12	2.20		1.06	1.83	

values are defined as follows:

$$L2 = \|u - u_h\|_{0,\Omega},$$
$$\widetilde{L2} = \|\epsilon_h\|_{0,\Omega},$$
$$EF_0 = \frac{\|\epsilon_h\|_{0,\Omega}}{\|u - u_h\|_{0,\Omega}},$$
$$H1 = |u - u_h|_{1,\Omega},$$
$$\widetilde{H1} = |\epsilon_h|_{1,\Omega},$$
$$EF_1 = \frac{|\epsilon_h|_{1,\Omega}}{|u - u_h|_{1,\Omega}},$$
$$H2 = |u - u_h|_{2,\Omega},$$
$$\widetilde{H2} = |u - R(u_h)|_{2,\Omega},$$
$$EF_2 = \frac{|R(u_h) - u_h|_{2,\Omega}}{|u - u_h|_{2,\Omega}}.$$

For each type of norm, we made a least squares fit of the data to a function of the form $F(N) = CN^{-p/2}$ to estimate the order of convergence p. All integrals were approximated using a 12-point order 7 quadrature formula applied to each triangle.

We note here the superconvergence of the second derivatives and effectivity ratios that are close to one. Despite the lack of a complete theory, error estimates $\widetilde{L2}$ and $\widetilde{H1}$ are also quite accurate, and the orders of convergence are optimal in all three norms (and superconvergent for the recovered second derivatives).

In our second example, we consider the nonlinear problem

$$\begin{aligned} -\nabla \cdot (a\nabla u) + \mathrm{e}^u &= f \qquad \text{in } \Omega = (0,1) \times (0,1), \\ u &= 0 \qquad \text{on } \partial\Omega, \end{aligned}$$

where a is the 2×2 diagonal matrix

$$a = \begin{pmatrix} .01 & \\ & 1 \end{pmatrix}.$$

The function f is chosen such that $u = x(1-x)^3 y^5 (1-y)$ is the exact solution. We repeat the same computations as in the first example, with uniform and adaptive meshes. The uniform meshes are identical to those of the first example. Some of the adaptive meshes are shown in Figure 4.2. The numerical results are summarized in Tables 4.3–4.4.

This problem is more difficult than the first in several respects. The diffusion is anisotropic, and the operator is nonlinear. The solution is smooth but generally has larger derivatives than the first example. Nonetheless, we see a similar behavior of the gradient recovery scheme and a posteriori error estimate.

In our third example, we consider the problem

$$\begin{aligned} -\Delta u &= 0 \qquad \text{in } \Omega, \\ u &= g \qquad \text{on } \partial\Omega_1, \\ u_n &= 0 \qquad \text{on } \partial\Omega_2, \end{aligned}$$

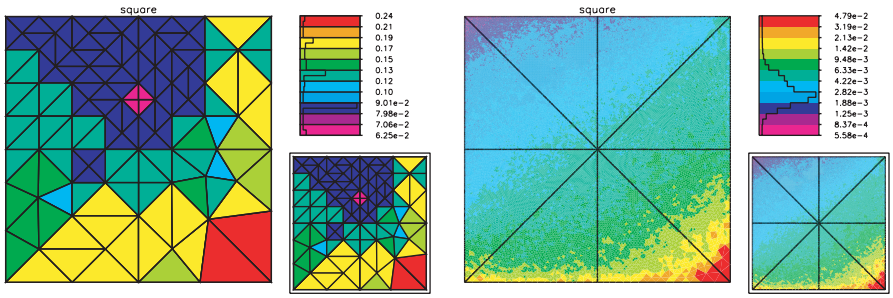


FIG. 4.2. *Left: Adaptive refinement with nt = 135. Right: Adaptive refinement with nt = 129345. Elements are colored according to size.*

TABLE 4.3
Error estimates for uniform refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	6.9e-4	3.7e-4	0.5	1.0e-2	5.6e-3	0.6	0.2	0.2	0.5
32	2.2e-4	1.3e-4	0.6	4.7e-3	3.9e-3	0.8	0.1	0.2	1.0
128	4.3e-5	3.1e-5	0.7	2.0e-3	2.0e-3	1.0	0.1	0.1	1.4
512	6.1e-6	5.3e-6	0.9	6.2e-4	7.0e-4	1.1	4.4e-2	0.1	1.5
2048	6.7e-7	7.4e-7	1.1	1.5e-4	2.0e-4	1.3	2.1e-2	3.7e-2	1.6
8192	6.4e-8	8.9e-8	1.4	3.0e-5	4.7e-5	1.6	1.0e-2	1.6e-2	1.7
32768	6.4e-9	1.0e-8	1.5	6.4e-6	1.0e-5	1.6	4.5e-3	6.3e-3	1.6
131072	7.1e-1	1.0e-9	1.5	1.5e-6	2.2e-6	1.5	2.2e-3	2.4e-3	1.4
Order	3.26	3.21		2.18	2.20		1.08	1.35	

TABLE 4.4
Error estimates for adaptive refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	6.9e-4	2.0e-4	0.3	1.0e-2	3.0e-3	0.3	0.2	0.2	0.5
32	2.2e-4	8.2e-5	0.4	4.7e-3	2.5e-3	0.5	0.1	0.2	1.0
135	1.5e-5	1.9e-5	1.2	8.6e-4	1.5e-3	1.7	0.1	0.1	1.8
524	2.0e-6	2.8e-6	1.4	1.7e-4	5.2e-4	3.0	2.1e-2	3.0e-2	1.6
2060	3.4e-7	4.1e-7	1.2	4.5e-5	1.4e-4	3.0	1.0e-2	9.1e-3	1.3
8119	4.6e-8	4.8e-8	1.1	1.2e-5	3.2e-5	2.6	5.2e-3	2.2e-3	1.1
32333	5.4e-9	5.2e-9	1.0	2.9e-6	6.4e-6	2.2	2.7e-3	4.9e-4	1.0
129345	5.8e-1	5.4e-1	0.9	7.0e-7	1.3e-6	1.8	1.3e-3	1.5e-4	1.0
Order	3.16	3.25		2.07	2.30		0.99	1.86	

where Ω is a circle of radius one centered at the origin, and with a crack along the positive x-axis $0 \le x \le 1$. The boundary $\partial\Omega_2$ is the bottom edge of the crack, and $\partial\Omega_1 = \partial\Omega \setminus \partial\Omega_2$. The function g is chosen such that the exact solution is $u = r^{1/4}\sin(\theta/4)$, the leading term of the singularity associated with the interior angle of 2π and change in boundary conditions at the origin. In Figure 4.3 we illustrate the initial mesh and several of the uniformly and adaptively refined meshes.

Convergence results for uniform and adaptive refinement are reported in Tables 4.5–4.6. The solution u is not smooth in this case ($u \in H^{5/4-\epsilon}(\Omega)$), and this is reflected in the results. In particular, both $H2$ and $\widetilde{H2}$ should be infinite and are computed as finite only because of numerical quadrature. Despite this singularity, the resulting error approximations ϵ_τ still provided useful information and formed a reliable basis for adaptive refinement.

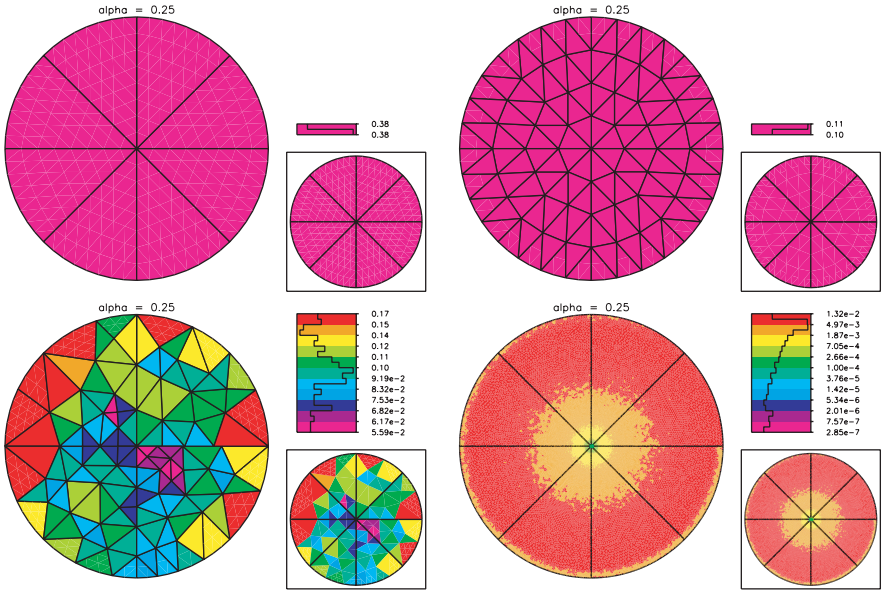


FIG. 4.3. *Top left: Initial mesh with nt = 8. Top right: Uniform refinement with nt = 128. Bottom left: Adaptive refinement with nt = 133. Bottom right: Adaptive refinement with nt = 133890. Elements are colored according to size.*

TABLE 4.5
Error estimates for uniform refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	0.1	3.9e-2	0.3	0.6	0.3	0.5	4.1	4.3	0.4
32	0.1	8.6e-3	0.1	0.5	0.1	0.3	6.6	6.7	0.5
128	0.1	2.9e-3	4.9e-2	0.4	0.1	0.3	11.0	11.0	0.5
512	4.0e-2	1.2e-3	2.9e-2	0.3	0.1	0.3	18.0	19.0	0.5
2048	2.7e-2	4.7e-4	1.7e-2	0.2	0.1	0.3	31.0	32.0	0.5
8192	1.9e-2	2.0e-4	1.0e-2	0.2	0.1	0.3	52.0	53.0	0.5
32768	1.3e-2	8.1e-5	6.1e-3	0.2	4.6e-2	0.3	87.0	90.0	0.5
131072	9.3e-3	3.4e-5	3.7e-3	0.1	3.8e-2	0.3	1.5e2	1.5e2	0.5
Order	0.53	1.29		0.27	0.27		-0.76	-0.76	

TABLE 4.6
Error estimates for adaptive refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	0.1	3.9e-2	0.3	0.6	0.3	0.5	4.1	4.3	0.4
31	0.1	8.9e-3	0.1	0.5	0.1	0.3	6.4	6.5	0.5
133	4.5e-2	2.3e-3	5.0e-2	0.3	0.1	0.3	14.0	14.0	0.5
533	2.3e-2	4.4e-4	1.9e-2	0.2	0.1	0.3	34.0	35.0	0.5
2078	8.5e-3	5.7e-5	6.7e-3	0.1	0.1	0.4	1.5e2	1.6e2	0.5
8237	2.3e-3	4.6e-6	2.0e-3	0.1	2.7e-2	0.4	1.0e3	1.1e3	0.5
32796	6.0e-4	5.8e-7	1.0e-3	3.5e-2	1.2e-2	0.3	7.7e3	8.1e3	0.5
130890	1.1e-4	7.4e-8	6.6e-4	1.5e-2	7.6e-3	0.5	9.8e4	1.1e5	0.6
Order	2.18	3.07		1.12	0.83		-3.29	-3.30	

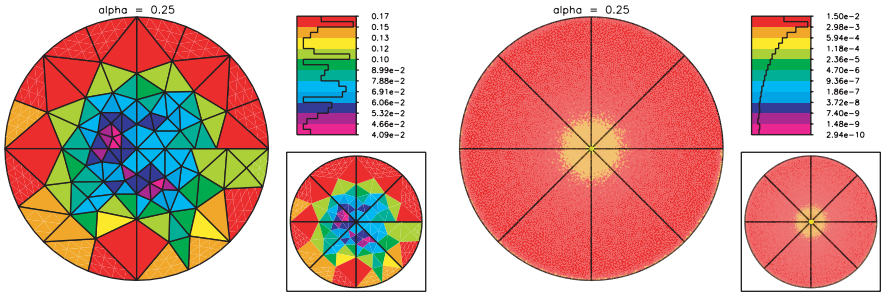


FIG. 4.4. *Left: Adaptive refinement with* $nt = 136$. *Right: Adaptive refinement with* $nt = 130586$. *Elements are colored according to size.*

For the case of uniform refinement, the 0.25 order of convergence of the gradient coincides with the smoothness of the solution. The effectivity ratios for all cases show a lack of asymptotic exactness. EF_1, although not approaching unity, still seems under control, reflecting the analysis (3.5)–(3.7). On the other hand, EF_0 seems very poorly controlled, indicating that $\|\epsilon_\tau\|_{0,\Omega}$ is not a very reliable estimate for $\|u-u_h\|_{0,\Omega}$.

For the adaptive meshes, the order of convergence improves and seems to be approaching order one for the gradient and order two for the solution. This is suboptimal for quadratic elements. This was not due to poor error indicators but rather to an overly aggressive refinement strategy. In the adaptive refinement procedure implemented in PLTMG, all elements are placed on a heap with the element having the largest error at the root. The root element is then selected for refinement. When an element is refined, it is removed from the heap, and its child elements are added to the heap. This of course requires the child elements to have error indicators. These are constructed using derivative values inherited from the parent and their own geometry information. Thus a single element might undergo several levels of refinement during a given adaptive step. Using old derivative information for new elements will generally fail to be optimal after sufficiently many levels of refinement.

In PLTMG, the amount of refinement allowed in a given refinement step is governed by the user by specifying a target number of vertices in the refined mesh. In this example, we chose a strategy that increased the number of vertices by roughly a factor of four in each refinement step in order to closely match the size of problems generated by uniform refinement. If there are too many levels of refinement of individual elements before the problem is resolved, the resulting mesh might be lower quality, as was the case in this example. On the other hand, frequently assembling and resolving the global finite element equations results in higher quality adaptive meshes but at a much greater cost. Since the appropriate compromise is likely to be highly problem-dependent, in PLTMG it is up to the user to choose the proper balance.

For this example, we solved this problem adaptively a second time, this time specifying that the number of vertices should be increased by a factor of roughly two between resolves rather than four. We report the results in Table 4.7. Here we see the near optimal rate of convergence for both $L2$ and $H1$. In other respects, the data are quite similar to Table 4.6. Meshes corresponding to the adaptive meshes in Figure 4.3 are shown in Figure 4.4.

It is interesting to note that a refinement factor of four caused no problems in the case of a similar experiment performed in the case $p = 1$ in [3]. For $p = 2$ the refinement is much sharper in the region of the singularity, and it was this increase in sharpness that required a less aggressive refinement strategy. Even higher order

TABLE 4.7
Error estimates for adaptive refinement.

nt	$L2$	$\widetilde{L2}$	EF_0	$H1$	$\widetilde{H1}$	EF_1	$H2$	$\widetilde{H2}$	EF_2
8	0.1	3.9e-2	0.3	0.6	0.3	0.5	4.1	4.3	0.4
31	0.1	8.9e-3	0.1	0.5	0.1	0.3	6.4	6.5	0.5
74	0.1	4.9e-3	0.1	0.4	0.1	0.4	9.5	9.8	0.5
136	4.5e-2	1.8e-3	4.1e-2	0.3	0.1	0.3	14.0	15.0	0.5
302	2.5e-2	5.6e-4	2.2e-2	0.3	0.1	0.3	34.0	35.0	0.4
534	1.3e-2	2.7e-4	2.1e-2	0.2	0.1	0.4	87.0	89.0	0.4
1179	5.8e-3	7.8e-5	1.4e-2	0.1	4.1e-2	0.3	2.5e2	2.6e2	0.5
2077	2.8e-3	3.5e-5	1.3e-2	0.1	2.7e-2	0.3	7.0e2	7.3e2	0.5
4617	1.1e-3	1.2e-5	1.1e-2	5.0e-2	1.9e-2	0.4	3.1e3	3.3e3	0.5
8204	3.7e-4	5.1e-6	1.4e-2	3.0e-2	1.1e-2	0.4	1.4e4	1.4e4	0.5
18388	1.4e-4	1.5e-6	1.1e-2	1.8e-2	6.2e-3	0.4	6.4e4	6.7e4	0.5
32669	4.0e-5	6.7e-7	1.7e-2	9.5e-3	4.2e-3	0.4	3.8e5	4.0e5	0.6
73439	1.1e-5	2.0e-7	1.8e-2	5.0e-3	1.9e-3	0.4	2.6e6	2.7e6	0.5
130586	2.6e-6	8.5e-8	3.2e-2	2.5e-3	1.0e-3	0.4	2.0e7	2.2e7	0.5
Order	3.56	2.96		1.77	1.73		-5.21	-5.22	

elements are likely to require even more conservative refinement strategies to create meshes corresponding to optimal orders of convergence. Perhaps this adds another dimension to already complex general discussions evaluating the relative merits of higher order methods.

REFERENCES

[1] R.E. BANK, *PLTMG: A Software Package for Solving Elliptic Partial Differential Equations, Users' Guide* 9.0, Technical report, Department of Mathematics, University of California at San Diego, 2004.

[2] R.E. BANK AND J. XU, *Asymptotically exact a posteriori error estimators, Part* I: *Grids with superconvergence*, SIAM J. Numer. Anal., 41 (2003), pp. 2294–2312.

[3] R.E. BANK AND J. XU, *Asymptotically exact a posteriori error estimators, Part* II: *General unstructured grids*, SIAM J. Numer. Anal., 41 (2003), pp. 2313–2332.

[4] J.H. BRAMBLE, J.A. NITSCHE, AND A.H. SCHATZ, *Maximum-norm interior estimates for Ritz-Galerkin methods*, Math. Comp., 29 (1975), pp. 677–688.

[5] L. DU AND N. YAN, *Gradient recovery type a posteriori error estimate for finite element approximation on non-uniform meshes*, Adv. Comput. Math., 14 (2001), pp. 175–193.

[6] E. HINTON AND J.S. CAMPBELL, *Local and global smoothing of discontinuous finite element functions using a least square method*, Internat. J. Numer. Methods Engrg., 8 (1974), pp. 461–480.

[7] I. HLAVÁČEK, M. KŘÍŽEK, AND V. PIŠTORA, *How to recover the gradient of linear elements on nonuniform triangulations*, Appl. Math., 41 (1996), pp. 241–267.

[8] Y.Q. HUANG AND J. XU, *Superconvergence for Quadratic Triangular Finite Elements on Mildly Structured Grids*, preprint, 2005.

[9] B. LI, *Lagrange interpolation and finite element superconvergence*, Numer. Methods, Partial Differential Equations, 20 (2004), pp. 33–59.

[10] J.T. ODEN AND H.J. BRAUCHLI, *On the calculation of consistent stress distributions in finite element applications*, Internat. J. Numer. Methods Engrg., 3 (1971), pp. 317–325.

[11] J.S. OVALL, *Asymptotically exact functional error estimators based on superconvergent gradient recovery*, Numer. Math., 102 (2006), pp. 543–558.

[12] N.-E. WIBERG AND F. ABDULWAHAB, *Patch recovery based on superconvergent derivatives and equilibrium*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 2703–2724.

[13] N.-E. WIBERG, F. ABDULWAHAB, AND S. ZIUKAS, *Enhanced superconvergence patch recovery incoorporating equilibrium and boundary conditions*, Internat. J. Numer. Methods Engrg., 37 (1994), pp. 3417–3440.

[14] J. XU AND Z. ZHANG, *Analysis of recovery type a posteriori error estimators for mildly structured grids*, Math. Comp., 73 (2003), pp. 1139–1152.

[15] N. Yan and A. Zhou, *Gradient recovery type a posteriori error estimates for finite element approximations on irregular meshes*, Comput. Methods Appl. Mech. Engrg., 190 (2001), pp. 4289–4299.

[16] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.

[17] O.C. Zienkiewicz and J.Z. Zhu, *The superconvergence patch recovery and a posteriori error estimates part* 1: *The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

[18] O.C. Zienkiewicz and J.Z. Zhu, *The superconvergence patch recovery and a posteriori error estimates part* 2: *Error estimates and adaptivity*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1365–1382.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2047–2071

A FAMILY OF C^0 FINITE ELEMENTS FOR KIRCHHOFF PLATES I: ERROR ANALYSIS*

L. BEIRÃO DA VEIGA†, J. NIIRANEN‡, AND R. STENBERG‡

Abstract. A new finite element formulation for the Kirchhoff plate model is presented. The method is a displacement formulation with the deflection and the rotation vector as unknowns, and it is based on ideas stemming from a stabilized method for the Reissner–Mindlin model [R. Stenberg, in *Asymptotic Methods for Elastic Structures*, P. Ciarlet, L. Trabucho, and J. M. Viano, eds., de Gruyter, Berlin, 1995] and a method to treat a free boundary [P. Destuynder and T. Nevers, *RAIRO Modél. Math. Anal. Numér.*, 22 (1988), pp. 217–242]. Optimal a priori and a posteriori error estimates are derived.

Key words. finite elements, Kirchhoff plate model, free boundary, a priori error analysis, a posteriori error analysis

AMS subject classifications. 65N30, 74K20, 74S05

DOI. 10.1137/06067554X

1. Introduction.

A conforming finite element method for the Kirchhoff plate-bending problem requires a C^1-continuity and hence leads to methods that are rarely used in practice. Instead, either a nonconforming method is used or the model is abandoned in favor of the Reissner–Mindlin model. For the latter, there exist several families of methods that have rigorously been shown to be free from locking and optimally convergent.

A natural idea is to consider the Kirchhoff model as the limit of the Reissner–Mindlin model when the plate thickness approaches zero and to use a good Reissner–Mindlin element with the thickness (after a scaling, see below) representing the parameter when penalizing the Kirchhoff constraint. In this approach, there are two obstacles. First, for a free boundary, this leads to a method which is not consistent. This inconsistency significantly reduces the convergence rate of the method. In the literature, this point is often ignored since mostly the clamped case is considered. A remedy to this was developed by Destuynder and Nevers, who showed that the consistency is obtained by adding a term penalizing the tangential Kirchhoff condition along the free boundary [7]. Even if this modification has been done, there remains a second drawback. In order for the solution to the penalized formulation to be close to the exact solution, the penalty parameter should be large. This, however, leads to an ill-conditioned discrete system.

The free boundary inconsistency of the limit problem is closely related to the strong boundary layer of the Reissner–Mindlin plate problem with free boundaries. For Reissner–Mindlin plates, the presence of free boundaries significantly reduces the regularity of the solution and hence decreases the convergence rate of finite element approximations [1, 10, 5]. In [5, 2], the regularity of the solution has been improved by modifying the boundary conditions for free boundaries. These modifications imitate the boundary conditions of the Kirchhoff model as well as couple the variational

*Received by the editors November 21, 2006; accepted for publication (in revised form) May 7, 2007; published electronically September 19, 2007.
http://www.siam.org/journals/sinum/45-5/67554.html
†Dipartimento di Matematica "F. Enriques," Università di Milano, via Saldini 50, 20133 Milano, Italy (beirao@mat.unimi.it).
‡Institute of Mathematics, Helsinki University of Technology, P. O. Box 1100, 02015 TKK, Finland (jarkko.niiranen@tkk.fi, rolf.stenberg@tkk.fi).

spaces for the deflection and the rotation through the tangential Kirchhoff constraint along free boundaries. Adopting the modified boundary conditions on the discrete level it has been proved in [5, 2] that a set of finite element methods maintain their optimal order of convergence in the free boundary case. However, it can be seen as a drawback that all of these methods follow the mixed formulation with the shear force as an additional unknown. For positive values of the thickness parameter t, as usual, the corresponding displacement formulations can be achieved by condensing the shear force from the formulation. For the limit case $t = 0$, however, this possibility is excluded due to the nominator t^2 of the factor penalizing the Kirchhoff condition. For this reason, applying these methods for Kirchhoff plates requires a mixed formulation with the additional shear force degrees of freedom.

Our aim in the present paper is to present a family of Kirchhoff plate-bending elements which follows the displacement formulation and for which the convergence rate is optimal even in the presence of free boundaries. The method is a formulation combining the ideas from the stabilized method for Reissner–Mindlin plates presented in [13] and the treatment of the free boundary presented in [7]. Altough the method resembles the one with the linked interpolation technique in [2] for Reissner–Mindlin plates, it has been independently derived for the Kirchhoff plate problem with free boundaries. The family includes "simple low-order" elements, and it is well-conditioned. In the second part [3] of this paper, we give the results of numerical tests and a more detailed and constructive motivation for the method (cf. [4] as well).

The paper is organized as follows. In the next section, we describe the plate-bending problem, and in section 3, we introduce the new family of finite elements. In section 4, an a priori error analysis is derived. This analysis leads to optimal results, with respect both to the regularity of the solution and to the polynomial degree used. In section 5, an a posteriori error analysis is performed. We derive a local error indicator which is shown to be both reliable and efficient.

2. The Kirchhoff plate-bending problem. We consider the problem of bending of an isotropic linearly elastic plate and assume that the undeformed plate midsurface is described by a given convex polygonal domain $\Omega \subset \mathbb{R}^2$. The plate is considered to be clamped on the part Γ_{C} of its boundary $\partial\Omega$, simply supported on the part $\Gamma_{\mathrm{S}} \subset \partial\Omega$, and free on $\Gamma_{\mathrm{F}} \subset \partial\Omega$. The deflection and transversal load are denoted by w and g, respectively.

In what follows, we indicate with $\mathcal{V}$ the set of all corner points in Γ_{F}. Moreover, $\boldsymbol{n}$ and $\boldsymbol{s}$ represent the unit outward normal and the unit counterclockwise tangent to the boundary, respectively. Finally, for points $x \in \mathcal{V}$, we introduce the following notation. We indicate with $\boldsymbol{n}_1$ and $\boldsymbol{s}_1$ the unit vectors corresponding, respectively, to $\boldsymbol{n}$ and $\boldsymbol{s}$ on one of the two edges forming the boundary angle at x; with $\boldsymbol{n}_2$ and $\boldsymbol{s}_2$ we indicate the ones corresponding to the other edge. Note that which of the two edges correspond to the subscript 1 or 2 is not relevant.

The classical Kirchhoff plate-bending model is then given by the biharmonic partial differential equation

$$\mathsf{D}\Delta^2 w = g \qquad \text{in } \Omega, \tag{2.1}$$

the boundary conditions

$$\begin{aligned}
&w = 0, && \frac{\partial w}{\partial \boldsymbol{n}} = 0 && \text{on } \Gamma_{\mathrm{C}},\\
&w = 0, && \boldsymbol{n}\cdot\boldsymbol{M}\boldsymbol{n} = 0 && \text{on } \Gamma_{\mathrm{S}},\\
&\boldsymbol{n}\cdot\boldsymbol{M}\boldsymbol{n} = 0, && \frac{\partial}{\partial \boldsymbol{s}}\big(\boldsymbol{s}\cdot\boldsymbol{M}\boldsymbol{n}\big) + (\mathbf{div}\,\boldsymbol{M})\cdot\boldsymbol{n} = 0 && \text{on } \Gamma_{\mathrm{F}},
\end{aligned} \tag{2.2}$$

and the corner conditions

$$\big(\boldsymbol{s}_1 \cdot \boldsymbol{M}\boldsymbol{n}_1\big)(x) = \big(\boldsymbol{s}_2 \cdot \boldsymbol{M}\boldsymbol{n}_2\big)(x) \qquad \forall x \in \mathcal{V}. \tag{2.3}$$

Here

$$\mathsf{D} = \frac{\mathsf{E}t^3}{12(1-\nu^2)} \tag{2.4}$$

is the bending rigidity, with E, ν being the Young modulus and the Poisson ratio for the material, respectively. Note that for the shear modulus G it holds that

$$\mathsf{G} = \frac{\mathsf{E}}{2(1+\nu)}. \tag{2.5}$$

The moment tensor is given by

$$\boldsymbol{M}(\nabla w) = \mathsf{D}\big((1-\nu)\boldsymbol{\varepsilon}(\nabla w) + \nu \mathrm{div}(\nabla w)\boldsymbol{I}\big), \tag{2.6}$$

with the symmetric gradient $\boldsymbol{\varepsilon}$, and the shear force by

$$\boldsymbol{Q} = -\mathbf{div}\, \boldsymbol{M}. \tag{2.7}$$

Note that the independence of the Poisson ratio ν in the differential equation (2.1) is a consequence of cancellations when substituting (2.6) and (2.7) into the equilibrium equation

$$-\mathrm{div}\, \boldsymbol{Q} = g. \tag{2.8}$$

For the analysis below, it will be convenient to perform a scaling of the problem by assuming that the load is given by $g = \mathsf{G}t^3 f$, with f fixed. Then the differential equation (2.1) becomes independent of the plate thickness:

$$\frac{1}{6(1-\nu)}\Delta^2 w = f \qquad \text{in } \Omega. \tag{2.9}$$

Furthermore, we use the following scaled moment tensor $\boldsymbol{m}$:

$$\boldsymbol{M}(\nabla w) = \mathsf{G}t^3 \boldsymbol{m}(\nabla w), \tag{2.10}$$

and the shear force $\boldsymbol{q}$ is defined by

$$\boldsymbol{Q} = \mathsf{G}t^3 \boldsymbol{q}. \tag{2.11}$$

The unknowns in our finite element method will be the approximations to the deflection and its gradient, the rotation $\boldsymbol{\beta} = \nabla w$. With this as a new unknown, our problem can be written as the system of partial differential equations

$$\nabla w - \boldsymbol{\beta} = \mathbf{0}, \tag{2.12}$$

$$-\mathrm{div}\, \boldsymbol{q} = f, \tag{2.13}$$

$$\boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{q} = \mathbf{0} \qquad \text{in } \Omega, \tag{2.14}$$

the boundary conditions

$$w = 0,\ \boldsymbol{\beta} = \mathbf{0} \qquad \text{on } \Gamma_{\mathrm{C}}, \tag{2.15}$$

$$w = 0,\ \boldsymbol{\beta}\cdot\boldsymbol{s} = 0,\ \boldsymbol{n}\cdot\boldsymbol{m}(\boldsymbol{\beta})\boldsymbol{n} = 0 \qquad \text{on } \Gamma_{\mathrm{S}}, \tag{2.16}$$

$$\frac{\partial w}{\partial \boldsymbol{s}} - \boldsymbol{\beta}\cdot\boldsymbol{s} = 0,\ \boldsymbol{n}\cdot\boldsymbol{m}(\boldsymbol{\beta})\boldsymbol{n} = 0,\ \frac{\partial}{\partial \boldsymbol{s}}\big(\boldsymbol{s}\cdot\boldsymbol{m}(\boldsymbol{\beta})\boldsymbol{n}\big) - \boldsymbol{q}\cdot\boldsymbol{n} = 0 \text{ on } \Gamma_{\mathrm{F}}, \tag{2.17}$$

and the corner conditions

$$\big(\boldsymbol{s}_1\cdot\boldsymbol{m}(\boldsymbol{\beta})\boldsymbol{n}_1\big)(x) = \big(\boldsymbol{s}_2\cdot\boldsymbol{m}(\boldsymbol{\beta})\boldsymbol{n}_2\big)(x) \qquad \forall x\in\mathcal{V}. \tag{2.18}$$

The operator $\boldsymbol{L}$ is defined as

$$\boldsymbol{L}\boldsymbol{\beta} = \mathbf{div}\,\boldsymbol{m}(\boldsymbol{\beta}), \tag{2.19}$$

and the scaled bending moment is considered as a function of the rotation:

$$\boldsymbol{m}(\boldsymbol{\beta}) = \frac{1}{6}\left(\boldsymbol{\varepsilon}(\boldsymbol{\beta}) + \frac{\nu}{1-\nu}\mathrm{div}\,\boldsymbol{\beta}\,\boldsymbol{I}\right). \tag{2.20}$$

In what follows, we will often write $\boldsymbol{m}$ instead of $\boldsymbol{m}(\boldsymbol{\beta})$. We further denote

$$a(\boldsymbol{\beta},\boldsymbol{\eta}) = (\boldsymbol{m}(\boldsymbol{\beta}),\boldsymbol{\varepsilon}(\boldsymbol{\eta})). \tag{2.21}$$

In order to neglect plate rigid movements and the related technicalities, we will in what follows assume that the one-dimensional measure of Γ_{C} is positive.

3. The finite element formulation. In this section, we will introduce our finite element method. Even if our method is stable for all choices of finite element spaces, we will, for simplicity, present it for triangular elements and for the polynomial degrees that yield an optimal convergence rate. Hence, let a regular family of triangular meshes on Ω be given. For the integer $k\geq 1$, we then define the discrete spaces

$$W_h = \{v\in W \ \mid\ v_{|K}\in P_{k+1}(K)\ \ \forall K\in\mathcal{C}_h\}, \tag{3.1}$$

$$\boldsymbol{V}_h = \{\boldsymbol{\eta}\in\boldsymbol{V} \ \mid\ \boldsymbol{\eta}_{|K}\in[P_k(K)]^2\ \ \forall K\in\mathcal{C}_h\}, \tag{3.2}$$

with

$$W = \left\{v\in H^1(\Omega)\,|\,v = 0 \text{ on } \Gamma_{\mathrm{C}}\cup\Gamma_{\mathrm{S}}\right\}, \tag{3.3}$$

$$\boldsymbol{V} = \left\{\boldsymbol{\eta}\in[H^1(\Omega)]^2\,|\,\boldsymbol{\eta} = \boldsymbol{0} \text{ on } \Gamma_{\mathrm{C}},\ \boldsymbol{\eta}\cdot\boldsymbol{s} = 0 \text{ on } \Gamma_{\mathrm{S}}\right\}. \tag{3.4}$$

Here $\mathcal{C}_h$ represents the set of all triangles K of the mesh, and $P_k(K)$ is the space of polynomials of degree k on K. In what follows, we will indicate with h_K the diameter of each element K, while h will indicate the maximum size of all of the elements in the mesh. Furthermore, we will indicate with E a general edge of the triangulation and with h_E the length of E. The set of all edges lying on the free boundary Γ_{F} we denote by $\mathcal{F}_h$.

Before introducing the method, we state the following result which trivially follows from classical scaling arguments and the coercivity of the form a.

LEMMA 3.1. *There exist positive constants C_I and C_I' such that*

$$C_I\sum_{K\in\mathcal{C}_h} h_K^2\|\boldsymbol{L}\boldsymbol{\phi}\|_{0,K}^2 \leq a(\boldsymbol{\phi},\boldsymbol{\phi}) \qquad \forall\boldsymbol{\phi}\in\boldsymbol{V}_h, \tag{3.5}$$

$$C_I'\sum_{E\in\mathcal{F}_h} h_E\,\|m_{ns}(\boldsymbol{\phi})\|_{0,E}^2 \leq a(\boldsymbol{\phi},\boldsymbol{\phi}) \qquad \forall\boldsymbol{\phi}\in\boldsymbol{V}_h, \tag{3.6}$$

where the operator $m_{ns}(\boldsymbol{\phi}) = \boldsymbol{s} \cdot \boldsymbol{m}(\boldsymbol{\phi})\boldsymbol{n}$, *with* $\boldsymbol{n}, \boldsymbol{s}$, *being the unit outward normal and the unit counterclockwise tangent to the edge* E, *respectively, and with* $\boldsymbol{m}$ *defined in* (2.20).

Let two real numbers γ and α be assigned: $\gamma > 2/C_I'$ and $0 < \alpha < C_I/4$. Then the discrete problem reads as follows.

Method 3.1. Find $(w_h, \boldsymbol{\beta}_h) \in W_h \times \boldsymbol{V}_h$, such that

$$\mathcal{A}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) = (f, v) \qquad \forall (v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h, \tag{3.7}$$

where the form $\mathcal{A}_h$ is defined as

$$\mathcal{A}_h(z, \boldsymbol{\phi}; v, \boldsymbol{\eta}) = \mathcal{B}_h(z, \boldsymbol{\phi}; v, \boldsymbol{\eta}) + \mathcal{D}_h(z, \boldsymbol{\phi}; v, \boldsymbol{\eta}), \tag{3.8}$$

with

$$\begin{aligned} \mathcal{B}_h(z, \boldsymbol{\phi}; v, \boldsymbol{\eta}) &= a(\boldsymbol{\phi}, \boldsymbol{\eta}) - \sum_{K \in \mathcal{C}_h} \alpha h_K^2 (\boldsymbol{L}\boldsymbol{\phi}, \boldsymbol{L}\boldsymbol{\eta})_K \\ &\quad + \sum_{K \in \mathcal{C}_h} \frac{1}{\alpha h_K^2} (\nabla z - \boldsymbol{\phi} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\phi}, \nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K \end{aligned} \tag{3.9}$$

and

$$\begin{aligned} \mathcal{D}_h(z, \boldsymbol{\phi}; v, \boldsymbol{\eta}) &= \langle m_{ns}(\boldsymbol{\phi}), [\nabla v - \boldsymbol{\eta}] \cdot \boldsymbol{s} \rangle_{\Gamma_{\mathrm{F}}} + \langle [\nabla z - \boldsymbol{\phi}] \cdot \boldsymbol{s}, m_{ns}(\boldsymbol{\eta}) \rangle_{\Gamma_{\mathrm{F}}} \\ &\quad + \sum_{E \in \mathcal{F}_h} \frac{\gamma}{h_E} \langle [\nabla z - \boldsymbol{\phi}] \cdot \boldsymbol{s}, [\nabla v - \boldsymbol{\eta}] \cdot \boldsymbol{s} \rangle_E \end{aligned} \tag{3.10}$$

for all $(z, \boldsymbol{\phi}), (v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h$. Here $\langle \cdot, \cdot \rangle_{\Gamma_{\mathrm{F}}}$ and $\langle \cdot, \cdot \rangle_E$ denote the L^2-inner products on Γ_{F} and E, respectively.

The bilinear form $\mathcal{B}_h$ constitutes the Reissner–Mindlin method of [13] with the thickness t set equal to zero, while the additional form $\mathcal{D}_h$ is introduced in order to avoid the convergence deterioration in the presence of free boundaries.

Furthermore, we introduce the discrete shear force

$$\boldsymbol{q}_{h|K} = \frac{1}{\alpha h_K^2} (\nabla w_h - \boldsymbol{\beta}_h - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}_h)_{|K} \quad \forall K \in \mathcal{C}_h. \tag{3.11}$$

We note that, due to (2.14) and (2.12), it holds that

$$\boldsymbol{q}_{|K} = \frac{1}{\alpha h_K^2} (\nabla w - \boldsymbol{\beta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta})_{|K} \quad \forall K \in \mathcal{C}_h, \tag{3.12}$$

and hence it follows that the definition (3.11) is consistent with the exact shear force.

For simplicity, in the rest of this section we assume that the deflection w belongs to $H^3(\Omega)$; this is a very reasonable assumption, as discussed at the end of this section. Note as well that, with some additional technical work involving the appropriate Sobolev spaces and their duals, such an assumption could probably be avoided. The following result states the consistency of the method.

THEOREM 3.2. *The solution* $(w, \boldsymbol{\beta})$ *of the problem* (2.14)–(2.18) *satisfies*

$$\mathcal{A}_h(w, \boldsymbol{\beta}; v, \boldsymbol{\eta}) = (f, v) \qquad \forall (v, \boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h. \tag{3.13}$$

Proof. The definition of the bilinear forms in Method 3.1, recalling (2.14) and the expression (3.12), give

$$\begin{aligned}\mathcal{B}_h(w,\boldsymbol{\beta};v,\boldsymbol{\eta}) &= a(\boldsymbol{\beta},\boldsymbol{\eta}) - \sum_{K\in\mathcal{C}_h} \alpha h_K^2(\boldsymbol{L}\boldsymbol{\beta},\boldsymbol{L}\boldsymbol{\eta})_K \\ &\quad + \sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}(\nabla w - \boldsymbol{\beta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}, \nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K \\ &= a(\boldsymbol{\beta},\boldsymbol{\eta}) + \sum_{K\in\mathcal{C}_h} \alpha h_K^2(\boldsymbol{q},\boldsymbol{L}\boldsymbol{\eta})_K + \sum_{K\in\mathcal{C}_h} (\boldsymbol{q}, \nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K \\ &= a(\boldsymbol{\beta},\boldsymbol{\eta}) + (\boldsymbol{q}, \nabla v - \boldsymbol{\eta}). \qquad (3.14)\end{aligned}$$

First, by the definition (2.21), then integrating by parts on each triangle, and finally using the regularity of the functions involved, and the boundary conditions (2.15), (2.16) on Γ_{C}, Γ_{S}, respectively, we get

$$\begin{aligned}a(\boldsymbol{\beta},\boldsymbol{\eta}) + (\boldsymbol{q}, \nabla v - \boldsymbol{\eta}) &= (\boldsymbol{m}(\boldsymbol{\beta}), \boldsymbol{\varepsilon}(\boldsymbol{\eta})) + (\boldsymbol{q}, \nabla v - \boldsymbol{\eta}) \\ &= -(\boldsymbol{L}\boldsymbol{\beta} + \boldsymbol{q}, \boldsymbol{\eta}) + \langle \boldsymbol{m}(\boldsymbol{\beta})\cdot\boldsymbol{n}, \boldsymbol{\eta}\rangle_{\Gamma_{\mathrm{F}}} - (\operatorname{div}\boldsymbol{q}, v) + \langle \boldsymbol{q}\cdot\boldsymbol{n}, v\rangle_{\Gamma_{\mathrm{F}}}. \qquad (3.15)\end{aligned}$$

Recalling (2.14) and (2.13), the identity above becomes

$$a(\boldsymbol{\beta},\boldsymbol{\eta}) + (\boldsymbol{q}, \nabla v - \boldsymbol{\eta}) = (f, v) + \langle \boldsymbol{m}(\boldsymbol{\beta})\cdot\boldsymbol{n}, \boldsymbol{\eta}\rangle_{\Gamma_{\mathrm{F}}} + \langle \boldsymbol{q}\cdot\boldsymbol{n}, v\rangle_{\Gamma_{\mathrm{F}}}, \qquad (3.16)$$

while using the boundary conditions of (2.17) on Γ_{F} and integration by parts along the boundary finally leads to

$$a(\boldsymbol{\beta},\boldsymbol{\eta}) + (\boldsymbol{q}, \nabla v - \boldsymbol{\eta}) = (f, v) - \langle m_{ns}(\boldsymbol{\beta}), [\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_{\Gamma_{\mathrm{F}}}. \qquad (3.17)$$

Due to (2.17), we have

$$\begin{aligned}\mathcal{D}_h(w,\boldsymbol{\beta};v,\boldsymbol{\eta}) &= \langle m_{ns}(\boldsymbol{\beta}), [\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_{\Gamma_{\mathrm{F}}} + \langle [\nabla w - \boldsymbol{\beta}]\cdot\boldsymbol{s}, m_{ns}(\boldsymbol{\eta})\rangle_{\Gamma_{\mathrm{F}}} \\ &\quad + \sum_{E\in\mathcal{F}_h} \frac{\gamma}{h_E}\langle [\nabla w - \boldsymbol{\beta}]\cdot\boldsymbol{s}, [\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_E \\ &= \langle m_{ns}(\boldsymbol{\beta}), [\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_{\Gamma_{\mathrm{F}}}. \qquad (3.18)\end{aligned}$$

The result now directly follows from (3.14), (3.17), and (3.18). ∎

Remark 3.1. If the Reissner–Mindlin method of [13] without the additional form $\mathcal{D}_h$ is employed by setting $t = 0$, then in the presence of a free boundary we obtain

$$\mathcal{B}_h(w,\boldsymbol{\beta};v,\boldsymbol{\eta}) = (f, v) + \langle m_{ns}(\boldsymbol{\beta}), [\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_{\Gamma_{\mathrm{F}}} \qquad \forall (v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h. \qquad (3.19)$$

Therefore, this would lead to an inconsistent method. We return to this in Remark 4.1 below.

4. Stability and a priori error estimates. For $(v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h$, we introduce the following mesh-dependent norms:

$$|(v,\boldsymbol{\eta})|_h^2 = \sum_{K\in\mathcal{C}_h} h_K^{-2}\|\nabla v - \boldsymbol{\eta}\|_{0,K}^2, \qquad (4.1)$$

$$\|v\|_{2,h}^2 = \|v\|_1^2 + \sum_{K\in\mathcal{C}_h} |v|_{2,K}^2 + \sum_{E\in\mathcal{I}_h} h_E^{-1}\left\| \left[\!\left[\frac{\partial v}{\partial \boldsymbol{n}} \right]\!\right] \right\|_{0,E}^2 + \sum_{E\subset\Gamma_{\mathrm{C}}} h_E^{-1}\left\| \frac{\partial v}{\partial \boldsymbol{n}} \right\|_{0,E}^2, \qquad (4.2)$$

$$|||(v,\boldsymbol{\eta})|||_h = \|\boldsymbol{\eta}\|_1 + \|v\|_{2,h} + |(v,\boldsymbol{\eta})|_h, \qquad (4.3)$$

where $[\![\cdot]\!]$ represents the jump operator and $\mathcal{I}_h$ denotes the edges lying in the interior of the domain Ω.

In [12], the following lemma is proved.

LEMMA 4.1. *There exists a positive constant C such that*

$$\|v\|_{2,h} \le C\big(\|\boldsymbol{\eta}\|_1 + \|v\|_1 + |(v,\boldsymbol{\eta})|_h\big) \qquad \forall (v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h. \tag{4.4}$$

Using the Poincaré inequality and the previous lemma, the following equivalence easily follows.

LEMMA 4.2. *There exists a positive constant C such that*

$$C|||(v,\boldsymbol{\eta})|||_h \le \|\boldsymbol{\eta}\|_1 + |(v,\boldsymbol{\eta})|_h \le |||(v,\boldsymbol{\eta})|||_h \quad \forall (v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h. \tag{4.5}$$

We now have the following stability estimate.

THEOREM 4.3. *Let $0 < \alpha < C_I/4$ and $\gamma > 2/C_I'$. Then there exists a positive constant C such that*

$$\mathcal{A}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) \ge C|||(v,\boldsymbol{\eta})|||_h^2 \quad \forall (v,\boldsymbol{\eta}) \in W_h \times \boldsymbol{V}_h. \tag{4.6}$$

Proof. Using the first inverse estimate of Lemma 3.1 we get

$$\begin{aligned}
&\mathcal{B}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) \\
&\quad = a(\boldsymbol{\eta},\boldsymbol{\eta}) - \sum_{K\in\mathcal{C}_h} \alpha h_K^2 \|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 + \sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 \\
&\quad \ge \left(1 - \frac{\alpha}{C_I}\right) a(\boldsymbol{\eta},\boldsymbol{\eta}) + \sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2. \qquad (4.7)
\end{aligned}$$

Next, using locally the arithmetic-geometric mean inequality with the constant γ/h_E then the second inverse inequality of Lemma 3.1, we get

$$\begin{aligned}
&\mathcal{D}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) \\
&\quad = \sum_{E\in\mathcal{F}_h} \Big(2\langle m_{ns}(\boldsymbol{\eta}), [\nabla v - \boldsymbol{\eta}]\cdot \boldsymbol{s}\rangle_E + \frac{\gamma}{h_E}\|[\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\|_{0,E}^2\Big) \\
&\quad \ge \sum_{E\in\mathcal{F}_h} \Big(-\frac{\gamma}{h_E}\|[\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\|_{0,E}^2 - \gamma^{-1}h_E\,\|m_{ns}(\boldsymbol{\eta})\|_{0,E}^2 + \frac{\gamma}{h_E}\|[\nabla v - \boldsymbol{\eta}]\cdot\boldsymbol{s}\|_{0,E}^2\Big) \\
&\quad = -\sum_{E\in\mathcal{F}_h} \gamma^{-1}h_E\,\|m_{ns}(\boldsymbol{\eta})\|_{0,E}^2 \\
&\quad \ge -\frac{\gamma^{-1}}{C_I'}\,a(\boldsymbol{\eta},\boldsymbol{\eta}) \ge -\frac{1}{2}\,a(\boldsymbol{\eta},\boldsymbol{\eta}). \qquad (4.8)
\end{aligned}$$

Joining (4.7) with (4.8) and using Korn's inequality we then obtain

$$\begin{aligned}
&\mathcal{B}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) + \mathcal{D}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta}) \\
&\quad \ge \left(\frac{1}{2} - \frac{\alpha}{C_I}\right) a(\boldsymbol{\eta},\boldsymbol{\eta}) + \sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2 \\
&\quad \ge C\Big(\|\boldsymbol{\eta}\|_1^2 + \sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2\Big). \qquad (4.9)
\end{aligned}$$

From the triangle inequality, again the inverse estimate of Lemma 3.1, and the boundedness of the bilinear form a, it follows that

$$\begin{aligned}
&\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v-\boldsymbol{\eta}\|_{0,K}^2\\
&\quad\le 2\left(\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v-\boldsymbol{\eta}-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2\right)\\
&\quad\le 2\left(\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v-\boldsymbol{\eta}-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+\sum_{K\in\mathcal{C}_h} \alpha h_K^2\|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2\right)\\
&\quad\le C\left(\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v-\boldsymbol{\eta}-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+a(\boldsymbol{\eta},\boldsymbol{\eta})\right)\\
&\quad\le C\left(\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v-\boldsymbol{\eta}-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2+\|\boldsymbol{\eta}\|_1^2\right),
\end{aligned}\tag{4.10}$$

which combined with (4.9) gives

$$\mathcal{A}_h(v,\boldsymbol{\eta};v,\boldsymbol{\eta})\ge C\big(\|\boldsymbol{\eta}\|_1^2+|(v,\boldsymbol{\eta})|_h^2\big).\tag{4.11}$$

The result then follows from the norm equivalence of Lemma 4.2. ∎

We can now derive the error estimates for the method. We note that the assumptions of the theorem are supposed to be valid for the further results below as well and hence are not repeated in what follows.

THEOREM 4.4. *Let $0<\alpha<C_I/4$ and $\gamma>2/C_I'$. Let $(w,\boldsymbol{\beta})$ be the exact solution of the problem, and let $(w_h,\boldsymbol{\beta}_h)$ be the approximate solution obtained with Method* 3.1. *Suppose that $w\in H^{s+2}(\Omega)$, with $1\le s\le k$. Then it holds that*

$$|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_h\le Ch^s\|w\|_{s+2}.\tag{4.12}$$

Proof. Step 1. Let $(w_I,\boldsymbol{\beta}_I)\in W_h\times\boldsymbol{V}_h$ be the usual Lagrange interpolants to w and $\boldsymbol{\beta}$, respectively. Using first the stability result of Theorem 4.3 and then the consistency result of Theorem 3.2, one has the existence of a pair

$$(v,\boldsymbol{\eta})\in W_h\times\boldsymbol{V}_h,\qquad |||(v,\boldsymbol{\eta})|||_h\le C\tag{4.13}$$

such that

$$\begin{aligned}
|||(w_h-w_I,\boldsymbol{\beta}_h-\boldsymbol{\beta}_I)|||_h&\le\mathcal{A}_h(w_h-w_I,\boldsymbol{\beta}_h-\boldsymbol{\beta}_I;v,\boldsymbol{\eta})\\
&=\mathcal{A}_h(w-w_I,\boldsymbol{\beta}-\boldsymbol{\beta}_I;v,\boldsymbol{\eta}),
\end{aligned}\tag{4.14}$$

where we recall that $\mathcal{A}_h=\mathcal{B}_h+\mathcal{D}_h$.

Step 2. For the $\mathcal{B}_h$-part, we have

$$\begin{aligned}
\mathcal{B}_h(w-w_I,\boldsymbol{\beta}-\boldsymbol{\beta}_I;v,\boldsymbol{\eta})&=a(\boldsymbol{\beta}-\boldsymbol{\beta}_I,\boldsymbol{\eta})-\sum_{K\in\mathcal{C}_h}\alpha h_K^2(\boldsymbol{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_I),\boldsymbol{L}\boldsymbol{\eta})_K\\
&+\sum_{K\in\mathcal{C}_h}\frac{1}{\alpha h_K^2}(\nabla(w-w_I)-(\boldsymbol{\beta}-\boldsymbol{\beta}_I)-\alpha h_K^2\boldsymbol{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_I),\nabla v-\boldsymbol{\eta}-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta})_K.
\end{aligned}\tag{4.15}$$

Due to the first inverse inequality of Lemma 3.1, we get

$$\left(\sum_{K\in\mathcal{C}_h} h_K^2\|\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2\right)^{1/2} \le C|||(v,\boldsymbol{\eta})|||_h \tag{4.16}$$

and

$$\left(\sum_{K\in\mathcal{C}_h} \frac{1}{\alpha h_K^2}\|\nabla v-\boldsymbol{\eta}-\alpha h_K^2\boldsymbol{L}\boldsymbol{\eta}\|_{0,K}^2\right)^{1/2} \le C|||(v,\boldsymbol{\eta})|||_h. \tag{4.17}$$

Using these bounds in (4.15) and recalling (4.13), we obtain

$$\begin{aligned}&\mathcal{B}_h(w-w_I,\boldsymbol{\beta}-\boldsymbol{\beta}_I;v,\boldsymbol{\eta})\\ &\quad\le C\left(|||(w-w_I,\boldsymbol{\beta}-\boldsymbol{\beta}_I)|||_h+\left(\sum_{K\in\mathcal{C}_h} h_K^2|\boldsymbol{\beta}-\boldsymbol{\beta}_I|_{2,K}^2\right)^{1/2}\right).\end{aligned} \tag{4.18}$$

Substituting the definition of the norm (4.3) in (4.18), using the triangle inequality, and finally applying the classical interpolation estimates, it easily follows that

$$\mathcal{B}_h(w-w_I,\boldsymbol{\beta}-\boldsymbol{\beta}_I;v,\boldsymbol{\eta}) \le Ch^s\big(\|w\|_{s+2}+\|\boldsymbol{\beta}\|_{s+1}\big). \tag{4.19}$$

Step 3. For the $\mathcal{D}_h$-part in (4.14), we have, by the definition (3.10),

$$\begin{aligned}\mathcal{D}_h(w-w_I,\boldsymbol{\beta}-\boldsymbol{\beta}_I;v,\boldsymbol{\eta}) &= \langle m_{ns}(\boldsymbol{\beta}-\boldsymbol{\beta}_I),[\nabla v-\boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_{\Gamma_F}\\ &\quad+\langle[\nabla(w-w_I)-(\boldsymbol{\beta}-\boldsymbol{\beta}_I)]\cdot\boldsymbol{s},m_{ns}(\boldsymbol{\eta})\rangle_{\Gamma_F}\\ &\quad+\sum_{E\in\mathcal{F}_h}\frac{\gamma}{h_E}\langle[\nabla(w-w_I)-(\boldsymbol{\beta}-\boldsymbol{\beta}_I)]\cdot\boldsymbol{s},[\nabla v-\boldsymbol{\eta}]\cdot\boldsymbol{s}\rangle_E\\ &=: T_1+T_2+T_3.\end{aligned} \tag{4.20}$$

Scaling arguments give

$$\|[\nabla v-\boldsymbol{\eta}]\cdot\boldsymbol{s}\|_{0,E}^2 \le \|\nabla v-\boldsymbol{\eta}\|_{0,E}^2 \le Ch_{K(E)}^{-1}\|\nabla v-\boldsymbol{\eta}\|_{0,K(E)}^2 \tag{4.21}$$

for all $E\in\mathcal{F}_h$, where $K(E)$ is the triangle with E as an edge. The l^2-Cauchy–Schwarz inequality, the bound (4.21), and the norm definition (4.3) now give

$$\begin{aligned}T_1 &\le \left(\sum_{E\in\mathcal{F}_h} h_{K(E)}\|m_{ns}(\boldsymbol{\beta}-\boldsymbol{\beta}_I)\|_{0,E}^2\right)^{1/2}\left(\sum_{E\in\mathcal{F}_h} h_{K(E)}^{-1}\|[\nabla v-\boldsymbol{\eta}]\cdot\boldsymbol{s}\|_{0,E}^2\right)^{1/2}\\ &\le C\left(\sum_{E\in\mathcal{F}_h} h_{K(E)}\|m_{ns}(\boldsymbol{\beta}-\boldsymbol{\beta}_I)\|_{0,E}^2\right)^{1/2}|||(v,\boldsymbol{\eta})|||_h.\end{aligned} \tag{4.22}$$

Recalling the bound (4.13), classical polynomial interpolation properties give

$$T_1 \le C\left(\sum_{E\in\mathcal{F}_h} h_{K(E)}\|m_{ns}(\boldsymbol{\beta}-\boldsymbol{\beta}_I)\|_{0,E}^2\right)^{1/2} \le Ch^s\|\boldsymbol{\beta}\|_{s+1}. \tag{4.23}$$

Again, by scaling we have

$$\|m_{ns}(\boldsymbol{\eta})\|_{0,E}^2 \leq h_{K(E)}^{-1}|\boldsymbol{\eta}|_{1,K(E)}^2 \qquad \forall E \in \mathcal{F}_h. \tag{4.24}$$

The l^2-Cauchy–Schwarz inequality, this bound, and the norm definition (4.3) give

$$\begin{aligned} T_2 &\leq \left(\sum_{E\in\mathcal{F}_h} h_{K(E)}^{-1}\|\nabla(w-w_I)-(\boldsymbol{\beta}-\boldsymbol{\beta}_I)\|_{0,E}^2\right)^{1/2}\left(\sum_{E\in\mathcal{F}_h} h_{K(E)}\|m_{ns}(\boldsymbol{\eta})\|_{0,E}^2\right)^{1/2} \\ &\leq C\left(\sum_{E\in\mathcal{F}_h} h_{K(E)}^{-1}\|\nabla(w-w_I)-(\boldsymbol{\beta}-\boldsymbol{\beta}_I)\|_{0,E}^2\right)^{1/2} |||(v,\boldsymbol{\eta})|||_h. \end{aligned} \tag{4.25}$$

Recalling the bound (4.13), classical polynomial interpolation estimates give

$$\begin{aligned} T_2 &\leq C\left(\sum_{E\in\mathcal{F}_h} h_{K(E)}^{-1}\|\nabla(w-w_I)-(\boldsymbol{\beta}-\boldsymbol{\beta}_I)\|_{0,E}^2\right)^{1/2} \\ &\leq Ch^s\big(\|\boldsymbol{\beta}\|_{s+1}+\|w\|_{s+2}\big). \end{aligned} \tag{4.26}$$

The bound for T_3 follows by combining the same techniques used for T_1 and T_2; we get

$$T_3 \leq Ch^s\big(\|\boldsymbol{\beta}\|_{s+1}+\|w\|_{s+2}\big). \tag{4.27}$$

Now, joining all of the bounds (4.14), (4.19), (4.20), (4.23), (4.26), and (4.27) we obtain

$$|||(w_h-w_I,\boldsymbol{\beta}_h-\boldsymbol{\beta}_I)|||_h \leq Ch^s\big(\|\boldsymbol{\beta}\|_{s+1}+\|w\|_{s+2}\big). \tag{4.28}$$

The triangle inequality and the classical polynomial interpolation estimates (recalling that $\boldsymbol{\beta}=\nabla w$) then yield

$$|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_h \leq Ch^s\big(\|\boldsymbol{\beta}\|_{s+1}+\|w\|_{s+2}\big) \leq Ch^s\|w\|_{s+2}. \tag{4.29}$$

Note that the result holds for real values of the regularity parameter s since the interpolation results used above are valid for real values of s. ∎

Remark 4.1. As noted in Remark 3.1, the limiting Reissner–Mindlin method (i.e., without the additional correction $\mathcal{D}_h$) is inconsistent. Regardless of the solution regularity and the polynomial degree k, the inconsistency term can be bounded only with the order $O(h^{1/2})$. As is well known (see, for example, [10]), the inconsistency error is a lower bound for the error of finite element methods. As a consequence, the numerical scheme will not converge with a rate better than $h^{1/2}$ if $\Gamma_{\mathrm{F}} \neq \emptyset$. This observation is also confirmed by the numerical tests shown in [3]. See [6] for other numerical tests regarding this issue. Note further that this boundary inconsistency term is connected not only to the formulation in [13] but is common to any other Kirchhoff method which follows a "Reissner–Mindlin limit" approach.

For the shear force, the practical norm to use is the discrete negative norm

$$\|\boldsymbol{r}\|_{-1,h} = \left(\sum_{K\in\mathcal{C}_h} h_K^2\|\boldsymbol{r}\|_{0,K}^2\right)^{1/2}. \tag{4.30}$$

Since we assume that $w \in H^{s+2}(\Omega)$, with $s \geq 1$, we have $\boldsymbol{q} \in [L^2(\Omega)]^2$, and from the estimates above the lemma immediately follows.

LEMMA 4.5. *It holds that*

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,h} \leq Ch^s \|w\|_{s+2}. \tag{4.31}$$

From this follows a norm estimate in the dual to the space

$$\boldsymbol{V}_* = \left\{ \boldsymbol{\eta} \in [H^1(\Omega)]^2 \mid \boldsymbol{\eta} = \boldsymbol{0} \text{ on } \Gamma_{\mathrm{C}},\ \boldsymbol{\eta} \cdot \boldsymbol{s} = 0 \text{ on } \Gamma_{\mathrm{F}} \cup \Gamma_{\mathrm{S}} \right\}, \tag{4.32}$$

i.e., in the norm

$$\|\boldsymbol{r}\|_{-1,*} = \sup_{\boldsymbol{\eta} \in \boldsymbol{V}_*} \frac{\langle \boldsymbol{r}, \boldsymbol{\eta} \rangle}{\|\boldsymbol{\eta}\|_1}. \tag{4.33}$$

We have the following result.

LEMMA 4.6. *It holds that*

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,*} \leq Ch^s \|w\|_{s+2}. \tag{4.34}$$

Proof. The proof is essentially an application of the "Pitkäranta–Verfürth trick" (see [11, 14]). By the definition of the norm $\|\cdot\|_{-1,*}$ there exists a function $\boldsymbol{\eta} \in \boldsymbol{V}_*$ such that

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,*} \leq (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\eta}), \qquad \|\boldsymbol{\eta}\|_1 \leq C. \tag{4.35}$$

Using a Clément-type interpolant we can find a piecewise linear function $\boldsymbol{\eta}_I \in \boldsymbol{V}_*$ such that it holds that

$$h_K^{s-1} \|\boldsymbol{\eta} - \boldsymbol{\eta}_I\|_{s,K} \leq C \|\boldsymbol{\eta}\|_{1,K} \leq C', \quad s = 0, 1, \tag{4.36}$$

for all $K \in \mathcal{C}_h$. Using the Cauchy–Schwarz inequality, the bound (4.36) with $s = 0$, and the definition (4.30), it follows that

$$\begin{aligned} (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\eta}) &= (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\eta} - \boldsymbol{\eta}_I) + (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\eta}_I) \\ &\leq C \|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,h} + (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\eta}_I). \end{aligned} \tag{4.37}$$

Note that $\boldsymbol{\eta}_I$ is in both $\boldsymbol{V}_h$ and $\boldsymbol{V}_*$; moreover, $\boldsymbol{L}\boldsymbol{\eta}_I = \boldsymbol{0}$ on each element K of $\mathcal{C}_h$. As a consequence, using (3.7), (3.11), (3.12), and Theorem 3.2, it follows that

$$\begin{aligned} (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\eta}_I) &= a(\boldsymbol{\beta} - \boldsymbol{\beta}_h, \boldsymbol{\eta}_I) + \langle [\nabla w_h - \boldsymbol{\beta}_h)] \cdot \boldsymbol{s}, M_{ns}(\boldsymbol{\eta}_I) \rangle_{\Gamma_{\mathrm{F}}} \\ &=: T_1 + T_2. \end{aligned} \tag{4.38}$$

Due to the continuity of the bilinear form and using bound (4.36) with $s = 1$, it immediately follows that

$$T_1 \leq C \|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_1 \leq C |||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_h. \tag{4.39}$$

Using first the Cauchy–Schwarz inequality, then the Agmon inequality, and finally the bound (4.36) with $s = 1$, Lemma 3.1, and the definition (4.3), we get

$$\begin{aligned} T_2 &\leq \left(\sum_{E \in \mathcal{F}_h} h_E^{-1} \|\nabla w_h - \boldsymbol{\beta}_h)\|_{0,E}^2 \right)^{1/2} \left(\sum_{E \in \mathcal{F}_h} h_E \|M_{ns}(\boldsymbol{\eta}_I)\|_{0,E}^2 \right)^{1/2} \\ &\leq \left(\sum_{K \in \mathcal{C}_h} h_K^{-2} \|\nabla w_h - \boldsymbol{\beta}_h)\|_{0,K}^2 \right)^{1/2} \|\boldsymbol{\eta}_I\|_1 \\ &\leq C |||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_h, \end{aligned} \tag{4.40}$$

where in the last inequality we implicitly used the relation $\nabla w - \boldsymbol{\beta} = \mathbf{0}$. Combining (4.35), (4.37) with (4.38), (4.39), and (4.40), it follows that

$$\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,*} \leq C\big(\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,h} + |||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_h\big). \tag{4.41}$$

Joining (4.41) and (4.31) and using Theorem 4.4 the proposition immediately follows. ∎

The regularity of the solution to the Kirchhoff plate problems for convex polygonal domains, with all three main types of boundary conditions, is very case-dependent. We refer, for example, to the work [9], in which a rather complete study is accomplished. Note that if $f \in H^{-1}(\Omega)$, in most cases of interest, the regularity condition $w \in H^3(\Omega)$ is indeed achieved.

Note further that with classical duality arguments and technical calculations it is possible to derive the error bound

$$\|w - w_h\|_1 \leq Ch^{s+1}\|w\|_{s+2}, \tag{4.42}$$

if the regularity estimate

$$\|w\|_3 \leq C\|f\|_{-1} \tag{4.43}$$

holds. Moreover, if $k \geq 2$ and the regularity estimate

$$\|w\|_4 \leq C\|f\|_0 \tag{4.44}$$

is satisfied, then it holds that

$$\|w - w_h\|_0 \leq Ch^{s+2}\|w\|_{s+2}. \tag{4.45}$$

5. A posteriori error estimates. In this section, we prove the reliability and the efficiency for an a posteriori error estimator for our method. To this end, we introduce

$$\tilde{\eta}_K^2 := h_K^4\|f + \operatorname{div} \boldsymbol{q}_h\|_{0,K}^2 + h_K^{-2}\|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K}^2, \tag{5.1}$$

$$\eta_E^2 := h_E^3\|[\![\boldsymbol{q}_h \cdot \boldsymbol{n}]\!]\|_{0,E}^2 + h_E\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}^2, \tag{5.2}$$

$$\eta_{S,E}^2 := h_E\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}^2, \tag{5.3}$$

$$\eta_{F,E}^2 := h_E\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}^2 + h_E^3\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h) - \boldsymbol{q}_h \cdot \boldsymbol{n}\right\|_{0,E}^2, \tag{5.4}$$

where h_E denotes the length of the edge E and $[\![\cdot]\!]$ represents the jump operator (which is assumed to be equal to the function value on boundary edges). Further, for a triangle $K \in \mathcal{C}_h$ we denote the sets of edges lying in the interior of Ω, on Γ_{S}, and on Γ_{F}, by $I(K)$, $S(K)$, and $F(K)$, respectively. By $\mathcal{S}_h$ we denote the set of all edges on Γ_{S} and by $\mathcal{I}_h$ the ones lying in the interior of the domain.

Given any element $K \in \mathcal{C}_h$, let the local error indicator be

$$\eta_K := \left(\tilde{\eta}_K^2 + \frac{1}{2}\sum_{E\in I(K)}\eta_E^2 + \sum_{E\in S(K)}\eta_{S,E}^2 + \sum_{E\in F(K)}\eta_{F,E}^2\right)^{1/2}. \tag{5.5}$$

Finally, the global error indicator is defined as

$$\eta := \left(\sum_{K\in\mathcal{C}_h}\eta_K^2\right)^{1/2}. \tag{5.6}$$

Remark 5.1. It is worth noting that, by the definition (3.11),

$$(\boldsymbol{q}_h + \boldsymbol{L}\boldsymbol{\beta}_h)_{|K} = \frac{1}{\alpha h_K^2}(\nabla w_h - \boldsymbol{\beta}_h)_{|K} \qquad \forall K \in \mathcal{C}_h, \tag{5.7}$$

which is the reason why there appear no terms of the kind $\|\boldsymbol{q}_h + \boldsymbol{L}\boldsymbol{\beta}_h\|_{0,K}$ in the error estimator. We note as well that scaling arguments give

$$\sum_{E\in\mathcal{F}_h} h_E^{-1}\|\nabla w_h - \boldsymbol{\beta}_h\|_{0,E}^2 \le C \sum_{K\in\mathcal{C}_h} h_K^{-2}\|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K}^2, \tag{5.8}$$

which is the reason why there appear no boundary terms of the kind $\|\nabla w_h - \boldsymbol{\beta}_h\|_{0,E}$.

5.1. Upper bound. In order to derive the reliability of the method we need the following saturation assumption.

Assumption 5.1. Given a mesh $\mathcal{C}_h$, let $\mathcal{C}_{h/2}$ be the mesh obtained by splitting each triangle $K \in \mathcal{C}_h$ into four triangles connecting the edge midpoints. Let $(w_{h/2}, \boldsymbol{\beta}_{h/2})$ be the discrete solution corresponding to the mesh $\mathcal{C}_{h/2}$. We assume that there exists a constant ρ, $0 < \rho < 1$, such that

$$\begin{aligned} &|||(w - w_{h/2}, \boldsymbol{\beta} - \boldsymbol{\beta}_{h/2})|||_{h/2} + \|\boldsymbol{q} - \boldsymbol{q}_{h/2}\|_{-1,*} \\ &\qquad \le \rho\big(|||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_h + \|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,*}\big), \end{aligned} \tag{5.9}$$

where by $|||\cdot|||_{h/2}$ we indicate the mesh-dependent norm with respect to the new mesh $\mathcal{C}_{h/2}$.

In what follows, we will need the following result.

LEMMA 5.1. *Let, for $v \in W_{h/2}$, the local seminorm be*

$$|v|_{2,h/2,K} = \left(\sum_{K'\in\mathcal{C}_{h/2},\ K'\subset K} |v|_{2,K'}^2\right)^{1/2}. \tag{5.10}$$

Then there is a positive constant C such that for all $v \in W_{h/2}$ there exists $v_I \in W_h$ with the bound

$$\|v - v_I\|_{0,K} + h_K^{1/2}\|v - v_I\|_{0,\partial K} \le C h_K^2 |v|_{2,h/2,K} \qquad \forall K \in \mathcal{C}_h. \tag{5.11}$$

Moreover, v_I interpolates v at all of the vertices of the triangulation $\mathcal{C}_{h/2}$.

Proof. We choose v_I as the only function in $H^1(\Omega)$ such that

$$\begin{aligned} &v_{I|K} \in P_2(K) \qquad \forall K \in \mathcal{C}_h, \\ &v_I(x) = v(x) \qquad \forall x \in \mathcal{V}_{h/2}, \end{aligned} \tag{5.12}$$

where $\mathcal{V}_{h/2}$ represents the set of all of the vertices of $\mathcal{C}_{h/2}$. Note that it is trivial to check that $v_I \in W_h$ for all $k \ge 1$. Observing that

$$|v|_{2,h/2,K} + \sum_{x\in\mathcal{V}_{h/2}\cap K} |v(x)|, \qquad v \in W_{h/2}, K \in \mathcal{C}_h, \tag{5.13}$$

is indeed a norm on the finite-dimensional space of the functions $v \in W_{h/2}$ restricted to K, the result follows applying the classical scaling argument. ∎

For simplicity, in what follows we will treat the case $\Gamma_{\mathrm{S}} = \emptyset$, the general case following with identical arguments as the ones that follow. We have the following preliminary result.

THEOREM 5.2. *It holds that*

$$|||(w_{h/2} - w_h, \boldsymbol{\beta}_{h/2} - \boldsymbol{\beta}_h)|||_{h/2} \leq C\eta. \tag{5.14}$$

Proof. Step 1. Due to the stability of the discrete formulation, proved in Theorem 4.3, there exists a couple $(v, \boldsymbol{\eta}) \in W_{h/2} \times \boldsymbol{V}_{h/2}$ such that

$$|||(v, \boldsymbol{\eta})|||_{h/2} \leq C \tag{5.15}$$

and

$$|||(w_{h/2} - w_h, \boldsymbol{\beta}_{h/2} - \boldsymbol{\beta}_h)|||_{h/2} \leq \mathcal{A}_{h/2}(w_{h/2} - w_h, \boldsymbol{\beta}_{h/2} - \boldsymbol{\beta}_h; v, \boldsymbol{\eta}). \tag{5.16}$$

Furthermore, we have

$$\mathcal{A}_{h/2}(w_{h/2}, \boldsymbol{\beta}_{h/2}; v, \boldsymbol{\eta}) = (f, v). \tag{5.17}$$

Step 2. Simple calculations and the definition (3.11) give

$$\begin{aligned}
\mathcal{B}_{h/2}(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) &= a(\boldsymbol{\beta}_h, \boldsymbol{\eta}) - \sum_{K \in \mathcal{C}_{h/2}} \alpha h_K^2 (\boldsymbol{L}\boldsymbol{\beta}_h, \boldsymbol{L}\boldsymbol{\eta})_K \\
&\quad + \sum_{K \in \mathcal{C}_{h/2}} \frac{1}{\alpha h_K^2} (\nabla w_h - \boldsymbol{\beta}_h - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}_h, \nabla v - \boldsymbol{\eta} - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\eta})_K \\
&= a(\boldsymbol{\beta}_h, \boldsymbol{\eta}) - \sum_{K \in \mathcal{C}_{h/2}} (\nabla w_h - \boldsymbol{\beta}_h, \boldsymbol{L}\boldsymbol{\eta})_K + \sum_{K \in \mathcal{C}_{h/2}} (\boldsymbol{q}_h, \nabla v - \boldsymbol{\eta})_K \\
&\quad + R_1(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) \\
&= \mathcal{B}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) + R_1(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}),
\end{aligned} \tag{5.18}$$

where $\boldsymbol{q}_h$ is defined as in (3.11), i.e., based on the coarser mesh, and

$$\begin{aligned}
R_1(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) &= \sum_{K \in \mathcal{C}_{h/2}} \frac{1}{\alpha h_K^2} (\nabla w_h - \boldsymbol{\beta}_h, \nabla v - \boldsymbol{\eta})_K \\
&\quad - \sum_{K \in \mathcal{C}_h} \frac{1}{\alpha h_K^2} (\nabla w_h - \boldsymbol{\beta}_h, \nabla v - \boldsymbol{\eta})_K.
\end{aligned} \tag{5.19}$$

The last term on the right-hand side is well defined since $\nabla v - \boldsymbol{\eta}$ is piecewise L^2-regular.

Let now $\mathcal{F}_{h/2}$ indicate the set of all edges of $\mathcal{C}_{h/2}$ lying on Γ_{F}. Adding and subtracting the difference between the two forms, it then follows that

$$\mathcal{D}_{h/2}(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) = \mathcal{D}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) + R_2(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}), \tag{5.20}$$

where

$$\begin{aligned}
R_2(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) &= \sum_{E \in \mathcal{F}_{h/2}} \frac{\gamma}{h_E} \langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, [\nabla v - \boldsymbol{\eta}] \cdot \boldsymbol{s} \rangle_E \\
&\quad - \sum_{E \in \mathcal{F}_h} \frac{\gamma}{h_E} \langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, [\nabla v - \boldsymbol{\eta}] \cdot \boldsymbol{s} \rangle_E
\end{aligned} \tag{5.21}$$

and where the first member on the right-hand side is indeed well defined due to the piecewise regularity of $(v, \boldsymbol{\eta})$. We will denote

$$R(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) = R_1(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) + R_2(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}). \tag{5.22}$$

Joining (5.17)–(5.21) then yields

$$\mathcal{A}_{h/2}(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) = \mathcal{A}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) + R(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}). \tag{5.23}$$

Step 3. Let $v_I \in W_h$ be the interpolant defined in Lemma 5.1, and let $\boldsymbol{\eta}_I \in \boldsymbol{V}_h$ be the piecewise linear interpolant to $\boldsymbol{\eta}$. First, we have

$$\mathcal{A}_h(w_h, \boldsymbol{\beta}_h; v_I, \boldsymbol{\eta}_I) = (f, v_I). \tag{5.24}$$

This, together with (5.17) and (5.23), gives

$$\begin{aligned}
&\mathcal{A}_{h/2}(w_{h/2} - w_h, \boldsymbol{\beta}_{h/2} - \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) \\
&\quad = \mathcal{A}_{h/2}(w_{h/2}, \boldsymbol{\beta}_{h/2}; v, \boldsymbol{\eta}) - \mathcal{A}_{h/2}(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) \\
&\quad = \mathcal{A}_{h/2}(w_{h/2}, \boldsymbol{\beta}_{h/2}; v, \boldsymbol{\eta}) - \mathcal{A}_h(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) - R(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}) \\
&\quad = (f, v - v_I) - \mathcal{A}_h(w_h, \boldsymbol{\beta}_h; v - v_I, \boldsymbol{\eta} - \boldsymbol{\eta}_I) - R(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta}). \qquad (5.25)
\end{aligned}$$

Step 4. Next, we bound the last terms above. Recalling that $\mathcal{C}_{h/2}$ is a subdivision of $\mathcal{C}_h$, the Cauchy–Schwarz inequality, (4.3), and (5.15) give

$$\begin{aligned}
|R_1(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta})| &\le 2 \left| \sum_{K \in \mathcal{C}_{h/2}} \frac{1}{\alpha h_K^2} (\nabla w_h - \boldsymbol{\beta}_h, \nabla v - \boldsymbol{\eta})_K \right| \\
&\le 2 \left(\sum_{K \in \mathcal{C}_{h/2}} \frac{1}{h_K^2} \|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K}^2 \right)^{1/2} \left(\sum_{K \in \mathcal{C}_{h/2}} \frac{1}{h_K^2} \|\nabla v - \boldsymbol{\eta}\|_{0,K}^2 \right)^{1/2} \\
&\le C \left(\sum_{K \in \mathcal{C}_{h/2}} \frac{1}{h_K^2} \|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K}^2 \right)^{1/2}. \qquad (5.26)
\end{aligned}$$

Using scaling and arguments similar to those already adopted in (5.26) it can be checked that

$$|R_2(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta})| \le C \left(\sum_{K \in \mathcal{C}_{h/2}} \frac{1}{h_K^2} \|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K}^2 \right)^{1/2}. \tag{5.27}$$

Combining (5.26) and (5.27) we get

$$|R(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta})| \le |R_1(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta})| + |R_2(w_h, \boldsymbol{\beta}_h; v, \boldsymbol{\eta})| \le C\eta. \tag{5.28}$$

Step 5. Next, we expand, substitute the expression (3.11) for $\boldsymbol{q}_h$, and regroup the terms:

$$
\begin{aligned}
&(f, v - v_I) - \mathcal{A}_h(w_h, \boldsymbol{\beta}_h; v - v_I, \boldsymbol{\eta} - \boldsymbol{\eta}_I) \\
&= (f, v - v_I) - \Big\{ a(\boldsymbol{\beta}_h, \boldsymbol{\eta} - \boldsymbol{\eta}_I) - \sum_{K \in \mathcal{C}_h} \alpha h_K^2 \big(\boldsymbol{L}\boldsymbol{\beta}_h, \boldsymbol{L}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big)_K \\
&\qquad + \sum_{K \in \mathcal{C}_h} \frac{1}{\alpha h_K^2} \big(\nabla w_h - \boldsymbol{\beta}_h - \alpha h_K^2 \boldsymbol{L}\boldsymbol{\beta}_h, \nabla(v - v_I) - (\boldsymbol{\eta} - \boldsymbol{\eta}_I) - \alpha h_K^2 \boldsymbol{L}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big)_K \\
&\qquad + \big\langle m_{ns}(\boldsymbol{\beta}_h), [\nabla(v - v_I) - (\boldsymbol{\eta} - \boldsymbol{\eta}_I)] \cdot \boldsymbol{s} \big\rangle_{\Gamma_{\mathrm{F}}} + \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, m_{ns}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big\rangle_{\Gamma_{\mathrm{F}}} \\
&\qquad + \sum_{E \in \mathcal{F}_h} \frac{\gamma}{h_E} \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, [\nabla(v - v_I) - (\boldsymbol{\eta} - \boldsymbol{\eta}_I)] \cdot \boldsymbol{s} \big\rangle_E \Big\} \\
&= (f, v - v_I) - \Big\{ a(\boldsymbol{\beta}_h, \boldsymbol{\eta} - \boldsymbol{\eta}_I) - \sum_{K \in \mathcal{C}_h} \alpha h_K^2 \big(\boldsymbol{L}\boldsymbol{\beta}_h + \boldsymbol{q}_h, \boldsymbol{L}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big)_K \\
&\qquad + \big(\boldsymbol{q}_h, \nabla(v - v_I) - (\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big) \\
&\qquad + \big\langle m_{ns}(\boldsymbol{\beta}_h), [\nabla(v - v_I) - (\boldsymbol{\eta} - \boldsymbol{\eta}_I)] \cdot \boldsymbol{s} \big\rangle_{\Gamma_{\mathrm{F}}} + \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, m_{ns}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big\rangle_{\Gamma_{\mathrm{F}}} \\
&\qquad + \sum_{E \in \mathcal{F}_h} \frac{\gamma}{h_E} \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, [\nabla(v - v_I) - (\boldsymbol{\eta} - \boldsymbol{\eta}_I)] \cdot \boldsymbol{s} \big\rangle_E \Big\} \\
&= \Big\{ (f, v - v_I) - \big(\boldsymbol{q}_h, \nabla(v - v_I) \big) - \big\langle m_{ns}(\boldsymbol{\beta}_h), [\nabla(v - v_I)] \cdot \boldsymbol{s} \big\rangle_{\Gamma_{\mathrm{F}}} \\
&\qquad - \sum_{E \in \mathcal{F}_h} \frac{\gamma}{h_E} \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, [\nabla(v - v_I)] \cdot \boldsymbol{s} \big\rangle_E \Big\} \\
&\qquad - \Big\{ a(\boldsymbol{\beta}_h, \boldsymbol{\eta} - \boldsymbol{\eta}_I) - \sum_{K \in \mathcal{C}_h} \alpha h_K^2 \big(\boldsymbol{L}\boldsymbol{\beta}_h + \boldsymbol{q}_h, \boldsymbol{L}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big)_K - \big(\boldsymbol{q}_h, \boldsymbol{\eta} - \boldsymbol{\eta}_I \big) \\
&\qquad - \big\langle m_{ns}(\boldsymbol{\beta}_h), [\boldsymbol{\eta} - \boldsymbol{\eta}_I] \cdot \boldsymbol{s} \big\rangle_{\Gamma_{\mathrm{F}}} + \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, m_{ns}(\boldsymbol{\eta} - \boldsymbol{\eta}_I) \big\rangle_{\Gamma_{\mathrm{F}}} \\
&\qquad - \sum_{E \in \mathcal{F}_h} \frac{\gamma}{h_E} \big\langle [\nabla w_h - \boldsymbol{\beta}_h] \cdot \boldsymbol{s}, [\boldsymbol{\eta} - \boldsymbol{\eta}_I] \cdot \boldsymbol{s} \big\rangle_E \Big\} \\
&=: A - B.
\end{aligned}
\tag{5.29}
$$

Step 6. In the part A above, integration by parts and using the fact that $v(x) = v_I(x)$ at the corner points $x \in \mathcal{V}$ yields

$$
\begin{aligned}
&(f, v - v_I) - \big(\boldsymbol{q}_h, \nabla(v - v_I) \big) - \big\langle m_{ns}(\boldsymbol{\beta}_h), [\nabla(v - v_I)] \cdot \boldsymbol{s} \big\rangle_{\Gamma_{\mathrm{F}}} \\
&\qquad = (f + \operatorname{div} \boldsymbol{q}_h, v - v_I) + \left\langle \frac{\partial}{\partial \boldsymbol{s}} m_{ns}(\boldsymbol{\beta}_h) - \boldsymbol{q}_h \cdot \boldsymbol{n}, v - v_I \right\rangle_{\Gamma_{\mathrm{F}}}.
\end{aligned}
\tag{5.30}
$$

The separate terms are then estimated as follows, using the Cauchy–Schwarz inequal-

ity and Lemma 5.1:

$$\begin{aligned}
\big|(f+\operatorname{div}\boldsymbol{q}_h, v-v_I)\big| &= \left|\sum_{K\in\mathcal{C}_h}(f_h+\operatorname{div}\boldsymbol{q}_h, v-v_I)_K\right| \\
&\le \left(\sum_{K\in\mathcal{C}_h} h_K^4\|f+\operatorname{div}\boldsymbol{q}_h\|_{0,K}^2\right)^{1/2}\left(\sum_{K\in\mathcal{C}_h} h_K^{-4}\|v-v_I\|_{0,K}^2\right)^{1/2} \\
&\le C\left(\sum_{K\in\mathcal{C}_h} h_K^4\|f+\operatorname{div}\boldsymbol{q}_h\|_{0,K}\right)^{1/2}\left(\sum_{K\in\mathcal{C}_h}|v|_{2,h/2,K}^2\right)^{1/2} \\
&\le C\left(\sum_{K\in\mathcal{C}_h}\tilde{\eta}_K^2\right)^{1/2} \qquad (5.31)
\end{aligned}$$

and

$$\begin{aligned}
\left|\left\langle\frac{\partial}{\partial\boldsymbol{s}}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}, v-v_I\right\rangle_{\Gamma_F}\right| &= \left|\sum_{E\in\mathcal{F}_h}\left\langle\frac{\partial}{\partial\boldsymbol{s}}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}, v-v_I\right\rangle_E\right| \\
&\le \left(\sum_{E\in\mathcal{F}_h} h_E^3\left\|\frac{\partial}{\partial\boldsymbol{s}}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E}^2\right)^{1/2}\left(\sum_{E\in\mathcal{F}_h} h_E^{-3}\|v-v_I\|_{0,E}^2\right)^{1/2} \\
&\le C\left(\sum_{E\in\mathcal{F}_h}\eta_{F,E}^2\right)^{1/2}\left(\sum_{K\in\mathcal{C}_h}|v|_{2,h/2,K}^2\right)^{1/2} \\
&\le C\left(\sum_{E\in\mathcal{F}_h}\eta_{F,E}^2\right)^{1/2}. \qquad (5.32)
\end{aligned}$$

The last term in A is readily estimated by scaling estimates and Lemma 5.1:

$$\begin{aligned}
&\left|\sum_{E\in\mathcal{F}_h}\frac{\gamma}{h_E}\big\langle[\nabla w_h-\boldsymbol{\beta}_h]\cdot\boldsymbol{s}, [\nabla(v-v_I)]\cdot\boldsymbol{s}\big\rangle_E\right| \\
&\quad\le \left(\sum_{E\in\mathcal{F}_h} h_E^{-1}\|\nabla w_h-\boldsymbol{\beta}_h\|_{0,E}^2\right)^{1/2}\left(\sum_{E\in\mathcal{F}_h} h_E^{-1}\|\nabla(v-v_I)\|_{0,E}^2\right)^{1/2} \\
&\quad\le C\left(\sum_{K\in\mathcal{C}_h} h_K^{-2}\|\nabla w_h-\boldsymbol{\beta}_h\|_{0,K}^2\right)^{1/2}\left(\sum_{E\in\mathcal{F}_h} h_E^{-3}\|v-v_I\|_{0,E}^2\right)^{1/2} \\
&\quad\le C\left(\sum_{K\in\mathcal{C}_h}\tilde{\eta}_K^2\right)^{1/2}\left(\sum_{K\in\mathcal{C}_h}|v|_{2,h/2,K}^2\right)^{1/2} \le C\left(\sum_{K\in\mathcal{C}_h}\tilde{\eta}_K^2\right)^{1/2}. \qquad (5.33)
\end{aligned}$$

Collecting (5.30)–(5.33) we obtain

$$\big|A\big| \le C\eta. \qquad (5.34)$$

Step 7. We will now estimate the term B. The following terms are directly estimated as the similar terms above:

$$\begin{aligned}&\left|\langle[\nabla w_h-\boldsymbol{\beta}_h]\cdot\boldsymbol{s},m_{ns}(\boldsymbol{\eta}-\boldsymbol{\eta}_I)\rangle_{\Gamma_{\mathrm{F}}}\right|+\left|\sum_{E\in\mathcal{F}_h}\frac{\gamma}{h_E}\langle[\nabla w_h-\boldsymbol{\beta}_h]\cdot\boldsymbol{s},[\boldsymbol{\eta}-\boldsymbol{\eta}_I]\cdot\boldsymbol{s}\rangle_E\right|\\ &\qquad\le C\eta.\end{aligned}\tag{5.35}$$

Since $\boldsymbol{\eta}_I$ is piecewise linear, it holds that $\boldsymbol{L}\boldsymbol{\eta}_{I|K}=\mathbf{0}$. The inverse estimate then gives

$$\begin{aligned}&\left|\sum_{K\in\mathcal{C}_h}\alpha h_K^2\big(\boldsymbol{L}\boldsymbol{\beta}_h+\boldsymbol{q}_h,\boldsymbol{L}(\boldsymbol{\eta}-\boldsymbol{\eta}_I)\big)_K\right|=\left|\sum_{K\in\mathcal{C}_h}\alpha h_K^2\big(\boldsymbol{L}\boldsymbol{\beta}_h+\boldsymbol{q}_h,\boldsymbol{L}\boldsymbol{\eta}\big)_K\right|\\ &\quad\le C\left(\sum_{K\in\mathcal{C}_h}\alpha h_K^2\|\boldsymbol{L}\boldsymbol{\beta}_h+\boldsymbol{q}_h\|_{0,K}^2\right)^{1/2}\|\boldsymbol{\eta}\|_1\\ &\quad\le C\eta,\end{aligned}\tag{5.36}$$

where we in the last step used (5.7). The final step in estimating the term B is to integrate by parts, use the Cauchy–Schwarz inequality, interpolation estimates, and again (5.7):

$$\begin{aligned}&\big|a(\boldsymbol{\beta}_h,\boldsymbol{\eta}-\boldsymbol{\eta}_I)-\big(\boldsymbol{q}_h,\boldsymbol{\eta}-\boldsymbol{\eta}_I\big)-\big\langle m_{ns}(\boldsymbol{\beta}_h),[\boldsymbol{\eta}-\boldsymbol{\eta}_I]\cdot\boldsymbol{s}\big\rangle_{\Gamma_{\mathrm{F}}}\big|\\ &\quad=\Bigg|-\sum_{K\in\mathcal{C}_h}(\boldsymbol{L}\boldsymbol{\beta}_h+\boldsymbol{q}_h,\boldsymbol{\eta}-\boldsymbol{\eta}_I)+\sum_{E\in\mathcal{I}_h}\langle[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!],\boldsymbol{\eta}-\boldsymbol{\eta}_I\rangle_E\\ &\qquad+\langle m_{nn}(\boldsymbol{\beta}_h),[\boldsymbol{\eta}-\boldsymbol{\eta}_I]\cdot\boldsymbol{n}\rangle_{\Gamma_{\mathrm{S}}\cup\Gamma_{\mathrm{F}}}\Bigg|\\ &\quad\le\sum_{K\in\mathcal{C}_h}\|\boldsymbol{L}\boldsymbol{\beta}_h+\boldsymbol{q}_h\|_{0,K}\|\boldsymbol{\eta}-\boldsymbol{\eta}_I\|_{0,K}+\sum_{E\in\mathcal{I}_h}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}\|\boldsymbol{\eta}-\boldsymbol{\eta}_I\|_{0,E}\\ &\qquad+\sum_{E\in\mathcal{S}_h\cup\mathcal{F}_h}\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}\|\boldsymbol{\eta}-\boldsymbol{\eta}_I\|_{0,E}\\ &\quad\le C\eta.\end{aligned}\tag{5.37}$$

Collecting (5.35)–(5.37) we obtain

$$\big|B\big|\le C\eta.\tag{5.38}$$

Step 8. The asserted estimate now follows from (5.16), (5.25), (5.28), (5.29), (5.34), and (5.38). ∎

We also have the following lemma for the shear force.

LEMMA 5.3. *It holds that*

$$\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|_{-1,*}\le C\big(|\!|\!|(w_{h/2}-w_h,\boldsymbol{\beta}_{h/2}-\boldsymbol{\beta}_h)|\!|\!|_{h/2}+\eta\big).\tag{5.39}$$

Proof. We start by observing that, referring to the definition (3.11) and its "$h/2$" counterpart, $\boldsymbol{q}_h$ and $\boldsymbol{q}_{h/2}$ are defined on different meshes and therefore with different h_K^2 coefficients. However, recalling that the size ratio between the two meshes is bounded, it is easy to check that an opportune splitting and the triangle inequality

give

$$\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|^2_{-1,h}\le C\left(\sum_{K\in\mathcal{C}_{h/2}}\|\nabla(w_{h/2}-w_h)-(\boldsymbol{\beta}_{h/2}-\boldsymbol{\beta}_h)\|^2_{0,K}\right.$$

$$\left.+\sum_{K\in\mathcal{C}_h}\|\nabla w_h-\boldsymbol{\beta}_h\|^2_{0,K}+\sum_{K\in\mathcal{C}_{h/2}}h_K^2\|\boldsymbol{L}\boldsymbol{\beta}_{h/2}-\boldsymbol{L}\boldsymbol{\beta}_h\|^2_{0,K}\right). \tag{5.40}$$

The first and the last term in (5.40) can be bounded in terms of the $|||\cdot|||_{h/2}$ norm, simply using the definition (4.3) and the inverse inequality

$$h_K^2\|\boldsymbol{L}\boldsymbol{\beta}_{h/2}-\boldsymbol{L}\boldsymbol{\beta}_h\|^2_{0,K}\le C\|\boldsymbol{\beta}_{h/2}-\boldsymbol{\beta}_h\|^2_{1,K}. \tag{5.41}$$

Therefore, recalling the definition (5.1), we get

$$\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|_{-1,h}\le C\big(|||(w_{h/2}-w_h,\boldsymbol{\beta}_{h/2}-\boldsymbol{\beta}_h)|||_{h/2}+\eta\big). \tag{5.42}$$

The transition from the $\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|_{-1,h}$ norm to the $\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|_{-1,*}$ norm is accomplished by using the "Pitkäranta–Verfürth trick" with steps almost identical to those used in Lemma 4.5, which are therefore omitted. ∎

Joining Theorem 5.2 and Lemma 5.3 gives the following a posteriori upper bound for the method.

THEOREM 5.4. *It holds that*

$$|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_h+\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*}\le C\eta. \tag{5.43}$$

Proof. Theorem 5.2 combined with Lemma 5.3 trivially gives

$$|||(w_{h/2}-w_h,\boldsymbol{\beta}_{h/2}-\boldsymbol{\beta}_h)|||_{h/2}+\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|_{-1,*}\le C\eta. \tag{5.44}$$

From the saturation assumption it follows that

$$|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h/2}+\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*}$$

$$\le\frac{1}{1-\rho}\big(|||(w_{h/2}-w_h,\boldsymbol{\beta}_{h/2}-\boldsymbol{\beta}_h)|||_{h/2}+\|\boldsymbol{q}_{h/2}-\boldsymbol{q}_h\|_{-1,*}\big), \tag{5.45}$$

and hence the assertion follows from (5.44). ∎

5.2. Lower bound. In this section, we prove the efficiency of the error estimator. Given any edge E of the triangulation, we define ω_E as the set of all of the triangles $K\in\mathcal{C}_h$ that have E as an edge. Given any $K\in\mathcal{C}_h$, we define ω_K as the set of all of the triangles in $\mathcal{C}_h$ that share an edge with K. We then have the following lemma [8].

LEMMA 5.5. *Given any edge E of the triangulation $\mathcal{C}_h$, let $P_k(E)$ be the space of polynomials of degree at most k on E. There exists a linear operator*

$$\Pi_E\ :\ P_k(E)\longrightarrow H^2_0(\omega_E) \tag{5.46}$$

such that for all $p_k\in P_k(E)$ it holds that

$$C_1\|p_k\|^2_{0,E}\le\langle p_k,\Pi_E(p_k)\rangle_E\le\|p_k\|^2_{0,E}, \tag{5.47}$$

$$\|\Pi_E(p_k)\|_{0,\omega_E}\le C_2h_E^{1/2}\|p_k\|_{0,E}, \tag{5.48}$$

where the positive constants C_i above depend only on k and the minimum angle of the triangles in $\mathcal{C}_h$.

Next, we define a local counterpart of the negative norm defined in (4.33) for the shear force.

$$\|\boldsymbol{r}\|_{-1,*,\omega_K} = \sup_{\substack{\boldsymbol{\eta}\in V_* \\ \boldsymbol{\eta}=\boldsymbol{0} \text{ in } \Omega\setminus\omega_K}} \frac{\langle \boldsymbol{r},\boldsymbol{\eta}\rangle}{\|\boldsymbol{\eta}\|_1}. \tag{5.49}$$

We then have the following reliability result.

THEOREM 5.6. *It holds that*

$$\eta_K \le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_K} + \|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_K} + h_K^2\|f-f_h\|_{0,\omega_K}\big), \tag{5.50}$$

where f_h is some approximation of the load f. Here $|||\cdot|||_{h,\omega_K}$ and $\|\cdot\|_{0,\omega_K}$ represent, respectively, the standard restrictions of the norms $|||\cdot|||_h$ and $\|\cdot\|_0$ to the domain ω_K.

Proof. The proof of the theorem consists of bounding separately all of the addenda of η_K in (5.5).

Step 1. We first bound the terms of $\tilde{\eta}_K^2$ in (5.1). Considering the right-hand side of (5.50), the triangle inequality immediately shows that it is sufficient to bound the term $h_K^2\|f_h+\operatorname{div}\boldsymbol{q}_h\|_{0,K}$.

Given any $K\in\mathcal{C}_h$, let b_K indicate the standard third-order polynomial bubble function on K, scaled such that $\|b_K\|_{L^\infty(K)}=1$. Given $K\in\mathcal{C}_h$, let now $\varphi_K\in H_0^2(K)$ be defined as

$$\varphi_K = (f_h+\operatorname{div}\boldsymbol{q}_h)\, b_K^2. \tag{5.51}$$

The standard scaling arguments then easily show that

$$\|f_h+\operatorname{div}\boldsymbol{q}_h\|_{0,K}^2 \le C(f_h+\operatorname{div}\boldsymbol{q}_h,\varphi_K)_K, \tag{5.52}$$

$$\|\varphi_K\|_{0,K} \le C\|f_h+\operatorname{div}\boldsymbol{q}_h\|_{0,K}. \tag{5.53}$$

For the first term in $\tilde{\eta}_K^2$, the equilibrium equation (2.13) and integration by parts give

$$\begin{aligned} h_K^2\|f_h+\operatorname{div}\boldsymbol{q}_h\|_{0,K}^2 &\le Ch_K^2(f_h+\operatorname{div}\boldsymbol{q}_h,\varphi_K)_K \\ &= Ch_K^2\big((f+\operatorname{div}\boldsymbol{q}_h,\varphi_K)_K + (f_h-f,\varphi_K)_K\big) \\ &= Ch_K^2\big((-\operatorname{div}\boldsymbol{q}+\operatorname{div}\boldsymbol{q}_h,\varphi_K)_K + (f_h-f,\varphi_K)_K\big) \\ &= Ch_K^2\big((\boldsymbol{q}_h-\boldsymbol{q},\nabla\varphi_K)_K + (f_h-f,\varphi_K)_K\big). \end{aligned} \tag{5.54}$$

We note, in particular, that $\nabla\varphi_K\in\boldsymbol{V}_*$ and $\nabla\varphi_K=\boldsymbol{0}$ in $\Omega\setminus K$. Therefore, the duality inequality and the Cauchy–Schwarz inequality followed by the inverse inequality and the bound (5.53) lead to the estimate

$$\begin{aligned} &Ch_K^2\big((\boldsymbol{q}_h-\boldsymbol{q},\nabla\varphi_K)_K + (f_h-f,\varphi_K)\big) \\ &\quad\le C\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,K}\, h_K^2\|\nabla\varphi_K\|_{1,K} + Ch_K^2\|f-f_h\|_{0,K}\|\varphi_K\|_{0,K} \\ &\quad\le C\big(\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,K} + h_K^2\|f-f_h\|_{0,K}\big)\|f_h+\operatorname{div}\boldsymbol{q}_h\|_{0,K}. \end{aligned} \tag{5.55}$$

Combining now (5.54) with (5.55) gives

$$h_K^2\|f_h+\operatorname{div}\boldsymbol{q}_h\|_{0,K} \le C\big(\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,K} + h_K^2\|f-f_h\|_{0,K}\big). \tag{5.56}$$

The second term of $\tilde{\eta}_K^2$ in (5.1) can be directly bounded by using the Kirchhoff condition (2.12) with the definitions (4.1)–(4.3):

$$\begin{aligned} h_K^{-1}\|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K} &= h_K^{-1}\|\nabla(w - w_h) - (\boldsymbol{\beta} - \boldsymbol{\beta}_h)\|_{0,K}^2 \\ &\le |||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_{h,K}. \end{aligned} \tag{5.57}$$

Step 2. We next bound the terms of η_E^2 in (5.2). Given now $E \in I(K)$, an edge of the element K lying in the interior of Ω, let

$$\boldsymbol{\varphi}_E = \Pi_E([\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]), \tag{5.58}$$

where, with a little abuse of notation, the operator Π_E is intended as applied on each single component. Then, from (5.47) with integration by parts, it follows that

$$\begin{aligned} h_E^{1/2}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}^2 &\le Ch_E^{1/2}\langle [\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!], \boldsymbol{\varphi}_E\rangle_E \\ &= Ch_E^{1/2}\big((\boldsymbol{L}\boldsymbol{\beta}_h, \boldsymbol{\varphi}_E)_{\omega_E} + (\boldsymbol{m}(\boldsymbol{\beta}_h), \nabla\boldsymbol{\varphi}_E)_{\omega_E}\big), \end{aligned} \tag{5.59}$$

where we recall that ω_E was defined at the start of this section. Integration by parts and the equation (2.14) immediately lead to the identity

$$(\boldsymbol{m}(\boldsymbol{\beta}), \nabla\boldsymbol{\varphi}_E)_{\omega_E} = -(\boldsymbol{L}\boldsymbol{\beta}, \boldsymbol{\varphi}_E)_{\omega_E} = (\boldsymbol{q}, \boldsymbol{\varphi}_E)_{\omega_E}, \tag{5.60}$$

which, applied to (5.59), gives

$$\begin{aligned} &h_E^{1/2}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}^2 \\ &\quad\le Ch_E^{1/2}\big((\boldsymbol{L}\boldsymbol{\beta}_h + \boldsymbol{q}, \boldsymbol{\varphi}_E)_{\omega_E} + (\boldsymbol{m}(\boldsymbol{\beta}_h) - \boldsymbol{m}(\boldsymbol{\beta}), \nabla\boldsymbol{\varphi}_E)_{\omega_E}\big) \\ &\quad= Ch_E^{1/2}\big((\boldsymbol{L}\boldsymbol{\beta}_h + \boldsymbol{q}_h, \boldsymbol{\varphi}_E)_{\omega_E} + (\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\varphi}_E)_{\omega_E} \\ &\qquad+(\boldsymbol{m}(\boldsymbol{\beta}_h) - \boldsymbol{m}(\boldsymbol{\beta}), \nabla\boldsymbol{\varphi}_E)_{\omega_E}\big). \end{aligned} \tag{5.61}$$

Next, we bound the three terms on the right-hand side of (5.61). For the first term, the identity (5.7), the Cauchy–Schwarz inequality, the definition (5.58), and the bound (5.48) give

$$\begin{aligned} &h_E^{1/2}(\boldsymbol{L}\boldsymbol{\beta}_h + \boldsymbol{q}_h, \boldsymbol{\varphi}_E)_{\omega_E} \\ &\quad\le C\left(\sum_{K\subset\omega_E} h_K^{-2}\|\nabla w_h - \boldsymbol{\beta}_h\|_{0,K}^2\right)^{1/2} \|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E} \\ &\quad\le C\,|||(w - w_h, \boldsymbol{\beta} - \boldsymbol{\beta}_h)|||_{h,\omega_E}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}. \end{aligned} \tag{5.62}$$

For the second term on the right-hand side of (5.61), we note that $\boldsymbol{\varphi}_E \in \boldsymbol{V}_*$ and $\boldsymbol{\varphi}_E = \boldsymbol{0}$ in $\Omega\backslash\omega_E$. Therefore, the duality inequality and the definition (5.58) combined with the bound (5.48) give

$$\begin{aligned} h_E^{1/2}(\boldsymbol{q} - \boldsymbol{q}_h, \boldsymbol{\varphi}_E)_{\omega_E} &\le h_E^{1/2}\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,*,\omega_E}\|\boldsymbol{\varphi}_E\|_{1,\omega_E} \\ &\le C\|\boldsymbol{q} - \boldsymbol{q}_h\|_{-1,*,\omega_E}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}. \end{aligned} \tag{5.63}$$

For the third term of (5.61), the Cauchy–Schwarz inequality, then the inverse inequality, and finally (5.58) combined with the bound (5.48) lead to the estimate

$$\begin{aligned} h_E^{1/2}(\boldsymbol{m}(\boldsymbol{\beta}_h) - \boldsymbol{m}(\boldsymbol{\beta}), \nabla\boldsymbol{\varphi}_E)_{\omega_E} &\le C\|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_{1,\omega_E}\, h_K^{-1/2}\|\boldsymbol{\varphi}_E\|_{0,\omega_E} \\ &\le C\|\boldsymbol{\beta} - \boldsymbol{\beta}_h\|_{1,\omega_E}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E}. \end{aligned} \tag{5.64}$$

Now, by combining (5.62), (5.63), and (5.64) with (5.61) it follows that

$$h_E^{1/2}\|[\![\boldsymbol{m}(\boldsymbol{\beta}_h)\boldsymbol{n}]\!]\|_{0,E} \le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E} + \|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E}\big). \tag{5.65}$$

The remaining term of η_E^2 is bounded with similar arguments; with the notation

$$\varphi_E = \Pi_E([\![\boldsymbol{q}_h\cdot\boldsymbol{n}]\!]), \tag{5.66}$$

the identity

$$-(\operatorname{div}\boldsymbol{q},\varphi_E)_{\omega_E} = (\boldsymbol{q},\nabla\varphi_E)_{\omega_E} \tag{5.67}$$

with (5.54) implies

$$\begin{aligned} h_E^{1/2}\|[\![\boldsymbol{q}\cdot\boldsymbol{n}]\!]\|_{0,E}^2 &\le Ch_E^{1/2}\langle[\![\boldsymbol{q}\cdot\boldsymbol{n}]\!],\varphi_E\rangle_E \\ &\le Ch_E^{1/2}\big((f-f_h,\varphi_E)_{\omega_E} + (\boldsymbol{q}_h-\boldsymbol{q},\nabla\varphi_E)_{\omega_E}\big). \end{aligned} \tag{5.68}$$

Finally, we note that $\nabla\varphi_E \in \boldsymbol{V}_*$ and $\nabla\varphi_E = \boldsymbol{0}$ in $\Omega\backslash\omega_E$. Therefore,

$$h_E^{3/2}\|[\![\boldsymbol{q}_h\cdot\boldsymbol{n}]\!]\|_{0,E} \le C\big(\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E} + h_K^2\|f-f_h\|_{0,\omega_E}\big). \tag{5.69}$$

Step 3. Third, we bound the only term of $\eta_{S,E}^2$ in (5.3) which appears in $\eta_{F,E}^2$ as well. Given now a triangulation edge E in $S(K)\cup F(K)$, let

$$\varphi_E = \Pi_E(m_{nn}(\boldsymbol{\beta}_h)). \tag{5.70}$$

Due to (5.47) and (2.19), integration by parts gives (here $\boldsymbol{\nabla}$ denotes the tensor-valued gradient applied to a vector-valued function)

$$\begin{aligned} h_E^{1/2}\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}^2 &\le h_E^{1/2}\langle m_{nn}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\rangle_E \\ &= h_E^{1/2}\langle \boldsymbol{m}_n(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\boldsymbol{n}\rangle_E \\ &= h_E^{1/2}\big((\boldsymbol{m}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\boldsymbol{\nabla}(\varphi_E\boldsymbol{n}))_{\omega_E} + (\boldsymbol{L}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\boldsymbol{n})_{\omega_E}\big), \end{aligned} \tag{5.71}$$

where $\boldsymbol{n}$ is, as usual, the chosen normal unit vector to E. For the first term, using the Cauchy–Schwarz inequality, then the inverse inequality, and finally the bound (5.48), we easily get

$$\begin{aligned} h_E^{1/2}(\boldsymbol{m}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\boldsymbol{\nabla}(\varphi_E\boldsymbol{n}))_{\omega_E} &\le h_E^{1/2}\|\boldsymbol{\beta}-\boldsymbol{\beta}_h\|_{1,\omega_E}\|\boldsymbol{\nabla}(\varphi_E\boldsymbol{n})\|_{0,\omega_E} \\ &\le C\|\boldsymbol{\beta}-\boldsymbol{\beta}_h\|_{1,\omega_E}\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}. \end{aligned} \tag{5.72}$$

For the second term in (5.71), recalling (2.14) we have

$$\begin{aligned} &h_E^{1/2}(\boldsymbol{L}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\boldsymbol{n})_{\omega_E} \\ &\quad = h_E^{1/2}(\boldsymbol{L}\boldsymbol{\beta}_h+\boldsymbol{q}_h,\varphi_E\boldsymbol{n})_{\omega_E} + h_E^{1/2}(\boldsymbol{q}-\boldsymbol{q}_h,\varphi_E\boldsymbol{n})_{\omega_E}. \end{aligned} \tag{5.73}$$

Observing now that $\varphi_E\boldsymbol{n} \in \boldsymbol{V}_*$ and $\varphi_E\boldsymbol{n} = \boldsymbol{0}$ in $\Omega\backslash\omega_E$, the two terms on the right-hand side of (5.73) can be bounded with the same arguments used above, respectively, in (5.62) and (5.63). Omitting the details, we therefore get

$$\begin{aligned} h_E^{1/2}(\boldsymbol{L}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\boldsymbol{n})_{\omega_E} &\le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E} \\ &\quad + \|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E}\big)\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}. \end{aligned} \tag{5.74}$$

From (5.71), (5.72), and (5.74) we get

$$h_E^{1/2}\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E} \le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E} + \|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E}\big). \tag{5.75}$$

Step 4. Finally, we bound the last term of $\eta_{F,E}^2$ in (5.4). Given now a triangulation edge E in $F(K)$, let

$$\varphi_E = \Pi_E\left(\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h) - \boldsymbol{q}_h\cdot\boldsymbol{n}\right). \tag{5.76}$$

Using (5.47) and recalling (2.17), we obtain

$$\begin{aligned} &h_E^{3/2}\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h) - \boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E}^2 \\ &\quad\le h_E^{3/2}\left(\left\langle\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\right\rangle_E + \langle[\boldsymbol{q}-\boldsymbol{q}_h]\cdot\boldsymbol{n},\varphi_E\rangle_E\right). \end{aligned} \tag{5.77}$$

For the first term, integration by parts on the edge and simple algebra give

$$\begin{aligned} h_E^{3/2}\left\langle\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\varphi_E\right\rangle_E &= h_E^{3/2}\langle m_{ns}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\nabla\varphi_E\cdot\boldsymbol{s}\rangle_E \\ &= h_E^{3/2}\big(\langle\boldsymbol{m}(\boldsymbol{\beta}-\boldsymbol{\beta}_h)\boldsymbol{n},\nabla\varphi_E\rangle_E - \langle m_{nn}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\nabla\varphi_E\cdot\boldsymbol{n}\rangle_E\big). \end{aligned} \tag{5.78}$$

Using again integration by parts, the first term in (5.78) can be written as

$$\begin{aligned} &h_E^{3/2}\langle\boldsymbol{m}(\boldsymbol{\beta}-\boldsymbol{\beta}_h)\boldsymbol{n},\nabla\varphi_E\rangle_E \\ &\quad= h_E^{3/2}\big(\boldsymbol{L}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\nabla\varphi_E)_{\omega_E} + \langle\boldsymbol{m}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\boldsymbol{\nabla}\nabla\varphi_E)_{\omega_E}\big). \end{aligned} \tag{5.79}$$

The second term in (5.77), again due to integration by parts and recalling (2.13), is instead equivalent to

$$\begin{aligned} h_E^{3/2}\langle[\boldsymbol{q}-\boldsymbol{q}_h]\cdot\boldsymbol{n},\varphi_E\rangle_E &= h_E^{3/2}\big(\boldsymbol{q}-\boldsymbol{q}_h,\nabla\varphi_E)_{\omega_E} \\ &\quad -(f_h+\operatorname{div}\boldsymbol{q}_h,\varphi_E)_{\omega_E} - (f-f_h,\varphi_E)_{\omega_E}\big). \end{aligned} \tag{5.80}$$

For the first term, due to (2.14) and (3.11), we now have

$$\begin{aligned} &h_E^{3/2}(\boldsymbol{q}-\boldsymbol{q}_h,\nabla\varphi_E)_{\omega_E} \\ &\quad= h_E^{3/2}\big(\boldsymbol{L}(\boldsymbol{\beta}_h-\boldsymbol{\beta}),\nabla\varphi_E)_{\omega_E} - \frac{1}{\alpha h_{\omega_E}^2}(\nabla w_h-\boldsymbol{\beta}_h,\nabla\varphi_E\big)_{\omega_E}, \end{aligned} \tag{5.81}$$

where h_{ω_E} is the size of the triangle ω_E. Combining all of the identities from (5.77) to (5.81), it follows that

$$\begin{aligned} &h_E^{3/2}\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h) - \boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E}^2 \\ &\quad\le h_E^{3/2}\Bigg((\boldsymbol{m}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\boldsymbol{\nabla}\nabla\varphi_E)_{\omega_E} - (m_{nn}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\nabla\varphi_E\cdot\boldsymbol{n}\rangle_E \\ &\qquad-\frac{1}{\alpha h_{\omega_E}^2}(\nabla w_h-\boldsymbol{\beta}_h,\nabla\varphi_E)_{\omega_E} - (f_h+\operatorname{div}\boldsymbol{q}_h,\varphi_E)_{\omega_E} \\ &\qquad-(f-f_h,\varphi_E)_{\omega_E}\Bigg). \end{aligned} \tag{5.82}$$

For the second term on the right-hand side of (5.82), recalling (2.17), using the Cauchy–Schwarz inequality and the bound (5.75), we have

$$\begin{aligned} h_E^{3/2}\langle m_{nn}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\nabla\varphi_E\cdot\boldsymbol{n}\rangle_E &\le h_E^{1/2}\|m_{nn}(\boldsymbol{\beta}_h)\|_{0,E}\, h_E\|\nabla\varphi_E\|_{0,E} \\ &\le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E}+\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E}\big)h_E\|\nabla\varphi_E\|_{0,E}, \end{aligned} \tag{5.83}$$

which, using the inverse inequality and the bound (5.48), gives

$$\begin{aligned} h_E^{3/2}\langle m_{nn}(\boldsymbol{\beta}-\boldsymbol{\beta}_h),\nabla\varphi_E\cdot\boldsymbol{n}\rangle_E &\le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E} \\ &\quad+\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E}\big)\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E}. \end{aligned} \tag{5.84}$$

The remaining terms on the right-hand side of (5.82) can all be bounded using the Cauchy–Schwarz inequality, the inverse inequality, and the bounds (5.56), (5.48) as already shown for the similar previous cases. Without showing all of the details, we finally get

$$\begin{aligned} &h_E^{3/2}\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E}^2 \\ &\quad\le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E}+\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E} \\ &\qquad+h_K^2\|f-f_h\|_{0,K}\big)\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E} \end{aligned} \tag{5.85}$$

or, trivially,

$$\begin{aligned} &h_E^{3/2}\left\|\frac{\partial}{\partial s}m_{ns}(\boldsymbol{\beta}_h)-\boldsymbol{q}_h\cdot\boldsymbol{n}\right\|_{0,E} \\ &\quad\le C\big(|||(w-w_h,\boldsymbol{\beta}-\boldsymbol{\beta}_h)|||_{h,\omega_E}+\|\boldsymbol{q}-\boldsymbol{q}_h\|_{-1,*,\omega_E}+h_K^2\|f-f_h\|_{0,K}\big). \end{aligned} \tag{5.86}$$

Recalling now the definitions for η_K in (5.1) and the local negative norm in (5.49), the proposition is proved. ∎

REFERENCES

[1] D. N. Arnold and R. S. Falk, *Asymptotic analysis of the boundary layer for the Reissner–Mindlin plate model*, SIAM J. Math. Anal., 27 (1996), pp. 486–514.

[2] L. Beirão da Veiga and F. Brezzi, *Reissner–Mindlin plates with free boundary conditions*, Atti dei Convegni Lincei, 210 (2004), pp. 43–54.

[3] L. Beirão da Veiga, J. Niiranen, and R. Stenberg, *A Family of C^0 Finite Elements for Kirchhoff Plates* II: *Numerical Results*, Helsinki University of Technology, Institute of Mathematics, Research Reports, A526 (2007), http://math.tkk.fi/reports/.

[4] L. Beirão da Veiga, J. Niiranen, and R. Stenberg, *A new family of C^0 finite elements for the Kirchhoff plate model*, in Topics on Mathematics for Smart Systems, Proceedings of the European Conference, B. Miara, G. Stavroulakis, and V. Valente, eds., World Scientific, River Edge, NJ, 2007, pp. 45–60.

[5] L. Beirão da Veiga, *Finite element methods for a modified Reissner–Mindlin free plate model*, SIAM J. Numer. Anal., 42 (2004), pp. 1572–1591.

[6] C. Chinosi, *PSRI elements for the Reissner–Mindlin free plate*, Comput. & Structures, 83 (2005), pp. 2559–2572.

[7] P. Destuynder and T. Nevers, *Une modification du modèle de Mindlin pour les plaques minces en flexion présentant un bord libre*, RAIRO Modél. Math. Anal. Numér., 22 (1988), pp. 217–242.

[8] C. Lovadina and R. Stenberg, *A posteriori error analysis of the linked interpolation technique for plate bending problems*, SIAM J. Numer. Anal., 43 (2005), pp. 2227–2249.

[9] H. Melzer and R. Rannacher, *Spannungskonzentrationen in Eckpunkten der vertikal belasteten Kirchhoffschen Platte*, Bauingenieur, 55 (1980), pp. 181–189.

[10] J. Pitkäranta and M. Suri, *Design principles and error analysis for reduced-shear plate-bending finite elements*, Numer. Math., 75 (1996), pp. 223–266.

[11] J. Pitkäranta, *Boundary subspaces for the finite element method with Lagrange multipliers*, Numer. Math., 33 (1979), pp. 273–289.

[12] J. Pitkäranta, *Analysis of some low-order finite element schemes for Mindlin–Reissner and Kirchhoff plates*, Numer. Math., 53 (1988), pp. 237–254.

[13] R. Stenberg, *A new finite element formulation for the plate bending problem*, in Asymptotic Methods for Elastic Structures, P. Ciarlet, L. Trabucho, and J. M. Viano, eds., de Gruyter, Berlin, 1995, pp. 209–221.

[14] R. Verfürth, *Error estimates for a mixed finite element approximation of the Stokes equations*, RAIRO Anal. Numer., 18 (1984), pp. 175–182.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2072–2097

ANALYSIS OF THE COUPLING OF PRIMAL AND DUAL-MIXED FINITE ELEMENT METHODS FOR A TWO-DIMENSIONAL FLUID-SOLID INTERACTION PROBLEM*

GABRIEL N. GATICA†, ANTONIO MÁRQUEZ‡, AND SALIM MEDDAHI§

Abstract. This paper deals with a time-harmonic fluid-solid interaction problem posed in the plane. More precisely, we apply the coupling of primal and dual-mixed finite element methods to compute both the pressure of the scattered wave in the linearized fluid and the elastic vibrations that take place in the solid elastic body. To this end, we solve a transmission problem holding between the cross-section of the infinitely long cylinder representing the obstacle and an annular region surrounding it. The novelty of our method lies in the use of a dual-mixed variational formulation in the obstacle, while maintaining the usual primal formulation in the fluid. In other words, we introduce a stress-pressure formulation of the problem instead of the traditional displacement-pressure encountered in the literature. As a consequence, one of the transmission conditions becomes essential, and hence we enforce it weakly by means of a Lagrange multiplier. Next, we apply the abstract framework developed in a recent work by A. Buffa, prove that our coupled variational formulation is well posed, and define the corresponding discrete scheme by using PEERS in the solid domain and standard Lagrange finite elements in the fluid domain. Then we show that the resulting Galerkin scheme is uniquely solvable and convergent and derive optimal error estimates. Finally, we illustrate our analysis with some results from computational experiments.

Key words. mixed finite elements, Helmholtz equation, elastodynamic equation

AMS subject classification. 65N30, 65N12, 65N15, 74F10, 74B05, 35J05

DOI. 10.1137/060660370

1. Introduction.

In this paper we develop a coupled primal/dual-mixed finite element method for a time-harmonic fluid-solid interaction problem in the plane. We consider an elastic body occupying a region Ω_s and assume that it is subject to a given incident wave that travels in the fluid surrounding it. Actually, we suppose here that the fluid occupies an annular region Ω_f whose exterior boundary Γ is located far from the obstacle (the solid body) and impose on this artificial closed curve a boundary condition that imitates the behavior of the scattered field at infinity. Thus, our model problem is posed in a bounded region. Concerning the numerical solution of this kind of fluid-solid interaction problem, we remark that they have deserved some attention recently from the FEM and BEM-FEM communities. However, to the best of our knowledge, all the methods rely on a displacement formulation of the linear elasticity equation posed in Ω_s (see, e.g., [8, 24, 27, 28] and the references cited therein). However, it is also known that the stress in the solid and the pressure in the fluid usually have more physical interest than the displacement and the velocity,

*Received by the editors May 19, 2006; accepted for publication (in revised form) February 28, 2007; published electronically September 26, 2007. This research was partially supported by CONICYT-Chile through the FONDAP Program in Applied Mathematics, by the Dirección de Investigación of the Universidad de Concepción through the Advanced Research Groups Program, and by the Ministry of Education of Chile through the MECESUP Project UCO0406.
http://www.siam.org/journals/sinum/45-5/66037.html

†GI²MA, Departamento de Ingeniería Matemática, Universidad de Concepción, Casilla 160-C, Concepción, Chile (ggatica@ing-mat.udec.cl).

‡Departamento de Construcción e Ingeniería de Fabricación, Universidad de Oviedo, Oviedo, España (amarquez@uniovi.es).

§Departamento de Matemáticas, Facultad de Ciencias, Universidad de Oviedo, Calvo Sotelo s/n, Oviedo, España (salim@orion.ciencias.uniovi.es).

respectively. Therefore, instead of that classical approach, our goal in this paper is to employ a dual-mixed variational formulation for plane elasticity in the obstacle and keep the usual primal formulation in the linearized fluid region. In this way, the stress tensor in Ω_s and the linearized fluid pressure p, which solves the Helmholtz equation in Ω_f, constitute our main unknowns. In addition, since one of the transmission conditions becomes essential, we enforce it weakly by means of a Lagrange multiplier. We remark that, as compared with the stress-velocity formulation, this turns out to be an advantageous feature of the present approach. In fact, if that formulation were employed, both transmission conditions would be essential and then two Lagrange multipliers would be required.

Now, it is important to point out that in the Helmholtz and elastodynamics equations there is a zero order term with a "wrong" sign that causes the loss of ellipticity of the operators arising from the corresponding primal formulations. Nevertheless, the compactness of the embedding $H^1(\Omega_s) \hookrightarrow L^2(\Omega_s)$ allows one to use successfully a Fredholm alternative to analyze its solvability. On the other hand, the usual dual-mixed formulation is more intricate since it does not fit in any classical theory for proving well-posedness. In particular, a strategy based on elaborated duality arguments is developed in [16] for dealing with this difficulty in the case of the Helmholtz equation. In the present paper we apply a different approach. First of all, in contrast to the usual dual-mixed formulation, the elastodynamic equation is used here to eliminate the original unknown given by the displacement field $\mathbf{u}$. This leads to a method that is equivalent (both at the continuous and the discrete levels) to the standard dual-mixed formulation (see [1]), which has the advantage of reducing the number of unknowns. However, the elimination of $\mathbf{u}$ yields to a formulation that, though having the typical mixed structure, does not satisfy the hypotheses of the Babuška–Brezzi theory. In addition, since the canonical embedding $\mathbf{H}(\mathbf{div}; \Omega_s) \hookrightarrow [L^2(\Omega_s)]^2$ is not compact, it is not possible to employ a Fredholm alternative, at least for the original form of the resulting variational formulation.

In order to circumvent the above difficulties, we take advantage of a recent technique essentially developed for electromagnetism (see [11] and the references cited therein). In fact, a successful strategy has been developed there to deal with a similar noncoercive bilinear form arising in the study of Maxwell equations. Actually, Buffa [11] succeeded in setting up this technique in a general framework. We extend here the range of application of this methodology by using it for the dual-mixed formulation described above. More precisely, we show that a judicious decomposition of $\mathbf{H}(\mathbf{div}; \Omega_s)$ renders suitable the application of a Fredholm alternative for the analysis of the whole coupled problem. The corresponding discrete scheme is defined with PEERS elements in the obstacle and the traditional first order Lagrange finite elements in the fluid domain. The stability and convergence of this Galerkin method also relies on a stable decomposition of the finite element space used to approximate the stress variable. Now, if the stress-velocity formulation were applied then, besides the second Lagrange multiplier on the interface, the discrete system would employ the Raviart–Thomas subspace in the fluid, which involves many more degrees of freedom than the first order Lagrange finite elements. Therefore, since the pressure and not the velocity is the variable of interest in the acoustic medium (fluid), this additional computational effort does not seem to be worthy at all.

The remainder of the paper is organized as follows. In sections 2 and 3 we give a brief description of the fluid-solid interaction problem and derive its coupled variational formulation. In section 4, we show that the resulting saddle point problem is

well posed. The corresponding Galerkin scheme is analyzed in section 5. Finally, in section 6 we provide results from numerical experiments that confirm our theoretical assertions. We end this section with some notation to be used below. Since in what follows we deal with complex valued functions, we let $\mathbb{C}$ be the set of complex numbers, use the symbol $\imath$ for $\sqrt{-1}$, and denote by $\overline{z}$ and $|z|$ the conjugate and modulus, respectively, of each $z \in \mathbb{C}$. In addition, given any Hilbert space U, U^2 and $U^{2\times 2}$ denote, respectively, the space of vectors and tensors of order 2 with entries in U. In particular, $\mathbf{I}$ is the identity matrix of $\mathbb{C}^{2\times 2}$, and given $\boldsymbol{\tau} := (\tau_{ij})$, $\boldsymbol{\zeta} := (\zeta_{ij}) \in \mathbb{C}^{2\times 2}$, we define as usual the transpose tensor $\boldsymbol{\tau}^{\mathtt{t}} := (\tau_{ji})$, the trace $\mathrm{tr}(\boldsymbol{\tau}) := \sum_{i=1}^{2} \tau_{ii}$, the deviator tensor $\boldsymbol{\tau}^{\mathtt{d}} := \boldsymbol{\tau} - \frac{1}{2}\,\mathrm{tr}(\boldsymbol{\tau})\,\mathbf{I}$, the tensor product $\boldsymbol{\tau} : \boldsymbol{\zeta} := \sum_{i,j=1}^{2} \tau_{ij}\,\zeta_{ij}$, and the conjugate tensor $\overline{\boldsymbol{\tau}} := (\overline{\tau}_{ij})$. Finally, in what follows we utilize the standard terminology for Sobolev spaces and norms, employ $\mathbf{0}$ to denote a generic null vector, and use C and c, with or without subscripts, bars, tildes or hats, to denote generic constants independent of the discretization parameters, which may take different values at different places.

2. The fluid-solid interaction problem. We consider an incident acoustic wave upon a bounded elastic body (obstacle) fully surrounded by a fluid, and are interested in determining both the response of both the body and the scattered wave. The obstacle is supposed to be an infinitely long cylinder parallel to the x_3-axis whose cross-section is Ω_s. The boundary of Ω_s is denoted by Σ. We assume that the incident wave and the volume force acting on the body exhibit a time-harmonic behavior with $e^{-\imath\,\omega\,t}$ ansatz and phasors p_i and $\mathbf{f}$, respectively, so that p_i satisfies the Helmholtz equation in $\mathbb{R}^2\backslash\Omega_s$. Hence, as we assume that the phenomenon is invariant under a translation in the x_3-direction, we may consider a bidimensional interaction problem posed in the frequency domain. In this way, and since we plan to employ a mixed variational formulation in the solid, our main unknowns become the phasor $\boldsymbol{\sigma} : \Omega_s \to \mathbb{C}^{2\times 2}$ of the Cauchy stress tensor, the phasor $\mathbf{u} : \Omega_s \to \mathbb{C}^2$ of the displacement field, and the phasor of the total (incident + scattered) pressure $p : \mathbb{R}^2\backslash\Omega_s \to \mathbb{C}$.

The fluid is assumed to be perfect, compressible, and homogeneous, with mass density ρ_f and wave number $\kappa_f := \frac{\omega}{v_0}$, where v_0 is the speed of sound in the linearized fluid. In addition, the solid is supposed to be isotropic and linearly elastic with mass density ρ_s and Lamé constants μ and λ, which means, in particular, that the corresponding constitutive equation is given by

$$\boldsymbol{\sigma} \,=\, \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}) \qquad \text{in } \Omega_s\,, \tag{2.1}$$

where $\boldsymbol{\varepsilon}(\mathbf{u}) := \frac{1}{2}\,(\nabla\mathbf{u} + (\nabla\mathbf{u})^{\mathtt{t}})$ is the strain tensor of small deformations, ∇ is the gradient tensor, and $\mathcal{C}$ is the elasticity operator given by Hooke's law, that is,

$$\mathcal{C}\,\boldsymbol{\zeta} \;:=\; \lambda\,\mathrm{tr}(\boldsymbol{\zeta})\,\mathbf{I} \,+\, 2\,\mu\,\boldsymbol{\zeta} \qquad \forall\,\boldsymbol{\zeta} \,\in\, [L^2(\Omega_s)]^{2\times 2}\,, \quad \boldsymbol{\zeta} \,=\, \boldsymbol{\zeta}^{\mathtt{t}}\,. \tag{2.2}$$

Consequently, under the hypotheses of small oscillations, in both the solid and the fluid, the unknowns $\boldsymbol{\sigma}$, $\mathbf{u}$, and p satisfy the following equations:

$$\begin{aligned} \boldsymbol{\sigma} \,&=\, \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}) && \text{in } \Omega_s\,, \\ \mathbf{div}(\boldsymbol{\sigma}) \,+\, \rho_s\,\omega^2\,\mathbf{u} \,&=\, -\,\mathbf{f} && \text{in } \Omega_s\,, \\ \Delta p \,+\, \kappa_f^2\,p \,&=\, 0 && \text{in } \mathbb{R}^2\backslash\Omega_s\,, \end{aligned} \tag{2.3}$$

together with the transmission conditions

$$\begin{aligned} \boldsymbol{\sigma}\boldsymbol{\nu} &= -p\boldsymbol{\nu} \quad \text{on } \Sigma\,, \\ \rho_f\,\omega^2\,\mathbf{u}\cdot\boldsymbol{\nu} &= \frac{\partial p}{\partial \boldsymbol{\nu}} \quad \text{on } \Sigma \end{aligned} \tag{2.4}$$

and the behavior at infinity given by

$$p - p_i \;=\; O(\mathtt{r}^{-1/2}) \tag{2.5}$$

and

$$\frac{\partial (p-p_i)}{\partial \mathtt{r}} \;-\; \imath\,\kappa_f\,(p-p_i) \;=\; o(\mathtt{r}^{-1/2})\,, \tag{2.6}$$

as $\mathtt{r} := \|\mathbf{x}\| \to +\infty$, uniformly for all directions $\frac{\mathbf{x}}{\|\mathbf{x}\|}$. Hereafter, **div** stands for the usual divergence operator div acting on each row of the tensor, $\|\mathbf{x}\|$ is the euclidean norm of a vector $\mathbf{x} := (x_1, x_2)^{\mathtt{t}} \in \mathbb{R}^2$, and $\boldsymbol{\nu}$ denotes the unit outward normal on Σ. The second and third equations of (2.3) correspond to the elastodynamic and acoustic equations in time-harmonic regime, respectively, whereas the transmission conditions given in (2.4) represent the equilibrium of forces and the equality of the normal displacements of the solid and fluid. Equation (2.6) is known as the Sommerfeld radiation condition.

On the other hand, it is important to remark, as a consequence of (2.5) and (2.6), that the outgoing waves are absorbed by the far field. Motivated by this fact, and aiming to obtain a suitable simplification of our model problem, we now introduce a sufficiently large circle Γ centered at the origin, define Ω_f as the annular domain bounded by Σ and Γ (see Figure 2.1), and consider the Robin boundary condition:

$$\frac{\partial p}{\partial \boldsymbol{\nu}} \;-\; \imath\,\kappa_f\,p \;=\; g \;:=\; \frac{\partial p_i}{\partial \boldsymbol{\nu}} \;-\; \imath\,\kappa_f\,p_i \quad \text{on } \Gamma\,, \tag{2.7}$$

where $\boldsymbol{\nu}$ denotes also the unit outward normal on Γ. Actually, in order to avoid introducing later a nonconforming Galerkin scheme, we may simply think of Γ as the polygonal curve resulting after joining with straight lines the points defining a uniform partition of the given circle. Alternatively, we could take Γ as an arbitrary closed curve sufficiently far away from Σ and consider the Dirichlet boundary condition

$$p \;=\; p_i \quad \text{on } \Gamma\,. \tag{2.8}$$

It is worth mentioning here that, irrespectively of the boundary condition chosen on Γ, the main idea is to reduce the original problem posed in $\mathbb{R}^2$ to an interaction problem on the bounded domain $\Omega_s \cup \Sigma \cup \Omega_f$. Nevertheless, we also remark that the Sommerfeld condition seems to be the most suitable choice since it shows a higher order of approximation (as $\mathtt{r} \to +\infty$) and constitutes the right condition guaranteeing the uniqueness of the exterior Helmholtz problem. Another possibility, which will be reported in a separate work, is to employ the boundary integral equation method in the unbounded region $\mathbb{R}^2 \backslash \overline{\Omega_s \cup \Omega_f}$. Further techniques, including Dirichlet-to-Neumann mappings, infinite elements, and PML approaches, are also available in the literature (see, e.g., [3, 7, 12, 13, 18, 19, 22] and the references therein).

Therefore, throughout the rest of the paper we assume the Robin boundary condition (2.7) and, given $\mathbf{f} \in [L^2(\Omega_s)]^2$ and $g \in H^{-1/2}(\Gamma)$, consider the following fluid-solid

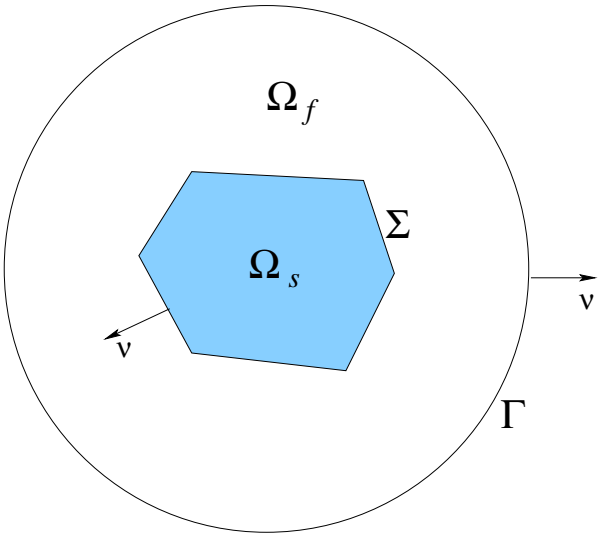


FIG. 2.1. *Geometry of the interaction problem.*

interaction problem: Find $\boldsymbol{\sigma} \in \mathbf{H}(\mathbf{div}; \Omega_s)$, $\mathbf{u} \in [L^2(\Omega_s)]^2$, and $p \in H^1(\Omega_f)$ such that the following hold in the distributional sense:

$$
\begin{aligned}
\boldsymbol{\sigma} &= \mathcal{C}\,\boldsymbol{\varepsilon}(\mathbf{u}) && \text{in } \Omega_s\,,\\
\mathbf{div}(\boldsymbol{\sigma}) + \kappa_s^2\,\mathbf{u} &= -\,\mathbf{f} && \text{in } \Omega_s\,,\\
\Delta p + \kappa_f^2\, p &= 0 && \text{in } \Omega_f\,,\\
\boldsymbol{\sigma}\boldsymbol{\nu} &= -\,p\boldsymbol{\nu} && \text{on } \Sigma\,,\\
\rho_f\,\omega^2\,\mathbf{u}\cdot\boldsymbol{\nu} &= \frac{\partial p}{\partial \boldsymbol{\nu}} && \text{on } \Sigma\,,\\
\frac{\partial p}{\partial \boldsymbol{\nu}} - \imath\,\kappa_f\, p &= g && \text{on } \Gamma\,,
\end{aligned}
\tag{2.9}
$$

where the wave number κ_s of the solid is defined by $\sqrt{\rho_s}\,\omega$.

3. The continuous variational formulation. In this section we employ primal and dual-mixed approaches in the fluid Ω_f and the solid Ω_s, respectively, to derive the full continuous variational formulation of (2.9). In fact, we first multiply the acoustic equation by $q \in H^1(\Omega_f)$, integrate by parts, and use the Robin boundary condition, to obtain

$$
\int_{\Omega_f} \nabla p\cdot\nabla q - \kappa_f^2 \int_{\Omega_f} pq + \left\langle \frac{\partial p}{\partial \boldsymbol{\nu}}, q \right\rangle_{\Sigma} - \imath\,\kappa_f \int_{\Gamma} pq = \langle g, q\rangle_{\Gamma}\,,
\tag{3.1}
$$

where, given $\mathcal{S} \in \{\Sigma, \Gamma\}$, $\langle\cdot,\cdot\rangle_{\mathcal{S}}$ stands for the duality pairing of $H^{-1/2}(\mathcal{S})$ and $H^{1/2}(\mathcal{S})$ with respect to the $L^2(\mathcal{S})$-inner product. Next, we use the second transmission condition and replace $\frac{\partial p}{\partial \boldsymbol{\nu}}$ by $\rho_f\,\omega^2\,\mathbf{u}\cdot\boldsymbol{\nu}$ on Σ, introduce the auxiliary unknown

$$
\boldsymbol{\varphi} := \mathbf{u}|_{\Sigma} \in [H^{1/2}(\Sigma)]^2\,,
$$

and divide by $\rho_f\,\omega^2$, whence (3.1) becomes

$$\frac{1}{\rho_f\,\omega^2}\int_{\Omega_f}\nabla p\cdot\nabla q\,-\,\frac{\kappa_f^2}{\rho_f\,\omega^2}\int_{\Omega_f}pq\,+\,\langle q\boldsymbol{\nu},\boldsymbol{\varphi}\rangle_\Sigma\,-\,\imath\,\frac{\kappa_f}{\rho_f\,\omega^2}\int_\Gamma pq \\ =\,\frac{1}{\rho_f\,\omega^2}\langle g,q\rangle_\Gamma\,, \tag{3.2}$$

where $\langle\,\cdot,\cdot\,\rangle_\Sigma$ denotes, from now on, the duality pairing of $[H^{-1/2}(\Sigma)]^2$ and $[H^{1/2}(\Sigma)]^2$ with respect to the $[L^2(\Sigma)]^2$-inner product.

On the other hand, in order to derive the mixed variational formulation in the solid Ω_s, we follow the usual procedure (see [1] and [31]) and introduce the rotation

$$\boldsymbol{\gamma}\,:=\,\frac{1}{2}(\nabla\mathbf{u}-(\nabla\mathbf{u})^{\mathtt{t}})\,\in\,[L^2(\Omega_s)]^{2\times2}_{\mathtt{asym}}$$

as a further unknown, where $[L^2(\Omega_s)]^{2\times2}_{\mathtt{asym}}$ denotes the space of asymmetric tensors with entries in $L^2(\Omega_s)$. In this way, the constitutive equation can be rewritten in the form

$$\mathcal{C}^{-1}\,\boldsymbol{\sigma}\,=\,\boldsymbol{\varepsilon}(\mathbf{u})\,=\,\nabla\mathbf{u}\,-\,\boldsymbol{\gamma}\,,$$

which, multiplying by $\boldsymbol{\tau}\in\mathbf{H}(\mathbf{div};\,\Omega_s)$ and integrating by parts, yields

$$\int_{\Omega_s}\mathcal{C}^{-1}\,\boldsymbol{\sigma}:\boldsymbol{\tau}\,+\,\int_{\Omega_s}\mathbf{u}\cdot\mathbf{div}(\boldsymbol{\tau})\,-\,\langle\boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{\varphi}\rangle_\Sigma\,+\,\int_{\Omega_s}\boldsymbol{\tau}:\boldsymbol{\gamma}\,=\,0\,. \tag{3.3}$$

Then from the elastodynamic equation we get

$$\mathbf{u}\,=\,-\,\frac{1}{\kappa_s^2}\,\big(\,\mathbf{f}\,+\,\mathbf{div}(\boldsymbol{\sigma})\big)\,, \tag{3.4}$$

which, replaced back into (3.3), gives

$$\int_{\Omega_s}\mathcal{C}^{-1}\,\boldsymbol{\sigma}:\boldsymbol{\tau}\,-\,\frac{1}{\kappa_s^2}\int_{\Omega_s}\mathbf{div}(\boldsymbol{\sigma})\cdot\mathbf{div}(\boldsymbol{\tau})\,-\,\langle\boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{\varphi}\rangle_\Sigma\,+\,\int_{\Omega_s}\boldsymbol{\tau}:\boldsymbol{\gamma} \\ =\,\frac{1}{\kappa_s^2}\int_{\Omega_s}\mathbf{f}\cdot\mathbf{div}(\boldsymbol{\tau})\,. \tag{3.5}$$

Finally, the symmetry of $\boldsymbol{\sigma}$ and the first transmission condition on Σ (see (2.4) or (2.9)) are imposed weakly through the relations

$$\int_{\Omega_s}\boldsymbol{\sigma}:\boldsymbol{\eta}\,=\,0\qquad\forall\,\boldsymbol{\eta}\,\in\,[L^2(\Omega_s)]^{2\times2}_{\mathtt{asym}} \tag{3.6}$$

and

$$\langle p\boldsymbol{\nu}\,+\,\boldsymbol{\sigma}\boldsymbol{\nu},\boldsymbol{\psi}\rangle_\Sigma\,=\,0\qquad\forall\,\boldsymbol{\psi}\,\in\,[H^{1/2}(\Sigma)]^2\,. \tag{3.7}$$

It is clear from (3.2) and (3.7), as already announced in the introduction, that the transmission conditions on Σ, say $\frac{\partial p}{\partial\boldsymbol{\nu}}\,=\,\rho_f\,\omega^2\,\mathbf{u}\cdot\boldsymbol{\nu}$ and $p\boldsymbol{\nu}\,=\,-\,\boldsymbol{\sigma}\boldsymbol{\nu}$, are natural and essential, respectively. In particular, the trace $\boldsymbol{\varphi}\,:=\,\mathbf{u}|_\Sigma\,\in\,[H^{1/2}(\Sigma)]^2$ constitutes the Lagrange multiplier associated with (3.7). If dual-mixed formulations were employed

in the solid and in the fluid, then both transmission conditions would be essential and therefore, besides $\mathbf{u}|_\Sigma$, the trace $p|_\Sigma$ would be required as a second Lagrange multiplier.

Now, adding (3.2) and (3.5), and subtracting (3.7) from (3.6), we arrive at the following variational formulation of (2.9): Find $((\boldsymbol{\sigma},p),(\boldsymbol{\varphi},\boldsymbol{\gamma})) \in \mathbf{H}\times\mathbf{Q}$ such that

$$\begin{aligned} A((\boldsymbol{\sigma},p),(\boldsymbol{\tau},q)) + B_1((\boldsymbol{\tau},q),(\boldsymbol{\varphi},\boldsymbol{\gamma})) &= F(\boldsymbol{\tau},q) \quad \forall\,(\boldsymbol{\tau},q)\in\mathbf{H}\,, \\ B_2((\boldsymbol{\sigma},p),(\boldsymbol{\psi},\boldsymbol{\eta})) &= 0 \qquad\quad \forall\,(\boldsymbol{\psi},\boldsymbol{\eta})\in\mathbf{Q}\,, \end{aligned} \tag{3.8}$$

where $\mathbf{H}$ and $\mathbf{Q}$ are the product spaces

$$\mathbf{H} := \mathbf{H}(\mathbf{div};\Omega_s)\times H^1(\Omega_f)\,, \quad \mathbf{Q} := [H^{1/2}(\Sigma)]^2\times[L^2(\Omega_s)]^{2\times 2}_{\mathtt{asym}}\,, \tag{3.9}$$

$F:\mathbf{H}\to\mathbb{C}$ is the linear functional

$$F(\boldsymbol{\tau},q) := \frac{1}{\kappa_s^2}\int_{\Omega_s}\mathbf{f}\cdot\mathbf{div}(\boldsymbol{\tau}) + \frac{1}{\rho_f\,\omega^2}\,\langle g,q\rangle_\Gamma \qquad \forall\,(\boldsymbol{\tau},q)\in\mathbf{H}\,, \tag{3.10}$$

and $A:\mathbf{H}\times\mathbf{H}\to\mathbb{C}$, $B_1:\mathbf{H}\times\mathbf{Q}\to\mathbb{C}$, and $B_2:\mathbf{H}\times\mathbf{Q}\to\mathbb{C}$ are the bilinear forms defined by

$$\begin{aligned} A((\boldsymbol{\zeta},r),(\boldsymbol{\tau},q)) &:= \int_{\Omega_s}\mathcal{C}^{-1}\,\boldsymbol{\zeta}:\boldsymbol{\tau} - \frac{1}{\kappa_s^2}\int_{\Omega_s}\mathbf{div}(\boldsymbol{\zeta})\cdot\mathbf{div}(\boldsymbol{\tau}) + \frac{1}{\rho_f\,\omega^2}\int_{\Omega_f}\nabla r\cdot\nabla q \\ &\quad - \frac{\kappa_f^2}{\rho_f\,\omega^2}\int_{\Omega_f} rq - \imath\,\frac{\kappa_f}{\rho_f\,\omega^2}\int_\Gamma rq \qquad \forall\,(\boldsymbol{\zeta},r),\,(\boldsymbol{\tau},q)\in\mathbf{H}\,, \end{aligned} \tag{3.11}$$

$$B_1((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) := \langle q\boldsymbol{\nu} - \boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{\psi}\rangle_\Sigma + \int_{\Omega_s}\boldsymbol{\tau}:\boldsymbol{\eta}\,, \tag{3.12}$$

and

$$B_2((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) := -\langle q\boldsymbol{\nu}+\boldsymbol{\tau}\boldsymbol{\nu},\boldsymbol{\psi}\rangle_\Sigma + \int_{\Omega_s}\boldsymbol{\tau}:\boldsymbol{\eta} \tag{3.13}$$

for all $(\boldsymbol{\tau},q)\in\mathbf{H}$, $(\boldsymbol{\psi},\boldsymbol{\eta})\in\mathbf{Q}$. It is easy to see that F, A, B_1, and B_2 are all bounded with constants depending on ω, ρ_f, ρ_s, κ_f, and κ_s, in the case of F and A, and constants independent of the physical parameters for B_1 and B_2. Concerning the form A, we also observe from (2.2) that the inverse operator $\mathcal{C}^{-1}$ reduces to

$$\mathcal{C}^{-1}\,\boldsymbol{\zeta} := \frac{1}{2\,\mu}\,\boldsymbol{\zeta} - \frac{\lambda}{4\,\mu\,(\lambda+\mu)}\,\mathrm{tr}(\boldsymbol{\zeta})\,\mathbf{I} \qquad \forall\,\boldsymbol{\zeta}\in[L^2(\Omega_s)]^{2\times 2}\,, \tag{3.14}$$

which implies that

$$\int_{\Omega_s}\mathcal{C}^{-1}\,\boldsymbol{\zeta}:\boldsymbol{\tau} = \frac{1}{2\,\mu}\int_{\Omega_s}\boldsymbol{\zeta}^{\mathtt{d}}:\boldsymbol{\tau}^{\mathtt{d}} + \frac{1}{4\,(\lambda+\mu)}\int_{\Omega_s}\mathrm{tr}(\boldsymbol{\zeta})\,\mathrm{tr}(\boldsymbol{\tau}) \tag{3.15}$$

for all $\boldsymbol{\zeta},\,\boldsymbol{\tau}\in[L^2(\Omega_s)]^{2\times 2}$, and hence

$$\int_{\Omega_s}\mathcal{C}^{-1}\,\boldsymbol{\zeta}:\overline{\boldsymbol{\zeta}} \;\geq\; \frac{1}{2\mu}\,\|\boldsymbol{\zeta}^{\mathtt{d}}\|^2_{[L^2(\Omega_s)]^{2\times 2}} \qquad \forall\,\boldsymbol{\zeta}\in[L^2(\Omega_s)]^{2\times 2}\,. \tag{3.16}$$

This estimate will be useful for our analysis below.

We end this section by commenting that, instead of eliminating the displacement $\mathbf{u}$ by means of (3.4), the classical method for the mixed formulation of the elasticity problem keeps this unknown in (3.3) and incorporates the equation

$$\int_{\Omega_s} \mathbf{v} \cdot \mathbf{div}(\boldsymbol{\sigma}) + \kappa_s^2 \int_{\Omega_s} \mathbf{u} \cdot \mathbf{v} = - \int_{\Omega_s} \mathbf{f} \cdot \mathbf{v} \,,$$

which arises after multiplying the elastodynamic equation by a test function $\mathbf{v} \in [L^2(\Omega_s]^2$. However, the resulting variational formulation does not fit into the framework of any of the available theories for proving well-posedness, and hence a different approach, such as the one proposed in the present paper, must be adopted. An alternative technique, which makes extensive use of duality arguments, was developed in [16] for the coupling of a mixed variational formulation and the boundary integral equation method when applied to an exterior Helmholtz problem in the plane.

4. Analysis of the continuous variational formulation. In this section we proceed analogously to [11] and employ a suitable decomposition of $\mathbf{H}(\mathbf{div}; \Omega_s)$ to show that (3.8) becomes a compact perturbation of a well-posed problem. First, we need to consider an elasticity problem in Ω_s with Neumann boundary conditions. Then this auxiliary problem yields the definition of an associated operator, which is employed to obtain the above-mentioned decomposition.

4.1. Preliminaries. Let $\mathbb{RM}(\Omega_s)$ be the space of rigid body motions in Ω_s, that is,

$$\mathbb{RM}(\Omega_s) := \left\{ \mathbf{v} : \Omega_s \to \mathbb{C}^2 \,:\, \mathbf{v}(\mathbf{x}) = \begin{pmatrix} a \\ b \end{pmatrix} + c \begin{pmatrix} x_2 \\ -x_1 \end{pmatrix} \right.$$
$$\left. \forall\, \mathbf{x} := \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \Omega_s \,,\; a,\, b \,,\, c \in \mathbb{C} \right\},$$

and let $\mathbf{M} : [L^2(\Omega_s)]^2 \to \mathbb{RM}(\Omega_s)$ be the $[L^2(\Omega_s)]^2$-orthogonal projector. Then, given $\boldsymbol{\tau} \in \mathbf{H}(\mathbf{div}; \Omega_s)$, we let $(\tilde{\boldsymbol{\sigma}}, \tilde{\mathbf{u}}, \tilde{\boldsymbol{\gamma}}) \in \mathbf{H}(\mathbf{div}; \Omega_s) \times (\mathbf{I} - \mathbf{M})([L^2(\Omega_s)]^2) \times [L^2(\Omega_s)]^{2\times 2}_{\mathtt{asym}}$ be the unique solution (see [1], [9, Theorem 9.2.30], [15]) of the dual-mixed variational formulation of the boundary value problem

$$\tilde{\boldsymbol{\sigma}} = \mathcal{C}\, \boldsymbol{\varepsilon}(\tilde{\mathbf{u}}) \,, \quad \mathbf{div}\, \tilde{\boldsymbol{\sigma}} = (\mathbf{I} - \mathbf{M})(\mathbf{div}\, (\boldsymbol{\tau})) \quad \text{in } \Omega_s \,, \quad \tilde{\boldsymbol{\sigma}}\, \boldsymbol{\nu} = \mathbf{0} \quad \text{on } \Sigma \,, \tag{4.1}$$

where $\tilde{\boldsymbol{\gamma}} := \frac{1}{2}(\nabla \tilde{\mathbf{u}} - (\nabla \tilde{\mathbf{u}})^{\mathtt{t}})$ denotes the auxiliary unknown named rotation, $\mathcal{C}\, \boldsymbol{\varepsilon}(\tilde{\mathbf{u}})$ is defined according to (2.2), and $\mathbf{I}$ stands for a generic identity operator. Owing to the regularity result for the elasticity problem with Neumann boundary conditions (see, e.g., [20, 21]), we know that there exists $\epsilon > 0$ such that

$$(\tilde{\boldsymbol{\sigma}}, \tilde{\mathbf{u}}, \tilde{\boldsymbol{\gamma}}) \in [H^{\epsilon}(\Omega_s)]^{2\times 2} \times [H^{1+\epsilon}(\Omega_s)]^2 \times [H^{\epsilon}(\Omega_s)]^{2\times 2} \tag{4.2}$$

and

$$\|\tilde{\boldsymbol{\sigma}}\|_{[H^{\epsilon}(\Omega_s)]^{2\times 2}} + \|\tilde{\mathbf{u}}\|_{[H^{1+\epsilon}(\Omega_s)]^2} + \|\tilde{\boldsymbol{\gamma}}\|_{[H^{\epsilon}(\Omega_s)]^{2\times 2}} \leq C\, \|\mathbf{div}(\boldsymbol{\tau})\|_{[L^2(\Omega_s)]^2} \,. \tag{4.3}$$

We now introduce the linear operator

$$\begin{aligned} \mathbf{P} : \mathbf{H}(\mathbf{div}; \Omega_s) &\to \mathbf{H}(\mathbf{div}; \Omega_s), \\ \boldsymbol{\tau} &\to \mathbf{P}(\boldsymbol{\tau}) := \tilde{\boldsymbol{\sigma}} \end{aligned} \tag{4.4}$$

and observe from (4.1) that

$$\begin{aligned} \mathbf{div}\,\mathbf{P}(\boldsymbol{\tau}) \,&=\, (\mathbf{I}-\mathbf{M})(\mathbf{div}\,(\boldsymbol{\tau})) \quad \text{in } \Omega_s\,, \\ \mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\tau}) \,&=\, \mathbf{M}(\mathbf{div}\,(\boldsymbol{\tau})) \quad \text{in } \Omega_s\,, \end{aligned} \tag{4.5}$$

and

$$\mathbf{P}(\boldsymbol{\tau})\,\boldsymbol{\nu} \,=\, \mathbf{0} \quad \text{on } \Sigma\,. \tag{4.6}$$

Then, using the continuous dependence result for (4.1), we find that

$$\|\mathbf{P}(\boldsymbol{\tau})\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)} \,\leq\, C\,\|\mathbf{div}(\boldsymbol{\tau})\|_{[L^2(\Omega_s)]^2} \qquad \forall\,\boldsymbol{\tau}\,\in\,\mathbf{H}(\mathbf{div};\,\Omega_s)\,,$$

which shows that $\mathbf{P}$ is bounded. Moreover, it is easy to see that $\mathbf{P}$ is actually a linear projection, and hence

$$\mathbf{H}(\mathbf{div};\,\Omega_s) \,=\, \mathbf{P}(\mathbf{H}(\mathbf{div};\,\Omega_s)) \,\oplus\, (\mathbf{I}-\mathbf{P})(\mathbf{H}(\mathbf{div};\,\Omega_s))\,. \tag{4.7}$$

Finally, it is clear from (4.2) and (4.3) that $\mathbf{P}(\boldsymbol{\tau}) \,\in\, [H^{\epsilon}(\Omega_s)]^{2\times 2}$ and

$$\|\mathbf{P}(\boldsymbol{\tau})\|_{[H^{\epsilon}(\Omega_s)]^{2\times 2}} \,\leq\, C\,\|\mathbf{div}(\boldsymbol{\tau})\|_{[L^2(\Omega_s)]^2} \qquad \forall\,\boldsymbol{\tau}\,\in\,\mathbf{H}(\mathbf{div};\,\Omega_s)\,. \tag{4.8}$$

4.2. Well-posedness of the continuous formulation. In order to show that our coupled problem (3.8) is well posed, we now employ the stable decomposition (4.7) to reformulate (3.8) in a more suitable form. We begin by observing, according to (4.5), (4.6), the symmetry of $\mathbf{P}(\boldsymbol{\tau})$, and the fact that $\nabla\mathbf{v} \,\in\, [L^2(\Omega_s)]^{2\times 2}_{\mathtt{asym}}$ for all $\mathbf{v} \,\in\, \mathbb{RM}(\Omega_s)$, that for all $\boldsymbol{\zeta}$, $\boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\,\Omega_s)$ there holds

$$\begin{aligned} \int_{\Omega_s} \mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})\cdot\mathbf{div}\,\mathbf{P}(\boldsymbol{\tau}) \,&=\, \int_{\Omega_s} \mathbf{M}(\mathbf{div}(\boldsymbol{\zeta}))\cdot\mathbf{div}\,\mathbf{P}(\boldsymbol{\tau}) \\ &=\, -\int_{\Omega_s} \nabla\mathbf{M}(\mathbf{div}(\boldsymbol{\zeta})) : \mathbf{P}(\boldsymbol{\tau}) \,+\, \langle\,\mathbf{P}(\boldsymbol{\tau})\boldsymbol{\nu}, \mathbf{M}(\mathbf{div}(\boldsymbol{\zeta}))\,\rangle_{\Sigma} \,=\, 0\,. \end{aligned} \tag{4.9}$$

Then, writing $\boldsymbol{\zeta} \,=\, \mathbf{P}(\boldsymbol{\zeta}) + (\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})$ and $\boldsymbol{\tau} \,=\, \mathbf{P}(\boldsymbol{\tau}) + (\mathbf{I}-\mathbf{P})(\boldsymbol{\tau})$ in (3.11), using the identity (4.9), and adding and subtracting the terms

$$\int_{\Omega_s} \mathcal{C}^{-1}\,\mathbf{P}(\boldsymbol{\zeta}) : \mathbf{P}(\boldsymbol{\tau})\,, \quad \int_{\Omega_s} \mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})\cdot\mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\tau})\,, \quad \text{and} \quad \frac{1}{\rho_f\,\omega^2}\int_{\Omega_f} rq\,,$$

we find that A can be decomposed as

$$A((\boldsymbol{\zeta},r),(\boldsymbol{\tau},q)) \,=\, A_0((\boldsymbol{\zeta},r),(\boldsymbol{\tau},q)) \,+\, K_0((\boldsymbol{\zeta},r),(\boldsymbol{\tau},q)) \tag{4.10}$$

for all $(\boldsymbol{\zeta},r)\,,(\boldsymbol{\tau},q) \,\in\, \mathbf{H}$, where $A_0 : \mathbf{H}\times\mathbf{H} \to \mathbb{C}$ and $K_0 : \mathbf{H}\times\mathbf{H} \to \mathbb{C}$ are the bounded and symmetric bilinear forms given by

(4.11)

$$\begin{aligned} A_0((\boldsymbol{\zeta},r),(\boldsymbol{\tau},q)) \,:=\, &-\int_{\Omega_s} \mathcal{C}^{-1}\,\mathbf{P}(\boldsymbol{\zeta}) : \mathbf{P}(\boldsymbol{\tau}) \,-\, \frac{1}{\kappa_s^2}\int_{\Omega_s} \mathbf{div}\,\mathbf{P}(\boldsymbol{\zeta})\cdot\mathbf{div}\,\mathbf{P}(\boldsymbol{\tau}) \\ &+\int_{\Omega_s} \mathcal{C}^{-1}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}) : (\mathbf{I}-\mathbf{P})(\boldsymbol{\tau}) + \int_{\Omega_s} \mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})\cdot\mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\tau}) \\ &+\frac{1}{\rho_f\,\omega^2}\int_{\Omega_f} \nabla r\cdot\nabla + \frac{1}{\rho_f\,\omega^2}\int_{\Omega_f} rq - \imath\,\frac{\kappa_f}{\rho_f\,\omega^2}\int_{\Gamma} rq\,, \end{aligned}$$

and

(4.12)

$$K_0((\boldsymbol{\zeta},r),(\boldsymbol{\tau},q)) := 2\int_{\Omega_s}\mathcal{C}^{-1}\,\mathbf{P}(\boldsymbol{\zeta}):\mathbf{P}(\boldsymbol{\tau}) + \int_{\Omega_s}\mathcal{C}^{-1}\,\mathbf{P}(\boldsymbol{\zeta}):(\mathbf{I}-\mathbf{P})(\boldsymbol{\tau}) + \int_{\Omega_s}\mathcal{C}^{-1}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}):\mathbf{P}(\boldsymbol{\tau}) - \left(1+\frac{1}{\kappa_s^2}\right)\int_{\Omega_s}\mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})\cdot\mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\tau}) - \frac{(1+\kappa_f^2)}{\rho_f\,\omega^2}\int_{\Omega_f} rq\,.$$

On the other hand, we easily deduce from (3.12) and (3.13) that B_1 and B_2 can be decomposed as

$$B_1((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) = B_0((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) + K_1((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) \tag{4.13}$$

and

$$B_2((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) = B_0((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) - K_1((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) \tag{4.14}$$

for all $(\boldsymbol{\tau},q)\in\mathbf{H}$, $(\boldsymbol{\psi},\boldsymbol{\eta})\in\mathbf{Q}$, where $B_0:\mathbf{H}\times\mathbf{H}\to\mathbb{C}$ and $K_1:\mathbf{H}\times\mathbf{H}\to\mathbb{C}$ are the bounded bilinear forms defined by

$$B_0((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) := -\langle\boldsymbol{\tau}\,\boldsymbol{\nu},\boldsymbol{\psi}\rangle_\Sigma + \int_{\Omega_s}\boldsymbol{\tau}:\boldsymbol{\eta} \tag{4.15}$$

and

$$K_1((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta})) := \langle q\boldsymbol{\nu},\boldsymbol{\psi}\rangle_\Sigma\,. \tag{4.16}$$

Next, we let $\mathbf{A}_0:\mathbf{H}\to\mathbf{H}$, $\mathbf{B}_0:\mathbf{H}\to\mathbf{Q}$, $\mathbf{K}_0:\mathbf{H}\to\mathbf{H}$, and $\mathbf{K}_1:\mathbf{H}\to\mathbf{Q}$ be the linear and bounded operators induced by the corresponding bilinear forms. In addition, we let $\mathbf{B}_0^*:\mathbf{Q}\to\mathbf{H}$ and $\mathbf{K}_1^*:\mathbf{Q}\to\mathbf{H}$ be the associated adjoint operators and denote by $\mathbf{F}$ the Riesz representant of F. Hence, using these notations and taking into account the decompositions (4.10), (4.13), and (4.14), our variational formulation (3.8) can be rewritten as the following operator equation: Find $((\boldsymbol{\sigma},p),(\boldsymbol{\varphi},\boldsymbol{\gamma}))\in\mathbf{H}\times\mathbf{Q}$ such that

$$\begin{pmatrix}\mathbf{A}_0 & \mathbf{B}_0^*\\ \mathbf{B}_0 & \mathbf{0}\end{pmatrix}\begin{pmatrix}(\boldsymbol{\sigma},p)\\(\boldsymbol{\varphi},\boldsymbol{\gamma})\end{pmatrix} + \begin{pmatrix}\mathbf{K}_0 & \mathbf{K}_1^*\\ -\mathbf{K}_1 & \mathbf{0}\end{pmatrix}\begin{pmatrix}(\boldsymbol{\sigma},p)\\(\boldsymbol{\varphi},\boldsymbol{\gamma})\end{pmatrix} = \begin{pmatrix}\mathbf{F}\\\mathbf{0}\end{pmatrix}. \tag{4.17}$$

Throughout the rest of this section we prove that the matrix operators on the left-hand side of (4.17) become invertible and compact, respectively.

The following two lemmas are needed to prove the continuous inf-sup condition for B_0. In particular, we notice that Lemma 4.1 provides a right inverse of the normal traces of functions in $\mathbf{H}(\mathbf{div};\Omega_s)$.

LEMMA 4.1. *There exists a bounded and linear operator* $\mathbf{S}:[H^{-1/2}(\Sigma)]^2\to\mathbf{H}(\mathbf{div};\Omega_s)$ *such that*

$$\mathbf{S}(\boldsymbol{\xi})\,\boldsymbol{\nu} = -\boldsymbol{\xi}\quad \text{on } \Sigma\qquad \forall\,\boldsymbol{\xi}\in[H^{-1/2}(\Sigma)]^2\,.$$

Proof. Given $\boldsymbol{\xi}:=(\xi_1,\xi_2)^{\mathtt{t}}\in[H^{-1/2}(\Sigma)]^2$, we let $\mathbf{S}(\boldsymbol{\xi}):=\nabla\mathbf{z}$ where $\mathbf{z}\in[H^1(\Omega_s)]^2$ is the unique weak solution, up to a constant vector in $\mathbb{C}^2$, of the boundary value problem with Neumann boundary conditions

$$-\Delta\,\mathbf{z} = \frac{1}{|\Omega_s|}\,(\langle\xi_1,1\rangle_\Sigma,\langle\xi_2,1\rangle_\Sigma)^{\mathtt{t}}\quad\text{in }\Omega_s\,,\quad \nabla\mathbf{z}\,\boldsymbol{\nu} = -\boldsymbol{\xi}\quad\text{on }\Sigma\,. \tag{4.18}$$

It follows easily that $\mathbf{S}(\boldsymbol{\xi})$ belongs to $\mathbf{H}(\mathbf{div};\Omega_s)$ and satisfies the required conditions. In particular, the fact that $\mathbf{div}(\mathbf{S}(\boldsymbol{\xi})) = -\frac{1}{|\Omega_s|}\,(\langle\xi_1,1\rangle_\Sigma,\langle\xi_2,1\rangle_\Sigma)^{\mathtt{t}}$ in Ω_s and the continuous dependence result for (4.18) imply the boundedness of $\mathbf{S}$. ∎

LEMMA 4.2. *There exists a bounded and linear operator* $\mathbf{T} : [L^2(\Omega_s)]^{2\times 2}_{\mathtt{asym}} \to \mathbf{H}(\mathbf{div};\Omega_s)$ *such that*

$$\mathbf{T}(\boldsymbol{\eta})\,\boldsymbol{\nu} = 0 \quad \text{on } \Sigma \quad \text{and} \quad \frac{1}{2}\left(\mathbf{T}(\boldsymbol{\eta}) - \mathbf{T}(\boldsymbol{\eta})^{\mathtt{t}}\right) = \boldsymbol{\eta} \qquad \forall\,\boldsymbol{\eta} \in [L^2(\Omega_s)]^{2\times 2}_{\mathtt{asym}}\,.$$

Proof. The proof is a slight variation of the proof of Lemma 4.4 in [17] (see also Lemma 4.2 in [6]). ∎

We are now in a position to prove that B_0 satisfies the continuous inf-sup condition.

LEMMA 4.3. *There exists* $C_1 > 0$ *such that*

$$\sup_{\substack{(\boldsymbol{\tau},q)\in\mathbf{H}\\(\boldsymbol{\tau},q)\neq\mathbf{0}}} \frac{|\,B_0((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta}))\,|}{\|(\boldsymbol{\tau},q)\|_{\mathbf{H}}} \geq C_1\,\|(\boldsymbol{\psi},\boldsymbol{\eta})\|_{\mathbf{Q}} \qquad \forall\,(\boldsymbol{\psi},\boldsymbol{\eta}) \in \mathbf{Q}\,. \tag{4.19}$$

Proof. Given $(\boldsymbol{\psi},\boldsymbol{\eta}) \in \mathbf{Q}$, we consider the linear operators $\mathbf{S}$ and $\mathbf{T}$ defined in the previous lemmas and observe from the definition of B_0 (cf. (4.15)) that

$$B_0((\mathbf{S}(\boldsymbol{\xi}),0),(\boldsymbol{\psi},\boldsymbol{\eta})) := \langle\boldsymbol{\xi},\boldsymbol{\psi}\rangle_\Sigma + \int_{\Omega_s}\mathbf{S}(\boldsymbol{\xi}) : \boldsymbol{\eta} \qquad \forall\,\boldsymbol{\xi} \in [H^{-1/2}(\Sigma)]^2$$

and

$$B_0((\mathbf{T}(\overline{\boldsymbol{\eta}}),0),(\boldsymbol{\psi},\boldsymbol{\eta})) := \int_{\Omega_s}\mathbf{T}(\overline{\boldsymbol{\eta}}) : \boldsymbol{\eta} = \int_{\Omega_s}\frac{1}{2}\left(\mathbf{T}(\overline{\boldsymbol{\eta}}) - \mathbf{T}(\overline{\boldsymbol{\eta}})^{\mathtt{t}}\right) : \boldsymbol{\eta} = \|\boldsymbol{\eta}\|^2_{[L^2(\Omega_s)]^{2\times 2}}\,.$$

Then, employing the boundedness of $\mathbf{S}$ and $\mathbf{T}$, respectively, we find that

$$\begin{aligned}
&\sup_{\substack{(\boldsymbol{\tau},q)\in\mathbf{H}\\(\boldsymbol{\tau},q)\neq\mathbf{0}}} \frac{|\,B_0((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta}))\,|}{\|(\boldsymbol{\tau},q)\|_{\mathbf{H}}} \geq \sup_{\substack{\boldsymbol{\xi}\in[H^{-1/2}(\Sigma)]^2\\\boldsymbol{\xi}\neq\mathbf{0}}} \frac{|\,B_0((\mathbf{S}(\boldsymbol{\xi}),0),(\boldsymbol{\psi},\boldsymbol{\eta}))\,|}{\|(\mathbf{S}(\boldsymbol{\xi}),0)\|_{\mathbf{H}}} \\
&= \sup_{\substack{\boldsymbol{\xi}\in[H^{-1/2}(\Sigma)]^2\\\boldsymbol{\xi}\neq\mathbf{0}}} \frac{\left|\,\langle\boldsymbol{\xi},\boldsymbol{\psi}\rangle_\Sigma + \displaystyle\int_{\Omega_s}\mathbf{S}(\boldsymbol{\xi}) : \boldsymbol{\eta}\,\right|}{\|\mathbf{S}(\boldsymbol{\xi})\|_{\mathbf{H}(\mathbf{div};\Omega_s)}} \geq \frac{1}{\|\mathbf{S}\|}\,\|\boldsymbol{\psi}\|_{[H^{1/2}(\Sigma)]^2} - \|\boldsymbol{\eta}\|_{[L^2(\Omega_s)]^{2\times 2}}
\end{aligned} \tag{4.20}$$

and

$$\begin{aligned}
&\sup_{\substack{(\boldsymbol{\tau},q)\in\mathbf{H}\\(\boldsymbol{\tau},q)\neq\mathbf{0}}} \frac{|\,B_0((\boldsymbol{\tau},q),(\boldsymbol{\psi},\boldsymbol{\eta}))\,|}{\|(\boldsymbol{\tau},q)\|_{\mathbf{H}}} \geq \frac{|\,B_0((\mathbf{T}(\overline{\boldsymbol{\eta}}),0),(\boldsymbol{\psi},\boldsymbol{\eta}))\,|}{\|(\mathbf{T}(\overline{\boldsymbol{\eta}}),0)\|_{\mathbf{H}}} \\
&= \frac{\|\boldsymbol{\eta}\|^2_{[L^2(\Omega_s)]^{2\times 2}}}{\|\mathbf{T}(\overline{\boldsymbol{\eta}})\|_{\mathbf{H}(\mathbf{div};\Omega_s)}} \geq \frac{1}{\|\mathbf{T}\|}\,\|\boldsymbol{\eta}\|_{[L^2(\Omega_s)]^{2\times 2}}\,.
\end{aligned} \tag{4.21}$$

The estimates (4.20) and (4.21) imply (4.19) and complete the proof. ∎

Our next goal, according to the well-known Babuška–Brezzi theory, is to prove that $\mathbf{A}_0$ is an isomorphism on the kernel of $\mathbf{B}_0$. To this end, we now introduce the decomposition

$$\mathbf{H}(\mathbf{div};\Omega_s) = \mathbf{H}_0(\mathbf{div};\Omega_s) \oplus \mathbb{C}\,\mathbf{I}\,, \tag{4.22}$$

where

$$\mathbf{H}_0(\mathbf{div};\Omega_s) := \left\{ \boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\Omega_s) : \quad \int_{\Omega_s} \mathrm{tr}(\boldsymbol{\tau}) = 0 \right\}. \tag{4.23}$$

This means that for any $\boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\Omega_s)$ there exist unique $\boldsymbol{\tau}_0 \in \mathbf{H}_0(\mathbf{div};\Omega_s)$ and $d \in \mathbb{C}$ given by $d := \frac{1}{2\,|\Omega_s|}\int_{\Omega_s} \mathrm{tr}(\boldsymbol{\tau})$, where $|\Omega_s|$ denotes the measure of Ω_s, such that $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\,\mathbf{I}$.

The inequalities provided by the following three lemmas will be crucial in our subsequent analysis. We notice that Lemma 4.5 corresponds to Lemma 2.2 in [14], whose proof, being short and simple, is recalled here for the sake of completeness.

LEMMA 4.4. *There exists $c_1 > 0$, depending only on Ω_s, such that*

$$c_1 \,\|\boldsymbol{\tau}_0\|^2_{[L^2(\Omega_s)]^{2\times 2}} \le \|\boldsymbol{\tau}^{\mathtt{d}}\|^2_{[L^2(\Omega_s)]^{2\times 2}} + \|\mathbf{div}(\boldsymbol{\tau})\|^2_{[L^2(\Omega_s)]^2} \quad \forall\, \boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\Omega_s)\,. \tag{4.24}$$

Proof. See Lemma 3.1 in [2] or Proposition 3.1 of Chapter IV in [10]. ∎

LEMMA 4.5. *There exists $c_2 > 0$, depending only on Ω_s, such that*

$$c_2\,\|\boldsymbol{\tau}\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} \le \|\boldsymbol{\tau}_0\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} \quad \forall\, \boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\Omega_s) \quad \textit{such that} \quad \boldsymbol{\tau}\boldsymbol{\nu} = \mathbf{0} \quad \textit{on } \Sigma\,. \tag{4.25}$$

Proof. Given $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\,\mathbf{I} \in \mathbf{H}(\mathbf{div};\Omega_s)$, with $\boldsymbol{\tau}_0 \in \mathbf{H}_0(\mathbf{div};\Omega_s)$ and $d \in \mathbb{C}$, and such that $\boldsymbol{\tau}\boldsymbol{\nu} = \mathbf{0}$ on Σ, we note that $d\,\boldsymbol{\nu} = -\,\boldsymbol{\tau}_0\,\boldsymbol{\nu}$ on Σ, and hence, using the trace theorem of $\mathbf{H}(\mathbf{div};\Omega_s)$,

$$|d|\,\|\boldsymbol{\nu}\|_{[H^{-1/2}(\Sigma)]^2} = \|\boldsymbol{\tau}_0\boldsymbol{\nu}\|_{[H^{-1/2}(\Sigma)]^2} \le \tilde{c}_2\,\|\boldsymbol{\tau}_0\|_{\mathbf{H}(\mathbf{div};\Omega_s)}\,.$$

This inequality and the fact that $\|\boldsymbol{\tau}\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} = \|\boldsymbol{\tau}_0\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} + 2\,d^2\,|\Omega_s|$ imply (4.25). ∎

LEMMA 4.6. *Let $\Xi := (\mathbf{I} - 2\,\mathbf{P}) : \mathbf{H}(\mathbf{div};\Omega_s) \to \mathbf{H}(\mathbf{div};\Omega_s)$. Then there exists $C > 0$, depending on μ, c_1, c_2, κ_s, ρ_f, and ω^2, such that for each $(\boldsymbol{\zeta}, r) \in \mathbf{H}$ there holds*

$$\begin{aligned} &\mathrm{Re}\left\{ A_0((\boldsymbol{\zeta}, r), (\Xi(\overline{\boldsymbol{\zeta}}), \overline{r})) \right\} \\ &\quad \ge C \left\{ \|\mathbf{P}(\boldsymbol{\zeta})\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} + \|(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})_0\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} + \|r\|^2_{H^1(\Omega_f)} \right\}, \end{aligned} \tag{4.26}$$

where $(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})_0$ is the $\mathbf{H}_0(\mathbf{div};\Omega_s)$-component of $(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})$.

Proof. Using that $\mathbf{P}$ is a projector, we easily observe that

$$\mathbf{P}\,\Xi(\boldsymbol{\zeta}) = -\,\mathbf{P}(\boldsymbol{\zeta}) \quad \text{and} \quad (\mathbf{I}-\mathbf{P})\,\Xi(\boldsymbol{\zeta}) = (\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}) \qquad \forall\, \boldsymbol{\zeta} \in \mathbf{H}(\mathbf{div};\Omega_s)\,, \tag{4.27}$$

and hence, according to the definition of A_0 (cf. (4.11)), we obtain

$$\begin{aligned} A_0((\boldsymbol{\zeta}, r), (\Xi(\overline{\boldsymbol{\zeta}}), \overline{r})) &:= \int_{\Omega_s} \mathcal{C}^{-1}\,\mathbf{P}(\boldsymbol{\zeta}) : \overline{\mathbf{P}(\boldsymbol{\zeta})} + \frac{1}{\kappa_s^2}\int_{\Omega_s} \mathbf{div}\,\mathbf{P}(\boldsymbol{\zeta}) \cdot \overline{\mathbf{div}\,\mathbf{P}(\boldsymbol{\zeta})} \\ &+ \int_{\Omega_s} \mathcal{C}^{-1}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}) : \overline{(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})} + \int_{\Omega_s} \mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}) \cdot \overline{\mathbf{div}\,(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta})} \\ &+ \frac{1}{\rho_f\,\omega^2}\int_{\Omega_f} \|\nabla r\|^2 + \frac{1}{\rho_f\,\omega^2}\int_{\Omega_f} |r|^2 - \imath\,\frac{\kappa_f}{\rho_f\,\omega^2}\int_{\Gamma} |r|^2 \end{aligned} \tag{4.28}$$

for all $(\boldsymbol{\zeta}, r) \in \mathbf{H}$. It follows that

$$\begin{aligned}\mathrm{Re}\Big\{ A_0((\boldsymbol{\zeta}, r), (\Xi(\overline{\boldsymbol{\zeta}}), \overline{r})) \Big\} &:= \int_{\Omega_s} \mathcal{C}^{-1}\, \mathbf{P}(\boldsymbol{\zeta}) : \overline{\mathbf{P}(\boldsymbol{\zeta})} + \frac{1}{\kappa_s^2} \int_{\Omega_s} \mathbf{div}\, \mathbf{P}(\boldsymbol{\zeta}) \cdot \overline{\mathbf{div}\, \mathbf{P}(\boldsymbol{\zeta})} \\ &+ \int_{\Omega_s} \mathcal{C}^{-1}\, (\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta}) : \overline{(\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta})} + \int_{\Omega_s} \mathbf{div}\, (\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta}) \cdot \overline{\mathbf{div}\, (\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta})} \\ &+ \frac{1}{\rho_f\, \omega^2}\, \|r\|^2_{H^1(\Omega_f)}\,,\end{aligned}$$

which, applying (3.16) and Lemmas 4.4 and 4.5, yields (4.26). Note that Lemma 4.5 cannot be applied to $(\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta})$ since its normal trace does not necessarily vanish on Σ. ∎

The weak coercivity of A_0 is established next.

LEMMA 4.7. *Let* $\mathbf{V}$ *be the kernel of* $\mathbf{B}_0$*, that is,*

$$\mathbf{V} := \{\, (\boldsymbol{\tau}, q) \in \mathbf{H} : \quad B_0((\boldsymbol{\tau}, q), (\boldsymbol{\psi}, \boldsymbol{\eta})) = 0 \qquad \forall\, (\boldsymbol{\psi}, \boldsymbol{\eta}) \in \mathbf{Q} \,\}\,.$$

Then there exists $C > 0$ *such that*

$$\sup_{\substack{(\boldsymbol{\tau}, q) \in \mathbf{V} \\ (\boldsymbol{\tau}, q) \neq \mathbf{0}}} \frac{|\, A_0((\boldsymbol{\zeta}, r), (\boldsymbol{\tau}, q))\,|}{\|(\boldsymbol{\tau}, q)\|_{\mathbf{H}}} \geq C\, \|(\boldsymbol{\zeta}, r)\|_{\mathbf{H}} \qquad \forall\, (\boldsymbol{\zeta}, r) \in \mathbf{V}\,. \tag{4.29}$$

In addition, there holds

$$\sup_{(\boldsymbol{\zeta}, r) \in \mathbf{V}} |\, A_0((\boldsymbol{\zeta}, r), (\boldsymbol{\tau}, q))\,| > 0 \qquad \forall\, (\boldsymbol{\tau}, q) \in \mathbf{V}\,,\, (\boldsymbol{\tau}, q) \neq \mathbf{0}\,. \tag{4.30}$$

Proof. From the definition of B_0 (cf. (4.15)) we find that $\mathbf{V} = V \times H^1(\Omega_f)$, where

$$V := \{\, \boldsymbol{\zeta} \in \mathbf{H}(\mathbf{div}; \Omega_s) : \quad \boldsymbol{\zeta} = \boldsymbol{\zeta}^{\mathtt{t}} \quad \text{in } \Omega_s \quad \text{and} \quad \boldsymbol{\zeta}\, \boldsymbol{\nu} = 0 \quad \text{on } \Sigma \,\}\,.$$

In addition, it is easy to see that $\Xi(\boldsymbol{\zeta}) \in V$ for each $\boldsymbol{\zeta} \in V$. In fact, the symmetry of $\Xi(\boldsymbol{\zeta})$ follows from that of $\boldsymbol{\zeta}$ and $\mathbf{P}(\boldsymbol{\zeta})$ (cf. (4.4)), whereas the identity (4.6) guarantees that $\Xi(\boldsymbol{\zeta})\, \boldsymbol{\nu} = \mathbf{0}$ on Σ. Then, applying Lemma 4.5 to $(\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta})$, we deduce from (4.26) that there exists $C > 0$, depending on μ, c_1, c_2, κ_s, ρ_f, and ω^2, such that

$$\begin{aligned}&\mathrm{Re}\Big\{ A_0((\boldsymbol{\zeta}, r), (\Xi(\overline{\boldsymbol{\zeta}}), \overline{r})) \Big\} \\ &\qquad \geq C \Big\{ \|\mathbf{P}(\boldsymbol{\zeta})\|^2_{\mathbf{H}(\mathbf{div}; \Omega_s)} + \|(\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta})\|^2_{\mathbf{H}(\mathbf{div}; \Omega_s)} + \|r\|^2_{H^1(\Omega_f)} \Big\}\end{aligned} \tag{4.31}$$

for each $(\boldsymbol{\zeta}, r) \in \mathbf{V}$. Therefore, using the stability of the decomposition (4.7), the fact that $\|(\boldsymbol{\zeta}, r)\|_{\mathbf{H}} = \|(\overline{\boldsymbol{\zeta}}, \overline{r})\|_{\mathbf{H}}$, and the boundedness of Ξ, we deduce from (4.31) that

$$\mathrm{Re}\Big\{ A_0((\boldsymbol{\zeta}, r), (\Xi(\overline{\boldsymbol{\zeta}}), \overline{r})) \Big\} \geq C\, \|(\boldsymbol{\zeta}, r)\|^2_{\mathbf{H}} \geq C\, \|(\Xi(\overline{\boldsymbol{\zeta}}), \overline{r})\|_{\mathbf{H}}\, \|(\boldsymbol{\zeta}, r)\|_{\mathbf{H}} \tag{4.32}$$

for each $(\boldsymbol{\zeta}, r) \in \mathbf{V}$, which implies the inf-sup condition (4.29). Finally, the symmetry of A_0 and the estimate (4.32) yield the inf-sup condition (4.30) and complete the proof. ∎

LEMMA 4.8. *The operators* $\mathbf{K}_0 : \mathbf{H} \to \mathbf{H}$ *and* $\mathbf{K}_1 : \mathbf{H} \to \mathbf{Q}$ *are compact.*

Proof. We begin by recalling (cf. (4.8)) that there exists $\epsilon > 0$ such that $\mathbf{P}(\boldsymbol{\tau}) \in [H^{\epsilon}(\Omega_s)]^{2\times 2}$ for all $\boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\Omega_s)$, which, according to the compact imbedding $H^{\epsilon}(\Omega_s) \overset{c}{\hookrightarrow} L^2(\Omega_s)$, yields the compactness of $\mathbf{P} : \mathbf{H}(\mathbf{div};\Omega_s) \to [L^2(\Omega_s)]^{2\times 2}$. It follows that $\mathbf{P}^* : [L^2(\Omega_s)]^{2\times 2} \to \mathbf{H}(\mathbf{div};\Omega_s)$, $\mathbf{P}^*\,\mathcal{C}^{-1}\,\mathbf{P}$, $(\mathbf{I}-\mathbf{P})^*\,\mathcal{C}^{-1}\,\mathbf{P}$, and $\mathbf{P}^*\,\mathcal{C}^{-1}\,(\mathbf{I}-\mathbf{P})$ are all compact, which shows that the operator associated to the first three terms defining K_0 (cf. (4.12)) becomes compact as well. Certainly, we have also used here that $\mathcal{C}^{-1} : [L^2(\Omega_s)]^{2\times 2} \to [L^2(\Omega_s)]^{2\times 2}$ (cf. (3.14)) is continuous. Next, because of the second identity in (4.5), the fourth term on the right-hand side of (4.12) constitutes a finite rank operator, whereas for the last one it suffices to apply the compact imbedding of $H^1(\Omega_f)$ into $L^2(\Omega_f)$.

Finally, the continuous mapping $q \in H^1(\Omega_f) \to q\boldsymbol{\nu} \in [L^2(\Sigma)]^2$ and the compact imbedding $H^{1/2}(\Sigma) \overset{c}{\hookrightarrow} L^2(\Sigma)$ imply the compactness of $\mathbf{K}_1 : \mathbf{H} \to \mathbf{Q}$ (cf. (4.16)). ∎

We are able now to establish the main result of this section.

THEOREM 4.1. *Assume that the homogeneous problem associated to* (3.8) *has only the trivial solution. Then, given* $\mathbf{f} \in [L^2(\Omega_s)]^2$ *and* $g \in H^{-1/2}(\Gamma)$, *there exists a unique solution* $((\boldsymbol{\sigma},p),(\boldsymbol{\varphi},\boldsymbol{\gamma})) \in \mathbf{H}\times\mathbf{Q}$ *to* (3.8) (*equivalently* (4.17)). *In addition, there exists* $C > 0$ *such that*

$$\|((\boldsymbol{\sigma},p),(\boldsymbol{\varphi},\boldsymbol{\gamma}))\|_{\mathbf{H}\times\mathbf{Q}} \;\leq\; C\left\{\,\|\mathbf{f}\|_{[L^2(\Omega_s)]^2} \;+\; \|g\|_{H^{-1/2}(\Gamma)}\,\right\}. \tag{4.33}$$

Proof. It suffices to observe that the left-hand side of (4.17) constitutes a Fredholm operator of index zero. In fact, Lemmas 4.3 and 4.7 imply that $\left(\begin{smallmatrix}\mathbf{A}_0 & \mathbf{B}_0^* \\ \mathbf{B}_0 & \mathbf{0}\end{smallmatrix}\right)$ is an isomorphism, and Lemma 4.8 yields the compactness of $\left(\begin{smallmatrix}\mathbf{K}_0 & \check{\mathbf{K}}_1^* \\ -\mathbf{K}_1 & \mathbf{0}\end{smallmatrix}\right)$. ∎

5. Analysis of the primal/dual-mixed finite element method. In this section we introduce a Galerkin approximation of (3.8) and prove its well-posedness.

5.1. Preliminaries. We first let $\{\mathcal{T}_h\}_{h>0} := \{\mathcal{T}_{h_s}\}_{h_s>0} \cup \{\mathcal{T}_{h_f}\}_{h_f>0}$, where $\{\mathcal{T}_{h_s}\}_{h_s\geq 0}$ and $\{\mathcal{T}_{h_f}\}_{h_f>0}$ are regular families of triangulations of the polygonal regions $\bar{\Omega}_s$ and $\bar{\Omega}_f$, respectively, by triangles T of diameter h_T with mesh sizes $h_s := \max\{\,h_T : T \in \mathcal{T}_{h_s}\,\}$, $h_f := \max\{\,h_T : T \in \mathcal{T}_{h_f}\,\}$, and $h := \max\{h_s,h_f\}$, and such that the vertices of $\{\mathcal{T}_{h_s}\}_{h_s>0}$ and $\{\mathcal{T}_{h_f}\}_{h_f>0}$ coincide on Σ. Also, for reasons that will become clear below, we introduce an independent partition $\{\hat{\Sigma}_1,\hat{\Sigma}_2,\ldots,\hat{\Sigma}_m\}$ of the interface Σ and denote $\hat{h} := \max\{\,|\hat{\Sigma}_j| : j \in \{1,\ldots,m\}\,\}$. Then we define the finite element subspaces $H_h^{\boldsymbol{\sigma}}$, H_h^p, $Q_{\hat{h}}^{\boldsymbol{\varphi}}$, and $Q_h^{\boldsymbol{\gamma}}$ for the unknowns $\boldsymbol{\sigma}$, p, $\boldsymbol{\varphi}$, and $\boldsymbol{\gamma}$ of (3.8), respectively, as follows:

$$H_h^{\boldsymbol{\sigma}} := \Big\{\boldsymbol{\tau}_h \in \mathbf{H}(\mathbf{div};\Omega_s) :\; \boldsymbol{\tau}_{h,i}|_T \in \mathbb{R}\mathbb{T}_0(T)^{\mathtt{t}} \oplus \mathbb{P}_0(T)\,\mathbf{curl}^{\mathtt{t}}\,b_T \quad \forall\, i\in\{1,2\}\,,\quad \forall\, T \in \mathcal{T}_{h_s}\Big\}, \tag{5.1}$$

$$H_h^p := \left\{q_h \in \mathbf{C}(\bar{\Omega}_f) :\; q_h|_T \in \mathbb{P}_1(T) \quad \forall\, T \in \mathcal{T}_{h_f}\right\}, \tag{5.2}$$

$$Q_{\hat{h}}^{\boldsymbol{\varphi}} := \left\{\boldsymbol{\psi}_{\hat{h}} \in [\mathbf{C}(\Sigma)]^2 :\; \boldsymbol{\psi}_{\hat{h}}|_{\hat{\Sigma}_j} \in [\mathbb{P}_1(\hat{\Sigma}_j)]^2 \quad \forall\, j\in\{1,\ldots,m\}\right\}, \tag{5.3}$$

$$Q_h^{\boldsymbol{\gamma}} := \left\{\begin{pmatrix} 0 & \eta_h \\ -\eta_h & 0\end{pmatrix} :\; \eta_h \in \mathbf{C}(\bar{\Omega}_s)\,,\quad \eta_h|_T \in \mathbb{P}_1(T) \quad \forall\, T \in \mathcal{T}_{h_s}\right\}, \tag{5.4}$$

where $\boldsymbol{\tau}_{h,i}$ is the ith row of $\boldsymbol{\tau}_h$, $\mathbb{RT}_0(T)$ is the local Raviart–Thomas space of order 0 (cf. [10, 30]), b_T is the usual cubic bubble function on $T \in \mathcal{T}_{h_s}$, $\mathbf{curl}^{\mathtt{t}}\, b_T := (\frac{\partial b_T}{\partial x_2}, -\frac{\partial b_T}{\partial x_1})$, $\mathbf{C}(\cdot)$ stands for the space of continuous functions on the corresponding domain, and, given an integer $\ell \geq 0$ and a subset $\mathcal{K}$ of $\mathbb{R}^2$, $\mathbb{P}_\ell(\mathcal{K})$ denotes the space of polynomials defined in $\mathcal{K}$ of total degree $\leq \ell$. In addition, in what follows we will also need the following spaces:

$$\tilde{H}_h^{\boldsymbol{\sigma}} := \{\tilde{\boldsymbol{\tau}}_h \in H_h^{\boldsymbol{\sigma}} : \ \tilde{\boldsymbol{\tau}}_h\, \boldsymbol{\nu} = \mathbf{0} \quad \text{on } \Sigma\}\,, \tag{5.5}$$

$$E_h^{\boldsymbol{\sigma}} := \{\boldsymbol{\tau}_h \in \mathbf{H}(\mathbf{div};\Omega_s) : \ \boldsymbol{\tau}_{h,i}|_T \in \mathbb{RT}_0(T)^{\mathtt{t}} \quad \forall\, i \in \{1,2\}\,, \quad \forall\, T \in \mathcal{T}_{h_s}\}\,, \tag{5.6}$$

and

$$Q_h^{\mathbf{u}} := \left\{\mathbf{v}_h \in [L^2(\Omega_s)]^2 : \ \mathbf{v}_h|_T \in [\mathbb{P}_0(T)]^2 \quad \forall\, T \in \mathcal{T}_{h_s}\right\}. \tag{5.7}$$

We remark that $H_h^{\boldsymbol{\sigma}} \times Q_h^{\mathbf{u}} \times Q_h^{\boldsymbol{\gamma}}$ constitutes the well-known PEERS introduced in [1] for a mixed finite element approximation of the linear elasticity problem in the plane.

Next, given $\delta \in (0,1]$, we let $\mathcal{E}_h : [H^\delta(\Omega_s)]^{2\times 2} \cap \mathbf{H}(\mathbf{div};\Omega_s) \to E_h^{\boldsymbol{\sigma}}$ be the usual equilibrium interpolation operator (see [10, 30]), which is characterized by the identities

$$\mathbf{div}(\mathcal{E}_h(\boldsymbol{\tau})) = \mathcal{P}_h(\mathbf{div}(\boldsymbol{\tau})) \quad \text{and} \quad \int_e \mathcal{E}_h(\boldsymbol{\tau})\boldsymbol{\nu} = \int_e \boldsymbol{\tau}\boldsymbol{\nu} \quad \text{for every edge } e \text{ of } \mathcal{T}_{h_s}\,, \tag{5.8}$$

where $\mathcal{P}_h : [L^2(\Omega_s)]^2 \to Q_h^{\mathbf{u}}$ is the $[L^2(\Omega_s)]^2$-orthogonal projector. Since $E_h^{\boldsymbol{\sigma}} \subseteq H_h^{\boldsymbol{\sigma}}$, we note that $\mathcal{E}_h$ can also be considered as acting from $[H^\delta(\Omega_s)]^2 \cap \mathbf{H}(\mathbf{div};\Omega_s)$ into $H_h^{\boldsymbol{\sigma}}$. In particular, the second identity in (5.8) implies that $\mathcal{E}_h(\boldsymbol{\tau}) \in \tilde{H}_h^{\boldsymbol{\sigma}}$ for each $\boldsymbol{\tau} \in [H^\delta(\Omega_s)]^2 \cap \mathbf{H}(\mathbf{div};\Omega_s)$ such that $\boldsymbol{\tau}\,\boldsymbol{\nu} = \mathbf{0}$ on Σ. On the other hand, it is well known (see, e.g., Theorem 3.16 of [23]) that $\mathcal{E}_h$ satisfies

$$\begin{aligned} &\|\boldsymbol{\tau} - \mathcal{E}_h(\boldsymbol{\tau})\|_{[L^2(\Omega_s)]^{2\times 2}} \\ &\leq C\, h^\delta \left\{|\boldsymbol{\tau}|_{[H^\delta(\Omega_s)]^{2\times 2}} + \|\mathbf{div}(\boldsymbol{\tau})\|_{[L^2(\Omega_s)]^2}\right\} \quad \forall\, \boldsymbol{\tau} \in [H^\delta(\Omega_s)]^{2\times 2} \cap \mathbf{H}(\mathbf{div};\Omega_s). \end{aligned} \tag{5.9}$$

Moreover, in order to establish the global approximation properties of our finite element subspaces, we now let $\Pi_h : H^1(\Omega_f) \to H_h^p$, $\mathcal{Q}_{\hat{h}} : [H^{1/2}(\Sigma)]^2 \to Q_{\hat{h}}^{\boldsymbol{\varphi}}$, and $\mathcal{R}_h : [L^2(\Omega_s)]^{2\times 2} \to Q_h^{\boldsymbol{\gamma}}$ be the corresponding orthogonal projectors with respect to the natural norms of each space. Then we have (see [4, 10, 30]) the following:

($\mathrm{AP}_h^{\boldsymbol{\sigma}}$) For each $\delta \in (0,1]$ and for each $\boldsymbol{\tau} \in [H^\delta(\Omega_s)]^{2\times 2}$, with $\mathbf{div}(\boldsymbol{\tau}) \in [H^\delta(\Omega_s)]^2$, there holds

$$\|\boldsymbol{\tau} - \mathcal{E}_h(\boldsymbol{\tau})\|_{\mathbf{H}(\mathbf{div};\Omega_s)} \leq C\, h^\delta \left\{\|\boldsymbol{\tau}\|_{[H^\delta(\Omega_s)]^{2\times 2}} + \|\mathbf{div}(\boldsymbol{\tau})\|_{[H^\delta(\Omega_s)]^2}\right\}.$$

(AP_h^p) For each $s \in [1,2]$ and for each $q \in H^s(\Omega_f)$, there holds

$$\|q - \Pi_h(q)\|_{H^1(\Omega_f)} \leq C\, h^{s-1}\, \|q\|_{H^s(\Omega_f)}\,.$$

($\mathrm{AP}_{\hat{h}}^{\boldsymbol{\varphi}}$) For each $t \in \left[\frac{1}{2}, \frac{3}{2}\right]$ and for each $\boldsymbol{\psi} \in [H^t(\Sigma)]^2$, there holds

$$\|\boldsymbol{\psi} - \mathcal{Q}_{\hat{h}}(\boldsymbol{\psi})\|_{[H^{1/2}(\Sigma)]^2} \leq C\, \hat{h}^{t-1/2}\, \|\boldsymbol{\psi}\|_{[H^t(\Sigma)]^2}\,.$$

($\mathrm{AP}_h^{\boldsymbol{\gamma}}$) For each $s \in [0,1]$ and for each $\boldsymbol{\eta} \in [H^s(\Omega_s)]^{2\times 2} \cap [L^2(\Omega_s)]^{2\times 2}_{\mathtt{asym}}$, there holds

$$\|\boldsymbol{\eta} - \mathcal{R}_h(\boldsymbol{\eta})\|_{[L^2(\Omega_s)]^{2\times 2}} \,\leq\, C\, h^s\, \|\boldsymbol{\eta}\|_{[H^s(\Omega_s)]^{2\times 2}}\,.$$

($\mathrm{AP}_h^{\mathbf{u}}$) For each $t \in [0,1]$ and for each $\mathbf{v} \in [H^t(\Omega_s)]^2$, there holds

$$\|\mathbf{v} - \mathcal{P}_h(\mathbf{v})\|_{[L^2(\Omega_s)]^2} \,\leq\, C\, h^t\, \|\mathbf{v}\|_{[H^t(\Omega_s)]^2}\,.$$

Note that ($\mathrm{AP}_h^{\boldsymbol{\sigma}}$) is actually a straightforward consequence of (5.8), (5.9), and ($\mathrm{AP}_h^{\mathbf{u}}$).

We now let

$$\mathbf{H}_h \,:=\, H_h^{\boldsymbol{\sigma}} \times H_h^p\,, \qquad \mathbf{Q}_{\hat{h},h} \,:=\, Q_{\hat{h}}^{\boldsymbol{\varphi}} \times Q_h^{\boldsymbol{\gamma}} \tag{5.10}$$

and define the primal/dual-mixed finite element scheme associated to our coupled problem (3.8) as follows: Find $((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\varphi}_{\hat{h}}, \boldsymbol{\gamma}_h)) \in \mathbf{H}_h \times \mathbf{Q}_{\hat{h},h}$ such that

(5.11)

$$\begin{aligned} A((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\tau}_h, q_h)) + B_1((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\varphi}_{\hat{h}}, \boldsymbol{\gamma}_h)) &= F(\boldsymbol{\tau}_h, q_h) \quad \forall\, (\boldsymbol{\tau}_h, q_h) \in \mathbf{H}_h\,, \\ B_2((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h)) &= 0 \qquad\qquad \forall\, (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h}\,. \end{aligned}$$

5.2. Well-posedness of the primal/dual-mixed finite element method. In this section we prove the well-posedness of our primal/dual-mixed finite element scheme (5.11). To this end, as established by a classical result on projection methods for Fredholm operators of index zero (see, e.g., Theorem 13.7 in [25]), it suffices to show that the Galerkin scheme associated to the isomorphism $\left(\begin{smallmatrix} \mathbf{A}_0 & \mathbf{B}_0^* \\ \mathbf{B}_0 & \mathbf{0} \end{smallmatrix}\right)$ is well posed. Therefore, in what follows we prove that A_0 and B_0 (cf. (4.11), (4.15)) satisfy the corresponding inf-sup conditions on the finite element subspace $\mathbf{H}_h \times \mathbf{Q}_{\hat{h},h}$, thus providing the discrete analogues of Lemmas 4.3 and 4.7.

We begin with the following preliminary estimate.

LEMMA 5.1. *There exists $C_1 > 0$, independent of h and $\hat{h}$, such that for each $(\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h}$ there holds*

$$\sup_{\substack{(\boldsymbol{\tau}_h, q_h) \in \mathbf{H}_h \\ (\boldsymbol{\tau}_h, q_h) \neq \mathbf{0}}} \frac{|\, B_0((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \,\geq\, C_1\, \|\boldsymbol{\eta}_h\|_{[L^2(\Omega_s)]^{2\times 2}}\,. \tag{5.12}$$

Proof. Given $(\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h}$, we apply Theorem 4.5 in [26] (see also Lemma 4.4 in [1]) and deduce the existence of $\boldsymbol{\zeta}_h \in H_h^{\boldsymbol{\sigma}}$ such that $\boldsymbol{\zeta}_h\, \boldsymbol{\nu} = \mathbf{0}$ on Σ, $\mathbf{div}(\boldsymbol{\zeta}_h) = \mathbf{0}$ in Ω_s, and

$$|\, B_0((\boldsymbol{\zeta}_h, 0), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,| \,=\, \left| \int_{\Omega_s} \boldsymbol{\zeta}_h : \boldsymbol{\eta}_h \right| \,\geq\, C\, \|\boldsymbol{\zeta}_h\|_{[L^2(\Omega_s)]^{2\times 2}}\, \|\boldsymbol{\eta}_h\|_{[L^2(\Omega_s)]^{2\times 2}}\,,$$

which yields (5.12) and finishes the proof. ∎

Next, we follow the analysis in [5] and introduce the subspace of $[H^{-1/2}(\Sigma)]^2$ given by the piecewise constant functions. In other words, if $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_n\}$ is the partition on Σ induced by the triangulation $\mathcal{T}_h$, we define

$$Q_h^{\boldsymbol{\xi}} \,:=\, \left\{ \boldsymbol{\xi}_h \in [L^2(\Sigma)]^2 :\; \boldsymbol{\xi}_h|_{\Sigma_j} \in [\mathbb{P}_0(\Sigma_j)]^2 \quad \forall j \in \{1, \ldots, n\} \right\}\,,$$

which satisfies the following approximation property (see [4, 29]):

($\mathrm{AP}_h^{\boldsymbol{\xi}}$) For each $s \in (-\frac{1}{2}, \frac{1}{2}]$ and for each $\boldsymbol{\xi} \in [H^s(\Sigma)]^2$ there exists $\boldsymbol{\xi}_h \in Q_h^{\boldsymbol{\xi}}$ such that

$$\|\boldsymbol{\xi} - \boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2} \,\leq\, C\, h^{s+1/2}\, \|\boldsymbol{\xi}\|_{[H^s(\Sigma)]^2}\,.$$

Also, we assume that $\{\Sigma_1, \Sigma_2, \ldots, \Sigma_n\}$ and $\{\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_m\}$ are uniformly regular, which means that there exist $C, \hat{C} > 0$, independent of h and $\hat{h}$, such that $|\Sigma_j| \geq C\, h$ for all $j \in \{1, \ldots, n\}$ and $|\hat{\Sigma}_j| \geq \hat{C}\, \hat{h}$ for all $j \in \{1, \ldots, m\}$, for all $h, \hat{h} > 0$. These conditions yield the inverse inequalities for the spaces $Q_h^{\boldsymbol{\xi}}$ and $Q_{\hat{h}}^{\boldsymbol{\varphi}}$ (see [4, 29]), respectively; that is, for any real numbers s and t with $-\frac{1}{2} \leq s \leq t \leq 0$, there exists $c > 0$ such that

$$\|\boldsymbol{\xi}_h\|_{[H^t(\Sigma)]^2} \;\leq\; c\, h^{s-t}\, \|\boldsymbol{\xi}_h\|_{[H^s(\Sigma)]^2} \qquad \forall\, \boldsymbol{\xi}_h \in Q_h^{\boldsymbol{\xi}}\,, \tag{5.13}$$

and for any real numbers s and t with $0 \leq s \leq t \leq 1$, there exists $c > 0$ such that

$$\|\boldsymbol{\psi}_{\hat{h}}\|_{[H^t(\Sigma)]^2} \;\leq\; c\, \hat{h}^{s-t}\, \|\boldsymbol{\psi}_{\hat{h}}\|_{[H^s(\Sigma)]^2} \qquad \forall\, \boldsymbol{\psi}_{\hat{h}} \in Q_{\hat{h}}^{\boldsymbol{\varphi}}\,. \tag{5.14}$$

The following lemma establishes a second preliminary estimate.

LEMMA 5.2. *There exist $C_0 \in\,]0, 1[$ and $C_2 > 0$, independent of h and $\hat{h}$, such that for all $h \leq C_0\, \hat{h}$ and for each $(\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h}$, there holds*

$$\sup_{\substack{(\boldsymbol{\tau}_h, q_h) \in \mathbf{H}_h \\ (\boldsymbol{\tau}_h, q_h) \neq \mathbf{0}}} \frac{|\, B_0((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \;\geq\; C_2\, \|\boldsymbol{\psi}_{\hat{h}}\|_{[H^{1/2}(\Sigma)]^2} \;-\; \|\boldsymbol{\eta}_h\|_{[L^2(\Omega_s)]^{2\times 2}}\,. \tag{5.15}$$

Proof. Given $\boldsymbol{\xi}_h \in Q_h^{\boldsymbol{\xi}}$, we consider the discrete analogue of (4.18) (cf. proof of Lemma 4.1) and let $\mathbf{z} \in [H^1(\Omega_s)]^2$ be the unique weak solution, up to an element in $[\mathbb{P}_0(\Omega_s)]^2$, of the boundary value problem with Neumann boundary conditions:

$$-\,\Delta \mathbf{z} \;=\; \frac{1}{|\Omega_s|} \int_{\Sigma} \boldsymbol{\xi}_h \quad \text{in } \Omega_s\,, \qquad \nabla \mathbf{z}\, \boldsymbol{\nu} \;=\; -\,\boldsymbol{\xi}_h \quad \text{on } \Sigma\,.$$

Since $Q_h^{\boldsymbol{\xi}} \subseteq [L^2(\Sigma)]^2$, the corresponding regularity result (see, e.g., [20, 21]) implies that $\mathbf{z} \in [H^{1+\delta}(\Omega_s)]^2$ for each $\delta \in \left[0, \frac{1}{2}\right]$, and

$$\|\mathbf{z}\|_{[H^{1+\delta}(\Omega_s)]^2} \;\leq\; C\, \|\boldsymbol{\xi}_h\|_{[H^{-1/2+\delta}(\Sigma)]^2}\,. \tag{5.16}$$

Then we let $\boldsymbol{\zeta} := \nabla \mathbf{z}$ in Ω_s and observe that

$$\mathbf{div}(\boldsymbol{\zeta}) \;=\; -\,\frac{1}{|\Omega_s|} \int_{\Sigma} \boldsymbol{\xi}_h \quad \text{in } \Omega_s \quad \text{and} \quad \boldsymbol{\zeta}\boldsymbol{\nu} \;=\; -\,\boldsymbol{\xi}_h \quad \text{on } \Sigma\,. \tag{5.17}$$

In addition, using (5.16) and (5.17), we obtain

$$\|\boldsymbol{\zeta}\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)} \leq C \|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2} \quad \text{and} \quad \|\boldsymbol{\zeta}\|_{[H^{\delta}(\Omega_s)]^{2\times 2}} \leq C \|\boldsymbol{\xi}_h\|_{[H^{-1/2+\delta}(\Sigma)]^2}. \tag{5.18}$$

Now, according to the characterization (5.8), there holds $\mathbf{div}(\mathcal{E}_h(\boldsymbol{\zeta})) = \mathbf{div}(\boldsymbol{\zeta}) = -\frac{1}{|\Omega_s|} \int_{\Sigma} \boldsymbol{\xi}_h$ in Ω_s and $\mathcal{E}_h(\boldsymbol{\zeta})\, \boldsymbol{\nu} = \boldsymbol{\zeta}\boldsymbol{\nu} = -\,\boldsymbol{\xi}_h$ on Σ. It follows from (5.9) and (5.18) that

$$\begin{aligned}
\|\mathcal{E}_h(\boldsymbol{\zeta})\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)} \;&\leq\; \|\boldsymbol{\zeta} - \mathcal{E}_h(\boldsymbol{\zeta})\|_{[L^2(\Omega_s)]^{2\times 2}} \;+\; \|\boldsymbol{\zeta}\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)} \\
&\leq\; C\, \left\{\, h^{\delta}\, |\boldsymbol{\zeta}|_{[H^{\delta}(\Omega_s)]^{2\times 2}} \;+\; \|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2} \,\right\} \\
&\leq\; C\, \left\{\, h^{\delta}\, \|\boldsymbol{\xi}_h\|_{[H^{-1/2+\delta}(\Sigma)]^2} \;+\; \|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2} \,\right\},
\end{aligned}$$

which, applying the inverse inequality (5.13), yields

$$\|\mathcal{E}_h(\boldsymbol{\zeta})\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)} \;\leq\; \hat{C}\, \|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2}\,.$$

Therefore, given $(\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h}$, we find that

$$\sup_{\substack{(\boldsymbol{\tau}_h, q_h)\in \mathbf{H}_h \\ (\boldsymbol{\tau}_h, q_h)\neq \mathbf{0}}} \frac{|\, B_0((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \geq \frac{|\, B_0((\mathcal{E}_h(\boldsymbol{\zeta}), 0), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,|}{\|\mathcal{E}_h(\boldsymbol{\zeta})\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)}}$$

$$= \frac{\Big|\, \langle \boldsymbol{\xi}_h\,,\, \boldsymbol{\psi}_{\hat{h}}\rangle_{\Sigma} + \displaystyle\int_{\Omega} \mathcal{E}_h(\boldsymbol{\zeta}) : \boldsymbol{\eta}_h \,\Big|}{\|\mathcal{E}_h(\boldsymbol{\zeta})\|_{\mathbf{H}(\mathbf{div};\,\Omega_s)}} \geq \frac{1}{\hat{C}}\, \frac{|\, \langle \boldsymbol{\xi}_h, \boldsymbol{\psi}_{\hat{h}}\rangle_{\Sigma}\,|}{\|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2}} - \|\boldsymbol{\eta}_h\|_{[L^2(\Omega_s)]^{2\times 2}}$$

for all $\boldsymbol{\xi}_h \in Q_h^{\boldsymbol{\xi}}$, and hence

(5.19)
$$\sup_{\substack{(\boldsymbol{\tau}_h, q_h)\in \mathbf{H}_h \\ (\boldsymbol{\tau}_h, q_h)\neq \mathbf{0}}} \frac{|\, B_0((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \geq \frac{1}{\hat{C}} \sup_{\substack{\boldsymbol{\xi}_h\in Q_h^{\boldsymbol{\xi}} \\ \boldsymbol{\xi}_h\neq \mathbf{0}}} \frac{|\, \langle \boldsymbol{\xi}_h, \boldsymbol{\psi}_{\hat{h}}\rangle_{\Sigma}\,|}{\|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2}} - \|\boldsymbol{\eta}_h\|_{[L^2(\Omega_s)]^{2\times 2}}\,.$$

Since the normal trace on Σ is well defined and continuous from $[H^{\delta}(\Omega_s)]^{2\times 2} \cap \mathbf{H}(\mathbf{div};\,\Omega_s)$ onto $[H^{-1/2+\delta}(\Sigma)]^2$ for $\delta \neq \frac{1}{2}$, we now apply the vector version of Lemma 3.3 in [5], making use of the approximation property $(\mathrm{AP}_h^{\boldsymbol{\xi}})$ and the inverse inequality (5.14), and deduce that there exist $C_0 \in\,]0,1[$ and $\tilde{C}_2 > 0$, independent of h and $\hat{h}$, such that for all $h \leq C_0\, \hat{h}$ there holds

$$\sup_{\substack{\boldsymbol{\xi}_h\in Q_h^{\boldsymbol{\xi}} \\ \boldsymbol{\xi}_h\neq \mathbf{0}}} \frac{|\, \langle \boldsymbol{\xi}_h, \boldsymbol{\psi}_{\hat{h}}\rangle_{\Sigma}\,|}{\|\boldsymbol{\xi}_h\|_{[H^{-1/2}(\Sigma)]^2}} \geq \tilde{C}_2\, \|\boldsymbol{\psi}_{\hat{h}}\|_{[H^{1/2}(\Sigma)]^2}\,. \tag{5.20}$$

In this way, (5.19) and (5.20) yield the required estimate and complete the proof. ∎

The discrete inf-sup condition for B_0 can be established now as a straightforward consequence of Lemmas 5.1 and 5.2.

LEMMA 5.3. *Let $C_0 \in\,]0,1[$ be the constant provided by Lemma 5.2. Then there exists $\beta > 0$, independent of h and $\hat{h}$, such that for all $h \leq C_0\, \hat{h}$ and for each $(\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h}$, there holds*

$$\sup_{\substack{(\boldsymbol{\tau}_h, q_h)\in \mathbf{H}_h \\ (\boldsymbol{\tau}_h, q_h)\neq \mathbf{0}}} \frac{|\, B_0((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\,|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \geq \beta\, \|(\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h)\|_{\mathbf{H}}\,.$$

Proof. It suffices to add (5.12) and (5.15), the latter multiplied by $C_1/2$. ∎

It remains to establish the discrete weak coercivity of A_0. To this end, we now introduce the following discrete approximation of the operator $\mathbf{P}$ (cf. (4.4)):

$$\begin{aligned} \mathbf{P}_h : \mathbf{H}(\mathbf{div};\,\Omega_s) &\to \tilde{H}_h^{\boldsymbol{\sigma}}, \\ \boldsymbol{\tau} &\to \mathbf{P}_h(\boldsymbol{\tau}) := \tilde{\boldsymbol{\sigma}}_h\,, \end{aligned} \tag{5.21}$$

where $(\tilde{\boldsymbol{\sigma}}_h, \tilde{\mathbf{u}}_h, \tilde{\boldsymbol{\gamma}}_h) \in \tilde{H}_h^{\boldsymbol{\sigma}} \times (Q_h^{\mathbf{u}} \cap \mathbb{RM}(\Omega_s)^{\perp}) \times Q_h^{\gamma}$ is the mixed finite element approximation of $(\tilde{\boldsymbol{\sigma}}, \tilde{\mathbf{u}}, \tilde{\boldsymbol{\gamma}})$ (cf. section 4.1) (see also [15] for details). In particular, we easily find that for each $\boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\,\Omega_s)$ there holds

$$\mathbf{P}_h(\boldsymbol{\tau})\,\boldsymbol{\nu} = \mathbf{0} \quad \text{on } \Sigma \qquad \text{and} \qquad \int_{\Omega_s} \mathbf{P}_h(\boldsymbol{\tau}) : \tilde{\boldsymbol{\eta}}_h = 0 \quad \forall\, \tilde{\boldsymbol{\eta}}_h \in Q_h^{\gamma}\,. \tag{5.22}$$

In addition, one can show that there exist $C > 0$, independent of h, such that

$$\begin{aligned} &\|\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}_h\|_{\mathbf{H}(\mathbf{div};\Omega_s)} + \|\tilde{\mathbf{u}} - \tilde{\mathbf{u}}_h\|_{[L^2(\Omega_s)]^2} + \|\tilde{\boldsymbol{\gamma}} - \tilde{\boldsymbol{\gamma}}_h\|_{[L^2(\Omega_s)]^{2\times 2}} \\ &\qquad \leq \tilde{C}\Big\{\|(\mathbf{I}-\mathcal{E}_h)(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{H}(\mathbf{div};\Omega_s)} + \|(\mathbf{I}-\mathcal{P}_h)(\tilde{\mathbf{u}})\|_{[L^2(\Omega_s)]^2} \\ &\qquad\qquad + \|(\mathbf{I}-\mathcal{R}_h)(\tilde{\boldsymbol{\gamma}})\|_{[L^2(\Omega_s)]^{2\times 2}}\Big\}. \end{aligned} \tag{5.23}$$

Hence, we are in a position to estimate

$$\|\mathbf{P}(\boldsymbol{\tau}_h) - \mathbf{P}_h(\boldsymbol{\tau}_h)\|_{\mathbf{H}(\mathbf{div};\Omega_s)} = \|\tilde{\boldsymbol{\sigma}} - \tilde{\boldsymbol{\sigma}}_h\|_{\mathbf{H}(\mathbf{div};\Omega_s)} \quad \text{for each} \quad \boldsymbol{\tau}_h \in H_h^{\boldsymbol{\sigma}}.$$

More precisely, we have the following result.

LEMMA 5.4. *Let $\epsilon > 0$ be the parameter defining the regularity of the solution of (4.1). Then there exists $C > 0$, independent of h, such that*

$$\|\mathbf{P}(\boldsymbol{\tau}_h) - \mathbf{P}_h(\boldsymbol{\tau}_h)\|_{\mathbf{H}(\mathbf{div};\Omega_s)} \leq C\, h^{\epsilon}\, \|\mathbf{div}(\boldsymbol{\tau}_h)\|_{[L^2(\Omega_s)]^2} \qquad \forall\, \boldsymbol{\tau}_h \in H_h^{\boldsymbol{\sigma}}. \tag{5.24}$$

Proof. It suffices to show that the right-hand side of (5.23) is bounded by $C\, h^{\epsilon}\, \|\mathbf{div}(\boldsymbol{\tau}_h)\|_{[L^2(\Omega_s)]^2}$. Indeed, using $(\mathrm{AP}_h^{\mathbf{u}})$, $(\mathrm{AP}_h^{\boldsymbol{\gamma}})$, and the regularity estimate (4.3), we easily find that

$$\begin{aligned} &\|(\mathbf{I}-\mathcal{P}_h)(\tilde{\mathbf{u}})\|_{[L^2(\Omega_s)]^2} + \|(\mathbf{I}-\mathcal{R}_h)(\tilde{\boldsymbol{\gamma}})\|_{[L^2(\Omega_s)]^{2\times 2}} \\ &\quad \leq C\Big\{ h\, \|\tilde{\mathbf{u}}\|_{[H^1(\Omega_s)]^2} + h^{\epsilon}\, \|\tilde{\boldsymbol{\gamma}}\|_{[H^{\epsilon}(\Omega_s)]^{2\times 2}}\Big\} \leq C\, h^{\epsilon}\, \|\mathbf{div}(\boldsymbol{\tau}_h)\|_{[L^2(\Omega_s)]^2}. \end{aligned} \tag{5.25}$$

Now, we observe that a straight application of $(\mathrm{AP}_h^{\boldsymbol{\sigma}})$ to $\|(\mathbf{I}-\mathcal{E}_h)(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{H}(\mathbf{div};\Omega_s)}$ yields the expression $h^{\delta}\, \|\mathbf{div}(\tilde{\boldsymbol{\sigma}})\|_{[H^{\delta}(\Omega_s)]^2}$, which does not provide the expected estimate, and hence a different procedure must be employed. In fact, it is clear from (4.1), with $\boldsymbol{\tau}_h$ instead of $\boldsymbol{\tau}$, that

$$\mathbf{div}(\tilde{\boldsymbol{\sigma}}) = (\mathbf{I}-\mathbf{M})(\mathbf{div}(\boldsymbol{\tau}_h)) \quad \text{in } \Omega_s, \tag{5.26}$$

which, applying (5.9), (4.3), and the boundedness of $\mathbf{M}$, leads to

$$\begin{aligned} &\|(\mathbf{I}-\mathcal{E}_h)(\tilde{\boldsymbol{\sigma}})\|_{[L^2(\Omega_s)]^{2\times 2}} \\ &\leq C\, h^{\epsilon}\Big\{|\tilde{\boldsymbol{\sigma}}|_{[H^{\epsilon}(\Omega_s)]^{2\times 2}} + \|\mathbf{div}(\tilde{\boldsymbol{\sigma}})\|_{[L^2(\Omega_s)]^2}\Big\} \leq C\, h^{\epsilon}\, \|\mathbf{div}(\boldsymbol{\tau}_h)\|_{[L^2(\Omega_s)]^2}. \end{aligned} \tag{5.27}$$

Next, it follows from (5.8) and (5.26) that

$$\begin{aligned} &\|\mathbf{div}(\tilde{\boldsymbol{\sigma}}) - \mathbf{div}(\mathcal{E}_h(\tilde{\boldsymbol{\sigma}}))\|_{[L^2(\Omega_s)]^2} \\ &= \|(\mathbf{I}-\mathcal{P}_h)(\mathbf{div}(\tilde{\boldsymbol{\sigma}}))\|_{[L^2(\Omega_s)]^2} = \|(\mathbf{I}-\mathcal{P}_h)(\mathbf{M}(\mathbf{div}(\boldsymbol{\tau}_h)))\|_{[L^2(\Omega_s)]^2}, \end{aligned}$$

whence $(\mathrm{AP}_h^{\mathbf{u}})$, the fact that all the norms in $\mathbb{RM}(\Omega_s)$ are equivalent (with constants certainly independent of h), and the boundedness of $\mathbf{M}$ imply that

$$\begin{aligned} \|\mathbf{div}(\tilde{\boldsymbol{\sigma}}) - \mathbf{div}(\mathcal{E}_h(\tilde{\boldsymbol{\sigma}}))\|_{[L^2(\Omega_s)]^2} &\leq C\, h\, \|\mathbf{M}(\mathbf{div}(\boldsymbol{\tau}_h))\|_{[H^1(\Omega_s)]^2} \\ &\leq C\, h\, \|\mathbf{M}(\mathbf{div}(\boldsymbol{\tau}_h))\|_{[L^2(\Omega_s)]^2} \leq C\, h\, \|\mathbf{div}(\boldsymbol{\tau}_h)\|_{[L^2(\Omega_s)]^2}. \end{aligned} \tag{5.28}$$

In this way, (5.27) and (5.28) give the required estimate for $\|(\mathbf{I}-\mathcal{E}_h)(\tilde{\boldsymbol{\sigma}})\|_{\mathbf{H}(\mathbf{div};\Omega_s)}$, which, together with (5.25) and (5.23), yields (5.24) and finishes the proof. ∎

We now recall from [14] the following result, which is the discrete analogue of Lemma 4.5, and whose proof applies the decomposition (4.22)–(4.23) and the approximation property $(\mathrm{AP}^{\boldsymbol{\varphi}}_{\hat{h}})$.

LEMMA 5.5. *Let* $V_{\hat{h}} := \{\, \boldsymbol{\tau} \in \mathbf{H}(\mathbf{div};\Omega_s) : \langle \boldsymbol{\tau}\boldsymbol{\nu}, \boldsymbol{\psi}_{\hat{h}} \rangle_\Sigma = 0 \ \text{for all}\ \boldsymbol{\psi}_{\hat{h}} \in Q^{\boldsymbol{\varphi}}_{\hat{h}} \,\}$. *Then there exist positive constants* C, h_0, *independent of* h, *such that for each* $\hat{h} \le h_0$ *there holds*

$$C\,\|\boldsymbol{\tau}\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} \ \le\ \|\boldsymbol{\tau}_0\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} \qquad \forall\, \boldsymbol{\tau} \in V_{\hat{h}}\,, \tag{5.29}$$

where $\boldsymbol{\tau} = \boldsymbol{\tau}_0 + d\,\mathbf{I}$, *with* $\boldsymbol{\tau}_0 \in \mathbf{H}_0(\mathbf{div};\Omega_s)$ *and* $d \in \mathbb{C}$.

Proof. See Lemma 4.4 in [14]. □

Actually, Lemma 4.4 in [14] establishes the estimate (5.29) for the subspace

$$V_{\hat{h}} := \{\, \boldsymbol{\tau}_h \in H^{\boldsymbol{\sigma}}_h : \langle \boldsymbol{\tau}_h\boldsymbol{\nu}, \boldsymbol{\psi}_{\hat{h}} \rangle_\Sigma = 0 \quad \forall\, \boldsymbol{\psi}_{\hat{h}} \in Q^{\boldsymbol{\varphi}}_{\hat{h}} \,\}\,.$$

Nevertheless, it is easy to see that the same proof is valid for $V_{\hat{h}}$ as defined here in Lemma 5.5.

The discrete weak coercivity of A_0 can be proved now.

LEMMA 5.6. *Let* $\mathbf{V}_{h,\hat{h}}$ *be the discrete kernel of* $\mathbf{B}_0$, *that is,*

$$\mathbf{V}_{h,\hat{h}} := \left\{ (\boldsymbol{\tau}_h, q_h) \in \mathbf{H}_h :\ B_0((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h)) = 0 \qquad \forall\, (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h) \in \mathbf{Q}_{\hat{h},h} \right\},$$

and let $h_0 > 0$ *be the constant provided by Lemma* 5.5. *Then there exist* C, $h_1 > 0$, *independent of* h *and* $\hat{h}$, *such that for each* $\hat{h} \le h_0$ *and for each* $h \le h_1$ *there holds*

$$\sup_{\substack{(\boldsymbol{\tau}_h, q_h) \in \mathbf{V}_{h,\hat{h}} \\ (\boldsymbol{\tau}_h, q_h) \neq \mathbf{0}}} \frac{|\, A_0((\boldsymbol{\zeta}_h, r_h), (\boldsymbol{\tau}_h, q_h))\,|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \ \ge\ C\,\|(\boldsymbol{\zeta}_h, r_h)\|_{\mathbf{H}} \qquad \forall\, (\boldsymbol{\zeta}_h, r_h) \in \mathbf{V}_{h,\hat{h}}\,. \tag{5.30}$$

In addition, for each $\hat{h} \le h_0$ *and for each* $h \le h_1$ *there holds*

$$\sup_{(\boldsymbol{\zeta}_h, r_h) \in \mathbf{V}_{h,\hat{h}}} |\, A_0((\boldsymbol{\zeta}_h, r_h), (\boldsymbol{\tau}_h, q_h))\,| \ >\ 0 \qquad \forall\, (\boldsymbol{\tau}_h, q_h) \in \mathbf{V}_{h,\hat{h}}\,,\ (\boldsymbol{\tau}_h, q_h) \neq \mathbf{0}\,. \tag{5.31}$$

Proof. Let us introduce the linear and bounded operator $\Xi_h := (\mathbf{I} - 2\,\mathbf{P}_h) : H^{\boldsymbol{\sigma}}_h \to H^{\boldsymbol{\sigma}}_h$. It follows from Lemma 5.4 that

$$\|\Xi(\boldsymbol{\zeta}_h) - \Xi_h(\boldsymbol{\zeta}_h)\|_{\mathbf{H}(\mathbf{div};\Omega_s)} \ \le\ C\,h^\epsilon\,\|\mathbf{div}(\boldsymbol{\zeta}_h)\|_{[L^2(\Omega_s)]^2} \ \le\ C\,h^\epsilon\,\|\boldsymbol{\zeta}_h\|_{\mathbf{H}(\mathbf{div};\Omega_s)}$$

for all $\boldsymbol{\zeta}_h \in H^{\boldsymbol{\sigma}}_h$, and hence, using the boundedness of A_0 and the inequality (4.26) (cf. Lemma 4.6), we find that for each $(\boldsymbol{\zeta}_h, r_h) \in \mathbf{H}_h$ there holds

$$\begin{aligned}
&\left|\, \mathrm{Re}\left\{ A_0((\boldsymbol{\zeta}_h, r_h), (\Xi_h(\overline{\boldsymbol{\zeta}}_h), \overline{r}_h)) \right\} \right| \\
&\quad \ge \left|\, \mathrm{Re}\left\{ A_0((\boldsymbol{\zeta}_h, r_h), (\Xi(\overline{\boldsymbol{\zeta}}_h), \overline{r}_h)) \right\} \right| - C\,h^\epsilon\,\|(\boldsymbol{\zeta}_h, r_h)\|^2_{\mathbf{H}} \\
&\quad \ge C \left\{ \|\mathbf{P}(\boldsymbol{\zeta}_h)\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} + \|(\mathbf{I} - \mathbf{P})(\boldsymbol{\zeta}_h)_0\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} + \|r_h\|^2_{H^1(\Omega_f)} \right\} \\
&\qquad - C\,h^\epsilon\,\|(\boldsymbol{\zeta}_h, r_h)\|^2_{\mathbf{H}}\,.
\end{aligned} \tag{5.32}$$

Now, from the definition of B_0 (cf. (4.15)) we see that $\mathbf{V}_{h,\hat{h}} = V_{h,\hat{h}} \times H^p_h$, where

$$V_{h,\hat{h}} := V_{\hat{h}} \ \cap\ \left\{ \boldsymbol{\tau}_h \in H^{\boldsymbol{\sigma}}_h :\ \int_{\Omega_s} \boldsymbol{\tau}_h : \boldsymbol{\eta}_h = 0 \quad \forall\, \boldsymbol{\eta}_h \in Q^\gamma_h \right\}.$$

Then it is clear that $(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}_h) \in V_{\hat{h}}$ for each $\boldsymbol{\zeta}_h \in V_{h,\hat{h}}$, and therefore Lemma 5.5 implies that for each $\hat{h} \leq h_0$ there holds

$$\|(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}_h)_0\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)} \geq c\,\|(\mathbf{I}-\mathbf{P})(\boldsymbol{\zeta}_h)\|^2_{\mathbf{H}(\mathbf{div};\Omega_s)}\,.$$

Thus, replacing the above estimate back into (5.32), and using the stability of the decomposition (4.7), we deduce the existence of $h_1 > 0$ such that for each $\hat{h} \leq h_0$ and for each $h \leq h_1$ there holds

$$\begin{aligned} &\left| \mathrm{Re}\left\{ A_0((\boldsymbol{\zeta}_h, r_h), (\Xi_h(\overline{\boldsymbol{\zeta}}_h), \overline{r}_h)) \right\} \right| \\ &\quad \geq C\,\|(\boldsymbol{\zeta}_h, r_h)\|^2_{\mathbf{H}} \geq C\,\|(\Xi_h(\overline{\boldsymbol{\zeta}}_h), \overline{r}_h)\|_{\mathbf{H}}\,\|(\boldsymbol{\zeta}_h, r_h)\|_{\mathbf{H}} \end{aligned} \tag{5.33}$$

for each $(\boldsymbol{\zeta}_h, r_h) \in \mathbf{V}_{h,\hat{h}}$, where the boundedness of Ξ_h has also been used in the last inequality. In this way, since (5.22) implies that $\Xi_h(\boldsymbol{\tau}_h) \in V_{h,\hat{h}}$ for each $\boldsymbol{\tau}_h \in V_{h,\hat{h}}$, the discrete inf-sup condition (5.30) follows straightforwardly from (5.33). In addition, the symmetry of A_0 and the estimate (5.33) yield the discrete inf-sup condition (5.31) and complete the proof. ∎

The following theorem establishes the well-posedness and convergence of the discrete scheme (5.11).

THEOREM 5.1. *Assume that the homogeneous problem associated to* (3.8) *has only the trivial solution. Let* $C_0 \in]0,1[$ *and* h_0, $h_1 > 0$ *be the constants provided by Lemmas* 5.2, 5.5, *and* 5.6, *respectively. Then there exist* $\tilde{h}_0 \in]0, h_0]$ *and* $\tilde{h}_1 \in]0, h_1]$ *such that for each* $\hat{h} \leq \tilde{h}_0$ *and for each* $h \leq \min\{\tilde{h}_1, C_0\,\hat{h}\}$, *the primal/dual-mixed finite element scheme* (5.11) *has a unique solution* $((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\varphi}_{\hat{h}}, \boldsymbol{\gamma}_h)) \in \mathbf{H}_h \times \mathbf{Q}_{\hat{h},h}$. *In addition, there exist* C_1, $C_2 > 0$, *independent of* h *and* $\hat{h}$, *such that*

$$\begin{aligned} &\|((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\varphi}_{\hat{h}}, \boldsymbol{\gamma}_h))\|_{\mathbf{H}\times\mathbf{Q}} \\ &\quad \leq C_1 \sup_{\substack{(\boldsymbol{\tau}_h, q_h)\in\mathbf{H}_h \\ (\boldsymbol{\tau}_h, q_h)\neq\mathbf{0}}} \frac{|F(\boldsymbol{\tau}_h, q_h)|}{\|(\boldsymbol{\tau}_h, q_h)\|_{\mathbf{H}}} \leq C_1 \left\{ \|\mathbf{f}\|_{[L^2(\Omega_s)]^2} + \|g\|_{H^{-1/2}(\Gamma)} \right\} \end{aligned} \tag{5.34}$$

and

$$\begin{aligned} &\|((\boldsymbol{\sigma}, p), (\boldsymbol{\varphi}, \boldsymbol{\gamma})) - ((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\varphi}_{\hat{h}}, \boldsymbol{\gamma}_h))\|_{\mathbf{H}\times\mathbf{Q}} \\ &\leq C_2 \inf_{((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\in\mathbf{H}_h\times\mathbf{Q}_{\hat{h},h}} \|((\boldsymbol{\sigma}, p), (\boldsymbol{\varphi}, \boldsymbol{\gamma})) - ((\boldsymbol{\tau}_h, q_h), (\boldsymbol{\psi}_{\hat{h}}, \boldsymbol{\eta}_h))\|_{\mathbf{H}\times\mathbf{Q}}\,. \end{aligned} \tag{5.35}$$

Furthermore, if there exists $\delta \in (0,1]$ *such that* $\boldsymbol{\sigma} \in [H^\delta(\Omega_s)]^{2\times 2}$, $\mathbf{div}(\boldsymbol{\sigma}) \in [H^\delta(\Omega_s)]^2$, $p \in H^{1+\delta}(\Omega_f)$, $\boldsymbol{\varphi} \in [H^{1/2+\delta}(\Sigma)]^2$, *and* $\boldsymbol{\gamma} \in [H^\delta(\Omega_s)]^{2\times 2}$, *then there holds*

$$\begin{aligned} &\|((\boldsymbol{\sigma}, p), (\boldsymbol{\varphi}, \boldsymbol{\gamma})) - ((\boldsymbol{\sigma}_h, p_h), (\boldsymbol{\varphi}_{\hat{h}}, \boldsymbol{\gamma}_h))\|_{\mathbf{H}\times\mathbf{Q}} \leq C_3\,\hat{h}^\delta\,\|\boldsymbol{\varphi}\|_{[H^{1/2+\delta}(\Sigma)]^2} \\ &+ C_3\,h^\delta \left\{ \|\boldsymbol{\sigma}\|_{[H^\delta(\Omega_s)]^{2\times 2}} + \|\mathbf{div}(\boldsymbol{\sigma})\|_{[H^\delta(\Omega_s)]^2} + \|p\|_{H^{1+\delta}(\Omega_f)} + \|\boldsymbol{\gamma}\|_{[H^\delta(\Omega_s)]^{2\times 2}} \right\}, \end{aligned} \tag{5.36}$$

with a constant $C_3 > 0$, *independent of* h *and* $\hat{h}$.

Proof. Thanks to Lemmas 5.3 and 5.6, the proof of the first part is a direct application of Theorem 13.7 in [25], whereas the rate of convergence (5.36) follows directly from the Cea estimate (5.35) and the approximation properties $(\mathrm{AP}^{\boldsymbol{\sigma}}_h)$, (AP^{p}_h), $(\mathrm{AP}^{\boldsymbol{\varphi}}_{\hat{h}})$, and $(\mathrm{AP}^{\boldsymbol{\gamma}}_h)$ (cf. section 5.1). ∎

6. Numerical results. In this section we present an example illustrating the performance of the primal/dual-mixed finite element scheme (5.11) on a finite sequence of uniform triangulations of the domain. We begin by introducing additional notation. The variable N stands for the number of degrees of freedom defining the finite element subspaces $\mathbf{H}_h$ and $\mathbf{Q}_{\hat{h},h}$, and the individual errors are denoted by

$$\mathsf{e}(\boldsymbol{\sigma}) := \|\boldsymbol{\sigma} - \boldsymbol{\sigma}_h\|_{\mathbf{H}(\mathbf{div};\Omega_s)}, \quad \mathsf{e}(p) := \|p - p_h\|_{H^1(\Omega_f)},$$

$$\mathsf{e}(\boldsymbol{\varphi}) := \|\boldsymbol{\varphi} - \boldsymbol{\varphi}_{\hat{h}}\|_{[H^{1/2}(\Sigma)]^2}, \quad \text{and} \quad \mathsf{e}(\boldsymbol{\gamma}) := \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_h\|_{[L^2(\Omega_s)]^{2\times 2}}.$$

Also, we let $r(\boldsymbol{\sigma})$, $r(p)$, $r(\boldsymbol{\varphi})$, and $r(\boldsymbol{\gamma})$ be the experimental rates of convergence given by

$$r(\boldsymbol{\sigma}) := \frac{\log(\mathsf{e}(\boldsymbol{\sigma})/\mathsf{e}'(\boldsymbol{\sigma}))}{\log(h/h')}, \quad r(p) := \frac{\log(\mathsf{e}(p)/\mathsf{e}'(p))}{\log(h/h')},$$

$$r(\boldsymbol{\varphi}) := \frac{\log(\mathsf{e}(\boldsymbol{\varphi})/\mathsf{e}'(\boldsymbol{\varphi}))}{\log(h/h')}, \quad \text{and} \quad r(\boldsymbol{\gamma}) := \frac{\log(\mathsf{e}(\boldsymbol{\gamma})/\mathsf{e}'(\boldsymbol{\gamma}))}{\log(h/h')},$$

where h and h' denote two consecutive mesh sizes with corresponding errors e and e'.

We consider the domains $\Omega_s :=]-0.3, 0.3[^2$ and $\Omega_f := \mathbf{B}(\mathbf{0}, 1) \setminus \Omega_s$, where $\mathbf{B}(\mathbf{0}, 1)$ is the unit circle, and take the parameters $\omega = 10$, $\rho_s = \rho_f = \lambda = \mu = 1$, whence $\kappa_f = 1$ and $\kappa_s = 10$. On the other hand, let K_0, K_1, and K_2 be the modified Bessel functions of the second kind and order 0, 1, and 2, respectively, and let $H_0^{(1)}$ be the Hankel function of the first kind and order 0. Then we choose the data $\mathbf{f}$ and g so that the exact solution of (2.9) is given by

$$\mathbf{u}(\mathbf{x}) = \begin{pmatrix} \dfrac{1}{2\pi}\,\psi(\mathbf{x}) - \dfrac{(x_1-1)^2}{r_1^2}\,\chi(\mathbf{x}) \\ \\ -\dfrac{(x_1-1)\,x_2}{r_1^2}\,\chi(\mathbf{x}) \end{pmatrix} \quad \forall\, \mathbf{x} \in \Omega_s, \quad p(\mathbf{x}) = H_0^{(1)}(\omega\,|\mathbf{x}|) \quad \forall\, \mathbf{x} \in \Omega_f,$$

where

$$\psi(\mathbf{x}) := K_0(\imath\,\omega\, r_1) + \frac{1}{\imath\,\omega\, r_1}\left(K_1(\imath\,\omega\, r_1) - \frac{1}{\sqrt{3}}\,K_1\left(\frac{\imath\,\omega\, r_1}{\sqrt{3}}\right)\right),$$

$$r_1 := \sqrt{(x_1-1)^2 + x_2^2}, \quad \text{and} \quad \chi(\mathbf{x}) := K_2(\imath\,\omega\, r_1) - \frac{1}{3}\,K_2\left(\frac{\imath\,\omega\, r_1}{\sqrt{3}}\right).$$

Actually, $\mathbf{u}$ is the fundamental solution, centered at $(1, 0)$, of the elastodynamic equation, which yields $\mathbf{f} = \mathbf{0}$ in Ω_s, and p is the fundamental solution, centered at the origin, of the Helmholtz equation in Ω_f. In this way, $(\mathbf{u}, p)$ is the solution of (2.9) with nonhomogeneous transmission conditions on Σ and suitable boundary conditions on Γ.

The numerical results shown below were obtained on a Pentium Xeon computer with dual processors, using a MATLAB code. According to the requirements established in our main theorem, Theorem 5.1, for the mesh sizes h and $\hat{h}$, and since the constant C_0 mentioned there is not explicitly known, we simply put a vertex of the

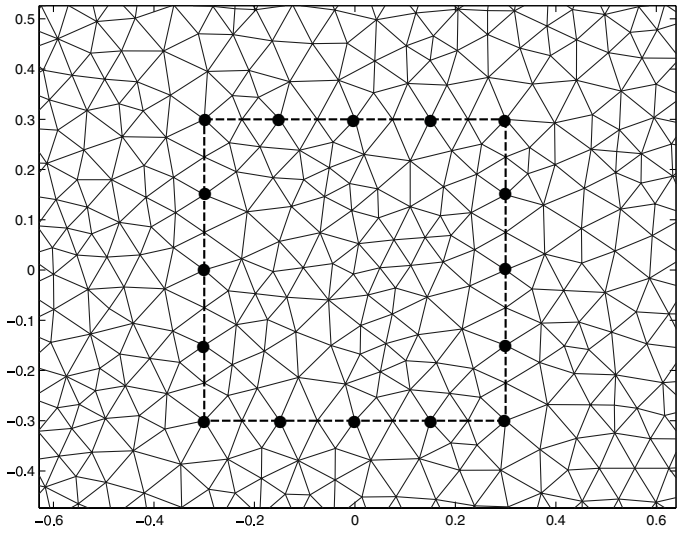


FIG. 6.1. *Zoom around* Σ *(dashed line) of a sample coarse mesh.*

TABLE 6.1
Mesh sizes h*, degrees of freedom* N*, individual errors, and rates of convergence.*

h	N	$\mathsf{e}(\boldsymbol{\sigma})$	$r(\boldsymbol{\sigma})$	$\mathsf{e}(p)$	$r(p)$	$\mathsf{e}(\boldsymbol{\varphi})$	$r(\boldsymbol{\varphi})$	$\mathsf{e}(\boldsymbol{\gamma})$	$r(\boldsymbol{\gamma})$
0.0982	1297	0.220E-00	—	0.103E-00	—	0.253E-01	—	0.942E-02	—
0.0654	2763	0.145E-00	1.02	0.654E-01	1.12	0.121E-01	1.81	0.556E-02	1.30
0.0490	4885	0.109E-00	0.99	0.464E-01	1.18	0.702E-02	1.90	0.320E-02	1.91
0.0321	11506	0.693E-01	1.06	0.288E-01	1.11	0.286E-02	2.10	0.159E-02	1.63
0.0245	19616	0.530E-01	1.00	0.223E-01	0.95	0.197E-02	1.39	0.119E-02	1.09
0.0164	43140	0.357E-01	0.97	0.143E-01	1.10	0.981E-03	1.72	0.664E-03	1.44
0.0123	78271	0.259E-01	1.11	0.108E-01	0.97	0.592E-03	1.75	0.440E-03	1.43
0.0082	175630	0.173E-01	0.99	0.710E-02	1.03	0.296E-03	1.70	0.255E-03	1.34
0.0061	310084	0.131E-01	0.96	0.530E-02	1.01	0.187E-03	1.59	0.181E-03	1.18

independent partition $\{\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_m\}$ every two vertices of $\mathcal{T}_h$ on Σ (see Figure 6.1, where the vertices of $\{\hat{\Sigma}_1, \hat{\Sigma}_2, \ldots, \hat{\Sigma}_m\}$ are marked with bullets). As we will see below, this choice works out well in the present example. In addition, there is no need to take sufficiently small values of $\hat{h}$ and h (as technically suggested by the inequalities $\hat{h} \leq \tilde{h}_0$ and $h \leq \tilde{h}_1$ in Theorem 5.1) since the resulting discrete schemes all become well posed for the degrees of freedom employed.

In Table 6.1 we present the convergence history of our example for a sequence of quasi-uniform triangulations of the computational domain $\bar{\Omega}_s \cup \bar{\Omega}_f$. We see there that the dominant error is given by $\mathsf{e}(\boldsymbol{\sigma})$, which is actually a quite frequent fact in many mixed finite element schemes. We also remark that the rate of convergence $O(h)$ predicted by Theorem 5.1 (when $\delta = 1$) is attained for all the unknowns. Moreover, we observe that at some stages the convergence of $\mathsf{e}(\boldsymbol{\varphi})$ and $\mathsf{e}(\boldsymbol{\gamma})$ is even faster than $O(h)$, which could mean either a superconvergence phenomenon of these unknowns or a special feature of this particular example. Finally, we display real and imaginary parts of some components of the exact and approximate solutions (for $N = 43,140$) in Figures 6.2 and 6.3, from which we notice that they are indistinguishable from each other.

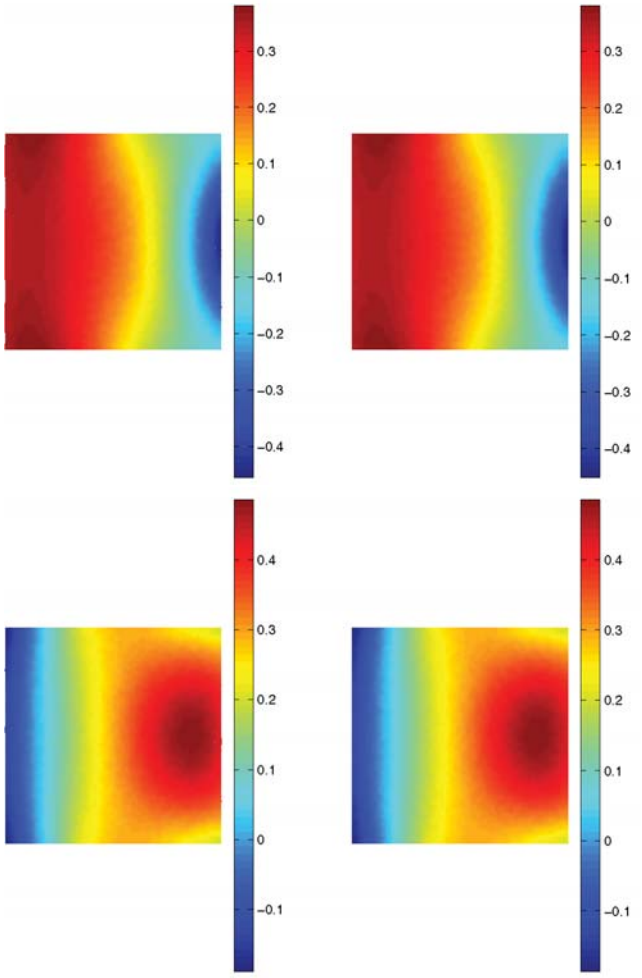


FIG. 6.2. *Approximate (left) and exact (right) real and imaginary parts of* $\boldsymbol{\sigma}_{11}$.

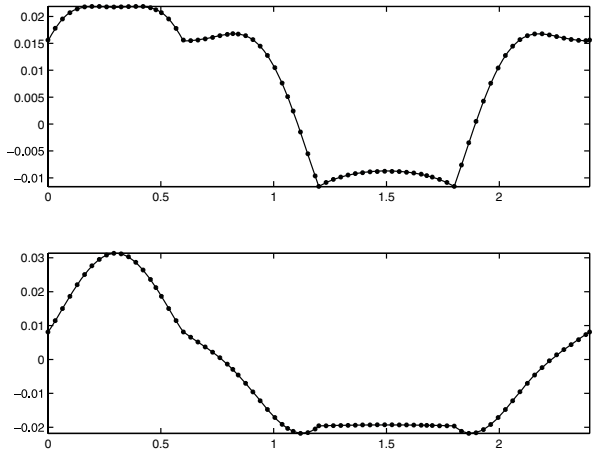


FIG. 6.3. *Approximate (dots) and exact (solid line) real and imaginary parts of* $\boldsymbol{\varphi}_1$.

REFERENCES

[1] D. N. Arnold, F. Brezzi, and J. Douglas, *PEERS: A new mixed finite element method for plane elasticity*, Japan J. Appl. Math., 1 (1984), pp. 347–367.

[2] D. N. Arnold, J. Douglas, and Ch. P. Gupta, *A family of higher order mixed finite element methods for plane elasticity*, Numer. Math., 45 (1984), pp. 1–22.

[3] J.-C. Autrique and F. Magoules, *Numerical analysis of a coupled finite-infinite element method for exterior Helmholtz problems*, J. Comput. Acoust., 14 (2006), pp. 21–43.

[4] I. Babuška and A. K. Aziz, *Survey lectures on the mathematical foundations of the finite element method*, in The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations, A. K. Aziz, ed., Academic Press, New York, 1972, pp. 1–359.

[5] I. Babuška and G. N. Gatica, *On the mixed finite element method with Lagrange multipliers*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 192–210.

[6] M. A. Barrientos, G. N. Gatica, and E. P. Stephan, *A mixed finite element method for nonlinear elasticity: Two-fold saddle point approach and a-posteriori error estimate*, Numer. Math., 91 (2002), pp. 197–222.

[7] J.-P. Berenger, *A perfectly matched layer for the absorption of electromagnetic waves*, J. Comput. Phys., 114 (1994), pp. 185–200.

[8] J. Bielak and R. C. MacCamy, *Symmetric finite element and boundary integral coupling methods for fluid-solid interaction*, Quart. Appl. Math., 49 (1991), pp. 107–119.

[9] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.

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

[11] A. Buffa, *Remarks on the discretization of some noncoercive operator with applications to heterogeneous Maxwell equations*, SIAM J. Numer. Anal., 43 (2005), pp. 1–18.

[12] L. Demkowicz and F. Ihlenburg, *Analysis of a coupled finite-infinite element method for exterior Helmholtz problems*, Numer. Math., 88 (2001), pp. 43–73.

[13] L. Demkowicz and J. Shen, *A few new (?) facts about infinite elements*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3572–3590.

[14] G. N. Gatica, *Analysis of a new augmented mixed finite element method for linear elasticity allowing* $\mathbb{RT}_0$ - $\mathbb{P}_1$ - $\mathbb{P}_0$ *approximations*, Math. Model. Numer. Anal., 40 (2006), pp. 1–28.

[15] G. N. Gatica, A. Márquez, and S. Meddahi, *A New Dual-Mixed Finite Element Method for the Plane Linear Elasticity Problem with Pure Traction Boundary Conditions*, Preprint 2006-23, Departamento de Ingeniería Matemática, Universidad de Concepción, 2006.

[16] G. N. Gatica and S. Meddahi, *On the coupling of MIXED-FEM and BEM for an exterior Helmholtz problem in the plane*, Numer. Math., 100 (2005), pp. 663–695.

[17] G. N. Gatica and W. L. Wendland, *Coupling of mixed finite elements and boundary elements for a hyperelastic interface problem*, SIAM J. Numer. Anal., 34 (1997), pp. 2335–2356.

[18] D. Givoli, *Recent advances in the DtN FE method*, Arch. Comput. Methods Engrg., 6 (1999), pp. 71–116.

[19] D. Givoli, J. B. Keller, and I. Patlashenko, *Discrete Dirichlet-to-Neumann maps for unbounded domains*, Comput. Methods Appl. Mech. Engrg., 164 (1998), pp. 173–185.

[20] P. Grisvard, *Elliptic Problems in Non-smooth Domains*, Monogr. Stud. Math. 24, Pitman, Boston, MA, 1985.

[21] P. Grisvard, *Probléms aux limites dans les polygones. Mode démploi*, EDF Bull. Direction Études Rech. Sér. C Math. Inform., 1 (1986), pp. 21–59.

[22] I. Harari and U. Albocher, *Studies of FE/PML for exterior problems of time-harmonic elastic waves*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3854–3879.

[23] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.

[24] F. Ihlenburg, *Finite Element Analysis of Acoustic Scattering*, Springer-Verlag, New York, 1998.

[25] R. Kress, *Linear Integral Equations*, Springer-Verlag, Berlin, 1989.

[26] M. Lonsing and R. Verfürth, *On the stability of BDMS and PEERS elements*, Numer. Math., 99 (2004), pp. 131–140.

[27] A. Márquez, S. Meddahi, and V. Selgas, *A new BEM-FEM coupling strategy for two-dimensional fluid-solid interaction problems*, J. Comput. Phys., 199 (2004), pp. 205–220.

[28] S. Meddahi and F.-J. Sayas, *Analysis of a new BEM-FEM coupling for two dimensional fluid-solid interaction*, Numer. Methods Partial Differential Equations, 21 (2005), pp. 1017–1042.

[29] S. Prössdorf and B. Silbermann, *Numerical Analysis for Integral and Related Operator Equations*, Birkhäuser-Verlag, Basel, 1991.

[30] J. E. Roberts and J. M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, Finite Element Methods (Part 1), P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, 1991, pp. 523–639.

[31] R. Stenberg, *A family of mixed finite elements for the elasticity problem*, Numer. Math., 53 (1988), pp. 513–538.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2098–2119

HIGH-ORDER RELAXATION SCHEMES FOR NONLINEAR DEGENERATE DIFFUSION PROBLEMS*

FAUSTO CAVALLI†, GIOVANNI NALDI†, GABRIELLA PUPPO‡, AND MATTEO SEMPLICE†

Abstract. Several relaxation approximations to partial differential equations have been recently proposed. Examples include conservation laws, Hamilton–Jacobi equations, convection-diffusion problems, and gas dynamics problems. The present paper focuses on diffusive relaxation schemes for the numerical approximation of nonlinear parabolic equations. These schemes are based on a suitable semilinear hyperbolic system with relaxation terms. High-order methods are obtained by coupling ENO and weighted essentially nonoscillatory (WENO) schemes for space discretization with implicit-explicit (IMEX) schemes for time integration. Error estimates and a convergence analysis are developed for semidiscrete schemes with a numerical analysis for fully discrete relaxed schemes. Various numerical results in one and two dimensions illustrate the high accuracy and good properties of the proposed numerical schemes, also in the degenerate case. These schemes can be easily implemented on parallel computers and applied to more general systems of nonlinear parabolic equations in two- and three-dimensional cases.

Key words. parabolic problems, relaxation schemes, high-order accuracy, porous media equation, WENO reconstruction

AMS subject classifications. 65M20, 65M12, 35K65

DOI. 10.1137/060664872

1. Introduction.

Relaxation approximations to nonlinear partial differential equations have been introduced in [22] (conservation laws) and [28] (degenerate diffusion) and later studied also in [2, 1, 21, 25, 29, 27]. The main idea is to approximate the original partial differential equation with a suitable semilinear hyperbolic system with stiff relaxation terms. As the relaxation parameter $\varepsilon \to 0$, the solution of the system converges to the solution of the original partial differential equation.

Moreover, appropriate numerical schemes for the relaxation system yield accurate numerical approximations to the original equation or system when the relaxation rate ϵ is sufficiently small. Numerically, the main advantage of solving the relaxation model over the original conservation law lies in the simple linear structure of characteristic fields and in the fact that the lower-order term is localized. In particular, the semilinear nature of the relaxation system gives a new way to develop numerical schemes that are simple, general, and Riemann solver free [20, 22].

The aim of this work is to analyze, from both a theoretical and a computational point of view, relaxation schemes for the numerical approximation of the following nonlinear degenerate diffusion problem:

$$\frac{\partial u}{\partial t} = D\Delta(p(u)), \quad x \in \mathbb{R}^d,\ t > 0, \tag{1.1}$$

with initial data $u(x,0) = u_0(x) \in L^1(\mathbb{R}^d)$. $D > 0$ is a diffusivity coefficient. As

*Received by the editors July 12, 2006; accepted for publication (in revised form) May 31, 2007; published electronically September 28, 2007. This work was partially supported by the MIUR/PRIN2005 project "Modellistica numerica per il calcolo scientifico ed applicazioni avanzate."
http://www.siam.org/journals/sinum/45-5/66487.html

†Dipartimento di Matematica, Università di Milano,Via Saldini 50, I-20133 Milano, Italy (cavalli@mat.unimi.it, giovanni.naldi@unimi.it, semplice@mat.unimi.it).

‡Dipartimento di Matematica, Politecnico di Torino, Corso Duca degli Abruzzi 24, 10129 Torino, Italy (gabriella.puppo@polito.it).

usual, we will assume $p : \mathbb{R} \to \mathbb{R}$ to be nondecreasing and Lipschitz continuous [36]. The equation is degenerate if $p(0) = 0$, and the set of points where $u(x)$ becomes 0 is called the interface.

In the case $p(u) = u^m$, $m > 1$, the previous equation is the *porous media equation*, which describes the flow of a gas through a porous interface according to some constitutive relation like Darcy's law in order to link the velocity of the gas and its pressure. With this choice of p, the diffusion coefficient mu^{m-1} vanishes at the points where $u = 0$. Thus the porous media equation is necessarily degenerate for compactly supported initial data [3], and the interfaces exhibit a finite speed of propagation. The degeneracy of the diffusion terms makes the dynamics of the interfaces difficult to study from both the theoretical and the numerical point of view. In general the numerical analysis of (1.2) is difficult for at least two reasons: the appearance of singularities for compactly supported solutions and the growth of the size of the support as time increases (*retention property*).

A common numerical technique to approximate (1.1) involves implicit discretization in time: It requires, at each time step, the solution of a nonlinear algebraic system, which can be singular on the interface. Another possibility is to linearize the nonlinear problem in order to take advantage of efficient linear solvers. For example, linear approximation schemes based on the so-called nonlinear Chernoff formula with a suitable relaxation parameter have been studied in [6, 26, 30, 31]. Other linear approximation schemes have been introduced by Jäger, Kačur, and Handlovičová [19, 23]. Also, a new scheme based on the maximum principle and on a perturbation and regularization approach was proposed by Pop and Yong in [33]. In the more general convection-diffusion case other approaches were investigated in the work of Evje and Karlsen [16], based on a suitable splitting of the convection and the diffusion operators, with a front tracking method for the advection term and implicit numerical integration of the latter. This approach limits the achievable order of accuracy and requires nonlinear solvers for the elliptic part.

A relaxation system to approximate degenerate parabolic equations was proposed originally in [28], inspired by kinetic schemes for the Carleman model. The convergence of the analytical solutions of the relaxation system to those of the partial differential equation is proven in [25], for the $u_t - (u^m)_{xx} = 0$ equation with $m > 0$ (porous media and fast-diffusion equations). The numerical integration of the relaxation system is performed at the macroscopic level, leading to the schemes proposed in [29, 27], where the kinetic derivation of the relaxation system is not relevant any more.

Natalini and coworkers proposed a kinetic approach to the numerical integration of conservation laws [2] and of convection-diffusion problems [7, 1]. Their work is based on a Bhatnagar–Gross–Krook (BGK) approximation which, despite being inspired by the work of Kurz and McKean as [28], is different at the kinetic level, as detailed in [24].

The aim of the present work is to obtain high-order numerical schemes in time and space for the integration of (1.1), following and developing the ideas of [27]. While in [27], the main focus is the development of suitable relaxation systems for several partial differential equations, here we will concentrate on the numerical analysis of the schemes resulting from the relaxation system. In particular we prove the convergence of the semidiscrete scheme, study the stability (linear and nonlinear) of the fully discrete scheme, and propose the construction of high-order extensions. In order to obtain higher-order methods, we couple ENO and weighted essentially nonoscillatory (WENO) schemes for space discretization and implicit-explicit (IMEX) schemes for time advancement. The schemes we obtain avoid both operator splitting techniques and implicit nonlinear solvers.

We point out that, despite the fact that high-order schemes may not reach their order of convergence due to the loss of regularity of the solution during the evolution, they are nevertheless interesting for error reduction when the number of grid points is fixed or until discontinuities develop (both cases arise, for example, in nonlinear filtering in image analysis [37]).

Note that the relaxation system we consider, following [27], can be obtained also as a three velocity model in the BGK approach of [24]. However, in the present case we are interested in the relaxed scheme, i.e., the $\varepsilon = 0$ limit of a numerical scheme for the relaxation system. For this reason, the numerical scheme we propose is different from those of [1], as described in the following section.

Our approach allows us to obtain numerical schemes for (1.1) that are easy to implement and suited for parallel coding, even in the multidimensional case and for more general and complex problems, such as oil recovery problems [15].

Equation (1.1) is a particular case of the more general convection diffusion equation

$$\frac{\partial u}{\partial t} + \operatorname{div} f(u) = D\Delta(p(u)), \quad x \in \mathbb{R}^d,\ t > 0. \tag{1.2}$$

The approach described in this paper can be extended to this more general case, introducing an additional equation to allow the relaxation of the convective term. However, this can be achieved in several ways, leading to different numerical schemes for the partial differential equation (1.2). The stability and efficiency of these schemes can differ wildly and will be the subject of further work [11].

The paper is organized as follows. Section 2 is devoted to the introduction of our relaxation schemes. The stability and error estimates of the semidiscrete scheme are provided in section 3. In section 4 we consider the fully discrete relaxed scheme with a nonlinear stability analysis and the extension to the multidimensional case. We also study parabolic problems in a bounded domain $\Omega \subset \mathbb{R}^d$ with Neumann boundary conditions. Finally, the implementation of the method as well as the results of several numerical experiments are discussed in section 5.

2. Relaxation approximation of nonlinear diffusion. The schemes proposed in the present work are based on the idea at the basis of the well-known relaxation schemes for hyperbolic conservation laws [22]. In the case of the nonlinear diffusion operator, an additional variable $\vec{v}(x,t) \in \mathbb{R}^d$ and a positive parameter ε are introduced, obtaining the following relaxation system:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \operatorname{div}(\vec{v}) = 0, \\ \dfrac{\partial \vec{v}}{\partial t} + \dfrac{D}{\varepsilon}\nabla p(u) = -\dfrac{1}{\varepsilon}\vec{v}. \end{cases} \tag{2.1}$$

Formally, in the small relaxation limit, $\varepsilon \to 0^+$, system (2.1) approximates to leading order (1.1). Next, we remove the nonlinear term from the second equation, as in standard relaxation schemes, introducing a variable $w(x,t) \in \mathbb{R}$ and rewriting the system as:

$$\begin{cases} \dfrac{\partial u}{\partial t} + \operatorname{div}(\vec{v}) = 0, \\ \dfrac{\partial \vec{v}}{\partial t} + \dfrac{D}{\varepsilon}\nabla w = -\dfrac{1}{\varepsilon}\vec{v}, \\ \dfrac{\partial w}{\partial t} + \operatorname{div}(\vec{v}) = -\dfrac{1}{\varepsilon}(w - p(u)). \end{cases} \tag{2.2}$$

Formally, as $\varepsilon \to 0^+$, $w \to p(u)$, $v \to -D\nabla p(u)$, and the original diffusion equation (1.1) is recovered. As a matter of fact, this convergence can be justified rigorously by the results of section 3 of [7], since the relaxation system (2.2) can be seen as a particular case of the BGK system in [7]. Hence we are guaranteed that the solutions of (2.2) converge to the solutions of the degenerate parabolic equation when $\varepsilon \to 0^+$.

For the numerical integration of (2.2) one has to deal with the stiff characteristic velocities due to the term $\nabla(w)/\varepsilon$. In [1], the authors propose two possible methods: either choose ε dependent of the space discretization h or consider $\varepsilon = 0$ and use a splitting technique. Instead, we introduce a suitable parameter φ and rewrite the system (2.2) as

$$\begin{cases} \dfrac{\partial u}{\partial t} + \operatorname{div}(\vec{v}) = 0, \\ \dfrac{\partial \vec{v}}{\partial t} + \varphi^2 \nabla w = -\dfrac{1}{\varepsilon}\vec{v} + \left(\varphi^2 - \dfrac{D}{\varepsilon}\right)\nabla w, \\ \dfrac{\partial w}{\partial t} + \operatorname{div}(\vec{v}) = -\dfrac{1}{\varepsilon}(w - p(u)). \end{cases} \tag{2.3}$$

We anticipate here that we intend to integrate implicitly the terms on the right-hand side of system (2.3), so that we can consider the case $\varepsilon = 0$ without being limited by the stiffness of the problem. In particular, in the relaxed case (i.e., $\varepsilon = 0$), the stiff source terms can be integrated by solving a system that is already in a suitable triangular form and does not require iterative solvers.

In the previous system the parameter ε has physical dimensions of time and represents the so-called relaxation time. Furthermore, w has the same dimensions as u, while each component of $\vec{v}$ has the dimension of u times a velocity; finally φ is a velocity. The inverse of ε gives the rate at which v decays onto $-\nabla p(u)$ in the evolution of the variable $\vec{v}$ governed by the stiff second equation of (2.3).

Equations (2.3) form a semilinear hyperbolic system with a stiff source term. The characteristic velocities of the hyperbolic part are given by $0, \pm\varphi$. The parameter φ allows one to "move" the stiff terms $\frac{D}{\varepsilon}\nabla p(u)$ to the right-hand side, without losing the hyperbolicity of the system.

We point out that degenerate parabolic equations often model physical situations with free boundaries or discontinuities: We expect that schemes for hyperbolic systems will be able to reproduce faithfully these details of the solution. One of the main properties of (2.3) consists in the semilinearity of the system; that is, all of the nonlinearities are in the (stiff) source terms, while the differential operator is linear. Hence, the solution of the convective part requires neither Riemann solvers nor the computation of the characteristic structure at each time step, since the eigenstructure of the system is constant in time. Moreover, the relaxation approximation does not exploit the form of the nonlinear function p, and hence it gives rise to a numerical scheme that, to a large extent, is independent of it, resulting in a very versatile tool.

3. The semidiscrete scheme. System (2.3) can be written in the form:

$$z_t + \operatorname{div} f(z) = \frac{1}{\varepsilon} g(z), \tag{3.1}$$

where

$$z = \begin{pmatrix} u \\ v \\ w \end{pmatrix}, \qquad f(z) = \begin{bmatrix} v^T \\ \Phi^2 w \\ v^T \end{bmatrix}, \qquad g(z) = \begin{pmatrix} 0 \\ -v + (\varphi^2\varepsilon - D)\nabla w \\ p(u) - w \end{pmatrix}, \tag{3.2}$$

and Φ^2 is the $d \times d$ identity matrix times the scalar φ^2. We start discretizing the system in time using, for simplicity, a uniform time step Δt. Let $z^n(x) = z(x, t^n)$, with $t^n = n\Delta t$. Since (3.1) involves both stiff and nonstiff terms, it is a natural idea to employ different time-discretization strategies for each of them, as in [4, 32]. In this work we integrate (3.1) with a Runge–Kutta IMEX scheme [32], obtaining the following semidiscrete formulation:

$$z^{n+1} = z^n - \Delta t \sum_{i=1}^{\nu} \tilde{b}_i \mathrm{div} f(z^{(i)}) + \frac{\Delta t}{\varepsilon} \sum_{i=1}^{\nu} b_i g(z^{(i)}), \tag{3.3}$$

where the $z^{(i)}$'s are the stage values of the Runge–Kutta scheme which are given by

$$z^{(i)} = z^n - \Delta t \sum_{k=1}^{i-1} \tilde{a}_{i,k} \mathrm{div} f(z^{(k)}) + \frac{\Delta t}{\varepsilon} \sum_{k=1}^{i} a_{i,k} g(z^{(k)}), \tag{3.4}$$

where $\tilde{b}_i$, $\tilde{a}_{ij}$ and b_i, a_{ij} denote the coefficients of the explicit and implicit Runge–Kutta schemes, respectively. We assume that the implicit scheme is of the diagonally implicit type. To find the $z^{(i)}$'s it is necessary in principle to solve a nonlinear system of equations which, however, can be easily decoupled. The system for the first stage $z^{(1)}$ at time t^n is

$$\begin{pmatrix} u^{(1)} \\ v^{(1)} \\ w^{(1)} \end{pmatrix} = \begin{pmatrix} u^n \\ v^n \\ w^n \end{pmatrix} + \frac{\Delta t}{\varepsilon} a_{11} \begin{pmatrix} 0 \\ -v^{(1)} + (\varphi^2\varepsilon - D)\nabla w^{(1)} \\ p(u^{(1)}) - w^{(1)} \end{pmatrix}. \tag{3.5}$$

The first equation yields $u^{(1)} = u^n$; substituting in the third equation, we immediately find $w^{(1)}$; and finally, substituting $w^{(1)}$ in the second equation, we compute $v^{(1)}$. In other words, the system can be written in triangular form. For the following stage values, by grouping the already computed terms in the vector $B^{(i)}$ given by

$$B^{(i)} = z^n - \Delta t \sum_{k=1}^{i-1} \tilde{a}_{i,k} \mathrm{div} f(z^{(k)}) + \frac{\Delta t}{\varepsilon} \sum_{k=1}^{i-1} a_{i,k} g(z^{(k)}), \tag{3.6}$$

the new stage values are given by

$$\begin{pmatrix} u^{(i)} \\ v^{(i)} \\ w^{(i)} \end{pmatrix} = B^{(i)} + \frac{\Delta t}{\varepsilon} a_{ii} \begin{pmatrix} 0 \\ -v^{(i)} + (\varphi^2\varepsilon - D)\nabla w^{(i)} \\ p(u^{(i)}) - w^{(i)} \end{pmatrix}, \tag{3.7}$$

which is again a triangular system. In the numerical tests, we will apply IMEX schemes of order 1, 2, and 3.

Following [22] we set $\varepsilon = 0$, thus obtaining the so-called *relaxed scheme*. The computation of the first stage reduces to

$$\begin{aligned} u^{(1)} &= u^n, \\ w^{(1)} &= p(u^{(1)}), \\ v^{(1)} &= -D\nabla w^{(1)}. \end{aligned} \tag{3.8}$$

For the following stages the first equation is

$$u^{(i)} = u^n - \Delta t \sum_{k=1}^{i-1} \tilde{a}_{i,k} \mathrm{div} v^{(k)}. \tag{3.9}$$

In the other equations the convective terms are dominated by the source terms, and thus $v^{(i)}$ and $w^{(i)}$ are given by

$$\begin{aligned} v^{(i)} &= -D \nabla w^{(i)}, \\ w^{(i)} &= p(u^{(i)}). \end{aligned} \tag{3.10}$$

We see that only the explicit part of the Runge–Kutta method is involved in the updating of the solution. Then, in the relaxed schemes we use only the explicit part of the tableaux. In particular we consider second- and third-order strongly stable Runge–Kutta (SSRK) schemes [17], namely,

`IMEX1` (1^{st} order)

0	
1	

`IMEX2` (2^{nd} order)

0	0	
1	0	
$\frac{1}{2}$	$\frac{1}{2}$	

`IMEX3` (3^{rd} order)

0	0	0	
1	0	0	
$\frac{1}{4}$	$\frac{1}{4}$	0	
$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{3}$	

In [10] we studied the increase in efficiency obtained by using suitable strongly stable Runge–Kutta schemes.

3.1. Convergence of the semidiscrete relaxed scheme. The aim of this section is to show the L^1 convergence of the solution of the semidiscrete in time relaxed scheme defined by (3.8), (3.9), and (3.10). We will extend the theorem proved in [6], where only the case of forward Euler time stepping was considered. In this section, for the sake of simplicity, we set $D = 1$.

Theorem 3.1 proves that the numerical solution of the relaxed scheme converges to the solution of (1.1). The proof does not make explicit use of the convergence of the solutions of the relaxation system (2.3) to the solutions of (1.1).

Eliminating v from (3.8) and (3.9) using (3.10), we rewrite the relaxed scheme as

$$\begin{aligned} u^{(1)} &= u^n, \\ w^{(1)} &= p(u^n) \end{aligned} \tag{3.11}$$

for the first stage, and

$$\begin{aligned} u^{(i)} &= u^n + \Delta t \sum_{k=1}^{i-1} \tilde{a}_{i,k} \Delta w^{(k)}, \\ w^{(i)} &= p(u^{(i)}) \end{aligned} \tag{3.12}$$

for subsequent stages. We recall that a Runge–Kutta scheme for the ordinary differential equation $y' = R(y)$ can also be written in the form [17]

$$\begin{aligned} y^{(1)} &= y^n, \\ y^{(i)} &= \sum_{k=1}^{i-1} \alpha_{ik} \left(y^{(k)} + \Delta t \frac{\beta_{ik}}{\alpha_{ik}} R(y^{(k)}) \right), \qquad i = 2, \ldots, \nu, \end{aligned} \tag{3.13}$$

where $y^{n+1} = y^{(\nu)}$. For consistency, $\sum_{k=1}^{i-1} \alpha_{ik} = 1$ for every $i = 1, \ldots, \nu$. Moreover we assumed that $\alpha_{ik} \geq 0$ and $\beta_{ik} \geq 0$ and that $\alpha_{ik} = 0$ implies $\beta_{ik} = 0$. Under these assumptions, each stage value $y^{(i)}$ can be written as a convex combination of forward Euler steps. This remark allows us to study the convergence of the Runge–Kutta scheme in terms of the convergence of the explicit forward Euler scheme applied to the nonlinear diffusion problem.

This latter was studied in [6] via a nonlinear semigroup argument. In the following we review the approach of [6], and next we extend the proof to the case of a ν-stages explicit Runge–Kutta scheme.

3.1.1. The forward Euler case. We wish to solve the evolution equation

$$\frac{du}{dt} + Lp(u) = 0, \qquad u(\cdot, t = 0) = u_0, \tag{3.14}$$

on the domain Ω, where $L = -\Delta$ and $p : \mathbb{R} \to \mathbb{R}$ is a nondecreasing locally Lipschitz function such that $p(0) = 0$. Under these hypotheses, the nonlinear operator $Au = Lp(u)$ with domain $D(A) = \{u \in L^1(\Omega) : p(u) \in D(L)\}$ is m-accretive in $L^1(\Omega)$; that is, for all $\varphi \in \mathrm{L}^1(\Omega)$ and for all $\lambda > 0$ there exists a unique solution $u \in D(A)$ such that $u + \lambda Lp(u) = \varphi$ and the application defined by $\varphi \mapsto u$ is a contraction [14].

Moreover $D(A)$ is dense in $L^1(\Omega)$, so it follows that

$$S_A(t)u_0 = \lim_{m\to\infty} \left(\mathbb{I} + \frac{t}{m}A\right)^{-m} u_0 \tag{3.15}$$

is a contraction semigroup on $L^1(\Omega)$ and $S_A(t)u_0$ is the generalized solution of (3.14) in the sense of Crandall–Liggett [14]. Let $S(t)$ be the linear contraction semigroup generated by $-L$; that is, $u(t) = S(t)u_0$ is the solution of the initial value problem $u_t = -L(u)$ and $u(\cdot, t = 0) = u_0$. The algorithm proposed in [6] is

$$\frac{u^{n+1} - u^n}{\tau} + \left[\frac{\mathbb{I} - S(\sigma_\tau)}{\sigma_\tau}\right] p(u^n) = 0, \tag{3.16}$$

where τ is the time step and $\sigma_\tau \downarrow 0$. This can be written as

$$u^{n+1} = F_E(\tau)u^n, \quad \text{where } F_E(\tau)\varphi = \varphi + \frac{\tau}{\sigma_\tau}\left[S(\sigma_\tau) - \mathbb{I}\right] p(\varphi). \tag{3.17}$$

Hence

$$u^n = (F_E(\tau))^n u_0. \tag{3.18}$$

The proof in [6] is based on the following argument. Note that formally $S(\sigma_\tau)\varphi \sim e^{-\sigma_\tau L}\varphi$. Let $t = \tau n$ and

$$\begin{aligned} u(t) &= \left[\mathbb{I} + \frac{t}{n\sigma_\tau}\left(S(\sigma_\tau) - \mathbb{I}\right) \circ p\right]^n u_0 \\ &= \left[\mathbb{I} + \frac{t}{n\sigma_\tau}\left(e^{-\sigma_\tau L} - \mathbb{I}\right) \circ p\right]^n u_0 \qquad \text{if } \sigma_\tau \to 0 \\ &= \left[\mathbb{I} - \frac{t}{n} L \circ p\right]^n u_0 \\ &\to S_A(u_0) \qquad \text{when } n \to \infty. \end{aligned} \tag{3.19}$$

The convergence proof requires that $\mu\frac{\tau}{\sigma_\tau} \leq 1$, where μ is the Lipschitz constant of $p(u)$. We point out that σ_τ is linked to the spatial approximation of the operator L, and in our scheme this requirement is reflected in the stability condition of the fully discrete scheme (see section 4).

3.1.2. Runge–Kutta schemes. Now we are going to prove convergence for the case of a ν-stages Runge–Kutta scheme.

Let $t > 0$ and $\tau = t/n$, with $n \geq 1$; let $\sigma_\tau : (0,\infty) \to (0,\infty)$ be a function such that $\lim_{\tau\to 0} \sigma_\tau = 0$.

$$\begin{aligned} &u^{(1)} = u^n, \\ &u^{(i)} = \sum_{k=1}^{i-1} \alpha_{ik} \left[u^{(k)} + \tau \frac{\beta_{ik}}{\alpha_{ik}} A(u^{(k)}) \right], \qquad i = 2, \dots, \nu, \end{aligned} \tag{3.20}$$

and proceeding as in (3.19), this becomes

$$\begin{aligned} &u^{(1)} = u^n, \\ &u^{(i)} = \sum_{k=1}^{i-1} \alpha_{ik} \left[u^{(k)} + \tau \frac{\beta_{ik}}{\alpha_{ik}} (S(\sigma_\tau) - \mathbb{I}) \circ p(u^{(k)}) \right], \qquad i = 2, \dots, \nu, \\ &u^{n+1} = u^{(\nu)}. \end{aligned} \tag{3.21}$$

We now extend (3.17) to the Runge–Kutta scheme defined by (3.21). Define, for $\phi \in L^1(\Omega)$,

$$\begin{aligned} &F^{(1)}(\tau)\phi = \phi, \\ &F^{(i)}(\tau)\phi = \sum_{k=1}^{i-1} \alpha_{ik} F^{(k)}(\tau)\phi + \frac{\tau\beta_{ik}}{\sigma_\tau} \left[S(\sigma_\tau) - \mathbb{I} \right] p(F^{(k)}(\tau)\phi), \\ &F(\tau)\phi = F^{(\nu)}(\tau)\phi, \end{aligned} \tag{3.22}$$

and therefore

$$u^n(t) = [F(\tau)]^n u_0. \tag{3.23}$$

Let $u(t)$ be the generalized solution of (3.14). The following theorem proves the convergence of the semidiscrete solution to $u(t)$.

THEOREM 3.1. *Assume $u^0 \in L^\infty(\Omega)$, and $\|u^0\|_\infty = M$; let p be a nondecreasing Lipschitz continuous function on $[-M, M]$ with Lipschitz constant μ. Assume that the following conditions hold:*

$$\begin{cases} \alpha_{ik} \geq 0, \\ \beta_{ik} \geq 0, \\ \alpha_{ik} = 0 \Rightarrow \beta_{ik} = 0, \\ \displaystyle\sum_{k=1}^{i-1} \alpha_{ik} = 1 \ \textit{(consistency)}, \\ \dfrac{\mu\tau}{\sigma_\tau} \leq \min \dfrac{\alpha_{ik}}{\beta_{ik}} \quad \textit{for } \tau > 0, \alpha_{ik} \neq 0 \qquad \textit{(stability)}, \end{cases} \tag{3.24}$$

and then $\lim_{n\to\infty} u^n(t) = u(t)$ in L^1. Moreover the convergence is uniform for t in any given bounded interval.

The proof follows the steps of [6]: First we show that u^n verifies a maximum principle (Lemma 3.2) and that F is a contraction (Lemma 3.3), and finally we apply the nonlinear Chernoff formula [8].

LEMMA 3.2. *If* (3.24) *is verified, then* $-M \le u^n \le M$ *for all* n.

Proof. We argue by induction on n: We assume that $-M \le u^n \le M$, and we show that $-M \le u^{n+1} \le M$. Let

$$u^{(i)} = F^{(i)}(\tau)u^n. \tag{3.25}$$

Since $u^{n+1} = u^{(\nu)}$, it suffices to prove that $-M \le u^{(i)} \le M$ for $i = 1, \dots, \nu$. We prove this by induction on i. When $i = 1$, the statement is true thanks to the induction hypothesis on n and being $F^{(1)} = \mathbb{I}$. Let's assume that $-M \le u^{(i-1)} \le M$ holds; we are going to show that

$$-M \le u^{(i)} = F^{(i)}(\tau)u^n \le M. \tag{3.26}$$

The function $s \mapsto \alpha_{ik}s - \frac{\tau\beta_{ik}}{\sigma_\tau}p(s)$ is nondecreasing thanks to (3.24) and the hypotheses on the function p. By the induction hypothesis on i, we have that for $k = 1, \dots, i-1$

$$-\alpha_{ik}M - \frac{\tau\beta_{ik}}{\sigma_\tau}p(-M) \le \alpha_{ik}u^{(k)} - \frac{\tau\beta_{ik}}{\sigma_\tau}p(u^{(k)}) \le \alpha_{ik}M - \frac{\tau\beta_{ik}}{\sigma_\tau}p(M). \tag{3.27}$$

Using again the induction hypothesis on i and recalling that p is nondecreasing, since S is a contraction in L^∞ [6] and $p(-M) \le p(u^{(k)}) \le p(M)$,

$$p(-M) \le S\left(p(u^{(k)})\right) \le p(M). \tag{3.28}$$

Multiplying the last equation by $\frac{\tau\beta_{ik}}{\sigma_\tau}$ and summing it to (3.27), we get

$$-\alpha_{ik}M \le \alpha_{ik}u^{(k)} + \frac{\tau\beta_{ik}}{\sigma_\tau}(S - \mathbb{I})p(u^{(k)}) \le \alpha_{ik}M, \qquad k = 1, \dots, i-1. \tag{3.29}$$

Summing for $k = 1, \dots, i-1$ and using the consistency relation of (3.24):

$$-M \le \sum_{k=1}^{i-1} \alpha_{ik}u^{(k)} + \frac{\tau\beta_{ik}}{\sigma_\tau}(S - \mathbb{I})p(u^{(k)}) \le M. \tag{3.30}$$

In particular this is valid when $i = \nu$, proving that $-M \le u^{(n+1)} \le M$. ∎

Now we can replace p by $\overline{p}$, where $\overline{p} = p$ in $-M \le x \le M$, $\overline{p} = p(M)$ for $x \ge M$, and $\overline{p} = p(-M)$ for $x \le -M$: The algorithm is the same, and in what follows we can assume that p is Lipschitz continuous with constant μ on all $\mathbb{R}$.

LEMMA 3.3. *If the hypotheses of Theorem* 3.1 *hold, then* $F(\tau)$ *is a contraction on* $L^1(\Omega)$, *i.e.,*

$$\|F(\tau)\phi - F(\tau)\psi\|_1 \le \|\phi - \psi\|_1 \qquad \forall \psi, \phi \in L^1. \tag{3.31}$$

Proof. We start showing that the result holds for a single forward Euler step. Recalling the definition of F_E from (3.17)

$$\begin{aligned} \|F_E(\tau)\phi - F_E(\tau)\psi\|_1 &\le \frac{\tau}{\sigma_\tau}\left\|S(\sigma_\tau)[p(\phi) - p(\psi)]\right\|_1 + \left\|(\phi - \psi) - \frac{\tau}{\sigma_\tau}[p(\phi) - p(\psi)]\right\|_1 \\ &\le \frac{\tau}{\sigma_\tau}\left\|p(\phi) - p(\psi)\right\|_1 + \left\|\left(\phi - \frac{\tau}{\sigma_\tau}p(\phi)\right) - \left(\psi - \frac{\tau}{\sigma_\tau}p(\psi)\right)\right\|_1 \\ &= \left\|\phi - \psi\right\|_1, \end{aligned} \tag{3.32}$$

where we used the contractivity of S. The last equality relies on the fact that p and the function $x \mapsto x - \frac{\tau}{\sigma_\tau} p(x)$ are nondecreasing, which in turn is guaranteed by the stability condition, which in this case reduces to $\mu\tau/\sigma_\tau \leq 1$ [6].

In the general case we have

$$\begin{aligned} \|F^{(i)}(\tau)\phi - F^{(i)}(\tau)\psi\|_1 &\leq \sum_{k=1}^{i-1} \alpha_{ik} \left\| F_E\left(\frac{\tau\beta_{ik}}{\alpha_{ik}}\right) F^{(k)}(\tau)\phi - F_E\left(\frac{\tau\beta_{ik}}{\alpha_{ik}}\right) F^{(k)}(\tau)\psi \right\|_1 \\ &\leq \sum_{k=1}^{i-1} \alpha_{ik} \left\| F^{(k)}(\tau)\phi - F^{(k)}(\tau)\psi \right\|_1 \qquad (3.33) \\ &\leq \|\phi - \psi\|_1 . \end{aligned}$$

In the second inequality we used the contractivity of F_E and the stability condition, while in the third one we apply an induction argument on the contractivity of $F^{(k)}$, the positivity constraint on α_{ik} and β_{ik}, as well as the consistency condition $\sum_k \alpha_{ik} = 1$. Setting $i = \nu$ yields the result. □

Proof of Theorem 3.1. Let ψ_τ and ψ be, respectively,

$$\psi_\tau = \left(I + \frac{\lambda}{\tau}(I - F(\tau))\right)^{-1} \phi \qquad \text{and} \qquad \psi = (I + \lambda A)^{-1}\phi. \qquad (3.34)$$

The function ψ exists since the operator A is m-accretive, whereas the existence of the function ψ_τ is guaranteed by the following fixed-point argument. Let

$$G(y) = \frac{1}{1+\eta}\phi + \frac{\eta}{\eta+1} F(\tau) y,$$

where $\phi \in \mathrm{L}^1$, $y \in \overline{D(A)}$, and $\eta \geq 0$. We have

$$\|G(y) - G(x)\| = \frac{\eta}{\eta+1}\|F(\tau)y - F(\tau)x\| \leq \frac{\eta}{\eta+1}\|y - x\|$$

since F is a contraction, as proved in Lemma 3.3. Thus G is also a contraction, and therefore it possesses a unique fixed point which coincides with ψ_τ.

We want to show that

$$\psi_\tau \to \psi \qquad \text{in } \mathrm{L}^1$$

as $\tau \to 0$ for each fixed $\lambda > 0$. Let

$$\phi_\tau = \psi + \frac{\lambda}{\tau}(\mathbb{I} - F(\tau))\psi.$$

We want to estimate $\psi_\tau - \psi$ in terms of $\phi_\tau - \phi$.

$$\phi_\tau - \phi = \left(1 + \frac{\lambda}{\tau}\right)(\psi - \psi_\tau) - \frac{\lambda}{\tau}(F(\tau)\psi - F(\tau)\psi_\tau),$$

Therefore

$$\left(1 + \frac{\lambda}{\tau}\right)(\psi - \psi_\tau) - (\phi_\tau - \phi) = \frac{\lambda}{\tau}(F(\tau)\psi - F(\tau)\psi_\tau),$$

and, by taking norms and using the fact that F is a contraction, we have

$$\left|\left(1+\frac{\lambda}{\tau}\right)\|\psi-\psi_\tau\|-\|\phi_\tau-\phi\|\right| \le \left\|\left(1+\frac{\lambda}{\tau}\right)(\psi-\psi_\tau)-(\phi_\tau-\phi)\right\| \le \frac{\lambda}{\tau}\|\psi-\psi_\tau\|.$$

In particular

$$\left(1+\frac{\lambda}{\tau}\right)\|\psi-\psi_\tau\|-\|\phi_\tau-\phi\| \le \frac{\lambda}{\tau}\|\psi-\psi_\tau\|,$$

and therefore $\|\psi-\psi_\tau\| \le \|\phi-\phi_\tau\|$.

Now we estimate $\|\phi-\phi_\tau\|$ in the simple case of a forward Euler scheme. Note that

$$\phi-\phi_\tau = \lambda A\psi - \frac{\lambda}{\tau}(\mathbb{I}-F(\tau))\psi,$$

and thus $\|\phi-\phi_\tau\|$ measures a sort of consistency error. For a single forward Euler step, $F=F_E$, where F_E is defined in (3.17). Thus

$$\|\phi-\phi_\tau\| = \lambda\left\|A\psi-\frac{1}{\sigma_\tau}(\mathbb{I}-S(\sigma_\tau))p(\psi)\right\| \to 0 \tag{3.35}$$

as $\tau\to 0$ since $\frac{\mathbb{I}-S(\sigma_\tau)}{\sigma_\tau}p(\psi)\to Lp(\psi)=A\psi$.

The more general case of a ν-stages Runge–Kutta scheme can be carried out by induction following the procedure already applied in the proofs of the previous lemmas.

We now use Theorem 3.2 of [8], which, specialized to our case, can be written as follows. Assume that $F(\tau):\mathrm{L}^1\to\mathrm{L}^1$ for $\tau>0$ is a family of contractions. Assume further that an m-accretive operator A is given, and let $S(t)$ be the semigroup generated by A. Assume further that the family $F(\tau)$ and the operator A are linked by the following formula:

$$\psi_\tau = \left(I+\frac{\lambda}{\tau}(I-F(\tau))\right)^{-1}\phi \to \psi = (I+\lambda A)^{-1}\phi \tag{3.36}$$

for each $\phi\in\mathrm{L}^1$. Then

$$\lim_{n\to\infty} F\left(\frac{t}{n}\right)^n \phi = S(t)\phi \qquad \forall\phi\in\mathrm{L}^1. \qquad \square$$

4. Fully discrete relaxed scheme. In order to complete the description of the scheme, we need to specify the space discretization. We will use discretizations based on finite differences, in order to avoid cell coupling due to the source terms.

Note that the IMEX technique reduces the integration to a cascade of relaxation and transport steps. The former are the implicit parts of (3.5) and (3.7), while the transport steps appear in the evaluation of the explicit terms $B^{(i)}$ in (3.6). Since (3.5) and (3.7) involve only local operations, the main task of the space discretization is the evaluation of $\mathrm{div}(f)$, where we will exploit the linearity of f in its arguments.

4.1. One-dimensional scheme. Let us introduce a uniform grid on $[a,b]\subset\mathbb{R}$, $x_j = a-\frac{h}{2}+jh$ for $j=1,\dots,n$, where $h=(b-a)/n$ is the grid spacing and n the number of cells. The fully discrete scheme may be written as

$$z_j^{n+1} = z_j^n - \Delta t\sum_{i=1}^{\nu}\tilde{b}_i\left(F^{(i)}_{j+1/2}-F^{(i)}_{j-1/2}\right)+\frac{\Delta t}{\varepsilon}\sum_{i=1}^{\nu}b_i g(z_j^{(i)}), \tag{4.1}$$

where $F^{(i)}_{j+1/2}$ are the numerical fluxes, which are the only items we still need to specify. For convergence it is necessary to write the scheme in conservation form. Thus, following [34], we introduce the function $\hat{F}$ such that

$$f(z(x,t)) = \frac{1}{h}\int_{x-h/2}^{x+h/2} \hat{F}(s,t)\mathrm{d}s \quad \Rightarrow \frac{\partial f}{\partial x}(z(x_j,t)) = \frac{1}{h}\left(\hat{F}(x_{j+1/2},t) - \hat{F}(x_{j-1/2},t)\right).$$

The numerical flux function $F_{j+1/2}$ must approximate $\hat{F}(x_{j+1/2})$.

In order to compute the numerical fluxes, for each stage value, we reconstruct boundary extrapolated data $z^{(i)\pm}_{j+1/2}$ with a nonoscillatory interpolation method from the point values $z^{(i)}_j$ of the variables at the center of the cells. Next we apply a monotone numerical flux to these boundary-extrapolated data.

To minimize numerical viscosity we choose the Godunov flux, which in the present case of a linear system of equations reduces to the upwind flux. In order to select the upwind direction we write the system in characteristic form. The characteristic variables relative to the eigenvalues $\varphi, -\varphi, 0$ (in one space dimension φ reduces to a scalar parameter) are, respectively,

$$U = \frac{\varphi w + v}{2\varphi}, \qquad V = \frac{\varphi w - v}{2\varphi}, \qquad W = u - w. \tag{4.2}$$

Note that $u = U + V + W$. Therefore the numerical flux in characteristic variables is $F_{j+1/2} = (\varphi U^-_{j+1/2}, -\varphi V^+_{j+1/2}, 0)$.

The accuracy of the scheme depends on the accuracy of the reconstruction of the boundary-extrapolated data. For a first-order scheme we use a piecewise constant reconstruction such that $U^-_{j+1/2} = U_j$ and $V^+_{j+1/2} = V_{j+1}$. For higher-order schemes, we use ENO or WENO reconstructions of appropriate accuracy [35].

For $\varepsilon \to 0$ we obtain the relaxed scheme. Recall from (3.10) that the relaxation steps reduce to

$$w^{(i)}_j = p(u^{(i)}_j), \qquad v^{(i)}_j = -D\widehat{\nabla} w^{(i)}_j, \tag{4.3}$$

where $\widehat{\nabla}$ is a suitable approximation of the one-dimensional gradient operator. Thus the transport steps need to be applied only to $u^{(i)}$

$$u^{(i)}_j = u^n_j - \lambda \sum_{k=1}^{i-1} \tilde{a}_{i,k}\left[\varphi\left(U^{(k)-}_{j+1/2} - U^{(k)-}_{j-1/2}\right) - \varphi\left(V^{(k)+}_{j+1/2} - V^{(k)+}_{j-1/2}\right)\right]. \tag{4.4}$$

Finally, taking the last stage value and going back to conservative variables,

$$\begin{aligned} u^{n+1}_j = u^n_j - \frac{\lambda}{2}\sum_{i=1}^{\nu} \tilde{b}_i \Big(& [v^{(i)-}_{j+1/2} + v^{(i)+}_{j+1/2} - (v^{(i)-}_{j-1/2} + v^{(i)+}_{j-1/2})] \\ & + \varphi[w^{(i)-}_{j+1/2} - w^{(i)+}_{j+1/2} - (w^{(i)-}_{j-1/2} - w^{(i)+}_{j-1/2})]\Big). \end{aligned} \tag{4.5}$$

We wish to emphasize that the scheme reduces to the time advancement of the single variable u. Although the scheme is based on a system of three equations, the construction is used only to select the correct upwinding for the fluxes of the relaxed scheme, and the computational cost of each time step remains moderate.

4.2. Nonlinear stability for the first-order scheme. The relaxed scheme in the first-order case reduces to:

$$u_j^{n+1} = u_j^n + \frac{\lambda}{2}\left(\partial_x p(u^n)|_{j+1} - \partial_x p(u^n)|_{j-1}\right) + \frac{\lambda}{2}\varphi\left(p(u_{j+1}^n) - 2p(u_j^n) + p(u_{j-1}^n)\right). \tag{4.6}$$

We wish to compute the restrictions on λ and φ so that the scheme is total variation nonincreasing. We select the centered finite difference formula to approximate the partial derivatives of $p(u)$; we drop the index n and write p_j for $p(u_j^n)$. Define $\Delta_{j+1/2} = \frac{p_{j+1}-p_j}{u_{j+1}-u_j}$, and observe that these quantities are always nonnegative since p is nondecreasing. We obtain

$$\begin{aligned}
\mathrm{TV}(u^{n+1}) &= \sum_j |u_j^{n+1} - u_{j-1}^{n+1}| \\
&\le \sum_j \left\{ \frac{\lambda}{4h}\Delta_{j+3/2}|u_{j+2} - u_{j+1}| + \frac{\lambda}{2}\varphi\Delta_{j+1/2}|u_{j+1} - u_j| \right. \\
&\qquad + \left(1 - \lambda\left(\frac{1}{2h} + \varphi\right)\Delta_{j-1/2}\right)|u_j - u_{j-1}| \\
&\qquad \left. + \frac{\lambda}{2}\varphi\Delta_{j-3/2}|u_{j-1} - u_{j-2}| + \frac{\lambda}{4h}\Delta_{j-5/2}|u_{j-2} - u_{j-3}| \right\}
\end{aligned} \tag{4.7}$$

provided that

$$\left(1 - \lambda\left(\frac{1}{2h} + \varphi\right)\Delta_{j-1/2}\right) \ge 0 \qquad \forall j. \tag{4.8}$$

Assuming that the data have compact support, we can rescale all sums and finally get $\mathrm{TV}(u^{n+1}) \le \mathrm{TV}(u^n)$. Taking into account the Lipschitz condition on p, the scheme is total variation stable provided that (4.8) is satisfied, i.e., that

$$\Delta t \le \frac{2h^2}{\mu}\frac{1}{1+2h\varphi} \simeq \frac{(2-\delta)}{\mu}h^2, \tag{4.9}$$

where δ vanishes as h does. We point out that the stability condition is of the parabolic type. Finally, we observe that, using one-sided approximations for the partial derivatives of p in the scheme (4.6), one gets a stability condition involving the relation $\varphi > 1/h$. This would reintroduce in the scheme the constraint due to the stiffness in the convective term that prompted the introduction of φ in (2.3).

4.3. Linear stability. We study the linear stability of the schemes based on (4.3), (4.4), and (4.5) in the case when $p(u) = u$, by von Neumann analysis. We substitute the discrete Fourier modes $u_j^n = \rho^n e^{i(jk/N)}$ into the scheme, where k is the wave number and N the number of cells. We set $\xi = k/N$ and compute the amplification factor $Z(\xi)$ such that $u_j^{n+1} = Z(\xi)u_j^n$. We can consider ξ as a continuous variable, since the amplification factors for various choices of N all lie on the curves obtained considering the variable $\xi \in [0, 2\pi]$.

First we consider the same scheme studied in the previous section, for comparison purposes. Using piecewise constant reconstructions in space and forward Euler time integration, the amplification factor is $Z(\xi) = 1 + M(\xi)$, where

$$M(\xi) = \frac{\lambda}{h}\left(\cos(\xi) - 1\right)\left(\cos(\xi) + 1 + h\varphi\right).$$

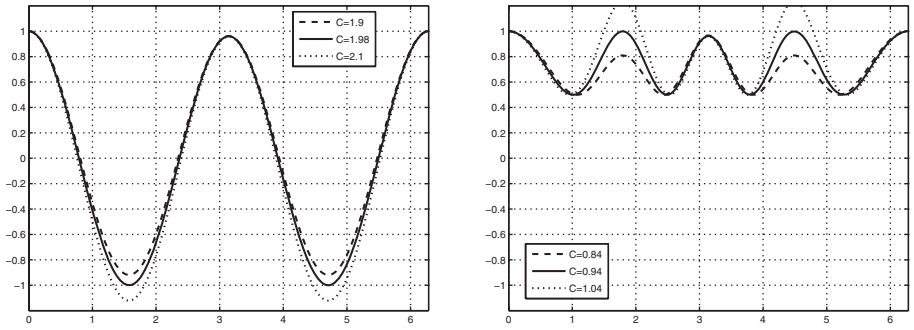


FIG. 4.1. *Amplification factor for upwind spatial reconstruction coupled with forward Euler (left) and for upwind second-order spatial reconstruction coupled with second-order time integration (right).*

$M(\xi)$ takes maximum value 0 and attains its minimum at the point ξ^* such that $\cos(\xi^*) = -\varphi h/2$. Stability requires that $M(\xi^*) \geq -2$, i.e.,

$$1 + \frac{\lambda}{h}\left(\frac{\varphi^2 h^2}{4} - 1\right) - \lambda\varphi\left(\frac{\varphi h}{2} + 1\right) \geq -1,$$

and, recalling that $\lambda = \Delta t/h$,

$$\Delta t \leq \frac{2h^2}{\left(1 + \frac{\varphi h}{2}\right)^2} \simeq 2\,(1 - \varphi h)\,h^2. \tag{4.10}$$

This gives a CFL condition of the form $\Delta t \leq 2(1-\delta)h^2$, where $\delta = O(h\varphi)$ (see Figure 4.1). These results are in very good agreement with those of the nonlinear analysis performed in the previous section.

Now we consider higher-order spatial reconstructions coupled with forward Euler time stepping. M takes the form

$$M(\xi, \gamma) = \frac{\lambda}{h}\left[f_1(\cos(\xi)) + \gamma f_2(\cos(\xi))\right],$$

where $\gamma = h\varphi$. Since γ is small, we compute the critical points ξ^* of $M(\xi, 0)$. For stability we thus require that $-2 \leq M(\xi^*, \gamma) \leq 0$.

We consider a piecewise linear and a WENO reconstruction. The first one is computed along characteristic variables using the upwind slope, while the gradient of $p(u)$ is computed with centered differences. The WENO reconstruction is fifth-order accurate and is obtained by setting to 1 the smoothness indicators, and the gradient of $p(u)$ is computed with the fourth-order centered difference formula.

For the piecewise linear reconstruction, we have that

$$M(\xi) = -\frac{\lambda}{h}\left[(\cos^2(\xi) - 1)(\cos(\xi) - 2) + h\varphi(\cos(\xi) - 1)^2\right],$$

and therefore

$$\Delta t \leq \frac{2h^2}{\frac{20+14\sqrt{7}}{27} + \frac{8+2\sqrt{7}}{9}\varphi h} \simeq 0.94(1 - 1.44\varphi h)h^2.$$

TABLE 4.1

	RK1	RK2	RK3
P-wise constant	2	2	2.51
P-wise linear	0.94	0.94	
WENO5	0.79	0.79	1

For the WENO reconstruction $M(\xi, \gamma)$ can be easily computed, and we get

$$\Delta t \leq 0.79(1 - 0.13\varphi h)h^2.$$

Now we wish to extend our results to the case of higher-order Runge–Kutta schemes. Since both the equation and the scheme are linear, the amplification factors for the Runge–Kutta schemes of orders 2 and 3 used here are, respectively,

$$Z_{(2)}(\xi) = 1 + M(\xi) + \frac{M(\xi)^2}{2},$$
$$Z_{(3)}(\xi) = 1 + M(\xi) + \frac{M(\xi)^2}{2} + \frac{M(\xi)^3}{6},$$

where $M(\xi)$ is the function appearing in the amplification factor relevant to the chosen spatial reconstruction. We have that

$$Z'_{(2)}(\xi) = M'(\xi)(1 + M(\xi)),$$
$$Z'_{(3)}(\xi) = M'(\xi)\left(1 + M(\xi) + \frac{M(\xi)^2}{2}\right),$$

and therefore the critical points are the points ξ^* such that $M'(\xi^*) = 0$.

In the Runge–Kutta 2 case the stability constraint $\|Z_{(2)}(\xi^*)\| \leq 1$ reduces to the CFL condition for the forward Euler scheme. For Runge–Kutta 3, $\|Z_{(3)}(\xi^*)\| \leq 1$, provided that

$$M(\xi^*) \geq \tilde{s} \simeq -2.51.$$

Notice that this is less restrictive than the Euler and second-order Runge–Kutta schemes for which the stability requirement is $M(\xi^*) \geq -2$.

For the third-order Runge–Kutta scheme with linearized WENO of order 5, we have

$$\Delta t \leq \frac{-\tilde{s}h^2}{2.51 + 0.33\varphi h} \simeq (1 - .1325\varphi h)h^2.$$

Table 4.1 summarizes the stability results obtained in this section by listing the values of the constant C that appears in the stability restriction $\Delta t \leq C(1 - C_1\varphi h)h^2$. Figures 4.1 and 4.2 contain the amplification factors $Z(\xi)$ for $\varphi = 1$ and $h = 10^{-2}$ for various choices of spatial reconstructions and time integration schemes. Each of them contains the curve corresponding to the value of C reported in Table 4.1 and two other close-by values.

4.4. Boundary conditions. Different boundary conditions can be implemented. Here we describe how to implement Neumann boundary conditions, considering for simplicity the one-dimensional case.

We first add g ghost points on each side of the computational domain $[a, b]$, where g depends on the order of the spatial reconstruction. We find a polynomial $q(x)$ of

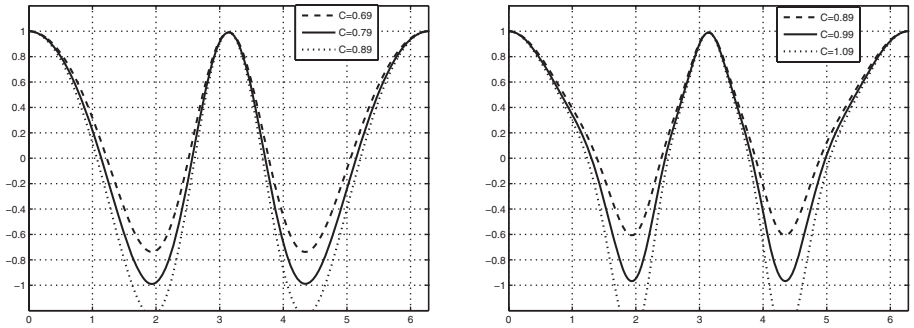


FIG. 4.2. *Amplification factors Z for WENO reconstructions of order* 5 *coupled with first-order (left) and third-order (right) time integration.*

degree d passing through the points (x_i, u_i) for $i = 1, \dots, d$ and having a prescribed derivative at the boundary point $x_{1/2} = a$. (The degree d is determined by the accuracy of the scheme that one wants to obtain and should match the degree of the reconstruction procedures used to obtain $U_j^{\pm}$ and $V_j^{\pm}$.) This polynomial is then used to set the values $u_{-i} = q(x_{-i})$ of the ghost points for $i = 0, 1, g-1$. One operates similarly at the right edge of the computational domain.

We also used periodic boundary conditions, which can be implemented with an obvious choice of the values u_i at the ghost points.

4.5. Multidimensional scheme. An appropriate numerical approximation of (2.3) in $\mathbb{R}^d$ that generalizes the scheme described in section 4.1 can be obtained by additive dimensional splitting. We consider the relaxed scheme, i.e., $\varepsilon = 0$, and for the sake of simplicity, let us focus on the square domain $[a, b] \times [a, b] \subset \mathbb{R}^2$. Here we shall describe the generalization of the scheme defined by (4.3), (4.2), (4.4), and (4.5) to the case of two space dimensions.

Without loss of generality, we consider a uniform grid in $[a, b] \times [a, b] \subset \mathbb{R}^2$ such that $\vec{x}_{i,j} = (x_i, y_j) = (a - h/2, a - h/2) + i(h, 0) + j(0, h)$ for $i, j = 1, 2, \dots, n$ and $h = (b - a)/n$.

In the present case, u and w are one-dimensional variables, while $\vec{v} = (v_{(1)}, v_{(2)})$ is now a field in $\mathbb{R}^2$. First we observe that the relaxation steps (4.3) are easily generalized for $d > 1$. For the transport steps, one has to evolve in time the system

$$\frac{\partial}{\partial t}\begin{pmatrix} u \\ v_{(1)} \\ v_{(2)} \\ w \end{pmatrix} + \frac{\partial}{\partial x}\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & \varphi^2 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}\begin{pmatrix} u \\ v_{(1)} \\ v_{(2)} \\ w \end{pmatrix} + \frac{\partial}{\partial y}\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \varphi^2 \\ 0 & 0 & 1 & 0 \end{bmatrix}\begin{pmatrix} u \\ v_{(1)} \\ v_{(2)} \\ w \end{pmatrix} = 0. \tag{4.11}$$

The semidiscretization in space of the above equation can be written as

$$\frac{\partial z_{i,j}}{\partial t} = -\frac{1}{h}\left(F_{i+1/2,j} - F_{i-1/2,j}\right) - \frac{1}{h}\left(G_{i,j+1/2} - G_{i,j-1/2}\right),$$

where F and G are the numerical fluxes in the x and the y direction, respectively, and can be written as

$$F_{i+1/2,j} = F(z^+_{i+1/2,j}, z^-_{i+1/2,j}), \qquad G_{i,j+1/2} = G(z^+_{i,j+1/2}, z^-_{i,j+1/2}).$$

The fluxes in the two directions are computed separately. We illustrate the computation of the flux F along the x direction. We note that only the field $v_{(1)}$ appears in

the differential operator along this direction. The third component of the flux is zero, and thus we have three independent characteristic variables, namely,

$$U_{(1)} = \frac{\varphi w + v_1}{2\varphi}, \qquad V_{(1)} = \frac{\varphi w - v_1}{2\varphi}, \qquad W = u - w,$$

which correspond, respectively, to the eigenvalues $\varphi, -\varphi, 0$. At this point the numerical fluxes can be easily evaluated by upwinding. We proceed similarly for the numerical flux G that depends on the characteristic variables $U_{(2)}, V_{(2)}, W$.

Denote by $U^{\pm}_{i+1/2,j}$ the reconstructions of $U_{(1)}(\cdot, y_j)$ at the point $(x_i + h/2, y_j)$. This involves a reconstruction of the restriction of $U_{(1)}$ to the line $y = y_i$ and can be obtained with any of the one-dimensional techniques mentioned in section 4.1. Similarly, denote by $U^{\pm}_{i,j+1/2}$ the reconstructions of $U_{(2)}(x_i, \cdot)$ at the point $(x_i, y_j + h/2)$. Now, (4.4) and (4.5) become, respectively,

$$u^{(l)}_{i,j} = u^n_{i,j} - \lambda \sum_{m=1}^{l-1} \tilde{a}_{l,m} \Big[\varphi \left(U^{(m)-}_{i+1/2,j} - U^{(m)-}_{i-1/2,j} \right) - \varphi \left(V^{(m)+}_{i+1/2,j} - V^{(m)+}_{i-1/2,j} \right) \varphi \left(U^{(m)-}_{i,j+1/2} - U^{(m)-}_{i,j-1/2} \right) - \varphi \left(V^{(m)+}_{i,j+1/2} - V^{(m)+}_{i,j-1/2} \right) \Big] \tag{4.12}$$

and

$$u^{n+1}_{i,j} = u^n_{i,j} - \lambda \sum_{l=1}^{\nu} \varphi \tilde{b}_l \Big[\left(U^{(l)-}_{i+1/2,j} - V^{(l)+}_{i+1/2,j} \right) - \left(U^{(l)-}_{i-1/2,j} - V^{(l)+}_{i-1/2,j} \right) \left(U^{(l)-}_{i,j+1/2} - V^{(l)+}_{i,j+1/2} \right) - \left(U^{(l)-}_{i,j-1/2} - V^{(l)+}_{i,j-1/2} \right) \Big]. \tag{4.13}$$

The generalization to $d > 2$ and rectangular domains is now trivial. We stress once again that no two-dimensional reconstruction is used, but only d one-dimensional reconstructions are needed. Finally, boundary conditions can be implemented directionwise with the same techniques used in the one-dimensional case.

5. Numerical results. We performed several numerical tests of our relaxed schemes. First we tested convergence for a linear diffusion equation with periodic and Neumann boundary conditions for initial data giving rise to smooth solutions. Next, numerical tests were also performed on the porous media equation $u_t = (u^m)_{xx}$, $m = 2, 3$, in both one and two dimensions.

5.1. Linear diffusion. For the first test we considered the linear problem

$$\begin{cases} \dfrac{\partial u}{\partial t}(x,t) = \dfrac{\partial^2 u}{\partial x^2} u(x,t), & x \in [0,1], \\ u(x,0) = u_0(x), & x \in [0,1]. \end{cases}$$

First we used periodic boundary conditions with $u_0(x) = \cos(2\pi x)$, so that $u(x,t) = \cos(2\pi x) e^{-4\pi^2 t}$ is an exact solution. Then we used Neumann boundary conditions $u_x(0) = u_x(1) = 1$ with initial data $u_0(x) = x + cos(2\pi x)$, so that $u(x,t) = x + \cos(2\pi x) e^{-4\pi^2 t}$ is an exact solution.

We tested the numerical schemes defined by (4.2), (4.3), (4.4), and (4.5) with various degrees of accuracy for the spatial reconstructions and time-stepping operators. We used ENO spatial reconstructions of degrees from 2 to 6 and WENO reconstructions of degrees 3 and 5. The time-stepping procedures chosen are IMEX

TABLE 5.1

L^1 *norms of the error and convergence rates for the linear diffusion equation with periodic boundary conditions, with smooth initial data.*

	$N = 40$	$N = 80$	$N = 160$	$N = 320$	$N = 640$
ENO2, RK1	2.012e-03	5.6378e-04	1.0736e-04	1.5539e-05	2.5065e-06
ENO3, RK2	1.9066e-06	2.3057e-07	5.6115e-08	8.6904e-09	1.1905e-09
ENO4, RK2	7.7517e-06	5.7082e-07	3.3507e-08	1.4978e-09	7.0725e-11
ENO5, RK3	1.3864e-08	6.0259e-10	2.2121e-11	7.4454e-13	2.3803e-14
ENO6, RK3	1.5538e-08	8.5661e-10	1.446e-11	1.7111e-13	1.5311e-15
WENO3, RK2	1.9799e-03	5.1278e0-4	1.4332e-04	2.1488e-05	7.512e-08
WENO5, RK3	1.5892e-07	4.8069e-09	1.59e-10	5.2337e-12	1.6758e-13

	$N = 40$	$N = 80$	$N = 160$	$N = 320$	$N = 640$
ENO2, RK1	1.3973	1.8354	2.3926	2.7886	2.6322
ENO3, RK2	5.9501	3.0477	2.0388	2.6909	2.8678
ENO4, RK2	3.8987	3.7634	4.0905	4.4836	4.4045
ENO5, RK3	6.8124	4.524	4.7677	4.8929	4.9671
ENO6, RK3	5.9907	4.181	5.8885	6.401	6.8043
WENO3, RK2	0.56648	1.949	1.8391	2.7376	8.1601
WENO5, RK3	2.9595	5.0471	4.918	4.925	4.9649

TABLE 5.2

L^1 *norms of the error and convergence rates for the linear diffusion equation with Neumann boundary conditions, with smooth initial data.*

	$N = 40$	$N = 80$	$N = 160$	$N = 320$	$N = 640$
ENO2, RK1	2.1965e-03	5.7152e-04	1.4301e-04	2.32e-05	4.743e-06
ENO3, RK2	2.0621e-06	2.2641e-07	6.7935e-08	8.8255e-09	1.2339e-09
ENO4, RK2	8.1764e-06	5.4431e-07	3.6974e-08	1.3686e-09	8.335e-11
ENO5, RK3	1.5484e-07	4.4163e-09	1.2405e-10	3.7803e-12	1.1669e-13
WENO3, RK2	1.9092e-03	4.4225e-04	1.2914e-04	4.5037e-06	7.4526e-08
WENO5, RK3	2.5048e-07	4.9279e-09	1.4776e-10	4.7482e-12	1.4948e-13

	$N = 40$	$N = 80$	$N = 160$	$N = 320$	$N = 640$
ENO2, RK1	1.4361	1.9424	1.9987	2.624	2.2902
ENO3, RK2	6.1004	3.1871	1.7367	2.9444	2.8385
ENO4, RK2	3.9763	3.909	3.8798	4.7558	4.0373
ENO5, RK3	5.6626	5.1317	5.1539	5.0362	5.0178
WENO3, RK2	1.2624	2.11	1.7759	4.8417	5.9172
WENO5, RK3	4.9122	5.6676	5.0597	4.9597	4.9893

Runge–Kutta schemes of section 3 of accuracy chosen to match the accuracy of the spatial reconstruction. Since stability forces the parabolic restriction $\Delta t \leq Ch^2$, an IMEX scheme of order m was coupled with a spatial ENO/WENO reconstruction of accuracy p such that $p \leq 2m$, obtaining a scheme of order p.

We computed the numerical solution of the diffusion equation with final time $t = 0.05$ with $N = 40, 80, 160, 320, 640$ grid points and computed the L^1 norm of the difference between the numerical and the exact solution. The results are in Table 5.1 for the periodic boundary conditions and Table 5.2 for the Neumann boundary conditions. One can see that the expected convergence rates are reached, even if the combination of the WENO reconstruction of accuracy 3 and the IMEX scheme of second order reach the predicted error reduction only on very fine grids.

5.2. Porous media equation. On the porous media equation (1.1) with $p(u) = u^m$ we performed a test proposed in [18]. We took $m = 2, 3$ and initial data of class

TABLE 5.3
L^1 *norms of the error and convergence rates for the porous media equation periodic boundary conditions, with initial data of class* C^1.

	$N = 60$	$N = 180$	$N = 540$	$N = 1620$
ENO2, RK1	2.6365e-04	1.9898e-05	2.049e-06	2.076e-07
ENO3, RK2	1.9605e-05	6.0423e-07	2.4141e-08	8.9729e-10
ENO4, RK2	1.2127e-05	2.967e-07	9.9925e-09	3.5781e-10
ENO5, RK3	4.694e-06	1.719e-07	6.3248e-09	2.4447e-10
ENO6, RK3	4.1099e-06	1.4711e-07	5.3992e-09	2.0849e-10
WENO3, RK2	1.5871e-04	1.0448e-05	4.3463e-07	8.8767e-09
WENO5, RK3	7.5662e-06	4.6049e-07	7.4746e-09	2.7985e-10

	$N = 60$	$N = 180$	$N = 540$	$N = 1620$
ENO2, RK1	2.8243	2.352	2.0692	2.084
ENO3, RK2	5.1899	3.1672	2.931	2.9968
ENO4, RK2	5.6271	3.3774	3.0865	3.0307
ENO5, RK3	6.491	3.0103	3.006	2.9611
ENO6, RK3	6.612	3.0311	3.0083	2.962
WENO3, RK2	3.2863	2.4765	2.8942	3.5418
WENO5, RK3	6.0565	2.5479	3.7509	2.9902

C^1 as follows:

$$u(x,0) = \begin{cases} \cos^2(\pi x/2), & |x| \le 1, \\ 0, & |x| > 1. \end{cases} \tag{5.1}$$

The computational domain is $\{|x| \le 3\} \subset \mathbb{R}$, and the boundary conditions are periodic; the CFL constant is taken as $C = 0.25$.

Since the initial data have compact support and are Lipschitz continuous, the solution will be of compact support for every $t \ge 0$ but will develop a discontinuity in u_x at some finite time $\tau > 0$ (see [3]).

As was shown in [3], the solution with the initial condition we chose has a front that does not move for $t < 0.034$. We therefore chose a final time of the simulation $t_{\text{fin}} = 0.03$ to prevent the formation of the singularity of u_x from affecting the order of convergence. We used as a reference solution the one obtained numerically with $N = 4860$ grid points and computed the L^1 norms of the errors of the solutions with $N = 60, 180, 540, 1620$ grid points. The results are presented in Table 5.3.

First of all one verifies that the degree of regularity of the solution poses a limit on the order of convergence of the schemes: Therefore the schemes we tested perform at best as third-order schemes, as confirmed by the data in Table 5.3. Still, high-order schemes yield a smaller error on a given grid. This can be of practical importance in problems where one does not have the freedom of choosing the number of grid points, as in digital image analysis, where nonlinear degenerate diffusion equations are sometimes used as filters for contour enhancement (see [5]).

In Figure 5.1 we show the numerical solution for the porous media equation with $p(u) = u^2$ and $p(u) = u^3$, with the initial data (5.1) and $t \in [0, 2]$. It can be appreciated that a front (i.e., a discontinuity of $\frac{\partial u}{\partial x}$) develops at a finite time and then it travels at finite speed.

We present a numerical simulation for the two-dimensional porous media equation (1.1) with $p(u) = u^2$. We chose an initial data $u_0(x, y)$ given by two bumps with periodic boundary conditions on $[-10, 10] \times [-10, 10]$. The large domain ensures that the compact support of the solution is still contained in the computational domain

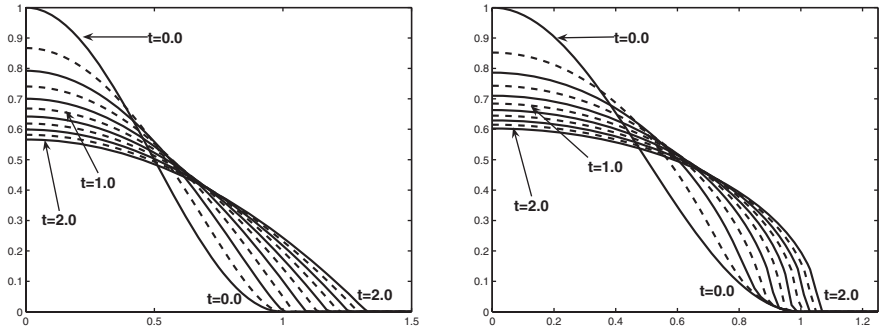


FIG. 5.1. *Snapshots of the numerical solutions for the porous media equation with* $p(u) = u^2$ *(left) and* $p(u) = u^3$ *(right). Initial data are chosen according to* (5.1)*, and the numerical solutions are represented at times* $t = 0, 0.2, \ldots, 2.0$*. The solutions are obtained with the spatial WENO reconstruction of order* 5 *and the RK*3 *time integrator.*

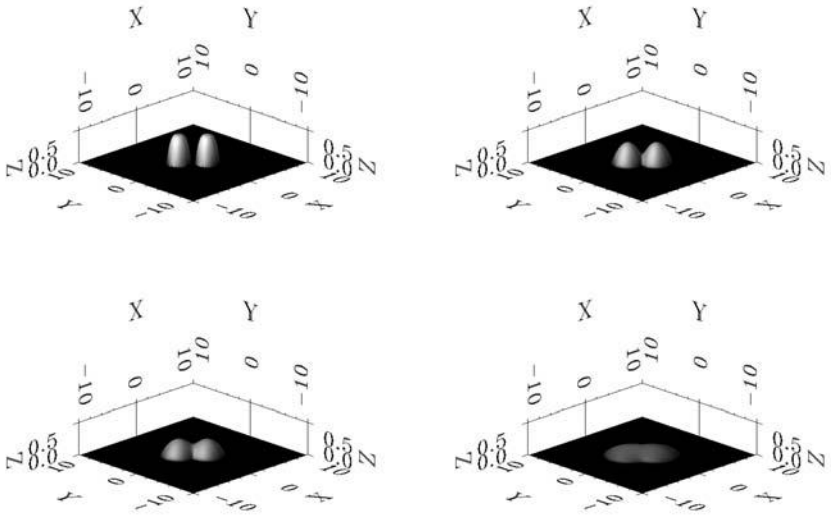


FIG. 5.2. *The numerical solution of the porous media equation on a square regular grid with compactly supported initial data. From top left to bottom right, we show the numerical solution at times* $t = 0, 0.5, 1.0, 4.0$*.*

at the final time of the calculation. The numerical approximation at different time levels is shown in Figure 5.2.

We can note that the symmetries of the initial data are preserved and the solution seems to be unaffected by the dimensional splitting of the two-dimensional scheme.

6. Conclusions. We have proposed and analyzed relaxed schemes for nonlinear degenerate parabolic equations.

We considered a relaxation system similar to the one used in [29, 27] but focused on the relaxed schemes that are obtained by taking the relaxation parameter $\varepsilon = 0$. By using suitable discretizations in space and time, namely, ENO/WENO nonoscillatory reconstructions for numerical fluxes and IMEX Runge–Kutta schemes for time

integration, we have obtained a class of high-order schemes. We proved a convergence theorem for the semidiscrete scheme using the nonlinear Chernoff formula; furthermore we obtained stability results for the fully discrete schemes. Our computational tests suggest that our schemes converge with the predicted rate and exhibit a high resolution of propagating fronts.

Finally, we point out that these schemes can be easily implemented on parallel computers. Some preliminary results and details are reported in [12]. In particular the schemes involve only linear matrix-vector operations, and the execution time scales linearly when increasing the number of processors.

Our numerical approach can be easily extended to more general problems. The case of degenerate reaction-diffusion equations will appear in [13]. The treatment of convection-diffusion equations requires the introduction of an additional equation to relax the convection terms. A preliminary study appears in [9], while some of these applications will appear in a forthcoming paper [11].

REFERENCES

[1] D. Aregba-Driollet, R. Natalini, and S. Tang, *Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems*, Math. Comp., 73 (2004), pp. 63–94.

[2] D. Aregba-Driollet and R. Natalini, *Discrete kinetic schemes for multidimensional systems of conservation laws*, SIAM J. Numer. Anal., 37 (2000), pp. 1973–2004.

[3] D. G. Aronson, *Regularity properties of flows through porous media: A counterexample*, SIAM J. Appl. Math., 19 (1970), pp. 299–307.

[4] U. Asher, S. Ruuth, and R. Spiteri, *Implicit-explicit Runge-Kutta methods for time dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[5] G. I. Barenblatt and J. L. Vázquez, *Nonlinear diffusion and image contour enhancement*, Interfaces Free Bound., 6 (2004), pp. 31–54.

[6] A. Berger, H. Brezis, and J. Rogers, *A numerical method for solving the problem* $u\_t - \Delta f(u) = 0$, RAIRO Numer. Anal., 13 (1979), pp. 297–312.

[7] F. Bouchut, F. R. Guarguaglini, and R. Natalini, *Diffusive BGK approximations for nonlinear multidimensional parabolic equations*, Indiana Univ. Math. J., 49 (2000), pp. 723–749.

[8] H. Brézis and A. Pazy, *Convergence and approximation of semigroups of nonlinear operators in Banach spaces*, J. Funct. Anal., 9 (1972), pp. 63–74.

[9] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *A comparison between relaxation and Kurganov-Tadmor schemes*, in Progress in Industrial Mathematics at ECMI 2006, L. L. Bonilla, M. Moscoso, G. Platero, and J. M. Vega, eds., Mathematics in Industry 12, Springer, 2007.

[10] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *Increasing efficiency through optimal RK time integration of diffusion equations*, in Proceedings of the HYP2006 Conference, 2006, to appear.

[11] F. Cavalli, G. Naldi, G. Puppo, and M. Semplice, *High Order Relaxation Approximation of Convection Diffusion Equations*, manuscript.

[12] F. Cavalli, G. Naldi, and M. Semplice, *Parallel algorithms for nonlinear diffusion by using relaxation approximation*, in ENUMATH 2005, A. Bermúdez de Castro, D. Gómez, P. Quintela, and P. Salgado, eds., Springer, Berlin, 2006, pp. 404–411.

[13] F. Cavalli and M. Semplice, *High order relaxed schemes for nonlinear reaction diffusion problems*, in Proceedings of the SIMAI 2006 Conference, 2006, submitted.

[14] M. Crandall and T. Liggett, *Generation of semi-groups of non linear transformations on general Banach spaces*, Amer. J. Math., 93 (1971), pp. 265–298.

[15] M. S. Espedal and K. H. Karlsen, *Numerical solution of reservoir flow models based on large time step operator splitting algorithms*, in Filtration in Porous Media and Industrial Application (Cetraro, 1998), Lecture Notes in Math. 1734, Springer, Berlin, 2000, pp. 9–77.

[16] S. Evje and K. H. Karlsen, *Viscous splitting approximation of mixed hyperbolic-parabolic convection-diffusion equations*, Numer. Math., 83 (1999), pp. 107–137.

[17] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[18] J. L. Graveleau and P. Jamet, *A finite difference approach to some degenerate nonlinear parabolic equations*, SIAM J. Appl. Math., 20 (1971), pp. 199–223.
[19] W. Jäger and J. Kačur, *Solution of porous medium type systems by linear approximation schemes*, Numer. Math., 60 (1991), pp. 407–427.
[20] S. Jin and C. D. Levermore, *Numerical schemes for hyperbolic conservation laws with stiff relaxation terms*, J. Comput. Phys., 126 (1996), pp. 449–467.
[21] S. Jin, L. Pareschi, and G. Toscani, *Diffusive relaxation schemes for multiscale discrete-velocity kinetic equations*, SIAM J. Numer. Anal., 35 (1998), pp. 2405–2439.
[22] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.
[23] J. Kačur, A. Handlovičová, and M. Kačurová, *Solution of nonlinear diffusion problems by linear approximation schemes*, SIAM J. Numer. Anal., 30 (1993), pp. 1703–1722.
[24] C. Lattanzio and R. Natalini, *Convergence of diffusive BGK approximations for nonlinear strongly parabolic systems*, Proc. Roy. Soc. Edinburgh Sect. A, 132 (2002), pp. 341–358.
[25] P. Lions and G. Toscani, *Diffusive limit for two-velocity Boltzmann kinetic models*, Rev. Mat. Iberoamericana, 13 (1997), pp. 473–513.
[26] E. Magenes, R. H. Nochetto, and C. Verdi, *Energy error estimates for a linear scheme to approximate nonlinear parabolic problems*, RAIRO Modél. Math. Anal. Numér., 21 (1987), pp. 655–678.
[27] G. Naldi, L. Pareschi, and G. Toscani, *Relaxation schemes for partial differential equations and applications to degenerate diffusion problems*, Surv. Math. Indust., 10 (2002), pp. 315–343.
[28] G. Naldi and L. Pareschi, *Numerical schemes for kinetic equations in diffusive regimes*, Appl. Math. Lett., 11 (1998), pp. 29–35.
[29] G. Naldi and L. Pareschi, *Numerical schemes for hyperbolic systems of conservation laws with stiff diffusive relaxation*, SIAM J. Numer. Anal., 37 (2000), pp. 1246–1270.
[30] R. H. Nochetto, A. Schmidt, and C. Verdi, *A posteriori error estimation and adaptivity for degenerate parabolic problems*, Math. Comp., 69 (2000), pp. 1–24.
[31] R. H. Nochetto and C. Verdi, *Approximation of degenerate parabolic problems using numerical integration*, SIAM J. Numer. Anal., 25 (1988), pp. 784–814.
[32] L. Pareschi and G. Russo, *Implicit-explicit Runge-Kutta schemes and applications to hyperbolic systems with relaxation*, J. Sci. Comput., 25 (2005), pp. 129–155.
[33] I. S. Pop and W. Yong, *A numerical approach to degenerate parabolic equations*, Numer. Math., 92 (2002), pp. 357–381.
[34] C. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes.* II, J. Comput. Phys., 83 (1989), pp. 32–78.
[35] C. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations (Cetraro, 1997), Lecture Notes in Math. 1697, Springer, Berlin, 1998, pp. 325–432.
[36] J. L. Vázquez, *An introduction to the mathematical theory of the porous medium equation*, in Shape Optimization and Free Boundaries (Montreal, PQ, 1990), NATO Sci. Ser. C Math. Phys. Sci. 380, Kluwer Academic Publishers, Dordrecht, 1992, pp. 347–389.
[37] J. Weickert, *Anisotropic Diffusion in Image Processing*, European Consortium for Mathematics in Industry, B. G. Teubner, Stuttgart, 1998.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2120–2140

NUMERICAL APPROXIMATION OF A QUASI-NEWTONIAN STOKES FLOW PROBLEM WITH DEFECTIVE BOUNDARY CONDITIONS*

VINCENT J. ERVIN† AND HYESUK LEE†

Abstract. In this article we study the numerical approximation of a quasi-Newtonian Stokes flow problem where only the flow rates are specified at the inflow and outflow boundaries. A variational formulation of the problem, using Lagrange multipliers to enforce the stated flow rates, is given. The existence and the uniqueness to the continuous, and discrete, variational formulations of the solution are shown. An error analysis for the numerical approximation is also given. Numerical computations are included which demonstrate the approximation scheme studied.

Key words. quasi-Newtonian Stokes flow, defective boundary condition, power law fluid

AMS subject classification. 65N30

DOI. 10.1137/060669012

1. Introduction.

In this paper we investigate the numerical approximation of a quasi-Newtonian Stokes flow problem with defective boundary conditions. Such problems arise in modeling viscoelastic fluid flow. Several examples are given in section 2.1. For well-posedness of a Newtonian fluid flow problem suitable boundary conditions are required to uniquely define the solution. Perhaps the simplest of these is to specify the velocity at each point on the boundary of the domain. Often what is assumed is that the flow is *fully developed* at the inflow and outflow boundaries, which justifies a parabolic flow profile at these boundaries. Typically a *no slip* (i.e., velocity $= \mathbf{0}$) is assumed along the other portions of the boundary of the domain. However, in many physical problems the assumption of fully developed flow at the inflow and outflow is either unreasonable or highly questionable. Usually what is known in physical fluid flow problems are the various inflow and outflow flow rates.

In [6] Formaggia et al. discuss the defective boundary condition problem for the time-dependent Navier–Stokes equation. They introduce a Lagrange multiplier approach to enforce flow constraints at the inflow and outflow portions of the boundary. For the steady-state Stokes problem, they show the existence and the uniqueness of the solution for flow rates imposed using the Lagrange multiplier formulation. Herein we extend this work to analyze a quasi-Newtonian Stokes flow problem subject to specified inflow and outflow flow rates. We establish the existence and the uniqueness of the solution for the continuous and discrete variational problems and present an error analysis for the numerical approximation.

Initially it is, perhaps, somewhat perplexing to note that, for the uniqueness of the solution to the variational problem for (i) the Dirichlet problem, we require that $\acute{d}$ (the dimension of the space) conditions be specified at each point on the boundary, whereas (ii) the defective boundary condition problem requires only that a single scalar be specified at inflow and outflow boundaries (and $\acute{d}$ conditions at

*Received by the editors September 4, 2006; accepted for publication (in revised form) May 14, 2007; published electronically September 28, 2007. This work was partially supported by the NSF under grant DMS-0410792.

http://www.siam.org/journals/sinum/45-5/66901.html

†Department of Mathematical Sciences, Clemson University, Clemson, SC 29634-0975 (vjervin@clemson.edu, hklee@clemson.edu).

other boundary points). This seeming anomaly is explained in Lemma 2.1 (see also [6, Proposition 2.1] and [12, p. 341]). Specifically, the variational formulation for the defective boundary condition problem implicity imposes that across each of the inflow and outflow boundaries the total stress normal to the boundaries is a constant, and the extra stress lying in the surface of the inflow and outflow boundaries is zero.

In [12] Heywood, Rannacher, and Turek also investigated the defective boundary condition problem for the time-dependent Navier–Stokes equations. They considered both the case of specified flow rates at the inflow and outflow boundaries and also the case of the mean specified pressure at the inflow and outflow boundaries. For the specified flow-rate problem, the formulation they considered (and proved the existence of a steady-state solution) involved the construction of suitable *flux-carrier* vector functions.

The numerical approximation of the quasi-Newtonian Stokes flow problem with homogeneous boundary conditions has been previously studied in several papers [2, 5, 8, 14, 17].

This paper is organized as follows. In sections 2.1 and 2.2 we describe the model problem, state our assumptions on the model, and introduce appropriate mathematical notation. We show in section 2.3 that the corresponding variational formulation, in which the flow rate boundary conditions are weakly imposed using Lagrange multipliers, is well-posed. A numerical approximation scheme is presented in section 3, and its solution shown to exist. A priori error estimates for the numerical approximation are derived in section 4. Numerical results are presented in section 5.

2. Mathematical model. Motivated by physical considerations we consider the numerical approximation of a three-field, quasi-Newtonian Stokes flow problem with fixed flow-rate boundary conditions.

2.1. Problem specification. Let Ω denote a bounded domain in $\mathbb{R}^{\acute{d}}$, $\acute{d} = 2$ or 3, whose boundary $\partial\Omega$ is decomposed into the union of Γ and several disjoint sections $S_1, S_2, \ldots, S_m$, $m \geq 2$. See Figure 2.1.

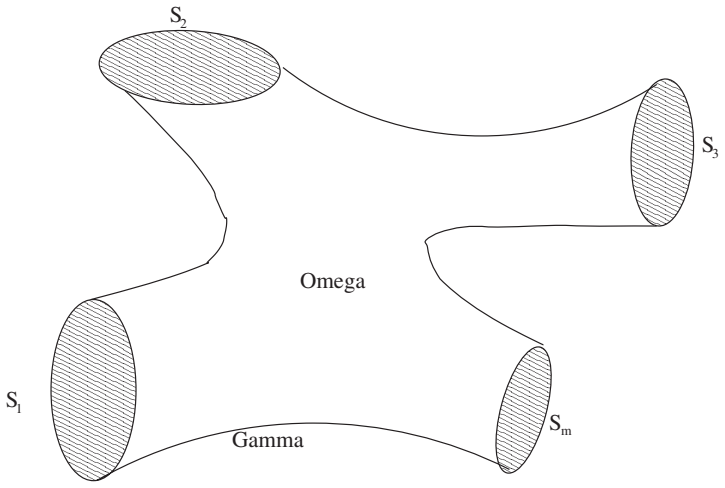


FIG. 2.1. *Illustration of the flow domain.*

We are interested in the numerical approximation of

$$\sigma = g(\mathbf{u}) \text{ in } \Omega, \tag{2.1}$$

$$-\nabla \cdot \sigma + \nabla p = \mathbf{f} \text{ in } \Omega, \tag{2.2}$$

$$\nabla \cdot \mathbf{u} = 0 \text{ in } \Omega, \tag{2.3}$$

$$\mathbf{u} = 0 \text{ on } \Gamma, \tag{2.4}$$

subject to the specified flow rates across the surfaces S_i:

$$\int_{S_i} \mathbf{u} \cdot \mathbf{n}\, ds = Q_i \text{ for } i = 1, \ldots, m. \tag{2.5}$$

We use $\mathbf{n}$ to denote the outward (from Ω) normal to the surface.

Because of the incompressibility condition (2.3) it follows that

$$\sum_{i=1}^{m} Q_i = 0. \tag{2.6}$$

Note that (2.1)–(2.5) can only determine the pressure p up to an arbitrary constant. Below, we fix this constant by requiring p to have mean value 0 over Ω.

The general form of the (algebraic) constitutive equation we assume in our analysis (see A1, A2, A3, in section 2.2) is motivated by the study of fluids having a *power law* constitutive equation, i.e.,

$$\sigma = \nu_0 |d(\mathbf{u})|^{r-2} d(\mathbf{u}), \quad \nu_0 > 0, \quad 1 < r < 2, \tag{2.7}$$

where σ denotes the extra stress tensor, $\mathbf{u}$ the fluid velocity, and $d(\mathbf{u}) := (\nabla \mathbf{u} + \nabla \mathbf{u}^T)/2$ the rate of deformation tensor.

The power law model has been used to model the viscosity of many polymeric solutions and melts over a considerable range of shear rates [11].

Other constitutive equations having a similar form to the power law model include [3, 14, 15]:

(i) *Ladyzhenskaya law* [13]

$$\sigma = \nu_0 + \nu_1 |\nabla \mathbf{u}|^{r-2} d(\mathbf{u}), \quad \nu_0 \geq 0, \nu_1 > 0, r > 1, \tag{2.8}$$

used in modeling fluids with large stresses; and

(ii) *Carreau law*

$$\sigma = \nu_0 \left(1 + |d(\mathbf{u})|^2\right)^{(r-2)/2} d(\mathbf{u}), \quad \nu_0 > 0, r \geq 1, \tag{2.9}$$

used in modeling viscoplastic flows and creeping flow of metals.

2.2. Notation/assumptions. We made the following assumptions regarding the constitutive equation (2.1) for the stress σ:

A1. $g(\mathbf{u})$ is (formally) uniquely invertible to obtain

$$d(\mathbf{u}) = \breve{g}(\sigma)\sigma, (\text{ or } \nabla \mathbf{u} = \breve{g}(\sigma)\sigma)$$

and the inverse is continuous. For $G(\sigma) := \breve{g}(\sigma)\sigma$,

A2.

$$(G(s) - G(t)) : (s - t) \geq c|s - t|^{r'} \qquad \forall s, t \in \mathbb{R}^{\acute{d} \times \acute{d}}, \tag{2.10}$$

A3.

$$|G(s) - G(t)| \leq M\left(|s| + |t|\right)^{r'-2}|s-t|, \qquad \forall s,t \in \mathbb{R}^{\acute{d}\times\acute{d}}. \tag{2.11}$$

For $r \in \mathbb{R}, r > 1$, we denote its unitary conjugate by r', satisfying $r^{-1} + r'^{-1} = 1$.

For problems of physical interest, $1 < r \leq 2$, e.g., shear thinning fluids. We therefore assume that $1 < r \leq 2$ and, consequently, $2 \leq r' < \infty$.

Properties A2 and A3 imply that $G(\cdot)$ is strongly monotone and Lipschitz continuous for bounded arguments [4].

We remark that differential constitutive models for viscoelastic fluids, such as the Oldroyd-B or Giesekus models, do not satisfy A1–A3.

Used in the analysis below are the following function spaces and norms:

$$T := \left(L^{r'}(\Omega)\right)^{\acute{d}\times\acute{d}}_{sym} = \left\{\tau = (\tau_{ij}); \tau_{ij} = \tau_{ji}; \tau_{ij} \in L^{r'}(\Omega); i,j = 1,\ldots,\acute{d}\right\},$$

with norm $\|\tau\|_T := (\int_\Omega |\tau|^{r'} d\Omega)^{1/r'}$, and

$$X := \left\{\mathbf{v} \in \left(W^{1,r}(\Omega)\right)^{\acute{d}} : \mathbf{v}|_\Gamma = \mathbf{0}\right\},$$

with $W^{k,p}(\Omega)$ denoting the usual Sobolev space notation. We take for the norm on X, $\|v\|_X := \left(\int_\Omega |d(\mathbf{v})|^r d\Omega\right)^{1/r}$, which is equivalent to the usual $\|\cdot\|_{W^{1,r}}$ norm by the Poincaré–Friedrichs lemma:

$$P := L^{r'}_0(\Omega) = \left\{q \in L^{r'}(\Omega) : \int_\Omega q d\Omega = 0\right\},$$

with norm $\|q\|_P := (\int_\Omega |q|^{r'} d\Omega)^{1/r'}$.

We use V_X to denote the subspace of X defined by

$$V_X := \left\{\mathbf{v} \in X : \int_\Omega q\nabla\cdot\mathbf{v} d\Omega + \sum_{i=1}^m \beta_i \int_{S_i} \mathbf{v}\cdot\mathbf{n} ds = 0 \ \forall (q,\beta) \in P \times \mathbb{R}^m\right\}$$

and let

$$V_T := \left\{\tau \in T : \int_\Omega \tau : d(\mathbf{v}) d\Omega = 0 \ \forall \mathbf{v} \in V_X\right\}.$$

For a Banach space Y, Y' denotes its dual space with associated norm $\|\cdot\|_{Y'}$. For σ, τ tensors and $\mathbf{u}$, $\mathbf{v}$ vectors, we use : and $\cdot$ to denote the scalar quantities $\sigma : \tau := \sum_{i=1}^{\acute{d}}\sum_{j=1}^{\acute{d}} \sigma_{ij}\tau_{ij}$ and $\mathbf{u}\cdot\mathbf{v} := \sum_{i=1}^{\acute{d}} \mathbf{u}_i\mathbf{v}_i$, respectively. We use $(\cdot,\cdot)$ to denote the L^2 inner product for functions (scalar, vector, or tensor) over Ω and $\langle\cdot,\cdot\rangle$ to denote the duality pairing between a function space and its dual space.

2.3. Lagrange multiplier approach. We consider the following variational formulation to (2.1)–(2.5): *Given* $\mathbf{f} \in X'$, $Q \in \mathbb{R}^m$, *determine* $(\sigma,\mathbf{u},p,\lambda) \in T \times X \times P \times \mathbb{R}^m$, *such that*

$$a(\sigma,\tau) - b(\tau,\mathbf{u}) = 0 \quad \forall \tau \in T, \tag{2.12}$$

$$b(\sigma,\mathbf{v}) - s(\mathbf{v},(p,\lambda)) = \langle \mathbf{f},\mathbf{v}\rangle \quad \forall \mathbf{v} \in X, \tag{2.13}$$

$$s(\mathbf{u},(q,\beta)) = \sum_{i=1}^m Q_i\beta_i \quad \forall (q,\beta) \in P \times \mathbb{R}^m, \tag{2.14}$$

where

$$a(\sigma,\tau) := \int_\Omega \breve{g}(\sigma)\sigma : \tau d\Omega, \tag{2.15}$$

$$b(\tau,\mathbf{u}) := \int_\Omega \tau : d(\mathbf{u})d\Omega, \tag{2.16}$$

$$s(\mathbf{v},(p,\lambda)) := \int_\Omega p\nabla\cdot\mathbf{v}d\Omega + \sum_{i=1}^m \lambda_i \int_{S_i} \mathbf{v}\cdot\mathbf{n}ds. \tag{2.17}$$

The *Lagrange multiplier* $\lambda \in \mathbb{R}^m$ is introduced to include the flow constraints (2.5) in the variational formulation, see [1, 6, 10].

Equivalence of the differential equations and variational formulations. The variational formulation is obtained by multiplying the differential equations by sufficiently smooth functions, integrating over the domain, and, where appropriate, applying Green's theorem. The constraint equations are imposed weakly using Lagrange multipliers. For a smooth solution the steps used in deriving the variational equations can be reversed to show that (2.1)–(2.5) are satisfied. In addition we have that a smooth solution of (2.12)–(2.14) satisfies the following additional boundary conditions (see [6]).

For $\mathbf{n}$ the outward normal on S_i, express the extra stress vector on S_i, $\sigma\cdot\mathbf{n}$, as

$$\sigma\cdot\mathbf{n} = s_n\mathbf{n} + \mathbf{s}_T,$$

where $s_n = (\sigma\cdot\mathbf{n})\cdot\mathbf{n}$ and $\mathbf{s}_T = \sigma\cdot\mathbf{n} - s_n\mathbf{n}$. The scalar s_n represents the magnitude of the extra stress in the outward normal direction to S_i, and $\mathbf{s}_T$ is the component of the extra stress vector which lies in the plane of S_i.

LEMMA 2.1. *Any smooth solution of* (2.12)–(2.14) *satisfies the additional boundary conditions*

$$-p + s_n|_{S_i} = \lambda_i \text{ and } \mathbf{s}_T|_{S_i} = \mathbf{0}, \quad i = 1,\dots,m. \tag{2.18}$$

Proof. The proof follows as in [6]. □

Remark. The equations (2.1)–(2.5) do not uniquely define a solution but rather a set of solutions. The variational formulation (2.12)–(2.14) chooses a solution from the solution set. Specifically, (2.12)–(2.14) chooses *the solution* which satisfies (2.18). A different variational formulation may result in a different selection for *the solution* from the solution set. (See, for example, [6].)

Unique solvability of (2.12)–(2.14). There are two main steps in showing that (2.12)–(2.14) is uniquely solvable. Step 1 involves showing that the (2.12)–(2.14) can be reduced to an equivalent problem involving only σ. Step 2 demonstrates that the stress is uniquely solvable. Used in step 1 is the following lemma.

LEMMA 2.2 (see [9, Remark 4.2, p. 61]). *Let* $(X,\|\cdot\|_X)$ *and* $(M,\|\cdot\|_M)$ *be two reflexive Banach spaces. Let* $(X',\|\cdot\|_{X'})$ *and* $(M',\|\cdot\|_{M'})$ *be their corresponding dual spaces. Let* $B : X \to M'$ *be a linear continuous operator and* $B' : M'' \to X'$ *the dual operator of* B. *Let* $V = ker(B)$ *be the kernel of* B; *we denote by* $V^o \subset X'$ *the polar set of* $V : V^o = \{x' \in X', \langle x', v\rangle = 0\ \forall v \in V\}$ *and* $\dot{B} : X/V \to M'$ *the quotient operator associated with* B. *The following three properties are equivalent:*

(i) $\exists c > 0$ *such that*

$$\inf_{q\in M}\sup_{v\in X}\frac{\langle Bv,q\rangle}{\|q\|_M\|v\|_X} \geq c;$$

(ii) *B' is an isomorphism from M'' onto V^o and*

$$\|B'q\|_{X'} \geq C_B\|q\|_{M''} \ \forall q \in M'';$$

(iii) *$\dot{B}$ is an isomorphism from X/V onto M' and*

$$\|\dot{B}\dot{v}\|_{M'} \geq C_B\|\dot{v}\|_{X/V} \ \forall \dot{v} \in X/V.$$

As the first part of step 1, we show that (p, λ) can be eliminated from (2.12)–(2.14). To do this we use the following *inf-sup* condition. (See also [18].)

LEMMA 2.3. *There exists $C_{PRX} > 0$ such that*

$$\inf_{(q,\beta)\in P\times\mathbb{R}^m} \sup_{\mathbf{u}\in X} \frac{s(\mathbf{u},(q,\beta))}{\|\mathbf{u}\|_X\|(q,\beta)\|_{P\times\mathbb{R}^m}} \geq C_{PRX}, \tag{2.19}$$

where $\|(q,\beta)\|_{P\times\mathbb{R}^m} := \|q\|_P + \|\beta\|_{\mathbb{R}^m}$.

Proof. Fix $(q,\beta) \in P \times \mathbb{R}^m$, and let

$$\hat{q} = \frac{|q|^{r'/r-1}q}{\|q\|_P^{r'-1}}, \quad \hat{\beta} = \frac{\beta}{\|\beta\|_{\mathbb{R}^m}}. \tag{2.20}$$

Note that $(q,\hat{q}) = \|q\|_P$, $\|\hat{q}\|_{P'} = 1$, $\hat{\beta}\cdot\beta = \|\beta\|_{\mathbb{R}^m}$, and $\|\hat{\beta}\|_{\mathbb{R}^m} = 1$.

Next, we introduce $\delta \in \mathbb{R}$ and $h \in W^{1-1/r,r}(\partial\Omega)$, a piecewise constant function, defined by

$$h = \begin{cases} \hat{\beta}_i/\mathtt{meas}(S_i) & \text{on} \quad S_i, i = 1,\dots,m, \\ 0 & \text{on} \quad \Gamma, \end{cases} \tag{2.21}$$

$$\delta = \left(\int_{\partial\Omega} h ds - \int_\Omega \hat{q} d\Omega\right)/\mathtt{meas}(\Omega). \tag{2.22}$$

From [7, p. 127], given $f \in L^r(\Omega), \mathbf{a} \in W^{1-1/r,r}(\partial\Omega), 1 < r < \infty$, satisfying

$$\int_\Omega f d\Omega = \int_{\partial\Omega} \mathbf{a}\cdot\mathbf{n} ds, \tag{2.23}$$

there exists $\mathbf{v} \in W^{1,r}(\Omega)$ such that

$$\nabla\cdot\mathbf{v} = f \text{ in } \Omega, \tag{2.24}$$

$$\mathbf{v} = \mathbf{a} \text{ on } \partial\Omega, \tag{2.25}$$

$$\text{with } \|\mathbf{v}\|_{W^{1,r}(\Omega)} \leq C\left(\|f\|_{L^r(\Omega)} + \|\mathbf{a}\|_{W^{1-1/r,r}(\partial\Omega)}\right). \tag{2.26}$$

Let $f = \hat{q} + \delta$, and, for $\{\mathbf{n}, \mathbf{t}_i, i = 1,\dots \acute{d}-1\}$ denoting an orthonormal system on $\partial\Omega$, let $\mathbf{a}$ be defined by

$$\begin{cases} \mathbf{a}\cdot\mathbf{n} &= h, \\ \mathbf{a}\cdot\mathbf{t}_i &= 0,\ i = 1,\dots \acute{d}-1. \end{cases}$$

Remark. The choice of the constant δ guarantees that the compatibility condition $\int_\Omega f d\Omega = \int_{\partial\Omega} \mathbf{a}\cdot\mathbf{n} ds$ is satisfied.

We have that

$$\|\hat{q}\|_{W^{0,r}(\Omega)} = 1 \text{ (by construction)}, \tag{2.27}$$

$$\|\mathbf{a}\|_{W^{1-1/r,r}(\partial\Omega)} \leq C_1\|\hat{\beta}\|_{\mathbb{R}^m} = C_1 \tag{2.28}$$

(by the equivalence of finite dimensional norms).

Also,

$$\int_\Omega \hat{q} d\Omega \leq \|\hat{q}\|_{P'} \|1\|_P = C_2, \tag{2.29}$$

$$\int_{\partial\Omega} h ds \leq \|\hat{\beta}\|_{\mathbb{R}^m} \|\mathbf{1}\|_{\mathbb{R}^m} = C_3, \tag{2.30}$$

and thus $\|\delta\|_{W^{0,r}(\Omega)} \leq C_4$.

With $\mathbf{u}$ denoting the solution of (2.24)–(2.26), we have that $\mathbf{u} \in X$ and satisfies

$$\|\mathbf{u}\|_{W^{1,r}(\Omega)} \leq C(1 + C_4 + C_1) \leq C_5. \tag{2.31}$$

Hence,

$$\begin{aligned} s(\mathbf{u}, (q, \beta)) &= (\nabla \cdot \mathbf{u}, q) + \sum_{i=1}^m \beta_i \int_{S_i} \mathbf{u} \cdot \mathbf{n} ds \\ &= (\hat{q} + \delta, q) + \hat{\beta} \cdot \beta \\ &= \|q\|_P + \|\beta\|_{\mathbb{R}^m} \\ &= \|(q, \beta)\|_{P \times \mathbb{R}^m}, \end{aligned}$$

as $(\delta, q) = 0$ for $q \in P (= L_0^{r'}(\Omega))$. Thus,

$$\sup_{\mathbf{u} \in X} \frac{s(\mathbf{u}, (q, \beta))}{\|(q, \beta)\|_{P \times \mathbb{R}^m} \|\mathbf{u}\|_{W^{1,r}(\Omega)}} \geq \frac{1}{C_5},$$

from which (2.19) directly follows. ∎

We now state and prove the existence and the uniqueness of the solution to (2.12)–(2.14).

THEOREM 2.1. *Given $\mathbf{f} \in X'$ and $Q \in \mathbb{R}^m$, there exists a unique $(\sigma, \mathbf{u}, p, \lambda) \in T \times X \times P \times \mathbb{R}^m$ satisfying* (2.12)–(2.14).

Proof. From Lemmas 2.3 and 2.2(i), (iii), with the associations $X = X$, $M = P \times \mathbb{R}^m$, $B : X \to (P \times \mathbb{R}^m)'$ defined by

$$B(\mathbf{v}) := s(\mathbf{v}, (\cdot, \cdot)),$$

$V = ker(B)$, we have that there exists $\dot{\mathbf{u}} \in X/V$ such that

$$s(\dot{\mathbf{u}}, (q, \beta)) = \sum_{i=1}^m Q_i \beta_i \quad \forall (q, \beta) \in P \times \mathbb{R}^m,$$

with $\|\dot{\mathbf{u}}\|_{X/V} \leq 1/C_s \|Q\|_{\mathbb{R}^m}$.

Note: $\|\dot{\mathbf{u}}\|_{X/V} := \inf_{\mathbf{v} \in \dot{\mathbf{u}}} \|\mathbf{v}\|_X$.

As the cosets in X/V are closed, we can choose $\mathbf{u}_s \in \dot{\mathbf{u}}$ such that

$$\|\mathbf{u}_s\|_X = \|\dot{\mathbf{u}}\|_{X/V} \leq 1/C_s \|Q\|_{\mathbb{R}^m}. \tag{2.32}$$

Let $\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{u}_s$. Then, solving (2.12)–(2.14) is equivalent to: *Find $\sigma \in X, \tilde{\mathbf{u}} \in V_X$, such that*

$$a(\sigma, \tau) - b(\tau, \tilde{\mathbf{u}}) = b(\tau, \mathbf{u}_s) \quad \forall \tau \in T, \tag{2.33}$$

$$b(\sigma, \mathbf{v}) = \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall \mathbf{v} \in V_X. \tag{2.34}$$

Now note that, for $\mathbf{v} \in X$ and $\tau = |d(\mathbf{v})|^{r-2}d(\mathbf{v}) \in T$, $\|\tau\|_T = \|\mathbf{v}\|_X^{r/r'}$ and

$$\frac{b(\tau, \mathbf{v})}{\|\tau\|_T} = \frac{\|\mathbf{v}\|_X^r}{\|\mathbf{v}\|_X^{r/r'}} = \|\mathbf{v}\|_X.$$

Thus

$$\inf_{\mathbf{v}\in X} \sup_{\tau \in T} \frac{b(\tau, \mathbf{v})}{\|\tau\|_T \|\mathbf{v}\|_X} \geq 1, \tag{2.35}$$

i.e., $b(\tau, \mathbf{v})$ satisfies an inf-sup condition over $X \times T$.

As above, there exists $\sigma_b \in T$ such that

$$\begin{aligned} b(\sigma_b, \mathbf{v}) &= \langle \mathbf{f}, \mathbf{v} \rangle \quad \forall \mathbf{v} \in X, \\ \text{with } \|\sigma_b\|_T &\leq \frac{1}{C_b}\|\mathbf{f}\|_{X'}. \end{aligned} \tag{2.36}$$

Let $\sigma = \tilde{\sigma} + \sigma_b$. Then, solving (2.33), (2.34) is equivalent to: *Find $\tilde{\sigma} \in V_T$, such that*

$$a(\tilde{\sigma} + \sigma_b, \tau) = b(\tau, \mathbf{u}_s) \quad \forall \tau \in V_T. \tag{2.37}$$

From assumptions A2 and A3 we have that $G(\tau) : V_T \to V_T'$ is a continuous, coercive, strictly monotone operator on a real, separable, reflexive Banach space [16]. Hence, there exists a unique $\tilde{\sigma} \in V_T$ satisfying (2.37). This then also uniquely determines $\sigma \in T$.

The inf-sup condition (2.35), together with (2.33), uniquely determines $\tilde{\mathbf{u}} \in V_X$ and hence also $\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{u}_s \in X$.

Finally, the inf-sup condition (2.19) and the equation (2.13) uniquely determine $p \in P$ and $\lambda \in \mathbb{R}^m$. ∎

We now establish a bound for $\|\sigma\|_T$, which we use below in section 4 in deriving a priori estimates for the numerical approximation. Estimates for $\mathbf{u}$, p, and λ can also be derived.

COROLLARY 2.1. *For $\sigma \in T$ satisfying* (2.12)–(2.14) *we have that there exists $C > 0$ such that*

$$\|\sigma\|_T \leq C(\|\mathbf{f}\|_{X'} + \|Q\|_{\mathbb{R}^m}^{r/r'}). \tag{2.38}$$

Proof. From (2.37), with the choice $\tau = \tilde{\sigma}$, we have

$$\begin{aligned} a(\tilde{\sigma} + \sigma_b, \tilde{\sigma}) &= b(\tilde{\sigma}, \mathbf{u}_s) \\ &\leq 2^{-r'}\epsilon\|\tilde{\sigma}\|_T^{r'} + C\|\mathbf{u}_s\|_X^r \\ &\leq 2^{-r'}\epsilon\left(\|\tilde{\sigma} + \sigma_b\|_T + \|\sigma_b\|_T\right)^{r'} + C\|\mathbf{u}_s\|_X^r \\ &\leq \epsilon\|\tilde{\sigma} + \sigma_b\|_T^{r'} + \epsilon\|\sigma_b\|_T^{r'} + C\|\mathbf{u}_s\|_X^r. \end{aligned} \tag{2.39}$$

Using assumption (2.10),

$$\begin{aligned} a(\tilde{\sigma} + \sigma_b, \tilde{\sigma}) &= \int_\Omega \breve{g}(\tilde{\sigma} + \sigma_b)(\tilde{\sigma} + \sigma_b) : \tilde{\sigma} d\Omega \\ &= \int_\Omega \breve{g}(\tilde{\sigma} + \sigma_b)(\tilde{\sigma} + \sigma_b) : (\tilde{\sigma} + \sigma_b) - \int_\Omega \breve{g}(\tilde{\sigma} + \sigma_b)(\tilde{\sigma} + \sigma_b) : \sigma_b d\Omega \\ &\geq \int_\Omega c|\tilde{\sigma} + \sigma_b|^{r'} d\Omega - \|G(\tilde{\sigma} + \sigma_b)\|_{L^r}\|\sigma_b\|_T \\ &\geq c\|\tilde{\sigma} + \sigma_b\|_T^{r'} - \frac{c}{2M^r}\|G(\tilde{\sigma} + \sigma_b)\|_{L^r}^r - C\|\sigma_b\|_T^{r'}. \end{aligned} \tag{2.40}$$

We use (2.11) to estimate the second term on the right-hand side (RHS) of (2.40):

$$\begin{aligned}\|G(\tilde{\sigma}+\sigma_b)\|_{L^r}^r = \int_\Omega |G(\tilde{\sigma}+\sigma_b)|^r d\Omega &\leq M^r \int_\Omega \left((|\tilde{\sigma}+\sigma_b|+0)^{r'-2} |\tilde{\sigma}+\sigma_b - 0| \right)^r d\Omega \\ &= M^r \int_\Omega |\tilde{\sigma}+\sigma_b|^{(r'-1)r} d\Omega \\ &= M^r \int_\Omega |\tilde{\sigma}+\sigma_b|^{r'} d\Omega \\ &= M^r \|\tilde{\sigma}+\sigma_b\|_T^{r'}. \end{aligned} \tag{2.41}$$

Combining (2.39)–(2.41) we have that

$$\left(\frac{c}{2}-\epsilon\right) \|\tilde{\sigma}+\sigma_b\|_T^{r'} \leq C \left(\|\sigma_b\|_T^{r'} + \|\mathbf{u}_s\|_X^r \right).$$

As $\sigma = \tilde{\sigma} + \sigma_b$, and using the estimates (2.32) and (2.36), we obtain (2.38). ∎

3. Discrete approximation. We now describe the discrete approximation problem corresponding to (2.12)–(2.14) and show that the problem is well-defined. Analogous to the continuous problem the existence and the uniqueness for the discrete problem rely on the approximating spaces satisfying suitable inf-sup conditions.

We begin by describing the finite element approximation framework used in the analysis.

Let $\Omega \subset \mathbb{R}^{\acute{d}} (\acute{d} = 2, 3)$ be a polygonal domain, and let $\mathcal{T}_h$ be a triangulation of Ω made of triangles (in $\mathbb{R}^2$) or tetrahedrals (in $\mathbb{R}^3$). Thus, the computational domain is defined by

$$\Omega = \cup K; \quad K \in T_h.$$

We assume that there exist constants c_1, c_2 such that

$$c_1 h \leq h_K \leq c_2 \rho_K,$$

where h_K is the diameter of triangle (tetrahedral) K, ρ_K is the diameter of the greatest ball (sphere) included in K, and $h = \max_{K \in \mathcal{T}_h} h_K$. Let $P_k(A)$ denote the space of polynomials on A of degree no greater than k. Then we define the finite element spaces as follows:

$$T_h := \left\{ \tau \in T \cap C(\bar{\Omega})^{2\times 2} : \tau|_K \in P_l(K) \; \forall K \in \mathcal{T}_h \right\}, \tag{3.1}$$

$$X_h := \left\{ \mathbf{v} \in X \cap C(\bar{\Omega})^2 : \mathbf{v}|_K \in P_k(K) \; \forall K \in \mathcal{T}_h \right\}, \tag{3.2}$$

$$P_h := \left\{ q \in P \cap C(\bar{\Omega}) : q|_K \in P_n(K) \; \forall K \in \mathcal{T}_h \right\}. \tag{3.3}$$

We assume that the velocity-stress and the pressure-velocity spaces satisfy the following (typical) discrete inf-sup condition: *There exist constants C_{XTh}, $C_{PXh} > 0$ such that*

$$\inf_{\mathbf{v} \in X_h} \sup_{\tau \in T_h} \frac{b(\tau, \mathbf{v})}{\|\tau\|_T \|\mathbf{v}\|_X} \geq C_{XTh}, \tag{3.4}$$

$$\inf_{q \in P_h} \sup_{\mathbf{v} \in X_h} \frac{\int_\Omega q \nabla \cdot \mathbf{v} dA}{\|q\|_P \|\mathbf{v}\|_X} \geq C_{PXh}. \tag{3.5}$$

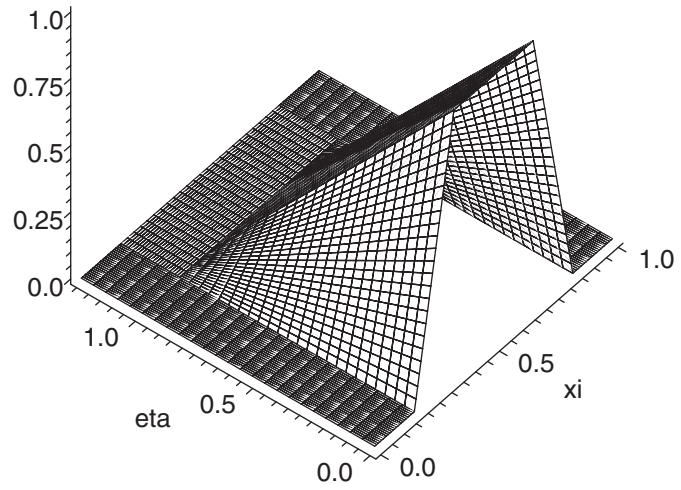


FIG. 3.1. *Plot of* g_i/β_i.

Discrete approximation problem. Given $\mathbf{f} \in X'$, and $Q \in \mathbb{R}^m$, determine $(\sigma_h, \mathbf{u}_h, p_h, \lambda_h) \in T_h \times X_h \times P_h \times \mathbb{R}^m$ such that

$$a(\sigma_h, \tau_h) - b(\tau_h, \mathbf{u}_h) = 0 \quad \forall \tau_h \in T_h, \tag{3.6}$$

$$b(\sigma_h, \mathbf{v}_h) - s(\mathbf{v}_h, (p_h, \lambda_h)) = \langle \mathbf{f}, \mathbf{v}_h \rangle \quad \forall \mathbf{v}_h \in X_h, \tag{3.7}$$

$$s(\mathbf{u}_h, (q_h, \beta_h)) = \sum_{i=1}^{m} Q_i \beta_i \quad \forall (q_h, \beta_h) \in P_h \times \mathbb{R}^m. \tag{3.8}$$

For the analysis a more general inf-sup condition than that given in (3.5) is needed. This is established using the following two lemmas. (See also [18].)

LEMMA 3.1. *There exists* $C_{RXh} > 0$ *such that*

$$\inf_{\beta \in \mathbb{R}^m} \sup_{\mathbf{v}_h \in X_h} \frac{\sum_{i=1}^{m} \beta_i \int_{S_i} \mathbf{v}_h \cdot \mathbf{n} ds}{\|\mathbf{v}_h\|_X \|\beta\|_{\mathbb{R}^m}} \geq C_{RXh}. \tag{3.9}$$

Outline of proof. From inspection of (3.9) we see that we would like to choose $\mathbf{v}_h \in X_h$ such that $\mathbf{v}_h \cdot \mathbf{n} = \beta_i$ on each S_i, and $\|\mathbf{v}_h\|_X \leq c\|\beta\|_{\mathbb{R}^m}$. This is done by constructing a suitable $\mathbf{v}_{h,i}$, with $\mathbf{v}_{h,i}|_{S_j} = \mathbf{0}, j \neq i$, and then letting $\mathbf{v}_h = \sum_i \mathbf{v}_{h,i}$.

We focus our attention on a single S_i. We will assume that on S_i, $\mathbf{n}(x) \cdot \mathbf{n}(y) \geq c > 0$ at all points $x, y \in S_i$ for which $\mathbf{n}$ is defined. (That is, on S_i the normal $\mathbf{n}$ does not vary by more than 90 degrees. If the normal does vary by more than 90 degrees, consider the surface as two surfaces.)

For ease of explanation, consider S_i as a straight line segment from $(0,0)$ to $(|S_i|, 0)$. Fix a *depth* d_i such that the rectangle $\mathcal{R}$ with vertices $(|S_i|/6, 0)$, $(5|S_i|/6, 0)$, $(5|S_i|/6, d_i)$, $(|S_i|/6, d_i)$ lies in Ω. Introduce the labeling of the following points: $A := (|S_i|/6, 0)$, $B := (5|S_i|/6, 0)$, $C := (5|S_i|/6, d_i)$, $D := (|S_i|/6, d_i)$, $E := (|S_i|/3, 0)$, $F := (2|S_i|/3, 0)$, $G := (2|S_i|/3, d_i)$, and $H := (|S_i|/3, d_i)$.

Let $\tilde{\mathbf{n}} = \mathbf{n}|_{(|S_i|/2,0)}$ and g_i be the continuous, piecewise bilinear, function defined by $g_i|_{E,F} = \beta_i$, and $g_i|_{A,B,C,D,G,H} = 0$. (See Figure 3.1. In Figure 3.1, xi $= x/|S_i|$, and eta $= y/d$.).

We define the function $\tilde{\mathbf{v}}_i$ as $\tilde{\mathbf{v}}_i|_{\Omega\setminus\mathcal{R}} = \mathbf{0}$, and $\tilde{\mathbf{v}}_i|_{\mathcal{R}} = g_i\tilde{\mathbf{n}}$. Then

$$\beta_i \int_{S_i} \tilde{\mathbf{v}}_i \cdot \mathbf{n} ds = \beta_i \int_E^F (\beta_i \tilde{\mathbf{n}}) \cdot \mathbf{n} ds \geq c_i \beta_i^2 |S_i|/3.$$

Also,

$$\begin{aligned}\|\tilde{\mathbf{v}}_i\|_X &= \left(\int_{\mathcal{R}} |\tilde{\mathbf{v}}_i|^r dA + \int_{\mathcal{R}} |\nabla \tilde{\mathbf{v}}_i|^r dA \right)^{1/r} \\ &= |\beta_i| \left((r+2) d_i |S_i|/(3(r+1)^2) + 6^{r-2} 2 d_i/(|S_i|^{r-1}(r+1)) \right. \\ &\qquad \left. + (6r+7)|S_i|/(18(r+1)d_i^{r-1}) \right)^{1/r}.\end{aligned}$$

Now there exists h_0 such that for all $h \leq h_0$ there exists $\mathbf{v}_{h,i} \in \mathcal{T}_h$ such that $\|\tilde{\mathbf{v}}_i - \mathbf{v}_{h,i}\|_\infty \leq c_i\beta_i/6$ and $\|\tilde{\mathbf{v}}_i - \mathbf{v}_{h,i}\|_X \leq |\beta_i|$. Then

$$\begin{aligned}\frac{\sum_{i=1}^m \beta_i \int_{S_i} \mathbf{v}_h \cdot \mathbf{n} ds}{\|\mathbf{v}_h\|_X} \geq \frac{\sum_{i=1}^m \beta_i \int_{S_i} \mathbf{v}_{h,i} \cdot \mathbf{n} ds}{\sum_{i=1}^m \|\mathbf{v}_{h,i}\|_X} &\geq \frac{\sum_{i=1}^m \left(\beta_i \int_{S_i} \tilde{\mathbf{v}}_i \cdot \mathbf{n} ds - c_i \beta_i^2 |S_i|/6 \right)}{\sum_{i=1}^m \left(\|\tilde{\mathbf{v}}_i\|_X + \|\tilde{\mathbf{v}}_i - \mathbf{v}_{h,i}\|_X \right)} \\ &\geq \frac{\sum_{i=1}^m c_i \beta_i^2 |S_i|/6}{\sum_{i=1}^m \hat{c}_i |\beta_i|} \geq C\|\beta\|,\end{aligned}$$

from which (3.9) then follows. □

LEMMA 3.2. *For h sufficiently small, there exists $C_{PRXh} > 0$ such that*

$$\inf_{(q_h,\beta)\in P_h\times\mathbb{R}^m} \sup_{\mathbf{v}_h \in X_h} \frac{s(\mathbf{v}_h, (q_h,\beta))}{\|\mathbf{v}_h\|_X \|(q,\beta)\|_{P\times\mathbb{R}^m}} \geq C_{PRXh}. \tag{3.10}$$

Proof. Let $(p_h, \beta) \in P_h \times \mathbb{R}^m$. From Lemma 3.1, there exists $\hat{\mathbf{u}}_h \in X_h$ such that

$$\|\hat{\mathbf{u}}_h\|_X = \|\beta\|_{\mathbb{R}^m} \quad \text{and} \quad \frac{\sum_{i=1}^m \beta_i \int_{S_i} \hat{\mathbf{u}}_h \cdot \mathbf{n} ds}{\|\hat{\mathbf{u}}_h\|_X} \geq c_1 \|\beta\|_{\mathbb{R}^m}. \tag{3.11}$$

Let $X_h^0 := \{\mathbf{v}_h \in X_h : \mathbf{v}_h|_{\partial\Omega} = \mathbf{0}\}$, and consider the (discrete) power law problem: *Determine $\tilde{\mathbf{u}}_h \in X_h^0, \tilde{p}_h \in P_h$ such that*

$$(|d(\tilde{\mathbf{u}}_h)|^{r-2} d(\tilde{\mathbf{u}}_h), d(\mathbf{v})) - (\tilde{p}_h, \nabla \cdot \mathbf{v}) = 0 \quad \forall \mathbf{v} \in X_h^0, \tag{3.12}$$

$$(q, \nabla \cdot \tilde{\mathbf{u}}_h) = (q, \|p_h\|_P^{1-r'/r} |p_h|^{r'/r-1} p_h - \nabla \cdot \hat{\mathbf{u}}_h) \quad \forall q \in P_h. \tag{3.13}$$

Note that $\|p_h\|_P^{1-r'/r} |p_h|^{r'/r-1} p_h - \nabla \cdot \hat{\mathbf{u}}_h \in L^r(\Omega)$.

The existence and the uniqueness of $\tilde{\mathbf{u}}_h \in X_h^0, \tilde{p}_h \in P_h$ satisfying (3.12), (3.13) follow analogous to the proof of Theorem 2.1. (See also [9, 2]).

From (3.12), (3.13) with the choices $\mathbf{v} = \tilde{\mathbf{u}}_h$ and $q = \tilde{p}_h$,

$$\begin{aligned}\|\tilde{\mathbf{u}}_h\|_X^r = (|d(\tilde{\mathbf{u}}_h)|^{r-2} d(\tilde{\mathbf{u}}_h), d(\tilde{\mathbf{u}}_h)) &= (\tilde{p}_h, \nabla \cdot \tilde{\mathbf{u}}_h) \\ &= (\tilde{p}_h, \|p_h\|_P^{1-r'/r} |p_h|_P^{r'/r-1} p_h - \nabla \cdot \hat{\mathbf{u}}_h) \\ &\leq \|\tilde{p}_h\|_P \left(\|p_h\|_P^{1-r'/r} \| |p_h|_P^{r'/r-1} p_h \|_{L^r} + \|\nabla \cdot \hat{\mathbf{u}}_h\|_{L^r} \right) \\ &\leq \|\tilde{p}_h\|_P \left(\|p_h\|_P + C\|\hat{\mathbf{u}}_h\|_X \right) \\ &= \|\tilde{p}_h\|_P \left(\|p_h\|_P + \|\beta\|_{\mathbb{R}^m} \right). \qquad (3.14)\end{aligned}$$

Also, from the inf-sup condition for spaces X_h^0 and P_h we have

$$\begin{aligned} c\|\tilde{p}_h\|_P &\le \sup_{\mathbf{v}\in X_h^0} \frac{(\tilde{p}_h, \nabla\cdot\mathbf{v})}{\|\mathbf{v}\|_X} \\ &= \sup_{\mathbf{v}\in X_h^0} \frac{(|d(\tilde{\mathbf{u}}_h)|^{r-2}d(\tilde{\mathbf{u}}_h), d(\mathbf{v}))}{\|\mathbf{v}\|_X} \\ &\le \sup_{\mathbf{v}\in X_h^0} \frac{(\||d(\tilde{\mathbf{u}}_h)|^{r-2}d(\tilde{\mathbf{u}}_h)\|_{L^{r'}}\|d(\mathbf{v})\|_{L^r}}{\|\mathbf{v}\|_X} \\ &= \||d(\tilde{\mathbf{u}}_h)|^{r-2}d(\tilde{\mathbf{u}}_h)\|_{L^{r'}} \\ &= \|\tilde{\mathbf{u}}_h\|_X^{r/r'}. \end{aligned} \tag{3.15}$$

Combining (3.14) and (3.15) we have the estimate

$$\|\tilde{\mathbf{u}}_h\|_X \le \left(\|p_h\|_P + C\|\beta\|_{\mathbb{R}^m}\right). \tag{3.16}$$

Let $\mathbf{u}_h = \tilde{\mathbf{u}}_h + \hat{\mathbf{u}}_h$. Then, using (3.13) and (3.11),

$$\begin{aligned} s(\mathbf{u}_h,(p_h,\beta)) &= \int_\Omega p_h\nabla\cdot\tilde{\mathbf{u}}_h d\Omega + \int_\Omega p_h\nabla\cdot\hat{\mathbf{u}}_h d\Omega + \sum_{i=1}^m \beta_i\int_{S_i}\tilde{\mathbf{u}}_h\cdot\mathbf{n}ds \\ &\quad + \sum_{i=1}^m \beta_i\int_{S_i}\hat{\mathbf{u}}_h\cdot\mathbf{n}ds \\ &= \int_\Omega p_h\|p_h\|_P^{1-r'/r}|p_h|^{r'/r-1}p_h d\Omega + \sum_{i=1}^m \beta_i\int_{S_i}\hat{\mathbf{u}}_h\cdot\mathbf{n}ds \\ &\ge c\left(\|p_h\|_P^2 + \|\beta\|_{\mathbb{R}^m}^2\right). \end{aligned} \tag{3.17}$$

Thus, using (3.11), (3.16), we have

$$\begin{aligned} \sup_{\mathbf{v}_h\in X_h}\frac{s(\mathbf{v}_h,(p_h,\beta))}{\|\mathbf{v}_h\|_X} &\ge \frac{s(\mathbf{u}_h,(p_h,\beta))}{\|\mathbf{u}_h\|_X} \\ &\ge C\left(\|p_h\|_P + \|\beta\|_{\mathbb{R}^m}\right), \end{aligned}$$

from which (3.10) immediately follows. ∎

We now state and prove the existence and the uniqueness of solutions to (3.6)–(3.8).

THEOREM 3.1. *Given $\mathbf{f}\in X'$ and $Q\in\mathbb{R}^m$, there exists a unique $(\sigma_h,\mathbf{u}_h,p_h,\lambda_h)\in T_h\times X_h\times P_h\times\mathbb{R}^m$ satisfying* (3.6)–(3.8). *In addition,*

$$\|\sigma_h\|_T \le C(\|\mathbf{f}\|_{X'} + \|Q\|_{\mathbb{R}^m}^{r/r'}). \tag{3.18}$$

Proof. With the inf-sup conditions given in (3.4) and (3.10) the proof of existence follows exactly as for the continuous problem in Theorem 2.1. Similarly, the norm estimate for σ_h follows as that for σ given in Corollary 2.1. ∎

4. A priori error estimate. In this section we derive an error estimate for the error in the approximation $(\sigma_h,\mathbf{u}_h,p_h,\lambda_h)$ satisfying (3.6)–(3.8) and $(\sigma,\mathbf{u},p,\lambda)$ satisfying (2.12)–(2.14).

The proof of the estimates gives in Theorem 4.1 follows along the same lines as the proofs for the existence and uniqueness, except for the error estimates we work

backwards. The procedure to establish the existence and uniqueness was to reduce the problem to an equivalent problem for σ (or σ_h) on a subspace of the solution space. To obtain the error estimates we begin by considering the determining equations for σ_h, u_h, over a subspace. Using the coercivity and continuity assumptions (2.10), (2.11), an error estimate for $\|\sigma - \sigma_h\|$ over the subspace is constructed. We then show that the estimate over the subspace can be extended to the entire solution space.

Useful in the analysis below is the following inf-sup condition which follows from (3.4) and (3.10).

LEMMA 4.1. *For h sufficiently small, there exists a constant $C_{XTPRh} > 0$ such that*

$$\inf_{\mathbf{v}\in X_h} \sup_{(\tau,q,\beta)\in T_h\times P_h\times\mathbb{R}^m} \frac{b(\tau,\mathbf{v}) - s(\mathbf{v},(q,\beta))}{\|(\tau,q,\beta)\|_{T\times P\times\mathbb{R}^m}\|\mathbf{v}\|_X} \geq C_{XTPRh}, \tag{4.1}$$

where $\|(\tau,q,\beta)\|_{T\times P\times\mathbb{R}^m} := \|\tau\|_T + \|q\|_P + \|\beta\|_{\mathbb{R}^m}$.

Proof. For $\mathbf{v} \in X_h$, from (3.4) there exists τ_v such that

$$b(\tau_v,\mathbf{v}) \geq \frac{C_{XTh}}{2}\|\tau_v\|_T\|\mathbf{v}\|_X. \tag{4.2}$$

We now consider two cases. First, if $s(\mathbf{v},(q,\beta)) = 0$ for all $(q,\beta) \in P_h\times\mathbb{R}^m$, then (4.1) follows immediately from (4.2). Otherwise, from the definition of $s(\mathbf{v},(q,\beta))$, there exists $(q_v,\beta_v) \in P_h \times \mathbb{R}^m$ such that $s(\mathbf{v},(q_v,\beta_v)) < 0$ and $\|(q_v,\beta_v)\|_{P_h\times\mathbb{R}^m} = \|\tau_v\|_T$. Thus,

$$\begin{aligned}\sup_{(\tau,q,\beta)\in T_h\times P_h\times\mathbb{R}^m} \frac{b(\tau,\mathbf{v}) - s(\mathbf{v},(q,\beta))}{\|(\tau,q,\beta)\|_{T\times P\times\mathbb{R}^m}} &\geq \frac{b(\tau_v,\mathbf{v}) - s(\mathbf{v},(q_v,\beta_v))}{\|(\tau_v,q_v,\beta_v)\|_{T\times P\times\mathbb{R}^m}}\\ &\geq \frac{C_{XTh}\|\tau_v\|_T\|\mathbf{v}\|_X}{2(\|\tau_v\|_T + \|\tau_v\|_T)},\end{aligned}$$

from which (4.2) then follows. ∎

THEOREM 4.1. *For $(\sigma,\mathbf{u},p,\lambda)$ satisfying (2.12)–(2.14) and $(\sigma_h,\mathbf{u}_h,p_h,\lambda_h)$ satisfying (3.6)–(3.8), for h sufficiently small, we have that there exists a constant $C > 0$ such that*

$$\begin{aligned}\|\sigma - \sigma_h\|_T^{r'} \leq C\Bigg(&\inf_{\tau_h\in T_h}\left(\|\sigma-\tau_h\|_T^r + \|\sigma-\tau_h\|_T^{r'}\right)\\ &+ \inf_{\mathbf{v}_h\in X_h}\|\mathbf{u}-\mathbf{v}_h\|_X^r + \inf_{q_h\in P_h}\|p-q_h\|_P^{r'}\Bigg),\end{aligned} \tag{4.3}$$

$$\|\mathbf{u}-\mathbf{u}_h\|_X \leq C\left(\|\sigma-\sigma_h\|_T + \inf_{\mathbf{v}_h\in X_h}\|\mathbf{u}-\mathbf{v}_h\|_X\right), \tag{4.4}$$

$$\|p-p_h\|_P + \|\lambda-\lambda_h\|_{\mathbb{R}^m} \leq C\left(\|\sigma-\sigma_h\|_T + \inf_{q_h\in P_h}\|p-q_h\|_P\right). \tag{4.5}$$

Proof. We have that $(\sigma_h,\mathbf{u}_h,p_h,\lambda_h)$ satisfies

$$a(\sigma_h,\tau_h) - b(\tau_h,\mathbf{u}_h) = 0 \quad \forall\tau_h \in T_h, \tag{4.6}$$

$$b(\sigma_h,\mathbf{v}_h) - s(\mathbf{v}_h,(p_h,\lambda_h)) = \langle\mathbf{f},\mathbf{v}_h\rangle \quad \forall\mathbf{v}_h \in X_h, \tag{4.7}$$

$$s(\mathbf{u}_h,(q_h,\beta_h)) = \sum_{i=1}^m Q_i\beta_i \quad \forall(q_h,\beta) \in P_h\times\mathbb{R}^m. \tag{4.8}$$

Introduce the affine subspaces $\tilde{X}_h \subset X_h$, $\tilde{K}_h$ defined by

$$\tilde{X}_h := \left\{ \mathbf{v}_h \in X_h : s(\mathbf{v}_h, (q_h, \beta)) = \sum_{i=1}^{m} Q_i \beta_i \quad \forall (q_h, \beta_h) \in P_h \times \mathbb{R}^m \right\}, \tag{4.9}$$

$$\tilde{K}_h := \{ \tau_h \in T_h : b(\tau_h, \mathbf{v}_h) - s(\mathbf{v}_h, (p_h, \lambda_h)) = \langle \mathbf{f}, \mathbf{v}_h \rangle \quad \forall \mathbf{v}_h \in \tilde{X}_h \}. \tag{4.10}$$

Note that $\sigma_h \in \tilde{K}_h$ and $\mathbf{u}_h \in \tilde{X}_h$.

From (2.10), (2.11) we have

$$\begin{aligned} c\|\sigma - \sigma_h\|_T^{r'} &\le a(\sigma, \sigma - \sigma_h) - a(\sigma_h, \sigma - \sigma_h) \\ &= a(\sigma, \sigma - \tau_h) - a(\sigma_h, \sigma - \tau_h) + a(\sigma, \tau_h - \sigma_h) - a(\sigma_h, \tau_h - \sigma_h) \\ &\le \int_\Omega M(|\sigma| + |\sigma_h|)^{r'-2} |\sigma - \sigma_h||\sigma - \tau_h| d\Omega + a(\sigma, \tau_h - \sigma_h) - a(\sigma_h, \tau_h - \sigma_h). \end{aligned} \tag{4.11}$$

Now, noting that $1 < r \le 2$, and hence $r'/r \ge 1$,

$$\begin{aligned} &\int_\Omega M(|\sigma| + |\sigma_h|)^{r'-2} |\sigma - \sigma_h||\sigma - \tau_h| d\Omega \\ &\quad \le \left(\int_\Omega M^r (|\sigma| + |\sigma_h|)^{(r'-2)r} |\sigma - \tau_h|^r d\Omega \right)^{1/r} \|\sigma - \sigma_h\|_T \\ &\quad \le \epsilon \|\sigma - \sigma_h\|_T^{r'} + C M^r \int_\Omega 2^{(r'-2)r} \left(|\sigma|^{(r'-2)r} + |\sigma_h|^{(r'-2)r} \right) |\sigma - \tau_h|^r d\Omega \\ &\quad \le \epsilon \|\sigma - \sigma_h\|_T^{r'} \\ &\qquad + C \left(\int_\Omega \left(|\sigma|^{(r'-2)r} + |\sigma_h|^{(r'-2)r} \right)^{r'/(r'-r)} d\Omega \right)^{(r'-r)/r'} \left(\int_\Omega |\sigma - \tau_h|^{r'} d\Omega \right)^{r/r'} \\ &\quad \le \epsilon \|\sigma - \sigma_h\|_T^{r'} + C \left(\|\sigma\|_T^{r'} + \|\sigma_h\|_T^{r'} \right)^{(r'-r)/r'} \|\sigma - \tau_h\|_T^r \\ &\quad \le \epsilon \|\sigma - \sigma_h\|_T^{r'} + C \|\sigma - \tau_h\|_T^r. \end{aligned} \tag{4.12}$$

With the choice $\tau_h \in \tilde{K}_h$, using (2.12) and (4.6),

$$\begin{aligned} a(\sigma, \tau_h - \sigma_h) - a(\sigma_h, \tau_h - \sigma_h) &= b(\tau_h - \sigma_h, \mathbf{u}) - b(\tau_h - \sigma_h, \mathbf{u}_h) \\ &= b(\tau_h - \sigma_h, \mathbf{u}) \quad (\text{since } \tau_h \text{ and } \sigma_h \text{ are in } \tilde{K}_h) \\ &= b(\tau_h - \sigma_h, \mathbf{u} - \mathbf{v}_h) \quad (\text{for } \mathbf{v}_h \in \tilde{X}_h) \\ &= \int_\Omega (\tau_h - \sigma_h) : d(\mathbf{u} - \mathbf{v}_h) d\Omega \\ &= \int_\Omega (\sigma - \sigma_h) : d(\mathbf{u} - \mathbf{v}_h) d\Omega \\ &\qquad + \int_\Omega (\tau_h - \sigma) : d(\mathbf{u} - \mathbf{v}_h) d\Omega \\ &\le \|\sigma - \sigma_h\|_T \|\mathbf{u} - \mathbf{v}_h\|_X + \|\sigma - \tau_h\|_T \|\mathbf{u} - \mathbf{v}_h\|_X \\ &\le \epsilon \|\sigma - \sigma_h\|_T^{r'} + C \left(\|\sigma - \tau_h\|_T^{r'} + \|\mathbf{u} - \mathbf{v}_h\|_X^r \right). \end{aligned} \tag{4.13}$$

Combining (4.11)–(4.13) gives an error bound for $\|\sigma - \sigma_h\|_T$ in terms of the *best approximations* of σ and $\mathbf{u}$ in the sets $\tilde{K}_h$ and $\tilde{X}_h$, respectively. Next we show that we can *lift* these best approximations from $\tilde{K}_h$ and $\tilde{X}_h$ to $T_h \times X_h$. This is done in two steps: first, *lifting* from $\tilde{K}_h$ to $\tilde{W}_h$ and then using the discrete inf-sup condition to go from $\tilde{W}_h$ to $T_h \times X_h$.

Let

$$\begin{aligned}\tilde{W}_h &:= \{(\tau_h, q_h) \in T_h \times P_h : b(\tau_h, \mathbf{v}_h) - s(\mathbf{v}_h, (q_h, \lambda_h)) \\ &= \langle \mathbf{f}, \mathbf{v}_h \rangle \ \ \forall \mathbf{v}_h \in X_h\}. \end{aligned} \tag{4.14}$$

Note that if (τ_h, q_h) is in $\tilde{W}_h$, then τ_h is in $\tilde{K}_h$. Hence,

$$\inf_{\tau_h \in K_h} \|\sigma - \tau_h\|_T \leq \inf_{(\tau_h, q_h) \in \tilde{W}_h} \|(\sigma, p) - (\tau_h, q_h)\|_{T \times P}. \tag{4.15}$$

From the inf-sup conditions (4.1) we have that there exist operators $\Pi_1 : T \to T_h$ and $\Pi_2 : P \to P_h$ such that

$$b(\tau - \Pi_1 \tau, \mathbf{v}_h) - s(\mathbf{v}_h, (q - \Pi_2 q, \lambda_h)) = 0 \quad \forall \mathbf{v}_h \in X_h \tag{4.16}$$

and

$$\|(\Pi_1 \tau, \Pi_2 q)\|_{T \times P} \leq \tilde{C} \|(\tau, q)\|_{T \times P} \ \ \forall (\tau, q) \in T \times P. \tag{4.17}$$

Consider $(\tau_h, q_h) \in T_h \times P_h$, and introduce $\tilde{\sigma} := \tau_h - \Pi_1(\tau_h - \sigma)$ and $\tilde{p} := q_h - \Pi_2(q_h - p)$. Then for all $\mathbf{v}_h \in X_h$

$$\begin{aligned} b(\tilde{\sigma}, \mathbf{v}_h) - s(\mathbf{v}_h, (\tilde{p}, \lambda_h)) &= b(\sigma, \mathbf{v}_h) - s(\mathbf{v}_h, (p, \lambda_h)) \\ &= \langle \mathbf{f}, \mathbf{v}_h \rangle, \end{aligned}$$

which implies $(\tilde{\sigma}, \tilde{p}) \in \tilde{W}_h$.

Also, using (4.17),

$$\begin{aligned} \|(\tilde{\sigma}, \tilde{p}) - (\tau_h, q_h)\|_{T \times P} &= \|(\Pi_1(\sigma - \tau_h), \Pi_2(p - q_h)\|_{T \times P} \\ &\leq \tilde{C} \|(\sigma - \tau_h, p - q_h)\|_{T \times P}. \end{aligned} \tag{4.18}$$

With $(\tilde{\sigma}, \tilde{p})$ as defined above, using (4.17), (4.18), and the triangle inequality,

$$\begin{aligned} \inf_{(\tau_h, q_h) \in \tilde{W}_h} \|(\sigma, p) - (\tau_h, q_h)\|_{T \times P} &\leq \inf_{(\tau_h, q_h) \in T_h \times P_h} \|(\sigma, p) - (\tilde{\sigma}, \tilde{p})\|_{T \times P} \\ &\leq \inf_{(\tau_h, q_h) \in T_h \times P_h} (\|(\sigma, p) - (\tau_h, q_h)\|_{T \times P} + \|(\tilde{\sigma}, \tilde{p}) - (\tau_h, q_h)\|_{T \times P}) \\ &\leq (1 + \tilde{C}) \inf_{(\tau_h, q_h) \in T_h \times P_h} \|(\sigma, p) - (\tau_h, q_h)\|_{T \times P}. \end{aligned} \tag{4.19}$$

Using an analogous argument with the inf-sup condition (3.10) it is straightforward to show that

$$\inf_{\mathbf{v}_h \in \tilde{X}_h} \|\mathbf{u} - \mathbf{v}_h\|_X \leq C \inf_{\mathbf{v}_h \in X_h} \|\mathbf{u} - \mathbf{v}_h\|_X. \tag{4.20}$$

Combining (4.11)–(4.13), (4.15), (4.19), and (4.20), we then have

$$\begin{aligned} \|\sigma - \sigma_h\|_T^{r'} \leq C \Bigg(&\inf_{\tau_h \in T_h} \left(\|\sigma - \tau_h\|_T^r + \|\sigma - \tau_h\|_T^{r'} \right) + \inf_{\mathbf{v}_h \in X_h} \|\mathbf{u} - \mathbf{v}_h\|_X^r \\ &+ \inf_{q_h \in P_h} \|p - q_h\|_P^{r'} \Bigg). \end{aligned}$$

To obtain the error estimate for the velocity we use (3.4). We have that

$$\begin{aligned} C_{XTh}\|\mathbf{u}_h - \mathbf{v}_h\|_X &\le \sup_{\tau_h \in T_h} \frac{b(\tau_h, \mathbf{u}_h - \mathbf{v}_h)}{\|\tau_h\|_T} \\ &= \sup_{\tau_h \in T_h} \frac{b(\tau_h, \mathbf{u}_h - \mathbf{u}) + b(\tau_h, \mathbf{u} - \mathbf{v}_h)}{\|\tau_h\|_T} \\ &\le \sup_{\tau_h \in T_h} \frac{a(\sigma_h, \tau_h) - a(\sigma, \tau_h)}{\|\tau_h\|_T} + \|\mathbf{u} - \mathbf{v}_h\|_X. \end{aligned} \tag{4.21}$$

Proceeding as in the estimate (4.12), we have that

$$\begin{aligned} a(\sigma_h, \tau_h) - a(\sigma, \tau_h) &= \int_\Omega (\breve{g}(\sigma_h)\sigma_h - \breve{g}(\sigma)\sigma) : \tau_h d\Omega \\ &\le \int_\Omega M \left(|\sigma_h| + |\sigma|\right)^{r'-2} |\sigma - \sigma_h| : \tau_h d\Omega \\ &\le C\|\sigma - \sigma_h\|_T \|\tau_h\|_T. \end{aligned} \tag{4.22}$$

Combining (4.21) and (4.22) yields

$$\|\mathbf{u}_h - \mathbf{v}_h\|_X \le C\left(\|\sigma - \sigma_h\|_T + \|\mathbf{u} - \mathbf{v}_h\|_X\right).$$

An application of the triangle inequality then establishes (4.4).

The error estimate for the pressure and the "Lagrange multipliers" is obtained using the inf-sup condition (3.10), the trace theorem, and the equivalence of norms in $\mathbb{R}^m$. We have that

$$\begin{aligned} C_{PRXh}\left(\|p_h - q_h\|_P + \|\lambda_h - \beta_h\|_{\mathbb{R}^m}\right) &\le \sup_{\mathbf{v}_h \in X_h} \frac{s(\mathbf{v}_h, (p_h - q_h, \lambda_h - \beta_h))}{\|\mathbf{v}_h\|_X} \\ &= \sup_{\mathbf{v}_h \in X_h} \frac{s(\mathbf{v}_h, (p_h - p, \lambda_h - \lambda)) + s(\mathbf{v}_h, (p - q_h, \lambda - \beta_h))}{\|\mathbf{v}_h\|_X} \\ &\le \sup_{\mathbf{v}_h \in X_h} \frac{b(\sigma, \mathbf{v}_h) - b(\sigma_h, \mathbf{v}_h)}{\|\mathbf{v}_h\|_X} \\ &\quad + \sup_{\mathbf{v}_h \in X_h} \frac{\int_\Omega (p - q_h)\nabla \cdot \mathbf{v}_h d\Omega + \sum_{i=1}^m (\lambda_i - \beta_{h,i}) \int_{S_i} \mathbf{v}_h \cdot \mathbf{n} ds}{\|\mathbf{v}_h\|_X} \\ &\le \sup_{\mathbf{v}_h \in X_h} \frac{\int_\Omega (\sigma - \sigma_h) : d(\mathbf{v}_h) d\Omega}{\|\mathbf{v}_h\|_X} + C\left(\|p - q_h\|_P + \|\lambda - \beta_h\|_{\mathbb{R}^m}\right) \\ &\le \|\sigma - \sigma_h\|_T + C\left(\|p - q_h\|_P + \|\lambda - \beta_h\|_{\mathbb{R}^m}\right). \end{aligned}$$

Estimate (4.5) then follows using the triangle inequality. ∎

Remark. As $\mathbf{u}_h$ exactly satisfies the specified flow rates, the error in the Lagrange multipliers does not appear in the error estimate for $\|\mathbf{u} - \mathbf{u}_h\|_X$.

COROLLARY 4.1. *For* $(\sigma, \mathbf{u}, p, \lambda) \in (W^{l+1,r'})^{\acute{d}\times\acute{d}} \times \left(W^{k+1,r}\right)^{\acute{d}} \times W^{n+1,r'} \times \mathbb{R}^m$ *satisfying* (2.12)–(2.14) *and* $(\sigma_h, \mathbf{u}_h, p_h, \lambda_h)$ *satisfying* (3.6)–(3.8) *(with* T_h, X_h, P_h *defined in* (3.1)–(3.3)*), for* h *sufficiently small, we have with* $\tilde{l} := \min\{(l+1)r/r', kr/r', n+1\}$ *that*

$$\|\sigma - \sigma_h\|_T + \|\mathbf{u} - \mathbf{u}_h\|_X + \|p - q_h\|_P + \|\lambda - \beta_h\|_{\mathbb{R}^m} \le Ch^{\tilde{l}}. \tag{4.23}$$

Proof. Estimate (4.23) follows from (4.3)–(4.5) and the approximating properties of continuous piecewise polynomials. (Note that, by assumption, $r \le r'$.) ∎

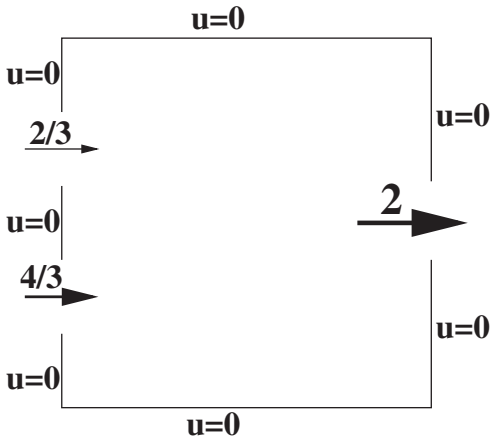


FIG. 5.1. *The flow problem.*

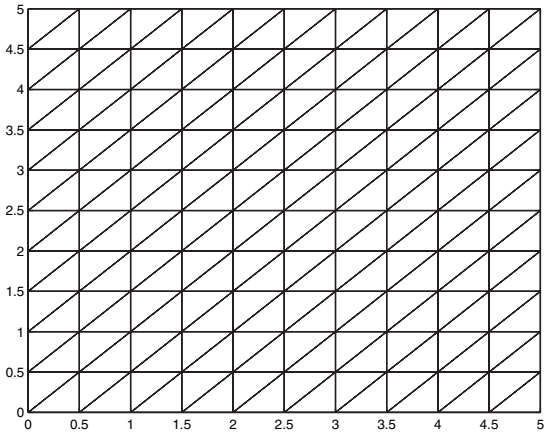


FIG. 5.2. *The second computational mesh,* $h = 1/2$.

5. Numerical computations. In this section we present numerical results, obtained using MATLAB, for a flow problem subject (only) to specified flow-rate conditions at the inflow and outflow boundaries. Along the other boundaries we impose the usual nonslip condition for the fluid velocity. In order to demonstrate the theoretical results derived in section 4 we consider a simple model problem of flow in a square domain $(0\,5) \times (0\,5)$, with inflow boundaries $x = 0$, $1 < y < 2$ and $x = 0$, $3 < y < 4$ and an outflow boundary at $x = 5$, $2 < y < 3$. The inflow rates were specified to be $4/3$ and $2/3$, respectively, with the outflow rate corresponding given as 2. (See Figure 5.1.)

Computations were performed on a sequence of four meshes, with each mesh a uniform refinement (each triangle subdivided into four similar/smaller triangles) of the preceding mesh. The second computational mesh is shown in Figure 5.2. The approximating nonlinear system was solved using a Newton method. For the approximation of the velocity and pressure we used continuous piecewise quadratic and continuous piecewise linear finite elements, respectively (i.e., the Taylor–Hood pair). For the approximation of the stress we used continuous piecewise linear finite elements.

TABLE 5.1
Norms of the velocity and stress for $r = 2$.

	$\|\nabla \mathbf{u}_h\|_{L^2}$	$\|\nabla(\mathbf{u}_h - \mathbf{u}_{2h})\|_{L^2}$	$\tilde{\alpha}_{\mathbf{u}}$	$\|\sigma_h\|_{L^2}$	$\|\sigma_h - \sigma_{2h}\|_{L^2}$	$\tilde{\alpha}_{\sigma}$
$h = 1$	6.014			2.920		
$h = 1/2$	5.763	3.192		3.836	2.601	
$h = 1/4$	5.662	1.880	0.76	3.889	1.565	0.73
$h = 1/8$	5.614	1.213	0.63	3.902	1.026	0.61

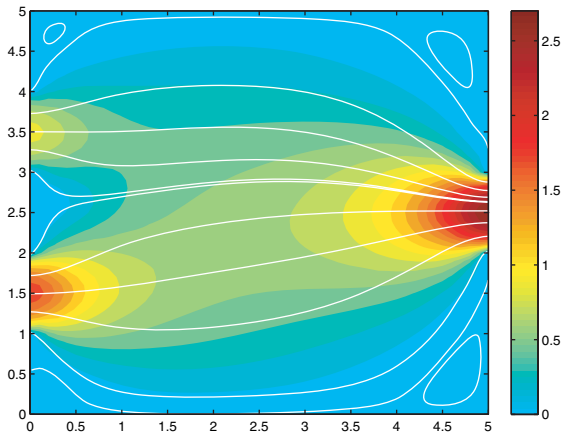


FIG. 5.3. *Plot of the magnitude of the velocity and streamlines for $r = 2$.*

TABLE 5.2
Norms of the velocity and stress for $r = 3/2$.

	$\|\nabla \mathbf{u}_h\|_{L^{3/2}}$	$\|\nabla(\mathbf{u}_h - \mathbf{u}_{2h})\|_{L^{3/2}}$	$\tilde{\alpha}_{\mathbf{u}}$	$\|\sigma_h\|_{L^3}$	$\|\sigma_h - \sigma_{2h}\|_{L^3}$	$\tilde{\alpha}_{\sigma}$
$h = 1$	8.680			1.758		
$h = 1/2$	8.664	4.710		1.995	1.279	
$h = 1/4$	8.531	2.607	0.85	2.063	0.871	0.55
$h = 1/8$	8.526	1.469	0.83	2.249	0.697	0.32

For the constitutive equation of the fluid we considered the power law equation (2.7), which, in the notation of (2.15), is rewritten as

$$d(\mathbf{u}) = \nu_0^{1-r'} |\sigma|^{r'-2} \sigma = \breve{g}(\sigma)\sigma. \tag{5.1}$$

Presented in Table 5.1 and Figure 5.3 are the results of the computations for the parameter $r = 2$ ($r' = 2$), and in Table 5.2 and Figure 5.4 are the results of the computations for the parameter $r = 3/2$ ($r' = 3$).

Assuming the convergence rate for the velocity is $\alpha_{\mathbf{u}}$, i.e., $\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^r} \sim C h^{\alpha_{\mathbf{u}}}$, we compute the experimental convergence rate for the velocity using

$$\|\nabla(\mathbf{u}_h - \mathbf{u}_{2h})\|_{L^r} \le \|\nabla(\mathbf{u} - \mathbf{u}_h)\|_{L^r} + \|\nabla(\mathbf{u} - \mathbf{u}_{2h})\|_{L^r} \sim \tilde{C} h^{\alpha_{\mathbf{u}}}.$$

$$\text{Therefore} \qquad \tilde{\alpha}_{\mathbf{u}} = \log\left(\|\nabla(\mathbf{u}_h - \mathbf{u}_{2h})\|_{L^r} / \|\nabla(\mathbf{u}_{2h} - \mathbf{u}_{4h})\|_{L^r}\right) / \log(2). \tag{5.2}$$

$$\text{Similarly} \qquad \tilde{\alpha}_{\sigma} = \log\left(\|\sigma_h - \sigma_{2h}\|_{L^{r'}} / \|\sigma_{2h} - \sigma_{4h}\|_{L^{r'}}\right) / \log(2). \tag{5.3}$$

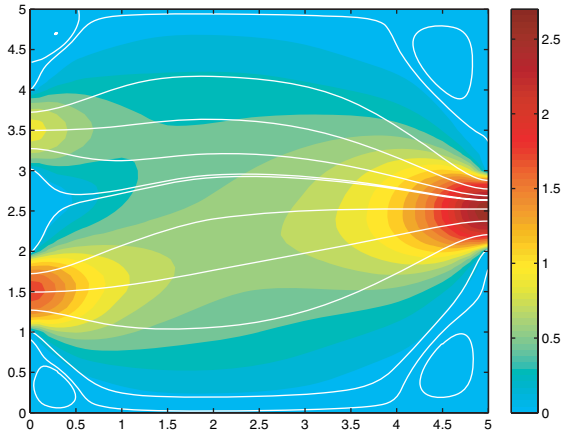


FIG. 5.4. *Plot of the magnitude of the velocity and streamlines for* $r = 3/2$.

The case $r = 2$ $(r' = 2)$. For the case $r = 2$ $(r' = 2)$ the constitutive equation describes a "Newtonian" fluid and the problem becomes a (linear) three-field Stokes problem with defective boundary conditions. As in this case $\sigma = \nu_0 d(\mathbf{u}) = \nu_0/2(\nabla \mathbf{u} + (\nabla \mathbf{u})^T)$, and we are constructing a piecewise linear approximation for the σ and a piecewise quadratic approximation for $\mathbf{u}$, we expect that $\tilde{\alpha}_{\mathbf{u}} \approx \tilde{\alpha}_\sigma$, as observed in Table 5.1. The fact that $\tilde{\alpha}_{\mathbf{u}} \approx \tilde{\alpha}_\sigma \neq 2$ is due to the lack of regularity of $\mathbf{u}$ and σ, attributable to the singular behavior of $\nabla \mathbf{u}$ and σ at the corners of the inflow and outflow boundaries.

The case $r = 3/2$ $(r' = 3)$. Note that for this case the velocity is in $(W^{1,3/2}(\Omega))^2$ and the stress in $(L^3(\Omega))^4_{sym}$. Also, we have that $\sigma = \nu_0 |d(\mathbf{u})|^{-0.5} d(\mathbf{u})$.

The a priori error estimates presented in Theorem 4.1 are dominated by the term $\|\sigma - \tau_h\|_T^r$ on the RHS of (4.3). If this term was not present, the a priori estimate would represent the *best approximation error* (for appropriately chosen approximation spaces for $\sigma_h, \mathbf{u}_h, p_h$). The computations in Table 5.2 are consistent with the approximations being *best approximations* (see below). This may be due to the fact that the behavior of the computational results are preasymptotic or that the estimates in Theorem 4.1 are not optimal.

At the end points of the inflow/outflow boundaries $\nabla \mathbf{u}$ will be singular. Assuming that at these points $\nabla \mathbf{u}$ has a point singularity of the form $\rho^{-s}, 0 < s < 1$, where ρ denotes the distance from the singular point, and $\mathbf{u}_I$ is a continuous piecewise quadratic interpolant of $\mathbf{u}$, then we expect that

$$\begin{aligned}
\|\nabla(\mathbf{u} - \mathbf{u}_I)\|_{L^r} &\sim \left(\int_{B(0,h)} (\rho^{-s})^r dA + \int_{B(0,R)\setminus B(0,h)} (h^2 \rho^{-s-2})^r dA \right)^{1/r} \\
&= \left(\int_{\theta=0}^{\pi} \int_{\rho=0}^{h} \rho \rho^{-sr} d\rho d\theta + \int_{\theta=0}^{\pi} \int_{\rho=h}^{R} \rho h^{2r} \rho^{-(s+2)r} d\rho d\theta \right)^{1/r} \\
&\sim C h^{(2-rs)/r}, \\
\text{(5.4)} \qquad \text{i.e.,} \quad \alpha_{\mathbf{u}} &= (2 - rs)/r.
\end{aligned}$$

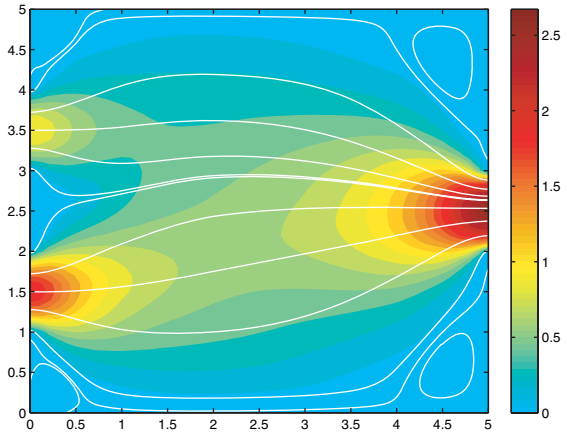


FIG. 5.5. *Plot of the magnitude of the velocity and streamlines for $r = 3/2$ with parabolic inflow boundary conditions and a "do nothing" outflow boundary condition.*

From (2.7), and for σ_I a continuous piecewise linear interpolant of σ, we would expect that

$$\begin{aligned}\|\sigma - \sigma_I\|_{L^{r'}} &\sim \left(\int_{B(0,h)} \left((\rho^{-s})^{r-2} \rho^{-s} \right)^{r'} dA \right. \\ &\qquad \left. + \int_{B(0,R)\setminus B(0,h)} \left(h^2 (\rho^{-s})^{r-2} \rho^{-s-2} \right)^{r'} dA \right)^{1/r'} \\ &\sim Ch^{(2-rs)/r'}, \\ \text{i.e.,} \quad \alpha_\sigma &= (2-rs)/r'. \end{aligned} \tag{5.5}$$

For $r = 3/2$, $r' = 3$, from (5.4),(5.5) we have that $\alpha_{\mathbf{u}}/\alpha_\sigma = r'/r = 2$, which is consistent with the computations in Table 5.2.

Comparing Figures 5.3 and 5.4, the flow fields corresponding to $r = 2$ and $r = 3/2$, respectively, we observe (as expected) the larger vortices in the upper and lower right-hand corners of Ω for the case $r = 3/2$. The magnitude of $\mathbf{u}(x, y) = [u_1(x, y), u_2(x, y)]$ plotted in Figures 5.3 and 5.4 was calculated via $|\mathbf{u}(x, y)| = (u_1(x, y)^r + u_2(x, y)^r)^{1/r}$.

For comparison, in Figure 5.5 is the flow field for the case $r = 3/2$, where we specify parabolic velocity inflow profiles (with inflow rates 4/3 and 2/3, respectively) and a "do nothing" (i.e., $\sigma - pI = \mathbf{0}$) outflow boundary condition. The flow field looks very similar to that in Figure 5.4 where the flow rates were imposed using the Lagrange multiplier approach. The values for the velocity seminorm and the stress norm in Figure 5.5 are 8.718 and 2.081, respectively, compared to 8.531 and 2.063 in Figure 5.4.

REFERENCES

[1] I. BABUŠKA, *The finite element method with Lagrange multiplers*, Numer. Math., 20 (1973), pp. 179–192.

[2] J. BARANGER, K. NAJIB, AND D. SANDRI, *Numerical analysis of a three-fields model for a quasi-Newtonian flow*, Comput. Methods Appl. Mech. Engrg., 109 (1993), pp. 281–292.

[3] R. B. BIRD, R. C. ARMSTRONG, AND O. HASSAGER, *Dynamics of Polymeric Liquids*, John Wiley and Sons, New York, 1987.

[4] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[5] V. J. Ervin and T. N. Phillips, *Residual a posteriori error estimator for a three-field model of a non-linear generalized Stokes problem*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2599–2610.
[6] L. Formaggia, J. F. Gerbeau, F. Nobile, and A. Quarteroni, *Numerical treatment of defective boundary conditions for the Navier–Stokes equations*, SIAM J. Numer. Anal., 40 (2002), pp. 376–401.
[7] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations*, Vol. 1, Springer-Verlag, New York, 1994.
[8] G. N. Gatica, M. Gonzlez, and S. Meddahi, *A low-order mixed finite element method for a class of quasi-Newtonian Stokes flows*. I. *A priori error analysis*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 881–892.
[9] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, Heidelberg, 1986.
[10] M. D. Gunzburger and S. L. Hou, *Treating inhomogeneous essential boundary conditions in finite element methods and the calculation of the boundary stresses*, SIAM J. Numer. Anal., 29 (1992), pp. 390–424.
[11] C. D. Han, *Multiphase Flow in Polymer Processing*, Academic Press, New York, 1981.
[12] J. G. Heywood, R. Rannacher, and S. Turek, *Artificial boundaries and flux and pressure conditions for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Fluids, 22 (1996), pp. 325–352.
[13] O. A. Ladyzhenskaya, *New equations for the description of the viscous incompressible fluids and solvability in the large of the boundary value problems for them*, in Boundary Value Problems of Mathematical Physics V, American Mathematical Society, Providence, RI, 1970.
[14] H. Manouzi and M. Farhloul, *Mixed finite element analysis of a non-linear three-fields Stokes model*, IMA J. Numer. Anal., 21 (2001), pp. 143–164.
[15] R. G. Owens and T. N. Phillips, *Computational Rheology*, Imperial College Press, London, 2002.
[16] M. Renardy and R. C. Rogers, *An Introduction to Partial Differential Equations*, Springer-Verlag, New York, 1993.
[17] D. Sandri, *A posteriori estimators for mixed finite element approximations of a fluid obeying the power law*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 329–340.
[18] R. Verfürth, *Finite element approximation of incompressible Navier-Stokes equations with slip boundary condition*, Numer. Math., 50 (1987), pp. 697–721.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2141–2176

EXPLICIT CONSTRUCTIONS OF QUASI-MONTE CARLO RULES FOR THE NUMERICAL INTEGRATION OF HIGH-DIMENSIONAL PERIODIC FUNCTIONS*

JOSEF DICK†

Abstract. In this paper, we give explicit constructions of point sets in the s-dimensional unit cube yielding quasi-Monte Carlo algorithms which achieve the optimal rate of convergence of the worst-case error for numerically integrating high-dimensional periodic functions. In the classical measure P_α of the worst-case error introduced by Korobov, the convergence, for every even integer $\alpha \geq 1$, is of $\mathcal{O}(N^{-\min(\alpha,d)}(\log N)^{s\alpha-2})$, where d is a parameter of the construction which can be chosen arbitrarily large and N is the number of quadrature points. This convergence rate is known to be the best possible up to some $\log N$ factors. We prove the result for the deterministic and also a randomized setting. The construction is based on a suitable extension of digital (t, m, s)-nets over the finite field $\mathbb{Z}_b$.

Key words. numerical integration, quasi-Monte Carlo method, digital net, digital sequence, lattice rule

AMS subject classifications. Primary, 11K38, 11K45, 65C05; Secondary, 65D30, 65D32

DOI. 10.1137/060658916

1. Introduction.

Korobov [11] and independently Hlawka [9] introduced a quadrature formula which is suited for numerically integrating high-dimensional periodic functions. More precisely, we want to approximate the high-dimensional integral $\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$ (where f is assumed to be periodic with period 1 in each coordinate) by a quasi-Monte Carlo rule, i.e., an equal weight quadrature rule $Q_{N,s}(f) = N^{-1}\sum_{n=0}^{N-1} f(\boldsymbol{x}_n)$, where $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_{N-1} \in [0,1]^s$ are the quadrature points. Specifically, Korobov and Hlawka suggested using a quadrature rule of the form $Q_{N,\boldsymbol{g},s}(f) = N^{-1}\sum_{n=0}^{N-1} f(\{n\boldsymbol{g}/N\})$, where for a vector of real numbers $\boldsymbol{x} = (x_1, \ldots, x_s)$ we define $\{\boldsymbol{x}\}$ as the fractional part of each component of $\boldsymbol{x}$, i.e., $\{x_j\} = x_j - \lfloor x_j \rfloor = x_j \pmod 1$, and where $\boldsymbol{g} \in \mathbb{Z}^s$ is an integer vector. The quadrature rule $Q_{N,\boldsymbol{g},s}$ is called *lattice rule*, and $\boldsymbol{g}$ is called the generating vector (of the lattice rule). The monographs [10, 12, 17, 25] deal partly or entirely with the approximation of such integrals. (Note that the assumption that the integrand f is periodic is not really a restriction since there are transformations which transform nonperiodic functions into periodic ones such that the smoothness of the integrand is preserved; see, for example, [25].)

To analyze the properties of a quadrature rule, one considers then the worst-case error $\sup_{f\in B_{\mathcal{H}}} |\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - Q_{N,s}(f)|$, where $B_{\mathcal{H}}$ denotes some class of functions. In the classical theory, the class ε_α^s of periodic functions has been considered where one demands that the absolute values of the Fourier coefficients of the function decay sufficiently fast (see [10, 12, 25, 17]). This leads us to the classical measure of the quality of lattice rules $P_\alpha = \sup_{f\in\varepsilon_\alpha^s} \left|\int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - Q_{N,s}(f)\right|$, which then for a

*Received by the editors May 4, 2006; accepted for publication (in revised form) June 14, 2007; published electronically September 28, 2007. The support of the Australian Research Council under its Centre of Excellence Program is gratefully acknowledged.
http://www.siam.org/journals/sinum/45-5/65891.html
†School of Mathematics and Statistics, UNSW, Sydney 2052, Australia (josi@maths.unsw.edu.au).

lattice rule with generating vector $\boldsymbol{g} = (g_1, \dots, g_s)$ can also be written as

$$P_\alpha = P_\alpha(\boldsymbol{g}, N) = \sum_{\substack{\boldsymbol{h} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\} \\ \boldsymbol{h} \cdot \boldsymbol{g} \equiv 0 \pmod N}} |\bar{\boldsymbol{h}}|^{-\alpha},$$

where $\boldsymbol{h} = (h_1, \dots, h_s)$, $\boldsymbol{h} \cdot \boldsymbol{g} = h_1 g_1 + \cdots + h_s g_s$, and $|\bar{\boldsymbol{h}}| = \prod_{j=1}^s \max(1, |h_j|)$. (Later on in this paper, we prefer to use the more contemporary notation of reproducing kernel Hilbert spaces, in our case so-called Korobov spaces (see section 2.3), but as is well understood (and as is also shown in section 2.3) the results also apply to the classical problem.)

By averaging over all generating vectors $\boldsymbol{g}$, several existence results for good lattice rules which achieve $P_\alpha = \mathcal{O}(N^{-\alpha}(\log N)^{\alpha s})$ have been shown; see [10, 11, 12, 18, 17, 25]. By a lower bound of Sharygin [24], this convergence is also known to be essentially the best possible, as he showed that the worst-case error is at least of order $N^{-\alpha}(\log N)^{s-1}$. But, except for dimension $s = 2$, no explicit generating vectors $\boldsymbol{g}$ which yield a small worst-case error are known. For $s \geq 3$, one relies on a computer search to find good generating vectors $\boldsymbol{g}$, and many such search algorithms have been introduced and analyzed, especially recently; see [11, 26, 27, 32].

On the other hand, one can of course also use some other quadrature rule $Q_{N,s}(f) = \sum_{n=0}^{N-1} \omega_n f(\boldsymbol{x}_n)$ to numerically integrate functions in the class ε_α^s. In this case, the worst-case error in the class ε_α^s for a quadrature rule with weights $\omega_0, \dots, \omega_{N-1}$ and points $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}\} \subset [0,1)^s$ is given by

$$P_\alpha(\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}\}) = \sum_{n,m=0}^{N-1} \omega_n \omega_m \sum_{\boldsymbol{h} \in \mathbb{Z}^s \setminus \{\boldsymbol{0}\}} \frac{\mathrm{e}^{2\pi \mathrm{i} \boldsymbol{h} \cdot (\boldsymbol{x}_n - \boldsymbol{x}_m)}}{|\bar{\boldsymbol{h}}|^\alpha}. \tag{1.1}$$

An explicit construction of such point sets was introduced by Niederreiter (see [14, Theorem 5.3]) and is called Kronecker sequence. Here the idea is to choose the quadrature points of the form $\{\boldsymbol{z}k\}$, $k = 1, 2, \dots$, where $\boldsymbol{z}$ is an s-dimensional vector of certain irrational numbers (for example, one can choose $\boldsymbol{z} = (\sqrt{p_1}, \dots, \sqrt{p_s})$, where $p_1, \dots, p_s$ are distinct prime numbers). Depending on the smoothness α, certain points will be used more than once; see [14]. In practice, problems can occur because of the finite precision of computers making it impossible to use points whose coordinates are all irrational numbers.

Another construction of quadrature rules is due to Smolyak [29] and is nowadays called a sparse grid; see also [7]. Those quadrature rules are sums over certain products of differences of one-dimensional quadrature rules. In principle, any one-dimensional quadrature rule can be chosen as a basis, leading to different quadrature rules. In many cases, the weights ω_n of such quadrature rules are not known explicitly but can be precomputed. But even if the underlying one-dimensional quadrature rule has only positive weights, it is possible that some weights in Smolyak's quadrature rules are negative, which can have a negative impact on the stability of the quadrature formula. In general, quadrature formulas for which all weights are equal and $\sum_{n=0}^{N-1} \omega_n = 1$, that is, $\omega_n = N^{-1}$ for all $n = 0, \dots, N-1$, are to be preferred. As mentioned above, such quadrature rules are called quasi-Monte Carlo rules, to which we now switch for the remainder of the paper.

As the weights for quasi-Monte Carlo rules are given by N^{-1}, the focus lies on the choice of the quadrature points. Constructions of quadrature points have been introduced with the aim to distribute the points as evenly as possible over the unit cube. An explicit construction of well-distributed point sets in the unit cube has been

introduced by Sobol [30]. A similar construction was established by Faure [6] before Niederreiter [16] (see also [17]) introduced the general concept of (t, m, s)-nets and (t, s)-sequences and the construction scheme of digital (t, m, s)-nets and digital (t, s)-sequences. For such point sets, it has been shown that the star discrepancy (which is a measure of the distribution properties of a point set) is $\mathcal{O}(N^{-1}(\log N)^{s-1})$; see [17]. From this result, it follows that those point sets yield quasi-Monte Carlo algorithms which achieve a convergence of $\mathcal{O}(N^{-2}(\log N)^{2s-2})$ for functions in the class ε_α^s for all $\alpha \geq 2$. This result holds in the deterministic and randomized setting.

For smoother functions, though, i.e., larger values of α in the class ε_α^s, one can expect higher order convergence. For example, if the partial derivatives up to order two are square integrable, then one would expect an integration error of $\mathcal{O}(N^{-4}(\log N)^{c(s)})$, for some $c(s) > 0$ depending only on s, in the function class ε_α^s, and, in general, if the mixed partial derivatives up to order $\alpha/2$ exist and are square integrable, then one would expect an integration error in ε_α^s of $\mathcal{O}(N^{-\alpha}(\log N)^{c(s,\alpha)})$, for some $c(s, \alpha) > 0$ depending only on s and α. But until now (t, m, s)-nets and (t, s)-sequences have only been shown to yield a convergence of at best $\mathcal{O}(N^{-2}(\log N)^{2s-2})$ (or $\mathcal{O}(N^{-3+\delta})$ for any $\delta > 0$ if one uses a randomization method called scrambling; see [22]) in ε_α^s, even if the integrands satisfy stronger smoothness assumptions.

In this paper, we show that a modification of digital (t, m, s)-nets and digital (t, s)-sequences introduced by Niederreiter [16, 17] yields point sets which achieve the optimal rate of convergence of the worst-case error $P_{2\alpha} = \mathcal{O}(N^{-2\min(\alpha,d)}(\log N)^{2s\alpha-2})$ for any integer $\alpha \geq 1$ and where $d \in \mathbb{N}$ is a parameter of the construction which can be chosen arbitrarily large. We, too, use the digital construction scheme introduced by Niederreiter [16, 17] for the construction of (t, m, s)-nets and (t, s)-sequences, but our analysis of the worst-case error shows that the t-value does not provide enough information about the point set. Hence we generalize the definition of digital (t, m, s)-nets and digital (t, s)-sequences to suit our needs. This leads us to the definition of digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences. For $\alpha = \beta = 1$, those definitions reduce to the case introduced by Niederreiter but are different for $\alpha > 1$. Subsequently, we prove that quasi-Monte Carlo rules based on digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences achieve the optimal rate of convergence. Further we give explicit constructions of digital $(t, \alpha, \min(\alpha, d), m, s)$-nets and digital $(t, \alpha, \min(\alpha, d), s)$-sequences, where $d \in \mathbb{N}$ is a parameter of the construction which can be chosen arbitrarily large.

Digital $(t, 2, 2, m, s)$-nets and digital $(t, 2, 2, s)$-sequences over $\mathbb{Z}_b$ (i.e., where $\alpha = \beta = 2$) can also be used for nonperiodic function spaces where one uses randomly shifted and then folded point sets using the baker's transformation (see [3]). Our analysis and error bounds for $\alpha = 2$ here also apply for the case considered in [3] (with different constants though), hence yielding useful constructions also for nonperiodic function spaces where one uses the baker's transformation. Using a digital (t, α, m, s)-net with a scrambling algorithm (see [22]), on the other hand, does not improve the performance in nonperiodic spaces compared to (t, m, s)-nets.

In the following we summarize some properties of the quadrature rules:

- The quadrature rules introduced in this paper are equal weight quadrature rules which achieve the optimal rate of convergence up to some $\log N$ factors, and we show the result for deterministic and randomly digitally shifted quadrature rules. The upper bound for the randomized quadrature rules even improves upon the best known upper bound (more precisely, the power of the $\log N$ factor) for lattice rules for the worst-case error in ε_α^s for all dimensions $s \geq 2$ and even integers $\alpha \geq 2$ (compare Corollary 6.5 to Theorem 2 in [18]).

- The construction of the underlying point set is explicit.
- They automatically adjust themselves to the optimal rate of convergence in the class $\varepsilon^s_{2\alpha}$ as long as α is an integer such that $\alpha \le d$, where d is a parameter of the construction which can be chosen arbitrarily large.
- The underlying point set is extensible in the dimension as well as in the number of points; i.e., one can always add some coordinates or points to an existing point set such that the quality of the point set is preserved.
- Tractability and strong tractability results (see [28]) can be obtained for weighted Korobov spaces.

The outline of the paper is as follows. In the next section we introduce the necessary tools, namely, Walsh functions, the digital construction scheme upon which the construction of the point set is based on and Korobov spaces. Further we also introduce the worst-case error in those Korobov spaces, and we give a representation of this worst-case error for digital nets in terms of the Walsh coefficients of the reproducing kernel. In section 3, we give the definition of digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences. Further we prove some propagation rules for those digital nets and sequences. In section 4, we give explicit constructions of digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences, and we prove some upper bounds on the t-value. We then show, in section 5, that quasi-Monte Carlo rules based on those digital nets and sequences achieve the optimal rate of convergence of the worst-case error in the Korobov spaces. The results are based on entirely deterministic point sets. Section 6 finally deals with randomly digitally shifted digital (t, α, β, m, s)-nets and (t, α, β, s)-sequences, and we show similar results for the mean square worst-case error in the Korobov space for this setting. The appendix is devoted to the analysis of the Walsh coefficients of the Walsh series representation of $B_{2\alpha}(|x-y|)$, where $B_{2\alpha}$ is the Bernoulli polynomial of degree 2α. In the last section, we give a concrete example of a digital $(t, \alpha, \alpha, m, s)$-net where we compute the t-value by hand.

2. Preliminaries. In this section we introduce the necessary tools for the analysis of the worst-case error and the construction of the point sets. In the following, let $\mathbb{N}$ denote the set of natural numbers, and let $\mathbb{N}_0$ denote the set of nonnegative integers.

2.1. Walsh functions. In the following, we define Walsh functions in base $b \ge 2$ which are the main tool of analyzing the worst-case error. First we give the definition for the one-dimensional case.

DEFINITION 2.1. *Let $b \ge 2$ be an integer and represent $k \in \mathbb{N}_0$ in base b, $k = \kappa_{a-1} b^{a-1} + \cdots + \kappa_0$, with $\kappa_i \in \{0, \ldots, b-1\}$. Further let $\omega_b = \mathrm{e}^{2\pi \mathrm{i}/b}$. Then the kth Walsh function ${}_b\mathrm{wal}_k : [0,1) \to \{1, \omega_b, \ldots, \omega_b^{b-1}\}$ in base b is given by*

$$ {}_b\mathrm{wal}_k(x) = \omega_b^{x_1\kappa_0 + \cdots + x_a\kappa_{a-1}}, $$

for $x \in [0,1)$ with base b representation $x = x_1 b^{-1} + x_2 b^{-2} + \cdots$ (unique in the sense that infinitely many of the x_i are different from $b-1$).

DEFINITION 2.2. *For dimension $s \ge 2$, $\boldsymbol{x} = (x_1, \ldots, x_s) \in [0,1)^s$, and $\boldsymbol{k} = (k_1, \ldots, k_s) \in \mathbb{N}_0^s$, we define ${}_b\mathrm{wal}_{\boldsymbol{k}} : [0,1)^s \to \{1, \omega_b, \ldots, \omega_b^{b-1}\}$ by*

$$ {}_b\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \prod_{j=1}^{s} {}_b\mathrm{wal}_{k_j}(x_j). $$

As we will always use Walsh functions in base b, we will in the following often write wal instead of ${}_b$wal.

We introduce some notation. By $\oplus$ we denote the digitwise addition modulo b; i.e., for $x = \sum_{i=w}^{\infty} x_i b^{-i}$ and $y = \sum_{i=w}^{\infty} y_i b^{-i}$ we define

$$x \oplus y = \sum_{i=w}^{\infty} z_i b^{-i},$$

where $z_i \in \{0, \dots, b-1\}$ is given by $z_i \equiv x_i + y_i \pmod b$, and let $\ominus$ denote the digitwise subtraction modulo b. In the same manner we also define a digitwise addition and digitwise subtraction for nonnegative integers based on the b-adic expansion. For vectors in $[0,1)^s$ or $\mathbb{N}_0^s$, the operations $\oplus$ and $\ominus$ are carried out componentwise. Throughout the paper, we always use base b for the operations $\oplus$ and $\ominus$. Further we call $x \in [0,1)$ a b-adic rational if it can be written in a finite base b expansion.

In the following proposition we summarize some basic properties of Walsh functions.

PROPOSITION 2.3.

1. *For all $k, l \in \mathbb{N}_0$ and all $x, y \in [0,1)$, with the restriction that if x, y are not b-adic rationals, then $x \oplus y$ is not allowed to be a b-adic rational, we have*

$$\mathrm{wal}_k(x) \cdot \mathrm{wal}_l(x) = \mathrm{wal}_{k \oplus l}(x), \quad \mathrm{wal}_k(x) \cdot \mathrm{wal}_k(y) = \mathrm{wal}_k(x \oplus y).$$

2. *We have*

$$\int_0^1 \mathrm{wal}_0(x)\,\mathrm{d}x = 1 \quad \textit{and} \quad \int_0^1 \mathrm{wal}_k(x) = 0 \ \textit{if } k > 0.$$

3. *For all $\boldsymbol{k}, \boldsymbol{l} \in \mathbb{N}_0^s$ we have the following orthogonality properties:*

$$\int_{[0,1)^s} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) \mathrm{wal}_{\boldsymbol{l}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \begin{cases} 1 & \textit{if } \boldsymbol{k} = \boldsymbol{l}, \\ 0 & \textit{otherwise.} \end{cases}$$

4. *For any $f \in \mathcal{L}_2([0,1)^s)$ and any $\boldsymbol{\sigma} \in [0,1)^s$ we have*

$$\int_{[0,1)^s} f(\boldsymbol{x} \oplus \boldsymbol{\sigma})\,\mathrm{d}\boldsymbol{x} = \int_{[0,1)^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.$$

5. *For any integer $s \geq 1$ the system $\{\mathrm{wal}_{\boldsymbol{k}} : \boldsymbol{k} = (k_1, \dots, k_s), k_1, \dots, k_s \geq 0\}$ is a complete orthonormal system in $\mathcal{L}_2([0,1)^s)$.*

The proofs of 1–3 are straightforward, and for a proof of the remaining items see [2] or [31] for more information.

2.2. The digital construction scheme. The construction of the point set used here is based on the digital construction scheme introduced by Niederreiter; see [17].

DEFINITION 2.4. *Let integers $m, s \geq 1$ and $b \geq 2$ be given. Let R_b be a commutative ring with identity such that $|R_b| = b$, and let $\mathbb{Z}_b = \{0, \dots, b-1\}$. Let $C_1, \dots, C_s \in R_b^{m \times m}$, with $C_j = (c_{j,k,l})_{1 \leq k,l \leq m}$. Further, let $\psi_l : \mathbb{Z}_b \to R_b$ for $l = 0, \dots, m-1$ and $\mu_{j,k} : R_b \to \mathbb{Z}_b$ for $j = 1, \dots, s$ and $k = 1, \dots, m$ be bijections.*

For $n = 0, \dots, b^m - 1$ let $n = \sum_{l=0}^{m-1} a_l(n) b^l$, with all $a_l(n) \in \mathbb{Z}_b$, be the base b digit expansion of n. Let $\vec{n} = (\psi_0(a_0(n)), \dots, \psi_{m-1}(a_{m-1}(n)))^T$, and let $\vec{y}_j = (y_{j,1}, \dots, y_{j,m})^T = C_j \vec{n}$ for $j = 1, \dots, s$. Then we define $x_{j,n} = \mu_{j,1}(y_{j,1}) b^{-1} + \cdots + \mu_{j,m}(y_{j,m}) b^{-m}$ for $j = 1, \dots, s$ and $n = 0, \dots, b^m - 1$, and the nth point $\boldsymbol{x}_n$ is then given by $\boldsymbol{x}_n = (x_{1,n}, \dots, x_{s,n})$. The point set $\{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{b^m-1}\}$ is called a digital net (over R_b) (with generating matrices $C_1, \dots, C_s$).

For $m = \infty$ we obtain a sequence $\{\boldsymbol{x}_0, \boldsymbol{x}_1, \dots\}$, which is called a digital sequence (over R_b) (with generating matrices $C_1, \dots, C_s$).

Niederreiter's concept of a digital (t, m, s)-net and a digital (t, s)-sequence will appear as a special case in section 3. Apart from sections 3 and 4, where we state the results using Definition 2.4 in the general form, we use only a special case of Definiton 2.4, where we assume that b is a prime number, we choose R_b the finite field $\mathbb{Z}_b$, and the bijections ψ_l and $\mu_{j,k}$ from $\mathbb{Z}_b$ to $\mathbb{Z}_b$ are all chosen to be the identity map.

We remark that, throughout the paper when Walsh functions wal, digitwise addition $\oplus$, digitwise subtraction $\ominus$, or digital nets are used in conjunction with each other, we always use the same base b for each of those operations.

2.3. Korobov space. Historically the function class ε_α^s has been used. In this paper, we use a more contemporary notation by replacing the function class ε_α^s with a reproducing kernel Hilbert space $\mathcal{H}_\alpha$ called the Korobov space. The worst-case error expression (1.1) will almost be the same for both function classes, and hence the results apply for both cases.

A reproducing kernel Hilbert space $\mathcal{H}$ over $[0,1)^s$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$ which allows a function $K : [0,1)^s \to \mathbb{R}$ such that $K(\cdot, \boldsymbol{y}) \in \mathcal{H}$, $K(\boldsymbol{x}, \boldsymbol{y}) = K(\boldsymbol{y}, \boldsymbol{x})$, and $\langle f, K(\cdot, \boldsymbol{y}) \rangle = f(\boldsymbol{y})$ for all $\boldsymbol{x}, \boldsymbol{y} \in [0,1)^s$ and all $f \in \mathcal{H}$. For more information on reproducing kernel Hilbert spaces, see [1]; for more information on reproducing kernel Hilbert spaces in the context of numerical integration, see, for example, [4, 28].

The Korobov space $\mathcal{H}_\alpha$ is a reproducing kernel Hilbert space of periodic functions. Its reproducing kernel is given by

$$K_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\boldsymbol{h} \in \mathbb{Z}^s} \frac{\mathrm{e}^{2\pi \mathrm{i} \boldsymbol{h} \cdot (\boldsymbol{x} - \boldsymbol{y})}}{|\bar{\boldsymbol{h}}|^{2\alpha}},$$

where $\alpha > 1/2$ and $|\bar{\boldsymbol{h}}| = \prod_{j=1}^s \max(1, |h_j|)$. The inner product in the space $\mathcal{H}_\alpha$ is given by

$$\langle f, g \rangle_\alpha = \sum_{\boldsymbol{h} \in \mathbb{Z}^s} |\bar{\boldsymbol{h}}|^{2\alpha} \hat{f}(\boldsymbol{h}) \hat{g}(\boldsymbol{h}), \tag{2.1}$$

where

$$\hat{f}(\boldsymbol{h}) = \int_{[0,1)^s} f(\boldsymbol{x}) \mathrm{e}^{-2\pi \mathrm{i} \boldsymbol{h} \cdot \boldsymbol{x}} \, \mathrm{d}\boldsymbol{x}$$

are the Fourier coefficients of f. The norm is given by $\|f\|_\alpha = \langle f, f \rangle_\alpha^{1/2}$.

Note that for α a natural number and any $x \in (0,1)$ we have

$$B_{2\alpha}(x) = \frac{(-1)^{\alpha+1} (2\alpha)!}{(2\pi)^{2\alpha}} \sum_{h \neq 0} \frac{\mathrm{e}^{2\pi \mathrm{i} h x}}{|h|^{2\alpha}},$$

where $B_{2\alpha}$ is the Bernoulli polynomial of degree 2α. Hence, for α a natural number we can write

$$K_\alpha(\boldsymbol{x}, \boldsymbol{y}) = \prod_{j=1}^s \left(1 + \sum_{h \neq 0} \frac{\mathrm{e}^{2\pi \mathrm{i} h (x_j - y_j)}}{|h|^{2\alpha}} \right) = \prod_{j=1}^s \left(1 - (-1)^\alpha \frac{(2\pi)^{2\alpha}}{(2\alpha)!} B_{2\alpha}(|x_j - y_j|) \right).$$

Let now

$$K_\alpha(x,y) = 1 + \sum_{h\neq 0} \frac{\mathrm{e}^{2\pi \mathrm{i} h(x-y)}}{|h|^{2\alpha}} = 1 - (-1)^\alpha \frac{(2\pi)^{2\alpha}}{(2\alpha)!} B_{2\alpha}(|x-y|). \tag{2.2}$$

Then we have

$$K_\alpha(\boldsymbol{x},\boldsymbol{y}) = \prod_{j=1}^{s} K_\alpha(x_j,y_j),$$

where $\boldsymbol{x} = (x_1,\ldots,x_s)$ and $\boldsymbol{y} = (y_1,\ldots,y_s)$. Hence the Korobov space is a tensor product of one-dimensional reproducing kernel Hilbert spaces.

Though $\alpha > 1/2$ can in general be any real number, we restrict ourselves to integers $\alpha \geq 1$ for most of this paper. The bounds on the integration error for $\mathcal{H}_\alpha$, with $\alpha \geq 1$ a real number, still apply when one replaces α with $\lfloor \alpha \rfloor$, as in this case the unit ball of $\mathcal{H}_\alpha$ given by $\{f \in \mathcal{H}_\alpha : \|f\|_\alpha \leq 1\}$ is contained in the unit ball $\{f \in \mathcal{H}_{\lfloor\alpha\rfloor} : \|f\|_{\lfloor\alpha\rfloor} \leq 1\}$ of $\mathcal{H}_{\lfloor\alpha\rfloor}$ as $\|f\|_{\lfloor\alpha\rfloor} \leq \|f\|_\alpha$. Hence it follows that integration in the space $\mathcal{H}_\alpha$ is easier than integration in the space $\mathcal{H}_{\lfloor\alpha\rfloor}$.

In general, the worst-case error $e(P,\mathcal{H})$ for multivariate integration in a normed space $\mathcal{H}$ over $[0,1]^s$ with norm $\|\cdot\|$ using a point set P is given by

$$e(P,\mathcal{H}) = \sup_{f\in\mathcal{H},\|f\|\leq 1} \left| \int_{[0,1]^s} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - Q_P(f) \right|,$$

where $Q_P(f) = N^{-1}\sum_{\boldsymbol{x}\in P} f(\boldsymbol{x})$ and $N = |P|$ is the number of points in P. If $\mathcal{H}$ is a reproducing kernel Hilbert space with reproducing kernel K, we will write $e(P,K)$ instead of $e(P,\mathcal{H})$. It is known that (see, for example, [28])

$$e^2(P,K) = \int_{[0,1)^{2s}} K(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} - \frac{2}{N}\sum_{n=0}^{N-1}\int_{[0,1)^s} K(\boldsymbol{x}_n,\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} + \frac{1}{N^2}\sum_{n,l=0}^{N-1} K(\boldsymbol{x}_n,\boldsymbol{x}_l), \tag{2.3}$$

where $P = \{\boldsymbol{x}_0,\ldots,\boldsymbol{x}_{N-1}\}$. Hence for the Korobov space $\mathcal{H}_\alpha$ we obtain

$$e^2(P,K_\alpha) = -1 + \frac{1}{N^2}\sum_{n,h=0}^{N-1} K_\alpha(\boldsymbol{x}_n,\boldsymbol{x}_h). \tag{2.4}$$

Therefore it follows that $e^2(P,K_\alpha) = P_{2\alpha}$, and hence our results also apply to the classical setting introduced by Korobov [11].

It follows from Proposition 2.3 that K_α can be represented by a Walsh series, i.e., let

$$K_\alpha(\boldsymbol{x},\boldsymbol{y}) = \sum_{\boldsymbol{k},\boldsymbol{l}\in\mathbb{N}_0^s} r_{b,\alpha}(\boldsymbol{k},\boldsymbol{l})\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})\overline{\mathrm{wal}_{\boldsymbol{l}}(\boldsymbol{y})}, \tag{2.5}$$

where

$$r_{b,\alpha}(\boldsymbol{k},\boldsymbol{l}) = \int_{[0,1)^{2s}} K_\alpha(\boldsymbol{x},\boldsymbol{y})\overline{\mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x})}\mathrm{wal}_{\boldsymbol{l}}(\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.$$

As the kernel K_α is a product of one-dimensional kernels, it follows that $r_{b,\alpha}(\boldsymbol{k},\boldsymbol{l}) = \prod_{j=1}^s r_{b,\alpha}(k_j,l_j)$, where $\boldsymbol{k} = (k_1,\ldots,k_s)$ and $\boldsymbol{l} = (l_1,\ldots,l_s)$ and

$$r_{b,\alpha}(k,l) = \int_0^1\int_0^1 K_\alpha(x,y)\overline{\mathrm{wal}_k(x)}\mathrm{wal}_l(y)\,\mathrm{d}x\,\mathrm{d}y.$$

For a digital net with generating matrices $C_1, \dots, C_s$, let $\mathcal{D} = \mathcal{D}(C_1, \dots, C_s)$ be the dual net given by

$$\mathcal{D} = \{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\} : C_1^T \vec{k}_1 + \cdots + C_s^T \vec{k}_s = \vec{0}\},$$

where for $\boldsymbol{k} = (k_1, \dots, k_s)$, with $k_j = \kappa_{j,0} + \kappa_{j,1} b + \cdots$, we set $\vec{k}_j = (\kappa_{j,0}, \dots, \kappa_{j,m-1})^T$. Further, for $\emptyset \neq u \subseteq \{1, \dots, s\}$ let $\mathcal{D}_u = \mathcal{D}((C_j)_{j \in u})$. We have the following theorem.

THEOREM 2.5. *Let $C_1, \dots, C_s \in \mathbb{Z}_b^{m \times m}$ be the generating matrices of a digital net P_{b^m}, and let $\mathcal{D}$ denote the dual net. Then for any $\alpha > 1/2$ the square worst-case error in $\mathcal{H}_\alpha$ is given by*

$$e^2(P_{b^m}, K_\alpha) = \sum_{\boldsymbol{k},\boldsymbol{l} \in \mathcal{D}} r_{b,\alpha}(\boldsymbol{k}, \boldsymbol{l}).$$

Proof. From (2.4) and (2.5) it follows that

$$e^2(P_{b^m}, K_\alpha) = -1 + \sum_{\boldsymbol{k},\boldsymbol{l} \in \mathbb{N}_0^s} r_{b,\alpha}(\boldsymbol{k}, \boldsymbol{l}) \frac{1}{b^{2m}} \sum_{\boldsymbol{x},\boldsymbol{y} \in P_{b^m}} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) \overline{\mathrm{wal}_{\boldsymbol{l}}(\boldsymbol{y})}.$$

In [4] it was shown that

$$\frac{1}{b^m} \sum_{\boldsymbol{x} \in P_{b^m}} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{k} \in \mathcal{D} \cup \{\boldsymbol{0}\}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence we have

$$e^2(P_{b^m}, K_\alpha) = -1 + \sum_{\boldsymbol{k},\boldsymbol{l} \in \mathcal{D} \cup \{\boldsymbol{0}\}} r_{b,\alpha}(\boldsymbol{k}, \boldsymbol{l}).$$

In the following we will show that $r_{b,\alpha}(\boldsymbol{0}, \boldsymbol{0}) = 1$ and $r_{b,\alpha}(\boldsymbol{0}, \boldsymbol{k}) = r_{b,\alpha}(\boldsymbol{k}, \boldsymbol{0}) = 0$ if $\boldsymbol{k} \neq \boldsymbol{0}$ from which the result then follows. Note that it is enough to show those identities for the one-dimensional case. We have $\mathrm{wal}_0(x) = 1$ for all $x \in [0, 1)$, and hence

$$\begin{aligned} r_{b,\alpha}(0, k) &= \int_0^1 \int_0^1 \left(1 + \sum_{h \in \mathbb{Z} \setminus \{0\}} |h|^{-2\alpha} \mathrm{e}^{2\pi \mathrm{i} h(x-y)}\right) \mathrm{wal}_k(y) \, \mathrm{d}x \, \mathrm{d}y \\ &= \int_0^1 \mathrm{wal}_k(y) \, \mathrm{d}y + \int_0^1 \sum_{h \in \mathbb{Z} \setminus \{0\}} |h|^{-2\alpha} \int_0^1 \mathrm{e}^{2\pi \mathrm{i} h x} \, \mathrm{d}x \, \mathrm{e}^{-2\pi \mathrm{i} h y} \mathrm{wal}_k(y) \, \mathrm{d}y \\ &= \int_0^1 \mathrm{wal}_k(y) \, \mathrm{d}y. \end{aligned}$$

It now follows from Proposition 2.3 that $r_{b,\alpha}(0, 0) = 1$ and $r_{b,\alpha}(0, k) = 0$ for $k > 0$. The result for $r_{b,\alpha}(k, 0)$ can be obtained in the same manner. Hence the result follows. ∎

In the following lemma we obtain a formula for the Walsh coefficients $r_{b,\alpha}$.

LEMMA 2.6. *Let $b \geq 2$ be an integer, and let $\alpha > 1/2$ be a real number. The Walsh coefficients $r_{b,\alpha}(k, l)$ for $k, l \in \mathbb{N}$ are given by*

$$r_{b,\alpha}(k, l) = \sum_{h \in \mathbb{Z} \setminus \{0\}} \frac{\overline{\beta_{h,k}} \beta_{h,l}}{|h|^{2\alpha}},$$

where $\beta_{h,k} = \int_0^1 \mathrm{e}^{-2\pi \mathrm{i} h x} \mathrm{wal}_k(x) \, \mathrm{d}x$.

Proof. We have

$$r_{b,\alpha}(k,l) = \int_0^1 \int_0^1 \sum_{h\in\mathbb{Z}\setminus\{0\}} |h|^{-2\alpha} \mathrm{e}^{2\pi \mathrm{i} h(x-y)} \overline{\mathrm{wal}_k(x)} \mathrm{wal}_l(y)\,\mathrm{d}x\,\mathrm{d}y$$
$$= \sum_{h\in\mathbb{Z}\setminus\{0\}} |h|^{-2\alpha} \int_0^1 \mathrm{e}^{2\pi \mathrm{i} hx} \overline{\mathrm{wal}_k(x)}\,\mathrm{d}x \int_0^1 \mathrm{e}^{-2\pi \mathrm{i} hy} \mathrm{wal}_l(y)\,\mathrm{d}y.$$

The result follows. □

It is difficult to calculate the exact value of $r_{b,\alpha}(k,l)$ in general, but for our purposes it is enough to obtain an upper bound. Note that $r_{b,\alpha}(k,k)$ is a nonnegative real number.

LEMMA 2.7. *Let $b \geq 2$ be an integer, and let $\alpha > 1/2$ be a real number. The Walsh coefficients $r_{b,\alpha}(k,l)$ for $k,l \in \mathbb{N}$ are bounded by*

$$|r_{b,\alpha}(k,l)|^2 \leq r_{b,\alpha}(k,k) r_{b,\alpha}(l,l).$$

Proof. Using Lemma 2.6 we obtain

$$|r_{b,\alpha}(k,l)|^2 \leq \left(\sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{|\beta_{h,k}||\beta_{h,l}|}{|h|^{2\alpha}} \right)^2 \leq \sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{|\beta_{h,k}|^2}{|h|^{2\alpha}} \sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{|\beta_{h,l}|^2}{|h|^{2\alpha}}$$
$$= r_{b,\alpha}(k,k) r_{b,\alpha}(l,l).$$

The result follows. □

In the following we will write $r_{b,\alpha}(k)$ instead of $r_{b,\alpha}(k,k)$ and also $r_{b,\alpha}(\boldsymbol{k})$ instead of $r_{b,\alpha}(\boldsymbol{k},\boldsymbol{k})$.

LEMMA 2.8. *Let $C_1,\dots,C_s \in \mathbb{Z}_b^{m\times m}$ be the generating matrices of a digital net P_{b^m}, and let $\mathcal{D}$ denote the dual net. Then for any natural number α the worst-case error in $\mathcal{H}_\alpha$ is bounded by*

$$e(P_{b^m}, K_\alpha) \leq \sum_{\boldsymbol{k}\in\mathcal{D}} \sqrt{r_{b,\alpha}(\boldsymbol{k})}.$$

Proof. From Theorem 2.5 and Lemma 2.7 it follows that

$$e^2(P_{b^m}, K_\alpha) \leq \sum_{\boldsymbol{k},\boldsymbol{l}\in\mathcal{D}} |r_{b,\alpha}(\boldsymbol{k},\boldsymbol{l})| \leq \left(\sum_{\boldsymbol{k}\in\mathcal{D}} \sqrt{r_{b,\alpha}(\boldsymbol{k},\boldsymbol{k})} \right)^2,$$

and hence the result follows. □

For $\alpha \geq 1$ a natural number we can write the reproducing kernel in terms of Bernoulli polynomials of degree 2α. Then for $k \geq 1$ we have

$$r_{b,\alpha}(k) = (-1)^{\alpha+1} \frac{(2\pi)^{2\alpha}}{(2\alpha)!} \int_0^1 \int_0^1 B_{2\alpha}(|x-y|) \overline{\mathrm{wal}_k(x)} \mathrm{wal}_k(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Note that the Bernoulli polynomials of even degree 2α are of the form

$$B_{2\alpha}(x) = c_\alpha x^{2\alpha} + c_{\alpha-1} x^{2(\alpha-1)} + \cdots + c_0 + c x^{2\alpha-1}$$

for some rational numbers $c_\alpha,\dots,c_0,c$, with $c_\alpha, c \neq 0$. Let

$$I_j(k) = \int_0^1 \int_0^1 |x-y|^j \overline{\mathrm{wal}_k(x)} \mathrm{wal}_k(y)\,\mathrm{d}x\,\mathrm{d}y. \tag{2.6}$$

As mentioned above, $r_{b,\alpha}(k)$ is a real number such that $r_{b,\alpha}(k) \geq 0$ for all $k \geq 1$ and $\alpha > 1/2$; hence, it follows that for any natural number α we have

$$r_{b,\alpha}(k) \leq \frac{(2\pi)^{2\alpha}}{(2\alpha)!} \left(|c_\alpha I_{2\alpha}(k)| + |c_{\alpha-1} I_{2(\alpha-1)}(k)| + \cdots + |c_0 I_0(k)| + |c I_{2\alpha-1}| \right).$$

Using Lemmas 8.2 and 8.5 from the appendix we obtain the following lemma.

LEMMA 2.9. *Let $b, \alpha \in \mathbb{N}$, with $b \geq 2$. For $k \in \mathbb{N}$, with $k = \kappa_1 b^{a_1 - 1} + \cdots + \kappa_\nu b^{a_\nu - 1}$, where $\nu \geq 1$, $\kappa_1, \ldots, \kappa_\nu \in \{1, \ldots, b-1\}$, and $1 \leq a_\nu < \cdots < a_1$, let $q_{b,\alpha}(k) = b^{-a_1 - \cdots - a_{\min(\nu,\alpha)}}$. Then for any natural number α and any natural number $b \geq 2$ there exists a constant $C_{b,\alpha} > 0$ which depends only on b and α such that*

$$r_{b,\alpha}(k) \leq C_{b,\alpha}^2 \, q_{b,\alpha}^2(k) \quad \text{for all } k \geq 1.$$

Let now $q_{b,\alpha}(0) = 1$. For $\boldsymbol{k} = (k_1, \ldots, k_s) \in \mathbb{N}_0^s$ we define $q_{b,\alpha}(\boldsymbol{k}) = \prod_{j=1}^s q_{b,\alpha}(k_j)$. We have the following lemma.

LEMMA 2.10. *Let $m \geq 1$, $b \geq 2$, and $\alpha \geq 2$ be natural numbers, and let $\mathcal{D}^*_{b^m,u} = \mathcal{D}_u \cap \{1, \ldots, b^m - 1\}^{|u|}$. Then we have*

$$\begin{aligned}
&\sum_{\boldsymbol{k} \in \mathcal{D}} \sqrt{r_{b,\alpha}(\boldsymbol{k})} \\
&\leq \sum_{\emptyset \neq u \subseteq \{1,\ldots,s\}} (1 + b^{-\alpha m} C_{b,\alpha}(\alpha + b^{-2}))^{s-|u|} C_{b,\alpha}^{|u|} (1 + \alpha + b^{-2})^{|u|} Q^*_{b,m,u,\alpha}(C_1, \ldots, C_s) \\
&\quad + (1 + b^{-\alpha m} C_{b,\alpha}(\alpha + b^{-2}))^s - 1,
\end{aligned}$$

where $C_{b,\alpha}$ is the constant from Lemma 2.9 and where

$$Q^*_{b,m,u,\alpha}(C_1, \ldots, C_s) = \sum_{\boldsymbol{k} \in \mathcal{D}^*_{b^m,u}} q_{b,\alpha}(\boldsymbol{k}).$$

Proof. Every $\boldsymbol{k} \in \mathbb{N}_0^s$ can be uniquely written in the form $\boldsymbol{k} = \boldsymbol{h} + b^m \boldsymbol{l}$, with $\boldsymbol{h} \in \{0, \ldots, b^m - 1\}^s$ and $\boldsymbol{l} \in \mathbb{N}_0^s$. Let $\mathcal{D}_{b^m} = \mathcal{D} \cap \{0, \ldots, b^m - 1\}^s$. Then we have

$$\sum_{\boldsymbol{k} \in \mathcal{D}} \sqrt{r_{b,\alpha}(\boldsymbol{k})} = \sum_{\boldsymbol{l} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} \sqrt{r_{b,\alpha}(b^m \boldsymbol{l})} + \sum_{\boldsymbol{h} \in \mathcal{D}_{b^m}} \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \sqrt{r_{b,\alpha}(\boldsymbol{h} + b^m \boldsymbol{l})}.$$

For the first sum we have

$$\sum_{\boldsymbol{l} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} \sqrt{r_{b,\alpha}(b^m \boldsymbol{l})} = -1 + \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \sqrt{r_{b,\alpha}(b^m \boldsymbol{l})} = -1 + \left(\sum_{l=0}^\infty \sqrt{r_{b,\alpha}(b^m l)} \right)^s.$$

By using Lemma 8.8 from the appendix and Lemma 2.9 we obtain that

$$\sum_{l=0}^\infty \sqrt{r_{b,\alpha}(b^m l)} = 1 + b^{-\alpha m} \sum_{l=1}^\infty \sqrt{r_{b,\alpha}(l)} \leq 1 + b^{-\alpha m} C_{b,\alpha} \sum_{l=1}^\infty q_{b,\alpha}(l).$$

We need to show that $\sum_{l=1}^\infty q_{b,\alpha}(l) \leq \alpha + b^{-2}$. Let $l = l_1 b^{c_1 - 1} + \cdots + l_\nu b^{c_\nu - 1}$ for some $\nu \geq 1$, with $1 \leq c_\nu < \cdots < c_1$ and $l_1, \ldots, l_\nu \in \{1, \ldots, b-1\}$. First we consider the sum over all those l for which $1 \leq \nu \leq \alpha$. This part of the sum is bounded by

$$\sum_{\nu=1}^\alpha (b-1)^\nu \sum_{c_1=\nu}^\infty \sum_{c_2=\nu-1}^{c_1-1} \cdots \sum_{c_\nu=1}^{c_{\nu-1}-1} b^{-c_1 - \cdots - c_\nu} \leq \sum_{\nu=1}^\alpha (b-1)^\nu \left(\sum_{c=1}^\infty b^{-c} \right)^\nu = \alpha.$$

If $\nu > \alpha$, we have $q_{b,\alpha}(l) = q_{b,\alpha}(l')$ for $l = l_1 b^{c_1-1} + \cdots + l_\nu b^{c_\nu - 1}$ and where $l' = l'(l) = l_1 b^{c_1-1} + \cdots + l_\alpha b^{c_\alpha - 1}$. Thus we only need to sum over all l' (i.e., natural numbers with exactly α digits) and for given l' multiply it with the number of l which yield the same l', which is $b^{c_\alpha - 1} - 1$ (and which we bound in the following by $b^{c_\alpha - 1}$). We have

$$\begin{aligned}
&(b-1)^\alpha \sum_{c_1=\alpha+1}^{\infty} \sum_{c_2=\alpha}^{c_1-1} \cdots \sum_{c_\alpha=2}^{c_{\alpha-1}-1} b^{-c_1-\cdots-c_\alpha} b^{c_\alpha-1} \\
&= b^{-1}(b-1)^\alpha \sum_{c_1=\alpha+1}^{\infty} \sum_{c_2=\alpha}^{c_1-1} \cdots \sum_{c_{\alpha-1}=3}^{c_{\alpha-2}-2} (c_{\alpha-1}-2) b^{-c_1-\cdots-c_{\alpha-1}} \\
&\le b^{-3}(b-1)^\alpha \left(\sum_{c=1}^{\infty} b^{-c} \right)^{\alpha-2} \sum_{c=1}^{\infty} c b^{-c} \\
&= \frac{1}{b^2}.
\end{aligned}$$

Thus we obtain $\sum_{l=1}^{\infty} q_{b,\alpha}(l) \le \alpha + b^{-2}$.

Further we have

$$\sum_{\boldsymbol{h} \in \mathcal{D}_{b^m}} \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \sqrt{r_{b,\alpha}(\boldsymbol{h} + b^m \boldsymbol{l})} = \sum_{\boldsymbol{h} \in \mathcal{D}_{b^m}} \prod_{j=1}^{s} \sum_{l=0}^{\infty} \sqrt{r_{b,\alpha}(h_j + b^m l)},$$

where $\boldsymbol{h} = (h_1, \ldots, h_s)$. By using Lemma 8.8 from the appendix and Lemma 2.9 we obtain

$$\sum_{l=0}^{\infty} \sqrt{r_{b,\alpha}(b^m l)} = 1 + b^{-\alpha m} C_{b,\alpha} \sum_{l=1}^{\infty} q_{b,\alpha}(l) \le 1 + b^{-\alpha m} C_{b,\alpha} (\alpha + b^{-2}).$$

Let now $0 < h_j < b^m$. From Lemma 2.9 we obtain

$$\sqrt{r_{b,\alpha}(h_j + b^m l)} \le C_{b,\alpha} q_{b,\alpha}(h_j + b^m l) \le C_{b,\alpha} q_{b,\alpha}(h_j) q_{b,\alpha}(l).$$

From above we have $\sum_{l=0}^{\infty} q_{b,\alpha}(l) \le 1 + \alpha + b^{-2}$ and hence

$$\sum_{l=0}^{\infty} \sqrt{r_{b,\alpha}(h_j + b^m l)} \le q_{b,\alpha}(h_j) C_{b,\alpha} \sum_{l=0}^{\infty} q_{b,\alpha}(l) \le C_{b,\alpha} (1 + \alpha + b^{-2}) q_{b,\alpha}(h_j).$$

Thus we obtain

$$\begin{aligned}
&\sum_{\boldsymbol{h} \in \mathcal{D}_{b^m}} \sum_{\boldsymbol{l} \in \mathbb{N}_0^s} \sqrt{r_{b,\alpha}(\boldsymbol{h} + b^m \boldsymbol{l})} \\
&= \sum_{\emptyset \ne u \subseteq \{1,\ldots,s\}} \sum_{\boldsymbol{h}_u \in \mathcal{D}^*_{b^m,u}} \prod_{j \in u} \sum_{l=0}^{\infty} \sqrt{r_{b,\alpha}(h_j + b^m l)} \prod_{j \notin u} \sum_{l=0}^{\infty} \sqrt{r_{b,\alpha}(b^m l)} \\
&\le \sum_{\emptyset \ne u \subseteq \{1,\ldots,s\}} (1 + b^{-\alpha m} C_{b,\alpha} (\alpha + b^{-2}))^{s-|u|} C_{b,\alpha}^{|u|} (1 + \alpha + b^{-2})^{|u|} \sum_{\boldsymbol{h}_u \in \mathcal{D}^*_{b^m,u}} \prod_{j \in u} q_{b,\alpha}(h_j),
\end{aligned}$$

where $\boldsymbol{h}_u = (h_j)_{j \in u}$. The result follows. ∎

In [24] it was shown that the square worst-case error for numerical integration in the Korobov space can at best be of $\mathcal{O}(N^{-2\alpha}(\log N)^{s-1})$, where N is the number of quadrature points. Hence Lemma 2.10 shows that it is enough to consider only $Q^*_{b,m,u,\alpha}(C_1,\dots,C_s)$ in order to investigate the convergence rate of digitally shifted digital nets.

3. $(t,\boldsymbol{\alpha},\boldsymbol{\beta},m,s)$-nets and $(t,\boldsymbol{\alpha},\boldsymbol{\beta},s)$-sequences. The t-value of a (t,m,s)-net is a quality parameter for the distribution properties of the net. A low t-value yields well-distributed point sets, and it has been shown (see, for example, [5, 17]) that a small t-value also guarantees a small worst-case error for integration in Sobolev spaces for which the partial first derivatives are square integrable.

In the following we will show how the definition of the t-value needs to be modified in order to obtain faster convergence rates for periodic Sobolev spaces for which the partial derivatives up to order α are square integrable. It is the aim of this definition to translate the problem of minimizing the worst-case error into an algebraical problem concerning the generating matrices. (This definition can therefore also be used in a computer search algorithm, where one could, for example, search for the polynomial lattice with the smallest $t(\alpha)$-value which, in turn, yields a small worst-case error for integration of periodic functions.)

For natural numbers $\alpha \geq 1$, Lemma 2.9 suggests defining the following metric $\mu_{b,\alpha}(\boldsymbol{k},\boldsymbol{l}) = \mu_{b,\alpha}(\boldsymbol{k} \ominus \boldsymbol{l})$ on $\mathbb{N}_0^s$ which is an extension of the metric introduced in [15]; see also [23] (for $\alpha = 1$ we basically obtain the metric in [15, 23]). Here $\mu_{b,\alpha}(0) = 0$ and for $k \in \mathbb{N}$, with $k = \kappa_\nu b^{a_\nu - 1} + \cdots + \kappa_1 b^{a_1 - 1}$, where $1 \leq a_\nu < \cdots < a_1$ and $\kappa_i \in \{1,\dots,b-1\}$, let $\mu_{b,\alpha}(k) = a_1 + \cdots + a_{\min(\alpha,\nu)}$. For a $\boldsymbol{k} \in \mathbb{N}_0^s$, with $\boldsymbol{k} = (k_1,\dots,k_s)$, let $\mu_{b,\alpha}(\boldsymbol{k}) = \mu_{b,\alpha}(k_1) + \cdots + \mu_{b,\alpha}(k_s)$. Then we have $q_{b,\alpha}(\boldsymbol{k}) = b^{-\mu_{b,\alpha}(\boldsymbol{k})}$. Hence in order to obtain a small worst-case error in the Korobov space $\mathcal{H}_\alpha$, we need digital nets for which $\min\{\mu_{b,\alpha}(\boldsymbol{k}) : \boldsymbol{k} \in \mathcal{D}\}$ is large. We can translate this property into a linear independence property of the row vectors of the generating matrices $C_1,\dots,C_s$. We have the following definition.

DEFINITION 3.1. *Let $m, \alpha \geq 1$ be natural numbers, let $0 < \beta \leq \alpha$ be a real number, and let $0 \leq t \leq \beta m$ be a natural number. Let R_b be a ring with b elements, and let $C_1,\dots,C_s \in R_b^{m\times m}$, with $C_j = (c_{j,1},\dots,c_{j,m})^T$. If for all $1 \leq i_{j,\nu_j} < \cdots < i_{j,1} \leq m$, where $0 \leq \nu_j \leq m$ for all $j = 1,\dots,s$, with*

$$i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha)} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha)} \leq \beta m - t,$$

the vectors

$$c_{1,i_{1,\nu_1}},\dots,c_{1,i_{1,1}},\dots,c_{s,i_{s,\nu_s}},\dots,c_{s,i_{s,1}}$$

are linearly independent over R_b, then the digital net which has generating matrices $C_1,\dots,C_s$ is called a digital (t,α,β,m,s)-net over R_b. Further we call a digital (t,α,α,m,s)-net over R_b a digital (t,α,m,s)-net over R_b.

If t is the smallest nonnegative integer such that the digital net generated by $C_1,\dots,C_s$ is a digital (t,α,β,m,s)-net, then we call the digital net a strict digital (t,α,β,m,s)-net or a strict digital (t,α,m,s)-net if $\alpha = \beta$.

A concrete example of a digital (t,α,β,m,s)-net, where we also calculate the exact t-value by hand, is given in section 7.

Remark 1. Using duality theory (see [19]) it follows that for every digital (t,α,β,m,s)-net we have $\min_{\boldsymbol{k}\in\mathcal{D}} \mu_{b,\alpha}(\boldsymbol{k}) > \beta m - t$, and for a strict digital (t,α,β,m,s)-net we have $\min_{\boldsymbol{k}\in\mathcal{D}} \mu_{b,\alpha}(\boldsymbol{k}) = \beta m - t + 1$. Hence digital (t,α,β,m,s)-nets with high quality have a large value of $\beta m - t$.

DEFINITION 3.2. *Let $\alpha \geq 1$ and $t \geq 0$ be integers, and let $0 < \beta \leq \alpha$ be a real number. Let R_b be a ring with b elements, and let $C_1, \ldots, C_s \in R_b^{\infty \times \infty}$, with $C_j = (c_{j,1}, c_{j,2}, \ldots)^T$. Further let $C_{j,m}$ denote the left upper $m \times m$ submatrix of C_j. If for all $m > t/\beta$ the matrices $C_{1,m}, \ldots, C_{s,m}$ generate a digital (t, α, β, m, s)-net, then the digital sequence with generating matrices $C_1, \ldots, C_s$ is called a digital (t, α, β, s)-sequence over R_b. Further we call a digital (t, α, α, s)-sequence over R_b a digital (t, α, s)-sequence over R_b.*

If t is the smallest nonnegative integer such that the digital sequence generated by $C_1, \ldots, C_s$ is a digital (t, α, β, s)-sequence, then we call the digital sequence a strict digital (t, α, β, s)-sequence or a strict digital (t, α, s)-sequence if $\alpha = \beta$.

Remark 2. Note that the definition of a digital $(t, 1, m, s)$-net coincides with the definition of a digital (t, m, s)-net and the definition of a digital $(t, 1, s)$-sequence coincides with the definition of a digital (t, s)-sequence as defined by Niederreiter [17]. Further note that the t-value depends on α and β, i.e., $t = t(\alpha, \beta)$ or $t = t(\alpha)$ if $\alpha = \beta$.

In the following theorem we establish some propagation rules.

THEOREM 3.3. *Let P be a digital (t, α, β, m, s)-net over a ring R_b, and let S be a digital (t, α, β, s)-sequence over a ring R_b. Then we have the following:*

(i) *P is a digital $(t', \alpha, \beta', m, s)$-net for all $1 \leq \beta' \leq \beta$ and all $t \leq t' \leq \beta' m$, and S is a digital (t', α, β', s)-sequence for all $1 \leq \beta' \leq \beta$ and all $t \leq t'$.*

(ii) *P is a digital $(t', \alpha', \beta', m, s)$-net for all $1 \leq \alpha' \leq m$, and S is a digital (t', α', β', s)-sequence for all $\alpha' \geq 1$, where $\beta' = \beta \min(\alpha, \alpha')/\alpha$ and $t' = \lceil t \min(\alpha, \alpha')/\alpha \rceil$.*

(iii) *Any digital (t, α, m, s)-net is a digital $(\lceil t\alpha'/\alpha \rceil, \alpha', m, s)$-net for all $1 \leq \alpha' \leq \alpha$, and every digital (t, α, s)-sequence is a digital $(\lceil t\alpha'/\alpha \rceil, \alpha', s)$-sequence for all $1 \leq \alpha' \leq \alpha$.*

Proof. Note that it follows from Definition 3.2 that we need to prove the result only for digital nets.

The first part follows trivially. To prove the second part choose an α' such that $\alpha' \geq 1$. Then choose arbitrary $1 \leq i_{j,\nu_j} < \cdots < i_{j,1} \leq m$, with $0 \leq \nu_j \leq m$, such that

$$i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha')} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha')} \leq m\beta \frac{\min(\alpha, \alpha')}{\alpha} - \left\lceil t \frac{\min(\alpha, \alpha')}{\alpha} \right\rceil.$$

We need to show that the vectors

$$c_{1,i_{1,\nu_1}}, \ldots, c_{1,i_{1,1}}, \ldots, c_{s,i_{s,\nu_s}}, \ldots, c_{s,i_{s,1}}$$

are linearly independent over R_b. This is certainly the case as long as

$$i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha)} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha)} \leq \beta m - t.$$

Indeed we have

$$\begin{aligned}
&i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha)} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha)} \\
&\leq \frac{\alpha}{\min(\alpha, \alpha')} (i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha')} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha')}) \\
&\leq m\beta - \frac{\alpha}{\min(\alpha, \alpha')} \left\lceil t \frac{\min(\alpha, \alpha')}{\alpha} \right\rceil \\
&\leq m\beta - t,
\end{aligned}$$

and hence the second part follows. The third part is just a special case of the second part. ∎

Remark 3. Note that by choosing $\alpha' = 1$ in part (iii) of Theorem 3.3 it follows that digital (t, α, m, s)-nets and digital (t, α, s)-sequences are also well-distributed point sets if the value of t is small; see [17].

4. Explicit constructions of digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences. In this section we show how suitable digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences can be constructed.

Let $d \geq 1$, and let $C_1, \ldots, C_{sd}$ be the generating matrices of a digital (t, m, sd)-net. Note that many explicit examples of such generating matrices are known; see, for example, [6, 17, 20, 30] and the references therein. For the construction of a (t, α, β, m, s)-net any of the above mentioned explicit constructions can be used, but, as will be shown below, the quality of the (t, α, β, m, s)-net obtained depends on the quality of the underlying digital (t, m, sd)-net on which our construction is based.

Let $C_j = (c_{j,1}, \ldots, c_{j,m})^T$ for $j = 1, \ldots, sd$; i.e., $c_{j,l}$ are the row vectors of C_j. Now let the matrix $C_j^{(d)}$ be made of the first rows of the matrices $C_{(j-1)d+1}, \ldots, C_{jd}$, then the second rows of $C_{(j-1)d+1}, \ldots, C_{jd}$, and so on, until $C_j^{(d)}$ is an $m \times m$ matrix, i.e., $C_j^{(d)} = (c_{j,1}^{(d)}, \ldots, c_{j,m}^{(d)})^T$, where $c_{j,l}^{(d)} = c_{u,v}$, with $l = (v-j)d + u$, $1 \leq v \leq m$, and $(j-1)d < u \leq jd$ for $l = 1, \ldots, m$ and $j = 1, \ldots, s$. In the following we will show that the matrices $C_1^{(d)}, \ldots, C_s^{(d)}$ are the generating matrices of a digital $(t, \alpha, \min(\alpha, d), m, s)$-net.

THEOREM 4.1. *Let $d \geq 1$ be a natural number, and let $C_1, \ldots, C_{sd}$ be the generating matrices of a digital (t', m, sd)-net over some ring R_b with b elements. Let $C_1^{(d)}, \ldots, C_s^{(d)}$ be defined as above. Then for any $\alpha \geq 1$ the matrices $C_1^{(d)}, \ldots, C_s^{(d)}$ are generating matrices of a digital $(t, \alpha, \min(\alpha, d), m, s)$-net over R_b, with*

$$t = \min(\alpha, d)\, t' + \left\lceil \frac{s(d-1)\min(\alpha, d)}{2} \right\rceil.$$

Proof. Let $C_j^{(d)} = (c_{j,1}^{(d)}, \ldots, c_{j,m}^{(d)})^T$ for $j = 1, \ldots, s$, and further let the integers $i_{1,1}, \ldots, i_{1,\nu_1}, \ldots, i_{s,1}, \ldots, i_{s,\nu_s}$ be such that $1 \leq i_{j,\nu_j} < \cdots < i_{j,1} \leq m$ and

$$i_{1,1} + \cdots + i_{1,\min(\nu_1,\alpha)} + \cdots + i_{s,1} + \cdots + i_{s,\min(\nu_s,\alpha)} \leq \min(\alpha, d)m - t.$$

We need to show that the vectors

$$c_{1,i_{1,1}}^{(d)}, \ldots, c_{1,i_{1,\nu_1}}^{(d)}, \ldots, c_{s,i_{s,1}}^{(d)}, \ldots, c_{s,i_{s,\nu_s}}^{(d)}$$

are linearly independent over R_b. For $j = 1, \ldots, s$ let $U_j = \{c_{j,i_{j,\nu_j}}^{(d)}, \ldots, c_{j,i_{j,1}}^{(d)}\}$. The vectors in the set U_j stem from the matrices $C_{(j-1)d+1}, \ldots, C_{jd}$. For $j = 1, \ldots, s$ and $d_j = (j-1)d + 1, \ldots, jd$ let e_{d_j} denote the largest index such that $(e_{d_j} - j)d + d_j \in \{i_{j,\nu_j}, \ldots, i_{j,1}\}$, and if for some d_j there is no such e_{d_j}, we set $e_{d_j} = 0$ (basically this means e_{d_j} is the largest integer such that $c_{d_j, e_{d_j}} \in U_j$).

Let $d \leq \alpha$; then we have $d((e_{(j-1)d+1} - 1)_+ + \cdots + (e_{jd} - 1)_+) + \sum_{l=1}^{L_j} l \leq i_{j,1} + \cdots + i_{j,\min(\nu_j,d)}$, where $(x)_+ = \max(x, 0)$ and $L_j = |\{(j-1)d+1 \leq d_j \leq jd : e_{d_j} > 0\}|$. Hence we have

$$\begin{aligned} &d((e_{(j-1)d+1} - 1)_+ + \cdots + (e_{jd} - 1)_+) + \sum_{l=1}^{L_j} l \\ &= d(e_{(j-1)d+1} + \cdots + e_{jd}) - L_j d + L_j(L_j + 1)/2 \\ &\geq d(e_{(j-1)d+1} + \cdots + e_{jd}) - \frac{d(d-1)}{2}. \end{aligned} \tag{4.1}$$

Thus it follows that

$$d(e_1+\cdots+e_{sd}) \le \sum_{j=1}^{s}(i_{j,1}+\cdots+i_{j,\min(\nu_j,\alpha)}) + s\frac{d(d-1)}{2} \le dm - t + s\frac{d(d-1)}{2},$$

and therefore

$$e_1+\cdots+e_{sd} \le m - \frac{t}{d} + s\frac{d-1}{2} \le m - t'.$$

Thus it follows from the (t', m, sd)-net property of the digital net generated by $C_1, \dots, C_{sd}$ that the vectors $c^{(d)}_{1,i_{1,1}}, \dots, c^{(d)}_{1,i_{1,\nu_1}}, \dots, c^{(d)}_{s,i_{s,1}}, \dots, c^{(d)}_{s,i_{s,\nu_s}}$ are linearly independent.

Let now $d > \alpha$. Then we have $d((e_{(j-1)d+1}-1)_+ + \cdots + (e_{jd}-1)_+) + \sum_{l=1}^{L_j} l \le i_{j,1}+\cdots+i_{j,\min(\nu_j,\alpha)} + (d-\alpha)i_{j,\min(\nu_j,\alpha)}$, where again $L_j = |\{(j-1)d+1 \le d_j \le jd : e_{d_j} > 0\}|$. Hence we can use inequality (4.1) again. Note that $i_{1,\min(\nu_1,\alpha)} + \cdots + i_{s,\min(\nu_s,\alpha)} \le m - t/\alpha$, and hence we have

$$\sum_{j=1}^{s}(i_{j,1}+\cdots+i_{j,\min(\nu_j,\alpha)}+(d-\alpha)i_{j,\min(\nu_j,\alpha)}) \le \alpha m - t + (d-\alpha)(m-t/\alpha) = dm - dt/\alpha.$$

Thus it follows that

$$\begin{aligned} d(e_1+\cdots+e_{sd}) &\le \sum_{j=1}^{s}(i_{j,1}+\cdots+i_{j,\min(\nu_j,\alpha)}+(d-\alpha)i_{j,\min(\nu_j,\alpha)}) + s\frac{d(d-1)}{2} \\ &\le dm - \frac{dt}{\alpha} + s\frac{d(d-1)}{2}, \end{aligned}$$

and therefore

$$e_1+\cdots+e_{sd} \le m - \frac{t}{\alpha} + s\frac{d-1}{2} \le m - t'.$$

Thus it follows from the (t', m, sd)-net property of the digital net generated by $C_1, \dots, C_{sd}$ that the vectors $c^{(d)}_{1,i_{1,1}}, \dots, c^{(d)}_{1,i_{1,\nu_1}}, \dots, c^{(d)}_{s,i_{s,1}}, \dots, c^{(d)}_{s,i_{s,\nu_s}}$ are linearly independent, and hence the result follows. □

In section 7 we use this construction method to construct a digital $(3, 2, 4, 2)$-net over $\mathbb{Z}_2$.

Note that the construction and Theorem 4.1 can easily be extended to (t, α, β, s)-sequences. Indeed, let $d \ge 1$, and let $C_1, \dots, C_{sd}$ be the generating matrices of a digital (t, sd)-sequence. Again many explicit generating matrices are known; see, for example, [6, 17, 20, 30]. Let $C_j = (c_{j,1}, c_{j,2}, \dots)^T$ for $j = 1, \dots, sd$; i.e., $c_{j,l}$ are the row vectors of C_j. Now let the matrix $C^{(d)}_j$ be made of the first rows of the matrices $C_{(j-1)d+1}, \dots, C_{jd}$, then the second rows of $C_{(j-1)d+1}, \dots, C_{jd}$, and so on, i.e.,

$$C^{(d)}_j = (c_{(j-1)d+1,1}, \dots, c_{jd,1}, c_{(j-1)d+1,2}, \dots, c_{jd,2}, \dots)^T.$$

The following theorem states that the matrices $C^{(d)}_1, \dots, C^{(d)}_s$ are the generating matrices of a digital $(t, \alpha, \min(\alpha, d), s)$-sequence.

THEOREM 4.2. *Let $d \ge 1$ be a natural number, and let $C_1, \dots, C_{sd}$ be the generating matrices of a digital (t', sd)-sequence over some ring R_b with b elements. Let*

$C_1^{(d)}, \dots, C_s^{(d)}$ *be defined as above. Then for any* $\alpha \geq 1$ *the matrices* $C_1^{(d)}, \dots, C_s^{(d)}$ *are generating matrices of a digital* $(t, \alpha, \min(\alpha, d), s)$*-sequence over* R_b*, with*

$$t = \min(\alpha, d)\, t' + \left\lceil \frac{s(d-1)\min(\alpha, d)}{2} \right\rceil.$$

The last result shows that (t, α, β, m, s)-nets indeed exist for any $0 < \beta \leq \alpha$ and for m arbitrarily large. We have even shown that digital (t, α, β, m, s)-nets exist which are extensible in m and s. This can be achieved by using an underlying (t', sd)-sequence which is itself extensible in m and s. If the t'-value of the original (t', m, s)-net or (t', s)-sequence is known explicitly, then we also know the t-value of the digital (t, α, β, m, s)-net or (t, α, β, s)-sequence. Furthermore it has also been shown how such digital nets can be constructed in practice.

In the following we investigate for which values of t, α, s, b digital (t, α, s)-sequences over $\mathbb{Z}_b$ exist. We need some further notation (see also [21, Definition 8.2.15]).

DEFINITION 4.3. *For given integers* $s, \alpha \geq 1$ *and prime number* b *let* $d_b(s, \alpha)$ *be the smallest value of* t *such that a* (t, α, s)*-sequence over* $\mathbb{Z}_b$ *exists.*

We have the following bound on $d_b(s, \alpha)$.

COROLLARY 4.4. *Let* $s, \alpha \geq 1$ *be integers and* b *be a prime number. Then we have*

$$\alpha \left(\frac{s}{b} - 1 - \log_b \frac{(b-1)s + b + 1}{2} \right) + 1$$
$$\leq d_b(s, \alpha) \leq \alpha(s-1)\frac{3b-1}{b-1} - \alpha \frac{(2b+4)\sqrt{s-1}}{\sqrt{b^2-1}} + 2\alpha + s\frac{\alpha(\alpha-1)}{2}.$$

Proof. The lower bound follows from part (iii) of Theorem 3.3 by choosing $\alpha' = 1$ and using a lower bound on the t-value for (t, s)-sequences (see [20]). The upper bound follows from Theorem 4.2 by choosing $d = \alpha$ and using Theorem 8.4.4 of [21]. □

5. A bound on the worst-case error in $\mathcal{H}_\alpha$ for digital (t, α, β, m, s)-nets and digital (t, α, β, s)-sequences. In this section we prove an upper bound on the worst-case error for integration in the Korobov space $\mathcal{H}_\alpha$ using digital (t, α, β, m, s)-nets and (t, α, β, s)-sequences.

LEMMA 5.1. *Let* $\alpha \geq 2$ *be a natural number, let* b *be prime, and let* $C_1, \dots, C_s \in \mathbb{Z}_b^{m \times m}$ *be the generating matrices of a digital* (t, α, β, m, s)*-net over* $\mathbb{Z}_b$*, with* $m > t/\beta$*. Then we have*

$$Q^*_{b,m,u,\alpha}(C_1, \dots, C_s) \leq 2b^{|u|\alpha} b^{-\beta m + t} (\beta m + 2)^{|u|\alpha - 1},$$

where $Q^*_{b,m,u,\alpha}$ *is defined in Lemma* 2.10*.*

Proof. We obtain a bound on $Q^*_{b,m,\{1,\dots,s\},\alpha}$; for all other subsets u, the bound can be obtained using the same arguments.

We first partition the set $\mathcal{D}^*_{b^m,\{1,\dots,s\}}$ into parts where the highest digits of k_j are prescribed, and we count the number of solutions of $C_1^T \vec{k}_1 + \cdots + C_s^T \vec{k}_s = \vec{0}$. For $j = 1, \dots, s$ let now $i_{j,\alpha} < \cdots < i_{j,1} \leq m$, with $i_{j,1} \geq 1$. Note that we now allow $i_{j,l} < 1$, in which case the contributions of those $i_{j,l}$ are to be ignored. This notation is adopted in order to avoid considering many special cases. Now we define

$$\mathcal{D}^*_{b^m,\{1,\dots,s\}}(i_{1,1}, \dots, i_{1,\alpha}, \dots, i_{s,1}, \dots, i_{s,\alpha})$$
$$= \{\boldsymbol{k} \in \mathcal{D}^*_{b^m,\{1,\dots,s\}} : k_j = \lfloor \kappa_{j,1} b^{i_{j,1}-1} + \cdots + \kappa_{j,\alpha} b^{i_{j,\alpha}-1} + l_j \rfloor, \text{ with } 0 \leq l_j < b^{i_{j,\alpha}-1}$$
$$\text{and } 1 \leq \kappa_{j,l} < b \text{ for } j = 1, \dots, s\},$$

where $\lfloor\cdot\rfloor$ just means that the contributions of $i_{j,l} < 1$ are to be ignored. Then we have

$$Q^*_{b,m,\{1,\dots,s\},\alpha}(C_1,\dots,C_s) = \sum_{i_{1,1}=1}^{m} \cdots \sum_{i_{1,\alpha}=1}^{i_{1,\alpha-1}-1} \cdots \sum_{i_{s,1}=1}^{m} \cdots \sum_{i_{s,\alpha}=1}^{i_{s,\alpha-1}-1} \frac{|\mathcal{D}^*_{b^m,\{1,\dots,s\}}(i_{1,1},\dots,i_{1,\alpha},\dots,i_{s,1},\dots,i_{s,\alpha})|}{b^{i_{1,1}+\cdots+i_{1,\alpha}+\cdots+i_{s,1}+\cdots+i_{s,\alpha}}}. \tag{5.1}$$

Some of the sums above can be empty, in which case we just set the corresponding summation index $i_{j,l} = 0$.

Note that by the (t,α,β,m,s)-net property we have

$$|\mathcal{D}^*_{b^m,\{1,\dots,s\}}(i_{1,1},\dots,i_{1,\alpha},\dots,i_{s,1},\dots,i_{s,\alpha})| = 0$$

as long as $i_{1,1}+\cdots+i_{1,\alpha}+\cdots+i_{s,1}+\cdots+i_{s,\alpha} \le \beta m - t$. Hence let now $0 \le i_{1,1},\dots,i_{s,\alpha} \le m$ be given such that $i_{1,1},\dots,i_{s,1} \ge 1$, $i_{j,\alpha} < \cdots < i_{j,1} \le m$ for $j = 1,\dots,s$ and where if $i_{j,l} < 1$ we set $i_{j,l} = 0$ and $i_{1,1}+\cdots+i_{1,\alpha}+\cdots+i_{s,1}+\cdots+i_{s,\alpha} > \beta m - t$. We now need to estimate $|\mathcal{D}^*_{b^m,\{1,\dots,s\}}(i_{1,1},\dots,i_{1,\alpha},\dots,i_{s,1},\dots,i_{s,\alpha})|$; that is, we need to count the number of $\boldsymbol{k} \in \mathcal{D}^*_{b^m,\{1,\dots,s\}}$ with $k_j = \lfloor \kappa_{j,1} b^{i_{j,1}-1} + \cdots + \kappa_{j,\alpha} b^{i_{j,\alpha}-1} + l_j \rfloor$ such that $C_1^T \vec{k}_1 + \cdots + C_s^T \vec{k}_s = \vec{0}$.

There are at most $(b-1)^{\alpha s}$ choices for $\kappa_{1,1},\dots,\kappa_{s,\alpha}$ (we write at most because if $i_{j,l} < 1$, then the corresponding $\kappa_{j,l}$ does not have any effect and therefore need not to be included). Let now $1 \le \kappa_{1,1},\dots,\kappa_{s,\alpha} < b$ be given, and define

$$\vec{g} = \kappa_{1,1} c^T_{1,i_{1,1}} + \cdots + \kappa_{1,\alpha} c^T_{1,i_{1,\alpha}} + \cdots + \kappa_{s,1} c^T_{s,i_{s,1}} + \cdots + \kappa_{s,\alpha} c^T_{s,i_{s,\alpha}},$$

where we set $c^T_{j,l} = 0$ if $l < 1$. Further let

$$B = (c^T_{1,1},\dots,c^T_{1,i_{1,\alpha}-1},\dots,c^T_{s,1},\dots,c^T_{s,i_{s,\alpha}-1}).$$

Now the task is to count the number of solutions $\vec{l}$ of $B\vec{l} = \vec{g}$. As long as the columns of B are linearly independent, the number of solutions can at most be 1. By the (t,α,β,m,s)-net property this is certainly the case if (we write $(x)_+ = \max(x,0)$)

$$\begin{aligned}(i_{1,\alpha}-1)_+ + \cdots + (i_{1,\alpha}-\alpha)_+ + \cdots + (i_{s,\alpha}-1)_+ + \cdots + (i_{s,\alpha}-\alpha)_+ &\\ \le \alpha(i_{1,\alpha}+\cdots+i_{s,\alpha}) &\\ \le \beta m - t,&\end{aligned}$$

that is, as long as

$$i_{1,\alpha}+\cdots+i_{s,\alpha} \le \frac{\beta m - t}{\alpha}.$$

Let now $i_{1,\alpha}+\cdots+i_{s,\alpha} > \frac{\beta m - t}{\alpha}$. Then by considering the rank of the matrix B and the dimension of the space of solutions of $B\vec{l} = \vec{0}$, it follows that the number of solutions of $B\vec{l} = \vec{g}$ is smaller or equal to $b^{i_{1,\alpha}+\cdots+i_{s,\alpha}-\lfloor(\beta m-t)/\alpha\rfloor}$. Thus we have

$$|\mathcal{D}^*_{b^m,\{1,\dots,s\}}(i_{1,1},\dots,i_{1,\alpha},\dots,i_{s,1},\dots,i_{s,\alpha})| \le \begin{cases} 0 & \text{if } \sum_{j=1}^{s}\sum_{l=1}^{\alpha} i_{j,l} \le \beta m - t, \\ (b-1)^{\alpha s} & \text{if } \sum_{j=1}^{s}\sum_{l=1}^{\alpha} i_{j,l} > \beta m - t \\ & \text{and } \sum_{j=1}^{s} i_{j,\alpha} \le \frac{\beta m - t}{\alpha}, \\ (b-1)^{\alpha s} b^{i_{1,\alpha}+\cdots+i_{s,\alpha}-\lfloor(\beta m-t)/\alpha\rfloor} & \text{if } \sum_{j=1}^{s}\sum_{l=1}^{\alpha} i_{j,l} > \beta m - t \\ & \text{and } \sum_{j=1}^{s} i_{j,\alpha} > \frac{\beta m - t}{\alpha}. \end{cases}$$

We estimate the sum (5.1) now. Let S_1 be the sum in (5.1) where $i_{1,1}+\cdots+i_{s,\alpha} > \beta m - t$ and $i_{1,\alpha}+\cdots+i_{s,\alpha} \le \frac{\beta m-t}{\alpha}$. For an $l > \beta m - t$ let $A_1(l)$ denote the number of admissible choices of $i_{1,1},\dots,i_{s,\alpha}$ such that $l = i_{1,1}+\cdots+i_{s,\alpha}$. Then we have

$$S_1 = (b-1)^{\alpha s} \sum_{l=\beta m-t+1}^{\alpha s m} \frac{A_1(l)}{b^l}.$$

We have $A_1(l) \le \binom{l+s\alpha-1}{s\alpha-1}$, and hence we obtain

$$S_1 \le (b-1)^{s\alpha} \sum_{l=\beta m-t+1}^{\infty} \binom{l+s\alpha-1}{s\alpha-1} \frac{1}{b^l} \le b^{s\alpha} b^{-\beta m+t-1} \binom{\beta m-t+s\alpha}{s\alpha-1},$$

where the last inequality follows from a result by Matoušek [13, Lemma 2.18]; see also [5, Lemma 6].

Let S_2 be the part of (5.1) for which $i_{1,1}+\cdots+i_{s,\alpha} > \beta m-t$ and $i_{1,\alpha}+\cdots+i_{s,\alpha} > \frac{\beta m-t}{\alpha}$; i.e., we have

$$\begin{aligned} S_2 &= (b-1)^{s\alpha} \sum_{i_{1,1}=1}^{m} \cdots \sum_{i_{1,\alpha}=1}^{i_{1,\alpha-1}-1} \cdots \sum_{i_{s,1}=1}^{m} \cdots \sum_{i_{s,\alpha}=1}^{i_{s,\alpha-1}-1} \frac{b^{-\lfloor(\beta m-t)/\alpha\rfloor}}{b^{i_{1,1}+\cdots+i_{1,\alpha-1}+\cdots+i_{s,1}+\cdots+i_{s,\alpha-1}}} \\ &\le \frac{m^s(b-1)^{s\alpha}}{b^{\lfloor(\beta m-t)/\alpha\rfloor}} \sum_{i_{1,1}=1}^{m} \cdots \sum_{i_{1,\alpha-1}=1}^{i_{1,\alpha-2}-1} \cdots \sum_{i_{s,1}=1}^{m} \cdots \sum_{i_{s,\alpha-1}=1}^{i_{s,\alpha-2}-1} \frac{1}{b^{i_{1,1}+\cdots+i_{1,\alpha-1}+\cdots+i_{s,1}+\cdots+i_{s,\alpha-1}}}, \end{aligned} \tag{5.2}$$

where in the first line above we have the additional conditions $i_{1,1}+\cdots+i_{s,\alpha} > \beta m-t$ and $i_{1,\alpha}+\cdots+i_{s,\alpha} > \frac{\beta m-t}{\alpha}$. From the last inequality and $i_{1,\alpha-l}+\cdots+i_{s,\alpha-l} > i_{1,\alpha}+\cdots+i_{s,\alpha}$ for $l = 1,\dots,\alpha-1$, it follows that $i_{1,1}+\cdots+i_{1,\alpha-1}+\cdots+i_{s,1}+\cdots+i_{s,\alpha-1} \ge \lfloor(\beta m-t)(1-\alpha^{-1})\rfloor + 1$. Let $A_2(l)$ denote the number of admissible choices of $i_{1,1},\dots,i_{1,\alpha-1},\dots,i_{s,1},\dots,i_{s,\alpha-1}$ such that $l = i_{1,1}+\cdots+i_{1,\alpha-1}+\cdots+i_{s,1}+\cdots+i_{s,\alpha-1}$. Note that we have $A_2(l) \le \binom{l+s(\alpha-1)-1}{s(\alpha-1)-1}$. Then we have

$$\begin{aligned} S_2 &\le \frac{m^s(b-1)^{s\alpha}}{b^{\lfloor(\beta m-t)/\alpha\rfloor}} \sum_{l=\lfloor(\beta m-t)(1-\alpha^{-1})\rfloor+1}^{\infty} \binom{l+s(\alpha-1)-1}{s(\alpha-1)-1} \frac{1}{b^l} \\ &\le \frac{m^s(b-1)^{s\alpha}}{b^{\lfloor(\beta m-t)/\alpha\rfloor}} \frac{b^{\lceil(\beta m-t)/\alpha\rceil}}{(1-b^{-1})^{s(\alpha-1)} b^{\beta m-t+1}} \binom{\lfloor(\beta m-t)(1-\alpha^{-1})\rfloor + s(\alpha-1)}{s(\alpha-1)-1}, \end{aligned}$$

where the last inequality follows again from a result by Matoušek [13, Lemma 2.18]; see also [5, Lemma 6]. Hence we have

$$S_2 \le m^s b^{s\alpha} b^{-\beta m+t} \binom{\lfloor(\beta m-t)(1-\alpha^{-1})\rfloor + s(\alpha-1)}{s(\alpha-1)-1}.$$

Note that we have $Q^*_{b,m,\alpha,\{1,\dots,s\}}(C_1,\dots,C_s) = S_1 + S_2$. Let $a \ge 1$ and $b \ge 0$ be integers; then we have

$$\binom{a+b}{b} = \prod_{i=1}^{b} \left(1+\frac{a}{i}\right) \le (1+a)^b.$$

Therefore we obtain $S_1 \le b^{s\alpha} b^{-\beta m+t-1}(\beta m-t+2)^{s\alpha-1}$ and $S_2 \le b^{s\alpha} b^{-\beta m+t} m^s (\beta m - t+2)^{s(\alpha-1)-1}$. Thus we have

$$Q^*_{b,m,\alpha,\{1,\dots,s\}}(C_1,\dots,C_s) \le 2b^{s\alpha} b^{-\beta m+t}(\beta m+2)^{s\alpha-1},$$

from which the result follows. □

The following theorem is an immediate consequence of Lemmas 2.10 and 5.1.

THEOREM 5.2. *Let b be prime, let $\alpha \ge 2$ be a natural number, and let $C_1,\dots,C_s \in \mathbb{Z}_b^{m\times m}$ be the generating matrices of a digital (t,α,β,m,s)-net over $\mathbb{Z}_b$, with $m > t/\beta$. Then the worst-case error in the Korobov space $\mathcal{H}_\alpha$ is bounded by*

$$e_{b,m,\alpha}(C_1,\dots,C_s) \le \frac{2\left(1+b^{-\alpha m}C_{b,\alpha}(\alpha+b^{-2})+C_{b,\alpha}(1+\alpha+b^{-2})(\beta m+2)^\alpha\right)^s}{b^{\beta m-t}(\beta m+2)} + (1+b^{-\alpha m}C_{b,\alpha}(\alpha+b^{-2}))^s - 1,$$

where $C_{b,\alpha} > 0$ is the constant in Lemma 2.9.

Remark 4. By the lower bound of Sharygin [24] we have that the worst-case error in the Korobov space $\mathcal{H}_\alpha$ is at most $\mathcal{O}(N^{-\alpha}(\log N)^{s-1})$. Hence it follows from Theorem 5.2 that for a digital (t,α,β,m,s)-net with $\beta > \alpha$ we must have $t = \mathcal{O}((\beta-\alpha)m)$. Thus in order to avoid having a t-value which grows with m, we added the restriction $\beta \le \alpha$ in Definition 3.1. Further, this also implies that a digital (t,α,β,s)-sequence with $t < \infty$ cannot exist if $\beta > \alpha$; hence, $\beta \le \alpha$ is in this case a consequence of the definition rather than a restriction.

Remark 5. Lemma 2.8 also holds for digital nets which are digitally shifted by an arbitrary digital shift $\boldsymbol{\sigma} \in [0,1)^s$, and hence it follows that Theorem 5.2 also holds in a more general form, namely, for all digital (t,α,β,m,s)-nets which are digitally shifted.

Theorem 5.2 shows that we can obtain the optimal convergence rate for natural numbers $\alpha \ge 2$ by using a digital (t,α,m,s)-net. The constructions previously proposed (for example, by Sobol, Faure, Niederreiter, or Niederreiter–Xing) have only been shown to be $(t,1,m,s)$-nets, and it has been proven that they achieve a convergence of the worst-case error of $\mathcal{O}(N^{-1}(\log N)^{s-1})$.

We can use Theorem 5.2 to obtain the following corollary.

COROLLARY 5.3. *Let b be prime, and let $C_1^{(d)},\dots,C_s^{(d)} \in \mathbb{Z}_b^{\infty\times\infty}$ be the generating matrices of a digital $(t(a),a,\min(a,d),s)$-sequence S over $\mathbb{Z}_b$ for any integer $a \ge 1$. Then for any real $\alpha \ge 1$ there is a constant $C'_{b,s,\alpha} > 0$, depending only on b,s, and α, such that the worst-case error in the Korobov space $\mathcal{H}_\alpha$ using the first $N = b^m$ points of S is bounded by*

$$e_{b,m,\alpha}(C_1^{(d)},\dots,C_s^{(d)}) \le C'_{b,s,\alpha} b^{t(\lfloor\alpha\rfloor)} \frac{(\log N)^{s\lfloor\alpha\rfloor-1}}{N^{\min(\lfloor\alpha\rfloor,d)}}.$$

Remark 6. The above corollary shows that digital $(t,\alpha,\min(\alpha,d),s)$-sequences constructed in section 4 achieve the optimal convergence (apart from maybe some $\log N$ factor) of $P_{2\alpha}$ of $\mathcal{O}(N^{-2\alpha}(\log N)^{2s\alpha-2})$ as long as α is an integer such that $1 \le \alpha \le d$. If $\alpha > d$, we obtain a convergence of $\mathcal{O}(N^{-2d}(\log N)^{2s\alpha-2})$.

6. A bound on the mean square worst-case error in $\mathcal{H}_\alpha$ for digital (t,α,β,m,s)-nets and digital (t,α,β,s)-sequences. To combine the advantages of random quadrature points with those of deterministic quadrature points, one

sometimes uses a combination of those two methods; see, for example, [5, 8, 13, 22]. The idea is to use a random element which preserves the essential properties of a deterministic point set. We call the expectation value of the square worst-case error of such randomized point sets the mean square worst-case error.

6.1. Randomization. In the following we introduce a randomization scheme called digital shift (see [4, 13]). Let $P_N = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}\} \subseteq [0,1)^s$, with $\boldsymbol{x}_n = (x_{1,n}, \dots, x_{s,n})$ and $x_{j,n} = x_{j,n,1}b^{-1} + x_{j,n,2}b^{-2} + \cdots$ for $n = 0, \dots, N-1$ and $j = 1, \dots, s$. Let $\sigma_{j,1}, \sigma_{j,2}, \ldots \in \{0,1\}$ be independently and identically distributed (i.i.d.) for $j = 1, \dots, s$. Then the randomly digitally shifted point set $P_{N,\boldsymbol{\sigma}} = \{\boldsymbol{z}_0, \dots, \boldsymbol{z}_{N-1}\}$, $\boldsymbol{z}_n = (z_{1,n}, \dots, z_{s,n})$ using a digital shift, is then given by

$$z_{j,n} = (x_{j,n,1} \oplus \sigma_{j,1})b^{-1} + (x_{j,n,2} \oplus \sigma_{j,2})b^{-2} + \cdots$$

for $j = 1, \dots, s$ and $n = 0, \dots, N-1$, where $x_{j,n,k} \oplus \sigma_{j,n} = x_{j,n,k} + \sigma_{j,n} \pmod b$ (note that all additions of the digits are carried out in the finite field $\mathbb{Z}_b$). Subsequently let $P_N = \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_{N-1}\}$, and let $P_{N,\boldsymbol{\sigma}}$ be the digitally shifted point set P_N using the randomization just described.

6.2. The mean square worst-case error in the Korobov space. In this section we will analyze the expectation value of $e^2(P_{N,\boldsymbol{\sigma}}, K_\alpha)$, which we denote by $\tilde{e}^2(P_N, K_\alpha) = \mathbb{E}[e^2(P_{N,\boldsymbol{\sigma}}, K_\alpha)]$, with respect to the random digital shift described above. We call $\tilde{e}^2(P_N, K_\alpha)$ the mean square worst-case error.

From (2.4) and the linearity of the expectation operator we have

$$\tilde{e}^2(P_N, K_\alpha) = \mathbb{E}[e^2(P_{N,\boldsymbol{\sigma}}, K_\alpha)] = -1 + \frac{1}{N^2} \sum_{n,l=0}^{N-1} \prod_{j=1}^{s} \mathbb{E}[K_\alpha(z_{j,n}, z_{j,l})].$$

In order to compute $\mathbb{E}[K_\alpha(z_{j,n}, z_{j,l})]$ we need the following lemma, which, in a very similar form, was already shown in [5, Lemma 3]. Hence we omit a proof.

LEMMA 6.1. *Let $x_1, x_2 \in [0,1)$, and let $z_1, z_2 \in [0,1)$ be the points obtained after applying an i.i.d. random digital shift to x_1 and x_2. Then we have*

$$\mathbb{E}[\mathrm{wal}_k(z_1)\overline{\mathrm{wal}_l(z_2)}] = \begin{cases} \mathrm{wal}_k(x_1)\overline{\mathrm{wal}_k(x_2)} & \text{if } k = l, \\ 0 & \text{otherwise.} \end{cases}$$

Recall that

$$K_\alpha(x_1, x_2) = \sum_{k,l=0}^{\infty} r_{b,\alpha}(k,l)\mathrm{wal}_k(x_1)\overline{\mathrm{wal}_l(x_2)},$$

where

$$r_{b,\alpha}(k,l) = \int_0^1 \int_0^1 K_\alpha(x_1, x_2)\overline{\mathrm{wal}_k(x_1)}\mathrm{wal}_l(x_2)\,\mathrm{d}x_1\,\mathrm{d}x_2.$$

Let z_1, z_2 be obtained by applying an i.i.d. random digital shift to x_1, x_2. Using Lemma 6.1 and the linearity of expectation we obtain

$$\mathbb{E}[K_\alpha(z_1, z_2)] = \sum_{k=0}^{\infty} r_{b,\alpha}(k)\mathrm{wal}_k(x_1)\overline{\mathrm{wal}_k(x_2)},$$

where $r_{b,\alpha}(k) = r_{b,\alpha}(k,k)$ and $r_{b,\alpha}(0) = 1$.

Therefore we obtain

$$\mathbb{E}[e^2(P_{N,\boldsymbol{\sigma}}, K_\alpha)] = -1 + \frac{1}{N^2} \sum_{n,l=0}^{N-1} \prod_{j=1}^{s} \sum_{k=0}^{\infty} r_{b,\alpha}(k) \mathrm{wal}_k(x_{j,n}) \overline{\mathrm{wal}_k(x_{j,l})}.$$

Further we have

$$\prod_{j=1}^{s} \sum_{k=0}^{b^m-1} r_{b,\alpha}(k) \mathrm{wal}_k(x_{j,n}) \overline{\mathrm{wal}_k(x_{j,l})} = 1 + \sum_{\boldsymbol{k} \in \{0,\ldots,b^m-1\}^s \setminus \{\boldsymbol{0}\}} r_{b,\alpha}(\boldsymbol{k}) \mathrm{wal}_k(\boldsymbol{x}_n \ominus \boldsymbol{x}_l),$$

where we write $r_{b,\alpha}(\boldsymbol{k}) = \prod_{j=1}^{s} r_{b,\alpha}(k_j)$ for $\boldsymbol{k} = (k_1, \ldots, k_s)$. We have shown the following theorem.

THEOREM 6.2. *Let $b \geq 2$ be a natural number, and let $\alpha > 1/2$ be a real number. Then the mean square worst-case error for integration in the Korobov space $\mathcal{H}_\alpha$ using the point set P_N randomized by a digital shift is given by*

$$\mathbb{E}[e^2(P_{N,\boldsymbol{\sigma}}, K_\alpha)] = \sum_{\boldsymbol{k} \in \mathbb{N}_0^s \setminus \{\boldsymbol{0}\}} r_{b,\alpha}(\boldsymbol{k}) \frac{1}{N^2} \sum_{n,l=0}^{N-1} \mathrm{wal}_k(\boldsymbol{x}_n \ominus \boldsymbol{x}_l).$$

In the following we closer investigate the mean square worst-case error for digital nets randomized with a digital shift.

Subsequently we will often write $\tilde{e}^2_{b,m,\alpha}(C_1, \ldots, C_s)$ to denote the mean square worst-case error $\mathbb{E}[e(P_{b^m,\boldsymbol{\sigma}}, K_\alpha)]$, where P_{b^m} is a digital net with generating matrices $C_1, \ldots, C_s$ and b^m points and $P_{b^m,\boldsymbol{\sigma}}$ is the digital net P_{b^m} randomized with a digital shift.

THEOREM 6.3. *Let $m \geq 1$, b be a prime number, and $\alpha > 1/2$ be a real number. The mean square worst-case error in the Korobov space $\mathcal{H}_\alpha$ using a randomly digitally shifted digital net over $\mathbb{Z}_b$ with generating matrices $C_1, \ldots, C_s \in \mathbb{Z}_b^{m \times m}$ is given by*

$$\tilde{e}^2_{b,m,\alpha}(C_1, \ldots, C_s) = \sum_{\boldsymbol{k} \in \mathcal{D}} r_{b,\alpha}(\boldsymbol{k}).$$

Proof. In [4] it was shown that

$$\frac{1}{b^{2m}} \sum_{n,l=0}^{b^m-1} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n \ominus \boldsymbol{x}_l) = \frac{1}{b^m} \sum_{n=0}^{b^m-1} \mathrm{wal}_{\boldsymbol{k}}(\boldsymbol{x}_n) = \begin{cases} 1 & \text{if } \boldsymbol{k} \in \mathcal{D} \cup \{\boldsymbol{0}\}, \\ 0 & \text{otherwise.} \end{cases}$$

Hence the result follows from Theorem 6.2. ∎

Remark 7. Theorems 2.5 and 6.3 now imply that

$$\tilde{e}_{b,m,\alpha}(C_1, \ldots, C_s) = \sqrt{\sum_{\boldsymbol{k} \in \mathcal{D}} r_{b,\alpha}(\boldsymbol{k})} \leq \sqrt{\sum_{\boldsymbol{k},\boldsymbol{l} \in \mathcal{D}} r_{b,\alpha}(\boldsymbol{k},\boldsymbol{l})} = e(P_{b^m}, K_\alpha);$$

i.e., the root mean square worst-case error is always smaller than the worst-case error; see also Remark 5.

Remark 5 and also the above Remark 7 imply that the bounds on the worst-case error also hold for the root mean square worst-case error. On the other hand, following the proofs for the bound on the worst-case error using the criterion for the root mean square worst-case error yields a better bound. We outline the results subsequently.

Following the proof of Lemma 2.10 we obtain

$$\sum_{\boldsymbol{k}\in\mathcal{D}} r_{b,\alpha}(\boldsymbol{k}) \le \sum_{\emptyset\ne u\subseteq\{1,\dots,s\}} (1+b^{-2\alpha m}\bar{C}_{b,\alpha}^2)^{s-|u|}(C_{b,\alpha}^2+\bar{C}_{b,\alpha}^2)^{|u|} \sum_{\boldsymbol{k}\in\mathcal{D}^*_{b^m,u}} q_{b,\alpha}^2(\boldsymbol{k}) + (1+b^{-2\alpha m}\bar{C}_{b,\alpha}^2)^s - 1,$$

where $C_{b,\alpha}$ is the constant from Lemma 2.9 and

$$\bar{C}_{b,\alpha} = C_{b,\alpha}\sqrt{b^{-1}+(b^2-b)^{-1}\prod_{c=3}^{\alpha+1}(b^{2c}-b^{2(c-1)})^{-1}}. \tag{6.1}$$

The sum $\sum_{\boldsymbol{k}\in\mathcal{D}^*_{b^m,u}} q_{b,\alpha}^2(\boldsymbol{k})$ can now be bounded using almost the same arguments as in the proof of Lemma 5.1. Doing this one can obtain that for a digital (t,α,β,m,s)-net we have

$$\sum_{\boldsymbol{k}\in\mathcal{D}^*_{b^m,u}} q_{b,\alpha}^2(\boldsymbol{k}) \le (2b)^{|u|\alpha} b^{-2(\beta m-t)+1}(\beta m-t+1)^{|u|\alpha-1}.$$

Hence we obtain the following theorem.

THEOREM 6.4. *Let b be prime, let $\alpha\ge 1$ be an integer, and let $C_1,\dots,C_s\in\mathbb{Z}_b^{m\times m}$ be the generating matrices of a digital (t,α,β,m,s)-net over $\mathbb{Z}_b$, with $m>t/\beta$. Then the mean square worst-case error in the Korobov space $\mathcal{H}_\alpha$ is bounded by*

$$\tilde{e}_{b,m,\alpha}^2(C_1,\dots,C_s) \le \frac{\left(1+b^{-2\alpha m}\bar{C}_{b,\alpha}^2+(2b)^\alpha(C_{b,\alpha}^2+\bar{C}_{b,\alpha}^2)(\beta m-t+1)^\alpha\right)^s-(1+b^{-2\alpha m}\bar{C}_{b,\alpha}^2)^s}{b^{2(\beta m-t)-1}(\beta m-t+1)} + (1+b^{-2\alpha m}\bar{C}_{b,\alpha}^2)^s-1,$$

where $C_{b,\alpha}>0$ is the constant in Lemma 2.9 and the constant $\bar{C}_{b,\alpha}>0$ is given by (6.1).

We can use Theorem 6.4 to obtain the following corollary.

COROLLARY 6.5. *Let b be prime, and let $C_1^{(d)},\dots,C_s^{(d)}\in\mathbb{Z}_b^{\infty\times\infty}$ be the generating matrices of a digital $(t(a),a,\min(a,d),s)$-sequence S over $\mathbb{Z}_b$ for any integer $a\ge 1$. Then for any real $\alpha\ge 1$ there is a constant $C''_{b,s,\alpha}>0$, depending only on b,s, and α, such that the root mean square worst-case error in the Korobov space $\mathcal{H}_\alpha$ using the first $N=b^m$ points of S is bounded by*

$$\tilde{e}_{b,m,\alpha}(C_1^{(d)},\dots,C_s^{(d)}) \le C''_{b,s,\alpha} b^{t(\lfloor\alpha\rfloor)}\frac{(\log N)^{(s\lfloor\alpha\rfloor-1)/2}}{N^{\min(\lfloor\alpha\rfloor,d)}}.$$

Remark 8. The above corollary shows that the digital $(t,\alpha,\min(\alpha,d),s)$-sequences constructed in section 4 achieve the optimal convergence of $P_{2\alpha}$ of $\mathcal{O}(N^{-2\alpha}(\log N)^{s\alpha-1})$ as long as α is an integer such that $1\le\alpha\le d$. (This convergence is the best possible for $\alpha=1$ by the lower bound in [24].) If $\alpha>d$, we obtain a convergence of $\mathcal{O}(N^{-2d}(\log N)^{s\alpha-1})$.

Using the construction of Theorem 4.1 or 4.2 it follows that $t(a)$ also depends on the choice of d. Hence choosing a large value of d also increases the constant factor $b^{t(\lfloor\alpha\rfloor)}$ in Corollaries 5.3 and 6.5.

7. Some examples of digital (t, α, m, s)-nets over $\mathbb{Z}_2$. In this section we give a simple example to show how the nets described in this paper can be constructed. We use the construction method outlined in section 4.

7.1. Example of a digital $(0, 2, m, 1)$-net over $\mathbb{Z}_2$. First we use the so-called Hammersley net as the underlying digital net, which is a $(0, m, 2)$-net over $\mathbb{Z}_2$. The generating matrices for this net are given by

$$C_1 = \begin{pmatrix} 1 & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & 1 \end{pmatrix} \text{ and } C_2 = \begin{pmatrix} 0 & \dots & 0 & 1 \\ \vdots & \iddots & \iddots & 0 \\ 0 & \iddots & \iddots & \vdots \\ 1 & 0 & \dots & 0 \end{pmatrix}. \tag{7.1}$$

Now we use the construction method of section 4 to construct the matrix $C_1^{(2)}$, i.e., $d = 2$ in this case. The first row of $C_1^{(2)}$ is the first row of C_1, the second row of $C_1^{(2)}$ is the first row of C_2, the third row of $C_1^{(2)}$ is the second row of C_1, the fourth row of $C_1^{(2)}$ is the second row of C_2, and so on. Assume that C_1, C_2 are $m \times m$ matrices, where m is even. Then we obtain

$$C_1^{(2)} = \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & 1 \\ 0 & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & 0 & 1 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & 1 & 0 & \dots & \dots & 0 \\ 0 & \dots & \dots & 0 & 1 & 0 & \dots & 0 \end{pmatrix}.$$

So, for example, if $m = 4$, we obtain

$$C_1^{(2)} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}. \tag{7.2}$$

The matrix $C_1^{(2)}$ is of course nonsingular, and therefore the point set one obtains consists of just equidistant points starting with 0.

Assume that m is even. Then the digital net which one obtains from $C_1^{(2)}$ is a digital $(0, 1, m, 1)$-net over $\mathbb{Z}_2$, and, at the same time, it is also a digital $(0, 2, m, 1)$-net. Note that using the bound from Theorem 4.1 we obtain a t-value of 1, but by closer investigation using Definition 3.1 one can see that the properties also hold for $t = 0$. Hence the t-value obtained from Theorem 4.1 is not necessarily strict even if the value of the underlying digital net is strict.

7.2. Example of a digital $(t, 2, 4, 2)$-net over $\mathbb{Z}_2$. Consider the digital $(1, 4, 4)$-net over $\mathbb{Z}_2$ with generating matrices given by C_1, C_2 above and

$$C_3 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \text{ and } C_4 = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

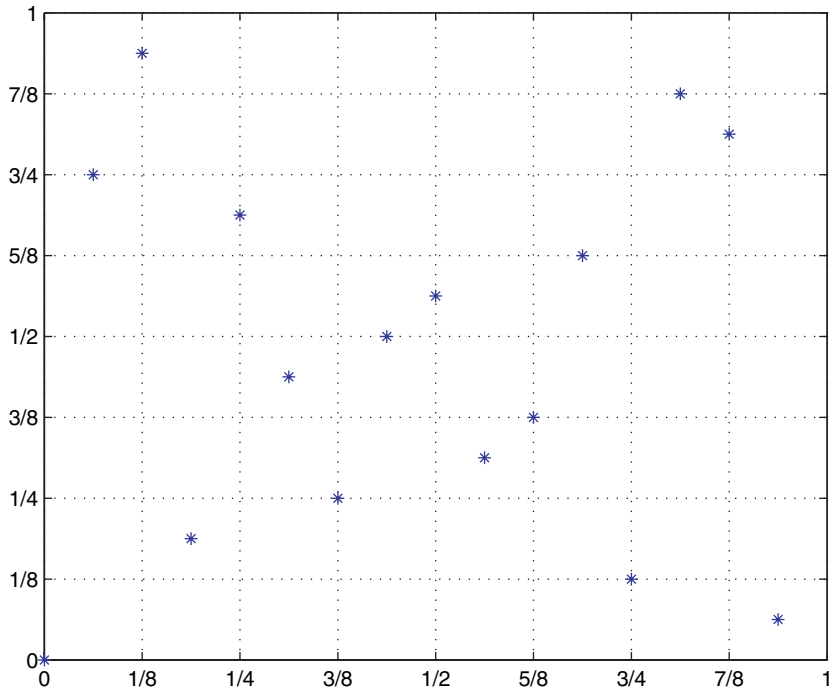


FIG. 7.1. *A digital* $(3,2,4,2)$*-net over* $\mathbb{Z}_2$.

Then $C_1^{(2)}$ is given by (7.2), and $C_2^{(2)}$ is given by

$$C_2^{(2)} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 \end{pmatrix}.$$

Using the digital construction scheme we obtain the points

$$(0,0),(\tfrac{1}{2},\tfrac{9}{16}),(\tfrac{1}{8},\tfrac{15}{16}),(\tfrac{5}{8},\tfrac{3}{8}),(\tfrac{1}{16},\tfrac{3}{4}),(\tfrac{9}{16},\tfrac{5}{16}),(\tfrac{3}{16},\tfrac{3}{16}),(\tfrac{11}{16},\tfrac{5}{8}),$$
$$(\tfrac{1}{4},\tfrac{11}{16}),(\tfrac{3}{4},\tfrac{1}{8}),(\tfrac{3}{8},\tfrac{1}{4}),(\tfrac{7}{8},\tfrac{13}{16}),(\tfrac{5}{16},\tfrac{7}{16}),(\tfrac{13}{16},\tfrac{7}{8}),(\tfrac{7}{16},\tfrac{1}{2}),(\tfrac{15}{16},\tfrac{1}{16}),$$

which are shown in Figure 7.1.

It can be checked that this digital net is a digital $(1,1,4,2)$-net, i.e., a digital $(1,4,2)$-net (the first two rows of $C_1^{(2)}$ and the first two rows of $C_2^{(2)}$ are linearly dependent, so the t-value cannot be 0 when $\alpha = 1$).

Now we investigate the t-value when $\alpha = 2$. First note that Theorem 4.1 yields a t-value of 4 for $\alpha = 2$ ($d = s = 2$). Further the t-value cannot be 2 in this case: We need to consider all cases where $i_{1,1} + i_{1,\min(\nu_1,2)} + i_{2,1} + i_{2,\min(\nu_2,2)} \le \alpha m - t = 2 \cdot 4 - 2 = 6$, with $0 \le \nu_1, \nu_2 \le 4$. But by choosing $i_{1,1} = i_{2,1} = 2$ and $i_{1,2} = i_{2,2} = 1$ we obtain the first two rows of C_1 and the first two rows of C_2, and as those 4 rows are linearly dependent it follows that the t-value cannot be 2. Now let us check whether a t-value of 3 is possible: We need to have $i_{1,1} + i_{1,\min(\nu_1,2)} + i_{2,1} + i_{2,\min(\nu_2,2)} \le 5$; hence, $\nu_1, \nu_2 \ge 2$ is not possible (because then we would have $i_{1,1} + i_{1,2} + i_{2,1} + i_{2,2} \ge 2+1+2+1 > 5$). Further the conditions are satisfied if either $\nu_1 = 0$ or $\nu_2 = 0$ as the matrices $C_1^{(2)}$

and $C_2^{(2)}$ are nonsingular. If $\nu_1 > 2$, then $i_{1,1} \geq 3$, $i_{1,2} \geq 2$, and hence $i_{1,1} + i_{1,2} \geq 5$, and we can only get $i_{1,1} + i_{1,\min(\nu_1,2)} + i_{2,1} + i_{2,\min(\nu_2,2)} \leq 5$ if $\nu_2 = 0$. Hence if either $\nu_1 > 2$ or $\nu_2 > 2$, the properties are also satisfied. Thus we are left with the following three cases: $(\nu_1, \nu_2) = (1, 1)$, $(\nu_1, \nu_2) = (1, 2)$, and $(\nu_1, \nu_2) = (2, 1)$.

Now let $\nu_1 = \nu_2 = 1$. Then we need to take one row of each matrix $C_1^{(2)}$ and $C_2^{(2)}$ such that the sum of their row indices is smaller or equal to 5 and check whether those two rows are linearly independent. It can be checked that this is always the case: Let $C_j^{(2)} = (c_{j,1}^\top, c_{j,2}^\top, c_{j,3}^\top, c_{j,4}^\top)$; i.e., $c_{j,k}$ denotes the kth row of $C_j^{(2)}$. Then the pairs of vectors $(c_{1,k}, c_{2,l})$, where $k + l \leq 5$, are always linearly independent for all admissible choices of k and l (i.e., $c_{1,k} \neq c_{2,l}$).

Consider now $\nu_1 = 1$ and $\nu_2 = 2$; i.e., we take one row from $C_1^{(2)}$ and two rows from $C_2^{(2)}$ such that the sum of the row indices does not exceed 5. Note that $i_{2,2}$ has to be 1; otherwise, $i_{2,1} + i_{2,2} \geq 5$ and $i_{1,1}$ cannot even be 1. As $i_{1,1} \geq 1$ and $i_{2,1} \geq 2$, the only choices left are $i_{1,1} = 1$, $i_{2,1} = 2, 3$ and $i_{1,1} = 2$, $i_{2,1} = 2$. So we need to check whether the triplets $(c_{1,1}, c_{2,1}, c_{2,2})$, $(c_{1,1}, c_{2,1}, c_{2,3})$, and $(c_{1,2}, c_{2,1}, c_{2,2})$ are all linearly independent, which upon inspection can be seen to be the case.

The case $\nu = 2$ and $\nu = 1$ can also be checked as the previous case. In this case all of the relevant sets of vectors are also always linearly independent; hence, a t-value of 3 is possible for $\alpha = 2$; i.e., the digital net above is a (strict) digital $(3, 2, 4, 2)$-net.

The classical t-value (i.e., $\alpha = 1$) of this digital net is not as good as, for example, the t-value of the Hammersley net (which is 0). On the other hand, it can be checked that for $\alpha = 2$ the t-value of the Hammersley net where $m = 4$ is 4, and hence for this case it is worse than the t-value of the digital net constructed above.

As a last example let us consider the Hammersley net again for arbitrary $m \geq 1$, i.e., with the $m \times m$ generating matrices given by (7.1). As, for example, the first row of C_1 and the last row of C_2 are the same (and therefore linearly dependent), we must have $\beta m - t < m + 1$ for all $\alpha \geq 1$ (for $\alpha = 1$ we can still choose $\beta = 1$ and $t = 0$, and hence the Hammersley net achieves the optimal t-value, but for $\alpha > 1$ we have seen in section 4 that there are better constructions). It is sensible to choose β such that we can have a t-value which is independent of m (for example, this is the case when one considers sequences and which is also the motivation for introducing those parameters; for digital nets it would of course also make sense to just state the value of $\beta m - t$ and m instead of t, β, and m). This means that $\beta \leq 1$, and as β indicates the convergence rate one can obtain, it follows that one cannot expect to obtain a convergence rate beyond $(b^m)^{-1+\delta}$ (for an arbitrary small $\delta > 0$) when using a Hammersley net.

8. Appendix: Some lemmas. We need the following lemmas.

LEMMA 8.1. *Let $j \geq 1$, $a \geq 0$, $b \geq 2$, and $0 \leq u, v < b^a$, with $u \neq v$. Then we have*

$$\int_{u/b^a}^{(u+1)/b^a} \int_{u/b^a}^{(u+1)/b^a} |x - y|^j \, \mathrm{d}x \, \mathrm{d}y = \frac{2}{b^{a(j+2)}(j+1)(j+2)}$$

and

$$\int_{u/b^a}^{(u+1)/b^a} \int_{v/b^a}^{(v+1)/b^a} |x - y|^j \, \mathrm{d}x \, \mathrm{d}y = \frac{2j!}{b^{a(j+2)}} \sum_{l=0}^{\lfloor j/2 \rfloor} \frac{|u - v|^{j-2l}}{(j-2l)!(2l+2)!}.$$

Proof. We have

$$\int_{u/b^a}^{(u+1)/b^a}\int_{u/b^a}^{(u+1)/b^a} |x-y|^j \,\mathrm{d}x\,\mathrm{d}y = \int_0^{1/b^a}\int_0^{1/b^a} |x-y|^j \,\mathrm{d}x\,\mathrm{d}y$$
$$= \frac{1}{b^{a(j+2)}}\int_0^1\int_0^1 |x-y|^j \,\mathrm{d}x\,\mathrm{d}y.$$

We divide the last double integral in two parts, and we have

$$\int_0^1\int_0^1 |x-y|^j \,\mathrm{d}x\,\mathrm{d}y = \int_0^1\int_0^y (y-x)^j \,\mathrm{d}x\,\mathrm{d}y + \int_0^1\int_y^1 (x-y)^j \,\mathrm{d}x\,\mathrm{d}y.$$

We calculate the first part and obtain

$$\int_0^1\int_0^y (y-x)^j \,\mathrm{d}x\,\mathrm{d}y = \frac{1}{j+1}\int_0^1 y^{j+1}\,\mathrm{d}y = \frac{1}{(j+1)(j+2)},$$

and the second part is given by

$$\int_0^1\int_y^1 (x-y)^j \,\mathrm{d}x\,\mathrm{d}y = \frac{1}{j+1}\int_0^1 (1-y)^{j+1}\,\mathrm{d}y = \frac{1}{(j+1)(j+2)}.$$

Hence we have

$$\int_0^1\int_0^1 |x-y|^j \,\mathrm{d}x\,\mathrm{d}y = \frac{2}{(j+1)(j+2)}.$$

For the second part we have

$$\int_{u/b^a}^{(u+1)/b^a}\int_{v/b^a}^{(v+1)/b^a} |x-y|^j \,\mathrm{d}x\,\mathrm{d}y = \int_0^{1/b^a}\int_{|u-v|/b^a}^{(|u-v|+1)/b^a} |x-y|^j \,\mathrm{d}x\,\mathrm{d}y$$
$$= \frac{1}{b^{a(j+2)}}\int_0^1\int_{|u-v|}^{|u-v|+1} (x-y)^j \,\mathrm{d}x\,\mathrm{d}y,$$

where now $|u-v|\geq 1$. We have

$$\int_0^1\int_{|u-v|}^{|u-v|+1} (x-y)^j \,\mathrm{d}x\,\mathrm{d}y = \frac{1}{j+1}\int_0^1 \left((|u-v|+1-y)^{j+1} - (|u-v|-y)^{j+1}\right)\,\mathrm{d}y$$
$$= \frac{2|u-v|^{j+2} - (|u-v|+1)^{j+2} - (|u-v|-1)^{j+2}}{(j+1)(j+2)}.$$

The result follows by simplifying the sum in the numerator. □

LEMMA 8.2. *Let $k\geq 1$ be given by $k=\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}$ for some $\nu\geq 1$, $\kappa_{a_1-1},\ldots,\kappa_{a_\nu-1}\in\{1,\ldots,b-1\}$, and $1\leq a_\nu<\cdots<a_1$. For any even $0\leq j<2\nu$ we have $I_j(k)=0$.*

Proof. The result for $j=0$ follows from Proposition 2.3 and (2.6). It was shown in [4, Appendix A] that

$$x = \frac{1}{2} + \sum_{c=1}^{\infty}\sum_{\tau=1}^{b-1} \frac{1}{b^c(\mathrm{e}^{-2\pi\mathrm{i}\tau/b}-1)}\mathrm{wal}_{\tau b^{c-1}}(x),$$

and hence

$$|x-y|^j = \left(\sum_{c=1}^{\infty}\sum_{\tau=1}^{b-1}\frac{1}{b^c(\mathrm{e}^{-2\pi\mathrm{i}\tau/b}-1)}(\mathrm{wal}_{\tau b^{c-1}}(y)-\mathrm{wal}_{\tau b^{c-1}}(x))\right)^j$$
$$= \sum_{c_1,\ldots,c_j=1}^{\infty}\frac{1}{b^{c_1+\cdots+c_j}}\prod_{i=1}^{j}\sum_{\tau=1}^{b-1}\frac{\mathrm{wal}_{\tau b^{c_i-1}}(y)-\mathrm{wal}_{\tau b^{c_i-1}}(x)}{\mathrm{e}^{-2\pi\mathrm{i}\tau/b}-1}.$$

Let

$$A_k(c_1,\ldots,c_j)=\int_0^1\int_0^1\prod_{i=1}^{j}\sum_{\tau=1}^{b-1}\frac{\mathrm{wal}_{\tau b^{c_i-1}}(y)-\mathrm{wal}_{\tau b^{c_i-1}}(x)}{\mathrm{e}^{-2\pi\mathrm{i}\tau/b}-1}\overline{\mathrm{wal}_k(x)}\mathrm{wal}_k(y)\,\mathrm{d}x\,\mathrm{d}y.$$

Then we have

$$I_j(k)=\sum_{c_1,\ldots,c_j=1}^{\infty}\frac{A_k(c_1,\ldots,c_j)}{b^{c_1+\cdots+c_j}}.$$

We have

$$\prod_{i=1}^{j}\sum_{\tau=1}^{b-1}\frac{\mathrm{wal}_{\tau b^{c_i-1}}(y)-\mathrm{wal}_{\tau b^{c_i-1}}(x)}{\mathrm{e}^{-2\pi\mathrm{i}\tau/b}-1}$$
$$=\sum_{\tau_1,\ldots,\tau_j=1}^{b-1}\prod_{i=1}^{j}(\mathrm{e}^{-2\pi\mathrm{i}\tau_j/b}-1)^{-1}\sum_{u\subseteq\{1,\ldots,j\}}(-1)^{|u|}\prod_{i\in u}\mathrm{wal}_{\tau_i b^{c_i-1}}(y)\prod_{i\notin u}\mathrm{wal}_{\tau_i b^{c_i-1}}(x)$$
$$=\sum_{\tau_1,\ldots,\tau_j=1}^{b-1}\prod_{i=1}^{j}(\mathrm{e}^{-2\pi\mathrm{i}\tau_j/b}-1)^{-1}\sum_{u\subseteq\{1,\ldots,j\}}(-1)^{|u|}\mathrm{wal}_{C_{u,\tau}}(y)\mathrm{wal}_{C_{\{1,\ldots,j\}\setminus u,\tau}}(x),$$

where $C_{u,\tau}=\sum_{i\in u}\tau_i b^{c_i-1}$, and hence

$$A_k(c_1,\ldots,c_j)$$
$$=\sum_{\tau_1,\ldots,\tau_j=1}^{b-1}\prod_{i=1}^{j}(\mathrm{e}^{-2\pi\mathrm{i}\tau_j/b}-1)^{-1}\sum_{u\subseteq\{1,\ldots,j\}}(-1)^{|u|}$$
$$\int_0^1\int_0^1\mathrm{wal}_{C_{u,\tau}}(y)\mathrm{wal}_{C_{\{1,\ldots,j\}\setminus u,\tau}}(x)\overline{\mathrm{wal}_k(x)}\mathrm{wal}_k(y)\,\mathrm{d}x\,\mathrm{d}y$$
$$=\sum_{\tau_1,\ldots,\tau_j=1}^{b-1}\prod_{i=1}^{j}(\mathrm{e}^{-2\pi\mathrm{i}\tau_j/b}-1)^{-1}$$
$$\sum_{u\subseteq\{1,\ldots,j\}}(-1)^{|u|}\int_0^1\mathrm{wal}_{C_{u,\tau}\oplus k}(y)\,\mathrm{d}y\int_0^1\mathrm{wal}_{C_{\{1,\ldots,j\}\setminus u,\tau}\ominus k}(x)\,\mathrm{d}x.$$

Note that if $\nu > j/2$, we have either $C_{u,\tau}\oplus k\neq 0$ or $C_{\{1,\ldots,j\}\setminus u,\tau}\ominus k\neq 0$, and hence $A_k(c_1,\ldots,c_j)=0$. The result now follows. □

Let $\sigma_p(n) = \sum_{h=1}^{n-1} h^p$. It is known that

$$\sigma_p(n) = \sum_{h=0}^{p} \frac{B_h}{h!} \frac{p!}{(p+1-h)!} n^{p+1-h}, \tag{8.1}$$

where $B_0, B_1, \ldots$ are the Bernoulli numbers (in particular, $B_0 = 1$, $B_1 = -1/2$, and $B_2 = 1/6$).

LEMMA 8.3. *Let $b \geq 2$, $1 \leq d \leq a$, and $k = \kappa_{d-1} b^{d-1} + \cdots + \kappa_0$, where $\kappa_{d-1} \in \{1, \ldots, b-1\}$, $\kappa_{d-2}, \ldots, \kappa_0 \in \{0, \ldots, b-1\}$, $m = m_{a-1} b^{a-1} + \cdots + m_0$, and $n = n_{a-1} b^{a-1} + \cdots + n_0$. Then we have*

$$\sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} \mathrm{wal}_k((n \ominus m)/b^a) = b^{2a-d} \left(\frac{1}{2} + \frac{1}{\mathrm{e}^{2\pi \mathrm{i} \kappa_{d-1}/b} - 1} \right) - \frac{b^a}{2},$$

$$\sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n) \mathrm{wal}_k((n \ominus m)/b^a) = b^{3a-2d} \left(\frac{1}{6} - \frac{1}{2 \sin^2(\kappa_{d-1} \pi/b)} \right) - \frac{b^a}{6},$$

and

$$\sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n) = \frac{1}{6}(b^{3a} - b^a).$$

Proof. In order to obtain a formula for the first sum, let $m' = m_{a-1} b^{a-1} + \cdots + m_{a-d+1} b^{a-d+1}$, $m'' = m_{a-d-1} b^{a-d-1} + \cdots + m_0$, $n' = n_{a-1} b^{a-1} + \cdots + n_{a-d+1} b^{a-d+1}$, and $n'' = n_{a-d-1} b^{a-d-1} + \cdots + n_0$. First consider the case where $m' > n'$ and arbitrary m'', n''. We have

$$\sum_{n_{a-d}=0}^{b-1} \sum_{m_{a-d}=0}^{b-1} \mathrm{e}^{2\pi \mathrm{i}(\kappa_0(n_{a-1}-m_{a-1}) + \cdots + \kappa_{d-1}(n_{a-d}-m_{a-d}))/b} = 0,$$

as $\sum_{m=0}^{b-1} \mathrm{e}^{2\pi \mathrm{i} \kappa m/b} = 0$ for all $\kappa = 1, \ldots, b-1$. Thus we only need to consider the case where $m' = n'$, for which case we have

$$\mathrm{e}^{2\pi \mathrm{i}(\kappa_0(n_{a-1}-m_{a-1}) + \cdots + \kappa_{d-1}(n_{a-d}-m_{a-d}))/b} = \mathrm{e}^{2\pi \mathrm{i} \kappa_{d-1}(m_{a-d}-n_{a-d})/b}.$$

This part is now given by

$$b^{d-1} \sum_{n''=0}^{b^{a-d}-1} \sum_{m''=0}^{b^{a-d}-1} \sum_{n_{a-d}=0}^{b-1} \sum_{m_{a-d}=0}^{b-1} \mathrm{e}^{2\pi \mathrm{i} \kappa_{d-1}(n_{a-d}-m_{a-d})/b}, \tag{8.2}$$

where we have the additional assumption $m_{a-d} b^{a-d} + m'' > n_{a-d} b^{a-d} + n''$. First consider the case where $m_{a-d} > n_{a-d}$. This part of (8.2) is given by

$$\begin{aligned} b^{d-1} \sum_{n''=0}^{b^{a-d}-1} \sum_{m''=0}^{b^{a-d}-1} \sum_{n_{a-d}=0}^{b-2} \sum_{m_{a-d}=n_{a-d}+1}^{b-1} & \mathrm{e}^{2\pi \mathrm{i} \kappa_{d-1}(n_{a-d}-m_{a-d})/b} \\ &= b^{d-1} b^{2(a-d)} \sum_{n_{a-d}=0}^{b-2} \sum_{m_{a-d}=n_{a-d}+1}^{b-1} \mathrm{e}^{2\pi \mathrm{i} \kappa_{d-1}(n_{a-d}-m_{a-d})/b} \\ &= \frac{b^{2a-d}}{\mathrm{e}^{2\pi \mathrm{i} \kappa_{d-1}/b} - 1}. \end{aligned}$$

Now consider the case where $m_{a-d} = n_{a-d}$. In this case we have the assumption that $m'' > n''$, and hence this part of (8.2) is given by

$$b^d \sum_{n''=0}^{b^{a-d}-2} \sum_{m''=n''+1}^{b^{a-d}-1} 1 = \frac{1}{2}\left(b^{2a-d} - b^a\right).$$

Thus (8.2) is given by

$$\frac{b^{2a-d}}{\mathrm{e}^{2\pi \mathrm{i}\kappa_{d-1}/b} - 1} + \frac{1}{2}\left(b^{2a-d} - b^a\right),$$

and the first result follows.

For the second sum let again $m' = m_{a-1}b^{a-1} + \cdots + m_{a-d+1}b^{a-d+1}$, $m'' = m_{a-d-1}b^{a-d-1} + \cdots + m_0$, $n' = n_{a-1}b^{a-1} + \cdots + n_{a-d+1}b^{a-d+1}$, and also $n'' = n_{a-d-1}b^{a-d-1} + \cdots + n_0$. First consider the case where $m' > n'$ and arbitrary m'', n''. We have

$$\begin{aligned}
&\sum_{n_{a-d}=0}^{b-1} \sum_{m_{a-d}=0}^{b-1} (m-n)\mathrm{e}^{2\pi \mathrm{i}(\kappa_0(n_{a-1}-m_{a-1})+\cdots+\kappa_{d-1}(n_{a-d}-m_{a-d}))/b} \\
&= \sum_{n_{a-d}=0}^{b-1} \sum_{m_{a-d}=0}^{b-1} (m_{a-d}-n_{a-d})\mathrm{e}^{2\pi \mathrm{i}(\kappa_0(n_{a-1}-m_{a-1})+\cdots+\kappa_{d-1}(n_{a-d}-m_{a-d}))/b} \\
&= 0,
\end{aligned}$$

as $\sum_{m=0}^{b-1} \mathrm{e}^{2\pi \mathrm{i}\kappa m/b} = 0$ for all $\kappa = 1, \ldots, b-1$.

Thus we are left with the case where $m' = n'$. We have

$$\mathrm{e}^{2\pi \mathrm{i}(\kappa_0(n_{a-1}-m_{a-1})+\cdots+\kappa_{d-1}(n_{a-d}-m_{a-d}))/b} = \mathrm{e}^{2\pi \mathrm{i}\kappa_{d-1}(m_{a-d}-n_{a-d})/b}.$$

Hence this part is given by

(8.3)

$$b^{d-1} \sum_{n_{a-d}=0}^{b-1} \sum_{m_{a-d}=0}^{b-1} \sum_{n''=0}^{b^{a-d}-1} \sum_{m''=0}^{b^{a-d}-1} (m''-n''+b^{a-d}(m_{a-d}-n_{a-d}))\mathrm{e}^{2\pi \mathrm{i}\kappa_{d-1}(n_{a-d}-m_{a-d})/b},$$

where we have the additional assumption $m_{a-d}b^{a-d} + m'' > n_{a-d}b^{a-d} + n''$. First consider the case where $m_{a-d} > n_{a-d}$. This part of (8.3) is given by

$$\begin{aligned}
&b^{d-1} \sum_{0 \le n_{a-d} < m_{a-d} < b} \sum_{m'',n''=0}^{b^{a-d}-1} (m''-n''+b^{a-d}(m_{a-d}-n_{a-d}))\mathrm{e}^{2\pi \mathrm{i}\kappa_{d-1}(n_{a-d}-m_{a-d})/b} \\
&\qquad = b^{d-1}b^{3(a-d)} \sum_{n_{a-d}=0}^{b-2} \sum_{m_{a-d}=n_{a-d}+1}^{b-1} (m_{a-d}-n_{a-d})\mathrm{e}^{2\pi \mathrm{i}\kappa_{d-1}(n_{a-d}-m_{a-d})/b} \\
&\qquad = -\frac{b^{3a-2d}}{2\sin^2(\kappa_{d-1}\pi/b)}.
\end{aligned}$$

Now consider the case where $m_{a-d} = n_{a-d}$. In this case we have the assumption that $m'' > n''$, and hence this part of (8.3) is given by

$$b^d \sum_{n''=0}^{b^{a-d}-2} \sum_{m''=n''+1}^{b^{a-d}-1} (m'' - n'') = \frac{b^d}{6}(b^{3(a-d)} - b^{a-d}).$$

(This result can be obtained using (8.1); see the proof of the third part below.) Thus (8.3) is given by

$$-\frac{b^{3a-2d}}{2\sin^2(\kappa_{d-1}\pi/b)} + \frac{b^d}{6}(b^{3(a-d)} - b^{a-d}),$$

and the second result follows.

The third result can easily be verified by using (8.1). Indeed we have

$$\sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n) = \sum_{n=0}^{b^a-2} \sum_{m=1}^{b^a-1-n} m = \sum_{n=0}^{b^a-2} \sigma_1(b^a - n) = \frac{1}{2} \sum_{n=0}^{b^a-2} ((b^a-n)^2 - (b^a-n)).$$

The last sum can be written as $\frac{1}{2}\sum_{n=1}^{b^a}(n^2-n) = \frac{1}{2}(\sigma_2(b^a+1) - \sigma_1(b^a+1))$, and by using (8.1) again the result follows. □

LEMMA 8.4. *Let $j \geq 0$, $\nu \geq 1$, $1 \leq a_\nu < \cdots < a_1 \leq a$, and $k = \kappa_{a_1-1}b^{a_1-1} + \cdots + \kappa_{a_\nu-1}b^{a_\nu-1}$, where $\kappa_{a_1-1}, \ldots, \kappa_{a_\nu-1} \in \{1, \ldots, b-1\}$. Then we have*

$$\sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n)^j = b^a \sigma_j(b^a) - \sigma_{j+1}(b^a) \leq \frac{b^{a(j+2)}}{(j+1)(j+2)}$$

and

$$\left| \sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n)^j \operatorname{wal}_k((n \ominus m)/b^a) \right| \leq C_{b,j} b^{(j+2)a - 2(a_1+\cdots+a_{\min(\nu, \lceil j/2 \rceil)})}$$

for some constant $C_{b,j} > 0$ which is independent of ν, a, and $a_1, \ldots, a_\nu$.

Proof. We have

$$\sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n)^j = \sum_{n=1}^{b^a-1} (b^a - n) n^j = b^a \sigma_j(b^a) - \sigma_{j+1}(b^a),$$

and by using (8.1) it follows that

$$\begin{aligned} &b^a \sigma_j(b^a) - \sigma_{j+1}(b^a) \\ &= b^{a(j+2)} \left(\sum_{h=0}^{j} B_h \left(\frac{j!}{h!(j+1-h)!} - \frac{(j+1)!}{h!(j+2-h)!} \right) b^{-ah} - B_{j+1} b^{-a(j+1)} \right) \\ &\leq b^{a(j+2)} B_0 \frac{1}{(j+1)(j+2)}, \end{aligned}$$

from which the first part follows as $B_0 = 1$.

For $j = 0, 1$ the second part immediately follows from Lemma 8.3. Let now $j \geq 2$, and assume the result holds for all $j-1, \ldots, 1, 0$.

Let $m = m_{a-1}b^{a-1} + \cdots + m_0$ and $n = n_{a-1}b^{a-1} + \cdots + n_0$. In order to obtain a bound on

$$\left| \sum_{n=0}^{b^a-2} \sum_{m=n+1}^{b^a-1} (m-n)^j \mathrm{e}^{2\pi \mathrm{i}(\kappa_{a_1-1}(n_{a-a_1}-m_{a-a_1})+\cdots+\kappa_{a_\nu-1}(n_{a-a_\nu}-m_{a-a_\nu}))/b} \right|, \tag{8.4}$$

we first sum over the digits m_{a-a_1} and n_{a-a_1}.

Let $m' = m_{a-1}b^{a-1}+\cdots+m_{a-a_1+1}b^{a-a_1+1}$, $n' = n_{a-1}b^{a-1}+\cdots+n_{a-a_1+1}b^{a-a_1+1}$, $m'' = m_{a-a_1-1}b^{a-a_1-1} + \cdots + m_0$, and $n'' = n_{a-a_1-1}b^{a-a_1-1} + \cdots + n_0$. We consider two cases, namely, where $m' > n'$ and where $m' = n'$.

For $m' = n'$ we have either $m_{a-a_1} > n_{a-a_1}$ or $m_{a-a_1} = n_{a-a_1}$ and $m'' > n''$, as $m > n$. First let $m_{a-a_1} > n_{a-a_1}$. We have b^{a_1-1} choices for $m' = n'$, and the sum over the digits m_{a-a_1}, n_{a-a_1} with $m_{a-a_1} > n_{a-a_1}$ can be written as one sum so that the part of (8.4) where $m' = n'$ is given by

$$\begin{aligned}
&b^{a_1-1} \left| \sum_{n''=0}^{b^{a-a_1}-1} \sum_{m''=0}^{b^{a-a_1}-1} \sum_{\tau=1}^{b-1} (b-\tau)(\tau b^{a-a_1} + m'' - n'')^j \mathrm{e}^{-2\pi \mathrm{i}\kappa_{a_1-1}\tau/b} \right| \\
&\leq b^{a_1-1} \sum_{n''=0}^{b^{a-a_1}-1} \sum_{m''=0}^{b^{a-a_1}-1} \sum_{\tau=1}^{b-1} (b-\tau)(\tau b^{a-a_1} + m'' - n'')^j \\
&\leq C''_{b,j} b^{a_1} b^{(j+2)(a-a_1)},
\end{aligned}$$

for some constant $C''_{b,j} > 0$ which depends only on b and j. Hence this part satisfies the bound. Now let $m_{a-a_1} = n_{a-a_1}$; then we have $m'' > n''$, and hence the part of (8.4) where $m' = n'$ and $m_{a-a_1} = n_{a-a_1}$ is given by

$$b^{a_1} \sum_{n''=0}^{b^{a-a_1}-1} \sum_{m''=n''+1}^{b^{a-a_1}-1} (m''-n'')^j \leq \frac{b^{a_1} b^{(j+2)(a-a_1)}}{(j+1)(j+2)},$$

where the inequality was already obtained in the first part of this proof. Hence also this part satisfies the bound.

Now we consider the part of (8.4) where $m' > n'$. We have

$$\begin{aligned}
&\sum_{m_{a-a_1}, n_{a-a_1}=0}^{b-1} (m'-n'+b^{a-a_1}(m_{a-a_1}-n_{a-a_1})+m''-n'')^j \mathrm{e}^{2\pi \mathrm{i}\kappa_{a_1-1}(n_{a-a_1}-m_{a-a_1})/b} \\
&= b(m'-n'+m''-n'')^j + \sum_{\tau=1}^{b-1} (b-\tau)[\mathrm{e}^{-2\pi \mathrm{i}\kappa_{a_1-1}\tau/b}(m'-n'+\tau b^{a-a_1}+m''-n'')^j \\
&\quad + \mathrm{e}^{2\pi \mathrm{i}\kappa_{a_1-1}\tau/b}(m'-n'-\tau b^{a-a_1}+m''-n'')^j] \\
&= b(m'-n'+m''-n'')^j + \sum_{u=0}^{j} \binom{j}{u} (m'-n'+m''-n'')^{j-u} b^{u(a-a_1)} E_u,
\end{aligned} \tag{8.5}$$

where

$$E_u = \sum_{\tau=1}^{b-1} (b-\tau)[\mathrm{e}^{-2\pi \mathrm{i}\kappa_{a_1-1}\tau/b}\tau^u + \mathrm{e}^{2\pi \mathrm{i}\kappa_{a_1-1}\tau/b}(-\tau)^u].$$

It can be checked that $E_0 = -b$ and $E_1 = 0$. Hence (8.5) is given by

$$\sum_{u=2}^{j} \binom{j}{u} (m' - n' + m'' - n'')^{j-u} b^{u(a-a_1)} E_u,$$

and hence the result follows from the induction assumption or the first part. □

LEMMA 8.5. *Let $k \ge 1$ be given by $k = \kappa_{a_1-1} b^{a_1-1} + \cdots + \kappa_{a_\nu-1} b^{a_\nu-1}$ for some $\nu \ge 1$, $1 \le a_\nu < \cdots < a_1$, and $\kappa_{a_1-1}, \ldots, \kappa_{a_\nu-1} \in \{1, \ldots, b-1\}$. Then for $j \ge 1$ we have*

$$|I_j(k)| \le \frac{\bar{C}_{b,j}}{b^{2(a_1+\cdots+a_{\min(\nu,\lceil j/2 \rceil)})}}$$

for some constant $\bar{C}_{b,j} > 0$ which depends only on b and j.

Proof. Let $k = \kappa_{a-1} b^{a-1} + \cdots + \kappa_0$, where now $a = a_1$, $u = u_{a-1} b^{a-1} + \cdots + u_0$, and $v = v_{a-1} b^{a-1} + \cdots + v_0$. Then we have

$$\begin{aligned} I_j(k) &= \int_0^1 \int_0^1 |x-y|^j \overline{\mathrm{wal}_k(x)} \mathrm{wal}_k(y) \,\mathrm{d}x \,\mathrm{d}y \\ &= \sum_{u=0}^{b^a-1} \sum_{v=0}^{b^a-1} \mathrm{e}^{2\pi \mathtt{i}(\kappa_0(u_{a-1}-v_{a-1})+\cdots+\kappa_{a-1}(u_0-v_0))} \int_{u/b^a}^{(u+1)/b^a} \int_{v/b^a}^{(v+1)/b^a} |x-y|^j \,\mathrm{d}x \,\mathrm{d}y. \end{aligned}$$

For $u = v$ we have $\mathrm{e}^{2\pi \mathtt{i}(\kappa_0(u_{a-1}-v_{a-1})+\cdots+\kappa_{a-1}(u_0-v_0))} = 1$. Using Lemma 8.1 it follows that this part in the above sum is given by

$$\frac{2}{b^{a(j+1)}(j+1)(j+2)}.$$

Hence it remains to calculate

$$\begin{aligned} &\sum_{u=0}^{b^a-1} \sum_{\substack{v=0 \\ u \ne v}}^{b^a-1} \mathrm{e}^{2\pi \mathtt{i}(\kappa_0(u_{a-1}-v_{a-1})+\cdots+\kappa_{a-1}(u_0-v_0))} \int_{u/b^a}^{(u+1)/b^a} \int_{v/b^a}^{(v+1)/b^a} |x-y|^j \,\mathrm{d}x \,\mathrm{d}y \\ &= 2 \sum_{u=0}^{b^a-2} \sum_{v=u+1}^{b^a-1} \mathrm{e}^{2\pi \mathtt{i}(\kappa_0(u_{a-1}-v_{a-1})+\cdots+\kappa_{a-1}(u_0-v_0))} \frac{2j!}{b^{a(j+2)}} \sum_{i=0}^{\lfloor j/2 \rfloor} \frac{|u-v|^{j-2i}}{(j-2i)!(2i+2)!} \\ &= \frac{4j!}{b^{a(j+2)}} \sum_{i=0}^{\lfloor j/2 \rfloor} \frac{1}{(j-2i)!(2i+2)!} \sum_{u=0}^{b^a-2} \sum_{v=u+1}^{b^a-1} \frac{\mathrm{e}^{2\pi \mathtt{i}(\kappa_0(u_{a-1}-v_{a-1})+\cdots+\kappa_{a-1}(u_0-v_0))}}{(v-u)^{2i-j}}, \end{aligned}$$

where we used Lemma 8.1. The absolute value of the inner double sum can now be bounded using Lemma 8.4, and hence the result follows. □

LEMMA 8.6. *Let $b \ge 2$ be an integer, and let $\alpha > 1/2$ be a real number. Then we have*

$$\sum_{k=1}^{\infty} r_{b,\alpha}(k) = 2\zeta(2\alpha),$$

where $\zeta(2\alpha) = \sum_{h=1}^{\infty} h^{-2\alpha}$.

Proof. Let $h \in \mathbb{Z} \setminus \{0\}$, and let $f_h(x) = \mathrm{e}^{2\pi \mathrm{i} h x}$. The Walsh coefficients $\hat{f}_h(k)$ of the function f_h are then given by $\hat{f}_h(k) = \int_0^1 f_h(x)\overline{\mathrm{wal}_k(x)}\,\mathrm{d}x$. It follows that $|\hat{f}_h(k)|^2 = |\beta_{h,k}|^2$, where $\beta_{h,k}$ was defined in Lemma 2.6. Using Parseval's equality we obtain

$$\sum_{k=1}^{\infty} |\beta_{h,k}|^2 = \sum_{k=1}^{\infty} |\hat{f}_h(k)|^2 = \int_0^1 |f_h(x)|^2\,\mathrm{d}x = \int_0^1 1\,\mathrm{d}x = 1.$$

Hence we have

$$\sum_{k=1}^{\infty} r_{b,\alpha}(k) = \sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{1}{|h|^{2\alpha}} \sum_{k=1}^{\infty} |\beta_{h,k}|^2 = \sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{1}{|h|^{2\alpha}} = 2\zeta(2\alpha).$$

The result follows. □

LEMMA 8.7. *Let* $k = \kappa_{a_1-1}b^{a_1-1} + \cdots + \kappa_{a_\nu-1}b^{a_\nu-1}$, *with* $1 \le a_\nu < \cdots < a_1$, *let* $\kappa_{a_1-1}, \ldots, \kappa_{a_\nu-1} \in \{1, \ldots, b-1\}$. *Then*

$$\beta_{h,\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}} = \sum_{\substack{h_1,\ldots,h_\nu\in\mathbb{Z},h_l\equiv\kappa_{a_l-1} \pmod b\\ h=h_1b^{a_1-1}+\cdots+h_\nu b^{a_\nu-1}}} \frac{b^\nu}{(2\pi\mathrm{i})^\nu} \prod_{l=1}^{\nu} \frac{1-\mathrm{e}^{2\pi\mathrm{i}h_l/b}}{h_l}.$$

Proof. First we consider $k = \kappa_{a-1}b^{a-1}$, with $\kappa_{a-1} \in \{1, \ldots, b-1\}$. Let $x = \frac{x_1}{b} + \frac{x_2}{b^2} + \cdots$, and then we have $\mathrm{wal}_k(x) = \mathrm{e}^{2\pi\mathrm{i}\kappa_{a-1}x_a/b}$. Note that $\mathrm{wal}_k(x)$ is constant in the intervals $[u/b^a, (u+1)/b^a)$ for $0 \le u < b^a$. Let $u = u_{a-1}b^{a-1} + \cdots + u_0$. Then for any $h \in \mathbb{Z} \setminus \{0\}$ we have

$$\begin{aligned}
\beta_{h,k} &= \sum_{u=0}^{b^a-1} \mathrm{e}^{2\pi\mathrm{i}\kappa_{a-1}u_0/b} \int_{u/b^a}^{(u+1)/b^a} \mathrm{e}^{-2\pi\mathrm{i}hx}\,\mathrm{d}x\\
&= \sum_{u=0}^{b^a-1} \mathrm{e}^{2\pi\mathrm{i}\kappa_{a-1}u_0/b} \frac{\mathrm{e}^{-2\pi\mathrm{i}h(u+1)/b^a} - \mathrm{e}^{-2\pi\mathrm{i}hu/b^a}}{-2\pi\mathrm{i}h}\\
&= \frac{1-\mathrm{e}^{-2\pi\mathrm{i}h/b^a}}{2\pi h\mathrm{i}} \sum_{u_0=0}^{b-1} \cdots \sum_{u_{a-1}=0}^{b-1} \mathrm{e}^{2\pi\mathrm{i}\kappa_{a-1}u_0/b} \mathrm{e}^{-2\pi\mathrm{i}h(u_{a-1}/b+\cdots+u_0/b^a)}\\
&= \frac{1-\mathrm{e}^{-2\pi\mathrm{i}h/b^a}}{2\pi h\mathrm{i}} \sum_{u_0=0}^{b-1} \mathrm{e}^{2\pi\mathrm{i}u_0(\kappa_{a-1}/b-h/b^a)} \sum_{u_1=0}^{b-1} \mathrm{e}^{-2\pi\mathrm{i}u_1h/b^{a-1}} \cdots \sum_{u_{a-1}=0}^{b-1} \mathrm{e}^{-2\pi\mathrm{i}u_{a-1}h/b}.
\end{aligned}$$

Let now $h \in \mathbb{Z}\setminus\{0\}$, let $h = h_{c-1}b^{c-1} + \cdots + h_0$, and set $h_c = h_{c+1} = \cdots = 0$. If $h > 0$, we assume that $h_i \in \{0, \ldots, b-1\}$, and if $h < 0$, we assume that $h_i \in \{-b+1, \ldots, 0\}$ for all $i \ge 0$. If $h_0 \ne 0$, then $\sum_{u_{a-1}=0}^{b-1} \mathrm{e}^{-2\pi\mathrm{i}u_{a-1}h/b} = 0$, and hence $\beta_{h,\kappa_{a-1}b^{a-1}} = 0$. If $h_0 = 0$, then $\sum_{u_{a-1}=0}^{b-1} \mathrm{e}^{-2\pi\mathrm{i}u_{a-1}h/b} = b$. In general, if for an $0 \le i < a-1$ we have $h_i \ne 0$, then $\beta_{h,\kappa_{a-1}b^{a-1}} = 0$. Further, if $h_i = 0$ for $0 \le i \le a-1$, then we also have $\beta_{h,\kappa_{a-1}b^{a-1}} = 0$. Hence, in order to obtain $\beta_{h,\kappa_{a-1}b^{a-1}} \ne 0$ we must have $h_0 = \cdots = h_{a-2} = 0$ and $\kappa_{a-1} - h_{a-1} \equiv 0 \pmod b$. In this case we have

$$\beta_{h,\kappa_{a-1}b^{a-1}} = \frac{1-\mathrm{e}^{-2\pi\mathrm{i}h_{a-1}/b}}{2\pi h\mathrm{i}} b^a,$$

where $h = h_{a-1}b^{a-1} + h_a b^a + \cdots$, with $h_{a-1} \equiv \kappa_{a-1} \pmod b$. We can also write

$$\beta_{hb^{a-1},\kappa_{a-1}b^{a-1}} = \frac{b(1-\mathrm{e}^{-2\pi \mathtt{i} h/b})}{2\pi \mathtt{i} h},$$

with $h \in \mathbb{Z}$ such that $h \equiv \kappa_{a-1} \pmod b$.

We can interpret $\beta_{h,k} = \int_0^1 \mathrm{e}^{-2\pi \mathtt{i} hx} \mathrm{wal}_k(x)\,\mathrm{d}x$ as the Fourier coefficients of the kth Walsh function; hence, it follows that

$$\mathrm{wal}_k(x) = \sum_{h\in\mathbb{Z}} \beta_{h,k} \mathrm{e}^{2\pi \mathtt{i} hx}.$$

Let now $k = \kappa_{a_1-1}b^{a_1-1} + \cdots + \kappa_{a_\nu-1}b^{a_\nu-1}$ for some $1 \le a_\nu < \cdots < a_1$. Then we have

$$\begin{aligned}
&\mathrm{wal}_{\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}}(x)\\
&= \mathrm{wal}_{\kappa_{a_1-1}b^{a_1-1}}(x)\cdots \mathrm{wal}_{\kappa_{a_\nu-1}b^{a_\nu-1}}(x)\\
&= \sum_{h_1\in\mathbb{Z}} \beta_{h_1,\kappa_{a_1-1}b^{a_1-1}} \mathrm{e}^{2\pi \mathtt{i} h_1 x} \cdots \sum_{h_\nu\in\mathbb{Z}} \beta_{h_\nu,\kappa_{a_\nu-1}b^{a_\nu-1}} \mathrm{e}^{2\pi \mathtt{i} h_\nu x}\\
&= \sum_{h_1,\ldots,h_\nu\in\mathbb{Z}} \beta_{h_1,\kappa_{a_1-1}b^{a_1-1}} \cdots \beta_{h_\nu,\kappa_{a_\nu-1}b^{a_\nu-1}} \mathrm{e}^{2\pi \mathtt{i}(h_1+\cdots+h_\nu)x}.
\end{aligned}$$

On the other hand, we have

$$\mathrm{wal}_{\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}}(x) = \sum_{h\in\mathbb{Z}\setminus\{0\}} \beta_{h,\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}} \mathrm{e}^{2\pi \mathtt{i} hx}.$$

On comparing the last two equations we obtain that $\beta_{h,\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}} = 0$ if either $b^{a_1-1} \not| h$ or $h \not\equiv \kappa_{a_1-1} \pmod{b^{a_1-1}}$. Now let $h \in \mathbb{Z}$ such that $b^{a_1-1} | h$ and $h \equiv \kappa_{a_1-1} \pmod{b^{a_1-1}}$. Then we have

$$\begin{aligned}
&\beta_{h,\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}}\\
&= \sum_{\substack{h_1,\ldots,h_\nu\in\mathbb{Z}, h_l\equiv\kappa_{a_l-1} \pmod b\\ h=h_1b^{a_1-1}+\cdots+h_\nu b^{a_\nu-1}}} \beta_{h_1b^{a_1-1},\kappa_{a_1-1}b^{a_1-1}} \cdots \beta_{h_\nu b^{a_\nu-1},\kappa_{a_\nu-1}b^{a_\nu-1}}\\
&= \sum_{\substack{h_1,\ldots,h_\nu\in\mathbb{Z}, h_l\equiv\kappa_{a_l-1} \pmod b\\ h=h_1b^{a_1-1}+\cdots+h_\nu b^{a_\nu-1}}} \frac{b^\nu}{(2\pi\mathtt{i})^\nu} \prod_{l=1}^{\nu} \frac{1-\mathrm{e}^{2\pi \mathtt{i} h_l/b}}{h_l},
\end{aligned}$$

and the result follows. □

LEMMA 8.8. *For $k \ge 1$, $b \ge 2$, $m \ge 1$, and $\alpha > 1/2$ we have*

$$r_{b,\alpha}(kb^m) = b^{-2\alpha m} r_{b,\alpha}(k).$$

Proof. First note that $\beta_{h,\kappa_{a_1-1}b^{m+a_1-1}+\cdots+\kappa_{a_\nu-1}b^{m+a_\nu-1}} = 0$ if $b^m \not| h$. Further it follows from the previous lemma that

$$\beta_{hb^m,\kappa_{a_1-1}b^{m+a_1-1}+\cdots+\kappa_{a_\nu-1}b^{m+a_\nu-1}} = \beta_{h,\kappa_{a_1-1}b^{a_1-1}+\cdots+\kappa_{a_\nu-1}b^{a_\nu-1}},$$

and hence by Lemma 2.6 we have

$$r_{b,\alpha}(kb^m) = \sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{|\beta_{hb^m,kb^m}|^2}{|hb^m|^{2\alpha}} = b^{-2\alpha m} \sum_{h\in\mathbb{Z}\setminus\{0\}} \frac{|\beta_{h,k}|^2}{|h|^{2\alpha}} = b^{-2\alpha m} r_{b,\alpha}(k).$$

The result follows. □

REFERENCES

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Amer. Math. Soc., 68 (1950), pp. 337–404.

[2] H. E. Chrestenson, *A class of generalized Walsh functions*, Pacific J. Math., 5 (1955), pp. 17–31.

[3] L. L. Cristea, J. Dick, G. Leobacher, and F. Pillichshammer, *The Tent Transformation Can Improve the Convergence Rate of Quasi-Monte Carlo Algorithms Using Digital Nets*, Numerische Mathematik, 105 (2007), pp. 413–455.

[4] J. Dick and F. Pillichshammer, *Multivariate integration in weighted Hilbert spaces based on Walsh functions and weighted Sobolev spaces*, J. Complexity, 21 (2005), pp. 149–195.

[5] J. Dick and F. Pillichshammer, *On the mean square weighted L_2 discrepancy of randomized digital (t, m, s)-nets over $\mathbb{Z}_2$*, Acta Arith., 117 (2005), pp. 371–403.

[6] H. Faure, *Discrèpances de suites associèes á un système de numèration (en dimension s)*, Acta Arith., 41 (1982), pp. 337–351.

[7] T. Gerstner and M. Griebel, *Numerical integration using sparse grids*, Numer. Algorithms, 18 (1998), pp. 209–232.

[8] F. J. Hickernell and R.-X. Yue, *The mean square discrepancy of scrambled (t, s)-sequences*, SIAM J. Numer. Anal., 38 (2000), pp. 1089–1112.

[9] E. Hlawka, *Zur angenäherten Berechnung mehrfacher Integrale*, Monatsh. Math., 66 (1962), pp. 140–151.

[10] L. K. Hua and Y. Wang, *Applications of Number Theory to Numerical Analysis*, Springer-Verlag, Berlin, 1981.

[11] N. M. Korobov, *The approximate computation of multiple integrals*, Dokl. Akad. Nauk, 124 (1959), pp. 1207–1210.

[12] N. M. Korobov, *Number-Theoretic Methods in Approximate Analysis*, Fizmatgiz, Moscow, 1963.

[13] J. Matoušek, *Geometric Discrepancy*, Algorithms Combin. 18, Springer-Verlag, Berlin, 1999.

[14] H. Niederreiter, *Quasi-Monte Carlo methods and pseudo-random numbers*, Bull. Amer. Math. Soc., 84 (1978), pp. 957–1041.

[15] H. Niederreiter, *Low-discrepancy point sets*, Monatsh. Math., 102 (1986), pp. 155–167.

[16] H. Niederreiter, *Point sets and sequences with small discrepancy*, Monatsh. Math., 104 (1987), pp. 273–337.

[17] H. Niederreiter, *Random Number Generation and Quasi-Monte Carlo Methods*, CBMS–NSF Regional Conf. Ser. in Appl. Math. 63, SIAM, Philadelphia, 1992.

[18] H. Niederreiter, *Improved error bounds for lattice rules*, J. Complexity, 9 (1993), pp. 60–75.

[19] H. Niederreiter and G. Pirsic, *Duality for digital nets and its applications*, Acta Arith., 97 (2001), pp. 173–182.

[20] H. Niederreiter and C. P. Xing, *Quasirandom points and global function fields*, Finite Fields and Applications, S. Cohen and H. Niederreiter, eds., London Math. Soc. Lecture Note Ser. 233, Cambridge University Press, Cambridge, 1996, pp. 269–296.

[21] H. Niederreiter and C. P. Xing, *Rational points on curves over finite fields*, London Math. Soc. Lecture Note Ser. 285, Cambridge University Press, Cambridge, 2001.

[22] A. B. Owen, *Scrambled net variance for integrals of smooth functions*, Ann. Statist., 25 (1997), pp. 1541–1562.

[23] M. Yu. Rosenbloom and M. A. Tsfasman, *Codes in the m-metric*, Problemy Peredachi Informatsii, 33 (1997), pp. 45–52.

[24] I. F. Sharygin, *A lower estimate for the error of quadrature formulas for certain classes of functions*, Zh. Vychisl. Mat. Mat. Fiz., 3 (1963), pp. 370–376.

[25] I. H. Sloan and S. Joe, *Lattice Methods for Multiple Integration*, Clarendon Press, Oxford, 1994.

[26] I. H. Sloan, F. Y. Kuo, and S. Joe, *Constructing randomly shifted lattice rules in weighted Sobolev spaces*, SIAM J. Numer. Anal., 40 (2002), pp. 1650–1665.

[27] I. H. Sloan, F. Y. Kuo, and S. Joe, *On the step-by-step construction of quasi-Monte Carlo integration rules that achieve strong tractability error bounds in weighted Sobolev spaces*, Math. Comp., 71 (2002), pp. 1609–1640.
[28] I. H. Sloan and H. Woźniakowski, *When are quasi-Monte Carlo algorithms efficient for high dimensional integrals?*, J. Complexity, 14 (1998), pp. 1–33.
[29] S. A. Smolyak, *Quadrature and interpolation formulas for tensor products of certain classes of functions*, Dokl. Akad. Nauk, 4 (1963), pp. 240–243.
[30] I. M. Sobol, *The distribution of points in a cube and the approximate evaluation of integrals*, Zh. Vychisl. Mat. Mat. Fiz., 7 (1967), pp. 784–802.
[31] J. L. Walsh, *A closed set of normal orthogonal functions*, Amer. J. Math., 55 (1923), pp. 5–24.
[32] X. Wang, I. H. Sloan, and J. Dick, *On Korobov lattice rules in weighted spaces*, SIAM J. Numer. Anal., 42 (2004), pp. 1760–1779.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2177–2200

A MASS AND MAGNETIZATION CONSERVATIVE AND ENERGY-DIMINISHING NUMERICAL METHOD FOR COMPUTING GROUND STATE OF SPIN-1 BOSE–EINSTEIN CONDENSATES*

WEIZHU BAO† AND HANQUAN WANG‡

Abstract. In this paper, a mass (or normalization) and magnetization conservative and energy-diminishing numerical method is presented for computing the ground state of spin-1 (or $F = 1$ spinor) Bose–Einstein condensates (BECs). We begin with the coupled Gross–Pitaevskii equations, and the ground state is defined as the minimizer of the energy functional under two constraints on the mass and magnetization. By constructing a continuous normalized gradient flow (CNGF) which is mass and magnetization conservative and energy-diminishing, the ground state can be computed as the steady state solution of the CNGF. The CNGF is then discretized by the Crank–Nicolson finite difference method with a proper way to deal with the nonlinear terms, and we prove that the discretization is mass and magnetization conservative and energy-diminishing in the discretized level. Numerical results of the ground state and their energy of spin-1 BECs are reported to demonstrate the efficiency of the numerical method.

Key words. spin-1 Bose–Einstein condensate, coupled Gross–Pitaevskii equations, ground state, continuous normalized gradient flow, mass and magnetization conservative, energy-diminishing

AMS subject classifications. 35Q55, 65T99, 65Z05, 65N12, 65N35, 81-08

DOI. 10.1137/070681624

1. Introduction.

Since its realization in dilute bosonic atomic gases [2, 13, 9], the atomic Bose–Einstein condensate (BEC) has been produced and studied extensively in the laboratory [28, 29, 16] and has provided a successful testing ground of theoretical studies of quantum many-body systems [28, 29]. In earlier BEC experiments, atoms were spatially confined with magnetic traps, which essentially freeze the atomic internal degrees of freedom [2, 13, 9]. Most studies were thus focused on scalar models, i.e., single-component quantum degenerate gases [12]. One of the most important recent developments in BECs was the study of spin-1 condensates (of atoms with hyperfine quantum number $F = 1$) [17, 27, 34, 10, 31], and they were realized in experiments recently using both ^{23}Na and ^{87}Rb [24, 35]. In fact, the emergence of the spin-1 BEC [19, 20, 24] has created opportunities for understanding degenerate gases with internal degrees of freedom [21, 22, 17, 18, 14, 25, 26, 32, 37].

At temperature T much smaller than the critical condensate temperature T_c [23], a spin-1 BEC is well described by the three-component wave function $\Psi = (\psi_1(\mathbf{x}, t), \psi_0(\mathbf{x}, t), \psi_{-1}(\mathbf{x}, t))^T$ whose evolution is governed by the coupled Gross–Pitaevskii equations (GPEs) [23, 17, 18, 38, 36]:

*Received by the editors February 1, 2007; accepted for publication (in revised form) June 25, 2007; published electronically September 28, 2007. This research was supported by National University of Singapore grant R-146-000-083-112.

http://www.siam.org/journals/sinum/45-5/68162.html

†Department of Mathematics and Center for Computational Science & Engineering, National University of Singapore, Singapore 117543, Singapore (bao@math.nus.edu.sg, http://www.math.nus.edu.sg/~bao/).

‡Current address: Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong (mahqwang@ust.hk).

$$i\hbar\partial_t\psi_1(\mathbf{x},t) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) + (c_0+c_2)\left(|\psi_1|^2+|\psi_0|^2\right) + (c_0-c_2)|\psi_{-1}|^2\right]\psi_1$$
$$+c_2\bar{\psi}_{-1}\psi_0^2, \tag{1.1}$$
$$i\hbar\partial_t\psi_0(\mathbf{x},t) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) + (c_0+c_2)\left(|\psi_1|^2+|\psi_{-1}|^2\right) + c_0|\psi_0|^2\right]\psi_0$$
$$+2c_2\psi_{-1}\bar{\psi}_0\psi_1, \tag{1.2}$$
$$i\hbar\partial_t\psi_{-1}(\mathbf{x},t) = \left[-\frac{\hbar^2}{2m}\nabla^2 + V(\mathbf{x}) + (c_0+c_2)\left(|\psi_{-1}|^2+|\psi_0|^2\right) + (c_0-c_2)|\psi_1|^2\right]\psi_{-1}$$
$$+c_2\psi_0^2\bar{\psi}_1. \tag{1.3}$$

Here $\mathbf{x} = (x,y,z)^T$ is the Cartesian coordinate vector, $\hbar$ is the Planck constant, m is the atomic mass, and $V(\mathbf{x})$ is the external trapping potential. When a harmonic trap potential is considered,

$$V(\mathbf{x}) = \frac{m}{2}(\omega_x^2x^2 + \omega_y^2y^2 + \omega_z^2z^2), \tag{1.4}$$

with ω_x, ω_y, and ω_z being the trap frequencies in the x-, y-, and z-directions, respectively. $\bar{f}$ and $\mathrm{Re}(f)$ denote the conjugate and the real part of the function f, respectively. $c_0 = 4\pi\hbar^2(a_0+2a_2)/3m$ and $c_2 = 4\pi\hbar^2(a_2-a_0)/3m$ denote constants of the mean-field (spin-independent) and spin-exchange interaction, respectively, with a_j the s-wave scattering lengths for the channel of total hyperfine spin j ($j = 0, 2$). The wave function is normalized according to

$$\|\Psi\|^2 := \int_{\mathbb{R}^3}|\Psi(\mathbf{x},t)|^2\,d\mathbf{x} = \int_{\mathbb{R}^3}\sum_{j=-1}^{1}|\psi_j(\mathbf{x},t)|^2\,d\mathbf{x} := \sum_{j=-1}^{1}\|\psi_j\|^2 = N, \tag{1.5}$$

where N is the total number of particles in the condensate.

By introducing the dimensionless variables $t \to t/\omega_m$, with $\omega_m = \min\{\omega_x,\omega_y,\omega_z\}$, and $\mathbf{x} \to \mathbf{x}\ a_s$, with $a_s = \sqrt{\frac{\hbar}{m\omega_m}}$, $\psi_j \to \sqrt{N}\psi_j/a_s^{3/2}$ ($j = -1, 0, 1$), we get the dimensionless coupled GPEs from (1.1)–(1.3) as [38, 39, 36]:

$$i\partial_t\psi_1(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\psi_1|^2+|\psi_0|^2\right) + (\beta_n-\beta_s)|\psi_{-1}|^2\right]\psi_1$$
$$+\beta_s\bar{\psi}_{-1}\psi_0^2, \tag{1.6}$$
$$i\partial_t\psi_0(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\psi_1|^2+|\psi_{-1}|^2\right) + \beta_n|\psi_0|^2\right]\psi_0$$
$$+2\beta_s\psi_{-1}\bar{\psi}_0\psi_1, \tag{1.7}$$
$$i\partial_t\psi_{-1}(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\psi_{-1}|^2+|\psi_0|^2\right) + (\beta_n-\beta_s)|\psi_1|^2\right]\psi_{-1}$$
$$+\beta_s\psi_0^2\bar{\psi}_1, \tag{1.8}$$

where $\beta_n = \frac{N\,c_0}{a_s^3\hbar\omega_m} = \frac{4\pi N(a_0+2a_2)}{3a_s}$, $\beta_s = \frac{N\,c_2}{a_s^3\hbar\omega_m} = \frac{4\pi N(a_2-a_0)}{3a_s}$, and $V(\mathbf{x}) = \frac{1}{2}(\gamma_x^2x^2 + \gamma_y^2y^2 + \gamma_z^2z^2)$, with $\gamma_x = \frac{\omega_x}{\omega_m}$, $\gamma_y = \frac{\omega_y}{\omega_m}$, and $\gamma_z = \frac{\omega_z}{\omega_m}$. Similar as those in the single-component BEC [29, 1, 7, 3, 6], in the disk-shaped condensation, i.e., $\omega_x \approx \omega_y$ and $\omega_z \gg \omega_x$ ($\Longleftrightarrow \gamma_x = 1$, $\gamma_y \approx 1$, and $\gamma_z \gg 1$, with $\omega_m = \omega_x$), the three-dimensional (3D) coupled GPEs (1.6)–(1.8) can be reduced to 2D coupled GPEs, and in the cigar-shaped condensation, i.e., $\omega_y \gg \omega_x$ and $\omega_z \gg \omega_x$ ($\Longleftrightarrow \gamma_x = 1$, $\gamma_y \gg 1$, and $\gamma_z \gg 1$,

with $\omega_m = \omega_x$), the 3D coupled GPEs (1.6)–(1.8) can be reduced to 1D coupled GPEs. Thus here we consider the dimensionless coupled GPEs in d dimensions ($d = 1, 2, 3$):

$$i\partial_t\psi_1(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\psi_1|^2+|\psi_0|^2\right) + (\beta_n-\beta_s)|\psi_{-1}|^2\right]\psi_1$$
$$+\beta_s\bar{\psi}_{-1}\psi_0^2, \tag{1.9}$$

$$i\partial_t\psi_0(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\psi_1|^2+|\psi_{-1}|^2\right) + \beta_n|\psi_0|^2\right]\psi_0$$
$$+2\beta_s\psi_{-1}\bar{\psi}_0\psi_1, \tag{1.10}$$

$$i\partial_t\psi_{-1}(\mathbf{x},t) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\psi_{-1}|^2+|\psi_0|^2\right) + (\beta_n-\beta_s)|\psi_1|^2\right]\psi_{-1}$$
$$+\beta_s\psi_0^2\bar{\psi}_1. \tag{1.11}$$

Here $V(\mathbf{x})$ is a real-valued potential whose shape is determined by the type of system under investigation; $\beta_n \propto N$ and $\beta_s \propto N$ correspond to the dimensionless mean-field (spin-independent) and spin-exchange interaction, respectively. Three important invariants of (1.9)–(1.11) are the *mass* (or normalization) of the wave function

$$N(\Psi(\cdot,t)) := \|\Psi(\cdot,t)\|^2 := \int_{\mathbb{R}^d}\sum_{j=-1}^{1}|\psi_j(\mathbf{x},t)|^2\,d\mathbf{x} \equiv N(\Psi(\cdot,0)) = 1, \qquad t \ge 0, \tag{1.12}$$

the *magnetization* (with $-1 \le M \le 1$)

$$M(\Psi(\cdot,t)) := \int_{\mathbb{R}^d}\left[|\psi_1(\mathbf{x},t)|^2 - |\psi_{-1}(\mathbf{x},t)|^2\right]d\mathbf{x} \equiv M(\Psi(\cdot,0)) = M, \tag{1.13}$$

and the energy per particle

$$E(\Psi(\cdot,t)) = \int_{\mathbb{R}^d}\Bigg\{\sum_{j=-1}^{1}\left(\frac{1}{2}|\nabla\psi_j|^2 + V(\mathbf{x})|\psi_j|^2\right) + (\beta_n-\beta_s)|\psi_1|^2|\psi_{-1}|^2$$
$$+\frac{\beta_n}{2}|\psi_0|^4 + \frac{\beta_n+\beta_s}{2}\left[|\psi_1|^4+|\psi_{-1}|^4+2|\psi_0|^2\left(|\psi_1|^2+|\psi_{-1}|^2\right)\right]$$
$$+\beta_s\left(\bar{\psi}_{-1}\psi_0^2\bar{\psi}_1 + \psi_{-1}\bar{\psi}_0^2\psi_1\right)\Bigg\}\,d\mathbf{x} \equiv E(\Psi(\cdot,0)), \qquad t \ge 0. \tag{1.14}$$

The ground state of a spin-1 BEC is defined as the minimizer of the following nonconvex minimization problem:

Find ($\Phi_g \in S$) such that

$$E_g := E(\Phi_g) = \min_{\Phi\in S} E(\Phi), \tag{1.15}$$

where the nonconvex set S is defined as

$$S = \left\{\Phi = (\phi_1,\phi_0,\phi_{-1})^T \mid \|\Phi\| = 1,\ \int_{\mathbb{R}^d}\left[|\phi_1(\mathbf{x})|^2 - |\phi_{-1}(\mathbf{x})|^2\right] = M,\ E(\Phi) < \infty\right\}. \tag{1.16}$$

When $\beta_n \ge 0$, $\beta_n \ge |\beta_s|$, and $\lim_{|\mathbf{x}|\to\infty} V(\mathbf{x}) = \infty$, the existence of a minimizer of the nonconvex minimization problem (1.15) follows from the standard theory [33].

For understanding the uniqueness question note that $E(\alpha \cdot \Phi_g) = E(\Phi_g)$ for all $\alpha = \left(e^{i\theta_1}, e^{i\theta_0}, e^{i\theta_{-1}}\right)^T$, with $\theta_1 + \theta_{-1} = 2\theta_0$. Thus additional constraints have to be introduced to show the uniqueness.

One of the fundamental problems in theoretical study of a spin-1 BEC is to find its ground state so as to compare the numerical results with experimental observations and to prepare initial data for studying the dynamics of a spin-1 BEC. Due to the facts that there are three components in the wave function Φ in (1.15) and that there are only two constraints in (1.16), it is not obvious that the most powerful and popular imaginary time method [11, 1, 3, 4, 5, 8, 7] used for computing the ground state of a single-component BEC could be extended to this case directly. The reason is that, in the projection step, we need to determine three parameters but have only two equations from the two constraints. However, in physics literatures, they still use the imaginary time method for computing the ground state of a spin-1 BEC by introducing a random variable to choose the three projection parameters in the projection step [38]. Of course, this is not a determinate and efficient way to compute the ground state of a spin-1 BEC due to the choice of the random variable. In fact, to our knowledge, there is no efficient and determinate numerical method for computing the ground state of a spin-1 BEC in the literature yet. The aim of this paper is to propose such a numerical method.

The paper is organized as follows. In section 2, we first introduce the Euler–Lagrange equations (or time-independent coupled GPEs) associated to the minimization problem (1.15) and then construct a continuous normalized gradient flow (CNGF) such that the ground state of a spin-1 BEC is the steady state solution of this CNGF. In section 3, the CNGF is discretized in space and time with a proper way to treat the nonlinear terms, and we prove that the discretization is mass and magnetization conservative and energy-diminishing. In section 4, numerical results are reported to demonstrate the efficiency of our numerical method. Finally, some conclusions are drawn in section 5.

2. A continuous normalized gradient flow. In this section, we will introduce the Euler–Lagrange equations associated to the minimization problem (1.15) and construct a continuous normalized gradient flow for computing the ground state of a spin-1 BEC.

2.1. Euler–Lagrange equations. In order to find the Euler–Lagrange equations associated to the minimization problem (1.15), we define the Lagrangian

$$\mathcal{L}(\Phi, \mu, \lambda) := E(\Phi) - \mu\left(\|\phi_1\|^2 + \|\phi_0\|^2 + \|\phi_{-1}\|^2 - 1\right) - \lambda\left(\|\phi_1\|^2 - \|\phi_{-1}\|^2 - M\right). \tag{2.1}$$

Differentiating (2.1) with respect to $\bar{\phi}_1$, $\bar{\phi}_0$, and $\bar{\phi}_{-1}$, respectively, we get the following Euler–Lagrange equations:

$$\begin{aligned}(\mu+\lambda)\,\phi_1(\mathbf{x}) = &\left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\phi_1|^2 + |\phi_0|^2\right) + (\beta_n-\beta_s)|\phi_{-1}|^2\right]\phi_1 \\ &+\beta_s\bar{\phi}_{-1}\phi_0^2 := H_1\phi_1,\end{aligned} \tag{2.2}$$

$$\begin{aligned}\mu\,\phi_0(\mathbf{x}) = &\left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\phi_1|^2 + |\phi_{-1}|^2\right) + \beta_n|\phi_0|^2\right]\phi_0 \\ &+2\beta_s\phi_{-1}\bar{\phi}_0\phi_1 := H_0\phi_0,\end{aligned} \tag{2.3}$$

$$(\mu-\lambda)\,\phi_{-1}(\mathbf{x}) = \left[-\frac{1}{2}\nabla^2 + V(\mathbf{x}) + (\beta_n+\beta_s)\left(|\phi_{-1}|^2+|\phi_0|^2\right) + (\beta_n-\beta_s)|\phi_1|^2\right]\phi_{-1}$$

$$+\beta_s\phi_0^2\bar{\phi}_1 := H_{-1}\phi_{-1}. \tag{2.4}$$

Here μ and λ are the Lagrange multipliers (or chemical potentials) of the coupled GPEs (2.2)–(2.4). In addition, (2.2)–(2.4) is also a nonlinear eigenvalue problem with two constraints:

$$\|\Phi\|^2 := \int_{\mathbb{R}^d} |\Phi(\mathbf{x})|^2\,d\mathbf{x} = \int_{\mathbb{R}^d}\sum_{j=-1}^{1}|\phi_j(\mathbf{x})|^2\,d\mathbf{x} := \sum_{j=-1}^{1}\|\phi_j\|^2 = 1, \tag{2.5}$$

$$\|\phi_1\|^2 - \|\phi_{-1}\|^2 := \int_{\mathbb{R}^d}\left[|\phi_1(\mathbf{x})|^2 - |\phi_{-1}(\mathbf{x})|^2\right]d\mathbf{x} = M. \tag{2.6}$$

In fact, the nonlinear eigenvalue problem (2.2)–(2.4) can be also obtained from the coupled GPEs (1.9)–(1.11) by plugging $\psi_j(\mathbf{x},t) = e^{-i\mu_j t}\phi_j(\mathbf{x})$ $(j=1,0,-1)$ with $\mu_1=\mu+\lambda$, $\mu_0=\mu$, and $\mu_{-1}=\mu-\lambda$. Thus it is also called time-independent coupled GPEs. In physics literatures, any eigenfunction Φ of the nonlinear eigenvalue problem (2.2)–(2.4) under the constraints (2.5) and (2.6) whose energy is larger than the energy of the ground state is called an excited state of the coupled GPEs (1.9)–(1.11).

When $V(\mathbf{x})$ is chosen as a harmonic oscillator potential, following the idea in [12, 29] for a single-component BEC, we have the following virial theorem for a spin-1 BEC.

LEMMA 2.1. *Suppose $\Phi \in S$ is an eigenfunction of the nonlinear eigenvalue problem* (2.2)–(2.4). *When $V(\mathbf{x})$ is chosen as a harmonic oscillator potential, i.e., it is a quadratic form in* $\mathbf{x}$, *we have*

$$2\,E_{\rm kin}(\Phi) - 2\,E_{\rm pot}(\Phi) + d\,E_{\rm int}(\Phi) = 0, \tag{2.7}$$

where $E_{\rm kin}$, $E_{\rm pot}$, and $E_{\rm int}$ are the kinetic energy, potential energy, and interaction energy, respectively, and are defined as

$$E_{\rm kin}(\Phi) = \frac{1}{2}\int_{\mathbb{R}^d}\sum_{j=-1}^{1}|\nabla\phi_j|^2\,d\mathbf{x}, \qquad E_{\rm pot}(\Phi) = \int_{\mathbb{R}^d}\sum_{j=-1}^{1}V(\mathbf{x})|\phi_j|^2\,d\mathbf{x}, \tag{2.8}$$

$$E_{\rm int}(\Phi) = \int_{\mathbb{R}^d}\Big[\frac{\beta_n+\beta_s}{2}\Big(|\phi_1|^4+|\phi_{-1}|^4+2|\phi_0|^2\left(|\phi_1|^2+|\phi_{-1}|^2\right)\Big)$$

$$+(\beta_n-\beta_s)|\phi_1|^2|\phi_{-1}|^2 + \frac{\beta_n}{2}|\phi_0|^4 + \beta_s\left(\bar{\phi}_{-1}\phi_0^2\bar{\phi}_1 + \phi_{-1}\bar{\phi}_0^2\phi_1\right)\Big]\,d\mathbf{x}. \tag{2.9}$$

Proof. Suppose $\Phi_e \in S$ is an eigenfunction of the nonlinear eigenvalue problem (2.2)–(2.4), and we define a trial function $\Phi_\varepsilon \in S$ as

$$\Phi_\varepsilon(\mathbf{x}) = (1+\varepsilon)^{d/2}\Phi_e\left((1+\varepsilon)\mathbf{x}\right), \qquad \mathbf{x}\in\mathbb{R}^d. \tag{2.10}$$

Plugging Φ_ε into the energy functional in (1.14), change of variables, we obtain

$$E(\Phi_\varepsilon(\mathbf{x})) = E\left((1+\varepsilon)^{d/2}\Phi_e\left((1+\varepsilon)\mathbf{x}\right)\right)$$

$$= (1+\varepsilon)^2 E_{\rm kin}\left(\Phi_e(\mathbf{x})\right) + \frac{1}{(1+\varepsilon)^2}E_{\rm pot}\left(\Phi_e(\mathbf{x})\right) + (1+\varepsilon)^d E_{\rm int}\left(\Phi_e(\mathbf{x})\right). \tag{2.11}$$

Differentiating (2.11) with respect to ε, we get

$$\frac{dE(\Phi_\varepsilon)}{d\varepsilon} = 2(1+\varepsilon)E_{\rm kin}\left(\Phi_e\right) - \frac{2}{(1+\varepsilon)^3}E_{\rm pot}\left(\Phi_e\right) + d(1+\varepsilon)^{d-1}E_{\rm int}\left(\Phi_e\right). \tag{2.12}$$

Since Φ_e is also a critical point of the energy functional $E(\Phi)$ over the set S, we get (2.7) from (2.12) by setting $\varepsilon = 0$ and noticing $\Phi_{\varepsilon=0}(\mathbf{x}) = \Phi_e(\mathbf{x})$ in (2.10). ∎

2.2. A continuous normalized gradient flow. In order to compute the ground state of a spin-1 BEC in (1.15) numerically, we construct the following CNGF:

$$\partial_t \phi_1(\mathbf{x},t) = \left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - (\beta_n+\beta_s)\left(|\phi_1|^2+|\phi_0|^2\right) - (\beta_n-\beta_s)|\phi_{-1}|^2\right]\phi_1$$

$$-\beta_s\bar{\phi}_{-1}\phi_0^2 + [\mu_\Phi(t)+\lambda_\Phi(t)]\,\phi_1 = -H_1\,\phi_1 + [\mu_\Phi(t)+\lambda_\Phi(t)]\,\phi_1, \tag{2.13}$$

$$\partial_t \phi_0(\mathbf{x},t) = \left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - (\beta_n+\beta_s)\left(|\phi_1|^2+|\phi_{-1}|^2\right) - \beta_n|\phi_0|^2\right]\phi_0$$

$$-2\beta_s\phi_{-1}\bar{\phi}_0\phi_1 + \mu_\Phi(t)\,\phi_0 = -H_0\,\phi_0 + \mu_\Phi(t)\,\phi_0, \tag{2.14}$$

$$\partial_t \phi_{-1}(\mathbf{x},t) = \left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - (\beta_n+\beta_s)\left(|\phi_{-1}|^2+|\phi_0|^2\right) - (\beta_n-\beta_s)|\phi_1|^2\right]\phi_{-1}$$

$$-\beta_s\phi_0^2\bar{\phi}_1 + [\mu_\Phi(t)-\lambda_\Phi(t)]\,\phi_{-1} = -H_{-1}\,\phi_{-1} + [\mu_\Phi(t)-\lambda_\Phi(t)]\,\phi_{-1}. \tag{2.15}$$

Here $\mu_\Phi(t)$ and $\lambda_\Phi(t)$ are chosen such that the above CNGF is mass (or normalization) and magnetization conservative, and they are given as

$$\mu_\Phi(t) = \frac{R_\Phi(t)D_\Phi(t) - M_\Phi(t)F_\Phi(t)}{N_\Phi(t)R_\Phi(t) - M_\Phi^2(t)}, \qquad \lambda_\Phi(t) = \frac{N_\Phi(t)F_\Phi(t) - M_\Phi(t)D_\Phi(t)}{N_\Phi(t)R_\Phi(t) - M_\Phi^2(t)}, \tag{2.16}$$

with

$$N_\Phi(t) = \int_{\mathbb{R}^d}\left[|\phi_{-1}(\mathbf{x},t)|^2 + |\phi_0(\mathbf{x},t)|^2 + |\phi_1(\mathbf{x},t)|^2\right]d\mathbf{x}, \tag{2.17}$$

$$M_\Phi(t) = \int_{\mathbb{R}^d}\left[|\phi_1(\mathbf{x},t)|^2 - |\phi_{-1}(\mathbf{x},t)|^2\right]\,d\mathbf{x}, \tag{2.18}$$

$$R_\Phi(t) = \int_{\mathbb{R}^d}\left[|\phi_1(\mathbf{x},t)|^2 + |\phi_{-1}(\mathbf{x},t)|^2\right]\,d\mathbf{x}, \tag{2.19}$$

$$\begin{aligned} D_\Phi(t) = \int_{\mathbb{R}^d}\Bigg\{ &\sum_{j=-1}^{1}\left(\frac{1}{2}|\nabla\phi_j|^2 + V(\mathbf{x})|\phi_j|^2\right) + 2(\beta_n-\beta_s)|\phi_1|^2|\phi_{-1}|^2 + \beta_n|\phi_0|^4 \\ &+(\beta_n+\beta_s)\Big[|\phi_1|^4 + |\phi_{-1}|^4 + 2|\phi_0|^2\left(|\phi_1|^2+|\phi_{-1}|^2\right)\Big] \\ &+2\beta_s\left(\bar{\phi}_{-1}\phi_0^2\bar{\phi}_1 + \phi_{-1}\bar{\phi}_0^2\phi_1\right)\Bigg\}\,d\mathbf{x}, \end{aligned} \tag{2.20}$$

$$\begin{aligned} F_\Phi(t) = \int_{\mathbb{R}^d}\Bigg\{ &\frac{1}{2}\left(|\nabla\phi_1|^2 - |\nabla\phi_{-1}|^2\right) + V(\mathbf{x})\left(|\phi_1|^2 - |\phi_{-1}|^2\right) \\ &+(\beta_n+\beta_s)\Big[|\phi_1|^4 - |\phi_{-1}|^4 + |\phi_0|^2\left(|\phi_1|^2 - |\phi_{-1}|^2\right)\Big]\Bigg\}\,d\mathbf{x}. \end{aligned} \tag{2.21}$$

For the above CNGF, we have the following.

THEOREM 2.2. *For any given initial data*

$$\Phi(\mathbf{x},0) = (\phi_1(\mathbf{x},0),\phi_0(\mathbf{x},0),\phi_{-1}(\mathbf{x},0))^T := \Phi^{(0)}(\mathbf{x}), \qquad \mathbf{x}\in\mathbb{R}^d, \tag{2.22}$$

satisfying

$$N_\Phi(t=0) := N_{\Phi^{(0)}} = 1, \qquad M_\Phi(t=0) := M_{\Phi^{(0)}} = M, \tag{2.23}$$

the CNGF (2.13)–(2.15) *is mass and magnetization conservative and energy-diminishing, i.e.,*

$$N_\Phi(t) \equiv N_\Phi(t=0) = 1, \qquad M_\Phi(t) \equiv M_\Phi(t=0) = M, \qquad t \geq 0, \tag{2.24}$$

$$E\left(\Phi(\cdot,t)\right) \leq E\left(\Phi(\cdot,s)\right) \qquad \text{for any } t \geq s \geq 0. \tag{2.25}$$

Proof. Differentiating (2.17) with respect to t and noticing (2.13)–(2.15), we have

$$\begin{aligned}
\frac{dN_\Phi(t)}{dt} &= \frac{d}{dt}\int_{\mathbb{R}^d}\sum_{j=-1}^{1}|\phi_j(\mathbf{x},t)|^2\,d\mathbf{x} = \int_{\mathbb{R}^d}\sum_{j=-1}^{1}\Big[\bar\phi_j\,\partial_t\phi_j + \phi_j\,\partial_t\bar\phi_j\Big]\,d\mathbf{x}\\
&= \int_{\mathbb{R}^d}\sum_{j=-1}^{1}\Big(-\bar\phi_j H_j\phi_j - \phi_j\bar H_j\bar\phi_j\Big)d\mathbf{x} + 2\left[\mu_\Phi(t)+\lambda_\Phi(t)\right]\|\phi_1\|^2\\
&\quad +2\mu_\Phi(t)\|\phi_0\|^2 + 2\left[\mu_\Phi(t)-\lambda_\Phi(t)\right]\|\phi_{-1}\|^2\\
&= 2\mu_\Phi(t)\left(\|\phi_1\|^2+\|\phi_0\|^2+\|\phi_{-1}\|^2\right) + 2\lambda_\Phi(t)\left(\|\phi_1\|^2-\|\phi_{-1}\|^2\right)\\
&\quad -\int_{\mathbb{R}^d}\sum_{j=-1}^{1}\bar\phi_j H_j\phi_j\,d\mathbf{x} - \int_{\mathbb{R}^d}\sum_{j=-1}^{1}\phi_j\bar H_j\bar\phi_j\,d\mathbf{x}.
\end{aligned} \tag{2.26}$$

From (2.13)–(2.15) and (2.20), integrating by parts, we have

$$D_\Phi(t) = \int_{\mathbb{R}^d}\sum_{j=-1}^{1}\bar\phi_j H_j\phi_j\,d\mathbf{x} = \int_{\mathbb{R}^d}\sum_{j=-1}^{1}\phi_j\bar H_j\bar\phi_j\,d\mathbf{x}. \tag{2.27}$$

Plugging (2.27) into (2.26) and noticing (2.16), (2.17), and (2.18), we obtain

$$\begin{aligned}
\frac{dN_\Phi(t)}{dt} &= 2\mu_\Phi(t)N_\Phi(t) + 2\lambda_\Phi(t)M_\Phi(t) - 2D_\Phi(t)\\
&= 2N_\Phi(t)\frac{R_\Phi(t)D_\Phi(t) - M_\Phi(t)F_\Phi(t)}{N_\Phi(t)R_\Phi(t) - M_\Phi^2(t)} + 2M_\Phi(t)\frac{N_\Phi(t)F_\Phi(t) - M_\Phi(t)D_\Phi(t)}{N_\Phi(t)R_\Phi(t) - M_\Phi^2(t)}\\
&\quad -2D_\Phi(t)\\
&= 2D_\Phi(t) - 2D_\Phi(t) \equiv 0, \qquad t \geq 0.
\end{aligned} \tag{2.28}$$

Thus the first part in (2.24) can be obtained from (2.28) immediately. Similarly, differentiating (2.18) with respect to t, noticing (2.13), (2.15), (2.16), and (2.21), and integrating by parts, we obtain

$$\begin{aligned}
\frac{dM_\Phi(t)}{dt} &= \frac{d}{dt}\int_{\mathbb{R}^d}\Big[|\phi_1(\mathbf{x},t)|^2 - |\phi_{-1}(\mathbf{x},t)|^2\Big]\,d\mathbf{x}\\
&= \int_{\mathbb{R}^d}\Big[\bar\phi_1\,\partial_t\phi_1 + \phi_1\,\partial_t\bar\phi_1 - \bar\phi_{-1}\,\partial_t\phi_{-1} - \phi_{-1}\,\partial_t\bar\phi_{-1}\Big]d\mathbf{x}\\
&= \int_{\mathbb{R}^d}\Big[-\bar\phi_1\,H_1\phi_1 - \phi_1\,\bar H_1\bar\phi_1 + \bar\phi_{-1}\,H_{-1}\phi_{-1} + \phi_{-1}\,\bar H_{-1}\bar\phi_{-1}\Big]d\mathbf{x}\\
&\quad +2\left[\mu_\Phi(t)+\lambda_\Phi(t)\right]\|\phi_1\|^2 - 2\left[\mu_\Phi(t)-\lambda_\Phi(t)\right]\|\phi_{-1}\|^2\\
&= 2\mu_\Phi(t)\left(\|\phi_1\|^2-\|\phi_{-1}\|^2\right) + 2\lambda_\Phi(t)\left(\|\phi_1\|^2+\|\phi_{-1}\|^2\right)\\
&\quad -\int_{\mathbb{R}^d}\Big[\bar\phi_1\,H_1\phi_1 - \bar\phi_{-1}\,H_{-1}\phi_{-1}\Big]\,d\mathbf{x} - \int_{\mathbb{R}^d}\Big[\phi_1\,\bar H_1\bar\phi_1 - \phi_{-1}\,\bar H_{-1}\bar\phi_{-1}\Big]d\mathbf{x}.
\end{aligned} \tag{2.29}$$

From (2.13)–(2.15) and (2.21), integrating by parts, we have

$$F_\Phi(t) = \int_{\mathbb{R}^d} \left[\bar\phi_1\ H_1\phi_1 - \bar\phi_{-1}\ H_{-1}\phi_{-1}\right] d\mathbf{x} = \int_{\mathbb{R}^d} \left[\phi_1\ \bar H_1\bar\phi_1 - \phi_{-1}\ \bar H_{-1}\bar\phi_{-1}\right] d\mathbf{x}. \tag{2.30}$$

Plugging (2.30) into (2.29) and noticing (2.16), (2.18), and (2.19), we obtain

$$\begin{aligned}\frac{dM_\Phi(t)}{dt} &= 2\mu_\Phi(t)M_\Phi(t) + 2\lambda_\Phi(t)R_\Phi(t) - 2F_\Phi(t)\\ &= 2M_\Phi(t)\frac{R_\Phi(t)D_\Phi(t) - M_\Phi(t)F_\Phi(t)}{N_\Phi(t)R_\Phi(t) - M_\Phi^2(t)} + 2R_\Phi(t)\frac{N_\Phi(t)F_\Phi(t) - M_\Phi(t)D_\Phi(t)}{N_\Phi(t)R_\Phi(t) - M_\Phi^2(t)}\\ &\quad -2F_\Phi(t)\\ &= 2F_\Phi(t) - 2F_\Phi(t) \equiv 0, \qquad t \ge 0. \end{aligned} \tag{2.31}$$

Thus the second part in (2.24) can be obtained from (2.31) immediately. Finally, differentiating (1.14) (with $\Psi = \Phi$) with respect to t and integrating by parts, we have

$$\begin{aligned}\frac{dE(\Phi(t))}{dt} &= \frac{d}{dt}\int_{\mathbb{R}^d}\left\{\sum_{j=-1}^{1}\left(\frac{1}{2}|\nabla\phi_j|^2 + V(\mathbf{x})|\phi_j|^2\right) + (\beta_n - \beta_s)|\phi_1|^2|\phi_{-1}|^2\right.\\ &\quad + \frac{\beta_n}{2}|\phi_0|^4 + \frac{\beta_n+\beta_s}{2}\left[|\phi_1|^4 + |\phi_{-1}|^4 + 2|\phi_0|^2\left(|\phi_1|^2 + |\phi_{-1}|^2\right)\right]\\ &\quad \left. + \beta_s\left(\bar\phi_{-1}\phi_0^2\bar\phi_1 + \phi_{-1}\bar\phi_0^2\phi_1\right)\right\} d\mathbf{x}\\ &= \int_{\mathbb{R}^d}\sum_{j=-1}^{1}\left[\partial_t\phi_j\bar H_j\bar\phi_j + \partial_t\bar\phi_j H_j\phi_j\right]\ d\mathbf{x}. \end{aligned} \tag{2.32}$$

Plugging (2.13)–(2.15) into (2.32) and noticing (2.28) and (2.31), we obtain

$$\begin{aligned}\frac{dE(\Phi(t))}{dt} &= \int_{\mathbb{R}^d}\left[-2|\partial_t\phi_{-1}|^2 + \left(\mu_\Phi(t) - \lambda_\Phi(t)\right)\partial_t|\phi_{-1}|^2 - 2|\partial_t\phi_0|^2 + \mu_\Phi(t)\partial_t|\phi_0|^2\right.\\ &\quad \left. -2|\partial_t\phi_1|^2 + \left(\mu_\Phi(t) + \lambda_\Phi(t)\right)\partial_t|\phi_1|^2\right] d\mathbf{x}\\ &= \mu_\Phi(t)\int_{\mathbb{R}^d}\partial_t\left[|\phi_1|^2 + |\phi_0|^2 + |\phi_{-1}|^2\right] d\mathbf{x} + \lambda_\Phi(t)\int_{\mathbb{R}^d}\partial_t\left[|\phi_1|^2 - |\phi_{-1}|^2\right] d\mathbf{x}\\ &\quad -2\int_{\mathbb{R}^d}\left[|\partial_t\phi_{-1}|^2 + |\partial_t\phi_0|^2 + |\partial_t\phi_1|^2\right] d\mathbf{x}\\ &= \mu_\Phi(t)\frac{dN_\Phi(t)}{dt} + \lambda_\Phi(t)\frac{dM_\Phi(t)}{dt} - 2\int_{\mathbb{R}^d}\left[|\partial_t\phi_{-1}|^2 + |\partial_t\phi_0|^2 + |\partial_t\phi_1|^2\right] d\mathbf{x}\\ &= -2\int_{\mathbb{R}^d}\left[|\partial_t\phi_{-1}|^2 + |\partial_t\phi_0|^2 + |\partial_t\phi_1|^2\right] d\mathbf{x} \le 0, \qquad t \ge 0. \end{aligned} \tag{2.33}$$

Thus the inequality (2.25) can be obtained from (2.33) immediately. ∎

Using an argument similar to that in [33], when $V(\mathbf{x}) \ge 0$ for all $\mathbf{x} \in \mathbb{R}^d$, $\beta_n \ge 0$, $\beta_n \ge |\beta_s|$, and $\Phi_0 \in S$, we may get that as $t \to \infty$, Φ approaches to a steady state solution, which is a critical point of the energy functional $E(\Phi)$ over the set S. When the initial data Φ_0 in (2.22) for the CNGF (2.13)–(2.15) are chosen properly, e.g., its energy is less than that of the first excited state, the ground state Φ_g can be obtained from the steady state solution of the CNGF (2.13)–(2.15), i.e.,

$$\Phi_g(\mathbf{x}) = \lim_{t\to\infty}\Phi(\mathbf{x}, t), \qquad \mathbf{x} \in \mathbb{R}^d. \tag{2.34}$$

3. Mass and magnetization conservative and energy-diminishing numerical discretization. In this section, we present a mass and magnetization conservative and energy-diminishing scheme to discretize the continuous normalized gradient flow (2.13)–(2.15) for computing the ground state of a spin-1 BEC.

3.1. Semidiscretization in time. Choose a time step $k = \Delta t > 0$, and set $t_n = n\Delta t$ for $n = 0, 1, 2, \ldots$. Let $\Phi^n(\mathbf{x}) = (\phi_1^n(\mathbf{x}), \phi_0^n(\mathbf{x}), \phi_{-1}^n(\mathbf{x}))^T$ be the approximation of $\Phi(\mathbf{x}, t_n)$, and denote $\Phi^{n+1/2}(\mathbf{x}) = (\phi_1^{n+1/2}(\mathbf{x}), \phi_0^{n+1/2}(\mathbf{x}), \phi_{-1}^{n+1/2}(\mathbf{x}))^T$ defined as

$$\phi_j^{n+1/2} := \phi_j^{n+1/2}(\mathbf{x}) = \frac{1}{2}\left[\phi_j^{n+1}(\mathbf{x}) + \phi_j^n(\mathbf{x})\right], \qquad j = -1, 0, 1. \tag{3.1}$$

Consider the following implicit semidiscretization scheme for the CNGF (2.13)–(2.15):

$$\begin{aligned}\frac{\phi_1^{n+1}(\mathbf{x}) - \phi_1^n(\mathbf{x})}{\Delta t} = &\left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - \frac{\beta_n + \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)\right.\\ &\left.-\frac{\beta_n - \beta_s}{2}\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2\right)\right]\phi_1^{n+1/2}\\ &-\frac{\beta_s}{2}\left[\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2\right]\bar{\phi}_{-1}^{n+1/2} + \left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2}\right]\phi_1^{n+1/2},\end{aligned} \tag{3.2}$$

$$\begin{aligned}\frac{\phi_0^{n+1}(\mathbf{x}) - \phi_0^n(\mathbf{x})}{\Delta t} = &\left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - \frac{\beta_n + \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2\right)\right.\\ &\left.-\frac{\beta_n}{2}\left(|\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)\right]\phi_0^{n+1/2} - \beta_s\left(\phi_{-1}^{n+1}\phi_1^{n+1} + \phi_{-1}^n\phi_1^n\right)\bar{\phi}_0^{n+1/2}\\ &+\mu_\Phi^{n+1/2}\phi_0^{n+1/2},\end{aligned} \tag{3.3}$$

$$\begin{aligned}\frac{\phi_{-1}^{n+1}(\mathbf{x}) - \phi_{-1}^n(\mathbf{x})}{\Delta t} = &\left[\frac{1}{2}\nabla^2 - V(\mathbf{x}) - \frac{\beta_n + \beta_s}{2}\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)\right.\\ &\left.-\frac{\beta_n - \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2\right)\right]\phi_{-1}^{n+1/2}\\ &-\frac{\beta_s}{2}\left[\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2\right]\bar{\phi}_1^{n+1/2} + \left[\mu_\Phi^{n+1/2} - \lambda_\Phi^{n+1/2}\right]\phi_{-1}^{n+1/2}.\end{aligned} \tag{3.4}$$

Here $\mu_\Phi^{n+1/2}$ and $\lambda_\Phi^{n+1/2}$ are chosen such that the above discretization is mass (or normalization) and magnetization conservative, and they are given as

$$\begin{aligned}\mu_\Phi^{n+1/2} &= \frac{R_\Phi^{n+1/2}D_\Phi^{n+1/2} - M_\Phi^{n+1/2}F_\Phi^{n+1/2}}{N_\Phi^{n+1/2}R_\Phi^{n+1/2} - \left(M_\Phi^{n+1/2}\right)^2},\\ \lambda_\Phi^{n+1/2} &= \frac{N_\Phi^{n+1/2}F_\Phi^{n+1/2} - M_\Phi^{n+1/2}D_\Phi^{n+1/2}}{N_\Phi^{n+1/2}R_\Phi^{n+1/2} - \left(M_\Phi^{n+1/2}\right)^2},\end{aligned} \tag{3.5}$$

with

$$N_\Phi^{n+1/2} = \int_{\mathbb{R}^d} \left[|\phi_{-1}^{n+1/2}(\mathbf{x})|^2 + |\phi_0^{n+1/2}(\mathbf{x})|^2 + |\phi_1^{n+1/2}(\mathbf{x})|^2 \right] d\mathbf{x}, \tag{3.6}$$

$$M_\Phi^{n+1/2} = \int_{\mathbb{R}^d} \left[|\phi_1^{n+1/2}(\mathbf{x})|^2 - |\phi_{-1}^{n+1/2}(\mathbf{x})|^2) \right] d\mathbf{x}, \tag{3.7}$$

$$R_\Phi^{n+1/2} = \int_{\mathbb{R}^d} \left[|\phi_1^{n+1/2}(\mathbf{x})|^2 + |\phi_{-1}^{n+1/2}(\mathbf{x})|^2) \right] d\mathbf{x}, \tag{3.8}$$

$$\begin{aligned} D_\Phi^{n+1/2} = \int_{\mathbb{R}^d} \Bigg\{ & \sum_{j=-1}^{1} \left(\frac{1}{2} |\nabla \phi_j^{n+1/2}|^2 + V(\mathbf{x}) |\phi_j^{n+1/2}|^2 \right) + \frac{\beta_n}{2} \left(|\phi_0^{n+1}|^2 + |\phi_0^n|^2 \right) |\phi_0^{n+1/2}|^2 \\ & + \frac{\beta_n - \beta_s}{2} \left[\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 \right) |\phi_1^{n+1}|^2 + \left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 \right) |\phi_{-1}^{n+1}|^2 \right] \\ & + \beta_s \mathrm{Re} \Big(\phi_{-1}^{n+1/2} [\left(\bar{\phi}_0^{n+1} \right)^2 + \left(\bar{\phi}_0^n \right)^2] \phi_1^{n+1/2} \\ & + \left(\bar{\phi}_0^{n+1/2} \right)^2 \left(\phi_{-1}^{n+1} \phi_1^{n+1} + \phi_{-1}^n \phi_1^n \right) \Big) \\ & + \frac{\beta_n + \beta_s}{2} \Big[\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2 \right) |\phi_1^{n+1/2}|^2 \\ & + \left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2 \right) |\phi_{-1}^{n+1/2}|^2 \\ & + \left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 \right) |\phi_0^{n+1/2}|^2 \Big] \Bigg\} d\mathbf{x}, \end{aligned} \tag{3.9}$$

$$\begin{aligned} F_\Phi^{n+1/2} = \int_{\mathbb{R}^d} \Bigg\{ & \frac{1}{2} \left(|\nabla \phi_1^{n+1/2}|^2 - |\nabla \phi_{-1}^{n+1/2}|^2 \right) + V(\mathbf{x}) \left(|\phi_1^{n+1/2}|^2 - |\phi_{-1}^{n+1/2}|^2 \right) \\ & + \frac{\beta_n - \beta_s}{2} \left[\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 \right) |\phi_1^{n+1/2}|^2 - \left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 \right) |\phi_{-1}^{n+1/2}|^2 \right] \\ & + \frac{\beta_n + \beta_s}{2} \Big[\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2 \right) |\phi_1^{n+1/2}|^2 \\ & - \left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2 \right) |\phi_{-1}^{n+1/2}|^2 \Big] \Bigg\} d\mathbf{x}. \end{aligned} \tag{3.10}$$

For the above semidiscretization (3.2)–(3.4), we have the following.

THEOREM 3.1. *For any given time step $\Delta t > 0$ and initial data $\Phi^{(0)}(\mathbf{x})$ in* (2.22) *satisfying* (2.23), *the semidiscretization* (3.2)–(3.4) *is mass and magnetization conservative and energy-diminishing, i.e.,*

$$N_\Phi^{n+1} := N_\Phi(t_{n+1}) \equiv N_\Phi(t_0 = 0) = N_{\Phi^{(0)}} = 1, \tag{3.11}$$

$$M_\Phi^{n+1} := M_\Phi(t_{n+1}) \equiv M_\Phi(t_0 = 0) = M_{\Phi^{(0)}} = M, \tag{3.12}$$

$$E\left(\Phi^{n+1}\right) \le E\left(\Phi^n\right) \le \cdots \le E\left(\Phi^0\right) = E(\Phi^{(0)}), \qquad n = 0, 1, 2, \ldots. \tag{3.13}$$

Proof. Multiplying (3.2) by $2\bar{\phi}_1^{n+1/2} = \bar{\phi}_1^{n+1} + \bar{\phi}_1^n$, integrating over $\mathbb{R}^d$, and integrating by parts, we have

$$\begin{aligned}\|\phi_1^{n+1}\|^2 = -2\Delta t \int_{\mathbb{R}^d} \Bigg[& \frac{1}{2}|\nabla\phi_1^{n+1/2}|^2 + \frac{\beta_n - \beta_s}{2}\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2\right)|\phi_1^{n+1/2}|^2 \\ & + \frac{\beta_n + \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)|\phi_1^{n+1/2}|^2 + V(\mathbf{x})|\phi_1^{n+1/2}|^2 \\ & + \frac{\beta_s}{2}\bar{\phi}_{-1}^{n+1/2}\left[\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2\right]\bar{\phi}_1^{n+1/2}\Bigg] d\mathbf{x} \\ & + 2\Delta t\left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2}\right]\|\phi_1^{n+1/2}\|^2 + \|\phi_1^n\|^2 + \int_{\mathbb{R}^d}\left[\bar{\phi}_1^{n+1}\phi_1^n - \bar{\phi}_1^n\phi_1^{n+1}\right] d\mathbf{x}. \end{aligned} \tag{3.14}$$

Summing (3.14) with its conjugate and then dividing both sides by 2, we obtain

$$\begin{aligned}\|\phi_1^{n+1}\|^2 = -2\Delta t \int_{\mathbb{R}^d} \Bigg\{ & \frac{1}{2}|\nabla\phi_1^{n+1/2}|^2 + \frac{\beta_n - \beta_s}{2}\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2\right)|\phi_1^{n+1/2}|^2 \\ & + \frac{\beta_n + \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)|\phi_1^{n+1/2}|^2 + V(\mathbf{x})|\phi_1^{n+1/2}|^2 \\ & + \frac{\beta_s}{2}\,\mathrm{Re}\left(\phi_{-1}^{n+1/2}\left[\left(\bar{\phi}_0^{n+1}\right)^2 + \left(\bar{\phi}_0^n\right)^2\right]\phi_1^{n+1/2}\right)\Bigg\} d\mathbf{x} \\ & + \|\phi_1^n\|^2 + 2\Delta t\left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2}\right]\|\phi_1^{n+1/2}\|^2. \end{aligned} \tag{3.15}$$

Applying the same procedure to (3.3) by multiplying $2\bar{\phi}_0^{n+1/2} = \bar{\phi}_0^{n+1} + \bar{\phi}_0^n$, we get

$$\begin{aligned}\|\phi_0^{n+1}\|^2 = -2\Delta t \int_{\mathbb{R}^d} \Bigg\{ & \frac{1}{2}|\nabla\phi_0^{n+1/2}|^2 + V(\mathbf{x})|\phi_0^{n+1/2}|^2 + \frac{\beta_n}{2}\left(|\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)|\phi_0^{n+1/2}|^2 \\ & + \frac{\beta_n + \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2\right)|\phi_0^{n+1/2}|^2 \\ & + \beta_s\,\mathrm{Re}\left(\left(\bar{\phi}_0^{n+1/2}\right)^2\left(\phi_{-1}^{n+1}\phi_1^{n+1} + \phi_{-1}^n\phi_1^n\right)\right)\Bigg\} d\mathbf{x} \\ & + \|\phi_0^n\|^2 + 2\Delta t\,\mu_\Phi^{n+1/2}\,\|\phi_0^{n+1/2}\|^2. \end{aligned} \tag{3.16}$$

Applying the same procedure to (3.4) by multiplying $2\bar{\phi}_{-1}^{n+1/2} = \bar{\phi}_{-1}^{n+1} + \bar{\phi}_{-1}^n$, we have

$$\begin{aligned}\|\phi_{-1}^{n+1}\|^2 = -2\Delta t \int_{\mathbb{R}^d} \Bigg\{ & \frac{1}{2}|\nabla\phi_{-1}^{n+1/2}|^2 + \frac{\beta_n - \beta_s}{2}\left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2\right)|\phi_{-1}^{n+1/2}|^2 \\ & + \frac{\beta_n + \beta_s}{2}\left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2\right)|\phi_{-1}^{n+1/2}|^2 + V(\mathbf{x})|\phi_{-1}^{n+1/2}|^2 \\ & + \frac{\beta_s}{2}\,\mathrm{Re}\left(\phi_{-1}^{n+1/2}\left[\left(\bar{\phi}_0^{n+1}\right)^2 + \left(\bar{\phi}_0^n\right)^2\right]\phi_1^{n+1/2}\right)\Bigg\} d\mathbf{x} \\ & + \|\phi_{-1}^n\|^2 + 2\Delta t\left[\mu_\Phi^{n+1/2} - \lambda_\Phi^{n+1/2}\right]\|\phi_{-1}^{n+1/2}\|^2. \end{aligned} \tag{3.17}$$

Summing (3.15), (3.16), and (3.17) and noticing (3.9), (3.6), and (3.7), we get

$$\begin{aligned}
N_\Phi^{n+1} &= \|\phi_1^{n+1}\|^2 + \|\phi_0^{n+1}\|^2 + \|\phi_{-1}^{n+1}\|^2 \\
&= N_\Phi^n - 2\Delta t\, D_\Phi^{n+1/2} + 2\Delta t \left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2}\right] \|\phi_1^{n+1/2}\|^2 \\
&\quad +2\Delta t\, \mu_\Phi^{n+1/2} \|\phi_0^{n+1/2}\|^2 + 2\Delta t \left[\mu_\Phi^{n+1/2} - \lambda_\Phi^{n+1/2}\right] \|\phi_{-1}^{n+1/2}\|^2 \\
&= N_\Phi^n - 2\Delta t\, D_\Phi^{n+1/2} + 2\Delta t\, \mu_\Phi^{n+1/2} \left[\|\phi_1^{n+1/2}\|^2 + \|\phi_0^{n+1/2}\|^2 + \|\phi_{-1}^{n+1/2}\|^2\right] \\
&\quad +2\Delta t\, \lambda_\Phi^{n+1/2} \left[\|\phi_1^{n+1/2}\|^2 - \|\phi_{-1}^{n+1/2}\|^2\right] \\
&= N_\Phi^n - 2\Delta t\, D_\Phi^{n+1/2} + 2\Delta t\, \mu_\Phi^{n+1/2}\, N_\Phi^{n+1/2} + 2\Delta t\, \lambda_\Phi^{n+1/2}\, M_\Phi^{n+1/2}. \qquad (3.18)
\end{aligned}$$

Plugging (3.5) into (3.18), we obtain

$$\begin{aligned}
N_\Phi^{n+1} &= N_\Phi^n - 2\Delta t\, D_\Phi^{n+1/2} + 2\Delta t\, N_\Phi^{n+1/2}\, \frac{R_\Phi^{n+1/2} D_\Phi^{n+1/2} - M_\Phi^{n+1/2} F_\Phi^{n+1/2}}{N_\Phi^{n+1/2} R_\Phi^{n+1/2} - \left(M_\Phi^{n+1/2}\right)^2} \\
&\quad +2\Delta t\, M_\Phi^{n+1/2}\, \frac{N_\Phi^{n+1/2} F_\Phi^{n+1/2} - M_\Phi^{n+1/2} D_\Phi^{n+1/2}}{N_\Phi^{n+1/2} R_\Phi^{n+1/2} - \left(M_\Phi^{n+1/2}\right)^2} \\
&= N_\Phi^n - 2\Delta t\, D_\Phi^{n+1/2} + 2\Delta t\, D_\Phi^{n+1/2} \\
&= N_\Phi^n, \qquad n = 0, 1, 2, \ldots. \qquad (3.19)
\end{aligned}$$

Thus the mass conservation in (3.11) can be obtained from (3.19) by induction. Subtracting (3.17) from (3.15) and noticing (3.10), (3.6), and (3.8), we have

$$\begin{aligned}
M_\Phi^{n+1} &= \|\phi_1^{n+1}\|^2 - \|\phi_{-1}^{n+1}\|^2 \\
&= \|\phi_1^n\|^2 - \|\phi_{-1}^n\|^2 - 2\Delta t\, F_\Phi^{n+1/2} + 2\Delta t \left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2}\right] \|\phi_1^{n+1/2}\|^2 \\
&\quad -2\Delta t \left[\mu_\Phi^{n+1/2} - \lambda_\Phi^{n+1/2}\right] \|\phi_{-1}^{n+1/2}\|^2 \\
&= M_\Phi^n - 2\Delta t\, F_\Phi^{n+1/2} + 2\Delta t \mu_\Phi^{n+1/2} \left[\|\phi_1^{n+1/2}\|^2 - \|\phi_{-1}^{n+1/2}\|^2\right] \\
&\quad +2\Delta t\, \lambda_\Phi^{n+1/2} \left[\|\phi_1^{n+1/2}\|^2 + \|\phi_{-1}^{n+1/2}\|^2\right] \\
&= M_\Phi^n - 2\Delta t\, F_\Phi^{n+1/2} + 2\Delta t\, \mu_\Phi^{n+1/2} M_\Phi^{n+1/2} + 2\Delta t\, \lambda_\Phi^{n+1/2} R_\Phi^{n+1/2}. \qquad (3.20)
\end{aligned}$$

Plugging (3.5) into (3.20), we obtain

$$\begin{aligned}
M_\Phi^{n+1} &= M_\Phi^n - 2\Delta t\, F_\Phi^{n+1/2} + 2\Delta t\, M_\Phi^{n+1/2}\, \frac{R_\Phi^{n+1/2} D_\Phi^{n+1/2} - M_\Phi^{n+1/2} F_\Phi^{n+1/2}}{N_\Phi^{n+1/2} R_\Phi^{n+1/2} - \left(M_\Phi^{n+1/2}\right)^2} \\
&\quad +2\Delta t\, R_\Phi^{n+1/2}\, \frac{N_\Phi^{n+1/2} F_\Phi^{n+1/2} - M_\Phi^{n+1/2} D_\Phi^{n+1/2}}{N_\Phi^{n+1/2} R_\Phi^{n+1/2} - \left(M_\Phi^{n+1/2}\right)^2} \\
&= M_\Phi^n - 2\Delta t\, F_\Phi^{n+1/2} + 2\Delta t\, F_\Phi^{n+1/2} \\
&= M_\Phi^n, \qquad n = 0, 1, 2, \ldots. \qquad (3.21)
\end{aligned}$$

Thus the magnetization conservation in (3.12) can be obtained from (3.21) by induction. To prove the energy-diminishing property (3.13), multiplying (3.2) by $\hat{\phi}_1^{n+1/2} :=$

$\bar{\phi}_1^{n+1} - \bar{\phi}_1^n$, integrating over $\mathbb{R}^d$, and integrating by parts, we have

$$\begin{aligned}\frac{\|\phi_1^{n+1} - \phi_1^n\|^2}{\Delta t} = &- \int_{\mathbb{R}^d} \left[\frac{1}{2} \nabla \phi_1^{n+1/2} \cdot \nabla \hat{\phi}_1^{n+1/2} + \frac{\beta_s}{2} \bar{\phi}_{-1}^{n+1/2} \left(\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2 \right) \hat{\phi}_1^{n+1/2} \right. \\ &+ V(\mathbf{x}) \phi_1^{n+1/2} \hat{\phi}_1^{n+1/2} + \frac{\beta_n - \beta_s}{2} \left(|\phi_{-1}^{n+1}|^2 + |\phi_{-1}^n|^2 \right) \phi_1^{n+1/2} \hat{\phi}_1^{n+1/2} \\ &\left. + \frac{\beta_n + \beta_s}{2} \left(|\phi_1^{n+1}|^2 + |\phi_1^n|^2 + |\phi_0^{n+1}|^2 + |\phi_0^n|^2 \right) \phi_1^{n+1/2} \hat{\phi}_1^{n+1/2} \right] d\mathbf{x} \\ &+ \left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2} \right] \int_{\mathbb{R}^d} \phi_1^{n+1/2} \hat{\phi}_1^{n+1/2} \, d\mathbf{x}. \end{aligned} \tag{3.22}$$

Summing (3.22) with its conjugate, we obtain

$$\begin{aligned}\frac{2}{\Delta t} \|\phi_1^{n+1} - \phi_1^n\|^2 = &- \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla \phi_1^{n+1}|^2 + V(\mathbf{x}) |\phi_1^{n+1}|^2 + \frac{\beta_n - \beta_s}{2} |\phi_{-1}^{n+1}|^2 |\phi_1^{n+1}|^2 \right. \\ &+ \frac{\beta_n + \beta_s}{2} \left(|\phi_1^{n+1}|^2 + |\phi_0^{n+1}|^2 \right) |\phi_1^{n+1}|^2 - \frac{1}{2} |\nabla \phi_1^n|^2 - V(\mathbf{x}) |\phi_1^n|^2 \\ &- \frac{\beta_n - \beta_s}{2} |\phi_{-1}^n|^2 |\phi_1^n|^2 - \frac{\beta_n + \beta_s}{2} \left(|\phi_1^n|^2 + |\phi_0^n|^2 \right) |\phi_1^n|^2 \\ &+ \frac{\beta_n + \beta_s}{2} \left[\left(|\phi_1^n|^2 + |\phi_0^n|^2 \right) |\phi_1^{n+1}|^2 - \left(|\phi_1^{n+1}|^2 + |\phi_0^{n+1}|^2 \right) |\phi_1^n|^2 \right] \\ &+ \frac{\beta_n - \beta_s}{2} \left[|\phi_{-1}^n|^2 |\phi_1^{n+1}|^2 - |\phi_{-1}^{n+1}|^2 |\phi_1^n|^2 \right] \\ &\left. + \beta_s \operatorname{Re} \left(\bar{\phi}_{-1}^{n+1/2} \left(\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2 \right) \left(\bar{\phi}_1^{n+1} - \bar{\phi}_1^n \right) \right) \right] d\mathbf{x} \\ &+ \left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2} \right] \left(\|\phi_1^{n+1}\|^2 - \|\phi_1^n\|^2 \right). \end{aligned} \tag{3.23}$$

Here we use

$$\begin{aligned}&\phi_1^{n+1/2} \left(\bar{\phi}_1^{n+1} - \bar{\phi}_1^n \right) + \bar{\phi}_1^{n+1/2} \left(\phi_1^{n+1} - \phi_1^n \right) \\ &= \frac{1}{2} \left[\left(\phi_1^{n+1} + \phi_1^n \right) \left(\bar{\phi}_1^{n+1} - \bar{\phi}_1^n \right) + \left(\bar{\phi}_1^{n+1} + \bar{\phi}_1^n \right) \left(\phi_1^{n+1} - \phi_1^n \right) \right] \\ &= |\phi_1^{n+1}|^2 - |\phi_1^n|^2, \end{aligned} \tag{3.24}$$

and

$$\nabla \phi_1^{n+1/2} \cdot \nabla \left(\bar{\phi}_1^{n+1} - \bar{\phi}_1^n \right) + \nabla \bar{\phi}_1^{n+1/2} \cdot \nabla \left(\phi_1^{n+1} - \phi_1^n \right) = |\nabla \phi_1^{n+1}|^2 - |\nabla \phi_1^n|^2. \tag{3.25}$$

Applying the same procedure to (3.3) by multiplying $\bar{\phi}_0^{n+1} - \bar{\phi}_0^n$, we get

$$\begin{aligned}\frac{2}{\Delta t} \|\phi_0^{n+1} - \phi_0^n\|^2 = &- \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla \phi_0^{n+1}|^2 + V(\mathbf{x}) |\phi_0^{n+1}|^2 + \frac{\beta_n}{2} |\phi_0^{n+1}|^4 \right. \\ &+ \frac{\beta_n + \beta_s}{2} \left(|\phi_1^{n+1}|^2 + |\phi_{-1}^{n+1}|^2 \right) |\phi_0^{n+1}|^2 - \frac{1}{2} |\nabla \phi_0^n|^2 - V(\mathbf{x}) |\phi_0^n|^2 \\ &- \frac{\beta_n}{2} |\phi_0^n|^4 - \frac{\beta_n + \beta_s}{2} \left(|\phi_1^n|^2 + |\phi_{-1}^n|^2 \right) |\phi_0^n|^2 \\ &+ \frac{\beta_n + \beta_s}{2} \left[\left(|\phi_1^n|^2 + |\phi_{-1}^n|^2 \right) |\phi_0^{n+1}|^2 - \left(|\phi_1^{n+1}|^2 + |\phi_{-1}^{n+1}|^2 \right) |\phi_0^n|^2 \right] \\ &\left. + \beta_s \operatorname{Re} \left(\left(\phi_{-1}^{n+1} \phi_1^{n+1} + \phi_{-1}^n \phi_1^n \right) \left((\bar{\phi}_0^{n+1})^2 - (\bar{\phi}_0^n)^2 \right) \right) \right] d\mathbf{x} \\ &+ \mu_\Phi^{n+1/2} \left(\|\phi_0^{n+1}\|^2 - \|\phi_0^n\|^2 \right). \end{aligned} \tag{3.26}$$

Applying the same procedure to (3.4) by multiplying $\bar{\phi}_{-1}^{n+1} - \bar{\phi}_{-1}^{n}$, we get

$$\begin{aligned}
\frac{2}{\Delta t}\|\phi_{-1}^{n+1} - \phi_{-1}^{n}\|^2 = &- \int_{\mathbb{R}^d} \Big[\frac{1}{2}|\nabla\phi_{-1}^{n+1}|^2 + V(\mathbf{x})|\phi_{-1}^{n+1}|^2 + \frac{\beta_n - \beta_s}{2}|\phi_{-1}^{n+1}|^2|\phi_1^{n+1}|^2 \\
&+ \frac{\beta_n + \beta_s}{2}\left(|\phi_{-1}^{n+1}|^2 + |\phi_0^{n+1}|^2\right)|\phi_{-1}^{n+1}|^2 - \frac{1}{2}|\nabla\phi_{-1}^{n}|^2 - V(\mathbf{x})|\phi_{-1}^{n}|^2 \\
&- \frac{\beta_n - \beta_s}{2}|\phi_{-1}^{n}|^2|\phi_1^{n}|^2 - \frac{\beta_n + \beta_s}{2}\left(|\phi_{-1}^{n}|^2 + |\phi_0^{n}|^2\right)|\phi_{-1}^{n}|^2 \\
&+ \frac{\beta_n + \beta_s}{2}\Big[\left(|\phi_{-1}^{n}|^2 + |\phi_0^{n}|^2\right)|\phi_{-1}^{n+1}|^2 - \left(|\phi_{-1}^{n+1}|^2 + |\phi_0^{n+1}|^2\right)|\phi_{-1}^{n}|^2\Big] \\
&+ \frac{\beta_n - \beta_s}{2}\Big[|\phi_1^{n}|^2|\phi_{-1}^{n+1}|^2 - |\phi_1^{n+1}|^2|\phi_{-1}^{n}|^2\Big] \\
&+ \beta_s\,\mathrm{Re}\left(\bar{\phi}_1^{n+1/2}\left(\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^{n}\right)^2\right)\left(\bar{\phi}_{-1}^{n+1} - \bar{\phi}_{-1}^{n}\right)\right)\Big]\,d\mathbf{x} \\
&+ \left[\mu_\Phi^{n+1/2} - \lambda_\Phi^{n+1/2}\right]\left(\|\phi_{-1}^{n+1}\|^2 - \|\phi_{-1}^{n}\|^2\right). \qquad (3.27)
\end{aligned}$$

Adding (3.23), (3.26), and (3.27) and noticing (3.19), (3.21), and (1.14) with $\Psi = \Phi^{n+1}$ and $\Psi = \Phi^n$, respectively, we have

$$\begin{aligned}
E(\Phi^{n+1}) &= E(\Phi^n) - \frac{2}{\Delta t}\left[\|\phi_1^{n+1} - \phi_1^n\|^2 + \|\phi_0^{n+1} - \phi_0^n\|^2 + \|\phi_{-1}^{n+1} - \phi_{-1}^n\|^2\right] \\
&\quad + \left[\mu_\Phi^{n+1/2} + \lambda_\Phi^{n+1/2}\right]\left(\|\phi_1^{n+1}\|^2 - \|\phi_1^n\|^2\right) + \mu_\Phi^{n+1/2}\left(\|\phi_0^{n+1}\|^2 - \|\phi_0^n\|^2\right) \\
&\quad + \left[\mu_\Phi^{n+1/2} - \lambda_\Phi^{n+1/2}\right]\left(\|\phi_{-1}^{n+1}\|^2 - \|\phi_{-1}^n\|^2\right) \\
&= E(\Phi^n) - \frac{2}{\Delta t}\left[\|\phi_1^{n+1} - \phi_1^n\|^2 + \|\phi_0^{n+1} - \phi_0^n\|^2 + \|\phi_{-1}^{n+1} - \phi_{-1}^n\|^2\right] \\
&\quad + \mu_\Phi^{n+1/2}\left[\|\phi_1^{n+1}\|^2 + \|\phi_0^{n+1}\|^2 + \|\phi_{-1}^{n+1}\|^2 - \|\phi_1^n\|^2 - \|\phi_0^n\|^2 - \|\phi_{-1}^n\|^2\right] \\
&\quad + \lambda_\Phi^{n+1/2}\left[\|\phi_1^{n+1}\|^2 - \|\phi_{-1}^{n+1}\|^2 - \|\phi_1^n\|^2 + \|\phi_{-1}^n\|^2\right] \\
&= E(\Phi^n) - \frac{2}{\Delta t}\left[\|\phi_1^{n+1} - \phi_1^n\|^2 + \|\phi_0^{n+1} - \phi_0^n\|^2 + \|\phi_{-1}^{n+1} - \phi_{-1}^n\|^2\right] \\
&\quad + \mu_\Phi^{n+1/2}\left[N_\Phi^{n+1} - N_\Phi^n\right] + \lambda_\Phi^{n+1/2}\left[M_\Phi^{n+1} - M_\Phi^n\right] \\
&= E(\Phi^n) - \frac{2}{\Delta t}\left[\|\phi_1^{n+1} - \phi_1^n\|^2 + \|\phi_0^{n+1} - \phi_0^n\|^2 + \|\phi_{-1}^{n+1} - \phi_{-1}^n\|^2\right] \\
&\le E(\Phi^n), \qquad n = 0, 1, 2, \ldots. \qquad (3.28)
\end{aligned}$$

Here we use

$$\begin{aligned}
&\beta_s\,\mathrm{Re}\Big(\left[\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2\right]\left[\bar{\phi}_{-1}^{n+1/2}\left(\bar{\phi}_1^{n+1} - \bar{\phi}_1^n\right) + \bar{\phi}_1^{n+1/2}\left(\bar{\phi}_{-1}^{n+1} - \bar{\phi}_{-1}^n\right)\right] \\
&\qquad + \left(\phi_{-1}^{n+1}\phi_1^{n+1} + \phi_{-1}^n\phi_1^n\right)\left((\bar{\phi}_0^{n+1})^2 - (\bar{\phi}_0^n)^2\right)\Big) \\
&= \beta_s\,\mathrm{Re}\Big(\left[\left(\phi_0^{n+1}\right)^2 + \left(\phi_0^n\right)^2\right]\left(\bar{\phi}_1^{n+1}\bar{\phi}_{-1}^{n+1} - \bar{\phi}_1^n\bar{\phi}_{-1}^n\right) \\
&\qquad + \left(\phi_{-1}^{n+1}\phi_1^{n+1} + \phi_{-1}^n\phi_1^n\right)\left((\bar{\phi}_0^{n+1})^2 - (\bar{\phi}_0^n)^2\right)\Big) \\
&= \beta_s\left[\phi_{-1}^{n+1}\phi_1^{n+1}(\bar{\phi}_0^{n+1})^2 + \bar{\phi}_{-1}^{n+1}\bar{\phi}_1^{n+1}(\phi_0^{n+1})^2\right] \\
&\qquad - \beta_s\left[\phi_{-1}^{n}\phi_1^{n}(\bar{\phi}_0^{n})^2 + \bar{\phi}_{-1}^{n}\bar{\phi}_1^{n}(\phi_0^{n})^2\right]. \qquad (3.29)
\end{aligned}$$

Thus (3.13) can be obtained from (3.28) immediately. ∎

3.2. A fully discretized method. For simplicity of notation, we introduce a fully discretized method for the CNGF (2.13)–(2.15) truncated into a bounded interval $\Omega = [a, b]$ (with $|a|$ and $|b|$ sufficiently large) in the case of one spatial dimension ($d = 1$) with homogeneous Dirichlet boundary conditions

$$\phi_j(a,t) = \phi_j(b,t) = 0, \qquad t \geq 0, \quad j = 1, -0, 1. \tag{3.30}$$

Generalizations to a higher dimension are straightforward for tensor product grids, and the results remain valid without modifications. For $d = 1$, we choose the spatial mesh size $h = \Delta x > 0$, with $\Delta x = (b-a)/L$ and L is an even positive integer. The grid points are defined as $x_l = a + l\,h$ for $l = 0, 1, \cdots, L$, and let $\Phi_l^n = (\phi_{1,l}^n, \phi_{0,l}^n, \phi_{-1,l}^n)^T$ be the numerical approximation of $\Phi(x_j, t_n)$ and Φ_h^n the solution vector at time $t = t_n$ with components Φ_l^n. In addition, denote $\Phi_l^{n+1/2} = (\phi_{1,l}^{n+1/2}, \phi_{0,l}^{n+1/2}, \phi_{-1,l}^{n+1/2})^T$, with $\phi_{j,l}^{n+1/2}$ defined as

$$\phi_{j,l}^{n+1/2} := \frac{1}{2}\left[\phi_{j,l}^{n+1} + \phi_{j,l}^{n}\right], \qquad j = -1, 0, 1, \quad l = 0, 1, 2, \ldots, L. \tag{3.31}$$

Here we propose a full discretization for the CNGF (2.13)–(2.15) in 1D, for $1 \leq l \leq L-1$ and $n \geq 0$, as

$$\begin{aligned}
\frac{\phi_{1,l}^{n+1} - \phi_{1,l}^{n}}{\Delta t} &= \frac{\phi_{1,l+1}^{n+1/2} - 2\phi_{1,l}^{n+1/2} + \phi_{1,l-1}^{n+1/2}}{2h^2} - \frac{\beta_n - \beta_s}{2}\left(|\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2\right)\phi_{1,l}^{n+1/2} \\
&\quad - \left[\frac{\beta_n + \beta_s}{2}\left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 + |\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2\right) + V(x_l)\right]\phi_{1,l}^{n+1/2} \\
&\quad - \frac{\beta_s}{2}\bar{\phi}_{-1,l}^{n+1/2}\left[\left(\phi_{0,l}^{n+1}\right)^2 + \left(\phi_{0,l}^{n}\right)^2\right] + \left[\mu_{\Phi,h}^{n+1/2} + \lambda_{\Phi,h}^{n+1/2}\right]\phi_{1,l}^{n+1/2},
\end{aligned} \tag{3.32}$$

$$\begin{aligned}
\frac{\phi_{0,l}^{n+1} - \phi_{0,l}^{n}}{\Delta t} &= \frac{\phi_{0,l+1}^{n+1/2} - 2\phi_{0,l}^{n+1/2} + \phi_{0,l-1}^{n+1/2}}{2h^2} - \frac{\beta_n}{2}\left(|\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2\right)\phi_{0,l}^{n+1/2} \\
&\quad - \left[\frac{\beta_n + \beta_s}{2}\left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 + |\phi_{-,l}^{n+1}|^2 + |\phi_{-,l}^{n}|^2\right) + V(x_l)\right]\phi_{0,l}^{n+1/2} \\
&\quad - \beta_s\left(\phi_{-1,l}^{n+1}\phi_{1,l}^{n+1} + \phi_{-1,l}^{n}\phi_{1,l}^{n}\right)\bar{\phi}_{0,l}^{n+1/2} + \mu_{\Phi,h}^{n+1/2}\phi_{0,l}^{n+1/2},
\end{aligned} \tag{3.33}$$

$$\begin{aligned}
\frac{\phi_{-1,l}^{n+1} - \phi_{-1,l}^{n}}{\Delta t} &= \frac{\phi_{-1,l+1}^{n+1/2} - 2\phi_{-1,l}^{n+1/2} + \phi_{-1,l-1}^{n+1/2}}{2h^2} - \frac{\beta_n - \beta_s}{2}\left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2\right)\phi_{-1,l}^{n+1/2} \\
&\quad - \left[\frac{\beta_n + \beta_s}{2}\left(|\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2 + |\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2\right) + V(x_l)\right]\phi_{-1,l}^{n+1/2} \\
&\quad - \frac{\beta_s}{2}\bar{\phi}_{1,l}^{n+1/2}\left[\left(\phi_{0,l}^{n+1}\right)^2 + \left(\phi_{0,l}^{n}\right)^2\right] + \left[\mu_{\Phi,h}^{n+1/2} - \lambda_{\Phi,h}^{n+1/2}\right]\phi_{-1,l}^{n+1/2}.
\end{aligned} \tag{3.34}$$

Again, here $\mu_{\Phi,h}^{n+1/2}$ and $\lambda_{\Phi,h}^{n+1/2}$ are chosen such that the above discretization is mass (or normalization) and magnetization conservative, and they are given as

$$\begin{aligned}
\mu_{\Phi,h}^{n+1/2} &= \frac{R_{\Phi,h}^{n+1/2} D_{\Phi,h}^{n+1/2} - M_{\Phi,h}^{n+1/2} F_{\Phi,h}^{n+1/2}}{N_{\Phi,h}^{n+1/2} R_{\Phi,h}^{n+1/2} - \left(M_{\Phi,h}^{n+1/2}\right)^2}, \\
\lambda_{\Phi,h}^{n+1/2} &= \frac{N_{\Phi,h}^{n+1/2} F_{\Phi,h}^{n+1/2} - M_{\Phi,h}^{n+1/2} D_{\Phi,h}^{n+1/2}}{N_{\Phi,h}^{n+1/2} R_{\Phi,h}^{n+1/2} - \left(M_{\Phi,h}^{n+1/2}\right)^2},
\end{aligned} \tag{3.35}$$

with

$$N_{\Phi,h}^{n+1/2} = \sum_{l=0}^{L-1} h \left[|\phi_{-1,l}^{n+1/2}|^2 + |\phi_{0,l}^{n+1/2}|^2 + |\phi_{1,l}^{n+1/2}|^2 \right], \tag{3.36}$$

$$M_{\Phi,h}^{n+1/2} = \sum_{l=0}^{L-1} h \left[|\phi_{1,l}^{n+1/2}|^2 - |\phi_{-1,l}^{n+1/2}|^2 \right], \tag{3.37}$$

$$R_{\Phi,h}^{n+1/2} = \sum_{l=0}^{L-1} h \left[|\phi_{1,l}^{n+1/2}|^2 + |\phi_{-1,l}^{n+1/2}|^2 \right], \tag{3.38}$$

$$\begin{aligned} D_{\Phi,h}^{n+1/2} = h \sum_{l=0}^{L-1} \Bigg\{ & \sum_{j=-1}^{1} \left(\frac{1}{2h^2} |\phi_{j,l+1}^{n+1/2} - \phi_{j,l}^{n+1/2}|^2 + V(x_l) |\phi_{j,l}^{n+1/2}|^2 \right) \\ & + \frac{\beta_n - \beta_s}{2} \left[\left(|\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2 \right) |\phi_{1,l}^{n+1}|^2 + \left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 \right) |\phi_{-1,l}^{n+1}|^2 \right] \\ & + \beta_s \mathrm{Re} \Big(\phi_{-1,l}^{n+1/2} \big[\left(\bar{\phi}_{0,l}^{n+1} \right)^2 + \left(\bar{\phi}_{0,l}^{n} \right)^2 \big] \phi_{1,l}^{n+1/2} + \left(\bar{\phi}_{0,l}^{n+1/2} \right)^2 \big(\phi_{-1,l}^{n+1} \phi_{1,l}^{n+1} \\ & + \phi_{-1,l}^{n} \phi_{1,l}^{n} \big) \Big) + \frac{\beta_n + \beta_s}{2} \left[\left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 + |\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2 \right) |\phi_{1,l}^{n+1/2}|^2 \right. \\ & + \frac{\beta_n}{2} \left(|\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2 \right) |\phi_{0,l}^{n+1/2}|^2 + \left(|\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2 + |\phi_{0,l}^{n+1}|^2 \right. \\ & \left. \left. + |\phi_{0,l}^{n}|^2 \right) |\phi_{-1,l}^{n+1/2}|^2 + \left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 + |\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2 \right) |\phi_{0,l}^{n+1/2}|^2 \right] \Bigg\}, \end{aligned} \tag{3.39}$$

$$\begin{aligned} F_{\Phi,h}^{n+1/2} = h \sum_{l=0}^{L-1} \Bigg\{ & \frac{1}{2h^2} \left(|\phi_{1,l+1}^{n+1/2} - \phi_{1,l}^{n+1/2}|^2 - |\phi_{-1,l+1}^{n+1/2} - \phi_{-1,l}^{n+1/2}|^2 \right) + V(x_l) |\phi_{1,l}^{n+1/2}|^2 \\ & + \frac{\beta_n - \beta_s}{2} \left[\left(|\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2 \right) |\phi_{1,l}^{n+1/2}|^2 - \left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 \right) |\phi_{-1,l}^{n+1/2}|^2 \right] \\ & + \frac{\beta_n + \beta_s}{2} \left[\left(|\phi_{1,l}^{n+1}|^2 + |\phi_{1,l}^{n}|^2 + |\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2 \right) |\phi_{1,l}^{n+1/2}|^2 - V(x_l) |\phi_{-1,l}^{n+1/2}|^2 \right. \\ & \left. - \left(|\phi_{-1,l}^{n+1}|^2 + |\phi_{-1,l}^{n}|^2 + |\phi_{0,l}^{n+1}|^2 + |\phi_{0,l}^{n}|^2 \right) |\phi_{-1,l}^{n+1/2}|^2 \right] \Bigg\}. \end{aligned} \tag{3.40}$$

The homogeneous Dirichlet boundary conditions (3.30) are discretized as

$$\phi_{1,0}^{n+1} = \phi_{1,L}^{n+1} = \phi_{0,0}^{n+1} = \phi_{0,L}^{n+1} = \phi_{-1,0}^{n+1} = \phi_{-1,L}^{n+1} = 0, \qquad n = 0, 1, 2, \ldots. \tag{3.41}$$

The initial conditions (2.22) in 1D are discretized as

$$\phi_{j,l}^{0} = \phi_j(x_l, 0) = \phi_j^{(0)}(x_l), \qquad j = -1, 0, 1, \quad l = 0, 1, 2, \ldots, L. \tag{3.42}$$

For the above full discretization (3.32)–(3.34), we have the following.

THEOREM 3.2. *For any given time step $\Delta t > 0$ and mesh size $h > 0$ as well as initial data $\Phi^{(0)}(\mathbf{x})$ in* (2.22) *satisfying* (2.23)*, the full discretization* (3.32)–(3.34) *for the*

CNGF (2.13)–(2.15) *is mass and magnetization conservative and energy-diminishing, i.e.,*

$$N_{\Phi,h}^n := h\sum_{l=0}^{L-1}\sum_{j=-1}^{1}|\phi_{j,l}^n|^2 \equiv N_{\Phi,h}^0 := h\sum_{l=0}^{L-1}\sum_{j=-1}^{1}|\phi_j^{(0)}(x_l)|^2,$$

$$M_{\Phi,h}^n := h\sum_{l=0}^{L-1}\left[|\phi_{1,l}^n|^2 - |\phi_{-1,l}^{n+1}|^2\right] \equiv M_{\Phi,h}^0 := h\sum_{l=0}^{L-1}\left[|\phi_1^{(0)}(x_l)|^2 - |\phi_{-1}^{(0)}(x_l)|^2\right],$$

$$E_{\Phi,h}^n \le E_{\Phi,h}^{n-1} \le \cdots \le E_{\Phi,h}^0, \quad n = 0,1,2,\ldots, \tag{3.43}$$

where the discretized energy functional is defined as

$$\begin{aligned} E_{\Phi,h}^n = h\sum_{l=0}^{L-1}\Bigg\{ &\sum_{j=-1}^{1}\left(\frac{1}{2h^2}|\phi_{j,l+1}^n - \phi_{j,l}^n|^2 + V(x_l)|\phi_{j,l}^n|^2\right) + (\beta_n - \beta_s)|\phi_{1,l}^n|^2|\phi_{-1,l}^n|^2 \\ &+\frac{\beta_n}{2}|\phi_{0,l}^n|^4 + \frac{\beta_n+\beta_s}{2}\left[|\phi_{1,l}^n|^4 + |\phi_{-1,l}^n|^4 + 2|\phi_{0,l}^n|^2\left(|\phi_{1,l}^n|^2 + |\phi_{-1,l}^n|^2\right)\right] \\ &+\beta_s\left(\bar{\phi}_{-1,l}^n\left(\phi_{0,l}^n\right)^2\bar{\phi}_{1,l}^n + \phi_{-1,l}^n\left(\bar{\phi}_{0,l}\right)^2\phi_{1,l}^n\right)\Bigg\}. \end{aligned} \tag{3.44}$$

Proof. The proof is similar as that for Theorem 3.1 except that we need to replace integrating over $\mathbb{R}^d$ by summation over $0 \le l \le L-1$ and notice

$$\sum_{l=0}^{L-1}\left(\phi_{j,l+1}^{n+1/2} - 2\phi_{j,l}^{n+1/2} + \phi_{j,l-1}^{n+1/2}\right)\, g_l = \sum_{l=0}^{L-1}\left(\phi_{j,l+1}^{n+1/2} - \phi_{j,l}^{n+1/2}\right)\left(g_{l+1} - g_l\right) \tag{3.45}$$

for any g_l $(l = 0,1,2,\ldots,L)$ with $g_0 = g_L = 0$. The details are omitted here. ∎

Remark 3.1. For solving the nonlinear system (3.32)–(3.34), different iterative numerical methods in the literature can be applied. Here we use an efficient way which is easy to be extended to 2D and 3D to solve it iteratively by treating the linear terms implicitly and the nonlinear terms explicitly at each iterative step. For (3.32), the iterative method reads

$$\begin{aligned} \frac{\phi_{1,l}^{n+1,m+1} - \phi_{1,l}^n}{\Delta t} = &\frac{\phi_{1,l+1}^{n+1/2,m+1} - 2\phi_{1,l}^{n+1/2,m+1} + \phi_{1,l-1}^{n+1/2,m+1}}{2h^2} - \alpha_1\,\phi_{1,l}^{n+1,m+1} \\ &+\alpha_1\,\phi_{1,l}^{n+1,m} - \frac{\beta_n - \beta_s}{2}\left(|\phi_{-1,l}^{n+1,m}|^2 + |\phi_{-1,l}^n|^2\right)\phi_{1,l}^{n+1/2,m} \\ &-\left[\frac{\beta_n+\beta_s}{2}\left(|\phi_{1,l}^{n+1,m}|^2 + |\phi_{1,l}^n|^2 + |\phi_{0,l}^{n+1,m}|^2 + |\phi_{0,l}^n|^2\right) + V(x_l)\right]\phi_{1,l}^{n+1/2,m} \\ &-\frac{\beta_s}{2}\bar{\phi}_{-1,l}^{n+1/2,m}\left[\left(\phi_{0,l}^{n+1,m}\right)^2 + \left(\phi_{0,l}^n\right)^2\right] + \left[\mu_{\Phi,h}^{n+1/2,m} + \lambda_{\Phi,h}^{n+1/2,m}\right]\phi_{1,l}^{n+1/2,m}, \end{aligned} \tag{3.46}$$

where $\phi_{1,l}^{n+1,m}$ is the approximation of $\phi_{1,l}^{n+1}$ at the mth iterative step, with $\phi_{1,l}^{n+1,0} = \phi_{1,l}^n$, $\phi_{1,l}^{n+1/2,m+1} := \frac{1}{2}[\phi_{1,l}^{n+1,m+1} + \phi_{1,l}^n]$ and $\phi_{1,l}^{n+1/2,m} := \frac{1}{2}[\phi_{1,l}^{n+1,m} + \phi_{1,l}^n]$ $(j = 0,1,2,\ldots,L)$, and α_1 is a stabilization factor such that the iterative method converges as fast as possible [4]. The other two equations (3.33) and (3.34) can be dealt with in a similar way.

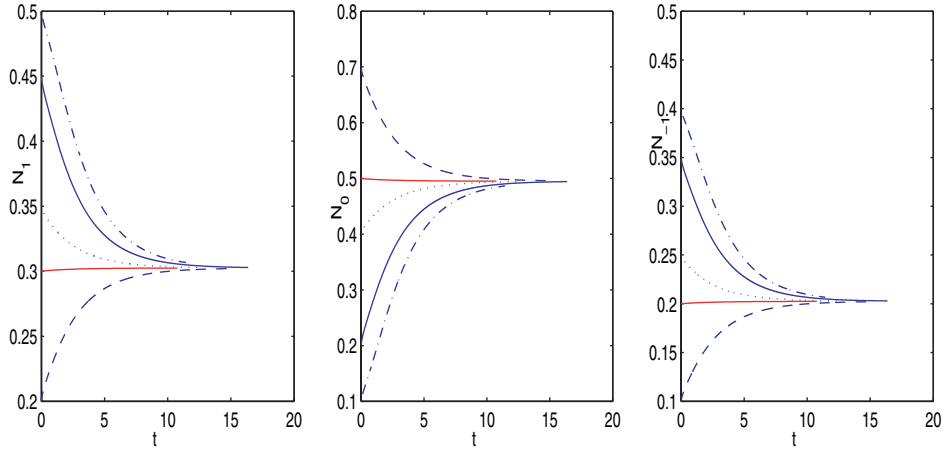


FIG. 1. *Time evolution of* $N_1 = \|\phi_1(\cdot,t)\|^2$ *(left),* $N_0 = \|\phi_0(\cdot,t)\|^2$ *(middle), and* $N_{-1} = \|\phi_{-1}(\cdot,t)\|^2$ *(right) for the full discretization* (3.32)–(3.34) *with* $\beta_n = 87.16$ *and* $\beta_s = -1.7481$ *to analyze the convergence of different initial data in* (4.2) *with* $\alpha = 0.1$ *(dotted-dashed line),* $\alpha = 0.2$ *(solid line),* $\alpha = 0.4$ *(dotted line),* $\alpha = 0.5$ *(horizontal line), and* $\alpha = 0.7$ *(dashed line), respectively.*

4. Numerical results. In this section, we will first study how to choose the initial data in (2.22) for computing the ground state and then test the energy-diminishing property and accuracy of our numerical method. Finally, we apply the method to compute the ground state of a spin-1 BEC with harmonic potential. In our computations, the ground state is reached by using the numerical method (3.32)–(3.34) when $\|\Phi_h^{n+1} - \Phi_h^n\| \le \varepsilon := 10^{-6}$.

In our computations, we choose $d = 1$, $V(x) = x^2/2$, $\beta_n = 0.08716N$, and $\beta_s = -0.0017481N$ in (2.13)–(2.15), with N the number of particles in the condensate. The values for the interaction strengths β_n and β_s correspond to the experimental setup with parameters as follows [34, 24, 25]: $\hbar = 1.054 \times 10^{-34}$[J s], $m = 1.443 \times 10^{-25}$[kg], $\omega_x = 2\pi$[Hz], $\omega_y = 2\pi \times 20\pi\sqrt{2}$[Hz], $\omega_z = 2\pi \times 20\pi\sqrt{2}$[Hz], $a_0 = 5.5$[nm]$= 5.5 \times 10^{-9}$[m], and $a_2 = 5.182$[nm]$= 5.182 \times 10^{-9}$[m], which implies $a_s = \sqrt{\hbar/m\omega_x} = 0.7624 \times 10^{-6}$, $\beta_n \approx \frac{4\pi(a_0+2a_2)N}{3a_s}\frac{\sqrt{\omega_y\omega_z}}{2\pi\omega_x} = 0.08716N$, and $\beta_s \approx \frac{4\pi(a_2-a_0)N}{3a_s}\frac{\sqrt{\omega_y\omega_z}}{2\pi\omega_x} = -0.0017481N$.

4.1. Choice of initial data and energy diminishing. Here we test that the converged solution is independent of different choices of the initial data in (2.22). In order to do so, we take $M = 0.1$ in (2.24) and choose the initial data in (2.22) as

$$\phi_1^{(0)}(x) = \sqrt{0.5(1+M-\alpha)}\frac{1}{\pi^{1/4}}e^{-x^2/2}, \qquad \phi_0^{(0)}(x) = \frac{\sqrt{\alpha}}{\pi^{1/4}}e^{-x^2/2}, \tag{4.1}$$

$$\phi_{-1}^{(0)}(x) = \sqrt{0.5(1-M-\alpha)}\frac{1}{\pi^{1/4}}e^{-x^2/2}, \qquad -\infty < x < \infty, \tag{4.2}$$

where α is a parameter to be determined. We solve the problem (2.13)–(2.15) by our discretization (3.32)–(3.34) on $[-16, 16]$ with time step $\Delta t = 0.01$ and mesh size $h = 1/16$ for different values of α in (4.2). Figure 1 plots the time evolution of $N_j(t) := \|\phi_j(\cdot,t)\|^2$ $(j = 1, 0, -1)$ for different choices of α in (4.2). In addition, Figure 2 shows the time evolution of mass N and magnetization M as well as energy E of our method for the problem with $\alpha = 0.1$ in the initial data (4.1)–(4.2).

From Figure 1 and additional results not shown here, we can see that the converged solution is independent of the choices of initial data in (2.22). In fact, other

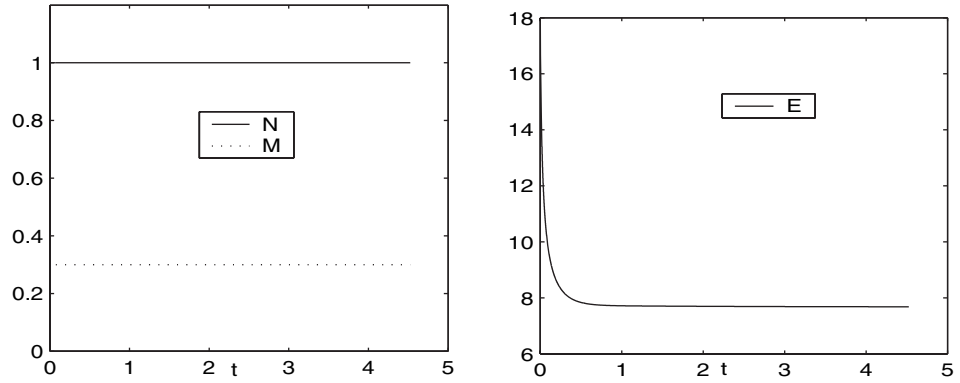


FIG. 2. *Time evolution of the mass N and magnetization M (left) and energy E (right) for the discretization* (3.32)–(3.34) *with* $\beta_n = 87.16$ *and* $\beta_s = -1.7481$ *and initial data* (4.2) *with* $\alpha = 0.1$.

TABLE 1
Spatial error analysis of the ground state for different mesh sizes h and number of particles N in the condensate with fixed magnetization $M = 0.3$.

N	$h = 1/2$	$h = 1/4$	$h = 1/8$	$h = 1/16$	$h = 1/32$
0	1.3336E-2	3.2999E-3	8.0E-4	2.0E-4	5.0E-5
10	4.3021E-3	1.1145E-3	2.5794E-4	6.0940E-5	1.1990E-5
100	1.9063E-3	5.1658E-4	1.3568E-4	2.9750E-5	6.7299E-6
1000	9.5683E-4	2.7421E-4	6.6909E-5	1.6079E-5	3.2299E-6
10000	8.9626E-4	1.8109E-4	4.4159E-5	1.0589E-5	2.1599E-6
30000	6.4606E-4	2.5697E-4	8.5889E-5	3.6030E-5	1.1980E-5

types of initial data are also tested. From our experiments, when $\beta_s \le 0$, for any $\phi_1^{(0)} \ge 0$, $\phi_0^{(0)} \ge 0$, and $\phi_{-1}^{(0)} \ge 0$ in (2.22) satisfying (2.23), we always get the unique positive ground state solution of (1.15). In addition, from Figure 2, the mass N and magnetization M are conserved (cf. Figure 2, left), and energy E is diminishing (cf. Figure 2, right) when time t increases, which confirm the results in Theorem 3.2.

4.2. Accuracy test. Here we test the accuracy of our numerical method (3.32)–(3.34) for computing the ground state of a spin-1 BEC. We choose $M = 0.3$ in (2.24) and $\alpha = 0.1$ in (4.1)–(4.2). For a given set of parameters, the "exact" ground state solution Φ_g is obtained by our numerical method with mesh size $h = 1/64$. Let Φ_g^h be the numerical solution obtained by our method with mesh size h. Table 1 lists the error $\|\Phi_g - \Phi_g^h\|$ for different mesh sizes h and number of particles N in the condensate.

From Table 1, we can see that the full discretization (3.32)–(3.34) is second order in space for computing the ground state of a spin-1 BEC.

4.3. Applications. Now we report the ground state of a spin-1 BEC computed by our numerical method (3.32)–(3.34) for different parameter regimes. In this subsection, the initial data are always taken as in (4.1)–(4.2) with $\alpha = 0.3$, and the bounded computational interval is taken as $[-32, 32]$. We choose mesh size $h = 1/16$ and time step $\Delta t = 0.01$ in (3.32)–(3.34) in our computation.

First, we report the energy of the ground state and study conservation law (2.7) of our numerical ground state. Table 2 shows the numerical kinetic energy $E_{\rm kin}^h := E_{\rm kin}(\Phi_g^h)$ (with Φ_g^h is the numerical ground state), potential energy $E_{\rm pot}^h := E_{\rm pot}(\Phi_g^h)$, interaction energy $E_{\rm int}^h := E_{\rm int}(\Phi_g^h)$, total energy $E_g^h := E(\Phi_g^h)$, and the error $e^h =$

TABLE 2

Different energies of the ground state for different numbers of particles N in the condensate with fixed magnetization $M = 0.2$.

N	$E^h_{\rm kin}$	$E^h_{\rm pot}$	$E^h_{\rm int}$	E^h_g	e^h
0	0.24997	0.25000	0.00000	0.49997	−0.000061
100	0.11046	0.62889	1.03689	1.77618	0.000016
200	0.08175	0.92923	1.69499	2.70597	0.000017
500	0.05321	1.64097	3.17555	5.41056	0.000040
1000	0.03779	2.57116	5.06673	8.01489	−0.000001
2000	0.02654	4.05694	8.06083	12.14431	0.000035
5000	0.01638	7.43899	14.84528	22.30065	0.000049
10000	0.01126	11.74132	23.46028	35.21286	0.000159
15000	0.00907	15.42019	30.82933	46.25859	0.007074
20000	0.00773	18.74087	37.46981	56.21798	0.003524
50000	0.00463	34.39498	68.78410	103.18371	0.003392
100000	0.00312	54.26870	108.5344	162.80622	0.003288

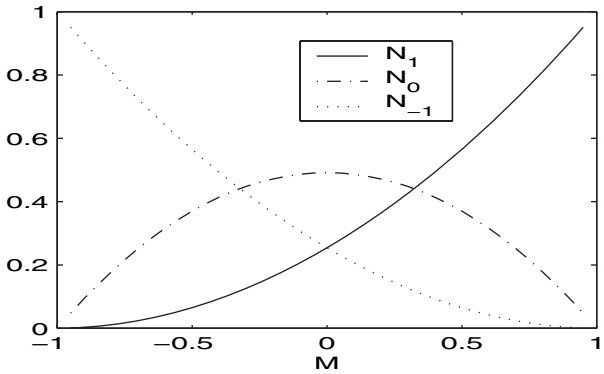


FIG. 3. *Mass of the three components of the ground state, i.e., $N_j = \|\phi_j\|^2$ $(j = 1, 0, -1)$, of a spin-1 BEC with a fixed number of particles $N = 1000$ for different magnetizations $-1 < M < 1$.*

$2E^h_{\rm kin} - 2E^h_{\rm pot} - E^h_{\rm int}$, with magnetization $M = 0.2$ for different numbers of particles N in the condensate.

From Table 2, we can see that, when the number of particles N in the condensate increases, the total energy, potential energy, and interaction energy increases, too, where the kinetic energy decreases. In addition, the relation (2.7) for different energies of the ground state is kept very well in our numerical results.

Second, we report the ground state wave functions for different magnetizations M and numbers of particles N in the condensate. Figures 3 and 4 plot the mass of the three components and wave functions of the ground states of a spin-1 BEC with a fixed number of particles $N = 1000$ in the condensate for different magnetizations M, respectively. In addition, Figure 5 depicts the wave functions of the ground state of a spin-1 BEC with fixed magnetization $M = 0.1$ for different numbers of particles N in the condensate.

From Figure 3, we can see that, for a fixed number of particles N in the condensate, when the magnetization M increases from -1 to 1, the mass N_1 increases from 0 to 1, the mass N_{-1} decreases from 1 to 0, and the mass N_0 increases from 0 to its maximum when $-1 \le M \le 0$, attains its maximum when $M = 0$, and decreases from its maximum to 0 when $0 \le M \le 1$. From Figures 4 and 5, we can see that the ground states are positive functions when $\beta_s \le 0$.

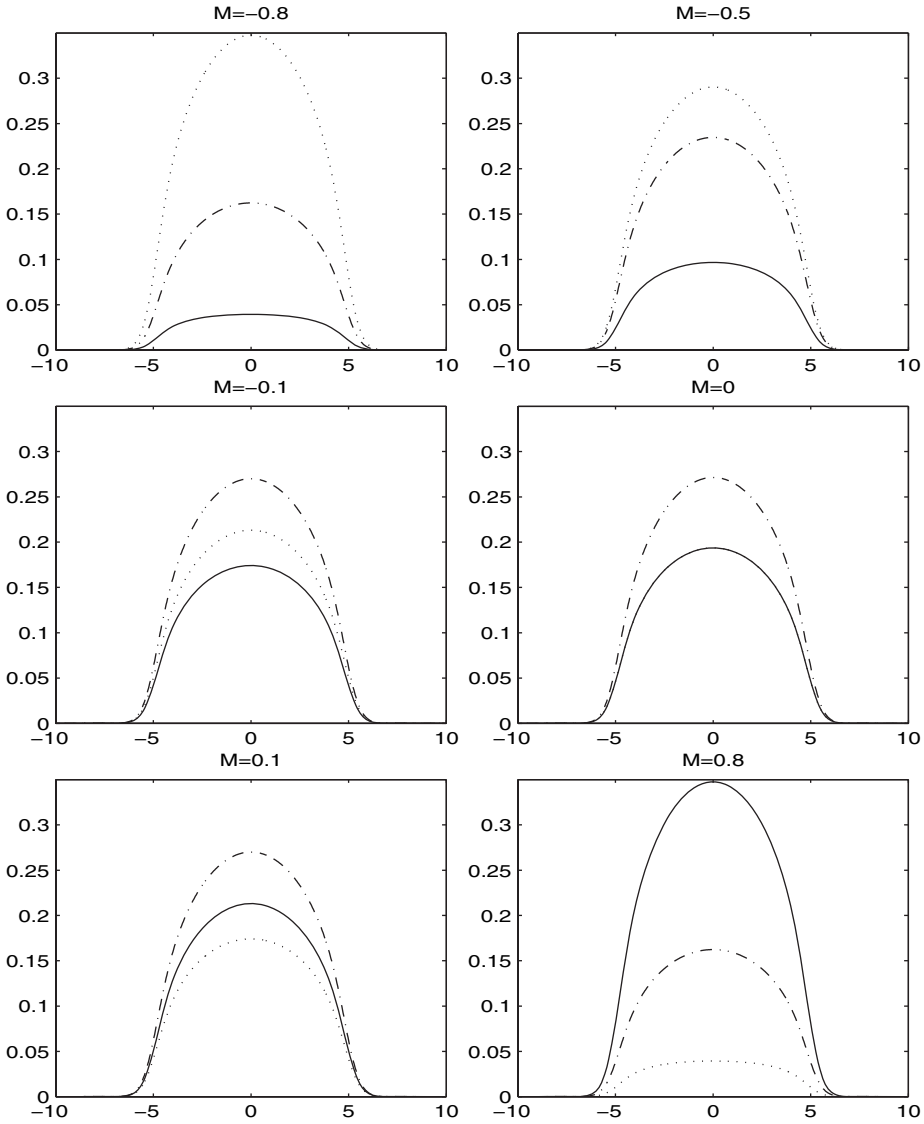


FIG. 4. *Wave functions of the ground state, i.e., $\phi_1(x)$ (solid line), $\phi_0(x)$ (dashed-dotted line), and $\phi_{-1}(x)$ (dotted line), of a spin-1 BEC with a fixed number of particles $N = 1000$ in the condensate for different magnetizations $M = -0.8, -0.5, -0.1, 0, 0.1, 0.8$.*

5. Conclusion. We have proposed an efficient and determinate numerical method for computing the ground state of a spin-1 BEC. By constructing a CNGF which is mass and magnetization conservative and energy-diminishing, the ground state of a spin-1 BEC can be computed as the steady state solution of the CNGF. The CNGF was then discretized in space by the finite difference method and in time by the Crank–Nicolson method with a proper way to deal with the nonlinear terms, and we proved rigorously that the discretization is mass and magnetization conservative and energy-diminishing in the discretized level. Numerical results were reported to demonstrate the efficiency of our new numerical method for computing the ground

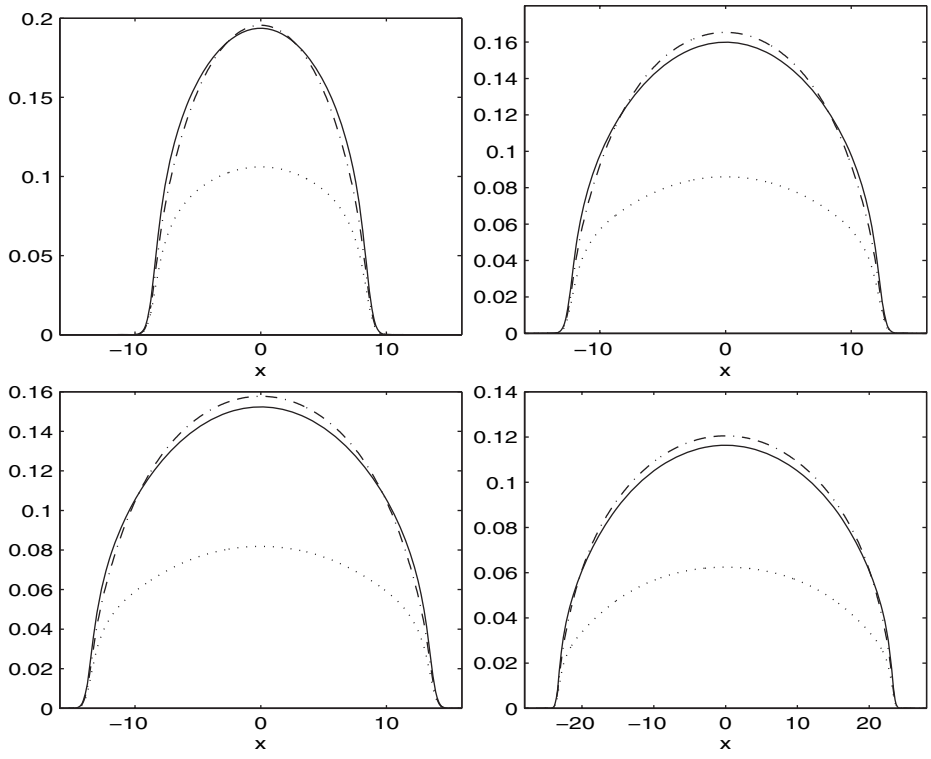


FIG. 5. *Wave functions of the ground state, i.e.* $\phi_1(x)$ *(solid line),* $\phi_0(x)$ *(dashed-dotted line) and* $\phi_{-1}(x)$ *(dotted line), of spin-*1 *BEC with fixed magnetization* $M = 0.1$ *for different number of particles* $N = 5000$ *(top left),* $N = 10000$ *(top right),* $N = 20000$ *(down left) and* $N = 100000$ *(down right), in the condensate.*

state of a spin-1 BEC. In the future we plan to study physically more complex systems based on our new numerical method and extend our method to compute the ground state of a spin-2 [15, 32] and spin-3 [30] Bose–Einstein condensates.

Acknowledgments. We thank I-Liang Chern and Libin Fu for the stimulating discussions. In addition, we also thank the referees for their valuable comments and suggestions to improve the paper.

REFERENCES

[1] A. AFTALION AND Q. DU, *Vortices in a rotating Bose-Einstein condensate: Critical angular velocities and energy diagrams in the Thomas-Fermi regime*, Phys. Rev. A, 64 (2001), article 063603.

[2] M. H. ANDERSON, J. R. ENSHER, M. R. MATTHEWA, C. E. WIEMAN, AND E. A. CORNELL, *Observation of Bose-Einstein condensation in a dilute atomic vapor*, Science, 269 (1995), pp. 198–201.

[3] W. BAO, *Ground states and dynamics of multicomponent Bose–Einstein condensates*, Multiscale Model. Simul., 2 (2004), pp. 210–236.

[4] W. BAO, I.-L. CHERN, AND F. Y. LIM, *Efficient and spectrally accurate numerical methods for computing ground and first excited states in Bose-Einstein condensates*, J. Comput. Phys., 219 (2006), pp. 836–854.

[5] W. BAO AND Q. DU, *Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow*, SIAM J. Sci. Comput., 25 (2004), pp. 1674–1697.

[6] W. Z. Bao, D. Jaksch, and P. A. Markowich, *Numerical solution of the Gross-Pitaevskii equation for Bose-Einstein condensation*, J. Comput. Phys., 187 (2003), pp. 318–342.
[7] W. Bao and W. Tang, *Ground state solution of trapped interacting Bose-Einstein condensate by directly minimizing the energy functional*, J. Comput. Phys., 187 (2003), pp. 230–254.
[8] W. Bao, H. Wang, and P. A. Markowich, *Ground state, symmetric and central vortex state in rotating Bose-Einstein condensate*, Commun. Math. Sci., 3 (2005), pp. 57–88.
[9] C. C. Bradley, C. A. Sackett, J. J. Tollett, and R. G. Hulet, *Evidence of Bose-Einstein condensation in an atomic gas with attractive interactions*, Phys. Rev. Lett., 75 (1995), pp. 1687–1690.
[10] M. Chang, Q. Qin, W. Zhang, L. You, and M. S. Chapman, *Coherent spinor dynamics in a spin-1 Bose condensate*, Nature Physics, 1 (2005), pp. 111–116.
[11] M. L. Chiofalo, S. Succi, and M. P. Tosi, *Ground state of trapped interacting Bose-Einstein condensates by an explicit imaginary-time algorithm*, Phys. Rev. E, 62 (2000), article 7438.
[12] F. Dalfovo, S. Giorgini, L. P. Pitaevskii, and S. Stringari, *Theory of Bose-Einstein condensation in trapped gases*, Rev. Mod. Phys., 71 (1999), pp. 463–512.
[13] K. B. Davis, M. O. Mewes, M. R. Andrews, N. J. van Druten, D. S. Durfee, D. M. Kurn, and W. Ketterle, *Bose-Einstein condensation in a gas of sodium atoms*, Phys. Rev. Lett., 75 (1995), pp. 3969–3973.
[14] E. V. Goldstein and P. Meystre, *Phase conjugation of multicomponent Bose-Einstein condensates*, Phys. Rev. A, 59 (1999), pp. 1509–1513.
[15] A. Görlitz, T. L. Gustavson, A. E. Leanhardt, E. Löw, A. P. Chikkatur, S. Gupta, S. Inouye, D. E. Pritchard, and W. Ketterle, *Soium Bose-Einstein condensates in the $F = 2$ state in a large-volume optical trap*, Phys. Rev. Lett., 90 (2003), article 090401.
[16] D. S. Hall, M. R. Matthews, J. R. Ensher, C. E. Wieman, and E. A. Cornell, *Dynamics of component separation in a binary mixture of Bose-Einstein condensates*, Phys. Rev. Lett., 81 (1998), pp. 1539–1542.
[17] T. L. Ho, *Spinor Bose condensates in optical traps*, Phys. Rev. Lett., 81 (1998), pp. 742–745.
[18] W. J. Huang and S. C. Gou, *Ground state energy of the $F = 1$ spinor Bose-Einstein condensates*, Phys. Rev. A, 59 (1999), pp. 4608–4613.
[19] T. Isoshima, K. Machida, and T. Ohmi, *Spin-domain formation in spinor Bose-Einstein condensation*, Phys. Rev. A, 60 (1999), pp. 4857–4863.
[20] T. Isoshima, M. Nakahara, T. Ohmi, and K. Machida, *Creation of a persistent current and vortex in a Bose-Einstein condensate of alkali-metal atoms*, Phys. Rev. A, 61 (2000), article 063610.
[21] K. Kasamatsu, M. Tsubota, and M. Ueda, *Vortices in multicomponent Bose-Einstein condensates*, Internat. J. Modern Phys. B, 19 (2005), p. 1835.
[22] T. Kita, T. Mizushima, and K. Machida, *Spinor Bose-Einstein condensates with many vortices*, Phys. Rev. A, 66 (2002), article 061601.
[23] L. Laudau and E. Lifschitz, *Quantum Mechanics: Non-relativistic Theory*, Pergamon Press, New York, 1977.
[24] H. J. Miesner, D. M. Stamper-Kurn, J. Stenger, S. Inouye, A. P. Chikkatur, and W. Ketterle, *Observation of metastable states in spinor Bose-Einstein condensates*, Phys. Rev. Lett., 82 (1999), pp. 2228–2231.
[25] T. Mizushima, N. Kobayashi, and K. Machida, *Coreless and singular vortex lattices in rotating spinor Bose-Einstein condensates*, Phys. Rev. A, 70 (2004), article 043613.
[26] T. Mizushima, K. Machida, and T. Kita, *Axisymmetric versus non-axisymmetric vortices in spinor Bose-Einstein condensates*, Phys. Rev. A, 66 (2002), article 053610.
[27] T. Ohmi and K. Machida, *Bose-Einstein condensation with internal degrees of freedom in alkali atom gases*, J. Phys. Soc. Japan, 67 (1998), pp. 1822–1825.
[28] C. J. Pethick and H. Smith, *Bose-Einstein Condensation in Dilute Gases*, Cambridge University Press, Cambridge, 2002.
[29] L. P. Pitaevskii and S. Stringari, *Bose-Einstein Condensation*, Clarendon Press, Oxford, 2003.
[30] L. Santos and T. Pfau, *Spin-3 chromium Bose Einstein condensates*, Phys. Rev. Lett., 96 (2006), article 190404.
[31] H. Schmaljohann, M. Erhard, J. Kronjager, K. Sengstock, and K. Bongs, *Dynamics and thermodynamics in spinor quantum gases*, Appl. Phys. B, 79 (2004), pp. 1001–1007.
[32] H. Schmaljohann, M. Erhard, J. Kronjager, M. Kottke, S. van Staa, L. Cacciapuoti, J. J. Arlt, K. Bongs, and K. Sengstock, *Dynamics of $F = 2$ spinor Bose-Einstein condensates*, Phys. Rev. Lett., 92 (2004), article 040402.
[33] L. Simon, *Asymptotics for a class of nonlinear evolution equations, with applications to geometric problems*, Ann. Math., 118 (1983), pp. 525–571.

[34] D. M. Stamper-Kurn and W. Ketterle, *Spinor condensates and light scattering from Bose-Einstein condensates*, in Proceedings of Les Houches 1999 Summer School (Session LXXII).
[35] J. Stenger, S. Inouye, D. M. Stamper-Kurn, H.-J. Miesner, A. P. Chikkatur, and W. Ketterle, *Spin domains in ground-state Bose Einstein condensates*, Nature, 396 (1998), pp. 345–348.
[36] H. Wang, *Quantized Vortex States and Dynamics in Bose-Einstein Condensates*, Ph.D. thesis, Department of Mathematics, National University of Singapore, 2006.
[37] S. Yi, Ö. E. Müstecaphoğlu, C. P. Sun, and L. You, *Single-mode approximation in a spinor-*1 *atomic condensate*, Phys. Rev. A, 66 (2002), article 011601.
[38] W. Zhang, S. Yi, and L. You, *Mean field ground state of a spin-*1 *condensate in a magnetic field*, New J. Phys., 5 (2003), pp. 77–89.
[39] W. Zhang and L. You, *An effective quasi-one-dimensional description of a spin-*1 *atomic condensate*, Phys. Rev. A, 71 (2005), article 025603.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2201–2227

METHODS FOR RELIABLE TOPOLOGY CHANGES FOR PERIMETER-REGULARIZED GEOMETRIC INVERSE PROBLEMS*

BENJAMIN HACKL†

Abstract. This paper is devoted to the incorporation of topological derivativelike expansions into level set methods for perimeter-regularized geometric inverse problems. The expansions are done up to the second order with respect to the Lebesgue measure of the symmetric difference. They provide simpler shape functionals, still including the perimeter, and therefore allow the construction of steepest descent- and Newton-type algorithms to force topology changes during the level set evolution. Numerous numerical examples are provided that show the strong and also the weak points of the newly developed algorithms.

Key words. geometric inverse problems, perimeter regularization, topological derivatives, level set methods

AMS subject classifications. 35R30, 49Q10, 74B05, 65J20

DOI. 10.1137/060652208

1. Introduction.

Identification of unknown geometries via minimizing appropriate objective functionals is a challenging task, appearing in various applications ranging from topology optimization (cf. Bendsøe and Sigmund [11], Bourdin and Chambolle [12]) over image processing (cf. Tsai and Osher [46]) to inverse problems (cf. Burger and Osher [18], Habib and Hyeonbae [29]). Later it got very common to use level set methods (cf. Osher and Fedkiw [40], Litman, Lesselier, and Santosa [36]) whose velocities depend on shape derivatives (cf. Delfour and Zolésio [22]) to solve such geometric problems. For several optimization problems, these level set methods were successfully applied to compute optimal geometries without a priori knowledge of the number of connected components (cf. Burger [13, 15], Dorn, Miller, and Rapport [23], Hintermüller and Ring [33], Ito, Kunisch, and Li [34], Santosa et al. [36, 41, 42]).

Level set methods are gradientlike methods that allow a simple and flexible geometry representation and evolution. Hence the topologies can only split, merge, and vanish during the level set evolution. A sudden appearance of a new component during the evolution is not possible. Due to the fact that level set methods just evolve the boundary of a geometry, they may easily get stuck in local minima. Theoretical constructed examples prove this, but it is even observed practically (cf. Allaire, Jouve, and Toader [3, 4], Burger, Hackl, and Ring [16]).

Recently a new concept called topological derivatives (cf. Eschenauer et al. [24, 25], Sokolowski and Żochowski [43, 44]) was developed. In this concept one considers the variation of an objective functional with respect to the introduction of infinitesimally small holes at a certain point. The topological derivative then indicates whether it is favorable to introduce a hole at this point or not. Already the definition of the topological derivative suggests an algorithm that was successfully applied to several problems (cf. Amstutz et al. [7, 8, 9], Guillaume and Idris [27], Guzina and Bonnet

*Received by the editors February 16, 2006; accepted for publication (in revised form) June 26, 2007; published electronically September 28, 2007. This work was supported by the Austrian National Science Foundation (FWF) under grant SFB F 013/08 and the Johann Radon Institute for Computational and Applied Mathematics (RICAM).

http://www.siam.org/journals/sinum/45-5/65220.html

†Industrial Mathematics Competence Center (IMCC), Altenbergerstr. 69, A-4040 Linz, Austria (hackl@mathconsult.co.al).

[28], Masmoudi et al. [9, 37]). Also algorithms based just on topological derivatives may get stuck in local minima. This is mainly due to their "disability" to reduce the number of connected components.

Hence several authors (cf. Allaire, de Gournay, and Toader [2], Burger, Hackl, and Ring [16], Hintermüller [32], Wang, Mei, and Wang [47]) tried successfully to combine classical level set methods with the concept of topological derivatives. There are basically two ideas about how to combine these methods. One idea is to add an additional source term to the right-hand side of the level set methods. This additional source term depends on the topological derivative and is defined in the whole domain. Therefore the modified level set methods allow also for the addition of new components. The second idea is to restart the level set evolution after some fixed time (or due to some stopping criteria). The initial value of the restart is determined via the topological derivative at the last time step. The rationale behind both ideas is to fulfill the combined necessary optimality condition for shape and topological derivatives (see Sokolowski and Żochowski [45]).

Nonetheless there are still some problems. First, in geometric inverse problems one usually uses perimeter regularization, which is not topological differentiable at all. Second, topological derivatives are just indicators. They do not provide information about the size and shape of the preferable topology change which actually decreases the cost functional.

By means of an ill-posed, PDE-constraint inverse shape identification problem we are going to construct local approximations of the perimeter-regularized objective functional such that:

- we can provide local error estimates of first (respectively, second) order with respect to the Lebesgue measure of the symmetric difference;
- the approximated shape functionals are either not PDE-constraint (first order approximation) or the PDE constraint becomes simpler (second order approximation); and
- in the limit, Lebesgue measure to zero, we retrieve the topological derivative of the original problem.

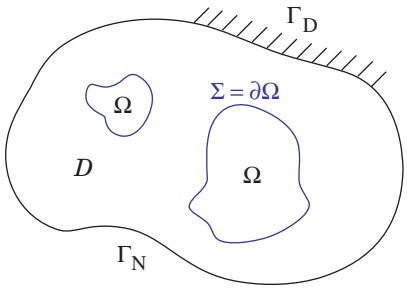
Minimizing the approximated shape functionals will allow one to construct algorithms of the steepest descent type (first order approximation) and the Newton type (second order approximation). Up to provided error estimates every minimizer of the approximated shape functional guarantees a descent of the original perimeter-regularized objective functional. This is in contrast to most methods relying on topological derivatives. Like in the functional analytic framework, the numerical minimization of the first order approximation is less expensive than the second order approximation but also provides less accurate error estimates close to the solution.

Model problem. Let $\mathcal{D} \subset \mathbb{R}^d$ be some bounded open Lipschitz set, and let $\Omega \in \mathcal{F}_L(\mathcal{D})$, with $\mathcal{F}_L(\mathcal{D}) = \big\{\Omega \subset \operatorname{cl}(\mathcal{D}) \,\big|\, \Omega \text{ open and measurable}\big\}$. Consider the partial differential equation

$$\begin{aligned} -\Delta u + c_\Omega\, u &= f && \text{in } \mathcal{D}, \\ \frac{\partial u}{\partial n} &= h && \text{on } \Gamma_N, \\ u &= g && \text{on } \Gamma_D, \end{aligned} \tag{1.1}$$

where $c_\Omega = \underline{c} + (\overline{c} - \underline{c})\chi_\Omega$ and χ_Ω is the characteristic function of Ω. Furthermore assume that the source term $f \in L^2(\mathcal{D})$, the Neumann boundary term h is in $H^{-\frac{1}{2}}(\Gamma_N)$, and the Dirichlet term g is in $H^{\frac{1}{2}}(\Gamma_D) = H^1(\mathcal{D})|_{\Gamma_D}$. For every set $\Omega \in \mathcal{F}_L(\mathcal{D})$ the

partial differential equation (1.1) provides an unique solution $u \in H^1(\mathcal{D})$.



Now let $\Gamma_M \subset \mathcal{D}$ be some open set or $\Gamma_M \subset \Gamma_N$ with positive $(d-1)$-dimensional Lebesgue measure. Denote by $\hat{u}$ measurements of $u|_{\Gamma_M}$ restricted to Γ_M with possible additional Gaussian noise bounded in the L^2-norm. The geometric inverse problem is to identify the set $\Omega \in \mathcal{F}_L(\mathcal{D})$ from measurements $\hat{u}$.

In general the solution to the above geometric inverse problem is not stable. A stabilized approximation of the original problem is the minimization of the perimeter-regularized least squares functional

$$J_\alpha(\Omega) = \frac{1}{2}\int_{\Gamma_M} |u - \hat{u}|^2 \, ds + \alpha \mathrm{Per}\big(\Omega\big), \tag{1.2}$$

where α acts as regularization parameter and is chosen in dependence of the noise level of the measurements $\hat{u}$. Note that the perimeter of a set is defined by

$$\mathrm{Per}\big(\Omega\big) = |\chi_\Omega|_{\mathrm{BV}(\mathcal{D})} = \sup_{\substack{\phi \in C_0^1(\mathcal{D},\mathbb{R}^d) \\ \|\phi\|_\infty \le 1}} \int_{\mathcal{D}} \chi_\Omega \mathrm{div}\phi \, dx \,.$$

The regularization property of the perimeter and the choice of the parameter α will not be dealt with in this paper (cf. Ben Ameur, Burger, and Hackl [10] for a detailed analysis). Only the appearance of the perimeter in the minimization functional will be in the focus in the following, since it prevents the application of known solution methods based on topological derivatives.

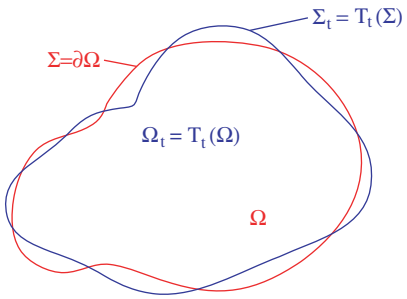
Notation. We denote by $L^p(\mathcal{D})$ the space of functions on Ω whose pth power is integrable, with $H^k(\mathcal{D})$ the Sobolev space of k-times differentiable functions whose derivatives up to order k are in $L^2(\mathcal{D})$. Note that the space $H^k(\mathcal{D})$ is a Hilbert space. Furthermore we abbreviate by $H^1_{D,0}(\mathcal{D}) \subset H^1(\mathcal{D})$ the function space with boundary values zero at the boundary $\Gamma_D \subset \partial\mathcal{D}$. We often use the notation $\preceq$, which means $\le$ up to a multiplicative constant that does not depend on the important properties. With ∂_u we denote the partial derivative with respect to u. Finally we denote with $\tilde{\Omega}\Delta\Omega = (\tilde{\Omega} \setminus \Omega) \cup (\Omega \setminus \tilde{\Omega}) = (\tilde{\Omega} \cup \Omega) \setminus (\tilde{\Omega} \cap \Omega)$ the symmetric difference of sets.

The paper is organized as follows: In section 2 we provide the shape and the topological derivative for the objective functional (1.2). Then, based on the proof of the topological derivative, we provide in section 3 the first and second order approximations, in volume and perimeter, of the objective functional (1.2). The first and second order approximations allow one to construct steepest descent (respectively, Newton-) type iterations which allow topology changes. Details about the numerical implementation of level set methods and the newly suggested steepest descent- as well as Newton-type iterations are provided in section 4. In section 5 we provide

some numerical examples to show the applicability and performance of the methods suggested in this paper. Finally we draw the conclusion in section 6.

2. Shape and topological derivatives. In this section we recall two different concepts of shape (geometry) perturbations and consider the sensitivity of the shape functional (1.2) with respect to these perturbations. The first perturbation is a pure boundary perturbation moving a shape (geometry) in a velocity field V. This approach results in the concept of shape derivatives. For a comprehensive introduction to this topic we refer to Delfour and Zolésio [22]. The second perturbation changes the topology of the shape (geometry) by introducing a fixed additional shape with varying size and position. The sensitivity of the shape functional (1.2) with respect to the size of the newly introduced shape results in the concept of topological derivatives, which were first introduced by Eschenauer et al. [24, 25] and by Céa et al. [19] in the context of topology optimization and made mathematically rigorous by Sokolowski and Żochowski [43, 44].

2.1. Shape derivatives. Shape derivatives for geometric problems allow one to characterize extrema and yield directions of steepest descent. They take the role of *Gateaux* and *Fréchet derivatives* in a functional analytic framework.



The basic idea is to define a perturbation of a domain Ω (piecewise C^2) via the time evolution of Ω in a vector field $V : \mathbb{R}_0^+ \times \mathbb{R}^d \to \mathbb{R}^d$, with

$$\begin{aligned} &\exists\, \tau > 0 \,\forall\, x \in \mathbb{R}^d \,:\, V(.,x) \in C([0,\tau], \mathbb{R}^d) \\ &\exists\, L > 0 \,\forall\, x, y \in \mathbb{R}^d \,: \\ &\quad \|V(.,y) - V(.,x)\|_{C([0,\tau],\mathbb{R}^d)} \leq L|y - x|\,. \end{aligned} \tag{2.1}$$

Then the perturbed domain is set to $\Omega_t(V) = T_t(\Omega, V)$, where $T_t(., V)$ is the solution map (the flow) of the dynamical system

$$\begin{aligned} \frac{dT_t(x,V)}{dt} &= V(t, T_t(x,V)), \\ T_0(x,V) &= x. \end{aligned} \tag{2.2}$$

With these perturbations we are able to define the Eulerian semiderivative of a shape functional $J(\Omega)$ as

$$J'(\Omega)[V] = \frac{d}{dt} J(T_t(\Omega, V))\big|_{t=0}.$$

When this semiderivative is linear and continuous for all velocities $V = t\theta$, with $\theta \in C^{0,1}(\mathcal{D}, \mathbb{R}^d)$, we call the shape functional $J(\Omega)$ shape differentiable.

A basic structure theorem (cf. Novruzi and Pierre [38], Delfour and Zolésio [22]) proves that the shape derivative depends only on $V|_{\partial\Omega}$. Furthermore, for smooth shapes, the perturbation vector field V can be decomposed into a normal and a tangential component on $\partial\Omega$, where the tangential component leaves Ω invariant. Hence the shape derivative is independent of the tangential component, and we obtain

$$J'(\Omega)[V] = J'(\Omega)[(V.n)n].$$

In the case that $J(\Omega)$ is an objective functional in a minimization problem a necessary condition for the shape Ω to be optimal is

$$\forall\, V \,:\, J'(\Omega)[V] \,=\, 0.$$

When the shape derivative is not zero we can construct a velocity V such that the objective functional decreases. This allows the construction of gradientlike descent algorithms as level set methods (see section 4.1).

To calculate the shape derivative of the shape functional J_α (1.2) we need the shape derivative of domain (respectively, boundary) integrals and the solution of the partial differential equation (1.1). These derivatives are well known in the literature (cf. Delfour and Zolésio [22] for shape derivatives of domain and boundary integrals and Hettlich and Rundell [31] for the shape derivative of (1.1)), and we just state them in the following theorems.

THEOREM 2.1 (shape derivative domain and boundary integrals). *Let Ω be an open, bounded measurable domain of class C^2 with boundary $\Sigma = \partial\Omega$, and assume that $V \in C^0\big([0,\tau], C^1_{\rm loc}(\mathbb{R}^d,\mathbb{R}^d)\big)$ fulfill (2.1). Furthermore assume $\varphi \in C\big(0,\tau, W^1_{\rm loc}(\mathbb{R}^d) \cap C^1(0,\tau, H^2_{\rm loc}(\mathbb{R}^d))\big)$. Then the semiderivative of the shape functionals*

$$J_D(\Omega_t) \,:=\, \int_{T_t(\Omega,V)} \varphi(t)\,dx, \qquad J_B(\Sigma_t) \,:=\, \int_{T_t(\Sigma,V)} \varphi(t)\,ds$$

at $t = 0$ are given by

$$J'_D(\Omega)[V(0)] = \int_\Omega \frac{d\varphi}{dt}(0)\,dx + \int_\Sigma \varphi(0)V(0).n\,ds,$$

$$J'_B(\Gamma)[V(0)] = \int_\Sigma \frac{d\varphi}{dt}(0) + \Big(\frac{\partial\varphi(0)}{\partial n} + \kappa\varphi(0)\Big)V(0).n\,ds\,,$$

where κ is the mean curvature of Σ.

THEOREM 2.2. *Let Ω be a domain with C^1 boundary and the velocity field V be as in the previous theorem. Then the solution u of (1.1) is shape differentiable and its shape derivative is characterized by the unique solution $u' = u'(\Omega)[V(0)]$ to the transmission problem*

$$\begin{array}{lll}
-\Delta u' + c_\Omega\, u' = 0 & & \text{in } \Omega \cup \mathcal{D} \setminus \overline{\Omega},\\
\Big[\!\Big[\dfrac{\partial u'}{\partial n}\Big]\!\Big] & = -\big[\!\big[c_\Omega\big]\!\big] V(0).n & \text{on } \partial\Omega,\\
\big[\!\big[u'\big]\!\big] & = 0 & \text{on } \partial\Omega,\\
\dfrac{\partial u'}{\partial n} & = 0 & \text{on } \Gamma_N,\\
u' & = 0 & \text{on } \Gamma_D,
\end{array} \tag{2.3}$$

where $[\![\,.\,]\!]$ denotes the jump across the interface $\partial\Omega$.

Summing up, the shape derivative of the objective functional J_α (1.2) is given as

$$J'_\alpha(\Omega)[V(0)] = \int_{\Gamma_M} u'[V(0)](u-\hat{u})\,ds + \alpha \int_{\partial\Omega} \kappa V(0).n\,ds.$$

We can simplify this shape derivative when we introduce the adjoint state w

$$\begin{aligned} &-\Delta w + c_\Omega\, w = -\chi_{\Gamma_M}(u-\hat{u}), && \text{respectively,} && 0, && \text{in } \mathcal{D},\\ &\frac{\partial w}{\partial n} = 0, && \text{respectively,} && -\chi_{\Gamma_M}(u-\hat{u}), && \text{on } \Gamma_N,\\ &w = 0 && && && \text{on } \Gamma_D. \end{aligned} \tag{2.4}$$

The shape derivative of the objective functional $J_\alpha(\Omega)$ finally gets

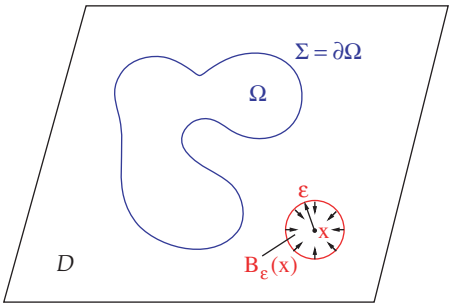
$$J'_\alpha(\Omega)[V(0)] = \int_{\partial\Omega} ([\![c_\Omega]\!]uw + \alpha\kappa)V(0).n\,ds. \tag{2.5}$$

First note that we just solve two partial differential equations, namely, (1.1) and (2.4), to calculate the shape derivative. Second, the first order necessary condition for optimal shapes Ω requires for all V : $J'_\alpha(\Omega)[V(0)] = 0$. Hence $(uw + \alpha\kappa) = 0$ on $\partial\Omega$. If this is not the case, we can construct a velocity such that $J'_\alpha(\Omega)[V(0)] < 0$. For example, take $V = -(uw + \alpha\kappa)n$, but other choices are possible (cf. Burger [14]).

2.2. Topological derivatives. In contrast to *shape derivatives* where one considers variations of a shape, *topological derivatives* aim for variations of the topology. The basic idea of the topological derivative is to add a small sphere with center x and radius ϵ to the domain Ω and consider the variation of the objective functional $J(\Omega \cup B_\epsilon(x))$ with respect to the radius of this sphere. Note that different shapes than spheres are possible and might result in different values of the derivatives.

DEFINITION 2.3 (topological derivative). *Let $J : \Omega \subset \mathbb{R}^d \to \mathbb{R}$ be an objective functional. Then the topological derivative is defined as the limit (if it exists)*

$$d_\tau J(x) := \lim_{\epsilon \to 0^+} \frac{J(\Omega \cup B_\epsilon(x)) - J(\Omega)}{|B_\epsilon(x)|}. \tag{2.6}$$



A negative topological derivative $d_\tau J(x) < 0$ indicates that it might be reasonable to add a small sphere at point x to reduce the objective functional. It is also possible to subtract material, i.e., take the set-minus instead of adding material, i.e., the union in (2.6).

In practice, topology changes forced by the topological derivative are neither spherical nor infinitesimally small. Most methods are based on a threshold criteria such as (4.3). This is opposed to functional analytic methods, where the finite step sizes are usually based on higher order estimates (remainder estimates) or line search methods, still guaranteeing descent in the objective functional. For a line searchlike idea in combination with topological derivatives, see Hintermüller [32]. In the framework of topological, asymptotic higher order estimates are used by Cedio-Fengya, Moskow, and Vogelius [20] and Habib and Hyeonbae [29]. To the best knowledge of the author, higher order estimates, although available, are not used for algorithms based on topological derivatives. One reason might be that higher order estimates for topological derivatives would just allow for a finite number of disjoint small holes (cf. Cedio-Fengya, Moskow, and Vogelius [20]).

Nonetheless the practical experience by most authors, using the topological derivative as an indicator, is rather positive. Therefore let us calculate the topological derivative for our objective functional (1.2). First of all consider the topological derivative of the perimeter $\mathrm{Per}\big(\Omega\big)$:

$$d_\tau \mathrm{Per}\big(\Omega\big) = \lim_{\epsilon\to 0} \frac{\mathrm{Per}\big(B_\epsilon(x)\big)}{|B_\epsilon(x)|} \simeq \lim_{\epsilon\to 0} \frac{\epsilon^{d-1}}{\epsilon^d} = \infty.$$

Hence the perimeter is not topologically differentiable. In practical applications one usually neglects this fact and calculates the topological derivative of the objective functional without the perimeter, i.e., for $J_0(\Omega)$. This derivative is already well known (see Amstutz [6]). Nonetheless we provide a detailed proof that is based on Hölder estimates, Sobolev embedding, and regularity results for elliptic partial differential equations. Parts of the proof will play a crucial role in the rest of the paper. Let us first state Sobolev's embedding theorem (cf. Adams [1])

THEOREM 2.4 (Sobolev embedding). *Let ω be a Lipschitz domain in $\mathbb{R}^d$, $m \in \mathbb{N}_0$.*
$d > 2m$: $H^m(\omega) \hookrightarrow L^q(\omega)$ for $2 \le q \le \frac{2d}{d-2m}$,
$d = 2m$: $H^m(\omega) \hookrightarrow L^q(\omega)$ for $2 \le q < \infty$,
$d < 2m$: $H^m(\omega) \hookrightarrow C\big(\mathrm{cl}(\omega)\big)$.

Based on the Caccioppoli inequality we cite an interior regularity result for elliptic partial differential equations (cf. Giaquinta [26]) that applies to (1.1) and (2.4).

THEOREM 2.5 (interior regularity). *Let $A_{ij} \in C^{0,1}(\mathcal{D})$ be strictly elliptic and $f \in L^2(\mathcal{D})$ and u be a weak solution to the linear elliptic equation*

$$-\partial_i\big(A_{ij}\partial_j u\big) = f;$$

then $u \in H^2_{\mathrm{loc}}(\mathcal{D})$ and even more

$$\|u\|^2_{H^2([\mathcal{D}]^\eta)} \le C(d)\left(\|f\|^2_{L^2(\mathcal{D})} + \frac{1}{\eta^2}\|\nabla u\|^2_{L^2(\mathcal{D})}\right), \tag{2.7}$$

where $\eta > 0$ and $[\mathcal{D}]^\eta := \mathrm{cl}\big(\mathrm{int}\{x \in \mathcal{D} | \mathrm{dist}(x, \partial\mathcal{D}) \ge \eta\}\big)$.

The strength of the above interior regularity result is that it is independent of the boundary conditions and the regularity of the domain $\mathcal{D}$ at the price that it is valid only in the interior of the domain $\mathcal{D}$. This allows us to deal with Lipschitz domains $\mathcal{D}$ and arbitrary mixed boundaries; i.e., $g \in H^{\frac{1}{2}}(\Gamma_D), h \in H^{-\frac{1}{2}}(\Gamma_N)$ are arbitrary in (1.1). Therefore we need not bother about corner singularities. The drawback for this convenience is that we need to restrict Ω to be a subset of $[\mathcal{D}]^\eta$.

In our application $u, w \in H^1(\mathcal{D})$ are solutions to the elliptic partial differential equations (1.1), (2.4). Therefore $\|u\|_{H^1(\mathcal{D})}$ and $\|w\|_{H^1(\mathcal{D})}$ can be estimated by

$$\begin{aligned}\|u\|^2_{H^1(\mathcal{D})} &\preceq \|f\|^2_{H^{-1}(\mathcal{D})} + \|g\|^2_{H^{\frac{1}{2}}(\Gamma_D)} + \|h\|^2_{H^{-\frac{1}{2}}(\Gamma_N)},\\ \|w\|_{H^1(\mathcal{D})} &\preceq \|u-\hat{u}\|_{L^2(\Gamma_M)}.\end{aligned}$$

When we plug this into the interior regularity result (2.7) we achieve

$$\begin{aligned}\|u\|^2_{H^2([\mathcal{D}]^\eta)} &\leq C(\eta,d,\mathcal{D})\big(\|f\|^2_{L^2(\mathcal{D})} + \|g\|^2_{H^{\frac{1}{2}}(\Gamma_D)} + \|h\|^2_{H^{-\frac{1}{2}}(\Gamma_N)}\big),\\ \|w\|_{H^2([\mathcal{D}]^\eta)} &\leq C(\eta,d,\mathcal{D})\|u-\hat{u}\|_{L^2(\Gamma_M)}.\end{aligned}$$

Note that, if it is not an option to restrict $\Omega \subset [\mathcal{D}]^\eta$, one needs to restrict instead $\mathcal{D} \in C^1$, $g \in H^{\frac{3}{2}}(\Gamma_D)$, $h \in H^{\frac{1}{2}}(\Gamma_N)$, and g, h compatible such that the outer regularity result applies (cf. Giaquinta [26]):

$$\begin{aligned}&\|u\|_{H^2(\mathcal{D})} \leq C(d,\mathcal{D})\big(\|f\|_{L^2(\mathcal{D})} + \|g\|_{H^{\frac{3}{2}}(\Gamma_D)} + \|h\|_{H^{\frac{1}{2}}(\Gamma_N)}\big),\\ &\|w\|_{H^2(\mathcal{D})} \leq C(d,\mathcal{D})\|u-\hat{u}\|_{L^2(\Gamma_M)}, \quad \text{respectively,} \quad \|w\|_{H^{\frac{3}{2}}(\mathcal{D})} \leq C(d,\mathcal{D})\|u-\hat{u}\|_{L^2(\Gamma_M)}.\end{aligned}$$

Finally, before we calculate the topological derivative, let us recall the direct (1.1) and the adjoint (2.4) partial differential equations but in their weak form.

Direct problem.

$$\langle \nabla u, \nabla v\rangle + \langle c_\Omega u, v\rangle = \langle f, v\rangle + \langle h, v\rangle_{H^{-\frac{1}{2}}(\Gamma_N)\times H^{\frac{1}{2}}(\Gamma_N)} \qquad \forall\, v \in H^1_{0,D}(\mathcal{D}). \tag{2.8}$$

Adjoint problem.

$$\langle \nabla v, \nabla w\rangle + \langle c_\Omega v, w\rangle = -\partial_u J_0(\Omega)[v] = -\langle u-\hat{u}, v\rangle_{L^2(\Gamma_M)} \qquad \forall\, v \in H^1_{0,D}(\mathcal{D}). \tag{2.9}$$

In the following we will often use the topologically perturbed domain $\tilde{\Omega}$ and the corresponding solution $\tilde{u}$ of (1.1).

PROPOSITION 2.6. *For $\eta > 0$ and every point $x \in [\mathcal{D}]^\eta \setminus \partial\Omega$ the topological derivative of the shape functional* (1.2) *with $\alpha = 0$ is given by*

$$d_\tau J_0(\Omega)(x) = -2(\chi_\Omega - \tfrac{1}{2})(\overline{c} - \underline{c})u(x)w(x).$$

Proof. Let $\tilde{\Omega}$, $\Omega \subset [\mathcal{D}]^\eta$ be arbitrary domains with positive Lebesgue measure, and consider the first order Taylor expansion of the objective functional J_0 with respect to the state u:

$$J_0(\tilde{\Omega}) - J_0(\Omega) = \partial_u J_0(\Omega)[\tilde{u}-u] + \mathcal{O}\big(\|\tilde{u}-u\|^2_{H^1(\mathcal{D})}\big).$$

$\|\tilde{u}-u\|_{H^1(\mathcal{D})} \preceq |\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{2d}}$: We subtract the two determining partial differential equations for u (respectively, $\tilde{u}$) and rearrange the terms to get

$$\begin{aligned}\langle \nabla(\tilde{u}-u), \nabla v\rangle + \langle c_{\tilde{\Omega}}(\tilde{u}-u), v\rangle &\overset{(1.1)}{=} -\langle (c_{\tilde{\Omega}} - c_\Omega)u, v\rangle\\ &\overset{\text{Hölder } \frac{1}{r}+\frac{1}{p}+\frac{1}{q}=1}{\leq} |\overline{c}-\underline{c}||\tilde{\Omega}\Delta\Omega|^{\frac{1}{r}}\|u\|_{L^p(\tilde{\Omega}\Delta\Omega)}\|v\|_{L^q(\tilde{\Omega}\Delta\Omega)}.\end{aligned}$$

To minimize r we use the interior regularity result $u \in H^2([\mathcal{D}]^\eta)$ and $v \in H^1(\mathcal{D})$ and apply Sobolev's embedding theorem to conclude $p \leq \infty$:

$$q \leq \begin{cases} \frac{2d}{d-2} & d=3\\ <\infty & d=2\end{cases} \quad\Rightarrow\quad r \geq \begin{cases} \frac{2d}{d+2} & d=3\\ >1 & d=2.\end{cases}$$

Finally the Lax Milgram lemma provides us with the desired result

$$\|\tilde{u} - u\|_{H^1(\mathcal{D})} \preceq |\overline{c} - \underline{c}||\tilde{\Omega}\Delta\Omega|^{\frac{1}{r}} \|u\|_{H^2([\mathcal{D}]^\eta)}.$$

$\partial_u J_0(\Omega)[\tilde{u} - u] = \langle (c_{\tilde{\Omega}} - c_\Omega)u, w\rangle + \mathcal{O}\big(|\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{d}}\big)$: We just use the definition of the adjoint problem (2.4) and afterwards (1.1) for $\tilde{u}$ and u to get

$$\begin{aligned}\partial_u J_0(\Omega)[\tilde{u} - u] &\overset{(2.4)}{=} -\langle \nabla(\tilde{u} - u), \nabla w\rangle - \langle c_\Omega(\tilde{u} - u), w\rangle \overset{(1.1)}{=} \langle (c_{\tilde{\Omega}} - c_\Omega)\tilde{u}, w\rangle \\ &= \langle (c_{\tilde{\Omega}} - c_\Omega)u, w\rangle + \langle (c_{\tilde{\Omega}} - c_\Omega)(\tilde{u} - u), w\rangle.\end{aligned}$$

For the term $\langle (c_{\tilde{\Omega}} - c_\Omega)(\tilde{u} - u), w\rangle$ we proceed as above, but this time we consider $(\tilde{u} - u) \in H^1_{0,D}(\mathcal{D})$ and $w \in H^2([\mathcal{D}]^\eta)$. Therefore we get, with r as above, the estimate

$$\partial_u J_0(\Omega)[\tilde{u} - u] = \langle (c_{\tilde{\Omega}} - c_\Omega)u, w\rangle + \mathcal{O}\big(|\tilde{\Omega}\Delta\Omega|^{\frac{1}{r}} |\overline{c} - \underline{c}| \|\tilde{u} - u\|_{H^1(\mathcal{D})} \|w\|_{H^2([\mathcal{D}]^\eta)}\big).$$

Summing up all of the estimates we get

$$J_0(\tilde{\Omega}) - J_0(\Omega) \leq \langle (c_{\tilde{\Omega}} - c_\Omega)u, w\rangle + \mathcal{O}\big(|\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{d}}\big). \tag{2.10}$$

Now set $\tilde{\Omega} = B_\epsilon(x) \cup \Omega$ and perform the limit according to the definition of the topological derivative. In general the limit can be deduced from the Lebesgue differentiation theorem (cf. Giaquinta [26]) almost everywhere, but due to the fact that u, $w \in H^2([\mathcal{D}^\eta]) \hookrightarrow C([\mathcal{D}]^\eta)$ it is even more obvious. ∎

Like the shape derivative (2.5), the topological derivative depends on the solution u of (1.1) and the adjoint w (2.4) only, which is standard for adjoint methods. Moreover, both derivatives are the same, which is not true in general but holds, up to a constant, for surprisingly many cases.

3. Topological expansions up to the first and second orders. In the previous section we introduced two concepts of geometry derivatives. *Shape derivatives* take the role of the Gateux and Fréchet derivatives in a functional analytic framework and allow Taylor expansions with remainder estimates in proper shape metrics such as the Courant metric. Shape derivatives are even suitable for perimeter-regularized objective functionals. This is different for *topological derivatives*. They provide Taylor expansion with respect to the size parameter ϵ (see Definition 2.3) but not with respect to shape metrics such as the L^1-metric (Lebesgue measure), and they are not suitable for perimeter-regularized problems. Therefore the goal of this section is to overcome these, remedy, and provide Taylor expansions of first and second order with respect to the L^1-metric. To allow also for perimeter-regularized problems we need to add to the Taylor expansion an additional dominating first order term depending on the perimeter of the symmetric difference of the two objects. The rational of this extra term is the general inequality

$$\big|\mathrm{Per}\big(\tilde{\Omega}\big) - \mathrm{Per}\big(\Omega\big)\big| \leq \mathrm{Per}\big(\tilde{\Omega}\Delta\Omega\big). \tag{3.1}$$

With a Taylor expansion at hand we are able to construct, like in the functional analytic framework, steepest descent- and Newton-type algorithms to reduce the shape objective functional $J_\alpha(\Omega)$.

3.1. First order topological expansion. A closer look at the proof of the topological derivative (Proposition 2.6), more precisely, (2.10) together with (3.1), shows that we already have the desired topological estimate.

PROPOSITION 3.1. *Let $\tilde{\Omega}, \Omega \in \mathcal{F}_L([\mathcal{D}]^\eta)$; then the objective functional $J_\alpha(\Omega)$ (1.2) with state u given by (1.1) has the first order topological expansion*

$$J_\alpha(\tilde{\Omega}) - J_\alpha(\Omega) \leq \langle (c_{\tilde{\Omega}} - c_\Omega)u, w\rangle + \alpha \mathrm{Per}\big(\tilde{\Omega}\Delta\Omega\big) + \mathcal{O}\big(|\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{2}}\big)\,. \tag{3.2}$$

Our aim is to minimize the objective functional $J_\alpha(\Omega)$ (1.2) with respect to the geometry Ω. Hence, when we already have an initial guess Ω_k we can improve it and calculate a new geometry Ω_{k+1} such that we reduce the objective functional $J_\alpha(\Omega_k)$, when we solve the auxiliary minimization problem

$$\Omega_{k+1} = \operatorname*{argmin}_{\Omega \in \mathcal{F}_L([\mathcal{D}]^\eta)} \langle (c_\Omega - c_{\Omega_k})u, w\rangle + \alpha \mathrm{Per}\big(\Omega\Delta\Omega_k\big) + c\big(|\Omega\Delta\Omega_k|^{<\frac{d+2}{2}}\big)\,. \tag{3.3}$$

The constant c is a consequence of the embedding and regularity results and can in principle be estimated. Algorithmically it seems more favorable to perform a trust region approach and vary c until the predicted decrease of the objective functional $J_\alpha(\Omega)$ is close to the actual decrease.

The minimization problem (3.3) already suggests a steepest descent-type algorithm to solve the original minimization problem of $J_\alpha(\Omega)$. We are more interested to solve (3.3) just once for every restart in the level set methods to force systematically reliable topology changes that decrease the objective functional $J_\alpha(\Omega)$. Therefore we prove that the minimization problem (3.3) has a solution which is not necessarily unique.

PROPOSITION 3.2. *Let $\alpha > 0$ and $\mathrm{Per}\big(\Omega_k\big) < \infty$; then the minimization problem (3.3) has a solution in the finite perimeter (Caccioppoli) sets.*

Proof. For every $\Omega \in \mathcal{F}_L([\mathcal{D}]^\eta)$ with a finite objective functional in the minimization problem (3.3) we conclude from $\alpha > 0$ that $\mathrm{Per}\big(\Omega\Delta\Omega_k\big) < \infty$. Therefore we can restrict $\Omega \in \mathcal{F}_L([\mathcal{D}]^\eta)$ to finite perimeter sets. Next we observe that $c_\Omega - c_{\Omega_k} = -2(\overline{c} - \underline{c})\big(\chi_{\Omega_k} - \frac{1}{2}\big)\chi_{\Omega\Delta\Omega_k}$. This allows one to replace the minimization problem (3.3) over Ω by a minimization problem over $\Omega\Delta\Omega_k$ and therefore an equivalent reformulation in $\mathrm{BV}(\mathcal{D}, \{0,1\})$:

$$\min_{p \in \mathrm{BV}(\mathcal{D},\{0,1\})} \Big\langle -2(\overline{c} - \underline{c})(\chi_{\Omega_k} - \tfrac{1}{2})uw, p\Big\rangle + \alpha |p|_{\mathrm{BV}(\mathcal{D})} + c\big(|p|^{<\frac{d+2}{2}}_{L^1(\mathcal{D})}\big)\,.$$

Now take a minimizing sequence $p_i \in \mathrm{BV}(\mathcal{D}, \{0,1\})$ which is due to $\alpha > 0$ uniformly bounded in $\|\,.\,\|_{\mathrm{BV}(\mathcal{D})}$. Therefore there exists a BV weak-* limit $p \in \mathrm{BV}(\mathcal{D})$. Due to the compact embedding $\mathrm{BV}(\mathcal{D}) \hookrightarrow^c L^1(\mathcal{D})$ (cf. Ambrosio, Fusco, and Pallara [5, Corollary 3.49]), the weak-* limit is also a strong limit in $L^1(\mathcal{D})$. Consequently we conclude $p \in \mathrm{BV}(\mathcal{D}, \{0,1\})$. Finally, the lower semicontinuity of $|\,.\,|_{\mathrm{BV}(\mathcal{D})}$ in the functions of bounded variation $\mathrm{BV}(\mathcal{D})$ guarantees that p is a solution to the above minimization problem formulated in $\mathrm{BV}(\mathcal{D}, \{0,1\})$, and therefore $\Omega_{k+1} = \big(\Omega_k \setminus \{p = 1\}\big) \cup \{(1 - \chi_{\Omega_k})p = 1\}$ is a minimizer for (3.3). ∎

Note that the minimizer to (3.3) might be Ω_k itself; i.e., no topology change is favorable to generate a guaranteed descent in the objective functional $J_\alpha(\Omega)$. This happens when the perimeter term dominates the first order term, i.e., when the topology changes get too small or when the topology is already the optimum of $J_\alpha(\Omega)$.

3.2. Second order topological expansion. The first order topological expansion in the L^1-metric was based mainly on a first order Taylor expansion of the objective functional $J_0(\Omega)$ with respect to the state u. We follow this strategy for the second order topological expansion. To allow a proper estimation of the second order terms we need to introduce an auxiliary partial differential equation.

Linearized problem.

$$\begin{aligned} -\Delta u^{\mathrm{lin}}(\tilde{\Omega}) + c_\Omega u^{\mathrm{lin}}(\tilde{\Omega}) &= -(c_{\tilde{\Omega}} - c_\Omega)u && \text{in } \mathcal{D}, \\ \frac{\partial u^{\mathrm{lin}}}{\partial n}(\tilde{\Omega}) &= 0 && \text{on } \Gamma_N, \\ u^{\mathrm{lin}}(\tilde{\Omega}) &= 0 && \text{on } \Gamma_D. \end{aligned}$$

We will indeed need the weak form of the linearized problem, which is given by

$$\langle \nabla u^{\mathrm{lin}}(\tilde{\Omega}), \nabla v\rangle + \langle c_\Omega u^{\mathrm{lin}}(\tilde{\Omega}), v\rangle = -\langle (c_{\tilde{\Omega}} - c_\Omega)u, v\rangle \qquad \forall\, v \in H^1_{0,D}(\mathcal{D}). \tag{3.4}$$

The reason why we denote the above partial differential equation linearized problem of (1.1) is that, when we consider (1.1) in a functional analytic setting with $c_\Omega \in L^2(\mathcal{D})$, the above formula would be its linearization.

PROPOSITION 3.3. *Let $\alpha > 0$ and $\mathrm{Per}(\Omega_k) < \infty$; then the objective functional $J_\alpha(\Omega)$ (1.2) with state u given by (1.1) has the second order topological expansion*

$$\begin{aligned} J_\alpha(\tilde{\Omega}) - J_\alpha(\Omega) \quad \le \quad & \big\langle (c_{\tilde{\Omega}} - c_\Omega)\big(u + u^{\mathrm{lin}}(\tilde{\Omega})\big), w\big\rangle + \tfrac{1}{2}\partial_u^2 J_0(\Omega)[u^{\mathrm{lin}}(\tilde{\Omega})]^2 \\ & + \alpha \mathrm{Per}\big(\tilde{\Omega}\Delta\Omega\big) + \mathcal{O}\big(|\tilde{\Omega}\Delta\Omega|^{<\frac{d+4}{d}}\big). \end{aligned} \tag{3.5}$$

Proof. First we start to do a Taylor expansion up to the second order for the unregularized objective functional $J_0(\Omega)$ with respect to the state u:

$$\begin{aligned} J_0(\tilde{\Omega}) - J_0(\Omega) \quad = \quad & \partial_u J_0(\Omega)[\tilde{u} - u] + \tfrac{1}{2}\partial_u^2 J_0(\Omega)[u^{\mathrm{lin}}(\tilde{\Omega})]^2 \\ & + \partial_u^2 J_0(\Omega)[\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})][u^{\mathrm{lin}}(\tilde{\Omega})] \\ & + \tfrac{1}{2}\partial_u^2 J_0(\Omega)[\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})]^2 + \mathcal{O}\big(\|\tilde{u} - u\|^3_{H^1(\mathcal{D})}\big). \end{aligned}$$

Again we estimate the higher order terms in the second and third rows by the L^1-metric $|\tilde{\Omega}\Delta\Omega|$, where we use $\|\partial_u^2 J_0(\Omega)\| \preceq 1$. Then we reformulate the terms in the first row. Several times we will use that $u, w \in H^2([\mathcal{D}]^\eta)$ (Theorem 2.5), Sobolev's embedding theorem, Hölder's inequality, and the Lax Milgram lemma.

$\|u^{\mathrm{lin}}(\tilde{\Omega})\|_{H^1(\mathcal{D})} \preceq |\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{2d}}$: With r, p, q as in the proof of Proposition 2.6 we have

$$\begin{aligned} \langle \nabla u^{\mathrm{lin}}(\tilde{\Omega}), \nabla v\rangle + \langle c_\Omega u^{\mathrm{lin}}(\tilde{\Omega}), v\rangle &= -\langle (c_{\tilde{\Omega}} - c_\Omega)u, v\rangle \\ &\overset{\text{Hölder } \frac{1}{r}+\frac{1}{p}+\frac{1}{q}=1}{\preceq} (\overline{c} - \underline{c})|\tilde{\Omega}\Delta\Omega|^{\frac{1}{r}}\|u\|_{L^p(\tilde{\Omega}\Delta\Omega)}\|v\|_{L^q(\tilde{\Omega}\Delta\Omega)} \\ &\overset{\text{Theorem 2.4}}{\preceq} (\overline{c} - \underline{c})|\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{2d}}\|u\|_{H^2([\mathcal{D}]^\eta)}\|v\|_{H^1(\mathcal{D})}. \end{aligned}$$

$\|\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})\|_{H^1(\mathcal{D})} \preceq |\tilde{\Omega}\Delta\Omega|^{<\frac{d+6}{2d}}$: First note that due to Sobolev's embedding theorem $H^1(\mathcal{D}) \hookrightarrow L^q(\mathcal{D})$, with

$$q \le \begin{cases} \frac{2d}{d-2} & d = 3 \\ < \infty & d = 2. \end{cases}$$

Then we apply Lax Milgram's lemma to the following term:

$$\begin{aligned} \langle \nabla(\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})), \nabla v\rangle + \langle c_\Omega(\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})), v\rangle &\overset{(1.1)\ \&\ (3.4)}{=} -\langle (c_{\tilde{\Omega}} - c_\Omega)(\tilde{u} - u), v\rangle \\ &\overset{\text{Hölder } \frac{1}{p}+\frac{2}{q}=1}{\preceq} |\tilde{\Omega}\Delta\Omega|^{\frac{1}{p}}\|\tilde{u} - u\|_{L^q(\tilde{\Omega}\Delta\Omega)}\|v\|_{L^q(\tilde{\Omega}\Delta\Omega)} \\ &\overset{\text{Theorem 2.4}}{\preceq} |\tilde{\Omega}\Delta\Omega|^{<\frac{2}{d}}\|\tilde{u} - u\|_{H^1(\mathcal{D})}\|v\|_{H^1(\mathcal{D})} \\ &\overset{\text{Proposition 2.6}}{\preceq} |\tilde{\Omega}\Delta\Omega|^{<\frac{d+6}{2d}}\|u\|_{H^2([\mathcal{D}]^\eta)}\|v\|_{H^1(\mathcal{D})}. \end{aligned}$$

$$\partial_u J_0(\Omega)[\tilde{u}-u] = \langle (c_{\tilde{\Omega}} - c_\Omega)(u + u^{\mathrm{lin}}(\tilde{\Omega})), w\rangle + \mathcal{O}\big(|\tilde{\Omega}\Delta\Omega|^{<\frac{d+4}{d}}\big):$$

$$\begin{aligned}\partial_u J_0(\Omega)[\tilde{u}-u] &\overset{(2.4)}{=} -\langle \nabla(\tilde{u}-u), \nabla w\rangle - \langle c_\Omega(\tilde{u}-u), w\rangle \overset{(1.1)}{=} \langle (c_{\tilde{\Omega}} - c_\Omega)\tilde{u}, w\rangle \\ &= \langle (c_{\tilde{\Omega}} - c_\Omega)(u + u^{\mathrm{lin}}(\tilde{\Omega})), w\rangle + \langle (c_{\tilde{\Omega}} - c_\Omega)(\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})), w\rangle.\end{aligned}$$

Finally we need to to estimate the term $\langle (c_{\tilde{\Omega}} - c_\Omega)(\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})), w\rangle$ following the same arguments and with the same choice for r, p, q as in Proposition 2.6:

$$\begin{aligned}\dots &\overset{\text{Hölder } \frac{1}{r}+\frac{1}{p}+\frac{1}{q}=1}{\leq} && |\tilde{\Omega}\Delta\Omega|^{\frac{1}{r}} |\overline{c} - \underline{c}| \|\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})\|_{L^q(\mathcal{D})} \|w\|_{L^p(\mathcal{D})} \\ &\overset{\text{Theorem 2.4}}{\preceq} && |\tilde{\Omega}\Delta\Omega|^{<\frac{d+2}{2d}} |\overline{c} - \underline{c}| \|\tilde{u} - u - u^{\mathrm{lin}}(\tilde{\Omega})\|_{H^1(\mathcal{D})} \|w\|_{H^2([\mathcal{D}]^\eta)}.\end{aligned}$$

The perimeter term we estimate as in the first order topological expansion. ∎

With the second order expansion at hand we can improve an initial geometry Ω_k, such that the objective functional $J_\alpha(\Omega)$ decreases, when we solve the partial differential equation constraint minimization problem

(3.6)
$$\Omega_{k+1} = \operatorname*{argmin}_{\Omega\in\mathcal{F}_L([\mathcal{D}]^\eta)} \langle (c_\Omega - c_{\Omega_k})(u + u^{\mathrm{lin}}(\Omega)), w\rangle + \tfrac{1}{2}\partial_u^2 J_\alpha(\Omega_k)[u^{\mathrm{lin}}(\Omega)]^2 + \alpha \mathrm{Per}\big(\Omega\Delta\Omega_k\big).$$

The constraint minimization problem is similar to a Newton-type step. Later we will use the solution to the minimization problem to force a topology change in level set methods. Before that, we prove that the constraint minimization problem (3.6) has a solution which is not necessarily unique.

PROPOSITION 3.4. *Let $\alpha > 0$ and $\mathrm{Per}\big(\Omega_k\big) < \infty$; then the minimization problem* (3.6) *has a solution in the finite perimeter (Caccioppoli) sets.*

Proof. First we note that $u^{\mathrm{lin}}(\Omega)$ is uniformly bounded in $H^1(\mathcal{D})$. Therefore also the objective functional of the constraint minimization problem (3.6) is uniformly bounded from below. Hence every $\Omega \in \mathcal{F}_L([\mathcal{D}]^\eta)$ with a finite objective functional of the constraint minimization problem (3.6) has a finite perimeter, i.e., $\mathrm{Per}\big(\Omega\big) < \infty$. As in Proposition 3.2 we conclude from $c_\Omega - c_{\Omega_k} = -2(\chi_{\Omega_k} - \frac{1}{2})\chi_{\Omega\Delta\Omega_k}$ that also $u^{\mathrm{lin}}(\Omega)$ (3.4) just depends on $\Omega\Delta\Omega_k$. Therefore we can again reformulate the minimization problem (3.6) over Ω to a minimization over $\Omega\Delta\Omega_k$ and then switch to $\mathrm{BV}(\mathcal{D}, \{0,1\})$:

$$\min_{p\in\mathrm{BV}(\mathcal{D},\{0,1\})} \big\langle -2(\overline{c} - \underline{c})(\chi_{\Omega_k} - \tfrac{1}{2}) w(u + u^{\mathrm{lin}}(p)), p\big\rangle + \tfrac{1}{2}\partial_u^2 J_\alpha(\Omega_k)[u^{\mathrm{lin}}(p)]^2 + \alpha|p|_{\mathrm{BV}}.$$

Note that $u^{\mathrm{lin}}(p)$ solves for all $v \in H^1_{0,D}(\mathcal{D})$:

$$\langle \nabla u^{\mathrm{lin}}(p), \nabla v\rangle + \langle c_{\Omega_k} u^{\mathrm{lin}}(p), v\rangle = \big\langle -2(\overline{c} - \underline{c})(\chi_{\Omega_k} - \tfrac{1}{2}) p u_{\Omega_k}, v\big\rangle.$$

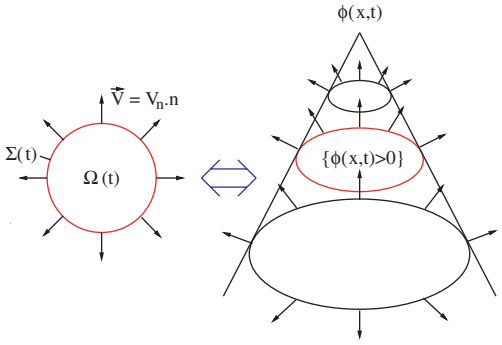
Following the arguments of Proposition 3.2 provides a minimizing sequence $p_i \in \mathrm{BV}(\mathcal{D}, \{0,1\})$ converging weak* to $p \in \mathrm{BV}(\mathcal{D})$ and strongly in $L^1(\mathcal{D})$. From the strong convergence of $p_i \to p$ in $L^1(\mathcal{D})$ and Lax Milgram's lemma, we conclude strong convergence of $u^{\mathrm{lin}}(p_i) \to u^{\mathrm{lin}}(p)$ in $H^1(\mathcal{D})$. Finally we deduce from the lower semicontinuity of $|\,.\,|_{\mathrm{BV}}$ that $p \in \mathrm{BV}(\mathcal{D}, \{0,1\})$ is a minimizer, and therefore $\Omega_{k+1} = \big(\Omega_k \setminus \{p = 1\}\big) \cup \{(1 - \chi_{\Omega_k})p = 1\}$ is a minimizer of (3.6). ∎

In principle, minimization problem (3.6) is as difficult as the original problem, but first we do not need to solve it too accurately; it is enough to correct the first order solution. Second, it might be much easier to construct efficient solvers for problems with the partial differential equation constraint that are linear in $\chi_{\Omega\Delta\Omega_k}$. For example, in imaging it is possible to reformulate problems in $\mathrm{BV}(\{0,1\})$ to problems in $\mathrm{BV}([0,1])$ (cf. Burger and Hintermüller [17]). Furthermore there is a well-developed theory for BV regularization for linear problems, linear in χ_Ω (cf. Osher et al. [39]).

4. Numerical solution. In this section we provide a brief introduction into level set methods (cf. Osher and Fedkiw [40]). Then we use the so-called phase I/II method (cf. Allaire et al. [2], Burger, Hackl, and Ring [16], Hintermüller [32]) to incorporate the steepest descent- (3.3) (respectively, the Newton-) type (3.6) step into level set methods. Furthermore we provide details about how we solved the minimization problems (3.3) and (3.6) numerically.

4.1. Level set methods. The main idea of level set methods is to represent an evolving front $\Sigma(t) = \partial\Omega(t)$ as the zero level set of a continuous function, i.e.,

$$\begin{aligned}\Omega(t) &= \big\{x \in \mathcal{D} \,\big|\, \phi(x,t) > 0\big\},\\ \Sigma(t) &= \big\{x \in \mathcal{D} \,\big|\, \phi(x,t) = 0\big\}.\end{aligned}$$



The geometric motion of the level set with normal velocity $\vec{V} = V_n.n$ can equivalently be described by the propagation of the level set function ϕ which solves the Hamilton–Jacobi equation

$$\frac{\partial\phi}{\partial t} - V_n|\nabla\phi| = 0 \qquad \text{in } \mathbb{R}^d \times \mathbb{R}^+. \tag{4.1}$$

The Hamilton–Jacobi equation (4.1) for ϕ is the analogon to the flow equation (2.2).

As already mentioned in section 2.1 the crucial point is an appropriate choice of the velocity, such that the objective functional $J_\alpha(\Omega)$ (1.2) decreases. This resembles the classical speed method in shape optimization (cf. Delfour and Zolésio [22]). The weak formulation via the level set methods allows for more general evolutions and for topological changes such as splitting, merging, and vanishing of domains.

From the shape derivative $J'_\alpha(\Omega)[V_n]$ of the objective functional (1.2)

$$J'_\alpha(\Omega)[V_n] = \int_{\partial\Omega} \big([\![c_\Omega]\!]uw + \alpha\kappa\big)V_n\,ds,$$

we can deduce normal velocities V_n such that the objective functional (1.2) decreases. The simplest choice would be $V_n = -\big([\![c_\Omega]\!]uw + \alpha\kappa\big)$. This choice results in a very regular velocity. According to Burger [14] a preconditioned velocity $V_n \in H^{-\frac{1}{2}}(\partial\Omega)$ is more appropriate and results into faster convergence. This $H^{-\frac{1}{2}}(\partial\Omega)$ velocity can be set to $V_n = [\![\frac{\partial\psi}{\partial n}]\!]$, where ψ solves the subproblem

$$\langle\psi, v\rangle = -\left\langle [\![c_\Omega]\!]uw + \alpha\kappa, \left[\!\!\left[\frac{\partial v}{\partial n}\right]\!\!\right]\right\rangle_{\partial\Omega} \qquad \forall\, v \in H^1_0(\mathcal{D}).$$

We solve the level set equation (4.1) with a standard fifth order weighted ENO scheme for the spatial and a third order explicit Runge–Kutta scheme for the time discretization (cf. Jiang and Peng [35]).

4.2. Phase I/II algorithm. In Burger, Hackl, and Ring [16] the topological derivatives were incorporated as an extra source term in the level set methods:

$$\frac{\partial \phi}{\partial t} - V_n|\nabla \phi| + \mathcal{S} = 0, \qquad V_n = V_n\big(J'_\alpha(\Omega)\big), \quad \mathcal{S} = \mathcal{S}\big(d_\tau J_0(\Omega)\big).$$

An inherent time step control in the level set methods guaranteed that the topological change was such that the objective function decreased. Another method suggested by Allaire et al. [2], Burger, Hackl, and Ring [16], and Hintermüller [32] is to restart the level set evolution after a fixed time (or due to clever criteria) with an initial level set function generated by the last time step plus the topological change due to the topological derivatives. This algorithm was phrased phase I/II algorithm by Hintermüller [32], where phase I corresponds to the algorithm for the topology change and phase II to the classical level set evolution. Let us put this into a more mathematical formulation: Let $(T_k)_{k\in\mathbb{N}_0}$ be a series of time steps, either fixed or generated due to a termination criterion in the level set evolution. Set $\phi_{-1}(T_{-1})$ to an initial guess (usually no material or material everywhere). Then the phase I/II algorithm is given by

$$\begin{aligned} \phi_k(t=0) \qquad\qquad &= \mathcal{S}(d_\tau J_0(\Omega_{k-1}), \phi_{k-1}(T_{k-1})), \\ \frac{\partial \phi_k}{\partial t} + V_n(J'_\alpha)|\nabla \phi_k| &= 0, \end{aligned}$$

where $\mathcal{S}(.,.)$ describes phase I; i.e., the algorithm that forces other topology changes than splitting, merging, and vanishing. Most implementations of phase I do not use higher order estimates for the topological derivative and force a topology change whenever the topological derivative is negative or smaller than a certain threshold criterion as (4.3). Therefore the objective functional $J_\alpha(\Omega)$ and even $J_0(\Omega)$ might increase. Exceptional to this practice is the line searchlike algorithm proposed by Hintermüller [32] that guarantees descent in $J_0(\Omega)$. An extension of the line search algorithm to problems with perimeter constraints $J_\alpha(\Omega)$ is possible with the methods developed in this paper.

The idea is to use the first (respectively, second) order topological expansion to construct phase I and therefore guarantee a decrease in the objective functional $J_\alpha(\Omega)$.

Phase I. $\mathcal{S}(J_\alpha, \phi_{k-1}(T_{k-1})) := b_{\Omega_k}$; Ω_k solution to (3.3) (respectively, (3.6)). With b_Ω we denoted the signed distance function defined by

$$b_\Omega(x) = \inf_{y\in\Omega} |x-y| - \inf_{y\in C(\Omega)} |x-y|.$$

In the following we describe briefly how we solve (3.3) (respectively, (3.6)) numerically.

Steepest descent-type topology changes. First we recall the auxiliary minimization problem (3.3) in its equivalent reformulation given, in Proposition 3.2:

(4.2)
$$\Omega_{k+1}\Delta\Omega_k = \operatorname*{argmin}_{\Omega\Delta\Omega_k\in\mathcal{F}_L([\mathcal{D}]^\eta)} \underbrace{\langle d_\tau J_0(\Omega_k), \chi_{\Omega\Delta\Omega_k}\rangle + \alpha \mathrm{Per}\big(\Omega\Delta\Omega_k\big) + c|\Omega\Delta\Omega_k|^{<\frac{d+2}{2}}}_{=:\mathcal{G}^1_{\Omega_k}(\Omega\Delta\Omega_k)},$$

where we used that $d_\tau J_0(\Omega_k) = -2(\overline{c}-\underline{c})(\chi_{\Omega_k}-\frac{1}{2})uw$. Ω_{k+1} is then given by $\Omega_{k+1} = \Omega_k \backslash (\Omega_{k+1}\Delta\Omega_k) \cup (C(\Omega_k)\cap(\Omega_{k+1}\Delta\Omega_k))$. As soon as the solutions to the direct problem u (1.1) and the adjoint problem w (2.4) are given, we can solve the minimization problem (4.2) without solving further partial differential equations. Furthermore note that the objective functional in the auxiliary minimization problem (3.3) over Ω was not continuous due to the term $\mathrm{Per}\big(\Omega\Delta\Omega_k\big)$. This is not the case any more for the objective functional $\mathcal{G}^1(\Omega\Delta\Omega_k)$, which is even shape differentiable:

$$\mathcal{G}^1_{\Omega_k}{}'(\Omega\Delta\Omega_k)[V] = \int_{\partial(\Omega\Delta\Omega_k)} \left(d_\tau J_0(\Omega_k) + c\left(< \frac{d+2}{2} \right) |\Omega\Delta\Omega_k|^{<\frac{d}{2}} + \alpha\kappa \right) V.n \, ds.$$

This allows us to use level set methods to solve the minimization problem (4.2). An obvious choice for the velocity is $V_n = -\big(d_\tau J_0(\Omega_k) + c\big(< \frac{d+2}{2}\big)|\tilde{\Omega}\Delta\Omega|^{<\frac{d}{2}} + \alpha\kappa\big)$. To provide an appropriate initial guess either we solve the minimization problem (3.3) for $\alpha = 0$ or we use the classical guess

$$\big\{\chi_{\Omega_k} d_\tau J_0(\Omega_k) < r \min(\underline{m}_{\chi_{\Omega_k}}, 0)\big\} \cup \big\{(1-\chi_{\Omega_k}) d_\tau J_0(\Omega_k) < r \min(\underline{m}_{1-\chi_{\Omega_k}}, 0)\big\}, \tag{4.3}$$

with $\underline{m}_\chi = -\| \min(\chi d_\tau J_0(\Omega_k), 0)\|_{L^\infty(\mathcal{D})}$ and $r \in [0,1]$.

Note that the solution to the minimization problem (4.2) with $\alpha = 0$ needs to fulfill

$$d_\tau J_0(\Omega_k)|_{\partial(\Omega\Delta\Omega_k)} + c\left(< \frac{d+2}{d} \right) |\Omega\Delta\Omega_k|^{<\frac{d}{2}} = 0.$$

Therefore we are searching for a level set of $-\big(> \frac{d}{c(d+2)}\big) d_\tau J_0(\Omega_k)$ whose enclosed volume to the power of $< \frac{d}{2}$ coincides with the value of the level set. The volume $|\Omega\Delta\Omega_k|$ enclosed by the level set $-\big(> \frac{d}{c(d+2)}\big) d_\tau J_0(\Omega_k)$ is monotonically (not necessarily continuously) decreasing to zero. Therefore a simple bisection algorithm can provide an outer (inner) approximation for the minimization problem (3.3) with $\alpha = 0$. This can be used as the initial guess for the above minimization problem with $\alpha \neq 0$.

Both the classical guess as well as the solution to the above minimization problem with $\alpha = 0$ are based on the level set of the topological derivative. In our numerical implementation we use the classical guess with $r = 0.7$. Both guesses might provide many topology changes at once and/or very rough topologies which might increase the original objective functional J_α. Therefore they are usually not very well suited as an initial guess for level set methods applied to the original problem (cf. Hackl [30]).

Newton-type topology changes. First we recall the minimization problem (3.6) but in its equivalent reformulation developed in Proposition 3.4:

$$\Omega_{k+1}\Delta\Omega_k = \operatorname*{argmin}_{\Omega\Delta\Omega_k \in \mathcal{F}_L([\mathcal{D}]^\eta)} \underbrace{\begin{array}{c} \langle -2(\overline{c}-\underline{c})(\chi_{\Omega_k}-\frac{1}{2})(u+u^{\mathrm{lin}}(\Omega\Delta\Omega_k))w, \chi_{\Omega\Delta\Omega_k}\rangle \\ +\frac{1}{2}\partial_u^2 J_\alpha(\Omega_k)[u^{\mathrm{lin}}(\Omega\Delta\Omega_k)]^2 + \alpha\mathrm{Per}\big(\Omega\Delta\Omega_k\big) \end{array}}_{=:\mathcal{G}^2_{\Omega_k}(\Omega)} . \tag{4.4}$$

Note that here $u^{\mathrm{lin}}(\Omega\Delta\Omega_k)$ is equivalently defined by for all $v \in H^1_{0,D}(\mathcal{D})$:

$$\langle \nabla u^{\mathrm{lin}}(\Omega\Delta\Omega_k), \nabla v\rangle + \langle c_\Omega u^{\mathrm{lin}}(\Omega\Delta\Omega_k), v\rangle = \langle 2(\overline{c}-\underline{c})(\chi_{\Omega_k}-\tfrac{1}{2})\chi_{\Omega\Delta\Omega_k} u_k, v\rangle.$$

Minimization problem (4.4) has a partial differential equation as a constraint. Hence it is as difficult to treat as the original minimization problem of $J_\alpha(\Omega)$ (1.2).

One might argue that there is no point in solving the Newton-type minimization problem but to just perform phase I with the steepest descent-type method and then proceed with the level set evolution of the original minimization problem with objective functional J_α. This argument is true because we do not have efficient solvers at hand that exploit the special structure of the constraint minimization problem where the partial differential equation constraint is $u^{\mathrm{lin}}(\Omega\Delta\Omega_k)$ is linear in $\chi_{\Omega\Delta\Omega_k}$. Algorithms that aim in this direction are known in imaging (cf. Burger and Hintermüller [17], Osher et al. [39]) but are not yet applicable to our problem. Furthermore note that the above argument would also apply to Newton's method in a functional analytic setting. There it is known that Newton-type iterations have quadratic convergence rates, whereas steepest descent-type iterations have linear rates. The better convergence comes at the price that one Newton iteration is usually computationally much more expansive than one steepest descent-type iteration. This effect is especially observed in minimization problems with partial differential equations as a constraint. Here the steepest descent-type methods often perform equally well, in terms of total computational time consumed, as Newton-type methods.

Our task is to solve the above shape optimization problem, and clearly we could use level set methods for that. To distinguish phase I with Newton-type iterations from phase II we decided to use a phase field approach to solve the above minimization problem. Again it might be much more efficient to apply this approach directly to the original shape optimization problem (see Hackl [30] for a comparison of these methods), but this is not the point of this paper. We focus on the expansion of first and second orders and show their applicability.

The phase field approach is based on the equivalent formulation of the above minimization problem in the space $\mathrm{BV}(\mathcal{D},\{0,1\})$ that we already met in the proof of Proposition 3.4. In the phase field approach the minimization problem, formulated in $\mathrm{BV}(\mathcal{D},\{0,1\})$, is relaxed, in the framework of Γ-convergence (cf. Bourdin and Chambolle [12]) to a Hilbert space problem, namely:

$$p_{k+1} = \operatorname*{argmin}_{p\in H_0^1(\mathcal{D})} \left\langle -2(\overline{c}-\underline{c})(\chi_{\Omega_k}-\tfrac{1}{2})w(u+u^{\mathrm{lin}}(p)),p\right\rangle + \tfrac{1}{2}\partial_u^2 J_0(\Omega_k)[u^{\mathrm{lin}}(p)]^2 + \left(\epsilon\|\nabla p\|^2_{L^2(\mathcal{D})} + \frac{\alpha^2}{\epsilon}\int_{\mathcal{D}} W_N(p)\,dx\right).$$

$W_N(.)$ is a normalized double well potential, with $W_N(0)=W_N(1)=0$, $W_N(s)>0, s\in\mathbb{R}\setminus\{0,1\}$, and $2\int_0^1\sqrt{W_N(s)}\,ds=1$.

The double well potential on one hand forces p to approach $\{0,1\}$ when $\epsilon\to 0$, whereas the H^1-seminorm of p requires smooth solutions. With $\epsilon\to 0$ the H^1-seminorm term forces the solution to switch smoothly from 0 to 1 in an ϵ-region. Altogether these two terms approximate with $\epsilon\to 0$ the perimeter term. The minimization problem is now posed in a Hilbert space setting, and one can use steepest descent- or Newton-type methods to solve this problem. For our numerical tests we chose

$$W_N(s) = \begin{cases} \left(\frac{4}{\pi}\right)^2 s(1-s) & s\in[0,1],\\ \infty & \text{otherwise,}\end{cases}$$

and perform a Gauss–Newton algorithm implemented as SQP. For details about the implementation of the phase field method see Hackl [30]. We will discuss only briefly the choice of ϵ.

From the theoretical point of view ϵ should be very small to approximate the original problem best, but numerically there needs to be a relation that connects $\frac{\epsilon}{\alpha}$ to the mesh size h. Practical experience led us to set this relation to

$$\frac{\epsilon}{\alpha} = \tau h, \qquad \tau \geq 2.$$

Furthermore when we start the optimization with a very small ϵ, then the double well potential provides a too-strict restriction, and the algorithm cannot perform topology changes other than merging, splitting, and vanishing. In this case the algorithm behaves more like classical level set methods, and therefore it is better to use a level set method, due to its clear, reliable, and simple implementation. Hence ϵ should be chosen large at the beginning of the iterations to allow for easy topology changes and get gradually closer to the smallest possible value for ϵ to get sharp interfaces.

Final remark. In our implementation for both the steepest descent- as well as the Newton-type phase I algorithms, we ensured that only the level set method (phase II) is responsible for evolving the boundary $\partial\Omega$. We did this by minimizing (4.2), (3.6) over $\Omega\Delta\Omega_k \in \mathcal{F}_L\big([\mathcal{D}]^\eta \cap [\partial\Omega_k]^\eta\big)$ instead of $\Omega\Delta\Omega_k \in \mathcal{F}_L([\mathcal{D}]^\eta)$. The main reason for that is that we believe that phase I should be responsible for topology changes only. The evolution of the boundary $\partial\Omega_k$ is usually much more efficient and reliable using level set methods (phase II). As a side effect of this restriction we have the equality $\mathrm{Per}\big(\Omega\big) - \mathrm{Per}\big(\Omega_k\big) = \mathrm{Per}\big(\Omega\Delta\Omega_k\big)$ instead of the inequality (3.1).

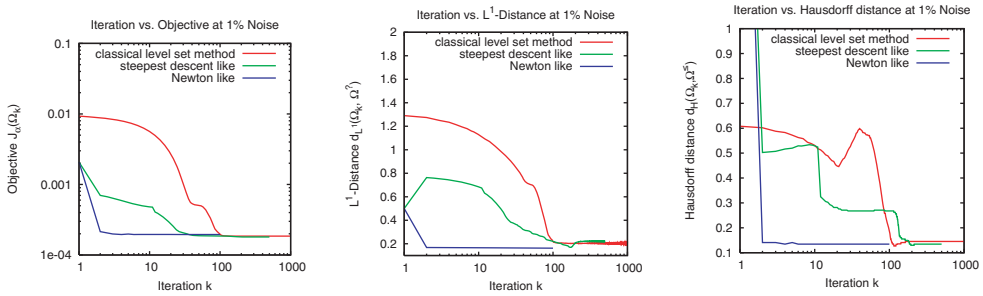
5. Numerical results. In this section we compare the classical level set method to the level set method proposed in this paper that incorporates steepest descent-type (4.2) and Newton-type (4.4) topology changes. For a comparison to other methods see Hackl [30]. We just restrict our attention to problems with more than one connected component, namely, to the identification of two ellipses and of an elliptic hole in another ellipse. The two ellipse case we consider for full measurements, i.e., $\Gamma_M = \mathcal{D}$, as well as for boundary measurements, i.e., $\Gamma_M = \Gamma_N$. Just the elliptic hole in an ellipse case we consider for full measurements only.

We perform all numerical tests on a fixed domain $\mathcal{D} = [-1, 1]^2$. To avoid inverse crime (cf. Colton and Kress [21, p. 133]), we generate the data on a different grid (finer mesh and higher order basis functions) and perturb it with 1% Gaussian noise, measured in the $\|\,.\,\|_{L^2(\Gamma_M)}$-norm. We use 1% noise because we expect the numerical error of our discretization to be of the same magnitude.

We provide graphs (Figures 5.1, 5.2, 5.3) that show the iteration number versus objective functional $J_\alpha(\Omega_k)$, L^1-distance $d_{L^1}(\Omega_k, \Omega^\dagger)$, and Hausdorff distance $d_H(\Omega_k, \Omega^\dagger)$:

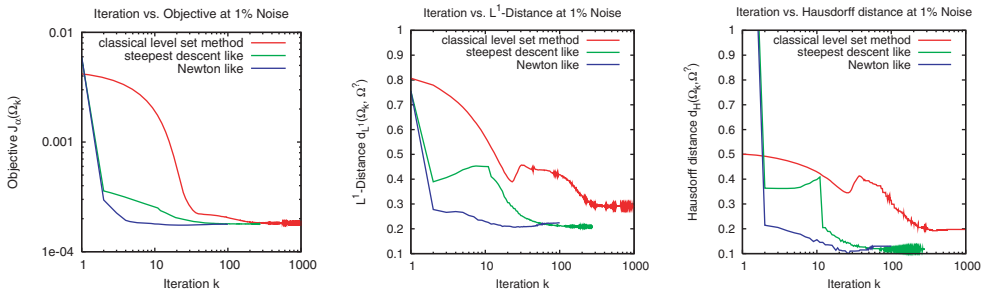
$$\begin{aligned} d_{L^1}(\Omega, \tilde{\Omega}) &:= |\Omega\Delta\tilde{\Omega}|, \\ d_H(\Omega, \tilde{\Omega}) &:= \max\big(\, \sup_{x\in\Omega} \inf_{y\in\tilde{\Omega}} |x-y|, \sup_{y\in\tilde{\Omega}} \inf_{x\in\Omega} |x-y| \,\big). \end{aligned}$$

To visualize the evolution of the geometry for each algorithm, we present a series of pictures (see Figures 5.4, 5.5, 5.6), starting with the first iteration up to the final solution (note the iteration numbers in each row do not coincide). The pictures are arranged such that each column represents the evolution for one algorithm: namely, the left column represents the classical level set method, the middle column represents the level set method with incorporated steepest descent-type topology change, and the right column represents the level set method with incorporated Newton type topology change.



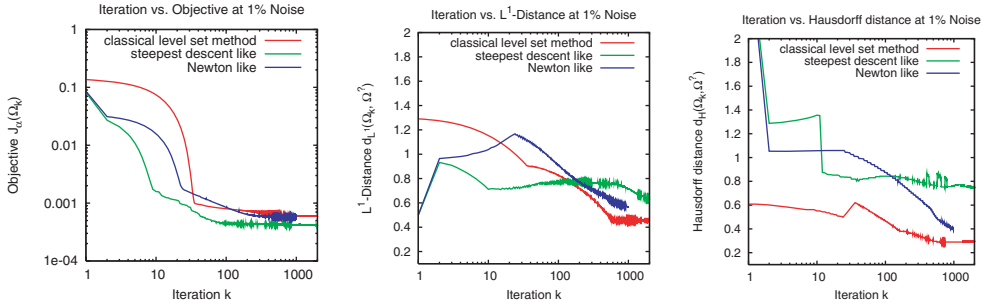
(a) Iteration vs. Objective (b) Iteration vs. L^1-Distance (c) Iteration vs. Hausdorff

FIG. 5.1. *Two ellipses,* $\Gamma_M = \mathcal{D}$*: Iteration vs.* $J_\alpha(\Omega_i)$, $d_{L^1}(\Omega_i, \Omega^\dagger)$, $d_H(\Omega_i, \Omega^\dagger)$.



(a) Iteration vs. Objective (b) Iteration vs. L^1-Distance (c) Iteration vs. Hausdorff

FIG. 5.2. *Ellipse with elliptic hole,* $\Gamma_M = \mathcal{D}$*: Iteration vs.* $J_\alpha(\Omega_i)$, $d_{L^1}(\Omega_i, \Omega^\dagger)$, $d_H(\Omega_i, \Omega^\dagger)$.

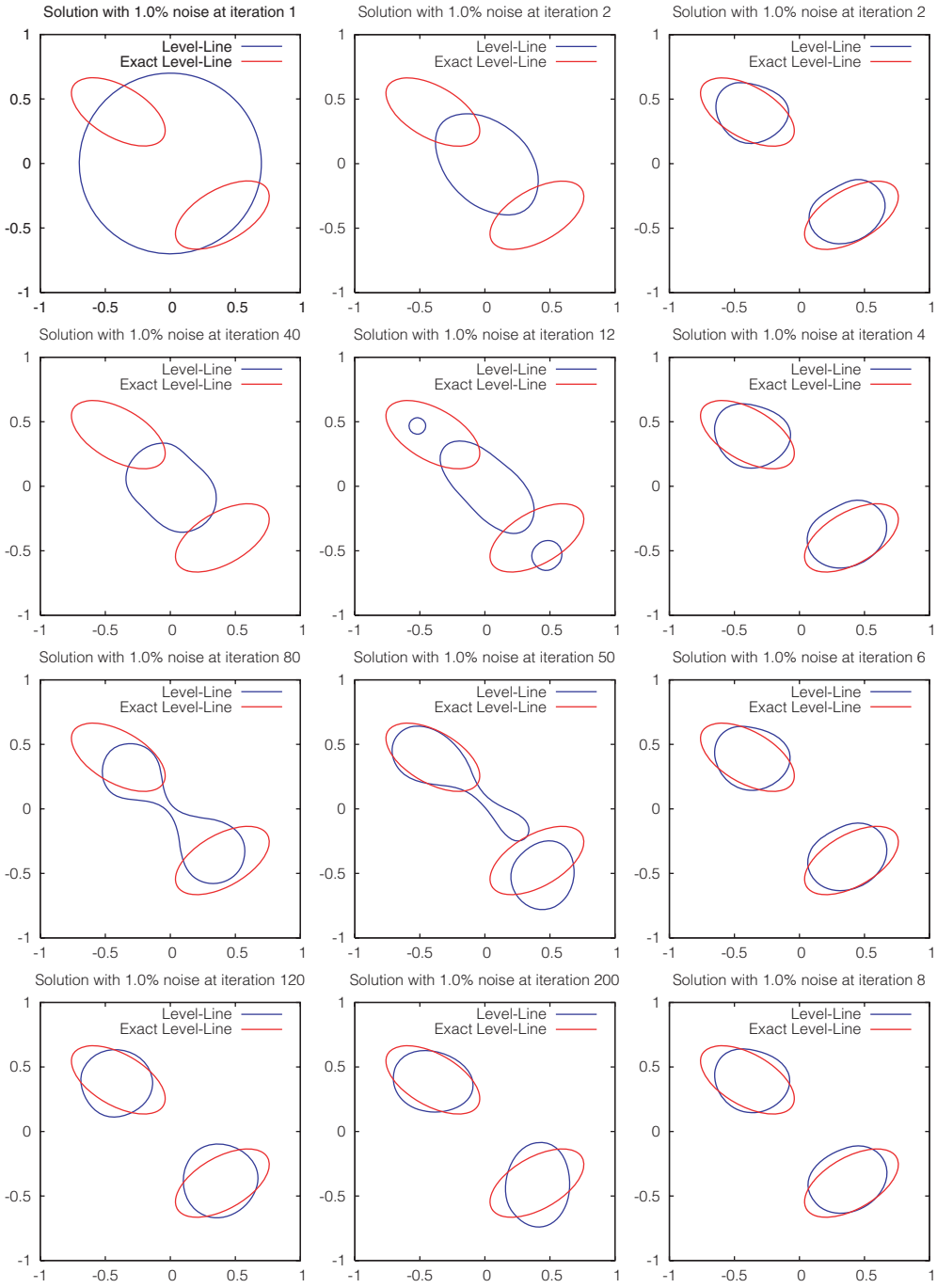


(a) Iteration vs. Objective (b) Iteration vs. L^1-Distance (c) Iteration vs. Hausdorff

FIG. 5.3. *Two ellipses,* $\Gamma_M = \Gamma_N$*: Iteration vs.* $J_\alpha(\Omega_i)$, $d_{L^1}(\Omega_i, \Omega^\dagger)$, $d_H(\Omega_i, \Omega^\dagger)$.

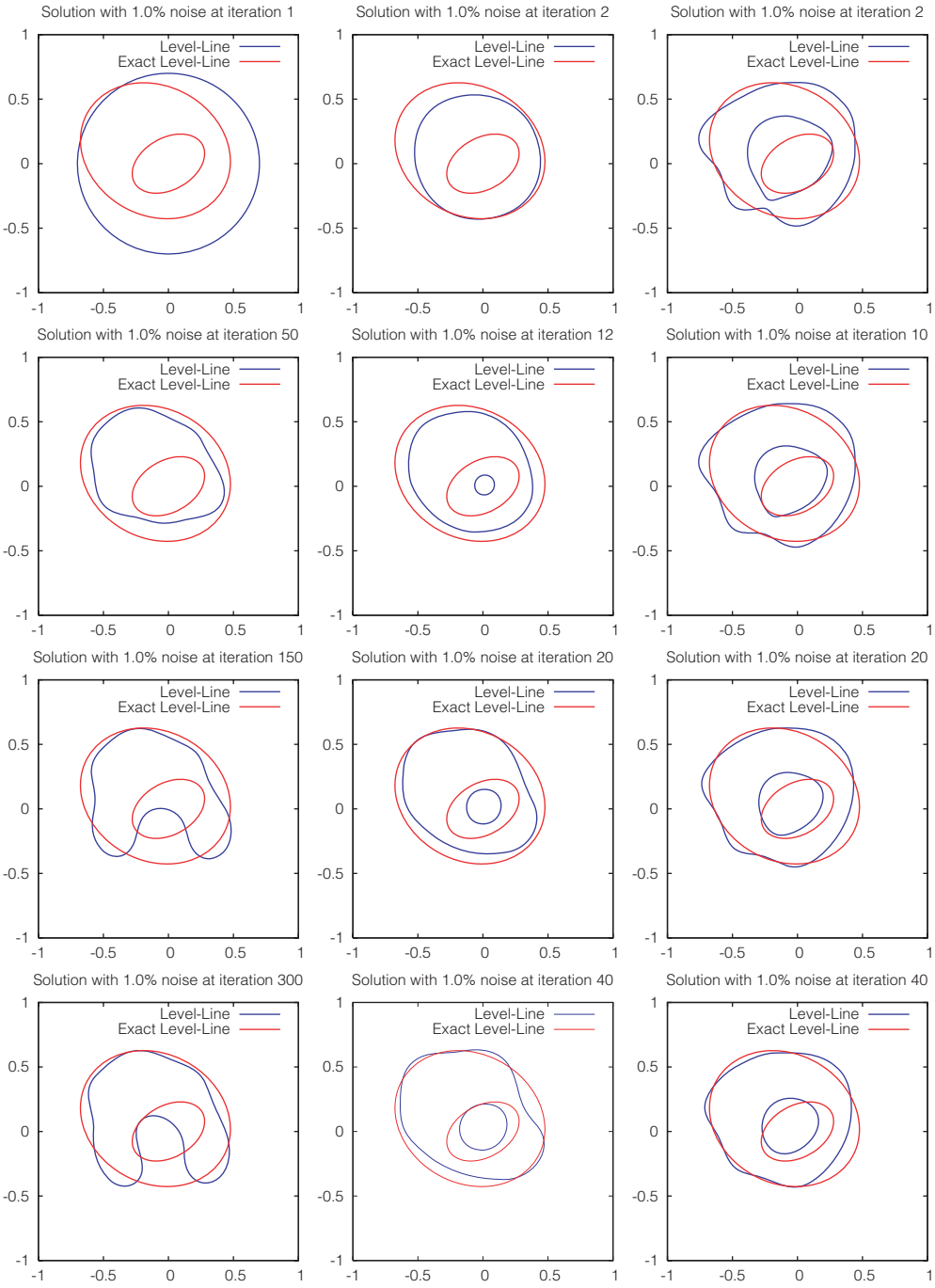
Finally we present in Table 5.1 the numbers of iterations needed for each method to get to the final solution, the corresponding values of the objective functional, the L^1-distance and the Hausdorff-distance, as well as the number of needed partial differential equation solver calls. For Newton-type phase I steps we solve a Newton system, implemented as SQP with an additional feasibility step. In our implementation one SQP call is approximately equivalent to 11 partial differential equation calls, when using also sparse direct solvers for the SQP system.

Our theoretic results for the topological derivative (Theorem 2.6) as well as for the steepest descent-type (3.3) and the Newton-type (3.6) topology changes are valid just inside the domain $\mathcal{D}$, and all constants depend on the distance to the boundary $\partial\mathcal{D}$.



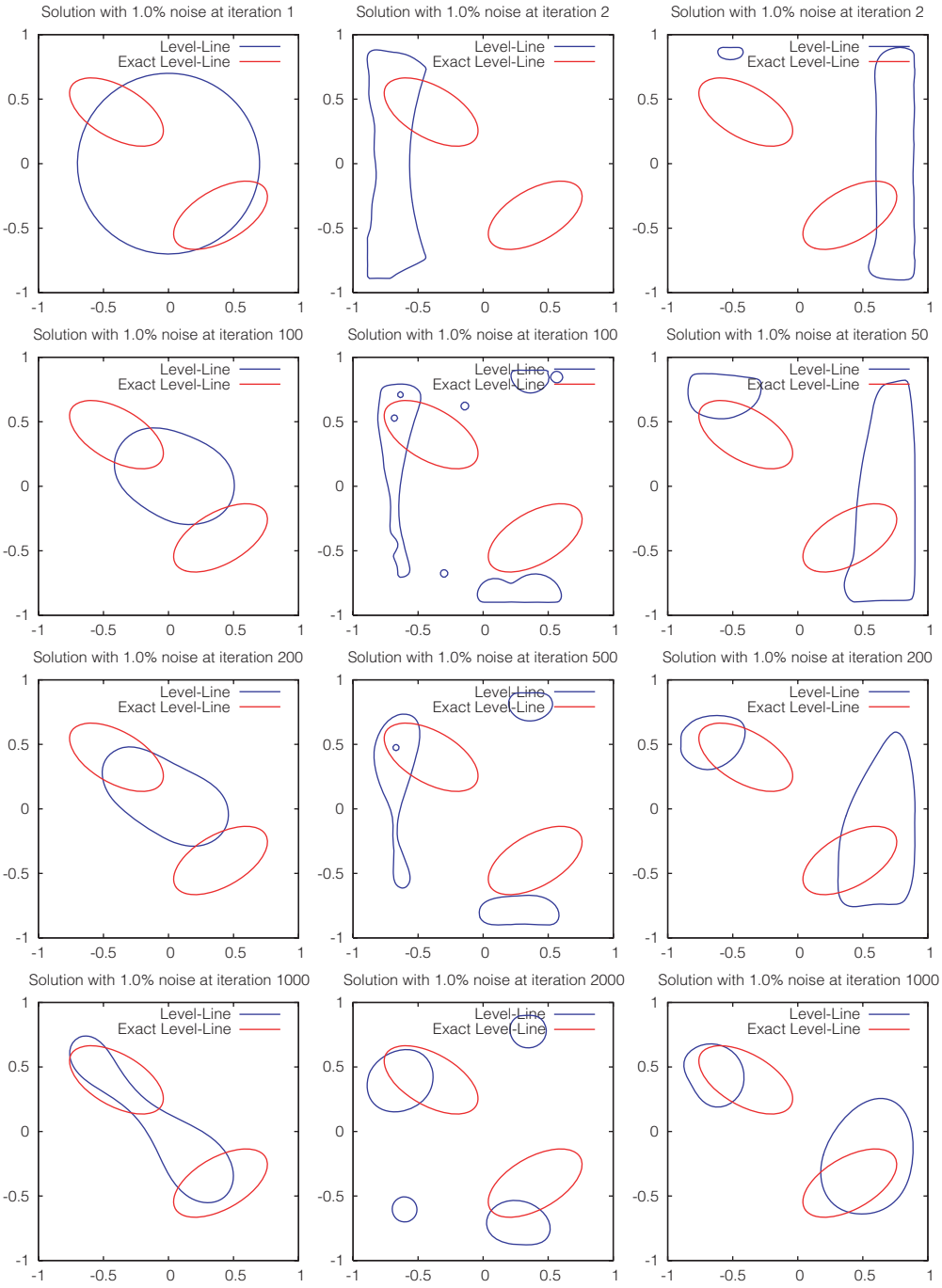
(a) Classical Level Set Methods (b) Steepest descent like topology changes (c) Newton type topology changes

FIG. 5.4. *Two ellipses, $\Gamma_M = \mathcal{D}$: Evolution of the algorithm.*



(a) Classical Level Set Methods (b) Steepest descent like topology changes (c) Newton type topology changes

FIG. 5.5. *Ellipse with elliptic hole, $\Gamma_M = \mathcal{D}$: Evolution of the algorithm.*



(a) Classical Level Set Methods (b) Steepest descent like topology changes (c) Newton type topology changes

FIG. 5.6. *Two ellipses, $\Gamma_M = \Gamma_N$: Evolution of the algorithm.*

TABLE 5.1
Iteration numbers k, objective $J_\alpha(\Omega_k)$, Hausdorff- and L^1-distance, and number of PDE calls for different algorithms.

Method	k	$J_\alpha(\Omega_k)$	$d_H(\Omega_k, \Omega^\dagger)$	$d_{L^1}(\Omega_k, \Omega^\dagger)$	# PDEs
Level set method	120	$3.13 \cdot 10^{-4}$	0.134	0.206	3×120
Steepest descent	200	$2.94 \cdot 10^{-4}$	0.130	0.207	$3 \times 200 + 20\times$H.-J.
Newton type	8	$3.08 \cdot 10^{-4}$	0.135	0.167	$3 \times 8 + 14\times$SQP system

(a) Two ellipses, $\Gamma_M = \mathcal{D}$.

Method	k	$J_\alpha(\Omega_k)$	$d_H(\Omega_k, \Omega^\dagger)$	$d_{L^1}(\Omega_k, \Omega^\dagger)$	# PDEs
Level set method	300	$2.19 \cdot 10^{-4}$	0.205	0.296	3×300
Steepest descent	40	$1.99 \cdot 10^{-4}$	0.127	0.230	$3 \times 40 + 4\times$H.-J.
Newton type	40	$1.99 \cdot 10^{-4}$	0.115	0.210	$3 \times 40 + 15\times$SQP system

(b) Ellipse with elliptic hole, $\Gamma_M = \mathcal{D}$.

Method	k	$J_\alpha(\Omega_k)$	$d_H(\Omega_k, \Omega^\dagger)$	$d_{L^1}(\Omega_k, \Omega^\dagger)$	# PDEs
Level set method	1000	$6.57 \cdot 10^{-4}$	0.290	0.455	3×1000
Steepest descent	2000	$5.33 \cdot 10^{-4}$	0.743	0.630	$3 \times 2000 + 200\times$H.-J.
Newton type	1000	$7.17 \cdot 10^{-4}$	0.395	0.565	$3 \times 1000 + 120\times$SQP system

(c) Two ellipses, $\Gamma_M = \Gamma_N$.

This was due to the use of the interior regularity result (Theorem 2.5). Therefore we restrict our algorithm to the domain $[\mathcal{D}]^{0.1} = [-0.9, 0.9]^2$. Furthermore we perform phase I just on the set $[\mathcal{D}]^{0.1} \cap [\partial\Omega_k]^{0.1}$ to guarantee a disjoint set of influence for phases I and II.

Our objective functional $J_\alpha(\Omega)$ incorporates perimeter regularization. To get a proper choice for the regularization parameter α we chose α such that the classical level set method, started at the exact solution (with 1% noisy data), does not iterate away from the exact solution too far. A too-large α would not allow us to achieve topology changes in our algorithms, and a too-small α does not regularize the problem enough. We found $\alpha = 10^{-5}\|\hat{u}\|^2_{L^2(\Gamma_M)}$ to be a proper choice for all examples.

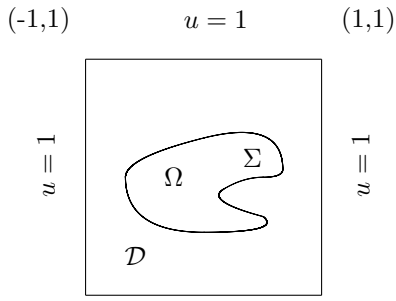
As an initial guess we take for all of our test examples a circle with radius $r = 0.7$ and centered at the origin $(0, 0)$ for the classical level set method, while we started with no material for the level set method incorporating steepest descent-type (respectively, Newton-type) topological changes.

For the level set method incorporating a steepest descent-type topology change we solve (4.2) after every 10th classical level set iterations, while for Newton-type topology changes we solve (4.4) after every 50th classical level set iteration.

In our implementation we did not implement any termination criteria except a maximal number of iterations. This is because our implementation of the L^1-distance and Hausdorff distance are not accurate enough to allow classical termination criteria from optimization terminate when $d_{L^1}(\Omega_{k+1}, \Omega_k)$ is small enough. The number of iterations needed to get to the optimum was estimated in a postprocessing step looking for the minimum in the L^1-distance $d_{L^1}(\Omega_k, \Omega^\dagger)$.

5.1. Full measurements $\Gamma_M = \mathcal{D}$. In this section we consider the identification of two ellipses and an ellipse with an elliptic hole from full measurements, i.e., $\Gamma_M = \mathcal{D}$. The full measurement case is mildly ill-posed, something like twice differentiation, and provides a lot of data. Therefore we can expect good results for all algorithms. Most challenging is probably the ellipse with an elliptic hole. For this case we expect that the classical level set method does not perform a topology change.

The partial differential equations (1.1), (2.4) are described by



$$
\begin{aligned}
-\Delta u + \chi_\Omega u &= 0 && \text{in } \mathcal{D},\\
u &= 1 && \text{on } \Gamma_D,\\
\frac{\partial u}{\partial n} &= 0 && \text{on } \Gamma_N,\\
-\Delta w + \chi_\Omega w &= -(u-\hat{u}) && \text{in } \mathcal{D},\\
w &= 0 && \text{on } \Gamma_D,\\
\frac{\partial w}{\partial n} &= 0 && \text{on } \Gamma_N.
\end{aligned}
$$

Measurements.

$$
\begin{aligned}
\hat{u} &= u|_{\mathcal{D}} + 1\%\ \text{noise} \quad \text{in } L^2(\mathcal{D}),\\
J_\alpha(\Omega) &= \tfrac{1}{2}\|u-\hat{u}\|^2_{L^2(\mathcal{D})} + \alpha \mathrm{Per}\big(\Omega\big).
\end{aligned}
$$

Due to the Dirichlet boundary conditions $u = 1$ on $\partial\mathcal{D}$ the solution $u(\Omega^\dagger)$ is close to the solution of the system without material $u(\emptyset) = 1$. Hence the subproblem (3.4) approximates the original problem (1.1) quite well, and we can expect that the level set method, incorporating Newton-type topology changes, performs very well. Indeed we will see that already the first solution to (3.6) provides a very good initial guess and predicts the correct topology of the desired geometry.

Two ellipses, $\boldsymbol{\Gamma_M = \mathcal{D}}$. We compare the classical level set method to a level set method incorporating a steepest descent-type topology change (4.2) and a level set method incorporating a Newton-type topology change (4.4). As expected all three methods perform very well and approximate the exact geometry quite accurately (Figure 5.4, last row).

Classical level set method (Figure 5.4, 1st column). The algorithm performs very well and even realizes the necessary topology change, by splitting. The number of iterations needed to approach the solution is moderate (see Table 5.1a). The distance to the exact geometry in both the L^1- and the Hausdorff metrics is reasonably small (see Figure 5.1).

Steepest descent-type topology change (Figure 5.4, 2nd column). The classical level set method incorporated into the steepest descent-type topology changes (4.2) does not predict the correct topology within the first solution to (4.2). It needs a second call (Figure 5.4, 2nd row, 2nd column) to generate a further topology change. Even when this topology change does not result in a substantial decrease in the objective functional (Figure 5.1a), the topological change can be observed in the jump of the Hausdorff distance $d_H(\Omega_k, \Omega^\dagger)$ (Figure 5.1c). Interestingly this topology change adds two new geometries at the correct position but does not try to reduce the wrong geometry. Further calls of (3.3) do not cause any changes of the geometry. The only topology change that occurs happens during the level set evolution where the above two geometries merge together. Even when the level set method incorporating a steepest descent-type topology change almost reaches the minimum of the objective functional $J_\alpha(\Omega)$ before the classical level set method (Figure 5.1a), it needs more iterations until it stays at the final geometry (see Table 5.1a).

Newton-type topology change (Figure 5.4, 3rd column). The first Newton-type topology change (3.6) already predicts the correct topology (Figure 5.4, 1st row, 3rd column). Also the objective functional $J_\alpha(\Omega)$ as well as the L^1- and Hausdorff distances (see Figure 5.1) get very close to their optimum and need just a few correction steps with the classical level set method. This is not too unexpected because the subproblem (3.4) approximates the nonlinear partial differential equation (1.1) very accurately. Although the number of iterations to approach the solution is very low,

note that one solution of (3.6) is more expansive (see Table 5.1a). Nonetheless the phase I/II with Newton-type topology changes is significantly faster than the other methods discussed above.

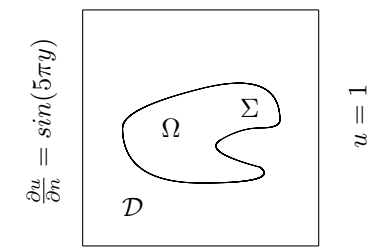
Elliptic hole in ellipse, $\boldsymbol{\Gamma_M = \mathcal{D}}$. This geometry is more challenging, and we expect that the classical level set method gets stuck in a local minima and does not predict the correct geometry. Here the power of the other two methods should show up. Again the subproblem (3.4) approximates the original partial differential equation (1.1) very well, and we can expect that the level set method incorporating Newton-type topology changes perform very well within the first solution of (3.6).

Classical level set method (Figure 5.5, 1st column). During the level set evolution the topology does not change although the objective functional $J_\alpha(\Omega)$ (Figure 5.2a) gets close to its minimum. Nonetheless the identified geometry does not look too bad (visually). The number of iterations doubled but is still moderate (see Table 5.1b).

Steepest-descent type topology change (Figure 5.5, 2nd column). As before the steepest descent-type topology change (3.3) does not generate the correct topology within its first call but needs two calls (Figure 5.5, 2nd row, 2nd column). Further solver calls of (3.3) do not force further changes. Hence after the second solution call of (3.3) we evolve the geometry just by the classical level set method. This can be observed in the L^1-distance as well as in the Hausdorff distance (Figures 5.2b, 5.2c) which have a jump at iteration 12. We are at the final solution at approximately 40 level set iterations, which makes an equivalent, due to 3 times calling a solver for (3.3), of 50 classical level set iterations in total (see Table 5.1b). In this test the steepest descent-type topology change results into the fastest solution.

Newton-type topology change (Figure 5.5, 3rd column). Again the first call to the Newton-type topology change (4.4) already predicts the topology correct (see Figure 5.5, 1st row, 3rd column). This time the objective functional $J_\alpha(\Omega)$ as well as the L^1- and Hausdorff distances (Figure 5.2) are not yet too close to their optimum. Hence we additionally need 20–40 classical level set iteration to end at the final solution. Summing up the level set iterations and the SQP equivalent for one solution of (4.4), the total cost is about 75–95 classical level set iterations. Therefore the algorithm is severely faster than the classical level set method but takes twice the time of the level set method incorporating steepest descent-type topology changes (see Table 5.1b).

5.2. Boundary neasurements $\boldsymbol{\Gamma_M = \Gamma_N}$. Finally we consider the identification of two ellipses from just one set of boundary measurements, i.e., $\Gamma_M = \Gamma_N$. To deal with boundary measurements we have to change slightly our boundary conditions for the partial differential equations (1.1), (2.4), namely,



$$
\begin{aligned}
-\Delta u + \chi_\Omega u &= 0 && \text{in } \mathcal{D},\\
u &= 1 && \text{on } \Gamma_D,\\
\frac{\partial u}{\partial n} &= h && \text{on } \Gamma_N,\\
-\Delta w + \chi_\Omega w &= 0 && \text{in } \mathcal{D},\\
w &= 0 && \text{on } \Gamma_D,\\
\frac{\partial w}{\partial n} &= -(u-\hat{u}) && \text{on } \Gamma_N.
\end{aligned}
$$

Measurements.

$$
\begin{aligned}
\hat{u} &= u|_{\Gamma_N} + 1\% \text{ noise} \quad \text{in } L^2(\Gamma_N),\\
J_\alpha(\Omega) &= \tfrac{1}{2}\|u-\hat{u}\|^2_{L^2(\Gamma_N)} + \alpha \mathrm{Per}\big(\Omega\big).
\end{aligned}
$$

Geometric inverse problems with boundary measurements are supposed to be severely ill-posed. Severely ill-posed problems are extremely challenging to every algorithm, and usually one cannot expect too good results for them. Especially topology changes are extremely difficult to achieve. We expect that the classical level set method is not able to perform the desired topology change, even when it managed it for the full measurements case.

Due to the Neumann boundary conditions the solution u to the above system is not close to the solution $u(\emptyset) = 1$ (solution without material). Hence the subproblem (3.4) does not approximate the original problem (1.1) very well (when starting with no material). As a consequence of this the first step of a Newton-type topology change (4.4) shall not perform as good as in the full measurement cases.

Two ellipses $\boldsymbol{\Gamma_M = \Gamma_N}$. Once more we compare the classical level set method to a level set method incorporating a steepest descent-type topology change (4.2) and a level set method incorporating a Newton-type topology change (4.4).

Classical level set method (Figure 5.6, 1st column). As predicted, the classical level set method does not split. Nonetheless the finally identified geometry does not look too bad (visually). The number of iterations needed, until it approaches its optimum, is very high, but this is not uncommon for severely ill-posed problems. Although we do not get the correct topology, the objective functional $J_\alpha(\Omega)$ (Figure 5.3a) gets close to its minimum.

Steepest descent-type topology change (Figure 5.6, 2nd column). As before the classical level set method incorporating steepest descent-type topology changes (4.2) does not predict the correct topology within the first solution to (4.2). Iterating further and calculating several times the solution to (4.2), the algorithm forces further topology changes, some of them correctly located, some of them not (Figure 5.6, 1st column). Nonetheless the objective functional $J_\alpha(\Omega)$ decreases, as predicted by the theory. After many iterations the algorithm stops with four nonconnected components, where two are correctly located and the others are not.

Newton-type topology change (Figure 5.6, 3rd column). Finally we consider the classical level set method incorporating Newton-type topology changes (4.4). Already the first calculation of the solution to (4.4) predicts the correct number of connected components, and later solution calls of (4.4) do not force any additional topology changes. As for the steepest descent-type topology changes, the first solution to (4.4) does not look too good, but it is enough for the classical level set method to approach the exact solution. For the final result presented in Figure 5.6, 3rd column, 4th row, the objective functional and also the L^1- and Hausdorff distances (Figure 5.3) would decrease further. Hence, iterating further would still improve the result. Nonetheless we terminated the algorithm, because the number of iterations is already very high, and we can already see from Figure 5.6, 3rd column, that the algorithm behaves better than the two other.

6. Conclusion. In this paper we presented a way to generalize the notion of topological derivatives such that we can also deal with perimeter-regularized objective functionals. The generalization allows one to formulate auxiliary minimization problems similar to steepest descent-type and Newton-type minimization problems, such that a descent in the objective functional is guaranteed. This is in contrast to classical topological derivatives, where one gets just an indicator of where to force topology changes, but the indicator does not guarantee a descent in the objective functional.

We incorporated this generalization of topological derivatives into the classical level set method and showed by means of some examples its applicability. While, in some cases, the classical level set method failed to predict the correct topology, the suggested level set methods with incorporated steepest descent-type and Newton-type topology changes succeed to get the correct topology or at least forced topology changes.

The numerical results for the specific example presented in this paper were quite promising, and an extension to more complicated problems might be of interest.

Acknowledgments. I thank Dr. Martin Burger (University of Linz) for many useful discussions. My gratitude goes also to the two anonymous referees whose comments improved the presentation of the paper severely.

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, New York, San Francisco, London, 1975.
[2] G. Allaire, F. J. F. de Gournay, and A.-M. Toader, *Structural optimization using topological and shape sensitivity via a level set method*, Control Cybernet., 34 (2005), pp. 59–80.
[3] G. Allaire, F. Jouve, and A.-M. Toader, *A level-set method for shape optimization*, C. R. Math. Acad. Sci. Paris, 334 (2002), pp. 1125–1130.
[4] G. Allaire, F. Jouve, and A.-M. Toader, *Structural optimization using sensitivity analysis and a level-set method*, J. Comput. Phys., 194 (2004), pp. 363–393.
[5] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, Oxford University Press, Oxford, 2000.
[6] S. Amstutz, *Sensitivity analysis with respect to a local perturbation of the material property*, Asymptot. Anal., 49 (2006), pp. 87–108.
[7] S. Amstutz and H. Andrä, *A new algorithm for topology optimization using a level-set method*, J. Comput. Phys., 216 (2006), pp. 573–588.
[8] S. Amstutz and N. Dominguez, *Topological sensitivity in the context of ultrasonic nondestructive testing*, Engineering Analysis with Boundary Elements (special issue).
[9] S. Amstutz, I. Horchani, and M. Masmoudi, *Crack detection by the topological gradient method*, Control Cybernet., 34 (2005), pp. 81–101.
[10] H. Ben Ameur, M. Burger, and B. Hackl, *Level set methods for geometric inverse problems in linear elasticity*, Inverse Problems, 20 (2004), pp. 673–696.
[11] M. P. Bendsøe and O. Sigmund, *Topology Optimization: Theory, Methods and Applications*, Springer, Berlin, 2002.
[12] B. Bourdin and A. Chambolle, *Design-dependent loads in topology optimization*, ESAIM Control Optim. Calc. Var., 9 (2003), pp. 19–48.
[13] M. Burger, *A level set method for inverse problems*, Inverse Problems, 17 (2001), pp. 1327–1355.
[14] M. Burger, *A framework for the construction of level set methods for shape optimization and reconstruction*, Interfaces Free Bound., 5 (2003), pp. 301–329.
[15] M. Burger, *Levenberg-Marquardt level set methods for inverse obstacle problems*, Inverse Problems, 20 (2004), pp. 259–282.
[16] M. Burger, B. Hackl, and W. Ring, *Incorporating topological derivatives into level sets methods*, J. Comput. Phys., 194 (2004), pp. 334–362.
[17] M. Burger and M. Hintermüller, *Projected gradient flows for BV/level set relaxation*, PAMM, 5 (2005), pp. 11–14.
[18] M. Burger and S. J. Osher, *A survey on level set methods for inverse problems and optimal design*, European J. Appl. Math., 16 (2005), pp. 263–301.
[19] J. Céa, S. Garreau, P. Guillaume, and M. Masmoudi, *The shape and topological optimizations connection*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 713–726.
[20] D. J. Cedio-Fengya, S. Moskow, and M. S. Vogelius, *Identification of conductivity imperfections of small diameter by boundary measurements. Coninuous dependence and computational reconstruction*, Inverse Problems, 14 (1998), pp. 553–598.
[21] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering Theory*, 2nd ed., Springer, Berlin, 1998.
[22] M. C. Delfour and J.-P. Zolésio, *Shapes and Geometries: Analysis, Differential Calculus, and Optimization*, Advances in Design and Control, SIAM, Philadelphia, 2001.

[23] O. Dorn, E. L. Miller, and C. M. Rappaport, *A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets*, Inverse Problems, 16 (2000), pp. 1119–1156.

[24] H. A. Eschenauer, V. V. Kobelev, and A. Schumacher, *Bubble method for topology and shape optimization of structures*, J. Struct. Optim., 8 (1994), pp. 42–51.

[25] H. A. Eschenauer and A. Schumacher, *Topology optimization procedure using hole positioning criteria - theory and applications*, in Topology Optimization in Structural Mechanics, CISM Courses and Lectures 374, Springer, New York, 1997, pp. 135–196.

[26] M. Giaquinta, *Introduction to Regularity Theory for Nonlinear Elliptic Systems*, Birkhäuser, Basel, Boston, Berlin, 1993.

[27] Ph. Guillaume and K. S. Idris, *Topological sensitivity and shape optimization for the Stokes equations*, SIAM J. Control Optim., 43 (2004), pp. 1–31.

[28] B. B. Guzina and M. Bonnet, *Topological derivative for the inverse scattering of elastic waves*, Quart. J. Mech. Appl. Math., 57 (2004), pp. 161–179.

[29] A. Habib and K. Hyeonbae, *Reconstuction of Small Inhomogeneities from Boundary Measurements*, Lecture Notes in Mathematics, Springer, Berlin, 2004.

[30] B. Hackl, *Shape Variations, Level Sets- and Phasfield- Methods for Perimeter Regularized Geometric Inverse Problems*, Ph.D. thesis, Johannes Kepler Universität, Linz, Austria, 2006.

[31] F. Hettlich and W. Rundell, *Recovery of the support of a source term in an elliptic differential equation*, Inverse Problems, 13 (1997), pp. 959–976.

[32] M. Hintermüller, *Shape and topological sensitivity*, in Reat-Time PDE-Constrained Optimization, L. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. van Bloemen Waanders, eds., SIAM, Philadelphia, 2007, pp. 253–276.

[33] M. Hintermüller and W. Ring, *A second order shape optimization approach for image segmentation*, SIAM J. Appl. Math., 64 (2003), pp. 442–467.

[34] K. Ito, K. Kunisch, and Z. Li, *Level-set function approach to an inverse interface problem*, Inverse Problems, 17 (2001), pp. 1225–1242.

[35] G.-S. Jiang and D. Peng, *Weighted ENO schemes for Hamilton-Jacobi equations*, SIAM J. Sci. Comput., 21 (2000), pp. 2126–2143.

[36] A. Litman, D. Lesselier, and F. Santosa, *Reconstruction of a two-dimensional binary obstacle by controlled evolution of a level-set*, Inverse Problems, 14 (1998), pp. 685–706.

[37] M. Masmoudi, J. Pommier, and B. Samet, *The topological asymptotic expansion for the Maxwell equations and some applications*, Inverse Problems, 21 (2005), pp. 547–564.

[38] A. Novruzi and M. Pierre, *Structure of shape derivatives*, J. Evol. Equ., 2 (2002), pp. 365–382.

[39] S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin, *An iterative regularization method for total variation-based image restoration*, Multiscale Model. Simul., 4 (2005), pp. 460–489.

[40] S. J. Osher and R. Fedkiw, *Level Set Methods and Dynamic Implicit Surfaces*, Springer, New York, Berlin, Heidelberg, 2003.

[41] S. J. Osher and F. Santosa, *Level set methods for optimization problems involving geometry and constraints. I: Frequencies of a two-density inhomogeneous drum*, J. Comput. Phys., 171 (2001), pp. 272–288.

[42] F. Santosa, *A level-set approach for inverse problems involving obstacles*, ESAIM Control Optim. Calc. Var., 1 (1996), pp. 17–33.

[43] J. Sokolowski and A. Żochowski, *On the topological derivative in shape optimization*, SIAM J. Control Optim., 37 (1999), pp. 1251–1272.

[44] J. Sokolowski and A. Żochowski, *Topological derivatives for elliptic problems*, Inverse Problems, 15 (1999), pp. 123–134.

[45] J. Sokolowski and A. Żochowski, *Optimality conditions for simultaneous topology and shape optimization*, SIAM J. Control Optim., 42 (2003), pp. 1198–1221.

[46] Y.-H. R. Tsai and S. J. Osher, *Level set methods and their application in image sciences*, Commun. Math. Sci., (2003), pp. 623–656.

[47] X. Wang, Y. Mei, and M. Y. Wang, *Incorporating topological derivatives into level set methods for structural topology optimization*, in Optimal Shape Design and Modeling, T. Lewinski et al., eds., Polish Academy of Sciences, Warsaw, 2004, pp. 145–157.

SIAM J. NUMER. ANAL.
Vol. 45, No. 5, pp. 2228–2258

CONVERGENCE ANALYSIS OF A MIXED FINITE VOLUME SCHEME FOR AN ELLIPTIC-PARABOLIC SYSTEM MODELING MISCIBLE FLUID FLOWS IN POROUS MEDIA*

CLAIRE CHAINAIS-HILLAIRET† AND JÉRÔME DRONIOU‡

Abstract. We study a finite volume discretization of a strongly coupled elliptic-parabolic PDE system describing miscible displacement in a porous medium. We discretize each equation by a finite volume scheme which allows a wide variety of unstructured grids (in any space dimension) and gives strong enough convergence for handling the nonlinear coupling of the equations. We prove the convergence of the scheme as the time and space steps go to 0. Finally, we provide numerical results to demonstrate the efficiency of the proposed numerical scheme.

Key words. finite volume methods, porous medium, miscible fluid flow, convergence analysis, numerical tests

AMS subject classifications. 65M12, 65M30, 65N12, 65N30, 76S05, 76R99

DOI. 10.1137/060657236

1. Introduction.

1.1. Miscible displacement in porous media.

The mathematical model for the single-phase miscible displacement of one fluid by another in a porous medium, in the case where the fluids are considered incompressible, is an elliptic-parabolic coupled system [2, 4]. Let Ω be a bounded domain of $\mathbb{R}^d$ ($d = 2$ or 3) representing the reservoir and let $(0,T)$ be the time interval. The unknowns of the problem are p the pressure in the mixture, $\mathbf{U}$ its Darcy velocity, and c the concentration of the invading fluid.

We denote by $\Phi(x)$ and $\mathbf{K}(x)$ the porosity and the absolute permeability tensor of the porous medium, $\mu(c)$ the viscosity of the fluid mixture, $\widehat{c}$ the injected concentration, and q^+ and q^- the injection and the production source terms. If we neglect gravity, the model reads

$$\begin{cases} \operatorname{div}(\mathbf{U}) = q^+ - q^- & \text{in } (0,T)\times\Omega, \\ \mathbf{U} = -\dfrac{\mathbf{K}(x)}{\mu(c)}\nabla p & \text{in } (0,T)\times\Omega, \end{cases} \tag{1}$$

$$\Phi(x)\partial_t c - \operatorname{div}(D(x,\mathbf{U})\nabla c - c\mathbf{U}) + q^- c = q^+\widehat{c} \quad \text{in } (0,T)\times\Omega, \tag{2}$$

where D is the diffusion-dispersion tensor including molecular diffusion and mechanical dispersion

$$D(x,\mathbf{U}) = \Phi(x)\Big(d_m\mathbf{I} + |\mathbf{U}|\Big(d_l E(\mathbf{U}) + d_t(\mathbf{I} - E(\mathbf{U}))\Big)\Big) \tag{3}$$

*Received by the editors April 14, 2006; accepted for publication (in revised form) February 15, 2007; published electronically October 5, 2007.
http://www.siam.org/journals/sinum/45-5/65723.html
†Laboratoire de Mathématiques, UMR CNRS 6620, Université Blaise Pascal, 63177 Aubière cedex, France (Claire.Chainais@math.univ-bpclermont.fr).
‡Département de Mathématiques, UMR CNRS 5149, CC 051, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier cedex 5, France (droniou@math.univ-montp2.fr).

with $\mathbf{I}$ the identity matrix, d_m the molecular diffusion, d_l and d_t the longitudinal and transverse dispersion coefficients, and $E(\mathbf{U}) = (\frac{\mathbf{U}_i\mathbf{U}_j}{|\mathbf{U}|^2})_{1\le i,j\le d}$. Laboratory experiments have found that the longitudinal dispersivity d_l is much greater than the transverse dispersivity d_t and that the diffusion coefficient is very small by comparison.

In reservoir simulation, the boundary $\partial\Omega$ is typically impermeable. Therefore, if $\mathbf{n}$ denotes the exterior normal to $\partial\Omega$, the system (1)–(2) is supplemented with no flow boundary conditions:

$$\begin{cases} \mathbf{U}\cdot\mathbf{n} = 0 & \text{on } (0,T)\times\partial\Omega, \\ D(x,\mathbf{U})\nabla c\cdot\mathbf{n} = 0 & \text{on } (0,T)\times\partial\Omega. \end{cases} \tag{4}$$

An initial condition is also prescribed:

$$c(x,0) = c_0(x) \text{ in } \Omega. \tag{5}$$

Because of the homogeneous Neumann boundary conditions on $\mathbf{U}$, the injection and production source terms have to satisfy the compatibility condition $\int_\Omega q^+(\cdot,x)\ dx = \int_\Omega q^-(\cdot,x)\ dx$ in $(0,T)$, and since the pressure is defined only up to an arbitrary constant, we normalize p by the following condition:

$$\int_\Omega p(\cdot,x)\ dx = 0 \quad \text{in } (0,T). \tag{6}$$

The viscosity μ is usually determined by the following mixing rule

$$\mu(c) = \mu(0)\Big(1 + \Big(M^{1/4} - 1\Big)c\Big)^{-4} \text{ in } [0,1], \tag{7}$$

where $M = \frac{\mu(0)}{\mu(1)}$ is the mobility ratio (μ can be extended to $\mathbb{R}$ by letting $\mu = \mu(0)$ on $(-\infty,0)$ and $\mu = \mu(1)$ on $(1,\infty)$). The porosity Φ and the permeability $\mathbf{K}$ are in general assumed to be bounded from above and from below by positive constants (or positive multiples of $\mathbf{I}$ for the tensor $\mathbf{K}$).

In [15], Feng proved the existence of a weak solution to the problem (1)–(7) in the two-dimensional case and with $d_l \ge d_t > 0$ and $d_m > 0$. This result has been generalized by Chen and Ewing in [3] to the three-dimensional case and with gravity effects and various boundary conditions. At high flow velocities the effects of mechanical dispersion are much greater than those of molecular diffusion. Therefore, Amirat and Ziani studied in [1] the asymptotic behavior of the weak solution as d_m goes to 0 and proved the existence of a weak solution in the case where $d_m = 0$.

From a numerical point of view, various methods have already been developed for this problem. In general the pressure equation is discretized by a finite element method. However, the key point is that equation (2) on c is a convection-dominated equation, which is not well adapted to the discretization by finite difference or finite element methods. Douglas, Ewing, and Wheeler [6] used a mixed finite element method for the pressure equation and a Galerkin finite element method for the concentration equation. In [19], Russell introduced a modified method of characteristic for the resolution of (2), while (1) is solved by a finite element method. Then, Ewing, Russell, and Wheeler [10] combined a mixed finite element method for (1) and a modified method of characteristic for (2). In [20, 21], the authors also used a mixed

finite element method for (1) but developed an Eulerian Lagrangian localized adjoint method for (2).

Convergence of numerical schemes to (1)–(7) (or connected problems) has already been studied (see, e.g., [5, 6, 11, 12, 17]). But, to the best of our knowledge, these proofs of convergence are based on a priori error estimates, which need regularity assumptions on the solution $(p, \mathbf{U}, c)$ to the continuous problem. Such regularity does not seem provable in general, such as if we take a discontinuous permeability tensor (which is expected in field applications; see [20]).

Finite volume methods are well adapted to the discretization of conservation laws; see, for instance, the reference book by Eymard, Gallouët, and Herbin [13]. They provide efficient numerical schemes for elliptic equations as well as for convection-dominated parabolic equations. However, because of the anisotropic diffusion in (1) (due to $\mathbf{K}(x)$) and of the dispersion terms in (2)–(3), the standard four-point finite volume schemes cannot be used here. Besides, as said above, (2) is convection-dominated and, therefore, a good approximation of $\mathbf{U}$ is needed in the discretization of (2) in order to obtain admissible numerical results. In [9], Droniou and Eymard recently proposed a mixed finite volume scheme which handles anisotropic heterogeneous diffusion problems on any grid and precisely provides, for equations such as (1), good approximations of $\mathbf{U}$; this scheme is therefore a natural candidate to discretize such coupled problems as (1)–(7), especially as it has been shown to behave well from a numerical point of view.

In this paper, we extend the mixed finite volume scheme of [9] to a system, presented in section 1.2, which generalizes (1)–(7). Section 2 contains the definition of the scheme and the statement of the main results: existence and uniqueness of an approximate solution and its convergence to the solution of the continuous problem as the time and space steps tend to 0. A priori estimates on the approximate solution are established in section 3, and in section 4 we prove the existence and uniqueness of the solution to our scheme. The proof of convergence is presented in section 5, under no regularity assumption on the solution to the continuous problem. Section 6 presents some numerical experiments to demonstrate the efficiency of our numerical scheme. Section 7 is an appendix containing a few technical results.

1.2. Formulation of the problem and assumptions. Let us now rewrite the problem (1)–(7) under the following synthesized and more general form (notice that, from now on, we use letters with bar accents to denote the exact solutions, and we use letters without bar accents to denote approximate solutions):

$$(8)\quad \begin{cases} \operatorname{div}(\bar{\mathbf{U}}) = q^+ - q^- \ \text{ in } (0,T)\times\Omega, & \bar{\mathbf{U}} = -A(\cdot,\bar c)\nabla \bar p \ \text{ in } (0,T)\times\Omega, \\ \displaystyle\int_\Omega \bar p(\cdot,x)\,dx = 0 \ \text{ in } (0,T), & \bar{\mathbf{U}}\cdot\mathbf{n} = 0 \ \text{ on } (0,T)\times\partial\Omega, \end{cases}$$

$$(9)\quad \begin{cases} \Phi\partial_t \bar c - \operatorname{div}(D(\cdot,\bar{\mathbf{U}})\nabla\bar c) + \operatorname{div}(\bar c\bar{\mathbf{U}}) + q^-\bar c = q^+\widehat{c} & \text{in } (0,T)\times\Omega, \\ \bar c(0,\cdot) = c_0 & \text{in } \Omega, \\ D(\cdot,\bar{\mathbf{U}})\nabla\bar c\cdot\mathbf{n} = 0 & \text{on } (0,T)\times\partial\Omega. \end{cases}$$

In what follows, we assume that Ω is a convex polygonal bounded domain of $\mathbb{R}^d$, $T>0$, and the following:

$$(10)\quad \begin{aligned} &(q^+,q^-)\in L^\infty(0,T;L^2(\Omega)) \text{ are nonnegative,} \\ &\int_\Omega q^+(\cdot,x)\,dx = \int_\Omega q^-(\cdot,x)\,dx \text{ a.e. in } (0,T), \end{aligned}$$

$A : \Omega \times \mathbb{R} \to M_d(\mathbb{R})$ is a Carathéodory function satisfying the following:

$$\exists \alpha_A > 0\,,\ \exists \Lambda_A > 0 \text{ such that, for a.e. } x \in \Omega, \text{ all } s \in \mathbb{R}, \text{ and all } \xi \in \mathbb{R}^d, \tag{11}$$
$$A(x,s)\xi \cdot \xi \geq \alpha_A |\xi|^2 \text{ and } |A(x,s)| \leq \Lambda_A,$$

$D : \Omega \times \mathbb{R}^d \to M_d(\mathbb{R})$ is a Carathéodory function satisfying the following:

$$\exists \alpha_D > 0\,,\ \exists \Lambda_D > 0 \text{ such that, for a.e. } x \in \Omega, \text{ all } \mathbf{W} \in \mathbb{R}^d, \text{ and all } \xi \in \mathbb{R}^d, \tag{12}$$
$$D(x,\mathbf{W})\xi \cdot \xi \geq \alpha_D (1 + |\mathbf{W}|)|\xi|^2 \text{ and } |D(x,\mathbf{W})| \leq \Lambda_D (1 + |\mathbf{W}|),$$

$$\Phi \in L^\infty(\Omega) \text{ and there exists } \Phi_* > 0 \text{ such that } \Phi_* \leq \Phi \leq \Phi_*^{-1} \text{ a.e. in } \Omega, \tag{13}$$

$$\widehat{c} \in L^\infty((0,T) \times \Omega) \text{ satisfies } 0 \leq \widehat{c} \leq 1 \text{ a.e. in } (0,T) \times \Omega, \tag{14}$$

$$c_0 \in L^\infty(\Omega) \text{ satisfies } 0 \leq c_0 \leq 1 \text{ a.e. in } \Omega. \tag{15}$$

Remark 1.1. Since $E(\mathbf{U}) = \left(\mathbf{U}_i \mathbf{U}_j / |\mathbf{U}|^2\right)_{1 \leq i,j \leq d}$ is the orthogonal projector on $\mathbb{R}\mathbf{U}$, the model in section 1.1 satisfies this assumptions with $\alpha_D = \phi_* \inf(d_m, d_l, d_t)$ and $\Lambda_D = \phi_*^{-1} \sup(d_m, d_l, d_t)$.

As Φ does not depend on t, the following definition (similar to the one in [15]) of weak solution to (8)–(9) makes sense.

DEFINITION 1.1. *Under assumptions* (10)–(15), *a weak solution to* (8)–(9) *is* $(\bar{p}, \bar{\mathbf{U}}, \bar{c})$ *such that* $\bar{p} \in L^\infty(0,T; H^1(\Omega))$, $\bar{\mathbf{U}} \in L^\infty(0,T; L^2(\Omega))^d$, $\bar{c} \in L^\infty(0,T; L^2(\Omega)) \cap L^2(0,T; H^1(\Omega))$,

$$\int_\Omega \bar{p}(t,\cdot) = 0 \text{ for a.e. } t \in (0,T)\,, \quad \bar{\mathbf{U}} = -A(\cdot,\bar{c})\nabla \bar{p} \text{ a.e. in } (0,T) \times \Omega,$$
$$\forall \varphi \in C^\infty([0,T] \times \overline{\Omega})\,,\ -\int_0^T \int_\Omega \bar{\mathbf{U}} \cdot \nabla \varphi = \int_0^T \int_\Omega (q^+ - q^-)\varphi\,,$$
$$\forall \psi \in C_c^\infty([0,T) \times \overline{\Omega})\,,\ -\int_0^T \int_\Omega \Phi \bar{c} \partial_t \psi + \int_0^T \int_\Omega D(\cdot, \bar{\mathbf{U}}) \nabla \bar{c} \cdot \nabla \psi - \int_0^T \int_\Omega \bar{c} \bar{\mathbf{U}} \cdot \nabla \psi$$
$$+ \int_0^T \int_\Omega q^- \bar{c} \psi - \int_\Omega \Phi c_0 \psi(0,\cdot) = \int_0^T \int_\Omega q^+ \widehat{c} \psi.$$

2. Scheme and main results. Let us first define the notion of admissible mesh of Ω and some notation associated with it.

DEFINITION 2.1. *Let* Ω *be a convex polygonal bounded domain in* $\mathbb{R}^d$. *An admissible mesh of* Ω *is given by* $\mathcal{D} = (\mathcal{M}, \mathcal{E})$, *where the following hold:*

(i) $\mathcal{M}$ *is a finite family of nonempty disjoint convex polygonal domains in* Ω *(the "control volumes") such that* $\overline{\Omega} = \cup_{K \in \mathcal{M}} \overline{K}$.

(ii) $\mathcal{E}$ *is a finite family of disjoint subsets of* $\overline{\Omega}$ *(the "edges" of the mesh), such that, for all* $\sigma \in \mathcal{E}$, *there exists an affine hyperplane* E *of* $\mathbb{R}^d$ *and* $K \in \mathcal{M}$ *verifying that* $\sigma \subset \partial K \cap E$ *and* σ *is a nonempty open convex subset of* E. *We assume that, for all* $K \in \mathcal{M}$, *there exists a subset* $\mathcal{E}_K$ *of* $\mathcal{E}$ *such that* $\partial K = \cup_{\sigma \in \mathcal{E}_K} \overline{\sigma}$. *We also assume that, for all* $\sigma \in \mathcal{E}$, *either* $\sigma \subset \partial\Omega$ *or* $\overline{\sigma} = \overline{K} \cap \overline{L}$ *for some* $(K,L) \in \mathcal{M} \times \mathcal{M}$.

The d-dimensional measure of a control volume K is denoted by $\mathrm{m}(K)$, and the $(d-1)$-dimensional measure of an edge σ by $\mathrm{m}(\sigma)$; in the integral signs, γ denotes the measure on the edges. If $\sigma \in \mathcal{E}_K$, then $\mathbf{n}_{K,\sigma}$ is the unit normal to σ outward to K. In the case where $\sigma \in \mathcal{E}$ satisfies $\overline{\sigma} = \overline{K} \cap \overline{L}$ for $(K,L) \in \mathcal{M} \times \mathcal{M}$, we denote $\sigma = K|L$ (K and L are then called "neighboring control volumes"). We define the set of interior (resp., boundary) edges as $\mathcal{E}_{\rm int} = \{\sigma \in \mathcal{E};\ \sigma \not\subset \partial\Omega\}$ (resp., $\mathcal{E}_{\rm ext} = \{\sigma \in \mathcal{E};\ \sigma \subset \partial\Omega\}$). For all $K \in \mathcal{M}$ and all $\sigma \in \mathcal{E}$, $\mathbf{x}_K$ and $\mathbf{x}_\sigma$ are the respective barycenters of K and σ.

The size of a mesh $\mathcal{D}$ is $\mathrm{size}(\mathcal{D}) = \sup_{K\in\mathcal{M}} \mathrm{diam}(K)$. The following quantity measures the regularity of the mesh

$$\mathrm{regul}(\mathcal{D}) = \sup\left\{ \max\left(\frac{\mathrm{diam}(K)^d}{\rho_K^d}, \mathrm{Card}(\mathcal{E}_K) \right) \ ;\ K \in \mathcal{M} \right\},$$

where, for $K \in \mathcal{M}$, ρ_K is the supremum of the radius of the balls contained in K. The definition of $\mathrm{regul}(\mathcal{D})$ implies that, if ω_d is the volume of the unit ball in $\mathbb{R}^d$, for all $K \in \mathcal{M}$,

$$\mathrm{diam}(K)^d \le \mathrm{regul}(\mathcal{D})\rho_K^d \le \frac{\mathrm{regul}(\mathcal{D})}{\omega_d}\mathrm{m}(K)\,. \tag{16}$$

Remark 2.1. We ask for very few geometrical constraints on the mesh of Ω. This is particularly important since, in real-world problems, meshes used in basin and reservoir simulations can be quite irregular and not admissible in the usual finite element or finite volume senses (see [14]).

Our scheme is based on the mixed finite volume scheme introduced in [9] and, for elliptic equations, [8]. Its main goal is to handle a wide variety of grids for heterogeneous and anisotropic operators while giving strong convergence of approximate gradients. Therefore, this scheme applied to (8) provides a strong approximation of $\bar{\mathbf{U}}$, which can then be used in the discretization of the convective term $\mathrm{div}(\bar{c}\bar{\mathbf{U}})$ in the parabolic equation.

The idea is to consider, besides unknowns which approximate the functions $(\bar{p}, \bar{c})$, unknowns which approximate the gradients of these functions, as well as unknowns which stand for the fluxes associated with the differential operators. Thus, if $\mathcal{D}$ is an admissible mesh of Ω and $k > 0$ is a time step (we always choose time steps such that $N_k = T/k$ is an integer), we consider, for all $n = 1, \dots, N_k$ and all $K \in \mathcal{M}$, unknowns $(p_K^n, \mathbf{v}_K^n)$ which stand for approximate values of $(\bar{p}, \nabla\bar{p})$ on $[(n-1)k, nk) \times K$ and numbers $F_{K,\sigma}^n$ (for $\sigma \in \mathcal{E}_K$) which stand for approximate values of $-\int_\sigma \bar{\mathbf{U}} \cdot \mathbf{n}_{K,\sigma}\, d\gamma$ on $[(n-1)k, nk)$. Similarly, the unknowns $(c_K^n, \mathbf{w}_K^n)$ approximate $(\bar{c}, \nabla\bar{c})$ on $[(n-1)k, nk) \times K$ and the numbers $G_{K,\sigma}^n$ (for $\sigma \in \mathcal{E}_K$) approximate $\int_\sigma D(\cdot, \bar{\mathbf{U}})\nabla\bar{c} \cdot \mathbf{n}_{K,\sigma}\, d\gamma$ on $[(n-1)k, nk)$.

The quantities $q_K^{+,n}$, $q_K^{-,n}$, and $\widehat{c}_K^n$ denote the mean values of q^+, q^-, and $\widehat{c}$ on $[(n-1)k, nk) \times K$, and Φ_K, c_K^0, $A_K(s)$, and $D_K(\xi)$ are the mean values of Φ, c_0, $A(\cdot, s)$, and $D(\cdot, \xi)$ on K. We also take positive numbers $(\nu_K)_{K\in\mathcal{M}}$. The scheme for (8) reads as follows: for all $n = 1, \dots, N_k$,

$$\begin{aligned} \mathbf{v}_K^n \cdot (\mathbf{x}_\sigma - \mathbf{x}_K) + \mathbf{v}_L^n \cdot (\mathbf{x}_L - \mathbf{x}_\sigma) + \nu_K \mathrm{m}(K) F_{K,\sigma}^n - \nu_L \mathrm{m}(L) F_{L,\sigma}^n \\ = p_L^n - p_K^n \quad \forall \sigma = K|L \in \mathcal{E}_{\rm int}, \end{aligned} \tag{17}$$

$$F_{K,\sigma}^n + F_{L,\sigma}^n = 0 \quad \forall \sigma = K|L \in \mathcal{E}_{\rm int}\,, \tag{18}$$

$$\mathbf{U}_K^n = -A_K(c_K^{n-1})\mathbf{v}_K^n \quad \forall K \in \mathcal{M}\,, \tag{19}$$

$$\mathrm{m}(K)\mathbf{U}_K^n = -\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n(\mathbf{x}_\sigma - \mathbf{x}_K) \quad \forall K \in \mathcal{M}\,, \tag{20}$$

$$-\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n = \mathrm{m}(K)q_K^{+,n} - \mathrm{m}(K)q_K^{-,n} \quad \forall K \in \mathcal{M}\,, \tag{21}$$

$$\sum_{K\in\mathcal{M}} \mathrm{m}(K)p_K^n = 0\,, \tag{22}$$

$$F_{K,\sigma}^n = 0 \quad \forall K \in \mathcal{M}\,,\ \forall \sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}\,. \tag{23}$$

Denoting by $(-F_{K,\sigma}^n)^+$ and $(-F_{K,\sigma}^n)^-$ the positive and negative parts of $-F_{K,\sigma}^n$, the scheme for (9) reads as follows: for all $n = 1, \ldots, N_k$,

$$\begin{aligned}\mathbf{w}_K^n \cdot (\mathbf{x}_\sigma - \mathbf{x}_K) + \mathbf{w}_L^n \cdot (\mathbf{x}_L - \mathbf{x}_\sigma) + \nu_K \mathrm{m}(K) G_{K,\sigma}^n - \nu_L \mathrm{m}(L) G_{L,\sigma}^n \\ = c_L^n - c_K^n \quad \forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}\,,\end{aligned} \tag{24}$$

$$G_{K,\sigma}^n + G_{L,\sigma}^n = 0 \quad \forall \sigma = K|L \in \mathcal{E}_{\mathrm{int}}\,, \tag{25}$$

$$\mathrm{m}(K)D_K(\mathbf{U}_K^n)\mathbf{w}_K^n = \sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n(\mathbf{x}_\sigma - \mathbf{x}_K) \quad \forall K \in \mathcal{M}\,, \tag{26}$$

$$\begin{aligned}\mathrm{m}(K)\Phi_K \frac{c_K^n - c_K^{n-1}}{k} - \sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n + \sum_{\substack{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}} \\ \sigma = K|L}} \left[(-F_{K,\sigma}^n)^+ c_K^n - (-F_{K,\sigma}^n)^- c_L^n\right] \\ + \mathrm{m}(K)q_K^{-,n} c_K^n = \mathrm{m}(K)q_K^{+,n}\widehat{c}_K^n \quad \forall K \in \mathcal{M}\,,\end{aligned} \tag{27}$$

$$G_{K,\sigma}^n = 0 \quad \forall K \in \mathcal{M}\,,\ \forall \sigma \in \mathcal{E}_K \cap \mathcal{E}_{\mathrm{ext}}\,. \tag{28}$$

Let us explain why each equation of this scheme is quite natural.

- If we take $\nu_K = 0$, (17) and (24) state that $\mathbf{v}_K^n$ ($\approx \nabla\bar{p}$) and $\mathbf{w}_K^n$ ($\approx \nabla\bar{c}$) are "discrete gradients" of p_K^n ($\approx \bar{p}$) and c_K^n ($\approx \bar{c}$). The penalization using the fluxes (i.e., with $\nu_K > 0$) is added to ensure the stability of the scheme.
- Equations (18) and (25) state the conservation of the fluxes, and (23) and (28) translate the no flow boundary conditions.
- Equations (21) and (27) come from the integration on a control volume and on a time step of the PDEs in (8) and (9). Notice that, as usual, we have chosen a time-implicit scheme for the convection-diffusion equation with an upwind discretization of the convective term.
- Equations (19) and (22) are expressions of $\bar{\mathbf{U}} = -A(\cdot,\bar{c})\nabla\bar{p}$ and $\int_\Omega \bar{p}(t,\cdot) = 0$.
- Equations (20) and (26) come from the reconstruction formula given in Lemma 7.1, since $F_{K,\sigma}^n$ and $G_{K,\sigma}^n$ are approximations of the fluxes of $-\bar{\mathbf{U}}$ and $D(\cdot,\bar{\mathbf{U}})\nabla\bar{c}$.

In the following, if $a = (a_K^n)_{n=1,\dots,N_k\,,\,K\in\mathcal{M}}$ is a family of numbers (or vectors), we use a to denote the piecewise constant function on $[0,T)\times\Omega$ which is equal to a_K^n on $[(n-1)k, nk)\times K$. Similarly, for a fixed n, $a^n = (a_K^n)_{K\in\mathcal{M}}$ is identified with the function on Ω which takes the constant value a_K^n on the control volume K. Hence p denotes both the family $(p_K^n)_{n=1,\dots,N_k\,,\,K\in\mathcal{M}}$ and the corresponding function on $[0,T)\times\Omega$. We also denote by F and G the families $(F_{K,\sigma}^n)_{n=1,\dots,N_k\,,\,K\in\mathcal{M}\,,\,\sigma\in\mathcal{E}_K}$ and $(G_{K,\sigma}^n)_{n=1,\dots,N_k\,,\,K\in\mathcal{M}\,,\,\sigma\in\mathcal{E}_K}$.

THEOREM 2.1. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$ and let $T>0$. Assume* (10)–(15) *hold. Let $\mathcal{D}$ be an admissible mesh of Ω and $k>0$ such that T/k is an integer. Then there exists a unique solution $(p,\mathbf{v},\mathbf{U},F,c,\mathbf{w},G)$ to* (17)–(28).

THEOREM 2.2. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$ and let $T>0$. Assume* (10)–(15) *hold. Let $\nu_0>0$ and $\beta\in(2-2d, 4-2d)$. Let $(\mathcal{D}_m)_{m\geq 1}$ be a sequence of admissible meshes of Ω such that* $\mathrm{size}(\mathcal{D}_m)\to 0$ *as $m\to\infty$ and* $(\mathrm{regul}(\mathcal{D}_m))_{m\geq 1}$ *is bounded; assume that there exists C_1 such that, for all $m\geq 1$,*

$$\forall K,L\in\mathcal{M}_m \text{ neighboring control volumes, } \mathrm{diam}(K)^{2-\beta-d}\leq C_1\mathrm{diam}(L)^{d-2}. \tag{29}$$

For all $K\in\mathcal{M}_m$, we take $\nu_K=\nu_0\mathrm{diam}(K)^\beta$. Let $k_m>0$ be such that $N_{k_m}=T/k_m$ is an integer and $k_m\to 0$ as $m\to\infty$, and denote by $(p^m,\mathbf{v}^m,\mathbf{U}^m,F^m,c^m,\mathbf{w}^m,G^m)$ the solution to (17)–(28) *with $\mathcal{D}=\mathcal{D}_m$ and $k=k_m$. Then, up to a subsequence, as $m\to\infty$,*

$$\begin{aligned}
&p^m\to\bar p && \text{weakly-}* \text{ in } L^\infty(0,T;L^2(\Omega)) \text{ and strongly in } L^p(0,T;L^q(\Omega))\\
&&& \text{for all } p<\infty \text{ and all } q<2;\\
&\mathbf{v}^m\to\nabla\bar p && \text{weakly-}* \text{ in } L^\infty(0,T;L^2(\Omega))^d \text{ and strongly in } L^2((0,T)\times\Omega)^d;\\
&\mathbf{U}^m\to\bar{\mathbf{U}} && \text{weakly-}* \text{ in } L^\infty(0,T;L^2(\Omega))^d \text{ and strongly in } L^2((0,T)\times\Omega)^d;\\
&c^m\to\bar c && \text{weakly-}* \text{ in } L^\infty(0,T;L^2(\Omega)) \text{ and strongly in } L^p(0,T;L^q(\Omega))\\
&&& \text{for all } p<\infty \text{ and all } q<2;\\
&\mathbf{w}^m\to\nabla\bar c && \text{weakly in } L^2((0,T)\times\Omega)^d,
\end{aligned}$$

where $(\bar p,\bar{\mathbf{U}},\bar c)$ is a weak solution to (8)–(9).

Remark 2.2. As usual in finite volume schemes, we do not assume the existence of a solution to the continuous problem; this existence is obtained as a byproduct of the proof of convergence. In particular, this means that, contrary to [5] or [11], the convergence of the mixed finite volume scheme is proved here under no regularity assumption on the solution to (8)–(9). The convergence occurs only up to a subsequence because, with such a lack of regularity, the uniqueness of the solution is not known (see [15]); in the case where the solution is unique (for instance, under suitable regularity assumptions), then the whole sequence converges.

Remark 2.3. Note that, since $4-\beta-2d\geq 0$, one way to satisfy (29) is to ask that $\mathrm{diam}(K)\leq C_2\mathrm{diam}(L)$ for all neighboring control volumes K and L of a mesh. But (29) allows more freedom on the meshes (for example, if $d=1$ and $\beta\in(0,1]$ or if $d=2$ and $\beta\in(-2,0)$, then (29) is always satisfied).

3. The a priori estimates. We prove a priori estimates on the solution to the scheme.

PROPOSITION 3.1. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$ and let $T>0$. Assume* (10)–(11) *hold. Let $\mathcal{D}$ be an admissible mesh of Ω such that* $\mathrm{regul}(\mathcal{D})\leq\theta$ *for some $\theta>0$, and let $k>0$ be such that $N_k=T/k$ is an integer. Let*

$(\nu_K)_{K\in\mathcal{M}}$ be a family of positive numbers such that, for some $\nu_0 > 0$ and $\beta \geq 2-2d$, $\nu_K \leq \nu_0 \mathrm{diam}(K)^\beta$ for all $K \in \mathcal{M}$. Then there exists C_3 only depending on d, Ω, θ, β, ν_0, α_A, and Λ_A such that, for any numbers $(c_K^{n-1})_{n=1,\dots,N_k,\, K\in\mathcal{M}}$, any solution $(p, \mathbf{v}, \mathbf{U}, F)$ to (17)–(23) satisfies

$$\|p\|^2_{L^\infty(0,T;L^2(\Omega))} + \|\mathbf{v}\|^2_{L^\infty(0,T;L^2(\Omega))^d} + \|\mathbf{U}\|^2_{L^\infty(0,T;L^2(\Omega))^d} + \sup_{n=1,\dots,N_k} \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2 \leq C_3 \|q^+ - q^-\|^2_{L^\infty(0,T;L^2(\Omega))}.$$

Proof. Let $n \in [1, N_k]$. Multiply (21) by p_K^n, sum over all control volumes, and gather by edges using (18). Thanks to (23), the terms involving boundary edges disappear, and this leads to

$$\sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} F^n_{K,\sigma}(p^n_L - p^n_K) = \sum_{K\in\mathcal{M}} \mathrm{m}(K)(q_K^{+,n} - q_K^{-,n})p_K^n = \int_\Omega (q^{+,n} - q^{-,n})p^n,$$

where $q^{+,n}(\cdot) - q^{-,n}(\cdot) = \frac{1}{k}\int_{(n-1)k}^{nk} q^+(t,\cdot) - q^-(t,\cdot)\,dt$. Substituting (17) into this equality and gathering by control volumes (still using (18) and (23)), we deduce

$$\begin{aligned}\int_\Omega (q^{+,n} - q^{-,n})p^n &= \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} F^n_{K,\sigma}\left(\mathbf{v}_K^n \cdot (\mathbf{x}_\sigma - \mathbf{x}_K) + \mathbf{v}_L^n \cdot (\mathbf{x}_L - \mathbf{x}_\sigma)\right) \\ &\quad + \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} F^n_{K,\sigma}\left(\nu_K \mathrm{m}(K) F^n_{K,\sigma} - \nu_L \mathrm{m}(L) F^n_{L,\sigma}\right) \\ &= \sum_{K\in\mathcal{M}} \mathbf{v}_K^n \cdot \sum_{\sigma\in\mathcal{E}_K} F^n_{K,\sigma}(\mathbf{x}_\sigma - \mathbf{x}_K) + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2. \end{aligned} \tag{30}$$

Thanks to (20), (19), and hypothesis (11), we find

$$\|q^{+,n} - q^{-,n}\|_{L^2(\Omega)} \|p^n\|_{L^2(\Omega)} \geq \alpha_A \|\mathbf{v}^n\|^2_{L^2(\Omega)^d} + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2. \tag{31}$$

We notice that (17) is exactly (61) for $(p^n, \mathbf{v}^n, F^n)$. Hence, since p^n satisfies (22), we can apply the discrete Poincaré–Wirtinger inequality given in Lemma 7.2 to get

$$\|p^n\|_{L^2(\Omega)} \leq C_4 \left(\|\mathbf{v}^n\|_{L^2(\Omega)^d} + \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{diam}(K)^{2d-2} \nu_K^2 \mathrm{m}(K) |F^n_{K,\sigma}|^2 \right)^{\frac{1}{2}} \right),$$

where C_4 depends only on d, Ω, and θ. By choice of ν_K, we have $\mathrm{diam}(K)^{2d-2}\nu_K \leq \nu_0 \mathrm{diam}(K)^{2d-2+\beta}$; but $2d-2+\beta \geq 0$, and thus $\mathrm{diam}(K)^{2d-2}\nu_K \leq \nu_0 \mathrm{diam}(\Omega)^{2d-2+\beta}$. Hence

$$\|p^n\|_{L^2(\Omega)} \leq C_5 \left(\|\mathbf{v}^n\|_{L^2(\Omega)^d} + \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2 \right)^{\frac{1}{2}} \right), \tag{32}$$

where C_5 depends only on d, Ω, θ, β, and ν_0. Substituting this into (31), we obtain

$$\alpha_A \|\mathbf{v}^n\|^2_{L^2(\Omega)^d} + \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2 \leq C_5 \|q^{+,n} - q^{-,n}\|_{L^2(\Omega)} \|\mathbf{v}^n\|_{L^2(\Omega)^d} + C_5 \|q^{+,n} - q^{-,n}\|_{L^2(\Omega)} \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2 \right)^{\frac{1}{2}}.$$

Using Young's inequality, this gives the desired bound on $\mathbf{v}$ and F and, coming back to (32), the bound on p. The bound on $\mathbf{U}$ derives from the one on $\mathbf{v}$, since A is bounded (see (11)). ∎

PROPOSITION 3.2. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$ and let $T>0$. Assume (10) and (12)–(15) hold. Let $\mathcal{D}$ be an admissible mesh of Ω, and let $k>0$ be such that $N_k = T/k$ is an integer. Let $(\nu_K)_{K\in\mathcal{M}}$ be a family of positive numbers. Assume that $F = (F^n_{K,\sigma})_{n=1,\dots,N_k,\, K\in\mathcal{M},\, \sigma\in\mathcal{E}_K}$ satisfies (18), (21), and (23), and let $\mathbf{U} = (\mathbf{U}^n_K)_{n=1,\dots,N_k,\, K\in\mathcal{M}}$ be a family of vectors in $\mathbb{R}^d$. Then there exists C_6 depending only on d, Ω, T, α_D, and Φ_* such that any solution $(c,\mathbf{w},G)$ to (24)–(28) satisfies*

$$\|c\|^2_{L^\infty(0,T;L^2(\Omega))} + \|\mathbf{w}\|^2_{L^2((0,T)\times\Omega)^d} + \| \, |\mathbf{U}|^{1/2} \, |\mathbf{w}| \, \|^2_{L^2((0,T)\times\Omega)}$$
$$+ \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |G^n_{K,\sigma}|^2 \le C_6 \|c_0\|^2_{L^2(\Omega)} + C_6 \|q^+\|^2_{L^\infty(0,T;L^2(\Omega))}.$$

Proof. Multiply (27) by c^n_K and sum over all control volumes. Noting that $(c^n_K - c^{n-1}_K)c^n_K \ge \frac{1}{2}\left((c^n_K)^2 - (c^{n-1}_K)^2\right)$ and using (25) to gather by edges (no boundary term remains thanks to (28)), we obtain, since $\Phi_K \ge 0$,

$$\frac{1}{2k}\sum_{K\in\mathcal{M}} \mathrm{m}(K)\Phi_K\left((c^n_K)^2 - (c^{n-1}_K)^2\right) + \sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} G^n_{K,\sigma}(c^n_L - c^n_K) + \sum_{K\in\mathcal{M}} \mathrm{m}(K) q^{-,n}_K (c^n_K)^2$$
$$+ \sum_{K\in\mathcal{M}} \sum_{\substack{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\rm int}\\ \sigma=K|L}} \left[(-F^n_{K,\sigma})^+ c^n_K - (-F^n_{K,\sigma})^- c^n_L\right] c^n_K \le \sum_{K\in\mathcal{M}} \mathrm{m}(K) |q^{+,n}_K \widehat{c}^n_K| \, |c^n_K|. \tag{33}$$

Let us denote by $\mathcal{T}$ the fourth term of the inequality. Gathering by edges and using (18), which implies $(-F^n_{L,\sigma})^+ = (-F^n_{K,\sigma})^-$ and $(-F^n_{L,\sigma})^- = (-F^n_{K,\sigma})^+$, yields

$$\mathcal{T} = \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \left[(-F^n_{K,\sigma})^+ (c^n_K(c^n_K - c^n_L)) + (-F^n_{K,\sigma})^- (c^n_L(c^n_L - c^n_K))\right].$$

But $c^n_K(c^n_K - c^n_L) \ge \frac{1}{2}\left((c^n_K)^2 - (c^n_L)^2\right)$ and $c^n_L(c^n_L - c^n_K) \ge \frac{1}{2}\left((c^n_L)^2 - (c^n_K)^2\right)$, hence

$$\mathcal{T} \ge \frac{1}{2} \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} \left[(-F^n_{K,\sigma})^+ - (-F^n_{K,\sigma})^-\right]\left((c^n_K)^2 - (c^n_L)^2\right)$$
$$\ge \frac{1}{2} \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} -F^n_{K,\sigma}\left((c^n_K)^2 - (c^n_L)^2\right),$$

which gives, gathering by control volumes and using (18), (23), and (21),

$$\mathcal{T} \ge \frac{1}{2}\sum_{K\in\mathcal{M}} (c^n_K)^2 \left(-\sum_{\sigma\in\mathcal{E}_K} F^n_{K,\sigma}\right) \ge \frac{1}{2}\sum_{K\in\mathcal{M}} \mathrm{m}(K)(c^n_K)^2(q^{+,n}_K - q^{-,n}_K).$$

Since

$$\frac{1}{2}\sum_{K\in\mathcal{M}} \mathrm{m}(K)(c^n_K)^2(q^{+,n}_K - q^{-,n}_K) + \sum_{K\in\mathcal{M}} \mathrm{m}(K) q^{-,n}_K (c^n_K)^2$$
$$= \frac{1}{2}\sum_{K\in\mathcal{M}} \mathrm{m}(K)(q^{+,n}_K + q^{-,n}_K)(c^n_K)^2 \ge 0$$

(because q^+ and q^- are nonnegative), we deduce from (33) that

$$\frac{1}{2k}\sum_{K\in\mathcal{M}} \mathrm{m}(K)\Phi_K\left((c_K^n)^2-(c_K^{n-1})^2\right)+\sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} G_{K,\sigma}^n(c_L^n-c_K^n)$$

$$\leq \sum_{K\in\mathcal{M}} \mathrm{m}(K)|q_K^{+,n}\widehat{c}_K^n|\,|c_K^n|. \tag{34}$$

Using (24) and gathering by control volumes, we get, thanks to (25), (28), and (26),

$$\sum_{\substack{\sigma\in\mathcal{E}_{\rm int}\\ \sigma=K|L}} G_{K,\sigma}^n(c_L^n-c_K^n)=\sum_{K\in\mathcal{M}}\mathbf{w}_K^n\cdot\sum_{\sigma\in\mathcal{E}_K}G_{K,\sigma}^n(\mathbf{x}_\sigma-\mathbf{x}_K)+\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)|G_{K,\sigma}^n|^2$$

$$=\sum_{K\in\mathcal{M}}\mathrm{m}(K)D_K(\mathbf{U}_K^n)\mathbf{w}_K^n\cdot\mathbf{w}_K^n+\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)|G_{K,\sigma}^n|^2. \tag{35}$$

We then use (12) and plug the corresponding lower bound into (34), which we multiply by k and sum over $n=1,\dots,N$ (for some $N\in[1,N_k]$); since $|\widehat{c}|\leq 1$, this leads to

$$\frac12\sum_{K\in\mathcal{M}}\mathrm{m}(K)\Phi_K\left((c_K^N)^2-(c_K^0)^2\right)+\alpha_D\sum_{n=1}^N k\sum_{K\in\mathcal{M}}\mathrm{m}(K)(1+|\mathbf{U}_K^n|)|\mathbf{w}_K^n|^2$$

$$+\sum_{n=1}^N k\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)|G_{K,\sigma}^n|^2\leq T\|q^+\|_{L^\infty(0,T;L^2(\Omega))}\|c\|_{L^\infty(0,T;L^2(\Omega))}. \tag{36}$$

This gives in particular, by (13) and the definition of $(c_K^0)_{K\in\mathcal{M}}$,

$$\frac{\Phi_*}{2}\sum_{K\in\mathcal{M}}\mathrm{m}(K)(c_K^N)^2\leq\frac{\Phi_*^{-1}}{2}\|c_0\|_{L^2(\Omega)}^2+\frac{T^2}{\Phi_*}\|q^+\|_{L^\infty(0,T;L^2(\Omega))}^2+\frac{\Phi_*}{4}\|c\|_{L^\infty(0,T;L^2(\Omega))}^2.$$

Since $\|c\|_{L^\infty(0,T;L^2(\Omega))}^2=\sup_{r=1,\dots,N_k}\sum_{K\in\mathcal{M}}\mathrm{m}(K)(c_K^r)^2$, this inequality, valid for all $1\leq N\leq N_k$, gives the estimate on $\|c\|_{L^\infty(0,T;L^2(\Omega))}$. Plugged into (36), it gives the desired bounds on $\mathbf{w}$, $|\mathbf{U}|^{1/2}\,|\mathbf{w}|$ and G. ∎

4. Existence and uniqueness of numerical solutions. In this section, we prove Theorem 2.1. Note first that (17)–(23) and (24)–(28) are decoupled systems: at time step n, the knowledge of c_K^{n-1} (or of c_K^0 if $n=1$) shows that (17)–(23) is a linear system for $(p^n,\mathbf{v}^n,\mathbf{U}^n,(F_{K,\sigma}^n)_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$; once this system is solved, $\mathbf{U}^n$ is known and (24)–(28) becomes a linear system for $(c^n,\mathbf{w}^n,(G_{K,\sigma}^n)_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$. Hence, to prove Theorem 2.1 we only need to show that these linear systems are solvable.

Let us first consider the system on $(c^n,\mathbf{w}^n,(G_{K,\sigma}^n)_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$. By (25) and (28), we can consider that there is only one flux by interior edge and this system therefore has $(d+1)\mathrm{Card}(\mathcal{M})+\mathrm{Card}(\mathcal{E}_{\rm int})$ unknowns, with as many remaining equations ((26) gives $d\mathrm{Card}(\mathcal{M})$ equations, (27) another $\mathrm{Card}(\mathcal{M})$ equations, and (24) the last $\mathrm{Card}(\mathcal{E}_{\rm int})$ equations). Hence, this first system is a square system. Assume that $(c^n,\mathbf{w}^n,(G_{K,\sigma}^n)_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$ is a solution with a null right-hand side, i.e., with $c^{n-1}=\widehat{c}^n=0$; then (34) and (35) show that this solution is null, and therefore that this system is invertible.

Without the relation (22) and since we can eliminate $\mathbf{U}^n$ by (19), the system on $(p^n,\mathbf{v}^n,\mathbf{U}^n,(F_{K,\sigma}^n)_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$ also has $(d+1)\mathrm{Card}(\mathcal{M})+\mathrm{Card}(\mathcal{E}_{\rm int})$ unknowns

and the same number of equations. However, it is not invertible since its kernel clearly contains $(\mathcal{C},0,0,0)$, where $\mathcal{C}\in\mathbb{R}^{\mathrm{Card}(\mathcal{M})}$ is any constant vector; in fact, the estimates in the preceding section show that these vectors fully describe the kernel of ((17)–(21), (23)): if $(p^n,\mathbf{v}^n,\mathbf{U}^n,(F^n_{K,\sigma})_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$ belongs to this kernel, then $(p^n-\mathcal{C},\mathbf{v}^n,\mathbf{U}^n,(F^n_{K,\sigma})_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})$, where $\mathcal{C}$ is a constant vector such that (22) holds with $p^n-\mathcal{C}$, satisfies (17)–(23) with $q^{+,n}-q^{-,n}=0$, and is therefore null by Proposition 3.1, which shows that $(p^n,\mathbf{v}^n,\mathbf{U}^n,(F^n_{K,\sigma})_{K\in\mathcal{M},\sigma\in\mathcal{E}_K})=(\mathcal{C},0,0,0)$.

Summing (21) over K and using (18) and (23), we obtain that a necessary condition for ((17)–(21), (23)) to have a solution is $\sum_{K\in\mathcal{M}}\mathrm{m}(K)q_K^{n,+}-\mathrm{m}(K)q_K^{n,-}=0$. Since the kernel of the square system ((17)–(21), (23)) has dimension 1, this condition is also sufficient, and is clearly satisfied by the data we consider thanks to (10). We can therefore always find a solution to ((17)–(21), (23)) and, in view of the kernel of this system, (22) then selects one and only one solution.

Remark 4.1. As said above, at each time step the scheme (17)–(28) can be decoupled in two successive linear systems, (17)–(23) and then (24)–(28), each one with size $(d+1)\mathrm{Card}(\mathcal{M})+\mathrm{Card}(\mathcal{E}_{\mathrm{int}})$. However, it is possible to proceed to an algebraic elimination which leads to smaller sparse linear systems, following [18] for the mixed finite element method and [9] for the mixed finite volume method for anisotropic diffusion problems.

The computation of $(p,\mathbf{v},\mathbf{U},F)$ at each time step reduces to the resolution of a linear system of size $\mathrm{Card}(\mathcal{E}_{\mathrm{int}})$, while the computation of $(c,\mathbf{w},G)$ demands the resolution of a linear system of size $\mathrm{Card}(\mathcal{M})+\mathrm{Card}(\mathcal{E}_{\mathrm{int}})$ (the size of this last system cannot be reduced to $\mathrm{Card}(\mathcal{E}_{\mathrm{int}})$ because of the upwind and implicit discretization of the convective term $\mathrm{div}(c\mathbf{U})$).

5. Proof of the convergence of the scheme. In this section, we prove Theorem 2.2. To simplify the notation, we drop the index m and thus prove the desired convergence as $\mathrm{size}(\mathcal{D})\to 0$ and $k\to 0$, with $\mathrm{regul}(\mathcal{D})$ bounded and (29) uniformly satisfied for all considered meshes. Under these assumptions, Propositions 3.1 and 3.2 give estimates which are uniform with respect to the meshes and time steps.

5.1. Compactness of the concentration. We prove the strong compactness of the concentration.

LEMMA 5.1. *Under the assumptions of Theorem* 2.2, c *is relatively compact in* $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$.

Proof. We first construct an affine interpolant $\widetilde{c}$ of c and prove, thanks to Aubin's theorem, the relative compactness of this interpolant in a weaker space. We then deduce the compactness of c in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$. ∎

Step 1. An affine interpolant of c.

We define $\widetilde{c}:[0,T)\times\Omega\to\mathbb{R}$ as, for all $n=1,\dots,N_k$ and all $t\in[(n-1)k,nk)$,

$$\widetilde{c}(t,\cdot)=\frac{t-(n-1)k}{k}c_K^n+\frac{nk-t}{k}c_K^{n-1}\quad\text{on } K.$$

The estimates of Proposition 3.2 and the definition of $(c_K^0)_{K\in\mathcal{M}}$ ensure the bound of $\|\widetilde{c}\|_{L^\infty(0,T;L^2(\Omega))}$. For all $n=1,\dots,N_k$ and all $t\in[(n-1)k,nk)$, we have $\partial_t\widetilde{c}(t,\cdot)=\frac{c_K^n-c_K^{n-1}}{k}$ on K. Hence, denoting by $\Phi_{\mathcal{D}}$ the piecewise constant function on Ω equal to Φ_K on K and taking $\varphi\in C_c^2(\Omega)$, we deduce from (27) that if φ_K is the mean value

of φ on K,

$$\begin{aligned}\int_\Omega \Phi_{\mathcal{D}}(x)\partial_t\widetilde{c}(t,x)\varphi(x)\,dx &= \sum_{K\in\mathcal{M}} \mathrm{m}(K)\Phi_K \frac{c_K^n - c_K^{n-1}}{k}\varphi_K \\ &= \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n\varphi_K - \sum_{K\in\mathcal{M}} \mathrm{m}(K) q_K^{-,n} c_K^n \varphi_K + \sum_{K\in\mathcal{M}} \mathrm{m}(K) q_K^{+,n}\widehat{c}_K^n\varphi_K \\ &\quad - \sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} \left[(-F_{K,\sigma}^n)^+ c_K^n - (-F_{K,\sigma}^n)^- c_L^n\right]\varphi_K. \end{aligned} \tag{37}$$

Let us denote by T_1, T_3, T_4, and T_2 the four terms on the right-hand side of this equality. In the following, C_i denote constants which do not depend on k, $\mathcal{D}$, n, K, or φ; we induce $C_c^2(\Omega)$ with the norm $\|\varphi\| = \sup_{x\in\Omega}\left(|\varphi(x)| + |\nabla\varphi(x)| + |D^2\varphi(x)|\right)$.

Since $\mathbf{x}_K$ is the barycenter of K and φ is regular we have $\varphi(\mathbf{x}_\sigma) - \varphi_K = \nabla\varphi(\mathbf{x}_K)\cdot(\mathbf{x}_\sigma - \mathbf{x}_K) + R_{K,\sigma}$ for all $\sigma\in\mathcal{E}_K$, with $|R_{K,\sigma}| \le C_7\|\varphi\|\mathrm{diam}(K)^2$. Hence,

$$\varphi_L - \varphi_K = \nabla\varphi(\mathbf{x}_K)\cdot(\mathbf{x}_\sigma - \mathbf{x}_K) + \nabla\varphi(\mathbf{x}_L)\cdot(\mathbf{x}_L - \mathbf{x}_\sigma) + R_{K,\sigma} - R_{L,\sigma}. \tag{38}$$

Using this equality and gathering by control volumes, we get

$$\begin{aligned} -T_1 &= \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} G_{K,\sigma}^n(\varphi_L - \varphi_K) \\ &= \sum_{K\in\mathcal{M}} \nabla\varphi(\mathbf{x}_K)\cdot\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n(\mathbf{x}_\sigma - \mathbf{x}_K) + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n R_{K,\sigma} \\ &= \sum_{K\in\mathcal{M}} \mathrm{m}(K)\nabla\varphi(\mathbf{x}_K)\cdot D_K(\mathbf{U}_K^n)\mathbf{w}_K^n + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n R_{K,\sigma}. \end{aligned} \tag{39}$$

On one hand, thanks to (12) and to the estimate on $\mathbf{U}$ in $L^\infty(0,T;L^2(\Omega))^d$ (which gives in particular an estimate in $L^\infty(0,T;L^1(\Omega))^d$), we have

$$\begin{aligned}\left|\sum_{K\in\mathcal{M}} \mathrm{m}(K)\nabla\varphi(\mathbf{x}_K)\cdot D_K(\mathbf{U}_K^n)\mathbf{w}_K^n\right| &\le C_8\|\varphi\|\sum_{K\in\mathcal{M}} \mathrm{m}(K)(1+|\mathbf{U}_K^n|)\,|\mathbf{w}_K^n| \\ &\le C_9\|\varphi\|\left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(1+|\mathbf{U}_K^n|)|\mathbf{w}_K^n|^2\right)^{\frac12}. \end{aligned} \tag{40}$$

On the other hand, using $|R_{K,\sigma}| \le C_7\|\varphi\|\mathrm{diam}(K)^2$ and the Cauchy–Schwarz inequality, we get

$$\begin{aligned}&\left|\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G_{K,\sigma}^n R_{K,\sigma}\right| \\ &\quad\le C_7\|\varphi\|\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K\mathrm{m}(K)|G_{K,\sigma}^n|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{diam}(K)^4}{\nu_K\mathrm{m}(K)}\right)^{\frac12} \\ &\quad\le C_{10}\|\varphi\|\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K\mathrm{m}(K)|G_{K,\sigma}^n|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}}\mathrm{diam}(K)^{4-2d-\beta}\mathrm{m}(K)\right)^{\frac12} \end{aligned} \tag{41}$$

because (16) and the definition of ν_K imply

$$\frac{\mathrm{diam}(K)^4}{\nu_K \mathrm{m}(K)} = \mathrm{m}(K)\frac{\mathrm{diam}(K)^{4-\beta}}{\nu_0 \mathrm{m}(K)^2} \le \frac{1}{\nu_0}\left(\frac{\mathrm{regul}(\mathcal{D})}{\omega_d}\right)^2 \mathrm{m}(K)\mathrm{diam}(K)^{4-2d-\beta}. \tag{42}$$

But $4-2d-\beta \ge 0$ and thus $\mathrm{diam}(K)^{4-2d-\beta} \le \mathrm{diam}(\Omega)^{4-2d-\beta}$. Using this in (41) and substituting the result along with (40) into (39), we deduce the final estimate:

$$\begin{aligned} |T_1| &\le C_{11}\|\varphi\| \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(1+|\mathbf{U}_K^n|)|\mathbf{w}_K^n|^2\right)^{\frac12} \\ &\quad + C_{11}\|\varphi\| \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|G_{K,\sigma}^n|^2\right)^{\frac12}. \end{aligned} \tag{43}$$

For $\sigma = K|L$, set $b_{K,\sigma}^n = (-F_{K,\sigma}^n)^+ c_K^n - (-F_{K,\sigma}^n)^- c_L^n$. By (18), we have $b_{K,\sigma}^n = -b_{L,\sigma}^n$. Hence, using (38) and gathering by control volumes, we get

$$\begin{aligned} T_2 &= \sum_{\sigma=K|L\in\mathcal{E}_{\mathrm{int}}} b_{K,\sigma}^n(\varphi_L - \varphi_K) \\ &= \sum_{K\in\mathcal{M}} \nabla\varphi(\mathbf{x}_K)\cdot \sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} b_{K,\sigma}^n(\mathbf{x}_\sigma - \mathbf{x}_K) + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} b_{K,\sigma}^n R_{K,\sigma}. \end{aligned}$$

But $b_{K,\sigma}^n = -F_{K,\sigma}^n c_K^n + (-F_{K,\sigma}^n)^-(c_K^n - c_L^n)$ and thus, by (23) and (20),

$$T_2 = -\sum_{K\in\mathcal{M}} c_K^n \nabla\varphi(\mathbf{x}_K)\cdot \sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n(\mathbf{x}_\sigma - \mathbf{x}_K) + T_5 = \sum_{K\in\mathcal{M}} \mathrm{m}(K) c_K^n \nabla\varphi(\mathbf{x}_K)\cdot\mathbf{U}_K^n + T_5$$

with

$$T_5 = \sum_{K\in\mathcal{M}} \nabla\varphi(\mathbf{x}_K)\cdot \sum_{\substack{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}\\ \sigma=K|L}} (-F_{K,\sigma}^n)^-(c_K^n - c_L^n)(\mathbf{x}_\sigma - \mathbf{x}_K) + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} b_{K,\sigma}^n R_{K,\sigma}.$$

Let us estimate T_5. The corresponding calculations will be useful later in the proof of the convergence of the concentration. We have

$$\begin{aligned} |T_5| &\le \|\varphi\| \sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} |F_{K,\sigma}^n|\,|c_K^n - c_L^n|\mathrm{diam}(K) \\ &\quad + C_7\|\varphi\| \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} |b_{K,\sigma}^n|\mathrm{diam}(K)^2. \end{aligned} \tag{44}$$

But (24) entails

$$|c_K^n - c_L^n| \le |\mathbf{w}_K^n|\mathrm{diam}(K) + |\mathbf{w}_L^n|\mathrm{diam}(L) + \nu_K \mathrm{m}(K)|G_{K,\sigma}^n| + \nu_L \mathrm{m}(L)|G_{L,\sigma}^n|$$

and thus, using $|F_{K,\sigma}^n| = |F_{L,\sigma}^n|$ whenever $\sigma = K|L$,

$$\begin{aligned} |b_{K,\sigma}^n| &\le |F_{K,\sigma}^n|\,|c_K^n| + |F_{K,\sigma}^n|\,|\mathbf{w}_K^n|\mathrm{diam}(K) + |F_{K,\sigma}^n|\,|\mathbf{w}_L^n|\mathrm{diam}(L) \\ &\quad + \nu_K \mathrm{m}(K)|F_{K,\sigma}^n|\,|G_{K,\sigma}^n| + \nu_L \mathrm{m}(L)|F_{L,\sigma}^n|\,|G_{L,\sigma}^n|. \end{aligned}$$

Substituting these two estimates into (44) and bounding $\mathrm{diam}(K)$ either by $\mathrm{diam}(\Omega)$ or $\mathrm{size}(\mathcal{D})$, we get

$$
\begin{aligned}
|T_5| \le{}& C_{12}\|\varphi\| \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} |F^n_{K,\sigma}|\ (|\mathbf{w}^n_K| + |c^n_K|)\,\mathrm{diam}(K)^2 \\
&+ C_{12}\|\varphi\| \sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} |F^n_{K,\sigma}|\,|\mathbf{w}^n_L|\mathrm{diam}(K)\mathrm{diam}(L) \\
&+ C_{12}\|\varphi\|\mathrm{size}(\mathcal{D}) \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|\,|G^n_{K,\sigma}| \\
&+ C_{12}\|\varphi\|\mathrm{size}(\mathcal{D}) \sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} \nu_L \mathrm{m}(L)|F^n_{L,\sigma}|\,|G^n_{L,\sigma}|. \qquad (45)
\end{aligned}
$$

We successively apply the Cauchy–Schwarz inequality, the fact that $\mathrm{regul}(\mathcal{D})$ is bounded, inequality (42), and the estimates on F from Proposition 3.1. This yields

$$
\begin{aligned}
&\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} |F^n_{K,\sigma}|\ (|\mathbf{w}^n_K| + |c^n_K|)\,\mathrm{diam}(K)^2 \\
&\le C_{13}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)\mathrm{diam}(K)^{4-2d-\beta}(|\mathbf{w}^n_K| + |c^n_K|)^2\right)^{\frac12} \\
&\le C_{14}\mathrm{size}(\mathcal{D})^{\frac{4-2d-\beta}{2}}\left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(|\mathbf{w}^n_K| + |c^n_K|)^2\right)^{\frac12}. \qquad (46)
\end{aligned}
$$

Then we note that (thanks to (23))

$$
\sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} \nu_L \mathrm{m}(L)|F^n_{L,\sigma}|\,|G^n_{L,\sigma}| = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|\,|G^n_{K,\sigma}| \qquad (47)
$$

and, with the estimates on F from Proposition 3.1, we get

$$
\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|\,|G^n_{K,\sigma}| \le C_{15}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|G^n_{K,\sigma}|^2\right)^{\frac12}. \qquad (48)
$$

Using the fact that $\nu_K = \nu_0\mathrm{diam}(K)^\beta$ and inequalities (16) and (29),we get

$$
\begin{aligned}
&\sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} |F^n_{K,\sigma}|\,|\mathbf{w}^n_L|\mathrm{diam}(K)\mathrm{diam}(L) \\
&\le \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}}\sum_{\substack{\sigma\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}\\ \sigma=K|L}} \frac{\mathrm{diam}(K)^2\mathrm{diam}(L)^2}{\nu_K \mathrm{m}(K)}|\mathbf{w}^n_L|^2\right)^{\frac12} \\
&\le C_{16}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\left(\sum_{L\in\mathcal{M}}|\mathbf{w}^n_L|^2\sum_{\substack{\sigma\in\mathcal{E}_L\cap\mathcal{E}_{\mathrm{int}}\\ \sigma=L|K}}\mathrm{diam}(K)^{2-\beta-d}\mathrm{diam}(L)^2\right)^{\frac12} \\
&\le C_{17}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\left(\sum_{L\in\mathcal{M}} \mathrm{m}(L)|\mathbf{w}^n_L|^2\right)^{\frac12}.
\end{aligned}
$$

Finally, gathering (45), (46), (47), (48), and this last inequality, it yields

$$
\begin{aligned}
|T_5| &= \left| T_2 - \sum_{K\in\mathcal{M}} \mathrm{m}(K) c_K^n \nabla\varphi(\mathbf{x}_K)\cdot \mathbf{U}_K^n \right| \\
&\le C_{18}\|\varphi\|\mathrm{size}(\mathcal{D})^{\frac{4-2d-\beta}{2}} \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(|\mathbf{w}_K^n| + |c_K^n|)^2 \right)^{\frac12} \\
&\quad + C_{18}\|\varphi\| \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F_{K,\sigma}^n|^2 \right)^{\frac12} \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)|\mathbf{w}_K^n|^2 \right)^{\frac12} \\
&\quad + C_{18}\|\varphi\|\mathrm{size}(\mathcal{D}) \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |G_{K,\sigma}^n|^2 \right)^{\frac12}. \qquad (49)
\end{aligned}
$$

Thanks to the $L^\infty(0,T;L^2(\Omega))$ estimates on c and $\mathbf{U}$, we also have

$$
\begin{aligned}
\left| \sum_{K\in\mathcal{M}} \mathrm{m}(K) c_K^n \nabla\varphi(\mathbf{x}_K)\cdot \mathbf{U}_K^n \right| &\le \|\varphi\| \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)|c_K^n|^2 \right)^{\frac12} \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)|\mathbf{U}_K^n|^2 \right)^{\frac12} \\
&\le C_{19}\|\varphi\|,
\end{aligned}
$$

and, using the bound on the fluxes $F_{K,\sigma}^n$ from Proposition 3.1, the final estimate on T_2 reads

$$
\begin{aligned}
|T_2| &\le C_{20}\|\varphi\| + C_{20}\|\varphi\| \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(|\mathbf{w}_K^n| + |c_K^n|)^2 \right)^{\frac12} \\
&\quad + C_{20}\|\varphi\| \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |G_{K,\sigma}^n|^2 \right)^{\frac12}. \qquad (50)
\end{aligned}
$$

The estimates on T_3 and T_4 are straightforward, thanks to the $L^\infty(0,T;L^2(\Omega))$-bound on c; plugging (43) and (50) into (37), we obtain, for all $n = 1,\dots,N_k$ and all $t\in[(n-1)k, nk)$,

$$
\begin{aligned}
&\left| \int_\Omega \Phi_{\mathcal{D}}(x)\partial_t\widetilde{c}(t,x)\varphi(x)\,dx \right| \\
&\le C_{21}\|\varphi\| \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(1+|\mathbf{U}_K^n|)|\mathbf{w}_K^n|^2 \right)^{\frac12} + C_{21}\|\varphi\| \left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |G_{K,\sigma}^n|^2 \right)^{\frac12} \\
&\quad + C_{21}\|\varphi\| + C_{21}\|\varphi\| \left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)(|\mathbf{w}_K^n| + |c_K^n|)^2 \right)^{\frac12}.
\end{aligned}
$$

Since this inequality is satisfied for all $\varphi\in C_c^2(\Omega)$ and $\Phi_{\mathcal{D}}$ does not depend on t, this gives an estimate on $\|\partial_t(\Phi_{\mathcal{D}}\widetilde{c})(t,\cdot)\|_{(C_c^2(\Omega))'}$ which, squared, leads to

$$
\begin{aligned}
\|\partial_t(\Phi_{\mathcal{D}}\widetilde{c})(t,\cdot)\|_{(C_c^2(\Omega))'}^2 &\le C_{22}\sum_{K\in\mathcal{M}} \mathrm{m}(K)(1+|\mathbf{U}_K^n|)|\mathbf{w}_K^n|^2 + C_{22} \\
&\quad + C_{22}\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|G_{K,\sigma}^n|^2 + C_{22}\sum_{K\in\mathcal{M}} \mathrm{m}(K)(|\mathbf{w}_K^n| + |c_K^n|)^2
\end{aligned}
$$

for all $n = 1, \dots, N_k$ and all $t \in [(n-1)k, nk)$. Integrating this last inequality on $t \in [(n-1)k, nk)$ and summing over $n = 1, \dots, N_k$, we prove, thanks to the estimates of Proposition 3.2, that $\partial_t(\Phi_{\mathcal{D}}\widetilde{c})$ is bounded in $L^2(0,T;(C_c^2(\Omega))')$.

Noting that $\Phi_{\mathcal{D}}\widetilde{c}$ is bounded in $L^\infty(0,T;L^2(\Omega))$ (because $\widetilde{c}$ is bounded in this space and $\Phi_{\mathcal{D}}$ is bounded in $L^\infty(\Omega)$), and since $L^2(\Omega)$ is continuously embedded in $(C_c^2(\Omega))'$ (via the natural embedding $f \to (\varphi \to \int_\Omega f\varphi)$), this shows that $\Phi_{\mathcal{D}}\widetilde{c}$ is bounded in $H^1(0,T;(C_c^2(\Omega))')$. But $C_c^2(\Omega)$ is compactly and densely embedded in $C_0(\Omega)$, and, by duality, $(C_0(\Omega))'$ (the space of bounded measures on Ω) is compactly embedded in $(C_c^2(\Omega))'$. Since $L^2(\Omega)$ is continuously embedded in $(C_0(\Omega))'$ (via an embedding which is compatible with the preceding one), the embedding of $L^2(\Omega)$ in $(C_c^2(\Omega))'$ is in fact compact. Hence, by Aubin's compactness theorem we deduce that $\Phi_{\mathcal{D}}\widetilde{c}$ is relatively compact in $C([0,T];(C_c^2(\Omega))')$.

Step 2. Conclusion.

For all $n = 1, \dots, N_k$ and $t \in [(n-1)k, nk)$, we have $\Phi_{\mathcal{D}}c(t,\cdot) = \Phi_{\mathcal{D}}\widetilde{c}(nk,\cdot)$ on Ω (these functions are both equal to $\Phi_K c_K^n$ on each $K \in \mathcal{M}$). We also know (see, e.g., [7]) that $H^1(0,T;(C_c^2(\Omega))')$ is continuously embedded in $C^{1/2}([0,T];(C_c^2(\Omega))')$ (the space of 1/2-Hölder continuous functions $[0,T] \to (C_c^2(\Omega))'$). Hence, $\Phi_{\mathcal{D}}\widetilde{c}$ is also bounded in $C^{1/2}([0,T];(C_c^2(\Omega))')$ and there exists C_{23} not depending on k or $\mathcal{D}$ such that, for all $n = 1, \dots, N_k$ and all $t \in [(n-1)k, nk)$,

$$\|\Phi_{\mathcal{D}}c(t,\cdot) - \Phi_{\mathcal{D}}\widetilde{c}(t,\cdot)\|_{(C_c^2(\Omega))'} = \|\Phi_{\mathcal{D}}\widetilde{c}(nk,\cdot) - \Phi_{\mathcal{D}}\widetilde{c}(t,\cdot)\|_{(C_c^2(\Omega))'} \le C_{23}\sqrt{k}.$$

This means that, as $k \to 0$, $\Phi_{\mathcal{D}}c - \Phi_{\mathcal{D}}\widetilde{c} \to 0$ in $L^\infty(0,T;(C_c^2(\Omega))')$; since $\Phi_{\mathcal{D}}\widetilde{c}$ is relatively compact in this space, we deduce that $\Phi_{\mathcal{D}}c$ is also relatively compact in this same space, and thus in particular in $L^1(0,T;(C_c^2(\Omega))')$.

Let $n = 1, \dots, N_k$ and $t \in [(n-1)k, nk)$. By (24), Lemma 7.3 gives, for all ω relatively compact in Ω and all $|\xi| < \operatorname{dist}(\omega, \mathbb{R}^d \backslash \Omega)$,

$$\begin{aligned}\|c(t,\cdot+\xi) - c(t,\cdot)\|_{L^1(\omega)} &\le C_{24}|\xi| \sum_{K\in\mathcal{M}} \mathrm{m}(K)|\mathbf{w}_K^n| \\ &\quad + C_{24}|\xi| \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \operatorname{diam}(K)^{d-1}\nu_K \mathrm{m}(K)|G_{K,\sigma}^n|.\end{aligned}$$

Integrating on $t \in [(n-1)k, nk)$ and summing over $n = 1, \dots, N_k$, this implies that

$$\begin{aligned}&\|c(\cdot,\cdot+\xi) - c\|_{L^1((0,T)\times\omega)} \\ &\le C_{24}|\xi|\,\|\mathbf{w}\|_{L^1((0,T)\times\Omega)^d} + C_{24}|\xi| \left(\sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \operatorname{diam}(K)^{2d-2}\nu_K \mathrm{m}(K)\right)^{\frac12} \\ &\qquad\qquad \times \left(\sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K)|G_{K,\sigma}^n|^2\right)^{\frac12} \\ &\le C_{25}|\xi| + C_{25}|\xi| \left(\sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \operatorname{diam}(K)^{2d-2+\beta}\mathrm{m}(K)\right)^{\frac12}\end{aligned}$$

thanks to the estimates of Proposition 3.2. But $2d-2+\beta \ge 0$ and $\operatorname{diam}(K)^{2d-2+\beta} \le \operatorname{diam}(\Omega)^{2d-2+\beta}$. Hence, we see that $\|c(\cdot,\cdot+\xi) - c\|_{L^1((0,T)\times\omega)} \to 0$ as $\xi \to 0$, independent of k or $\mathcal{D}$.

Since $\Phi_{\mathcal{D}}$ is bounded in $L^\infty(\Omega)$ and c is bounded in $L^\infty(0,T;L^2(\Omega))$, we have

$$\begin{aligned}
&\|(\Phi_{\mathcal{D}}c)(\cdot,\cdot+\xi)-\Phi_{\mathcal{D}}c\|_{L^1((0,T)\times\omega)}\\
&\qquad=\|\Phi_{\mathcal{D}}(\cdot+\xi)(c(\cdot,\cdot+\xi)-c)+(\Phi_{\mathcal{D}}(\cdot+\xi)-\Phi_{\mathcal{D}})c\|_{L^1((0,T)\times\omega)}\\
&\qquad\le C_{26}\|c(\cdot,\cdot+\xi)-c\|_{L^1((0,T)\times\omega)}+C_{27}\|\Phi_{\mathcal{D}}(\cdot+\xi)-\Phi_{\mathcal{D}}\|_{L^2(\omega)},
\end{aligned}$$

where C_{26} and C_{27} do not depend on $\mathcal{D}$ or k. But it is classical that $\Phi_{\mathcal{D}}\to\Phi$ in $L^2(\Omega)$ as $\mathrm{size}(\mathcal{D})\to 0$ and thus $\|\Phi_{\mathcal{D}}(\cdot+\xi)-\Phi_{\mathcal{D}}\|_{L^2(\omega)}\to 0$ as $\xi\to 0$, independent of $\mathcal{D}$. We therefore obtain $\|(\Phi_{\mathcal{D}}c)(\cdot,\cdot+\xi)-\Phi_{\mathcal{D}}c\|_{L^1((0,T)\times\omega)}\to 0$ as $\xi\to 0$, independent of k or $\mathcal{D}$. Since $\Phi_{\mathcal{D}}c$ is relatively compact in $L^1(0,T;(C_c^2(\Omega))')$, Lemma 7.5 then shows that $\Phi_{\mathcal{D}}c$ is relatively compact in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$.

Up to a subsequence as $k\to 0$ and $\mathrm{size}(\mathcal{D})\to 0$, $\Phi_{\mathcal{D}}c\to f$ in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$. Using again the fact that $\Phi_{\mathcal{D}}\to\Phi$ in $L^2(\Omega)$ we also have, up to another subsequence, $\Phi_{\mathcal{D}}\to\Phi$ a.e. on Ω; moreover, $\Phi_{\mathcal{D}}\ge\Phi_*>0$ and thus $\frac{1}{\Phi_{\mathcal{D}}}$ stays bounded on Ω (independent of $\mathcal{D}$) and converges a.e. to $\frac{1}{\Phi}$. The Lebesgue dominated convergence theorem then shows that $c=\frac{1}{\Phi_{\mathcal{D}}}\Phi_{\mathcal{D}}c\to\frac{1}{\Phi}f$ in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$, which concludes the proof. ∎

In what follows, we extract a sequence such that c converges in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$ to some $\bar c$.

5.2. Convergence of the pressure. Let us now turn to the convergence of $(p,\mathbf{v},\mathbf{U})$. By Proposition 3.1, we can assume, up to a subsequence, that $p\to\bar p$ weakly-$*$ in $L^\infty(0,T;L^2(\Omega))$ and that $\mathbf{v}\to\bar{\mathbf{v}}$ weakly-$*$ in $L^\infty(0,T;L^2(\Omega))^d$. Since $\int_\Omega p(t,\cdot)=0$ for all $t\in(0,T)$, it is quite clear that $\int_\Omega\bar p(t,\cdot)=0$ for a.e. $t\in(0,T)$. By choice of ν_K and thanks to the estimate on F in Proposition 3.1 and the fact that $2d-2+\beta>0$, we have

$$\begin{aligned}
&\sum_{n=1}^{N_k}k\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{diam}(K)^{d-1}\nu_K\mathrm{m}(K)|F^n_{K,\sigma}|\\
&\le\left(\sum_{n=1}^{N_k}k\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(K)\right)^{\frac12}\left(\sum_{n=1}^{N_k}k\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{diam}(K)^{2d-2}\nu_K^2\mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\\
&\le C_{28}\left(\sup_{n=1,\dots,N_k}\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{diam}(K)^{2d-2+\beta}\nu_K\mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\\
&\le C_{29}\,\mathrm{size}(\mathcal{D})^{\frac{2d-2+\beta}{2}}. \qquad (51)
\end{aligned}$$

Hence, Lemma 7.4 shows that $\bar p\in L^2(0,T;H^1(\Omega))$ and that $\nabla\bar p=\bar{\mathbf{v}}$, so that $\bar p\in L^\infty(0,T;H^1(\Omega))$. Let $A_{\mathcal{D}}:\Omega\times\mathbb{R}\to M_d(\mathbb{R})$ be the function defined by $A_{\mathcal{D}}(x,s)=A_K(s)$ whenever $s\in\mathbb{R}$ and x belongs to $K\in\mathcal{M}$. We also define $\check c:(0,T)\times\Omega\to\mathbb{R}$ by $\check c=c_K^{n-1}$ on $[(n-1)k,nk)\times K$ ($n=1,\dots,N_k$ and $K\in\mathcal{M}$); noticing that $\check c=c_K^0\in[0,1]$ on $[0,k[\times K$ and that $\check c=c(\cdot-k,\cdot)$ on $[k,T[\times\Omega$, it is clear that $\check c\to\bar c$ in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$ as $k\to 0$ and $\mathrm{size}(\mathcal{D})\to 0$. We have $\mathbf{U}=-A_{\mathcal{D}}(\cdot,\check c)\mathbf{v}$ and thus, for all $\mathbf{Z}\in L^2((0,T)\times\Omega)^d$, $\int_0^T\int_\Omega\mathbf{Z}\cdot\mathbf{U}=\int_0^T\int_\Omega-A_{\mathcal{D}}(\cdot,\check c)^T\mathbf{Z}\cdot\mathbf{v}$. Applying Lemma 7.6 (with $-A^T$ instead of A, $u^m=\check c$, and $\mathbf{Z}^m$ constant equal to $\mathbf{Z}$), and since $\mathbf{v}$ converges to $\nabla\bar p$ weakly in $L^2((0,T)\times\Omega)^d$, we obtain that $\int_0^T\int_\Omega\mathbf{Z}\cdot\mathbf{U}\to\int_0^T\int_\Omega-A(\cdot,\bar c)^T\mathbf{Z}\cdot\nabla\bar p$, which proves that $\mathbf{U}\to\bar{\mathbf{U}}=-A(\cdot,\bar c)\nabla\bar p$ weakly in $L^2((0,T)\times\Omega)^d$ (since $\mathbf{U}$ is bounded in $L^\infty(0,T;L^2(\Omega))^d$, the convergence also holds weakly-$*$ in this space).

Let us now prove that $\bar{p}$ is the weak solution to (8) with $\bar{c}$ fixed as given above. Let $\varphi \in C^\infty([0,T]\times\overline{\Omega})$ and define $\varphi^n(x) = \frac{1}{k}\int_{(n-1)k}^{nk}\varphi(t,x)\,dt$ for $n = 1,\dots,N_k$. Multiply (21) by $\varphi^n(\mathbf{x}_K)$, sum over all control volumes, and, using (18) and (23), gather by edges; this gives

$$\sum_{K\in\mathcal{M}} \mathrm{m}(K)(q_K^{+,n} - q_K^{-,n})\varphi^n(\mathbf{x}_K) = \sum_{\sigma=K|L\in\mathcal{E}_{\rm int}} F_{K,\sigma}^n(\varphi^n(\mathbf{x}_L) - \varphi^n(\mathbf{x}_K)).$$

However, since φ is regular, we have

$$\begin{aligned}\varphi^n(\mathbf{x}_L) - \varphi^n(\mathbf{x}_K) &= \nabla\varphi^n(\mathbf{x}_K)\cdot(\mathbf{x}_\sigma - \mathbf{x}_K) + \nabla\varphi^n(\mathbf{x}_L)\cdot(\mathbf{x}_L - \mathbf{x}_\sigma)\\ &\quad + R_{K,\sigma}^n - R_{L,\sigma}^n\\ &\text{with } |R_{K,\sigma}^n| \le C_{30}\mathrm{diam}(K)^2\,,\end{aligned} \tag{52}$$

where C_{30} does not depend on n, $\sigma = K|L$, k, or $\mathcal{D}$. Therefore,

$$\begin{aligned}\sum_{K\in\mathcal{M}} \mathrm{m}(K)(q_K^{+,n} - q_K^{-,n})\varphi^n(\mathbf{x}_K) &= \sum_{K\in\mathcal{M}} \nabla\varphi^n(\mathbf{x}_K)\cdot\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n(\mathbf{x}_\sigma - \mathbf{x}_K)\\ &\quad + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n R_{K,\sigma}^n\\ &= -\sum_{K\in\mathcal{M}} \mathrm{m}(K)\nabla\varphi^n(\mathbf{x}_K)\cdot\mathbf{U}_K^n + \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n R_{K,\sigma}^n.\end{aligned} \tag{53}$$

If $\varphi_{k,\mathcal{D}}$ and $\Psi_{k,\mathcal{D}}$ denote the functions on $[0,T)\times\Omega$ which are equal to $\varphi^n(\mathbf{x}_K)$ and to $\nabla\varphi^n(\mathbf{x}_K)$ on $[(n-1)k, nk)\times K$, it is clear that $\varphi_{k,\mathcal{D}}\to\varphi$ and $\Psi_{k,\mathcal{D}}\to\nabla\varphi$ uniformly on $(0,T)\times\Omega$ as $k\to 0$ and $\mathrm{size}(\mathcal{D})\to 0$; multiplying (53) by k and summing over $n = 1,\dots,N_k$, we obtain

$$\int_0^T\int_\Omega (q^+ - q^-)\varphi_{k,\mathcal{D}} = -\int_0^T\int_\Omega \Psi_{k,\mathcal{D}}\cdot\mathbf{U} + \sum_{n=1}^{N_k} k\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n R_{K,\sigma}^n. \tag{54}$$

Adapting the proof of (41) to F by using Proposition 3.1, we get

$$\begin{aligned}\left|\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} F_{K,\sigma}^n R_{K,\sigma}^n\right| &\le C_{31}\left(\sum_{K\in\mathcal{M}} \mathrm{m}(K)\mathrm{diam}(K)^{4-2d-\beta}\right)^{\frac{1}{2}}\\ &\le C_{32}\ \mathrm{size}(\mathcal{D})^{\frac{4-2d-\beta}{2}}.\end{aligned} \tag{55}$$

Hence, by the weak convergence of $\mathbf{U}$, we can pass to the limit in (54) and find $\int_0^T\int_\Omega(q^+ - q^-)\varphi = -\int_0^T\int_\Omega \nabla\varphi\cdot\bar{\mathbf{U}}$; since this equation is satisfied for all $\varphi\in C^\infty([0,T]\times\overline{\Omega})$, this concludes the proof that $\bar{p}$ is the weak solution to (8) for the given $\bar{c}$ (limit of c).

We now want to prove the strong convergence of $\mathbf{v}$ to $\nabla\bar{p}$ in $L^2((0,T)\times\Omega)^d$. To do so, we use (20) and (19) in (30), which we then multiply by k and sum over $n = 1,\dots,N_k$; this leads to

$$\int_0^T\int_\Omega (q^+ - q^-)p = \int_0^T\int_\Omega A(\cdot,\breve{c})\mathbf{v}\cdot\mathbf{v} + \sum_{n=1}^{N_k} k\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K \mathrm{m}(K)|F_{K,\sigma}^n|^2. \tag{56}$$

Dropping the last term (which is nonnegative), the weak convergence of p gives, since $\bar{p}$ is a solution to (8),

$$\limsup_{k\to 0\,,\ \mathrm{size}(\mathcal{D})\to 0} \int_0^T \int_\Omega A(\cdot,\check{c})\mathbf{v}\cdot\mathbf{v} \le \int_0^T \int_\Omega (q^+ - q^-)\bar{p} = \int_0^T \int_\Omega A(\cdot,\bar{c})\nabla\bar{p}\cdot\nabla\bar{p} \tag{57}$$

(the last equality is obtained using $\bar{p}$ as a test function in (8), which is possible since the weak formulation of (8) is in fact valid with test functions in $L^1(0,T;H^1(\Omega))$). We now write, thanks to (11),

$$\begin{aligned}\alpha_A \int_0^T \int_\Omega |\mathbf{v}-\nabla\bar{p}|^2 &\le \int_0^T \int_\Omega A(\cdot,\check{c})(\mathbf{v}-\nabla\bar{p})\cdot(\mathbf{v}-\nabla\bar{p})\\ &= \int_0^T \int_\Omega A(\cdot,\check{c})\mathbf{v}\cdot\mathbf{v} - \int_0^T \int_\Omega A(\cdot,\check{c})\mathbf{v}\cdot\nabla\bar{p} - \int_0^T \int_\Omega A(\cdot,\check{c})\nabla\bar{p}\cdot\mathbf{v}\\ &\quad + \int_0^T \int_\Omega A(\cdot,\check{c})\nabla\bar{p}\cdot\nabla\bar{p}. \end{aligned}\tag{58}$$

Up to a subsequence, we can assume that $\check{c}\to\bar{c}$ a.e. on $(0,T)\times\Omega$, and (11) then gives $A(\cdot,\check{c})\nabla\bar{p}\to A(\cdot,\bar{c})\nabla\bar{p}$ and $A(\cdot,\check{c})^T\nabla\bar{p}\to A(\cdot,\bar{c})^T\nabla\bar{p}$ strongly in $L^2((0,T)\times\Omega)^d$. Hence, the weak convergence of $\mathbf{v}$ to $\nabla\bar{p}$ allows us to pass to the limit in the second and third terms on the right-hand side of (58); the last term on this right-hand side obviously converges and (57) therefore gives

$$\begin{aligned}\limsup_{k\to 0\,,\ \mathrm{size}(\mathcal{D})\to 0} \alpha_A \int_0^T \int_\Omega |\mathbf{v}-\nabla\bar{p}|^2 &\le \limsup_{k\to 0\,,\ \mathrm{size}(\mathcal{D})\to 0} \int_0^T \int_\Omega A(\cdot,\check{c})\mathbf{v}\cdot\mathbf{v}\\ &\quad - \int_0^T \int_\Omega A(\cdot,\bar{c})\nabla\bar{p}\cdot\nabla\bar{p} \le 0\,,\end{aligned}$$

which concludes the proof of the strong convergence of $\mathbf{v}$ to $\nabla\bar{p}$ in $L^2((0,T)\times\Omega)^d$. The strong convergence of $\mathbf{U}$ in the same space is then a consequence of Lemma 7.6, of the equality $\mathbf{U} = -A_{\mathcal{D}}(\cdot,\check{c})\mathbf{v}$, and of the strong convergence of $\mathbf{v}$.

We conclude by proving that, up to subsequence and as $k\to 0$ and $\mathrm{size}(\mathcal{D})\to 0$, $p(t)\to\bar{p}(t)$ in $L^1_{\mathrm{loc}}(\Omega)$ for a.e. $t\in(0,T)$. Since p is bounded in $L^\infty(0,T;L^2(\Omega))$, and thus in $L^\infty(0,T;L^1_{\mathrm{loc}}(\Omega))$, this a.e. convergence and Vitali's theorem imply the convergence in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$, and, using once again the bound on p in $L^\infty(0,T;L^2(\Omega))$, we deduce the strong convergences stated in Theorem 2.2.

As $\mathbf{v}$ converges in $L^2(0,T;L^2(\Omega))^d$, we can assume that, up to a subsequence, $\mathbf{v}(t)\to\nabla\bar{p}(t)$ in $L^2(\Omega)^d$ for a.e. $t\in(0,T)$. Take a t_0 for which this convergence holds, and such that $\int_\Omega \bar{p}(t_0)=0$; we now prove, using the method of proof by contradiction, that $p(t_0)\to\bar{p}(t_0)$ in $L^1_{\mathrm{loc}}(\Omega)$ (along the same subsequence as the one chosen for $\mathbf{v}$, which thus does not depend on t_0). If this convergence does not hold, then we can assume, up to a new subsequence, that, for some $\eta>0$, $\mathrm{d}_1(p(t_0),\bar{p}(t_0))\ge\eta$, where d_1 is the distance in $L^1_{\mathrm{loc}}(\Omega)$. By (17), $(p(t_0),\mathbf{v}(t_0),F^{n(t_0,k)})\in L_\nu(\mathcal{D})$ (where $n(t_0,k)$ is such that $(n(t_0,k)-1)k\le t_0<n(t_0,k)k$) and Proposition 3.1 proves, with the help of the Cauchy–Schwarz inequality, that $M_1(\mathcal{D},\nu,F^{n(t_0,k)})$ (defined in Lemma 7.3) stays bounded; hence, since $p(t_0)$ is bounded in $L^2(\Omega)$ (see again Proposition 3.1), Lemma 7.3 and Kolmogorov's compactness theorem show that, up to a subsequence, $p(t_0)$ converges to some P strongly in $L^1_{\mathrm{loc}}(\Omega)$ and weakly in $L^2(\Omega)$. By (22), it is

clear that $\int_\Omega P = 0$ (use the weak convergence in $L^2(\Omega)$). Applying Lemma 7.4 to the functions constant in time $(u, \mathbf{r}) = (p(t_0), \mathbf{v}(t_0))$ and to the fluxes $H = F^{n(t_0,k)}$, the estimates in Proposition 3.1 allow us to see that (64) is satisfied and thus that $\nabla P = \nabla \bar{p}(t_0)$ (because $\mathbf{v}(t_0) \to \nabla \bar{p}(t_0)$); hence, since $\int_\Omega \bar{p}(t_0) = 0$, we deduce that $P = \bar{p}(t_0)$, and therefore that $p(t_0) \to \bar{p}(t_0)$ in $L^1_{\rm loc}(\Omega)$. Since the subsequence along which this convergence holds has been extracted from a sequence which satisfies $\mathrm{d}_1(p(t_0), \bar{p}(t_0)) \geq \eta$, this gives the contradiction we sought.

Remark 5.1. From the strong convergence of $\mathbf{v}$ and the a.e. convergence of $\check{c}$, we have $\int_0^T \int_\Omega A(\cdot, \check{c})\mathbf{v}\cdot\mathbf{v} \to \int_0^T \int_\Omega A(\cdot, \bar{c})\nabla\bar{p}\cdot\nabla\bar{p} = \int_0^T \int_\Omega (q^+ - q^-)\bar{p}$. Hence, (56) implies

$$\lim_{k\to 0\,,\ \mathrm{size}(\mathcal{D})\to 0} \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |F^n_{K,\sigma}|^2 = 0. \tag{59}$$

5.3. Convergence of the concentration. Let us now turn to the convergence of $(c, \mathbf{w})$. By the estimates of Proposition 3.2, the convergence of c to $\bar{c}$ holds not only in $L^1(0,T; L^1_{\rm loc}(\Omega))$, but also in $L^\infty(0,T; L^2(\Omega))$ weak-$*$ and strongly in $L^p(0,T; L^q(\Omega))$ for all $p < \infty$ and $q < 2$. Up to a subsequence, we can assume that $\mathbf{w} \to \bar{\mathbf{w}}$ weakly in $L^2((0,T)\times\Omega)^d$. Thanks to the estimates on G from Proposition 3.2, the analogue of (51) reads

$$\sum_{k=1}^{N_k} k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{diam}(K)^{d-1} \nu_K \mathrm{m}(K) |G^n_{K,\sigma}| \leq C_{33} \mathrm{size}(\mathcal{D})^{\frac{2d-2+\beta}{2}} \to 0 \ \text{ as } \mathrm{size}(\mathcal{D}) \to 0.$$

Hence, by (24) and Lemma 7.4, we have $\bar{c} \in L^2(0,T; H^1(\Omega))$ and $\bar{\mathbf{w}} = \nabla\bar{c}$. We now prove that $\bar{c}$ is a solution to (9), with $\bar{\mathbf{U}}$ the strong limit of $\mathbf{U}$ found in section 5.2. Let $\psi \in C^\infty_c([0,T)\times\overline{\Omega})$ and, for $n = 1, \ldots, N_k$, $\psi^n(x) = \frac{1}{k}\int_{(n-1)k}^{nk} \psi(t,x)\,dt$. We multiply (27) by $k\psi^n(\mathbf{x}_K)$ and sum over all $K \in \mathcal{M}$ and over $n = 1, \ldots, N_k$; this gives $T_6 + T_7 + T_8 + T_9 = T_{10}$. Let us study the limit of each of these terms as $k \to 0$ and $\mathrm{size}(\mathcal{D}) \to 0$.

5.3.1. Limit of T_6. We have, since $\psi^{N_k} = \psi^{N_k+1} = 0$ for k small enough (the support of ψ does not touch $t = T$),

$$\begin{aligned} T_6 &= \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K)\Phi_K \frac{c^n_K - c^{n-1}_K}{k} \psi^n(\mathbf{x}_K) \\ &= \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K)\Phi_K c^n_K \frac{\psi^n(\mathbf{x}_K) - \psi^{n+1}(\mathbf{x}_K)}{k} - \sum_{K\in\mathcal{M}} \mathrm{m}(K)\Phi_K c^0_K \psi^1(\mathbf{x}_K) \\ &= \int_0^T \int_\Omega \Phi c \zeta_{k,\mathcal{D}} - \int_\Omega \Phi_{\mathcal{D}} c_0 \pi_{k,\mathcal{D}}, \end{aligned}$$

where $\Phi_{\mathcal{D}} = \Phi_K$ on K (as before), $\zeta_{k,\mathcal{D}} = \frac{\psi^n(\mathbf{x}_K) - \psi^{n+1}(\mathbf{x}_K)}{k}$ on $[(n-1)k, nk)\times K$, and $\pi_{k,\mathcal{D}} = \psi^1_K$ on K ($n = 1, \ldots, N_k$ and $K \in \mathcal{M}$). By regularity of ψ, it is clear that $\zeta_{k,\mathcal{D}} \to -\partial_t\psi$ uniformly on $(0,T)\times\Omega$ and $\pi_{k,\mathcal{D}} \to \psi(0,\cdot)$ uniformly on Ω; we also recall that $\Phi_{\mathcal{D}} \to \Phi$ strongly in $L^2(\Omega)$. The weak-$*$ convergence of c in $L^\infty(0,T; L^2(\Omega))$ then implies $T_6 \to -\int_0^T \int_\Omega \Phi\bar{c}\partial_t\psi - \int_\Omega \Phi c_0 \psi(0,\cdot)$.

5.3.2. Limit of T_7. Making use of manipulations which should be, at this stage, familiar to the reader, we get, using (52) with $\varphi = \psi$ and letting $\Psi_{k,\mathcal{D}}$ be the function on $[0,T)\times\Omega$ equal to $\nabla\psi^n(\mathbf{x}_K)$ on $[(n-1)k, nk)\times K$,

$$\begin{aligned}
T_7 &= -\sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G^n_{K,\sigma}\psi^n(\mathbf{x}_K)\\
&= \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \nabla\psi^n(\mathbf{x}_K)\cdot\sum_{\sigma\in\mathcal{E}_K} G^n_{K,\sigma}(\mathbf{x}_\sigma-\mathbf{x}_K) + \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G^n_{K,\sigma}R^n_{K,\sigma}\\
&= \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K) D_K(\mathbf{U}^n_K)\mathbf{w}^n_K\cdot\nabla\psi^n(\mathbf{x}_K) + \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G^n_{K,\sigma}R^n_{K,\sigma}\\
&= \int_0^T\int_\Omega \mathbf{w}\cdot D(\cdot,\mathbf{U})^T\Psi_{k,\mathcal{D}} + \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G^n_{K,\sigma}R^n_{K,\sigma}. \qquad (60)
\end{aligned}$$

However, thanks to the estimates on G from Proposition 3.2, the analogue of (55) reads

$$\left|\sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K} G^n_{K,\sigma}R^n_{K,\sigma}\right| \le C_{34}\ \mathrm{size}(\mathcal{D})^{\frac{4-2d-\beta}{2}} \to 0 \quad \text{as } \mathrm{size}(\mathcal{D})\to 0.$$

Since $\mathbf{U}\to\bar{\mathbf{U}}$ strongly in $L^2((0,T)\times\Omega)^d$, hypothesis (12) classically implies that $D(\cdot,\mathbf{U})\to D(\cdot,\bar{\mathbf{U}})$ strongly in $L^2((0,T)\times\Omega)^{d\times d}$. Since $\Psi_{k,\mathcal{D}}\to\nabla\psi$ uniformly on $(0,T)\times\Omega$, we deduce that $D(\cdot,\mathbf{U})^T\Psi_{k,\mathcal{D}}\to D(\cdot,\bar{\mathbf{U}})^T\nabla\psi$ in $L^2((0,T)\times\Omega)^d$ and the weak convergence of $\mathbf{w}$ to $\nabla\bar{c}$ allows us to pass to the limit in (60), and we get $T_7\to\int_0^T\int_\Omega D(\cdot,\bar{\mathbf{U}})\nabla\bar{c}\cdot\nabla\psi$.

5.3.3. Limit of T_8. The term T_8 is built by writing $-kT_2$ (introduced in the proof of Lemma 5.1) with $\psi^n(\mathbf{x}_K)$ instead of ψ_K and summing over n, that is,

$$T_8 = \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}}\sum_{\sigma=K|L\in\mathcal{E}_K\cap\mathcal{E}_{\mathrm{int}}} \left[(-F^n_{K,\sigma})^+c^n_K - (-F^n_{L,\sigma})^-c^n_L\right]\psi^n(\mathbf{x}_K).$$

In the proof of Lemma 5.1, the estimate (49) on T_2 has been proved for test functions φ in $C^2_c(\Omega)$, but it is also valid for test functions in $C^2(\overline{\Omega})$; in the same way, it is still valid if we use, in the definition of T_2, $\varphi(\mathbf{x}_K)$ rather than the mean value of φ on K (because (52) is similar to (38) without requiring $\mathbf{x}_K$ to be the barycenter of K). Therefore,

$$\begin{aligned}
&\left|T_8 + \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K)c^n_K\mathbf{U}^n_K\cdot\nabla\psi^n(\mathbf{x}_K)\right|\\
&\quad\le C_{35}\ \mathrm{size}(\mathcal{D})^{\frac{4-2d-\beta}{2}} \sum_{n=1}^{N_k} k\left(\sum_{K\in\mathcal{M}}\mathrm{m}(K)(|\mathbf{w}^n_K|+|c^n_K|)^2\right)^{\frac12}\\
&\qquad + C_{35}\sum_{n=1}^{N_k} k\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)|F^n_{K,\sigma}|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}}\mathrm{m}(K)|\mathbf{w}^n_K|^2\right)^{\frac12}\\
&\qquad + C_{35}\ \mathrm{size}(\mathcal{D})\sum_{n=1}^{N_k} k\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)|G^n_{K,\sigma}|^2\right)^{\frac12}
\end{aligned}$$

and, using the Cauchy–Schwarz inequality, the estimates of Proposition 3.2, and (59), this right-hand side tends to 0 as $k \to 0$ and size$(\mathcal{D}) \to 0$. With the same $\Psi_{k,\mathcal{D}}$ as before, we have $\sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K) c_K^n \mathbf{U}_K^n \cdot \nabla\psi^n(\mathbf{x}_K) = \int_0^T \int_\Omega c\mathbf{U} \cdot \Psi_{k,\mathcal{D}}$, and we therefore can pass to the limit (using the weak convergence of c in $L^2((0,T)\times\Omega)$, the strong convergence of $\mathbf{U}$ in $L^2((0,T)\times\Omega)^d$, and the uniform convergence of $\Psi_{k,\mathcal{D}}$ on $(0,T)\times\Omega$) to obtain $T_8 \to -\int_0^T \int_\Omega \bar{c}\bar{\mathbf{U}} \cdot \nabla\psi$.

5.3.4. Limits of T_9 and T_{10}. We have, with $\psi_{k,\mathcal{D}}$ equal to $\psi^n(\mathbf{x}_K)$ on $[(n-1)k, nk) \times K$,

$$T_9 = \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K) q_K^{-,n} c_K^n \psi^n(\mathbf{x}_K) = \int_0^T \int_\Omega q^- c\psi_{k,\mathcal{D}} \to \int_0^T \int_\Omega q^- \bar{c}\psi.$$

It is also easy to pass to the limit in

$$T_{10} = \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K) q_K^{+,n} \widehat{c}_K^n \psi^n(\mathbf{x}_K) = \int_0^T \int_\Omega q^+ \widehat{c}_{k,\mathcal{D}} \psi_{k,\mathcal{D}}$$

once we notice that, as for $\Phi_\mathcal{D}$, the function $\widehat{c}_{k,\mathcal{D}}$ equal to $\widehat{c}_K^n$ on $[(n-1)k, nk) \times K$ converges to $\widehat{c}$ in $L^2((0,T)\times\Omega)$. Hence, $T_{10} \to \int_0^T \int_\Omega q^+ \widehat{c}\psi$.

Gathering the preceding convergences in $T_6 + T_7 + T_8 + T_9 = T_{10}$, we deduce that $\bar{c}$ is a weak solution to (9) with the function $\bar{\mathbf{U}}$ being the limit of $\mathbf{U}$.

6. Numerical results. In this section, we illustrate the behavior of the mixed finite volume scheme by applying it to the system (1)–(7), which describes the miscible displacement of one fluid by another in a porous medium. Some of the tests cases come from [20], where an ELLAM-MFEM scheme is used, and our results compare very well to the ones in this reference. In practice, for the implementation of the numerical scheme we have used the hybrid method mentioned in Remark 4.1.

In all the test cases, the spatial domain is $\Omega = (0,1000) \times (0,1000)$ ft^2 and the time period is $[0, 3600]$ days. The injection well is located at the upper-right corner $(1000, 1000)$ with an injection rate $q^+ = 30$ ft^2/day and an injection concentration $\hat{c} = 1.0$. The production well is located at the lower-left corner $(0,0)$ with a production rate $q^- = 30$ ft^2/day. The viscosity of the oil is $\mu(0) = 1.0$ cp, the porosity of the medium is specified as $\Phi(x) = 0.1$, and the initial concentration is $c_0(x) = 0$.

Remark 6.1. Although this does not entirely satisfy the assumptions of our theoretical study, the wells can be considered as Dirac masses; from the point of view of numerical tests, we saw no difference between using Dirac masses for q^+ and q^- or approximations of such masses by functions with small support (which would be admissible in the theoretical study).

The mesh of the domain is partitioned into 928 triangles of maximal edge length 50 ft. We take as time step $k = 36$ days, but the scheme still works with greater time steps (indeed, the discretization is implicit in time and does not require any stability condition). In fact, if we use the same time step $k = 360$ days as in [20], we obtain numerical results close to the ones in this reference but, since the computational times are in any case very short (less than 3 seconds per time step on a personal computer), we choose the smaller time step $k = 36$ days to show more accurate results with respect to the exact solution. As noticed in [9], the choice of ν_K has very little impact on the numerical outcomes and any small value for the penalization gives good results; we therefore take $\nu_K \mathrm{m}(K) = 10^{-6}$ for all K. Note that for $10^{-10} \le \nu_K \mathrm{m}(K) \le 10^{-2}$,

the numerical results are similar. For each test case, we present the surface plot and/or the contour plot of the concentration c, the interesting physical quantity, at $t = 3$ years (≈ 30 time steps) and $t = 10$ years (≈ 100 time steps).

Remark 6.2. Notice that our scheme preserves the discrete mass, that is, for $n = 1, \dots, N_k$,

$$\int_\Omega \phi(x)c^n(x)\,dx + \int_{(n-1)k}^{nk}\int_\Omega q^-(t,x)c^n(x)\,dxdt = \int_\Omega \phi(x)c^{n-1}(x)\,dx + \int_{(n-1)k}^{nk}\int_\Omega q^+(t,x)\widehat{c}^n(x)\,dxdt$$

(this is obtained by summing (27) over all $K \in \mathcal{M}$ and using (25) and (28) to cancel the terms involving $G^n_{K,\sigma}$ and (18) to cancel the terms involving $F^n_{K,\sigma}$). This is of essential importance in the applications.

Remark 6.3. We also notice that, in all the following numerical tests, the computed values of the concentration remain in $[0, 1]$. This is, however, only a numerical verification, not a proof (but, thanks to assumption (11), these bounds are not needed to prove the convergence of the mixed finite volume scheme—and in fact, since the computed c remains in $[0, 1]$, the implementation of the scheme does not require extending μ outside of $[0, 1]$). The mixed finite volume method has many advantages: it works on very general meshes (which can be useful in petroleum engineering; see [14]); it ensures strong convergence of the discrete gradients (and therefore convergence of the scheme for the fully coupled system with minimal regularity assumptions on the data); it can be easily implemented. But the counterpart is that, though the *continuous* concentration remains in $[0, 1]$ (see [3] or [15]), we did not prove such bounds for the *approximate* concentration; they are just verified in numerical experiments (such is also the case for other numerical methods; see, e.g., [12, 17, 20]).

Test 1. For this test case, we assume that the porous medium is homogeneous and isotropic: the permeability tensor is diagonal and constant, $\mathbf{K} = 80\mathbf{I}$. The mobility ratio between the resident and the injected fluids is $M = 1$, so that the viscosity is constant, $\mu(c) = 1.0$ cp.

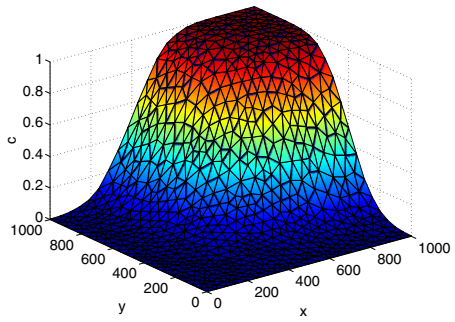
We assume that $\Phi d_m = 1.0$ ft^2/day, $\Phi d_l = 5.0$ ft, and $\Phi d_t = 0.5$ ft. This means that the diffusion effects will be considerably greater than the dispersion effects, which is in fact unrealistic.

The surface plot and the contour plot of the concentration c at $t = 3$ years and $t = 10$ years are shown in Figure 1. As expected, the Darcy velocity is radial and the contour plots are circular until the invading fluid reaches the production well (see at $t = 3$ years). When the production well is reached, the invading fluid continues to fill the whole domain until $c = 1$.

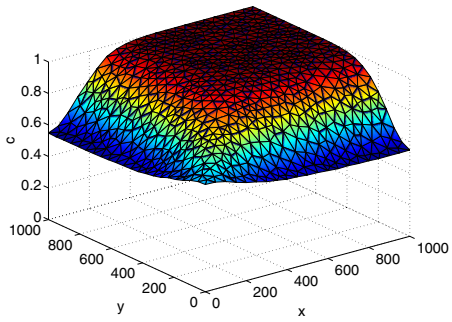
Test 2. The permeability tensor is still diagonal and constant, $\mathbf{K} = 80\mathbf{I}$. The adverse mobility ratio is $M = 41$ and the viscosity $\mu(c)$ now really depends on c.

We assume that there is no molecular diffusion $\Phi d_m = 0.0$ ft^2/day and that $\Phi d_l = 5.0$ ft and $\Phi d_t = 0.5$ ft. This means that we take into account dispersion effects, which is realistic.

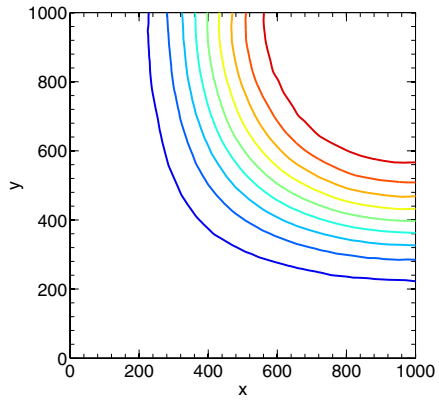
This test case is presented in [20] and permits us to see the macroscopic fingering phenomenon. Indeed, the viscosity $\mu(c)$ rapidly changes across the fluid interface. It induces rapid changes of the Darcy velocity $\mathbf{U}$, and the difference between the longitudinal and the transverse dispersivity coefficients implies that the fluid flow is much faster along the diagonal direction. Such effects can be seen on the surface and contour plots in Figure 2.



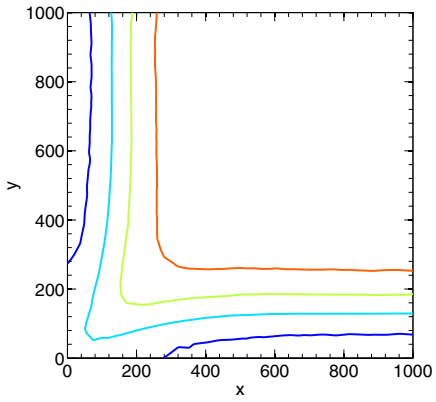
Surface plot at $t = 3$ years



Surface plot at $t = 10$ years



Contour plot at $t = 3$ years



Contour plot at $t = 10$ years

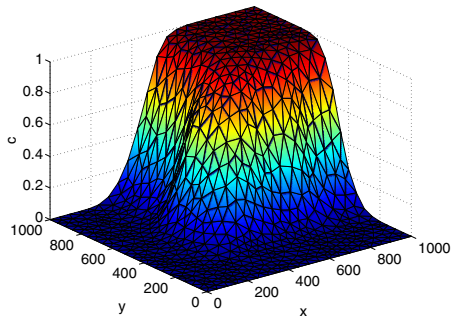
FIG. 1. *Concentration of the invading component in Test* 1.

Remark 6.4. Although this test (as well as Tests 3 and 4) does not satisfy our theoretical assumptions (because $d_m = 0$), we present its results to show that the mixed finite volume scheme is robust and can numerically handle more general cases than the ones admitted in the theoretical study, and also to compare it with other existing schemes for the same equations (note that there is no theoretical study of convergence whatsoever in [20] or [21]).

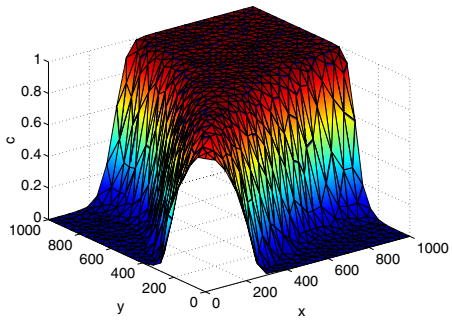
Test 3. In this test case, we consider that the permeability tensor is still diagonal but discontinuous: $\mathbf{K} = 80\mathbf{I}$ on the subdomain $(0, 1000) \times (0, 500)$ and $\mathbf{K} = 20\mathbf{I}$ on the subdomain $(0, 1000) \times (500, 1000)$. The adverse mobility ratio, the molecular diffusion, the longitudinal and the transverse dispersivities are the same as in Test 2.

The lower half domain has a larger permeability than the upper half domain. Therefore, when the invading fluid reaches the lower half domain, it "prefers" to pass through this domain rather than through the domain with lower permeability. As expected, we also notice that the upper half domain is, overall, less invaded than in Test 2. These effects are illustrated by the contour plots of c in Figure 3.

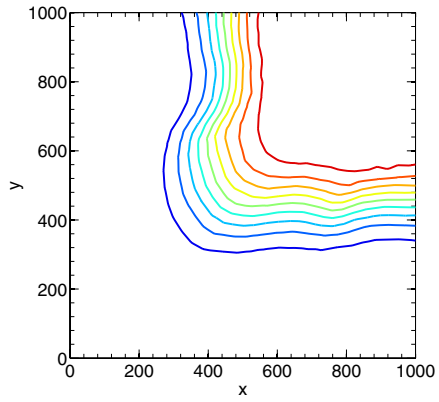
Test 4. In this last test case, the permeability tensor has the form $\mathbf{K} = \kappa(x)\mathbf{I}$ with $\kappa(x) = 80$ except on the four square subdomains $(200, 400) \times (200, 400)$, $(600, 800) \times (200, 400)$, $(200, 400) \times (600, 800)$, and $(600, 800) \times (600, 800)$, where $\kappa(x) = 20$. The



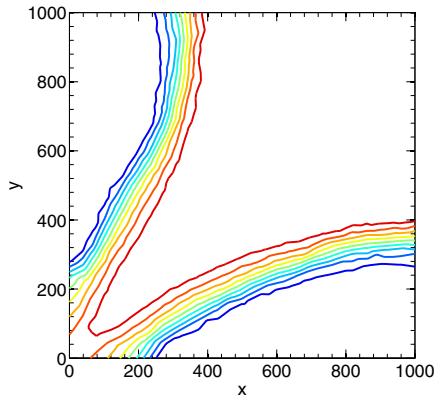
Surface plot at $t = 3$ years



Surface plot at $t = 10$ years



Contour plot at $t = 3$ years



Contour plot at $t = 10$ years

FIG. 2. *Concentration of the invading component in Test* 2.

adverse mobility ratio is $M = 41$, and we take $\Phi d_m = 0.0$ ft^2/day, $\Phi d_l = 5.0$ ft, and $\Phi d_t = 0.5$ ft.

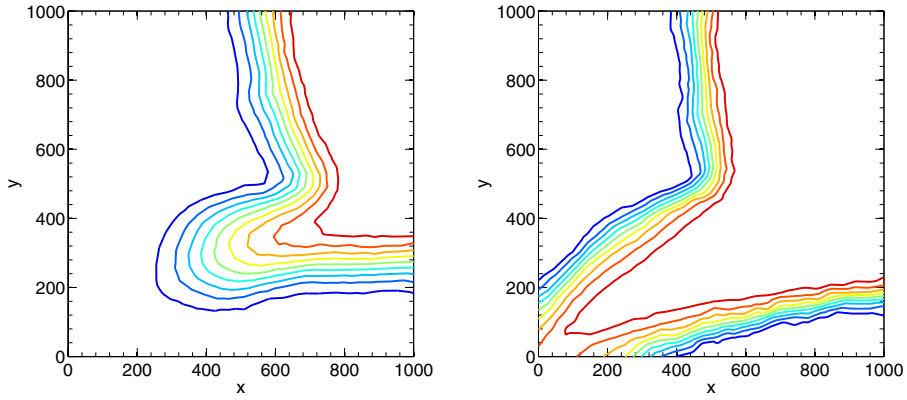
Figure 4 shows the contour plot of the concentration at $t = 3$ years and $t = 10$ years. The subdomains where the permeability is lower can easily be seen in the figures. We note that the area occupied by the invading fluid at $t = 10$ years is in this case larger than in Test 2, where the permeability was homogeneous.

7. Appendix.

7.1. A magical lemma. The proof of the following lemma (a very simple application of Stokes's formula) can be found in [9].

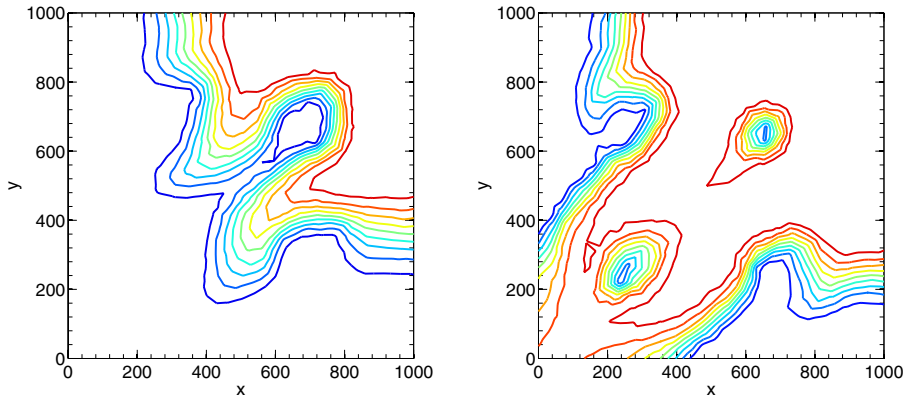
LEMMA 7.1. *Let K be a nonempty polygonal convex domain in $\mathbb{R}^d$. For $\sigma \in \mathcal{E}_K$, we define $\mathbf{x}_\sigma$ as the center of gravity of σ, and $\mathbf{n}_{K,\sigma}$ as the unit normal to σ outward to K. Then, for all vector $\mathbf{e} \in \mathbb{R}^d$ and for all point $\mathbf{x}_K \in \mathbb{R}^d$, we have* $\mathrm{m}(K)\mathbf{e} = \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma)\mathbf{e}\cdot\mathbf{n}_{K,\sigma}(\mathbf{x}_\sigma - \mathbf{x}_K)$, *where* $\mathrm{m}(K)$ *is the d-dimensional measure of K and* $\mathrm{m}(\sigma)$ *is the $(d-1)$-dimensional measure of σ.*

7.2. Lemmas on discrete gradients. For $\mathcal{D}$ an admissible mesh of Ω and $\nu = (\nu_K)_{K\in\mathcal{M}}$ a family of positive numbers, we denote by $L_\nu(\mathcal{D})$ the space of $(u, \mathbf{r}, H)$, with $u = (u_K)_{K\in\mathcal{M}}$ a family of numbers, $\mathbf{r} = (\mathbf{r}_K)_{K\in\mathcal{M}}$ a family of vectors, and



Contour plot at $t = 3$ years Contour plot at $t = 10$ years

FIG. 3. *Concentration of the invading component in Test* 3.



Contour plot at $t = 3$ years Contour plot at $t = 10$ years

FIG. 4. *Concentration of the invading component in Test* 4.

$H = (H_{K,\sigma})_{K\in\mathcal{M},\,\sigma\in\mathcal{E}_K}$ a family of numbers, such that, for all $\sigma = K|L \in \mathcal{E}_{\rm int}$,

$$\mathbf{r}_K \cdot (\mathbf{x}_\sigma - \mathbf{x}_K) + \mathbf{r}_L \cdot (\mathbf{x}_L - \mathbf{x}_\sigma) + \nu_K \mathrm{m}(K) H_{K,\sigma} - \nu_L \mathrm{m}(L) H_{L,\sigma} = u_L - u_K \tag{61}$$

(note that u and $\mathbf{r}$ are also identified with the corresponding functions on Ω constant on each control volume K). The following lemmas are the counterparts for Neumann boundary conditions of lemmas stated in [9] or [8] in the case of Dirichlet boundary conditions.

LEMMA 7.2. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$, $\mathcal{D}$ an admissible mesh of Ω such that* $\mathrm{regul}(\mathcal{D}) \le \theta$ *for some $\theta > 0$, and $\nu = (\nu_K)_{K\in\mathcal{M}}$ a family of positive numbers. Then there exists C_{36} depending only on d, Ω, and θ such that, for all $(u, \mathbf{r}, H) \in L_\nu(\mathcal{D})$ satisfying $\int_\Omega u = 0$,*

$$\|u\|_{L^2(\Omega)} \le C_{36} \left(\|\mathbf{r}\|_{L^2(\Omega)^d} + M_2(\mathcal{D}, \nu, H) \right)$$

with $M_2(\mathcal{D}, \nu, H) = \left(\sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \mathrm{diam}(K)^{2d-2} \nu_K^2 \mathrm{m}(K) |H_{K,\sigma}|^2 \right)^{\frac{1}{2}}$.

Proof. Let w be the weak solution of $-\Delta w = u$ on Ω with homogeneous Neumann boundary conditions on $\partial\Omega$ (such a w exists thanks to the fact that $\int_\Omega u = 0$) and null mean value. Since Ω is convex, it is well known (see [16]) that $w \in H^2(\Omega)$ and that there exists C_{37} depending only on d and Ω such that $\|w\|_{H^2(\Omega)} \le C_{37}\|u\|_{L^2(\Omega)}$.

We multiply each equation of (61) by $\int_\sigma \nabla w\cdot \mathbf{n}_{K,\sigma}\,d\gamma$, sum over the interior edges, gather by control volumes, and use that $\int_\sigma \nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma = 0$ whenever $\sigma\in\mathcal{E}_{\rm ext}$; this gives

$$\begin{aligned}(62)\qquad &\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathbf{r}_K\cdot(\mathbf{x}_\sigma-\mathbf{x}_K)\int_\sigma \nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma\\ &\qquad+\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)H_{K,\sigma}\int_\sigma\nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma = \|u\|^2_{L^2(\Omega)}.\end{aligned}$$

Since $\mathrm{regul}(\mathcal{D})\le\theta$, [8, Lemma 8.1] gives C_{38} depending only on d, Ω, and θ such that

$$\left|\int_\sigma \nabla w\,d\gamma\cdot\mathbf{n}_{K,\sigma}\right|^2 \le \left|\int_\sigma\nabla w\,d\gamma\right|^2 \le \frac{C_{38}\mathrm{m}(\sigma)}{\mathrm{diam}(K)}\|w\|^2_{H^2(K)}.$$

Using the Cauchy–Schwarz inequality, we deduce, since $\mathrm{Card}(\mathcal{E}_K)\le\mathrm{regul}(\mathcal{D})\le\theta$ for all $K\in\mathcal{M}$,

$$\begin{aligned}&\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathbf{r}_K\cdot(\mathbf{x}_\sigma-\mathbf{x}_K)\int_\sigma\nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma\\ &\le\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{m}(K)|\mathbf{r}_K|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{diam}(K)^2}{\mathrm{m}(K)}\left|\int_\sigma\nabla w\,d\gamma\cdot\mathbf{n}_{K,\sigma}\right|^2\right)^{\frac12}\\ &\le(C_{38}\theta)^{\frac12}\left(\sum_{K\in\mathcal{M}}\mathrm{m}(K)|\mathbf{r}_K|^2\right)^{\frac12}\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{diam}(K)\mathrm{m}(\sigma)}{\mathrm{m}(K)}\|w\|^2_{H^2(K)}\right)^{\frac12}.\end{aligned}$$

We have, if $\sigma\in\mathcal{E}_K$, $\mathrm{m}(\sigma)\le\omega_{d-1}\mathrm{diam}(K)^{d-1}$ (where ω_{d-1} is the volume of the unit ball in $\mathbb{R}^{d-1}$); thus, by (16), $\frac{\mathrm{diam}(K)\mathrm{m}(\sigma)}{\mathrm{m}(K)}\le\frac{\mathrm{regul}(\mathcal{D})\omega_{d-1}}{\omega_d}$ and we obtain

$$(63)\quad\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathbf{r}_K\cdot(\mathbf{x}_\sigma-\mathbf{x}_K)\int_\sigma\nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma\le\frac{\theta^{\frac32}\sqrt{C_{38}\omega_{d-1}}}{\sqrt{\omega_d}}\|\mathbf{r}\|_{L^2(\Omega)^d}\|w\|_{H^2(\Omega)}.$$

The Cauchy–Schwarz inequality also gives

$$\begin{aligned}&\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)H_{K,\sigma}\int_\sigma\nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma\\ &\le\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{diam}(K)^{2d-2}\nu_K^2\mathrm{m}(K)|H_{K,\sigma}|^2\right)^{\frac12}\\ &\qquad\qquad\times\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{m}(K)}{\mathrm{diam}(K)^{2d-2}}\left|\int_\sigma\nabla w\cdot\mathbf{n}_{K,\sigma}\,d\gamma\right|^2\right)^{\frac12}\\ &\le\sqrt{C_{38}}\,M_2(\mathcal{D},\nu,H)\left(\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\frac{\mathrm{m}(\sigma)\mathrm{m}(K)}{\mathrm{diam}(K)^{2d-1}}\|w\|^2_{H^2(K)}\right)^{\frac12}.\end{aligned}$$

Since $\frac{\mathrm{m}(\sigma)\mathrm{m}(K)}{\mathrm{diam}(K)^{2d-1}} \le \omega_{d-1}\omega_d$, this inequality and (63) plugged in (62) conclude the proof, since $\|w\|_{H^2(\Omega)} \le C_{37}\|u\|_{L^2(\Omega)}$. ∎

LEMMA 7.3. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$, $\mathcal{D}$ an admissible mesh of Ω such that* $\mathrm{regul}(\mathcal{D}) \le \theta$ *for some $\theta > 0$, and $\nu = (\nu_K)_{K\in\mathcal{M}}$ a family of positive numbers. Let ω be relatively compact in Ω. Then there exists C_{39} depending only on d, Ω, ω, and θ such that, for all $(u, \mathbf{r}, H) \in L_\nu(\mathcal{D})$ and all $|\xi| < \mathrm{dist}(\omega, \mathbb{R}^d\backslash\Omega)$,*

$$\|u(\cdot+\xi) - u\|_{L^1(\omega)} \le C_{39}\left(\|\mathbf{r}\|_{L^1(\Omega)^d} + M_1(\mathcal{D},\nu,H)\right)|\xi|,$$

where $M_1(\mathcal{D},\nu,H) = \sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\mathrm{diam}(K)^{d-1}\nu_K\mathrm{m}(K)|H_{K,\sigma}|$.

We leave to the reader the proof of Lemma 7.3, counterpart of Lemma 3.2 in [9].

LEMMA 7.4. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$ and let $T > 0$. Let $(\mathcal{D}_m)_{m\ge1}$ be a sequence of admissible meshes of Ω such that* $\mathrm{size}(\mathcal{D}_m) \to 0$ *as $m \to \infty$ and* $(\mathrm{regul}(\mathcal{D}_m))_{m\ge1}$ *is bounded. We also take, for all $m \ge 1$, $k_m > 0$ such that $N_{k_m} = T/k_m$ is an integer and $k_m \to 0$ as $m \to \infty$, and $\nu_m = (\nu_{m,K})_{K\in\mathcal{M}_m}$ a family of positive numbers.*

For all $m \ge 1$ and all $n = 1, \dots, N_{k_m}$, we take $(u^{m,n}, \mathbf{r}^{m,n}) = (u_K^{m,n}, \mathbf{r}_K^{m,n})_{K\in\mathcal{M}_m}$ and a family $H^{m,n} = (H_{K,\sigma}^{m,n})_{K\in\mathcal{M}_m,\,\sigma\in\mathcal{E}_K}$ such that $(u^{m,n}, \mathbf{r}^{m,n}, H^{m,n}) \in L_{\nu_m}(\mathcal{D}_m)$. We let $(u^m, \mathbf{r}^m)$ be the functions on $[0,T)\times\Omega$ equal to $(u_K^{m,n}, \mathbf{r}_K^{m,n})$ on $[(n-1)k, nk)\times K$ (for $n = 1, \dots, N_{k_m}$ and $K \in \mathcal{M}_m$).

Assume that, as $m \to \infty$, $u^m \to \bar{u}$ weakly in $L^2((0,T)\times\Omega)$, $\mathbf{r}^m \to \bar{\mathbf{r}}$ weakly in $L^2((0,T)\times\Omega)^d$, and

$$\sum_{n=1}^{N_{k_m}} k_m \sum_{K\in\mathcal{M}_m}\sum_{\sigma\in\mathcal{E}_K}\mathrm{diam}(K)^{d-1}\nu_{m,K}\mathrm{m}(K)|H_{K,\sigma}^{m,n}| \to 0. \tag{64}$$

Then $\bar{u} \in L^2(0,T;H^1(\Omega))$ and $\nabla\bar{u} = \bar{\mathbf{r}}$.

Proof. We first simplify the notation by dropping the index m; hence, we denote $\mathcal{D} = \mathcal{D}_m$, $k = k_m$, $u = u^m$, $\mathbf{r} = \mathbf{r}^m$, $H_{K,\sigma}^n = H_{K,\sigma}^{m,n}$, and we are interested in the convergence of quantities as $k \to 0$ and $\mathrm{size}(\mathcal{D}) \to 0$.

To prove the lemma, we just need to show that $\nabla\bar{u} = \bar{\mathbf{r}}$ in the sense of the distributions on $(0,T)\times\Omega$. Let $\varphi \in C_c^\infty((0,T)\times\Omega)$ and $\mathbf{e} \in \mathbb{R}^d$; we multiply each equation (61) on $(u^n, \mathbf{r}^n, H^n)$ by $\int_{(n-1)k}^{nk}\int_\sigma \varphi\mathbf{e}\cdot\mathbf{n}_{K,\sigma}\,d\gamma$. We then sum over all the edges and, using $\mathbf{n}_{K,\sigma} = -\mathbf{n}_{L,\sigma}$ if $\sigma = K|L \in \mathcal{E}_{\mathrm{int}}$, we gather by control volumes. Thanks to the fact that $\int_{(n-1)k}^{nk}\int_\sigma \varphi\mathbf{e}\cdot\mathbf{n}_{K,\sigma}\,d\gamma = 0$ if $\sigma \in \mathcal{E}_{\mathrm{ext}}$, we can freely introduce the terms corresponding to boundary edges (which are otherwise not present). Finally summing over $n = 1, \dots, N_k$, we obtain

$$\begin{aligned}&\sum_{n=1}^{N_k}\sum_{K\in\mathcal{M}}\mathbf{r}_K^n\cdot\sum_{\sigma\in\mathcal{E}_K}\int_{(n-1)k}^{nk}\int_\sigma \varphi\mathbf{e}\cdot\mathbf{n}_{K,\sigma}\,d\gamma(\mathbf{x}_\sigma - \mathbf{x}_K)\\ &+\sum_{n=1}^{N_k}\sum_{K\in\mathcal{M}}\sum_{\sigma\in\mathcal{E}_K}\nu_K\mathrm{m}(K)H_{K,\sigma}^n\int_{(n-1)k}^{nk}\int_\sigma \varphi\mathbf{e}\cdot\mathbf{n}_{K,\sigma}\,d\gamma = -\int_0^T\int_\Omega u\mathrm{div}(\varphi\mathbf{e}).\end{aligned} \tag{65}$$

By convergence of u, this right-hand side tends, as $k \to 0$ and $\mathrm{size}(\mathcal{D}) \to 0$, to $-\int_0^T\int_\Omega \bar{u}\mathrm{div}(\varphi\mathbf{e})$. Let us denote by T_{11} and T_{12} the two terms on the left-hand side of this equality.

We have, since φ is bounded and $\mathrm{m}(\sigma) \le \omega_{d-1}\mathrm{diam}(K)^{d-1}$ if $\sigma \in \mathcal{E}_K$,

$$|T_{12}| \le \|\varphi\|_\infty \omega_{d-1} \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \sum_{\sigma\in\mathcal{E}_K} \nu_K \mathrm{m}(K) |H^n_{K,\sigma}| \mathrm{diam}(K)^{d-1}$$

and thus, by assumption, $T_{12} \to 0$ as $k \to 0$ and $\mathrm{size}(\mathcal{D}) \to 0$. We now compare T_{11} with

$$T_{13} = \sum_{n=1}^{N_k} \sum_{K\in\mathcal{M}} \mathbf{r}^n_K \cdot \int_{(n-1)k}^{nk} \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma) \left(\frac{1}{\mathrm{m}(K)} \int_K \varphi\, \mathbf{e} \right) \cdot \mathbf{n}_{K,\sigma} (\mathbf{x}_\sigma - \mathbf{x}_K).$$

Since φ is regular, we have C_{40} depending only on φ such that

$$|T_{11} - T_{13}| \le C_{40} \mathrm{size}(\mathcal{D}) \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} |\mathbf{r}^n_K| \sum_{\sigma\in\mathcal{E}_K} \mathrm{m}(\sigma) \mathrm{diam}(K).$$

Using the fact that $\mathrm{regul}(\mathcal{D})$ stays bounded and that $\mathrm{m}(\sigma) \le \omega_{d-1}\mathrm{diam}(K)^{d-1}$, we get

$$|T_{11} - T_{13}| \le C_{41} \mathrm{size}(\mathcal{D}) \sum_{n=1}^{N_k} k \sum_{K\in\mathcal{M}} \mathrm{m}(K) |\mathbf{r}^n_K| = C_{41} \mathrm{size}(\mathcal{D}) \|\mathbf{r}\|_{L^1((0,T)\times\Omega)^d}.$$

Since $\mathbf{r}$ is bounded in $L^2((0,T)\times\Omega)^d$, this shows that $T_{11} - T_{13} \to 0$ as $\mathrm{size}(\mathcal{D}) \to 0$. Using Lemma 7.1 with $\frac{1}{\mathrm{m}(K)} \int_K \varphi(t,\cdot)\, \mathbf{e}$ instead of $\mathbf{e}$, we get

$$T_{13} = \sum_{n=1}^{N_k} \sum_{K\in\mathcal{M}} \mathbf{r}^n_K \cdot \int_{(n-1)k}^{nk} \int_K \varphi\, \mathbf{e} = \int_0^T \int_\Omega \mathbf{r} \cdot \varphi\, \mathbf{e} \longrightarrow \int_0^T \int_\Omega \bar{\mathbf{r}} \cdot \varphi\, \mathbf{e}$$

as $k \to 0$ and $\mathrm{size}(\mathcal{D}) \to 0$. Hence, the limit of (65) as $k \to 0$ and $\mathrm{size}(\mathcal{D}) \to 0$ gives $\int_0^T \int_\Omega \bar{\mathbf{r}} \cdot \varphi\, \mathbf{e} = -\int_0^T \int_\Omega \bar{u} \mathrm{div}(\varphi \mathbf{e})$, which concludes the proof. ∎

7.3. A compactness lemma. The following lemma, whose proof is inspired by classical proofs of Kolmogorov's or Aubin's compactness theorems, mixes a weak time-compactness and a space-equicontinuity property to obtain a strong time-space compactness.

LEMMA 7.5. *Let Ω be a bounded domain in $\mathbb{R}^d$, let $T > 0$, and let $A \subset L^1(0,T; L^1_{\mathrm{loc}}(\Omega))$. If A is relatively compact in $L^1(0,T;(C^2_c(\Omega))')$ and if, for all ω relatively compact in Ω,*

$$\sup_{u\in A} \|u(\cdot,\cdot+\xi) - u\|_{L^1((0,T)\times\omega)} \to 0 \quad \text{as } |\xi| \to 0,$$

then A is relatively compact in $L^1(0,T; L^1_{\mathrm{loc}}(\Omega))$.

Proof. Let ω be relatively compact in Ω and take $(\rho_\mu)_{0<\mu<\mathrm{dist}(\omega,\mathbb{R}^d\setminus\Omega)}$ smoothing kernels on $\mathbb{R}^d$ such that $\mathrm{supp}(\rho_\mu)$ is included in the ball of center 0 and radius μ. For $u \in A$, let $u_\mu = u * \rho_\mu$ (the convolution being only on the space variable), which is defined on $(0,T)\times\omega$.

We first prove that, for all μ, $A_\mu = \{u_\mu,\ u \in A\}$ is relatively compact in $L^1((0,T)\times\omega)$. Let $(u^n_\mu)_{n\ge1}$ be a sequence in A_μ. Since $(u^n)_{n\ge1}$ lies in A, it is relatively

compact in $L^1(0,T;(C_c^2(\Omega))')$ and we can assume, up to a subsequence, that it converges in this space. We then have, for all $(t,x)\in(0,T)\times\omega$, since $\mathrm{supp}(\rho_\mu(x-\cdot))\subset\Omega$ by choice of μ,

$$|u_\mu^n(t,x)-u_\mu^m(t,x)| = \left|\int_\Omega (u^n(t,y)-u^m(t,y))\rho_\mu(x-y)\,dx\right|$$
$$\le \|u^n(t,\cdot)-u^m(t,\cdot)\|_{(C_c^2(\Omega))'}\|\rho_\mu(x-\cdot)\|_{C_c^2(\Omega)}.$$

Hence, integrating on $x\in\omega$ and $t\in(0,T)$, we find C_μ depending on μ but not on n or m such that $\|u_\mu^n-u_\mu^m\|_{L^1((0,T)\times\omega)}\le C_\mu\|u^n-u^m\|_{L^1(0,T;(C_c^2(\Omega))')}$, which shows that $(u_\mu^n)_{n\ge1}$ converges in $L^1((0,T)\times\omega)$ since $(u^n)_{n\ge1}$ converges in $L^1(0,T;(C_c^2(\Omega))')$. Hence, for all $\mu\in(0,\mathrm{dist}(\omega,\mathbb{R}^d\backslash\Omega))$, A_μ is relatively compact in $L^1((0,T)\times\omega)$.

Let us now conclude. It is sufficient to show that $\sup_{u\in A}\|u-u_\mu\|_{L^1((0,T)\times\omega)}$ goes to 0 as $\mu\to0$. Indeed, once this is done, we get $A\subset A_\mu+B_{L^1((0,T)\times\omega)}(0,\delta(\mu))$ with $\delta(\mu)\to0$ as $\mu\to0$, which clearly shows, since A_μ is precompact in $L^1((0,T)\times\omega)$, that A is also precompact (and thus relatively compact) in this space. Let $u\in A$, $t\in(0,T)$, and $x\in\omega$; we have $|u(t,x)-u_\mu(t,x)|\le\int_{B(0,\mu)}|u(t,x)-u(t,x-y)|\rho_\mu(y)\,dy$ and thus, integrating on $x\in\omega$ and $t\in(0,T)$,

$$\|u-u_\mu\|_{L^1((0,T)\times\omega)}\le\int_{B(0,\mu)}\int_0^T\int_\omega|u(t,x)-u(t,x-y)|\,dtdx\,\rho_\mu(y)\,dy$$
$$\le\sup_{|y|\le\mu}\int_0^T\int_\omega|u(t,x)-u(t,x-y)|\,dtdx,$$

and the proof is concluded. ∎

7.4. A technical lemma. The proof of the following technical lemma is left to the reader.

LEMMA 7.6. *Let Ω be a convex polygonal bounded domain in $\mathbb{R}^d$, let $T>0$, and let $A:\Omega\times\mathbb{R}\to M_d(\mathbb{R})$ be a Carathéodory bounded matrix-valued function. Let $(\mathcal{D}_m)_{m\ge1}$ be a sequence of admissible meshes of Ω such that $\mathrm{size}(\mathcal{D}_m)\to0$ as $m\to\infty$, and let $k_m>0$ be such that $N_{k_m}=T/k_m$ is an integer and $k_m\to0$ as $m\to\infty$.*

Let $u^m=(u_K^{m,n})_{n=1,\dots,N_{k_m},\,K\in\mathcal{M}}$ be a function on $(0,T)\times\Omega$, constant on each $[(n-1)k,nk)\times K$ ($n=1,\dots,N_{k_m}$, $K\in\mathcal{M}_m$). We assume that $u^m\to\bar u$ in $L^1(0,T;L^1_{\mathrm{loc}}(\Omega))$ as $m\to\infty$. Let $\mathbf{Z}^m\in L^2((0,T)\times\Omega)^d$, which converges to $\bar{\mathbf{Z}}$ in $L^2((0,T)\times\Omega)^d$ as $m\to\infty$. Define $A_{\mathcal{D}_m}:\Omega\times\mathbb{R}\to M_d(\mathbb{R})$ by $A_{\mathcal{D}_m}(x,s)=\frac{1}{\mathrm{m}(K)}\int_K A(y,s)\,dy$ whenever x belongs to $K\in\mathcal{M}_m$.

Then $A_{\mathcal{D}_m}(\cdot,u^m)\mathbf{Z}^m\to A(\cdot,\bar u)\bar{\mathbf{Z}}$ in $L^2((0,T)\times\Omega)^d$ as $m\to\infty$.

REFERENCES

[1] Y. Amirat and A. Ziani, *Asymptotic behavior of the solutions of an elliptic-parabolic system arising in flow in porous media*, Z. Anal. Anwendungen, 23 (2004), pp. 335–351.

[2] J. Bear, *Dynamics of Fluids in Porous Media*, American Elsevier, New York, 1972.

[3] Z. Chen and R. Ewing, *Mathematical analysis for reservoir models*, SIAM J. Math. Anal., 30 (1999), pp. 431–453.

[4] J. Douglas, *Numerical methods for the flow of miscible fluids in porous media*, in Numerical Methods in Coupled Systems, R. W. Lewis, P. Bettess, and E. Hinton eds., John Wiley, New York, 1984, pp. 405–439.

[5] J. Douglas, R. E. Ewing, and M. F. Wheeler, *The approximation of the pressure by a mixed method in the simulation of miscible displacement*, RAIRO Anal. Numér., 17 (1983), pp. 17–33.

[6] J. Douglas, R. E. Ewing, and M. F. Wheeler, *A time-discretization procedure for a mixed finite element approximation of miscible displacement in porous media*, RAIRO Anal. Numér., 17 (1983), pp. 249–265.

[7] J. Droniou, *Intégration et Espaces de Sobolev à valeurs vectorielles*, available at http://www-gm3.univ-mrs.fr/polys/.

[8] J. Droniou, *Finite volume schemes for fully non-linear elliptic equations in divergence form*, M2AN Math. Model. Numer. Anal, 40 (2006), pp. 1069–1100.

[9] J. Droniou and R. Eymard, *A mixed finite volume scheme for anisotropic diffusion problems on any grid*, Numer. Math., 105 (2006), pp. 35–71.

[10] R. Ewing, T. Russel, and M. Wheeler, *Simulation of miscible displacement using mixed methods and a modified method of characteristics*, in Proceedings of the 7th SPE Symposium on Reservoir Simulation, Dallas, TX, Paper SPE 12241, Society of Petroleum Engineers, 1983, pp. 71–81.

[11] R. E. Ewing and T. F. Russell, *Efficient time-stepping methods for miscible displacement problems in porous media*, SIAM J. Numer. Anal., 19 (1982), pp. 1–67.

[12] R. E. Ewing, T. F. Russell, and M. F. Wheeler, *Convergence analysis of an approximation of miscible displacement in porous media by mixed finite elements and a modified method of characteristics*, Comput. Methods Appl. Mech. Engrg., 47 (1984), pp. 73–92.

[13] R. Eymard, T. Gallouët, and R. Herbin, *Finite volume methods*, in Handbook of Numerical Analysis, Vol. VII, P. G. Ciarlet and J. L. Lions, eds., North-Holland, Amsterdam, pp. 713–1020.

[14] I. Faille, *A control volume method to solve an elliptic equation on a two-dimensional irregular mesh*, Comput. Methods Appl. Mech. Engrg., 100 (1992), pp. 275–290.

[15] X. Feng, *On existence and uniqueness results for a coupled system modeling miscible displacement in porous media*, J. Math. Anal. Appl., 194 (1995), pp. 883–910.

[16] P. Grisvard, *Elliptic Problems in Nonsmooth Domains*, Pitman, Boston, MA, 1985.

[17] P. Lin and D. Yang, *An iterative perturbation method for the pressure equation in the simulation of miscible displacement in porous media*, SIAM J. Sci. Comput., 19 (1998), pp. 893–911.

[18] J. E. Roberts and J.-M. Thomas, *Mixed and hybrid methods*, in Handbook of Numerical Analysis, Vol. II, North-Holland, Amsterdam, 1991, pp. 523–639.

[19] T. F. Russell, *Finite elements with characteristics for two-component incompressible miscible displacement*, in Proceedings of the 6th SPE Symposium on Reservoir Simulation, New Orleans, Paper SPE 10500, Society of Petroleum Engineers, 1982, pp. 123–135.

[20] H. Wang, D. Liang, R. E. Ewing, S. L. Lyons, and G. Qin, *An approximation to miscible fluid flows in porous media with point sources and sinks by an Eulerian–Lagrangian localized adjoint method and mixed finite element methods*, SIAM J. Sci. Comput., 22 (2000), pp. 561–581.

[21] H. Wang, D. Liang, R. E. Ewing, S. L. Lyons, and G. Qin, *An improved numerical simulator for different types of flows in porous media*, Numer. Methods Partial Differential Equations, 19 (2003), pp. 343–362.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2259–2297

ANALYSIS OF DIRECT THREE-DIMENSIONAL PARABOLIC PANEL METHODS*

PHILIPPE PONCET†

Abstract. Adherence boundary conditions for time dependent partial differential equations, via Chorin algorithm, can be reduced to a parabolic problem with Robin–Fourier boundary conditions in the three-dimensional context. In the spirit of panel methods, one establishes an integral formulation whose key point is the estimation of the potential density, introducing a kind of panel method for tangential kinematic boundary conditions. This paper discusses explicit estimations of this density in the general case of an arbitrarily shaped three-dimensional body, which leads to a fast numerical scheme. An error analysis is also provided, involving body smoothness, the Hölder exponent of the density, and whether the body presents torsion or not.

Key words. numerical analysis, boundary value problems, integral representation, fundamental solutions, parabolic equations, growth of solutions

AMS subject classification. 65M, 35K20, 65L10, 35B05, 35C15

DOI. 10.1137/050625849

1. Introduction.

Numerical techniques aimed at solving partial differential equations involving kinematic boundary conditions, such as the Navier–Stokes equations, have been viewed from many perspectives. These kinematic boundary conditions usually rely on zero velocity field on bodies when considering viscous flows. Despite the fact that these conditions are all mathematically of homogeneous Dirichlet type, fixing velocity value at boundaries, such vectorial boundary conditions have very different meanings physically, depending on the velocity component: while the zero normal component of velocity field on a body is linked to a no-slip-through property, or impermeability, the zero tangential components come from an adherence property, or no-slip condition, not required for ideal fluids relevant to the Euler equations.

The present article focuses on the integral formulation of adherence properties, which is related to a parabolic problem via the Chorin algorithm (instead of elliptic for classical panel methods). We present for the first time the numerical analysis of an ad hoc density evaluation, commonly known as the fastest way to ensure adherence properties.

It is now generally recognized that integral methods provide powerful tools to enforce numerically such boundary conditions. Concerning the normal conditions of velocity, the most common discrete integral technique, known as the "panel method," was pioneered by Hess [18, 19] in the 1970s and consists in using a formulation close to electromagnetism [26], that is, in finding a potential of the form

$$\phi(x) = \int_{\partial\Omega} K(x,y)q(y)\mathrm{d}\sigma(y). \tag{1.1}$$

In this equation (1.1), K is a Green function, whose expression is $(4\pi|x-y|)^{-1}$ in the full space $\mathbb{R}^3$, and q is the density function, defined over domain boundary and

*Received by the editors March 3, 2005; accepted for publication (in revised form) January 17, 2007; published electronically October 10, 2007.
http://www.siam.org/journals/sinum/45-6/62584.html
†Laboratoire MIP, Department GMM, INSA, 135 avenue de Rangueil, 31077 Toulouse Cedex 4, France (Philippe.Poncet@insa-toulouse.fr).

solution of the following integral equation:

$$-\frac{q(x)}{2}+\int_{\partial\Omega} n_x\cdot\nabla K(x,y)\,q(y)\,\mathrm{d}\sigma(y)=g(x), \tag{1.2}$$

where n_x denotes the normal field to $\partial\Omega$ and g is a given function depending on the problem considered. The velocity is then obtained by differentiation of this potential. Proofs of regularity and well-posedness properties of related discrete operators can be found in the existing literature (see [34], for example). This usually leads to solving a large linear system of the size of the boundary discretization [21], which can be nevertheless efficiently preconditioned [16]. The order of convergence can be under control and possibly high [3, 4, 35]. In order to speed up the computation, a way to proceed is to use multipole methods [15, 33, 17], which are by definition well adapted for Lagrangian or pointwise formulations [2]. Another way is to provide an estimate of density, which limits convergence order but dramatically decreases computational time since only potential evaluation remains to be computed [13].

The tangential part of kinematic boundary conditions is a completely different matter. In the fluid dynamics context, these conditions are related physically to viscous effects, modeled by the Laplacian operator in the Navier–Stokes equations, which makes them of parabolic type, as opposed to the Euler equations, which are hyperbolic. Since the 1980s, several numerical schemes aimed at splitting apart linearity and nonlinearity have been proposed and implemented in various fields of physics and mathematical physics, in order to use well-fitted numerical techniques taking into account the linearity, or lack thereof.

These splitting techniques, also known as fractional step algorithms, can be basically of first order, or second order when based on the Strang formula, or higher order by using more general Trotter permutation formulae. Splitting the Navier–Stokes equations [22] over a time step leads one to consider successively the Euler equation with only its natural no-slip-though boundary condition, and then a Stokes equation with full no-slip conditions [9]. In its vorticity formulation, the Stokes problem can be reduced to a heat equation, possibly vectorial for three-dimensional configurations, with kinematic boundary conditions [8] relying only on tangential components. From a physical point of view, this heat equation takes into account both near-wall adherence properties and viscous effects in the whole fluid.

Lighthill's model states that these kinematic no-slip conditions for a fluid result from vorticity production on solid boundaries [29]. This production of vorticity has been viewed from many perspectives, involving Dirichlet conditions [7, 37] or Neumann conditions [23, 24], usually constrained by Kelvin's theorem to satisfy conservation of circulation. It has been shown that Neumann conditions are well adapted for nonstationary flows for two-dimensional problems [8, 25] or three-dimensional problems in the half-space [9], i.e., without curvature. It has been recently put forward that three-dimensional vortical boundary conditions involve the Robin–Fourier condition [11].

By using linearity of the heat equation, it can be split without any approximation into an equation with a generally nonzero initial condition and homogeneous boundary conditions, and another equation with zero initial condition and Robin–Fourier boundary conditions, which can be written as

$$\begin{cases} \dfrac{\partial\omega}{\partial t}-\nu\Delta\omega=0 \text{ in } \ \Omega\times]0,T[,\\ \omega(x,0)=0 \ \text{ on } \ \Omega,\\ \nu\,\mathcal{L}_x\omega(x,t)=F(x,t) \ \text{ on } \ \partial\Omega\times]0,T[, \end{cases} \tag{1.3}$$

where Ω is an open set of $\mathbb{R}^3$, F is a boundary source, ν is the diffusion coefficient, and $\mathcal{L}$, the differential operator defining Robin–Fourier boundary conditions, can be written as

$$\mathcal{L}_x = \beta(x)\mathrm{Id} + n(x)\cdot\nabla, \tag{1.4}$$

where $n(x)$ denotes the inward normal field to $\partial\Omega$ and β denotes a continuous and bounded function from $\partial\Omega$ to $\mathbb{R}$, bounded by β_0 (i.e., $|\beta(x)| < \beta_0$ for all $x \in \partial\Omega$). Throughout the present paper, Ω is supposed to be an open set such that $\partial\Omega$ is a two-dimensional submanifold of $\mathbb{R}^3$, of class $\mathcal{C}^{2+\lambda}$, and the source F is supposed to be a bounded and continuous function on $\partial\Omega \times [0,T]$. The solution of this heat equation can be found in its integral formulation:

$$\omega(x,t) = \int_0^t \int_{\partial\Omega} G_{\xi,\tau}(x,t)\widetilde{\mu}(\xi,\tau)\mathrm{d}\sigma(\xi)\mathrm{d}\tau, \tag{1.5}$$

where σ is a measure on $\partial\Omega$ induced by the Lebesgue measure and $G_{\xi,\tau}$ is the parametrix [20, 27], which is, in the case of an isotropic heat equation, simply the following three-dimensional Gaussian function:

$$G_{\xi,\tau}(x,t) = \widehat{G}(x-\xi,\nu(t-\tau)) \ \text{ with } \ \widehat{G}(x,\eta) = \frac{e^{-x^2/4\eta}}{(4\pi\eta)^{3/2}}, \tag{1.6}$$

whose standard deviation is $\sqrt{2\nu(t-\tau)}$. The density field $\widetilde{\mu}$ defined on $\partial\Omega$ is the solution of the following Volterra-type integral equation:

$$-\frac{1}{2}\widetilde{\mu}(x,t) + \nu\int_0^t \int_{\partial\Omega} \mathcal{L}_x G_{\xi,\tau}(x,t)\,\widetilde{\mu}(\xi,\tau)\mathrm{d}\sigma(\xi)\mathrm{d}\tau = F(x,t), \tag{1.7}$$

which admits a unique continuous and bounded solution over $\partial\Omega \times [0,T]$ under some minimalistic hypothesis of smoothness, discussed in [14]. Existence, uniqueness, and regularity of solutions of the heat equation and this integral equation have been intensively treated in the literature, many results being summarized in [14] and [28].

Joint equations (1.5)–(1.7) are similar in spirit to (1.1)–(1.2), with both providing a potential aimed at satisfying boundary conditions, and could be named "parabolic panel method." Nevertheless, such a panel method involves fully time dependent densities, which is a function of two variables (instead of one in (1.5)). Moreover, the integrodifferential operator in (1.7) is twice integrated, in time and space. These two remarks make joint equations (1.5)–(1.7) much more difficult to handle than classical panel methods and lead to a much higher degree of computational complexity.

Fast algorithms estimating density $\widetilde{\mu}$ are consequently of fundamental interest in order to make the integral method usable in practice, especially in a three-dimensional context. A way to obtain a fast algorithm is to estimate analytically the density $\widetilde{\mu}$ as a function of the source F, viscosity ν, time t, the coefficients of operator $\mathcal{L}$, and local invariants of $\partial\Omega$ such as its curvature.

Carrying out density estimation from (1.7) has been performed for two-dimensional bodies [24] and for the three-dimensional case of the half-plane [9] (which comes directly via a tensorialization for the two-dimensional case). Nevertheless, the existing literature either considers pure Neumann boundary conditions or simply neglects curvature effects, sometimes involving some hypothesis on nondependency on time (usually not mathematically valid) and in any case not followed by any mathematical analysis on the order of the method.

The present paper provides such an analysis for the more general problem of Robin–Fourier boundary conditions and a class of noncompact domains, in establishing, proving, and illustrating that the early behavior of the density can be explicitly given by the following formula:

$$\widetilde{\mu}(x,t) = \frac{-2F(x,t)}{1 + 2\left(\bar{\kappa}(x) - \beta(x)\right)\sqrt{\nu t/\pi}} + \mathcal{O}(t^{\gamma}), \tag{1.8}$$

where $\bar{\kappa}(x)$ is the mean curvature of $\partial\Omega$ in x, and where γ can reach different values among $]1/2, 3/2]$ in the present study, depending on regularity of $\partial\Omega$ and whether or not $\partial\Omega$ presents torsion. This result can then be used directly as a numerical scheme in formula (1.5), its order being led by the value of γ.

One can notice that Neumann boundary conditions can lead to qualitatively good results at high Reynolds numbers since the relative curvature $\bar{\kappa}\sqrt{\nu}$ tends toward 0. Nevertheless, the Dirichlet part of the Robin–Fourier boundary conditions is linked to the boundary curvatures [11], and is of the same order of the mean curvature effect, as shown in formula (1.8) above. Studying the full Robin–Fourier conditions is consequently of fundamental importance for engineering concerns on viscous flows, especially since the research community finds new interests in micro- and nanotechnologies, which involve small scales where viscous effects are potentially dominant. In this context, neglecting curvature can lead to dramatic errors in numerical simulation of fluids, especially when considering nonstationary dynamics whose prediction requires direct numerical simulation. Moreover, even for macroscopic devices, some new generation Lagrangian schemes such as vortex in cell (VIC) (see [10]) or smooth particle hydrodynamics (SPH) (see [6]) are very stable and can be used with large time steps. Consequently, this enlarges numerical viscous scales, which are of order $\sqrt{\nu\delta t}$ (where δt is the time step), and makes the curvature effects orders of magnitude stronger than standard numerical methods, such as spectral or finite element schemes, whose time scales are limited due to strong stability conditions related to transport terms.

The outline of the paper is as follows. Section 2 provides various preliminary properties. Section 2.1 gives a few well-known properties of fundamental solutions of the heat equation, results more or less already established in the literature. A few conditions are then set in section 2.2 in order to provide a good environment for differential and integral calculus for the following sections. In section 2.3, we show that integral calculus restricted to a local area is an accurate approximation of the global calculus at any order of time. This section also extends a classical result on Hölder continuity of the double heat layer to a class of noncompact manifolds. Section 3 provide convergence results and error estimations of geometrical approximation when the integrodifferential operator is computed on the local quadratic osculating manifold instead of the manifold itself. Section 4 shows that the heat layer of unit density on the best quadratic approximation of the surface can be determined at its main order with error estimation. Finally, section 5 presents the achievement of the present work, and Theorem 5.1 shows that the value of the approximated heat layer obtained in section 5 is a valid value at the first order in time. Theorem 5.3 shows that sufficiently smooth torsionless manifolds allow one to reach order 3/2. The link between these results is displayed in Figure 1. Sections 6 and 7 give several examples illustrating some statements of previous sections, in cylindrical and toroidal geometries, respectively. These examples show that estimates given by Theorems 5.1 and 5.3 are optimal and describe their application to kinematic boundary conditions.

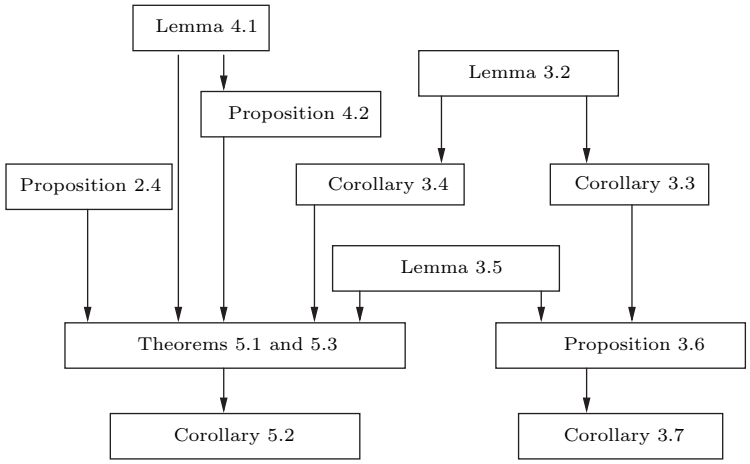


FIG. 1. *Relations between results leading to Theorems* 5.1 *and* 5.3 *(Lemma* 3.1 *and Proposition* 2.3 *are often involved in different proofs).*

2. Preliminary work. This section presents well-known results on parabolic problems in section 2.1 and the hypotheses that are required for the present study in section 2.2. Accuracy of domain restriction is then analyzed in section 2.3, with a specific proof valid for noncompact manifolds.

2.1. Well-known results. One first introduces Friedman's notion of Hölder continuity for parabolic problems: a function ω is said to be ϑ-Hölder continuous over $\Omega \times [0,T]$ if there exist two constants C and ϑ, independent of x, y, t, and s, such that

$$|\omega(x,t) - \omega(y,s)| \leqslant C \left(|x-y|^{\vartheta} + |\nu(t-s)|^{\vartheta/2} \right) \tag{2.1}$$

for all (x,t) and (y,s) in $\overline{\Omega} \times [0,T]$, where $|\cdot|$ denotes both the Euclidean norm of $\mathbb{R}^3$ and the absolute value, depending on the context.

Throughout the paper, $|\cdot|$ will be used for the Euclidean norm in $\mathbb{R}^n$ (absolute value when $n = 1$), $\|\cdot\|$ for norms of functions, and $|||\cdot|||$ for linear operators (not necessarily Euclidean for double and triple norms, but $\mathbb{L}^\infty$ norm most times and $\mathbb{L}^1$ occasionally).

The integral operator is of fundamental interest for the present study, since its value at the leading order is the effect of curvature. It is a double layer of density f with respect to the heat kernel and depends on surface location and time (on $\partial\Omega \times [0,T]$):

$$\widetilde{\mathcal{H}}(x,t)f = \nu \int_0^t \int_{\partial\Omega} \mathcal{L}_x G_{\xi,\tau}(x,t)\, f(\xi,\tau) \mathrm{d}\sigma(\xi) \mathrm{d}\tau \tag{2.2}$$

for any density f bounded and continuous on $\partial\Omega \times [0,T]$.

Furthermore, from [31] and Theorem 4 of Chapter 5 of [14], one gets the following result.

COROLLARY 2.1. *Under the notation above, if f is a continuous and bounded function on $\partial\Omega \times [0,T]$, and $\partial\Omega$ is a compact two-dimensional submanifold of $\mathbb{R}^3$ of class $\mathcal{C}^{1+\lambda}$, then $\widetilde{\mathcal{H}}(x,t)f$, as a function of x and t, is ϑ-Hölder continuous on $\partial\Omega \times [0,T]$ for any exponent satisfying $\vartheta < 2\min(\lambda,1)/3$.*

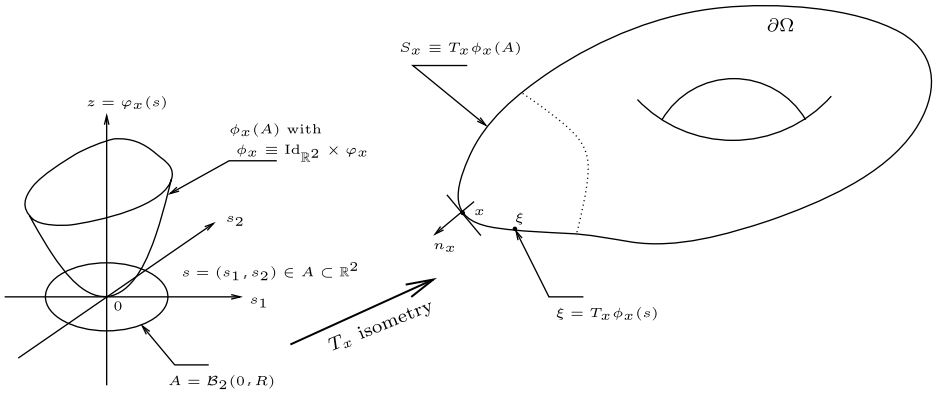


FIG. 2. *Geometrical setup and main notation.*

Nevertheless, for a manifold of class at least $\mathcal{C}^2$, this corollary gives that $\widetilde{\mathcal{H}}(x,t)f$ is ϑ-Hölder continuous for any exponent less than $2/3$. In our context, one can go further than this result. Indeed, Corollary 2.1 is independent of the dimension of space and is valid for manifolds presenting less regularity (i.e., manifolds of class $\mathcal{C}^{1+\lambda}$) than our present considerations. In the present context, set up in section 2.2, the following proposition (more general than Corollary 2.1) holds.

PROPOSITION 2.2. *Let $\widetilde{\mathcal{H}}$ be the integrodifferential operator defined by formula* (2.2). *Under conditions* (C1)–(C5), *with $\beta : \partial\Omega \to \mathbb{R}$ and $f : \partial\Omega \times [0,T] \to \mathbb{R}$ two bounded and, respectively, ϑ_β- and ϑ_f-Hölder continuous functions, we have that $\widetilde{\mathcal{H}}(\cdot,\cdot)f$ is ϑ^*-Hölder continuous with any $\vartheta^* = \min(1-\varepsilon, \vartheta_\beta, \vartheta_f)$ for all $\varepsilon > 0$.*

One can see with this proposition that the maximum Hölder exponent is bounded by the density's, which forbids us from deducting regularity of $\widetilde{\mu}$ from the regularity of $\widetilde{\mathcal{H}}$.

Moreover, Proposition 2.2 shows that the regularity of the density depends on the manifold considered, on the regularity of the source, and on the regularity of the coefficient defining the Dirichlet part of the boundary condition (i.e., the three are involved and arise at the same order in the regularity analysis). The regularity consequently has to be analyzed case by case and will not be discussed in the present paper. To proceed, one can refer to [36] and references therein or [5, 1].

2.2. Geometrical setup and conditions. In order to provide pertinent computations, one assumes that the following condition is satisfied:

(C1) $\partial\Omega$ is a two-dimensional differentiable submanifold of $\mathbb{R}^3$, of class $\mathcal{C}^{2+\lambda}$.

This condition means that $\partial\Omega$ is locally the graph of a function φ_x of class $\mathcal{C}^{2+\lambda}(\Pi_x, \mathbb{R}^3)$, where Π_x is the tangential plan of $\partial\Omega$ in x. One denotes by $T_x : \mathbb{R}^3 \to \mathbb{R}^3$ the affine operator that changes 0 into x and $\mathbb{R}^2 \times \{0\}$ into Π_x.

Consequently, for any $x \in \partial\Omega$, there exists a function φ_x defined in a neighborhood A_x of 0 satisfying $\varphi_x(0) = 0$ and $\nabla\varphi_x(0) = 0$ such that

$$T_x\left(A_x \times \varphi_x(A_x)\right) \subset \partial\Omega.$$

One can notice that the bijective application T_x is a composition of a translation and a rotation; thus its Jacobian is identically equal to 1, and consequently one can integrate indifferently over $\partial\Omega$ or over $T_x^{-1}\partial\Omega$.

This sets up notation for the local description of $\partial\Omega$ around x as a graph of functions φ_x of class $C^{2+\lambda}(\mathbb{R}^2, \mathbb{R})$, as displayed in Figure 2. One recalls that being

of class $C^{m+\lambda}(\mathbb{R}^2,\mathbb{R})$, with $0<\lambda\leqslant 1$, means that φ_x is of class $C^m(\mathbb{R}^2,\mathbb{R})$ and all its mth order partial derivatives are λ-Hölder continuous. This means that for any $x\in\partial\Omega$, there exist two constants C_x and C'_x such that locally one has

$$\left|\varphi_x(s)-\frac{1}{2}\,{}^t sK_x s\right| \;\leqslant\; C_x\,|s|^{2+\lambda} \quad\text{and}\quad \left|\nabla\varphi_x(s)-K_x s\right| \;\leqslant\; C'_x\,|s|^{1+\lambda}, \tag{2.3}$$

where K_x is the Hessian matrix of φ_x in $0=T_x^{-1}(x)$.

One then requires the following conditions on the globality of bounds defined above:

(C2) There exists $R>0$ such that domain definition A_x of applications φ_x contains $\mathcal{B}_2(0,R)$ for all $x\in\partial\Omega$.

(C3) The spectral radius of the Hessian K_x of ϕ_x in 0 is bounded independently of $x\in\partial\Omega$, and its upper bound is denoted ρ_0.

(C4) There exist two constants C and C' such that for any $x\in\partial\Omega$, one has $C_x\leqslant C$ and $C'_x\leqslant C'$.

Note that these hypotheses do not imply a lack of generality but provide some restrictions on smoothness (condition (C1)) and mapping orientation and size (conditions (C2) and (C4)). One can also notice that $\partial\Omega$ being the boundary of an open set Ω and condition (C1) imply that $\partial\Omega$ is an oriented manifold and thus the existence of an inward normal field n to $\partial\Omega$. Condition (C3) eliminates, for example, spiraloids or clothoidal surfaces whose curvature tends to infinity.

For convenience, one sets up the notation $\phi_x=\mathrm{Id}_{\mathbb{R}^2}\times\varphi_x$, which gives that

$$S_x\equiv T_x\phi_x\left(\mathcal{B}_2(0,R)\right)\subset\partial\Omega. \tag{2.4}$$

This local parameterization of $\partial\Omega$ gives the following integration formula:

$$\int_{S_x} f(\xi)\mathrm{d}\sigma(\xi)=\int_{\mathcal{B}_2(0,R)} f\circ T_x\circ\phi_x(s)\,|N_x(s)|\,\mathrm{d}s, \tag{2.5}$$

where

$$N_x(s)=\frac{\partial\phi_x}{s_1}(s)\wedge\frac{\partial\phi_x}{s_2}(s)$$

is the normal field induced by the parameterization ϕ_x, its Euclidean norm being the Jacobian of the parameterization and satisfying

$$|N_x(s)|=\sqrt{1+|\nabla\varphi_x(s)|}, \tag{2.6}$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^3$. Note that T_x is not involved in this Jacobian because $|\partial_{s_1}T_x\circ\phi_x\wedge\partial_{s_1}T_x\circ\phi_x|=|\partial_{s_1}\phi_x\wedge\partial_{s_1}\phi_x|=|N_x(s)|$.

A few direct consequences of conditions (C1)–(C4) can then be stated.

PROPOSITION 2.3. *Under conditions* (C1)–(C4), *one has the following:*

(P1) *N_x, the Jacobian of the parameterization by ϕ_x, is bounded independently of x, and its bound is denoted M_1.*

(P2) *There exists a constant M_2 such that*

$$|(x-\xi)\cdot n_x|\leqslant M_2|x-\xi|^2\quad\forall x,\xi\in\partial\Omega. \tag{2.7}$$

Proof. Let x be a point of $\partial\Omega$. By (2.3), one gets

$$|\nabla\varphi_x(s)| \leqslant |K_x s| + C'_x|s|^{1+\lambda} \leqslant \rho_0 R + C' R^{1+\lambda}.$$

Setting $M_1 = \sqrt{1 + (\rho_0 + C'R^\lambda)^2 R^2}$ finishes the proof of assertion (P1).

Furthermore, for all ξ in $\partial\Omega \cap \mathcal{B}_3(x,R) \subset T_x\phi_x(\mathcal{B}_2(0,R))$ where R is defined by condition (C2), one has $\xi = T_x\phi_x(s)$ with $\phi_x = \mathrm{Id}_{\mathbb{R}^2} \times \varphi_x$, and $T_x^{-1}n(x) = -e_3$. Consequently, one has $(\xi - x)\cdot n(x) = \varphi_x(s) - \varphi_x(0) = \varphi_x(s)$ and by condition (C3) and by definition of class $\mathcal{C}^{2+\lambda}$, one gets

$$|(\xi - x)\cdot n(x)| = |\varphi_x(s)| \leqslant \frac{\rho_0}{2}|s|^2 + C|s|^{2+\lambda} \leqslant \left(\frac{\rho_0}{2} + C\,R^\lambda\right)|s|^2.$$

For any ξ not in $\partial\Omega \cap \mathcal{B}_3(x,R)$, one has $|(\xi - x)\cdot n_x| \leqslant |\xi - x| \leqslant (\xi - x)^2/R$. Setting $M_2 = \max(C\,R^\lambda + \rho_0/2, 1/R)$ finishes the proof. □

Part (P1) of the proposition will often be useful for calculus through the maps, while part (P2) implies that the operator $\widetilde{\mathcal{H}}(x,t)f$ is bounded over $\partial\Omega \times [0,T]$, which in itself is useful for showing that the $\widetilde{\mathcal{H}}$ is Hölder continuous.

Moreover, one needs a last condition of measure growth of $\partial\Omega$:

(C5) There exist two positive constants C and k such that for any $x \in \partial\Omega$, the measure of the part of $\partial\Omega$ in the spherical strips $\mathcal{B}_3(x,(n+1)R) \setminus \mathcal{B}_3(x,nR)$ does not grow faster than $e^{k\,n^2}$; that is,

$$\sigma\left(\left\{\xi \in \partial\Omega \;/\; nR \leqslant |\xi - x| < (n+1)R\right\}\right) \leqslant C\,e^{k\,n^2} \quad \forall x \in \partial\Omega. \tag{2.8}$$

Condition (C5) is satisfied as soon as $\partial\Omega$ is compact, and provides useful majorations and controllable error estimates of map restriction when $\partial\Omega$ is not compact.

As a concluding note on conditions (C1)–(C5), let us remark that they are not very restrictive and do not lead to a lack of generality on the kind of surfaces considered. Their most significant effect is that they limit size and orientation of the maps and provide a good environment for integral calculus in the next sections.

2.3. Error estimation of the restriction to a map. The definition of $\widetilde{\mathcal{H}}$ involves a Gaussian function whose standard deviation tends to zero when t tends to zero (smaller than $\sqrt{2\nu t}$). Even if a Gaussian is not compactly supported, it decreases quickly, and its significant values are very localized. Consequently, thanks to limiting the final time T of the heat equation (which is not a limitation in practice), we can consider the integral over one map,

$$S_x \equiv T_x\phi_x(\mathcal{B}_2(0,R)) \subset \partial\Omega \subset \mathbb{R}^3, \tag{2.9}$$

instead of the whole surface $\partial\Omega$, and provide an error analysis thanks to the fast decreasing of Gaussian functions. One can introduce the heat layer restricted to $S_x \subset \partial\Omega$ as

$$\widetilde{\mathcal{H}}_{S_x}(x,t)f = \nu \int_0^t \int_{S_x} \mathcal{L}_x G_{\xi,\tau}(x,t)\, f(\xi,\tau)\mathrm{d}\sigma(\xi)\mathrm{d}\tau$$

for any continuous and bounded function $f : \partial\Omega \times [0,T] \to \mathbb{R}$, and the error due to the restriction on $S_x \subset \partial\Omega$ as

$$\widetilde{\mathcal{H}}_{_{err}}(x,t)f = \left(\widetilde{\mathcal{H}}(x,t) - \widetilde{\mathcal{H}}_{S_x}(x,t)\right) f = \nu \int_0^t \int_{\partial\Omega\setminus S_x} \mathcal{L}_x G_{\xi,\tau}(x,t)\, f(\xi,\tau)\mathrm{d}\sigma(\xi)\mathrm{d}\tau.$$

By setting $\eta = \nu(t-\tau)$, one gets

$$\frac{|\widetilde{\mathcal{H}}_{err}(x,t)f|}{\|f\|_\infty} \leqslant \int_0^{\nu t} \int_{\partial\Omega\setminus S_x} \left|\beta(x) - \frac{(x-\xi)\cdot n(x)}{2\eta}\right| \widehat{G}(x-\xi,\eta)\mathrm{d}\sigma(\xi)\mathrm{d}\eta.$$

One can now show that the error due to the restriction to a map can then be neglected at any order. The following integral calculus features the parametric approach of the present paper.

First, by condition (C2) there exists R, such that for all $x \in \partial\Omega$, one has $s = T_x^{-1}(\xi)$ for all ξ in the neighborhood S_x of x and the inequality

$$|s|^2 \leqslant |s|^2 + \varphi_x(s)^2 = |T_x\phi_x(s) - T_x(0)|^2 = |\xi - x|^2 < R^2$$

and, consequently, the inclusions

$$\partial\Omega \cap \mathcal{B}_3(x,R) \subset T_x\phi_x(\mathcal{B}_2(0,R)) = S_x \subset \partial\Omega$$

which imply directly

$$\partial\Omega \setminus S_x = \partial\Omega \setminus T_x\phi_x(\mathcal{B}_2(0,R)) \subset \partial\Omega \setminus \mathcal{B}_3(x,R) \subset \partial\Omega. \tag{2.10}$$

Error estimate $|\widetilde{\mathcal{H}}_{err}(x,t)f|$ is thus majorated by

$$\|f\|_\infty \int_0^{\nu t} \int_{\partial\Omega\setminus\mathcal{B}_3(x,R)} \left|\beta(x) - \frac{(x-\xi)\cdot n(x)}{2\eta}\right| \widehat{G}(x-\xi,\eta)\mathrm{d}\sigma(\xi)\mathrm{d}\eta, \tag{2.11}$$

which implies, due to part (P2) of Proposition 2.3, the following majoration:

$$\|f\|_\infty \int_0^{\nu t} \left(|\beta(x)| + M_2\frac{|x-\xi|^2}{2\eta}\right) \int_{\partial\Omega\setminus\mathcal{B}_3(x,R)} \widehat{G}(x-\xi,\eta)\mathrm{d}\sigma(\xi)\mathrm{d}\eta. \tag{2.12}$$

Second, and this notation will be used throughout the paper, one has for all positive x and α,

$$x^\alpha e^{-x} \leqslant L_\alpha e^{-x/2} \quad \text{with} \quad L_\alpha = (2\alpha)^\alpha e^{-\alpha}. \tag{2.13}$$

It follows from (2.13) that expression (2.12) is majorated by

$$\|f\|_\infty (\beta_0 + L_1M_2) \int_0^{\nu t} \int_{\partial\Omega\setminus\mathcal{B}_3(x,R)} \frac{e^{-|x-\xi|^2/8\eta}}{(4\pi\eta)^{3/2}}\, \mathrm{d}\sigma(\xi)\, \mathrm{d}\eta. \tag{2.14}$$

If $\partial\Omega$ is compact, then (2.14) is majorated by

$$\|f\|_\infty (\beta_0 + L_1M_2)\, \sigma(\partial\Omega) \int_0^{\nu t} \frac{e^{-R^2/8\eta}}{(4\pi\eta)^{3/2}}\, \mathrm{d}\eta, \tag{2.15}$$

which is finally majorated by

$$\|f\|_\infty 2\sqrt{2}\, L_{3/2}\, \frac{\beta_0 + M_2L_1}{R^3\pi^{3/2}}\, e^{-R^2/16\nu t}\, \nu t\, \sigma(\partial\Omega). \tag{2.16}$$

If $\partial\Omega$ is not compact, there is more work to be done, and one has to consider condition (C5) on maximal admissible growth of measures of $\partial\Omega$ contained in successive spherical strips, which leads to the same result. Indeed, one has the disjoint decomposition

$$\partial\Omega = (\partial\Omega \cap \mathcal{B}_3(x,R)) \cup \left[\bigcup_{n\in\mathbb{N}^*} (\partial\Omega \cap (\mathcal{B}_3(0,(n+1)R) \setminus \mathcal{B}_3(0,nR))) \right];$$

thus

$$\partial\Omega \setminus \mathcal{B}_3(x,R) = \bigcup_{n\in\mathbb{N}^*} (\partial\Omega \cap (\mathcal{B}_3(0,(n+1)R) \setminus \mathcal{B}_3(0,nR)))\,,$$

and one can build a majoration of (2.14) by the use of condition (C5):

$$\|f\|_\infty C \int_0^{\nu t} (\beta_0 + L_1 M_2) \left(\sum_{j\in\mathbb{N}^*} \frac{e^{-j^2R^2/8\eta}}{8\pi^{3/2}\eta^{3/2}}\, e^{j^2 k} \right) \mathrm{d}\eta \tag{2.17}$$

with C depending only on $\partial\Omega$. Equation (2.17) is itself majorated for νt sufficiently small, i.e., for $\nu t \leqslant R^2/8k$, where k depends only on $\partial\Omega$, by

$$\|f\|_\infty C\ (\beta_0 + L_1 M_2) \int_0^{\nu t} \frac{1}{8\pi^{3/2}\eta^{3/2}} \left(\sum_{j\in\mathbb{N}^*} \exp\left\{ -j^2 \left(\frac{R^2}{8\eta} - k \right) \right\} \right) \mathrm{d}\eta. \tag{2.18}$$

Now noticing that $j \leqslant j^2$ and that $k - R^2/8\eta \leqslant -R^2/16\eta$ for $\nu t \leqslant R^2/16k$, one gets another majoration of (2.18) by

$$\|f\|_\infty C\ (\beta_0 + L_1 M_2) \int_0^{\nu t} \frac{1}{8\pi^{3/2}\eta^{3/2}} \frac{1}{1 - e^{-R^2/16\eta}}\, e^{-R^2/16\eta}\ \mathrm{d}\eta. \tag{2.19}$$

Noticing also that $e^{-R^2/16\eta} \leqslant 1/2$ for $\eta \leqslant R^2/16\ln 2$, one gets

$$\|f\|_\infty C\ (\beta_0 + L_1 M_2) \int_0^{\nu t} \frac{e^{-R^2/16\eta}}{4\pi^{3/2}\eta^{3/2}}\ \mathrm{d}\eta, \tag{2.20}$$

which gives the final majoration for $\nu t \leqslant R^2/16k^a st$ with $k^* = \max(\ln 2,\, k)$, i.e., for t sufficiently small:

$$\|f\|_\infty 16\, C\, L_{3/2}\, \frac{\beta_0 + L_1 M_2}{\pi^{3/2} R^3}\, e^{-R^2/32\nu t} \nu t. \tag{2.21}$$

Equation (2.16) holds when $\partial\Omega$ is a compact manifold, and is extended to the noncompact case by means of (2.21). From these equations, one gets the following proposition.

PROPOSITION 2.4. *Under conditions* (C1)–(C5) *and previous notation, there exists a constant C independent of x and t such that*

$$\frac{|\widetilde{\mathcal{H}}_{err,\mathcal{B}_2(\zeta,R)}(x,t)f|}{\|f\|_\infty} \leqslant \frac{|\widetilde{\mathcal{H}}_{err,A}(x,t)f|}{\|f\|_\infty} \leqslant C\,\|f\|_\infty\, e^{-R^2/32\nu t} \nu t = \mathcal{O}(t^\infty) \tag{2.22}$$

for t sufficiently small and for all $x \in \partial\Omega$ and any $f \in \mathbb{L}^\infty(\partial\Omega)$.

$$
\begin{array}{ccc}
-\frac{1}{2}\widetilde{\mu}(x,t)+\widetilde{\mathcal{H}}(x,t)\widetilde{\mu}=F(x,t) & \xrightarrow{\text{param.}} & \mu_x^\flat(s,t)=\widetilde{\mu}(T_x\phi_x(s),t) \\
\Big\downarrow \text{restr. on } S_x & & \\
-\frac{1}{2}\mu^\natural(x,t)+\widetilde{\mathcal{H}}_{S_x}(x,t)\mu^\natural=F(x,t) & \xrightarrow{\text{param.}} & \mu_x(s,t)=\mu^\natural(T_x\phi_x(s),t) \\
\Updownarrow & & \\
-\frac{1}{2}\mu_x(0,t)+\mathcal{H}(x,t)\mu_x=F(x,t) & & \\
\Big\downarrow \text{approx. on } \overline{S_x} & & \\
-\frac{1}{2}\overline{\mu}_x(0,t)+\overline{\mathcal{H}}(x,t)\overline{\mu}_x=F(x,t) & \xrightarrow{\text{param.}} & \overline{\mu}_x(s,t)=\mu^\sharp(T_x\phi_x(s),t).
\end{array}
$$

FIG. 3. *Different integrodifferential operators and associated integral equations and densities involved in the surface potential analysis.*

3. Geometrical approximation. It has been shown in section 2.3 that only a local analysis of the heat layer is required to obtain the development in early time of the heat layer at any order, since the Gaussian kernel has significant values very locally.

The local parameterizations of the manifold and its quadratic osculating manifold are set up in sections 3.1 and 3.2, respectively, and we give a few approximation lemmas in section 3.3 that will be useful in proving convergence in section 3.4 and providing error estimates of truncations to get the final results.

The process of parameterization and approximation can be split into two different steps as shown in Figure 3, which should be read as follows. First one considers the original integral equation of solution $\widetilde{\mu}$, whose parameterization is denoted $\mu_x^\flat$ in a neighborhood of origin whose image is a neighborhood $S_x \subset \partial\Omega$ of x.

The map-restricted heat layer $\widetilde{\mathcal{H}}_{S_x}$ is introduced in section 3.1 and defines a new integral equation whose solution is denoted $\mu^\natural$ and its parameterization μ_x, which is itself the solution of an integral equation involving $\mathcal{H}$, the restricted heat layer acting on parameterizations, i.e., such that

$$\widetilde{\mathcal{H}}_{S_x}(x,t)\mu^\natural=\mathcal{H}(x,t)\mu_x.$$

The portion of surface S_x is then approximated by its best quadratic approximant $\overline{S_x}$ in section 3.2, which induces a heat layer on the surface approximant denoted $\overline{\mathcal{H}}$, and another integral equation involving this operator acting on parameterizations, whose solution is denoted $\overline{\mu}_x : A \to \mathbb{R}$. This solution corresponds to a density $\mu^\sharp : \partial\Omega \to \mathbb{R}$.

While the relation between densities $\widetilde{\mu}$ and $\mu^\natural$ is quite obvious by means of proposition 2.4, the link between $\mu^\natural$ and $\mu^\sharp$ is less obvious and requires an analysis of approximation relations between the associated integral operators. The convergence of densities is established in section 3.4, especially in Proposition 3.6.

3.1. Local parameterization of the surface. The two-dimensional manifold $\partial\Omega$ is described by its maps, which are themselves defined by means of the description of $\partial\Omega$ by graphs of functions φ_x, given in (2.4), which reads as follows:

$$T_x\phi_x \equiv T_x(\mathrm{Id}_{\mathbb{R}^2}\times\varphi_x) : A=\mathcal{B}(0,R)\subset\mathbb{R}^2 \longrightarrow T_x\phi_x(A)\equiv S_x\subset\partial\Omega. \tag{3.1}$$

A density $\widetilde{f} : \partial\Omega \to \mathbb{R}$ can be restricted on a map as

$$\begin{array}{rccc} \widetilde{f}: & S_x \subset \partial\Omega \subset \mathbb{R}^3 \times [0,T] & \longrightarrow & \mathbb{R} \times [0,T], \\ & (\xi, t) & \longmapsto & \widetilde{f}(\xi, t). \end{array} \tag{3.2}$$

One can introduce the local parameterization f in space of $\widetilde{f}$ through $T_x\phi_x$ in a neighborhood S_x of x in $\partial\Omega$, so that $\xi = T_x\phi_x(s)$, which gives

$$\begin{array}{rccl} f: & A \times [0,T] & \longrightarrow & \mathbb{R} \times [0,T], \\ & (s,t) & \longmapsto f(s,t) & = \widetilde{f}(T_x \circ \phi_x(s), t) \\ & & & = \widetilde{f}\left(T_x\left(s_1, s_2, \varphi_x(s)\right), t\right), \end{array} \tag{3.3}$$

where $A \equiv \mathcal{B}_2(0,R) \subset \mathbb{R}^2$.

One also considers the double heat layer restricted to a map acting on parameterizations of densities, defined consequently by

$$\mathcal{H}(x,t)f = \widetilde{\mathcal{H}}_{S_x}(x,t)\widetilde{f} = \nu \int_0^t \int_A \mathcal{L}_x G_{T_x\phi_x(s),\tau}(x,t) f(s,\tau) |N_x(s)| \mathrm{d}s\mathrm{d}\tau, \tag{3.4}$$

where

$$N_x(s) = \frac{\partial \phi_x}{s_1}(s) \wedge \frac{\partial \phi_x}{s_2}(s)$$

and $|N(s)|$ denotes its Euclidean norm in $\mathbb{R}^3$. This Jacobian calculus holds since the Jacobian of T_x is identically equal to 1 and gives

$$|N_x(s)| = \sqrt{1 + |\nabla\varphi_x(s)|^2}, \tag{3.5}$$

which is bounded with respect to x over $\partial\Omega$ by Proposition 2.3(P1). One can notice that the Gaussian functions are spherically symmetric; thus the equalities

$$\begin{aligned} G_{T_x\phi_x(s),\tau}(x,t) &= G_{\phi_x(s),\tau}(T_x^{-1}x, t) = G_{\phi_x(s),\tau}(0,t) \\ &= G_{0,0}(\phi_x(s), t-\tau) = \widehat{G}\left(\phi_x(s), \nu(t-\tau)\right) \\ &= \frac{e^{-(s^2+\varphi_x(s)^2)/4\nu(t-\tau)}}{(4\pi\nu(t-\tau))^{3/2}} \end{aligned}$$

hold because $\phi_x = \mathrm{Id}_{\mathbb{R}^2} \times \varphi_x$.

Since one has $\varphi_x(0) = 0$ and $\nabla\varphi_x(0) = 0$, this leads, by means of (2.3), to

$$\varphi_x(s) = \frac{1}{2}\, {}^t s\, K_x\, s\, + \mathcal{O}(s^{2+\lambda}), \tag{3.6}$$

where K_x is the symmetric 2×2 curvature matrix of $\partial\Omega$ in x, in the spirit of the two-dimensional approach of [24].

3.2. Local approximation of the surface. In order to provide a local analysis on an explicitly known surface, and to estimate the error related to geometrical approximation, one considers the second order approximation of $\partial\Omega$ in a neighborhood S_x of $x \in \partial\Omega$. This means we introduce the quadratic form

$$\overline{\varphi}_x(s) = \frac{1}{2}\, {}^t s\, K_x\, s$$

and all the related quantities, as displayed in Figure 4:

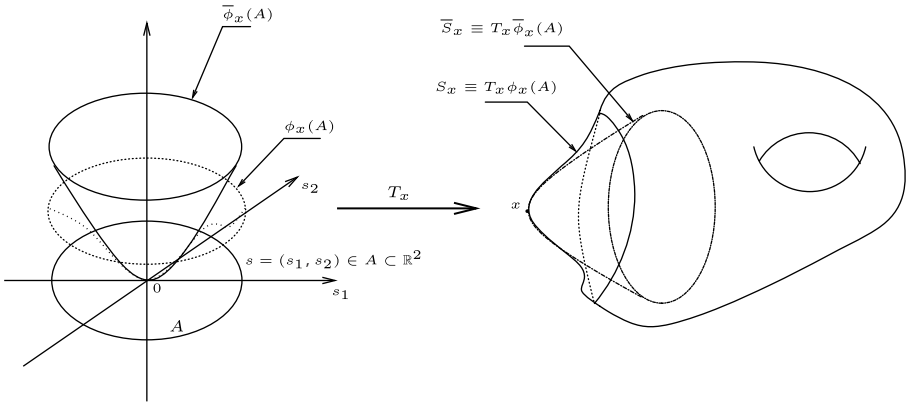


FIG. 4. *Context and notation of surface approximation.*

- $\overline{\phi}_x = T_x \circ (\mathrm{Id}_{\mathbb{R}^2} \times \overline{\varphi}_x)$,
- $\overline{S_x} = T_x\overline{\phi}_x(A)$, the quadratic approximant of S_x,
- $\overline{\mathcal{H}}(x,t)$, the integral operator defined over $\overline{S_x}$,
- $\overline{N}_x(s) = \partial_{s_1}\overline{\phi}_x \wedge \partial_{s_2}\overline{\phi}_x$, the Jacobian of this parameterization, which verifies

$$|\overline{N}_x(s)| = \sqrt{1 + |\nabla\varphi_x(s)|^2} = \sqrt{1 + |K_x s|^2}, \tag{3.7}$$

- $\overline{\mu}_x : A \to \mathbb{R}$, the parameterized solution of the integral equation

$$-\overline{\mu}_x(0,t) + 2\overline{\mathcal{H}}(x,t)\overline{\mu}_x = F(x,t), \tag{3.8}$$

  which defines the density $\mu^\sharp : \partial\Omega \to \mathbb{R}$ by $\mu^\sharp(x,t) = \overline{\mu}_x(0,t)$.

Furthermore, the boundary operator $\mathcal{L}_x$ defining the Robin–Fourier condition can be written as $\beta(x)\mathrm{Id} + n(x)\cdot\nabla$. The gradient of the Gaussian kernel satisfies

$$\nabla_x G_{\xi,\tau}(x,t) = -\frac{x-\xi}{2\nu(t-\tau)} G_{\xi,\tau}(x,t);$$

thus

$$\mathcal{L}_x G_{\xi,\tau}(x,t) = \left(\beta(x) - \frac{(x-\xi)\cdot n(x)}{2\nu(t-\tau)}\right) G_{\xi,\tau}(x,t). \tag{3.9}$$

Moreover, the gradient $\nabla\varphi_x(0) = 0$ implies $T_x^{-1}n(x) = -e_3$, and since the scalar product is conserved by rotation/translation (i.e., by operator T_x), one gets

$$\begin{aligned} -(x-\xi)\cdot n(x) &= (T_x\phi_x(s) - T_x\phi_x(0))\cdot n(x) \\ &= (\mathrm{Id}_{\mathbb{R}^2} \times \varphi_x)(s)\cdot(-be_3) = -\varphi_x(s) \end{aligned}$$

for all $\xi = T_x\phi_x(s) = T_x(\mathrm{Id}_{\mathbb{R}^2} \times \varphi_x)(s) \in S_x$, i.e., for all $s \in A = \mathcal{B}_2(0,R)$.

For a function $f : A \to \mathbb{R}$, one has consequently

$$\mathcal{H}(x,t)f = \tag{3.10}$$

$$\nu\int_0^t\int_A \left(\beta(x) - \frac{\varphi_x(s)}{2\nu(t-\tau)}\right) G_{0,\tau}\left(\phi_x(s),t\right) f(T_x\phi_x(s),\tau)\ \sqrt{1+|\nabla\varphi_x(s)|^2}\ \mathrm{d}s\mathrm{d}\tau,$$

and similarly, one gets

(3.11)

$$\overline{\mathcal{H}}(x,t)f = \nu \int_0^t \int_A \left(\beta(x) - \frac{\overline{\varphi}_x(s)}{2\nu(t-\tau)}\right) G_{0,\tau}\left(\overline{\phi}_x(s),t\right) f(T_x\phi_x(s),\tau)\ \sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}\ \mathrm{d}s\mathrm{d}\tau.$$

For practical considerations in the estimation of the integrals defined above and their related quantities, one will often need a purely computational result, which is given to increase readability.

LEMMA 3.1 (integral majoration). *There exists a constant C_0 independent of t such that for $R \geqslant 0$ and $k > 0$, the majoration*

$$\Gamma^{R,k}_{\alpha,\eta}(\tau) = \int_0^\tau \int_R^\infty \frac{r^\alpha + u^{\alpha/2}}{u^\eta} e^{-r^2/ku}\, r\, \mathrm{d}r\, \mathrm{d}u \leqslant C_0\, e^{-R^2/2k\tau}\, \tau^{2-\eta+\alpha/2}$$

holds when $\eta - \alpha/2 < 2$, i.e., when the integral is convergent.

Consequently, $\Gamma^{R,k}_{\alpha,\eta}(\tau) = \mathcal{O}(\tau^{2-\eta+\alpha/2})$ if $R = 0$, and $\Gamma^{R,k}_{\alpha,\eta}(\tau) = \mathcal{O}(\tau^\infty)$ otherwise.

Proof. One has

$$r\, \frac{r^\alpha + u^{\alpha/2}}{u^\eta} e^{-r^2/ku} \leqslant C\, r\, u^{\alpha/2-\eta} e^{-r^2/2ku}$$

with $C = 1 + L_{\alpha/2} k^{\alpha/2}$, and consequently

$$\begin{aligned}\int_R^\infty \frac{r^\alpha + u^{\alpha/2}}{u^\eta} e^{-r^2/ku}\, r\, \mathrm{d}r &\leqslant C\, k\, u^{\alpha/2+1-\eta} \int_R^\infty \frac{2r}{2ku} e^{-r^2/2ku}\, \mathrm{d}r \\ &= C\, k\, u^{\alpha/2+1-\eta} e^{-R^2/2ku}.\end{aligned}$$

Integral (3.1) is then majorated by

$$\Gamma^{R,k}_{\alpha,\eta}(\tau) \leqslant\ C\, k \int_0^\tau u^{\alpha/2+1-\eta} e^{-R^2/2ku} \mathrm{d}u \leqslant\ C_0\, e^{-R^2/2k\tau}\, \tau^{2-\eta+\alpha/2}$$

with

$$C_0 = \frac{k\, C}{2-\eta+\alpha/2} > 0.$$

The comment on order is then obvious. ∎

The question now is how great an error is made in the integral operator when the surface $\partial\Omega$ is approximated by its best osculating quadratic surface, and how the error on the integral operator is related to the error on the solution μ_x.

3.3. Approximation lemma. In this section we exhibit in Lemma 3.5 the convergence rate of $\overline{\mathcal{H}}$ toward $\mathcal{H}$ in time variable when the surface $\partial\Omega$ is replaced by its best quadratic approximation. This lemma requires preliminary results in Lemmas 3.2 and 3.1. The result of Lemma 3.5 is used in the next sections and also implies the convergence of $\overline{\mu}_x$ toward μ_x with the same rate, as explained in Proposition 3.6.

One can define the usual norms considered herein. The simple norm $|\cdot|$ denotes the Euclidean norm of $\mathbb{R}^3$, and the maximum double norm denotes the usual $\mathbb{L}^\infty$ norm over $\partial\Omega$ or $\partial\Omega \times [0,T]$, depending on the context.

The triple norm of linear operator applies to the integrodifferential operators defined above. Indeed, for any $(x,t) \in \partial\Omega \times [0,T]$, $\mathcal{H}(x,t)$ is a linear operator that verifies $|\mathcal{H}(x,t)f| \leqslant \|f\|_\infty |\mathcal{H}(x,t)1|$ for any bounded function f, with equality for f identically equal to 1. One then gets

$$|||\mathcal{H}(x,t)|||_\infty = \sup_{f \in \mathbb{L}^\infty(A)\setminus\{0\}} \frac{|\mathcal{H}(x,t)f|}{\|f\|_\infty} = \sup_{\|f\|_\infty = 1} |\mathcal{H}(x,t)f| = |\mathcal{H}(x,t)1|. \tag{3.12}$$

The triple norm $|||\overline{\mathcal{H}}(x,t)|||_\infty$ is itself a function of x and t, whose $\mathbb{L}^\infty$ norm over $\partial\Omega$ is defined as

$$||||\mathcal{H}(\cdot,t)||||_\infty = ||\ x \mapsto |||\mathcal{H}(x,t)|||_\infty\ ||_\infty = \sup_{x\in\partial\Omega} |||\overline{\mathcal{H}}(x,t)|||_\infty = \sup_{x\in\partial\Omega} |\mathcal{H}(x,t)1|. \tag{3.13}$$

LEMMA 3.2. *Let $f_t : A\times[0,T] \to \mathbb{R}$ be a function, possibly depending on $t \in [0,T]$, such that there exist two constants $\alpha \geqslant 0$ and C_f independent of t satisfying*

$$|f_t(s,\tau)| \leqslant C_f \left(|s|^\alpha + |\nu(t-\tau)|^{\alpha/2} \right)$$

for all $(s,\tau) \in A\times[0,t]$. Under the hypothesis (C1)–(C4) *of section* 2.2, *with $\mathcal{H}$ defined by formula* (3.10), *there exists a constant C_0 such that for any $x \in \partial\Omega$, one has*

$$\mathcal{H}(x,t)f_t = C_0\, t^{(\alpha+1)/2}.$$

Proof. The portion of surface S_x is the image by the isometry T_x of the graph of φ_x over the two-dimensional ball $A = \mathcal{B}_2(0,R)$, with

$$\varphi_x(s) = \frac{1}{2}\, {}^t s\, K_x\, s\ + \mathcal{O}(s^{2+\lambda})$$

since $\partial\Omega$ is a manifold of class $\mathcal{C}^{2+\lambda}$. One has, through the map $T_x\phi_x$,

$$\mathcal{H}(x,t)f_t = \nu \int_0^t \int_A \left(\beta(x) - \frac{\varphi_x(s)}{2\nu(t-\tau)} \right) G_{0,\tau}\left(\phi_x(s),t\right) f(s,\tau)\, |N_x(s)|\, \mathrm{d}s\mathrm{d}\tau \tag{3.14}$$

with

$$|N_x(s)| = \sqrt{1+|\nabla\varphi_x(s)|^2} \leqslant M_1$$

and

$$|\varphi_x(s)| \leqslant \left| \frac{1}{2}\, {}^t s\, K_x\, s \right| + C\,|s|^{2+\lambda} \leqslant \left(\frac{\rho_0}{2} + C\,R^\lambda \right)\, s^2, \tag{3.15}$$

where M_1, ρ_0, and C are the constants introduced, respectively, in Proposition 2.3(P1), condition (C3), and condition (C4).

One gets consequently, assuming $C_1 = (\rho_0/2 + C\,R^\lambda)$ and $u = \nu t$,

$$|\mathcal{H}(x,t)f_t| \leqslant$$
$$M_1 \int_0^{\nu t} \int_A \left(\beta_0 + C_1 \frac{|s|^2}{2u} \right) \frac{1}{(4\pi u)^{3/2}} \exp\left(-\frac{|s|^2 + ({}^t s K_x s)^2}{4u} \right)\, f_t(s,\tau)\, \mathrm{d}s\mathrm{d}\tau. \tag{3.16}$$

Now noticing that $({}^t sKs)^2 \geqslant 0$, assuming $r = |s|$, and applying the Hölder-like hypothesis on f_t, one has an axisymmetric expression that integrates into

$$|\mathcal{H}(x,t)f_t| \leqslant \frac{2\,C_f\,M_1}{\sqrt{\pi}} \int_0^{\nu t} \int_0^R \left(\beta_0 + C_1 \frac{r^2}{2u}\right) \frac{r^\alpha + u^{\alpha/2}}{(4u)^{3/2}}\, e^{-r^2/4u}\, r\,\mathrm{d}r\mathrm{d}\tau. \tag{3.17}$$

Using relation (2.13) that gives $x\,e^{-x} \leqslant L_1 e^{-x/2}$, the result follows from Lemma 3.1:

$$|\mathcal{H}(x,t)f_t| \leqslant \frac{C_f\,M_1}{4\sqrt{\pi}}\,(\beta_0 + 2L_1C_1)\,\Gamma^{0,8}_{\alpha,3/2}(t) = \mathcal{O}(t^{1/2+\alpha/2}).$$

Note that $C_0 = C_f\,M_1\,(\beta_0 + 2L_1C_1)\,/4\sqrt{\pi}$ depends neither on $x \in \partial\Omega$ nor on $t \in [0,T]$. ∎

One first has to notice that this result holds when $\alpha = 0$, which provides a useful majoration of $\mathcal{H}(x,t)f$ when f is only bounded, that is to say, majoration of $\mathcal{H}(x,t)1$, and directly gives that

$$|||\mathcal{H}(x,t)|||_\infty = O(t^{1/2}) \tag{3.18}$$

for any $(x,t) \in \partial\Omega \times [0,T]$.

One can also notice that the final majorant in the proof is not dependent on x, and applying this again to the unity function, one can extend result (3.18) into

$$||||\mathcal{H}(\cdot,t)||||_\infty = ||\, x \mapsto |||\mathcal{H}(x,t)|||_\infty\,||_\infty = O(t^{1/2}). \tag{3.19}$$

Another important result is that the lemma also holds for $\overline{\mathcal{H}}(x,t)$, which is a special case with the function $\varphi_x = \overline{\varphi}_x$ of class $\mathcal{C}^\infty$ so a fortiori $\mathcal{C}^3$, with a constant C set to 0 in (3.15). This gives that there exists a constant C such that

$$||||\overline{\mathcal{H}}(\cdot,t)||||_\infty \leqslant C\,t^{1/2} \tag{3.20}$$

and consequently the following corollary.

COROLLARY 3.3. *Let $\overline{\mathcal{H}}$ be the operator defined by formula* (3.11). *Then, under the hypothesis of Lemma* 3.2, *for t sufficiently small, one has $||||\overline{\mathcal{H}}(\cdot,t)||||_\infty < 1/2$.*

From Lemma 3.2, one also has the following property valid on $S_x \subset \partial\Omega$.

COROLLARY 3.4. *Let $Z_{x,t}$ be a function over $S_x \times [0,T]$ such that there exist two constants $\alpha \geqslant 0$ and C_Z independent of x, y, t, and τ satisfying*

$$|Z_{x,t}(y,\tau)| \leqslant C_Z \left(|x-y|^\alpha + |\nu(t-\tau)|^{\alpha/2}\right)$$

for all $(y,\tau) \in S_x \times [0,t]$ and $(x,t) \in \partial\Omega \times [0,T]$. Under the hypothesis (C1)–(C4) *of section* 2.2, *there exists a constant C_0 such that for any $x \in \partial\Omega$, one has*

$$\widetilde{\mathcal{H}}_{S_x}(x,t)Z_{x,t} = C_0\,t^{(\alpha+1)/2}.$$

Proof. Let $Z^\natural_{x,t}$ be the parameterization of Z through the map $T_x\phi_x$:

$$Z^\natural_{x,t}(s,\tau) = Z_{x,t}(T_x\phi_x(s),\tau),$$

which gives $\widetilde{\mathcal{H}}_{S_x}(x,t)Z_{x,t} = \mathcal{H}(x,t)Z^\natural_{x,t}$. One then has

$$|Z^\natural_{x,t}(s,\tau)| \leqslant C_Z \left(|T_x\phi_x(s) - x|^\alpha + |\nu(t-\tau)|^{\alpha/2}\right) \leqslant C_0 \left(|s|^\alpha + |\nu(t-\tau)|^{\alpha/2}\right)$$

since T_x is an isometry and $|\phi_x(s)|^2 = |s|^2 + |\varphi_x(s)|^2 \leqslant (1 + C_1^2R^2)|s|^2$, which leads to set $C_0 = C_Z\sqrt{1 + C_1^2R^2}$. One can then apply Lemma 3.2. ∎

3.4. Geometrical convergence. Now we present the following stability lemma of fundamental importance.

LEMMA 3.5 ($\mathcal{H}-\overline{\mathcal{H}}$ estimation). *Under the hypothesis* (C1)–(C4) *of section* 2.2, $\partial\Omega$ *being a two-dimensional manifold of class* $\mathcal{C}^{2+\lambda}$, $0<\lambda\leqslant 1$, *there exists a constant* C_0 *independent of* t *such that*

$$||||\mathcal{H}(\cdot,t)-\overline{\mathcal{H}}(\cdot,t)||||_\infty=\sup_{x\in\partial\Omega}|||\mathcal{H}(x,t)-\overline{\mathcal{H}}(x,t)|||_\infty\leqslant\ C_0\,t^{(1+\lambda)/2}.$$

Proof. The proof of this lemma uses more or less the same technique as the proof of Lemma 3.2. One chooses an $x\in\partial\Omega$ and gets for a bounded function over $S_x\subset\partial\Omega$

$$|\mathcal{H}(x,t)f-\overline{\mathcal{H}}(x,t)f|\leqslant\|f\|_\infty|\mathcal{H}(x,t)1-\overline{\mathcal{H}}(x,t)1|$$

with equality for $f\equiv 1$, and consequently

$$|||\mathcal{H}(x,t)-\overline{\mathcal{H}}(x,t)|||_\infty=|\mathcal{H}(x,t)1-\overline{\mathcal{H}}(x,t)1|.$$

It remains to build an accurate majoration of this quantity independent of x.

Operators $\mathcal{H}(x,t)$ and $\overline{\mathcal{H}}(x,t)$ are defined by formulas (3.10) and (3.11), which give

$$|||\mathcal{H}(x,t)-\overline{\mathcal{H}}(x,t)|||_\infty=\left|\int_0^{\nu t}\int_A\gamma(s,u)\widehat{G}\left(\phi_x(s),u\right)-\overline{\gamma}(s,u)\widehat{G}\left(\overline{\phi}_x(s),u\right)\,\mathrm{d}s\mathrm{d}u\right|,$$

where $u=\nu(t-\tau)$, and

$$\gamma(s,u)=\left(\beta_x-\frac{\varphi_x(s)}{2u}\right)\sqrt{1+|\nabla\varphi_x(s)|^2}$$

and

$$\overline{\gamma}(s,u)=\left(\beta_x-\frac{\overline{\varphi}_x(s)}{2u}\right)\sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}.$$

Moreover, $\widehat{G}\left(\overline{\phi}_x(s),u\right)$ and $\widehat{G}\left(\phi_x(s),u\right)$ are both majorated by $e^{-|s|^2/4u}/(4\pi u)^{3/2}$, and thus

$$|||\mathcal{H}(x,t)-\overline{\mathcal{H}}(x,t)|||_\infty\leqslant\int_0^{\nu t}\int_{\mathbb{R}^2}\left|\gamma(s,u)-\overline{\gamma}(s,u)\right|\frac{e^{-|s|^2/4u}}{(4\pi u)^{3/2}}\,\mathrm{d}s\mathrm{d}u. \tag{3.21}$$

The graph approximation, by means of condition (C4), satisfies

$$\left|\varphi_x(s)-\overline{\varphi}_x(s)\right|\ \leqslant\ C\,|s|^{2+\lambda}. \tag{3.22}$$

Moreover, the gradients satisfy $\nabla\overline{\varphi}_x(s)=K_xs$, and there exists a constant C' such that $|\nabla\varphi_x(s)-K_xs|\leqslant C'\,|s|^{1+\lambda}$; thus

$$|\nabla\varphi_x(s)-\nabla\overline{\varphi}_x(s)|^2\leqslant C'|s|^{2+2\lambda}. \tag{3.23}$$

One can do the following decomposition relying on the triangular inequality:

$$\left|\gamma(s,u)-\overline{\gamma}(s,u)\right|\leqslant K_1+K_2$$

with

$$K_1 = \left| \left(\beta_x - \frac{\varphi_x(s)}{2u}\right)\sqrt{1+|\nabla\varphi_x(s)|^2} - \left(\beta_x - \frac{\varphi_x(s)}{2u}\right)\sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}\ \right|$$

and

$$K_2 = \left| \left(\beta_x - \frac{\varphi_x(s)}{2u}\right)\sqrt{1+|\nabla\overline{\varphi}_x(s)|^2} - \left(\beta_x - \frac{\overline{\varphi}_x(s)}{2u}\right)\sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}\ \right|.$$

The first part K_1 can be majorated by

$$K_1 \leqslant \left(\beta_0 + \left|\frac{\varphi_x(s)}{2u}\right|\right)\left|\sqrt{1+|\nabla\varphi_x(s)|^2}\ -\ \sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}\ \right|, \tag{3.24}$$

and the second by

$$K_2 \leqslant \frac{1}{2u}\left|\varphi_x(s) - \overline{\varphi}_x(s)\right|\ \sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}, \tag{3.25}$$

which, noticing that $|\nabla\varphi_x(s)| = |K_x s| \leqslant \rho_0 R \leqslant \rho_0 R + C' R^{1+\lambda}$, gives (see proof of Proposition 2.3(P1))

$$K_2 \leqslant M_1 \frac{\left|\varphi_x(s) - \overline{\varphi}_x(s)\right|}{2u} \leqslant M_1 C'\,\frac{|s|^{2+\lambda}}{2u}. \tag{3.26}$$

Furthermore, the triangular inequality

$$|\nabla\varphi_x(s)|^2 \leqslant |\nabla\overline{\varphi}_x(s)|^2 + |\nabla\varphi_x(s) - \nabla\overline{\varphi}_x(s)|^2$$

implies

$$\sqrt{1+|\nabla\varphi_x(s)|^2} \leqslant \sqrt{1+|\nabla\overline{\varphi}_x(s)|^2 + |\nabla\varphi_x(s) - \nabla\overline{\varphi}_x(s)|^2}. \tag{3.27}$$

Now one can notice that the epigraph of $-\sqrt{x}$ is convex; thus for any $a > 0$ one has $\sqrt{a+x} \leqslant \sqrt{a} + x/2\sqrt{a}$ for all $x \in [-a, +\infty[$. This implies

$$\sqrt{1+|\nabla\varphi_x(s)|^2} \leqslant \sqrt{1+|\nabla\overline{\varphi}_x(s)|^2} + \frac{|\nabla\varphi_x(s) - \nabla\overline{\varphi}_x(s)|^2}{2\sqrt{1+|\nabla\overline{\varphi}_x(s)|^2}}. \tag{3.28}$$

This leads to a majoration of (3.24) using (3.23):

$$K_1 \leqslant \frac{C'}{2}\left(\beta_0 + \left|\frac{\varphi_x(s)}{2u}\right|\right)|s|^{2+2\lambda}. \tag{3.29}$$

Since one has $|\varphi_\varsigma(s)| \leqslant (\rho_0 + C\,R^\lambda)|s|^2$, one finally has the majoration

$$\left|\gamma(s,u) - \overline{\gamma}(s,u)\right| \leqslant \frac{C'}{2}\left(\beta_0 + (\rho_0 + C\,R^\lambda)\frac{|s|^2}{2u}\right)|s|^{2+2\lambda} + M_1 C'\,\frac{|s|^{2+\lambda}}{2u}. \tag{3.30}$$

Integral (3.21) can then be majorated by applying Lemma 3.1 and majoration obtained by (3.30). Indeed, one can set $r = |s|$ and apply Lemma 3.1 with $R = 0$. This gives

$$|||\mathcal{H}(x,t)-\overline{\mathcal{H}}(x,t)|||_\infty \leqslant \frac{C'\beta_0}{8\sqrt{\pi}}\Gamma^{0,4}_{2+2\lambda,3/2}(t)+\frac{C'\beta_0}{16\sqrt{\pi}}(\rho_0+C\,R^\lambda)\Gamma^{0,4}_{4+2\lambda,5/2}(t) + \frac{M_1C'}{8\sqrt{\pi}}\Gamma^{0,4}_{2+\lambda,5/2}(t)=\mathcal{O}(t^{(1+\lambda)/2}). \qquad \square \tag{3.31}$$

PROPOSITION 3.6 (geometrical convergence). *Under the previous notation and conditions* (C1)–(C4), *with* $\mathcal{H}$ *and* $\overline{\mathcal{H}}$ *defined, respectively, by formulas* (3.10) *and* (3.11), *and for* t *sufficiently small, the densities* $\mu^\natural$ *and* $\mu^\sharp$ *defined in section* 3 *satisfy*

$$\frac{\|\mu^\natural(\cdot,t)-\mu^\sharp(\cdot,t)\|_\infty}{\|\mu^\natural(\cdot,t)\|_\infty} \leqslant \frac{2}{1-2||||\overline{\mathcal{H}}(\cdot,t)||||_\infty}\,||||\mathcal{H}(\cdot,t)-\overline{\mathcal{H}}(\cdot,t)||||_\infty. \tag{3.32}$$

Proof. Let $\mu^\natural$, introduced in section 3, be the solution of the integral equation involving the localized integrodifferential operator (see also Figure 3),

$$-\frac{1}{2}\mu^\natural(x,t)+\widetilde{\mathcal{H}}_{S_x}\mu^\natural = F(x,t), \tag{3.33}$$

whose solution $\mu^\natural$, bounded over $\partial\Omega$, has a local parameterization μ_x on A by $\mu_x(s,t)=\mu^\natural(T_x\phi_x(s),t)$. The integral equation (3.33) is then equivalent to

$$-\frac{1}{2}\mu_x(0,t)+\mathcal{H}(x,t)\mu_x = F(x,t). \tag{3.34}$$

The S_x-localized integrodifferential operator $\mathcal{H}$ acting on parameterized densities can then be defined on an approximated surface $\overline{S_x}$, which leads to the integral equation

$$-\frac{1}{2}\overline{\mu}_x(0,t)+\overline{\mathcal{H}}(x,t)\overline{\mu}_x = F(x,t) \tag{3.35}$$

of solution $\overline{\mu}_x : A\times[0,T]\to\mathbb{R}$, which is the parameterization of the density $\mu^\sharp(x,t)=\overline{\mu}_x(0,t)$. The functions μ_x and $\overline{\mu}_x$ are both defined on the same domain A but represent densities on two different surfaces, $S_x\subset\partial\Omega$ and $\overline{S_x}$.

One can then write, subtracting (3.34) and (3.35),

$$\begin{aligned}\frac{\mu_x(s,t)-\overline{\mu}_x(s,t)}{2} &= \mathcal{H}(x,t)\mu_x-\overline{\mathcal{H}}(x,t)\overline{\mu}_x\\ &= \mathcal{H}(x,t)\mu_x-\overline{\mathcal{H}}(x,t)\mu_x+\overline{\mathcal{H}}(x,t)\mu_x-\overline{\mathcal{H}}(x,t)\overline{\mu}_x\end{aligned}$$

and

$$\frac{|\mu_x(s,t)-\overline{\mu}_x(s,t)|}{2} \leqslant |||\mathcal{H}(x,t)-\overline{\mathcal{H}}(x,t)|||_\infty\,\|\mu_x\|_\infty+|||\overline{\mathcal{H}}(x,t)|||_\infty\,\|\mu_x-\overline{\mu}_x\|_\infty, \tag{3.36}$$

where the $\mathbb{L}^\infty$-norm of densities is taken over A. One can notice that

$$\|\mu_x\|_\infty=\sup_{s\in A}|\mu_x(s)|=\sup_{\xi\in S_x}|\mu^\natural(\xi)|\leqslant\sup_{\xi\in\partial\Omega}|\mu^\natural(\xi)|=\|\mu^\natural\|_\infty$$

and also that $\|\mu_x-\overline{\mu}_x\|_\infty\leqslant\|\mu^\natural-\mu^\sharp\|_\infty$.

Taking the maximum of x over $\partial\Omega$ for the triple norms in (3.36) and applying it to $s = 0$ gives

(3.37)
$$|\mu^\natural(x,t) - \mu^\sharp(x,t)| \leqslant 2\,||||\mathcal{H}(\cdot,t) - \overline{\mathcal{H}}(\cdot,t)||||_\infty\,\|\mu^\natural\|_\infty + 2\,||||\overline{\mathcal{H}}(\cdot,t)||||_\infty\,\|\mu^\natural - \mu^\sharp\|_\infty.$$

The right-hand side of (3.37) is independent of x; thus one can take the maximum of x over $\partial\Omega$ on the left-hand side and get

$$\left(1 - 2\,||||\overline{\mathcal{H}}(\cdot,t)||||_\infty\right)\,\|\mu^\natural - \mu^\sharp\|_\infty \leqslant 2\,||||\mathcal{H}(\cdot,t) - \overline{\mathcal{H}}(\cdot,t)||||_\infty\,\|\mu^\natural\|_\infty. \tag{3.38}$$

Finally, by Corollary 3.3 of Lemma 3.2, one has $||||\overline{\mathcal{H}}(\cdot,t)||||_\infty < 1/2$, which finishes the proof. ∎

This proposition leads directly to the order of convergence when $\partial\Omega$ is sufficiently smooth, when used with Lemma 3.5

COROLLARY 3.7. *Under the hypothesis of Proposition* 3.6, *if* $\partial\Omega$ *is a manifold of class* $\mathcal{C}^{2+\lambda}$, $0 < \lambda \leqslant 1$, *then there exists a constant* C_0 *independent of* t *such that*

$$\|\mu^\natural(\cdot,t) - \mu^\sharp(\cdot,t)\|_\infty \leqslant C\,t^{(1+\lambda)/2}. \tag{3.39}$$

This means that a manifold of class $\mathcal{C}^3$ allows us to reach order 1 in density convergence when locally approximating the surface by its best quadratic osculating surface.

4. Leading order of $\overline{\mathcal{H}}$. We have shown in last section that the best parabolic approximation of the surface leads at least to first order in the approximation of the density, which is enough to carry out the main contribution of the curvature effect to the solution.

LEMMA 4.1. *Under conditions* (C1)–(C4) *and previous notation, we consider the integral*

$$\mathcal{H}_0(x,t) = \int_0^{\nu t}\int_{\mathbb{R}^2}\left(\beta_x - \frac{{}^t s K_x s}{4u}\right)\frac{e^{-(|s|^2 + {}^t s K_x s/2)/4u}}{(4\pi u)^{3/2}}\,\mathrm{d}s\,\mathrm{d}u. \tag{4.1}$$

Then $\overline{\mathcal{H}}(x,t)1 = \mathcal{H}_0(x,t) + \mathcal{O}\left(t^{3/2}\right)$, *with* $\overline{\mathcal{H}}(x,t)$ *defined by* (3.11).

Proof. We first introduce the integral

$$H_1(x,t) = \int_0^{\nu t}\int_{\mathbb{R}^2}\left(\beta_x - \frac{{}^t s K_x s}{4u}\right)\frac{e^{-(|s|^2 + ({}^t s K_x s)^2/4)/4u}}{(4\pi u)^{3/2}}\sqrt{1 + |K_x s|^2}\,\mathrm{d}s\,\mathrm{d}u \tag{4.2}$$

with ${}^t s K_x s/2 = \overline{\varphi}_x(s)$. We then get

$$|\overline{\mathcal{H}}(x,t)1 - H_1(x,t)| \leqslant \frac{M_1}{4\sqrt{\pi}}\int_0^{\nu t}\int_R^{+\infty}\left(\beta_0 + \frac{|\overline{\varphi}_x(s)|}{2u}\right)\frac{e^{-|s|^2/4u}}{u^{3/2}}\,r\,\mathrm{d}r,$$

where $|\overline{\varphi}_x(s)| \leqslant \rho_0|s|^2/2$, which gives

$$|\overline{\mathcal{H}}(x,t)1 - H_1(x,t)| \leqslant \frac{M_1}{4\sqrt{\pi}}(\beta_0 + \rho_0 L_1)\,\Gamma^{R,8}_{0,3/2}(t). \tag{4.3}$$

One can now focus on the estimation of $\mathcal{H}_1(x,t) - H_0(x,t)$. The difference between the two operators lies in the Jacobian $N_x(s)$, which is not present in the definition of $\mathcal{H}_0(x,t)$. One has

$$|\mathcal{H}_1(x,t) - H_0(x,t)| \leqslant \frac{\beta_0 + \rho_0 L_1}{4\sqrt{\pi}}\int_0^{\nu t}\int_0^{+\infty}\frac{e^{-|s|^2/8u}}{u^{3/2}}\left(\sqrt{1 + \rho_0^2 r^2} - 1\right)r\,\mathrm{d}r,$$

which gives, since $\sqrt{1+\alpha} \leqslant 1 + \alpha/2$ for any $\alpha \geqslant -1$,

$$|\mathcal{H}_1(x,t) - H_0(x,t)| \leqslant \frac{\rho_0^2}{8\sqrt{\pi}}(\beta_0 + \rho_0 L_1)\,\Gamma^{0,8}_{2,3/2}(t). \tag{4.4}$$

Applying triangular inequality and noticing that $\Gamma^{R,8}_{0,3/2}(t) + \Gamma^{0,8}_{2,3/2}(t) = \mathcal{O}\left(t^{3/2}\right)$ finishes the proof the lemma. □

Now that the integrals $\overline{\mathcal{H}}(x,t)1$ and $\mathcal{H}_0(x,t)$ are linked and are an approximation of one another at an order higher than surface approximation (see Lemma 3.5), one can focus on the estimation of $\mathcal{H}_0(x,t)$.

PROPOSITION 4.2. *Under conditions* (C1)–(C4) *and* $\mathcal{H}_0(x,t)$ *defined by* (4.1), *one has uniformly*

$$\mathcal{H}_0(x,t) = \left(\beta_x - \frac{\mathsf{tr}(K_x)}{2}\right)\sqrt{\frac{\nu t}{\pi}} + \mathcal{O}\left(t^{3/2}\right). \tag{4.5}$$

Proof. We begin to write the curvature matrix K_x as

$$K_x = \begin{bmatrix} \kappa_1 & \kappa_0 \\ \kappa_0 & \kappa_2 \end{bmatrix}.$$

By means of the cylindrical change of variable $s = (r\cos\theta,\, r\sin\theta)$, the integral (4.1) becomes

$$\mathcal{H}_0(x,t) = \int_0^{2\pi}\int_0^{\nu t}\int_0^{\infty}\left(\beta_x - \frac{r^2 m(\theta)}{4u}\right)\frac{e^{-(r^2+r^4 m(\theta)^2/4)/4u}}{(4\pi u)^{3/2}}\, r\,\mathrm{d}r\,\mathrm{d}u\,\mathrm{d}\theta, \tag{4.6}$$

where $m(\theta) = \kappa_1\cos^2\theta + \kappa_2\sin^2\theta + \kappa_0\cos\theta\sin\theta$. Now posing $(r,u) = (\xi\zeta, \zeta^2/4)$, associated with a Jacobian $\zeta^2/2$, one gets

$$\mathcal{H}_0(x,t) = \frac{1}{2\pi^{3/2}}\int_0^{2\pi}\int_0^{2\sqrt{\nu t}}\int_0^{\infty}\gamma(\xi,\zeta,\theta,x,t)\,\mathrm{d}\xi\,\mathrm{d}\zeta\,\mathrm{d}\theta$$

with

$$\gamma(\xi,\zeta,\theta,x,t) = (\xi\beta_x - \xi^3 m(\theta))\exp\left\{-\xi^2\left(1 + \frac{\xi^2\zeta^2 m(\theta)^2}{4}\right)\right\}, \tag{4.7}$$

which is infinitely differentiable in variables ξ, ζ, and θ.

Since for any positive constants a and b, the function $f(\zeta) = e^{-a-b\zeta^2}$ satisfies $f'(0) = 0$ and f'' is bounded over $\mathbb{R}^+$, there exists $\zeta_0 \in [0,\zeta]$ such that

$$f(\zeta) = f(0) + \zeta^2 f''(\zeta_0)/2.$$

In order to make a Taylor development of γ in ζ valid, one has to exhibit the bounds of f'' with respect to coefficients a, b, and c. Indeed, one has

$$f''(\zeta) = (4b^2\zeta^2 - 2b)e^{-a-b\zeta^2};$$

thus

$$|f''(\zeta)| \leqslant 2be^{-a}(1 + 2b\zeta^2)e^{-b\zeta^2} \leqslant 4be^{-a}. \tag{4.8}$$

Applying this result with $a = \xi^2$ and $b = \xi^4 m(\theta)^2/4$ to the exponential part of (4.7), and using majoration (2.13), one gets

$$\left|\exp\left\{-\xi^2\left(1+\frac{\xi^2\zeta^2 m(\theta)^2}{4}\right)\right\}-1\right| \leqslant \zeta^2\xi^4 m(\theta)^2 e^{-\xi^2} \leqslant L_2\,\zeta^2\, m(\theta)^2 e^{-\xi^2/2}.$$

Consequently, one has

$$\left|\gamma(\xi,\zeta,\theta,x,t) - (\xi\beta_x - \xi^3 m(\theta))e^{-\xi^2}\right| \leqslant L_2\left(\beta_0\xi + |m(\theta)|\xi^3\right)\zeta^2 m(\theta)^2 e^{-\xi^2/2}, \tag{4.9}$$

which is itself majorated, using again formula (2.13), by

$$\sqrt{2}\,L_2\left(L_{1/2}\beta_0 + 2L_{3/2}|m(\theta)|\right)\zeta^2 m(\theta)^2 e^{-\xi^2/4}.$$

Now noticing that

$$\begin{aligned}|m(\theta)| &\leqslant |\kappa_1| + |\kappa_0| + |\kappa_0| + |\kappa_2| \\ &\leqslant 2\max\left(|\kappa_1| + |\kappa_0|, |\kappa_0| + |\kappa_2|\right) = 2|||K_x|||_1 \leqslant 2\sqrt{2}|||K_x|||_2 \leqslant 2\sqrt{2}\rho_0,\end{aligned}$$

where $|||\cdot|||$ is the standard norm for linear operators, and setting

$$C = 8\sqrt{2}\,L_2\left(L_{1/2}\beta_0 + 4L_{3/2}\rho_0\right)\rho_0^2, \tag{4.10}$$

one gets

$$\left|\gamma(\xi,\zeta,\theta,x,t) - (\xi\beta_x - \xi^3 m(\theta))e^{-\xi^2}\right| \leqslant C\,\zeta^2\, e^{-\xi^2/4}, \tag{4.11}$$

whose right-hand side is variable separated and integrates obviously into

$$\int_0^{2\pi}\int_0^{2\sqrt{\nu t}}\int_0^{\infty} \zeta^2\, e^{-\xi^2/4}\,\mathrm{d}\xi\,\mathrm{d}\zeta\,\mathrm{d}\theta = \frac{16\pi^{3/2}}{3}(\nu t)^{3/2}, \tag{4.12}$$

and consequently

$$\left|\mathcal{H}_0(x,t) - \frac{\sqrt{\nu t}}{\pi^{3/2}}\int_0^{2\pi}\int_0^{\infty}(\xi\beta_x - \xi^3 m(\theta))e^{-\xi^2}\,\mathrm{d}\xi\,\mathrm{d}\theta\right| \leqslant \frac{8C}{3}(\nu t)^{3/2}. \tag{4.13}$$

Moreover, one has

$$\int_0^{2\pi} m(\theta)\,\mathrm{d}\theta = \pi\kappa_1 + \pi\kappa_2 = \pi\,\mathrm{tr}(K_x), \tag{4.14}$$

and thus

$$\left|\mathcal{H}_0(x,t) - \sqrt{\frac{\nu t}{\pi}}\int_0^{\infty}(2\xi\beta_x - \xi^3(\kappa_1+\kappa_2))e^{-\xi^2}\,\mathrm{d}\xi\right| \leqslant \frac{8C}{3}(\nu t)^{3/2} \tag{4.15}$$

with

$$\int_0^{\infty}(2\xi\beta_x - \xi^3(\kappa_1+\kappa_2))e^{-\xi^2}\,\mathrm{d}\xi = \beta_x - \frac{\kappa_1+\kappa_2}{2},$$

which concludes the proof. □

5. Curvature effect on the whole surface. From all the previous sections we can give now the following result.

THEOREM 5.1. *Let Ω be an open set of $\mathbb{R}^3$ such that $\partial\Omega$ is a two-dimensional manifold of class $\mathcal{C}^{2+\lambda}$, $0 < \lambda \leqslant 1$, satisfying conditions* (C1)–(C5), *and $\mathcal{L}_x$ is a Robin–Fourier differential operator $\mathcal{L}_x = \beta(x)\mathrm{Id} + n(x)\cdot\nabla$ with β bounded over $\partial\Omega$ and $n(x)$ the inward normal to $\partial\Omega$ in x. Let also $\widetilde{\mathcal{H}}$ be the following integrodifferential operator:*

$$\widetilde{\mathcal{H}}(x,t)f = \nu \int_0^t \int_{\partial\Omega} \mathcal{L}_x G_{\xi,\tau}(x,t)\, f(\xi,\tau)\mathrm{d}\sigma(\xi)\mathrm{d}\tau \tag{5.1}$$

for all continuous and bounded functions $f : \partial\Omega \times [0,T] \to \mathbb{R}$. If the density $\widetilde{\mu}$ is α-Hölder continuous, then

$$\widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = \left[\widetilde{\mu}(x,t)\ \left(\beta(x) - \bar{\kappa}(x)\right)\right]\sqrt{\frac{\nu t}{\pi}} + \mathcal{O}(t^{(1+\gamma)/2}), \tag{5.2}$$

where $\gamma = \min(\alpha,\lambda)$ and $\bar{\kappa}(x)$ is the mean curvature of $\partial\Omega$ in x.

Proof. Let $\widetilde{\mu}$ be a density that is supposed α-Hölder continuous with $\alpha \in]0,\,1]$. One can introduce the function $Z_{x,t}$ defined by

$$Z_{x,t}\ :\ (y,\tau) \mapsto Z_{x,t}(y,\tau) = \widetilde{\mu}(y,\tau) - \widetilde{\mu}(x,t) \tag{5.3}$$

with the following property:

$$|Z_{x,t}(y,\tau)| = |\widetilde{\mu}(y,\tau) - \widetilde{\mu}(x,t)| \leqslant C_Z\left(|x-y|^\alpha + |\nu(t-\tau)|^{\alpha/2}\right). \tag{5.4}$$

One then gets

$$\widetilde{\mathcal{H}}_{S_x}(x,t)Z_{x,t} = \widetilde{\mathcal{H}}_{S_x}(x,t)\widetilde{\mu} - \widetilde{\mu}(x,t)\widetilde{\mathcal{H}}_{S_x}(x,t)1.$$

By Corollary 3.4 of Lemma 3.2, one gets that there exists a constant C_1 independent of x and t such that

$$|\widetilde{\mathcal{H}}_{S_x}(x,t)Z_{x,t}| = |\widetilde{\mathcal{H}}_{S_x}(x,t)\widetilde{\mu} - \widetilde{\mu}(x,t)\widetilde{\mathcal{H}}_{S_x}(x,t)1| \leqslant C_1 t^{(\alpha+1)/2} \tag{5.5}$$

with $\widetilde{\mathcal{H}}_{S_x}(x,t)1 = \mathcal{H}(x,t)1$, and thus

$$\left|\widetilde{\mathcal{H}}_{S_x}(x,t)1 - \overline{\mathcal{H}}(x,t)1\right| \leqslant C_2 t^{(1+\lambda)/2} \tag{5.6}$$

by Lemma 3.5,

$$\left|\overline{\mathcal{H}}(x,t)1 - \mathcal{H}_0(x,t)\right| \leqslant C_3 t^{3/2} \tag{5.7}$$

by Lemma 4.1, and also

$$\left|\mathcal{H}_0(x,t) - (\beta_x - \bar{\kappa}(x))\sqrt{\nu t/\pi}\right| \leqslant C_4 t^{3/2} \tag{5.8}$$

by Proposition 4.2, where $\bar{\kappa}(x) = \mathrm{tr}K_x/2$ is the mean curvature of $\partial\Omega$ in x.

Joining together (5.5), (5.6), (5.7), and (5.8) gives

$$\left|\widetilde{\mathcal{H}}_{S_x}(x,t)\widetilde{\mu} - (\beta_x - \bar{\kappa}(x))\,\widetilde{\mu}(x,t)\sqrt{\nu t/\pi}\right| \leqslant C_5 t^{(\gamma+1)/2} \tag{5.9}$$

with $\gamma = \min(\alpha, \lambda)$ and $C_5 = C_1 T^{(\alpha-\gamma)/2} + C_2 T^{(\lambda-\gamma)/2} + (C_3 + C_4) T^{2-\gamma/2}$. Noticing that

$$\widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = \widetilde{\mathcal{H}}_{S_x}(x,t)\widetilde{\mu} + \mathcal{O}(t^\infty) \tag{5.10}$$

by Proposition 2.4 gives the final result. □

First of all, it is necessary to remark that Theorem 5.1 gives the early behavior of the solution, as mentioned in section 1.

COROLLARY 5.2. *Let $\partial\Omega$ be a two-dimensional manifold of class $\mathcal{C}^3$ and let F be a bounded function over $\partial\Omega \times [0,T]$ such that the solution $\widetilde{\mu}$ of the integral equation*

$$-\frac{1}{2}\widetilde{\mu}(x,t) + \widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = F(x,t) \tag{5.11}$$

is bounded and $(1-\varepsilon)$-Hölder continuous with the Hölder exponent satisfying $0 \leqslant \varepsilon < 1$. Under the notation and hypothesis of Theorem 5.1*, one has*

$$\widetilde{\mu}(x,t) = \frac{-2F(x,t)}{1 + 2\left(\bar{\kappa}(x) - \beta(x)\right)\sqrt{\nu t/\pi}} + \mathcal{O}(t^{1-\varepsilon/2}). \tag{5.12}$$

Furthermore, one can build a result in a smoother context concerning torsion-free surfaces.

THEOREM 5.3. *Let Ω be an open set of $\mathbb{R}^3$ such that $\partial\Omega$ is a two-dimensional manifold of class $\mathcal{C}^{3+\lambda^*}$ without torsion, with $0 < \lambda^* \leqslant 1$, satisfying conditions* (C1)–(C5). *Let β and F be two functions bounded, respectively, over $\partial\Omega$ and $\partial\Omega \times [0,T]$ such that the solution $\widetilde{\mu}$ of the integral equation* (5.11) *is of class $\mathcal{C}^{1+\alpha}(\partial\Omega \times [0,T])$ and bounded. Then the solution $\widetilde{\mu}$ of* (5.11) *satisfies*

$$\widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = \left[\widetilde{\mu}(x,t)\ \left(\beta(x) - \bar{\kappa}(x)\right)\right]\sqrt{\frac{\nu t}{\pi}} + \mathcal{O}(t^{1+\gamma/2}), \tag{5.13}$$

where $\gamma = \min(\alpha, \lambda^*)$ and $\bar{\kappa}(x)$ is the mean curvature of $\partial\Omega$ in x.

Proof. In the context of a manifold smoother than $\mathcal{C}^3$, there exist two constants C_x and C'_x such that

$$\left|\varphi_x(s) - \frac{1}{2}\,{}^t s K_x s - \frac{1}{6}\,T_x^{ijk} s_i s_j s_k\right| \ \leqslant\ C_x\,|s|^{3+\lambda} \tag{5.14}$$

and

$$\left|\nabla\varphi_x(s) - K_x s - \frac{1}{2}\,{}^t s(\mathcal{T}_x : e_\cdot)s\right| \ \leqslant\ C'_x\,|s|^{2+\lambda}, \tag{5.15}$$

where $\mathcal{T}_x$ is the torsion tensor of φ_x in 0 defined as

$$\mathcal{T}_x^{ijk} = \frac{\partial^3\varphi_x}{\partial s_i \partial s_j \partial s_k}(0)$$

and $\mathcal{T}_x^{ijk} s_i s_j s_k$ is its associated cubic form. The torsion tensor can be contracted with the vectors of canonical basis e_k, such that the kth component of $\mathcal{T}_x : e_\cdot$ is the matrix $\mathcal{T}_x : e_k$ associated to the quadratic form ${}^t s(\mathcal{T}_x : e_k)s$. This can be equivalently stated, using again the Einstein notation, as

$${}^t s(\mathcal{T}_x : e_\cdot)s = (\mathcal{T}_x : e_\cdot)^{ij} s_i s_j = \mathcal{T}_x^{ijk} s_i s_j e_k.$$

This statement is also equivalent to the symmetric formulation $(E^*)^{\otimes 3} \equiv E \otimes E^* \otimes E^*$ with $E = \mathbb{R}^2$.

Constants C_x and C'_x can be assumed to be bounded independently of x over $\partial\Omega$, and in the case of a torsion-free surface (for example, cylinders and spheres are torsion-free), the torsion tensor is identically equal to 0, which leads to majorations of the same kind as those of equations (2.3) with majorants $C|s|^{3+\lambda^*}$ and $C'|s|^{2+\lambda^*}$, respectively, for the gradients. In this context, Lemma 3.5 holds with $\lambda = 1 + \lambda^* \in]0, 2]$, which gives that

$$\widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = \mathcal{H}(x,t)\mu_x^\flat + \mathcal{O}(t^\infty) = \overline{\mathcal{H}}(x,t)\mu_x^\flat + \mathcal{O}(t^{1+\lambda^*/2}) \tag{5.16}$$

with $\mu_x^\flat(s,\tau) = \widetilde{\mu}(T_x\phi_x(s),\tau)$.

Moreover, thanks to greater regularity of the density through the maps, one can write the following development:

$$\begin{aligned}\mu_x^\flat(s,\tau) = \mu_x^\flat(0,t) + R_{x,t}(s,\tau) - \nu(t-\tau)\frac{\partial \mu_x^\flat}{\partial \tau}\Big|_{0,t} \\ + \mathcal{O}\left(|s|^{1+\alpha} + |s|u^{\alpha/2}\right) + \mathcal{O}\left(|s|^\alpha u + u^{1+\alpha/2}\right)\end{aligned} \tag{5.17}$$

with $R_{x,t}(s,\tau) = s \cdot \nabla\mu_x^\flat(0,t)$ and $u = \nu|t-\tau|$.

By oddness, one has $\overline{\mathcal{H}}(x,t)R_{x,t} = 0$, and one can easily establish upon proof of Lemma 3.2 that

$$\left|\overline{\mathcal{H}}(x,t)[|s|^a(\nu(t-\tau))^b]\right| \leqslant \frac{2\,M_1}{\sqrt{\pi}} \int_0^{\nu t}\int_0^R \left(\beta_0 + \frac{\rho_0}{2}\,\frac{r^2}{2u}\right) \frac{r^a u^b e^{-r^2/4u}}{8u^{3/2}}\, r\,\mathrm{d}r\mathrm{d}\tau, \tag{5.18}$$

which gives

$$\left|\overline{\mathcal{H}}(x,t)[|s|^a(\nu(t-\tau))^b]\right| \leqslant C_1\,\Gamma^{0,8}_{a,3/2-b} = \mathcal{O}(t^{1/2+b+a/2}) \tag{5.19}$$

with $C_1 = M_1(\beta_0 + L_1\rho_0)/4\sqrt{\pi}$. This implies that $|\overline{\mathcal{H}}(x,t)[\nu(t-\tau)]| = \mathcal{O}(t^{3/2})$; thus (5.17) reads as

$$\mu_x^\flat(s,\tau) \;= \mu_x^\flat(0,t) + \mathcal{O}(t^\eta) \tag{5.20}$$

with $\eta = \min(3/2, 1+\alpha/2, 3/2+\alpha/2) = 1+\alpha/2$. Combining this with (5.16) gives

$$\widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = \overline{\mathcal{H}}(x,t)\mu_x^\flat + \mathcal{O}(t^{1+\lambda^*/2}) = \mu_x^\flat(0,t)\overline{\mathcal{H}}(x,t)1 + \mathcal{O}(t^{1+\lambda^*/2}) + \mathcal{O}(t^{1+\alpha/2}). \tag{5.21}$$

By Lemma 4.1 and Proposition 4.2, one has

$$\left|\overline{\mathcal{H}}(x,t)1 - (\beta_x - \bar{\kappa}(x))\,\sqrt{\nu t/\pi}\right| = \mathcal{O}(t^{3/2}), \tag{5.22}$$

and consequently (5.21) leads to

$$\widetilde{\mathcal{H}}(x,t)\widetilde{\mu} = \widetilde{\mu}(x,t)\overline{\mathcal{H}}(x,t)1 + \mathcal{O}(t^{1+\gamma/2}) = \widetilde{\mu}(x,t)\,(\beta_x - \bar{\kappa}(x))\,\sqrt{\nu t/\pi} + \mathcal{O}(t^{1+\gamma/2}) \tag{5.23}$$

with $\gamma = \min(\lambda^*, \alpha)$. □

Note that Theorem 5.3 is not an extension of Theorem 5.1, because a function can be more regular than Hölder continuous (even with an exponent 1) and less than $\mathcal{C}^1$ (for example, $\sqrt{t}$). Moreover, one can notice that this result is useful only as a

kind of regularity estimation of the integrodifferential operator $\widetilde{\mathcal{H}}$, since the density obtained as the solution of the original integral equation (5.11) exhibits a square-root singularity in the general case (see Corollary 5.2) and consequently cannot be $\mathcal{C}^1$.

Furthermore, a direct corollary of this result is that if the density is $\mathcal{C}^2(\partial\Omega\times[0,T])$, which is in practice a restrictive condition, and if the manifold $\partial\Omega$ is $\mathcal{C}^4$, then the error is of order 3/2. This fact is illustrated in section 6.2 with a constant density and a cylinder.

6. Cylindrical examples. In this section we provide a few canonical examples showing the contribution of the curvature effect on the solution.

Example 6.1 puts the problem of boundary source in the more general context of enforcing boundary conditions. Examples 6.2 and 6.3 describe, respectively, the cases of the spanwise and azimuthal components of the cylinder. A numerical application of kinematic boundary conditions is then provided in section 6.4.

6.1. Splitting the full heat equation. The method presented herein can be very useful when a scheme that does not control boundary conditions is used. Indeed, the problem

$$\left\{\begin{array}{l} \dfrac{\partial\omega}{\partial t}-\nu\Delta\omega=g \ \text{ in } \ \Omega\times]0,T[,\\ \omega(x,0)=\omega_0(x) \ \text{ on } \ \Omega,\\ \nu\,\mathcal{L}_x\omega(x,t)=F(x,t) \ \text{ on } \ \partial\Omega\times]0,T[\end{array}\right. \tag{6.1}$$

can be solved in the inner part of Ω by a numerical method not consistent on boundaries (or leading to a prohibitive computational cost when consistent), i.e., approximating the problem

$$\left\{\begin{array}{l} \dfrac{\partial\omega}{\partial t}-\nu\Delta\omega=g \ \text{ in } \ \Omega\times]0,T[,\\ \omega(x,0)=\omega_0(x) \ \text{ on } \ \Omega \end{array}\right. \tag{6.2}$$

with arbitrary boundary conditions. Then one can measure the error on boundaries

$$q(x,t)=\nu\mathcal{L}_x\omega(x,t)$$

and use the present integral method to give explicitly (without the cost of another partial differential equation to solve) the solution of

$$\left\{\begin{array}{l} \dfrac{\partial\omega}{\partial t}-\nu\Delta\omega=0 \ \text{ in } \ \Omega\times]0,T[,\\ \omega(x,0)=0 \ \text{ on } \ \Omega,\\ \nu\,\mathcal{L}_x\omega(x,t)=F(x,t)-q(x,t) \ \text{ on } \ \partial\Omega\times]0,T[, \end{array}\right. \tag{6.3}$$

which is then approximated (at the appropriate order coming from Theorem 5.1 or Theorem 5.3) by

$$\omega(x,t)\simeq\int_0^t\int_{\partial\Omega}\frac{-2F(\xi,\tau)+2q(\xi,\tau)}{1+2\left(\bar\kappa(\xi)-\beta(\xi)\right)\sqrt{\nu\tau/\pi}}\,\frac{e^{-(x-\xi)^2/4\nu(t-\tau)}}{\left(4\pi\nu(t-\tau)\right)^{3/2}}\,\mathrm{d}\sigma(\xi)\mathrm{d}\tau, \tag{6.4}$$

which can itself be computed by a time quadrature (for t sufficiently small, typically for a time step when used in a numerical context) over a surface integral, using, for example, the midpoint rule:

$$\omega(x,t) \simeq \int_{\partial\Omega} \frac{-2t\,F(\xi,t/2)+2t\,q(\xi,t/2)}{1+2\,(\bar\kappa(\xi)-\beta(\xi))\,\sqrt{\nu t/2\pi}}\,\frac{e^{-(x-\xi)^2/2\nu t}}{(2\pi\nu t)^{3/2}}\,\mathrm{d}\sigma(\xi). \tag{6.5}$$

Note that the time quadrature based on an implicit Euler scheme is singular, while an explicit Euler scheme does not take curvature into account, despite the fact of being only first order.

Moreover, in order to provide more accuracy, this technique can be repeated on smaller intervals of time. The drawback in segmenting the time interval is that it reduces standard deviation of the Gaussian, which can possibly make the scheme underresolved, especially for three-dimensional computations. This approach, which couples the present integral method and a particle strength exchange (PSE) scheme (see [12]), has been successfully used for three-dimensional flow computations in [32], where boundary effects are the dominant effect.

Furthermore, one can also notice that (6.2)–(6.3) can be naturally parallelized if (6.2) is solved for homogeneous boundary conditions (i.e., $q=0$). Using this density estimation also provides a way to correct lack of regularity at the grid interface when performing domain decomposition at minimal cost.

6.2. The spanwise invariant cylinder. Let $\mathcal{B}_2(0,r)$ be the open ball of center 0 and radius r in $\mathbb{R}^2$, let $\Omega=\mathbb{R}^2\setminus\mathcal{B}_2(0,r)\times\mathbb{R}$ be an infinitely long cylindrical body of $\mathbb{R}^3$, and let $\mathcal{C}=\partial\mathcal{B}_2(0,r)$.

One considers the heat equation and the spanwise component of its solution, which is related to a pure Neumann boundary condition. One then gets the spanwise heat layer of unity:

$$\widetilde{\mathcal{H}}_z(x,t)1=\nu\int_0^t\int_{\partial\Omega} n_x\cdot\nabla G_{\xi,\tau}(x,t)\mathrm{d}\sigma(\xi)\mathrm{d}\tau.$$

Since the configuration is axisymmetric and spanwise invariant, one can set $x=(r,0,0)$ and $\xi=(r\cos\theta,r\sin\theta,0)$ without loss of generality. By means of integration in the spanwise direction, one gets

$$\begin{aligned}\widetilde{\mathcal{H}}_z(x,t)1&=-\nu\int_0^t\int_{\mathcal{C}}\frac{1}{4\pi\nu(t-\tau)}\,\frac{n_x\cdot(x-\xi)}{2\nu(t-\tau)}\,\exp\left(-\frac{(x-\xi)^2}{4\nu(t-\tau)}\right)\mathrm{d}\sigma(\xi)\mathrm{d}\tau\\&=-\frac{1}{\pi}\int_0^{\nu t}\int_{-\pi}^{\pi}\frac{r}{8u^2}(1-\cos\theta)\,\exp\left(-\frac{\left[(\cos\theta-1)^2+\sin^2\theta\right]r^2}{4u}\right)r\mathrm{d}\theta\mathrm{d}u\end{aligned}$$

with $u=\nu(t-\tau)$. Noticing the symmetry around $\theta=0$, one gets by parity

$$\widetilde{\mathcal{H}}_z(x,t)1=-\frac{r^2}{2\pi}\int_0^{\nu t}\int_0^1\frac{y}{u^2}\exp\left(-\frac{y\,r^2}{u}\right)\frac{1}{\sqrt{y(1-y)}}\,\mathrm{d}y\mathrm{d}u$$

with $y=(1-\cos\theta)/2$, which integrates successively into

$$\widetilde{\mathcal{H}}_z(x,t)1=-\frac{1}{2\pi}\int_0^1\frac{e^{-y\,r^2/\nu t}}{\sqrt{y(1-y)}}\,\mathrm{d}y=-\frac{1}{2}e^{-r^2/2\nu t}\,\mathrm{I}_0\left(\frac{r^2}{2\nu t}\right),$$

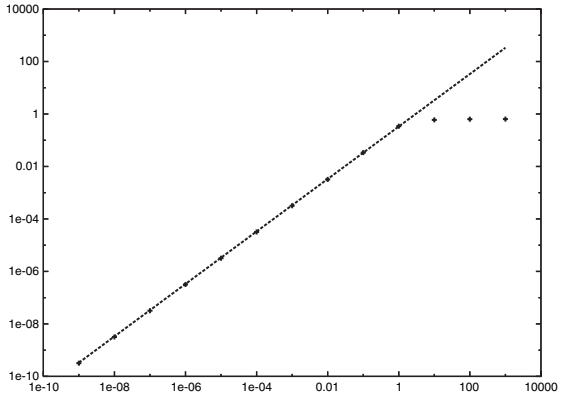


FIG. 5. *Numerical values of heat layer $\widetilde{\mathcal{H}}_z(x,t)f_x$ with respect to time t (dashed line represents the function $t/3$).*

where I_0 is the modified Bessel function of first kind (i.e., the solution of $x^2y'' + xy' - x^2y = 0$ with $y(0) = 1$ and $y'(0) = 0$). Since the mean curvature of $\partial\Omega$ is $\bar{\kappa} = 1/2r$, and noticing that

$$\lim_{x\to+\infty} \sqrt{x}\, e^{-x}\mathrm{I}_0(x) = \frac{1}{\sqrt{2\pi}},$$

one finally has that $\widetilde{\mathcal{H}}_z(x,t)1$ is equivalent to $-\bar{\kappa}\sqrt{\nu t/\pi}$ when t tends to 0, as expected by Lemma 4.2 and Theorem 5.1. One can notice that the present computation is performed on the exact surface $\partial\Omega$ instead of an approximation.

Moreover, in the present case, the density is smooth and the hypothesis of Theorem 5.3 holds; thus one can expect a full 3/2 order of convergence. Indeed, a cylinder presents no torsion, and one can show that

$$\lim_{x\to+\infty} \left(\sqrt{x}\, e^{-x}\mathrm{I}_0(x) - \frac{1}{\sqrt{2\pi x}}\right) x^{3/2} = \frac{1}{8\sqrt{2\pi}},$$

which proves the 3/2 order of heat layer error on both the exact surface and its quadratic approximation. This example illustrate the statement of Theorem 5.3, i.e., that no torsion implies no error at first order in time for a constant source.

In order to exhibit the limit convergence order of Theorem 5.1, one has to choose an example for which Theorem 5.3 is not valid. Since the cylinder is torsion-free, one has to choose a density which is 1-Hölder continuous without being differentiable. The Euclidean norm satisfies this condition, and using the notation already set above, one considers the density

$$f_x(\xi,\tau) = |x - \xi| = 2r\sqrt{y};$$

thus evaluating the heat layer at point x gives

$$\widetilde{\mathcal{H}}_z(x,t)f_x = -\frac{r}{\pi}\int_0^1 \frac{e^{-y\,r^2/\nu t}}{\sqrt{1-y}}\,\mathrm{d}y. \tag{6.6}$$

Integrating this integral symbolically is more difficult than the previous ones (though possible using erf functions). Figure 5 shows a certified 15-digit evaluation of the integral expression (6.6) with respect to time t. This actually exhibits a first order convergence since $f_x(x,t) = 0$, and consequently $\widetilde{\mathcal{H}}_z(x,t)f_x = f_x(x,t) + \mathcal{O}(t)$.

6.3. Vectorial kinematic boundary conditions and integral formulation of Chorin's algorithm in the cylinder case. In this section, one applies the present density estimation to Chorin's algorithm in the parabolic context (initially proposed in an hyperbolic context [7]), in the case of a circular cylinder.

Let $u : \Omega \times [0,T] \to \mathbb{R}^3$ be a velocity field satisfying the Stokes equations, and let ω be its associated vorticity field defined by $\omega = \mathrm{curl} u$ and satisfying the diffusion problem with kinematic boundary conditions:

$$\left\{ \begin{array}{l} \dfrac{\partial \omega}{\partial t} - \nu \Delta \omega = 0 \ \text{ in } \ \Omega \times]0,T[, \\ u(x,t) = 0 \ \text{ on } \ \partial\Omega \times]0,T[, \\ \omega(x,0) = \omega_0(x) \ \text{ on } \ \Omega. \end{array} \right. \tag{6.7}$$

While computing ω from u is obvious, building u from ω is more difficult. In the full space $\Omega = \mathbb{R}^3$, one can use the three-dimensional Green kernel $G(x) = (4\pi|x|)^{-1}$ through the Biot–Savart law

$$u = \nabla G \,\hat{*}\, \omega = \int_{\mathbb{R}^3} \nabla G(\cdot - x) \wedge \omega(x)\, \mathrm{d}x. \tag{6.8}$$

In the presence of boundaries, one can use symmetrizations of Biot–Savart laws around $\partial\Omega$ or more rigorously consider the single-layer integral formulation of (6.7) shown in [10]. Nevertheless, integral techniques are much less useful for this problem than the one presented herein, since the Green kernel and its gradient decrease much more slowly than Gaussian functions. A more competitive approach to computing the velocity from the vorticity is to introduce the Poisson equation on stream function with appropriate boundary conditions (see [30]) and eventually additional quantities such as potential stream to uncouple components of the three-dimensional Poisson equation (see [11]). In any case, it is possible to consider the operator $\mathcal{A}$, in the appropriate functional space, associating a velocity field $u = \mathcal{A}\omega$ to a vorticity field ω, satisfying $\mathrm{curl} u = \omega$ and $\mathrm{div} u = 0$ on Ω and $u \cdot n_x = 0$ on $\partial\Omega$.

One considers the infinitely long circular cylinder Ω of axis e_z and radius R, whose other tangential vector is denoted e_θ and whose normal vector is denoted e_r (that is, the standard cylindrical coordinates).

Applying Chorin's method (whose convergence is proved in [8] in the case of the Stokes equation and its rotational formulation), this problem can be reduced to two parabolic problems with Neumann and Robin–Fourier (see [11]) boundary conditions. The first has homogeneous boundary conditions, which in cylindrical coordinates reads as

$$\left\{ \begin{array}{l} \dfrac{\partial \omega^1}{\partial t} - \nu \Delta \omega^1 = 0 \ \text{ in } \ \Omega \times]0,T[, \\ \omega^1(x,0) = \omega_0(x) \ \text{ on } \ \Omega, \\ \nu \dfrac{\partial \omega^1_z}{\partial n} = 0 \ \text{ on } \ \partial\Omega \times]0,T[, \\ \nu \dfrac{\omega^1_\theta}{R} + \dfrac{\partial \omega^1_\theta}{\partial n} = 0 \ \text{ on } \ \partial\Omega \times]0,T[, \\ \omega^1_r = 0 \ \text{ on } \ \partial\Omega \times]0,T[, \end{array} \right. \tag{6.9}$$

whose divergence-free and no-slip-through associated velocity field $u^1 = \mathcal{A}\omega^1$ presents a priori nonzero tangential values (u^1 is usually called spurious velocity). One then considers a second diffusion problem with zero initial condition,

$$\left\{\begin{array}{l} \dfrac{\partial \omega^2}{\partial t} - \nu \Delta \omega^2 = 0 \ \text{ in } \ \Omega \times]0,T[, \\ \omega^2(x,0) = 0 \ \text{ on } \ \Omega, \\ \nu \dfrac{\partial \omega_z^2}{\partial n} = -\dfrac{\partial u_\theta}{\partial t} \ \text{ on } \ \partial\Omega \times]0,T[, \\ \nu \dfrac{\omega_\theta^2}{R} + \dfrac{\partial \omega_\theta^2}{\partial n} = \dfrac{\partial u_z}{\partial t} \ \text{ on } \ \partial\Omega \times]0,T[, \\ \omega_r^2 = 0 \ \text{ on } \ \partial\Omega \times]0,T[, \end{array}\right. \tag{6.10}$$

which allows us to link asymptotically (see [8]) the solution of (6.7) to the solutions of (6.9)–(6.10):

$$\omega^1(t) + \omega^2(t) = \omega(t) + \mathcal{O}(t). \tag{6.11}$$

One can immediately notice that problem (6.10) is a heat equation with a zero initial condition and a vectorial Robin–Fourier boundary condition, on which one can apply component by component the density estimation presented herein, as long as boundary conditions are expressed in a basis in which the fundamental solution of heat equation is still Gaussian. One can thus introduce the matrix

$$K = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \ N = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \text{ and } \ \Gamma_\theta = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6.12}$$

so that in the canonical basis $\omega = (\omega_x, \omega_y, \omega_z)$, the boundary operator of equation (6.10) reads as

$$\mathcal{L}_x \omega = \Gamma_{-\theta} K \, \Gamma_\theta \frac{\omega}{R} + \Gamma_{-\theta} N \, \Gamma_\theta \frac{\partial \omega}{\partial n} \tag{6.13}$$

when expressed at point $x = (R\cos\theta, R\sin\theta, z)$. This allows us to write the boundary conditions of (6.10) under the form

$$\nu \, \mathcal{L}_x \omega = \frac{\partial u}{\partial t} \wedge n_x. \tag{6.14}$$

One can use Theorem 5.1 and get by linearity

$$\widetilde{\mathcal{H}}(x,t) f \simeq \Gamma_{-\theta} \left(K - \frac{N}{2} \right) \Gamma_\theta \frac{f(x,t)}{R} \tag{6.15}$$

because the mean curvature of the cylinder in $\kappa = 1/2R$. Since the density of the potential giving ω satisfies the integral equation $\widetilde{\mu} - 2\widetilde{\mathcal{H}}\widetilde{\mu} = -2\partial_t u \wedge n$, one gets

$$\widetilde{\mu}(x,t) \simeq -2\Gamma_{-\theta} \left(\mathrm{Id} - \frac{\sqrt{\nu t/\pi}}{R}(2K - N) \right)^{-1} \Gamma_\theta \, \frac{\partial u}{\partial t} \wedge n_x \tag{6.16}$$

at the appropriate order, depending only on regularity of u since the cylinder is a $\mathcal{C}^\infty$ differentiable manifold.

Setting $\varepsilon = \sqrt{\nu t/\pi}/R$, (6.16) reads as

$$\widetilde{\mu}(x,t) \simeq -2\Gamma_{-\theta} P^{-1} \Gamma_\theta \frac{\partial u}{\partial t} \wedge n_x, \quad \text{where} \quad P = \begin{bmatrix} 1-2\varepsilon & 0 & 0 \\ 0 & 1-\varepsilon & 0 \\ 0 & 0 & 1+\varepsilon \end{bmatrix}. \tag{6.17}$$

This matrix P is obvious to inverse, as long as t is sufficiently small. One can notice that curvature effects are of different signs for ω_θ and ω_z due to the Dirichlet part in the azimuthal direction.

Note also that if one considers only the z direction (for a two-dimensional problem), the density estimation (6.17) can be written as

$$\widetilde{\mu}_z(x,t) \simeq \frac{+2}{1 + \frac{1}{R}\left(\sqrt{\nu t/\pi}\right)} \frac{\partial u_\theta}{\partial t}(x,t). \tag{6.18}$$

6.4. Numerical example of kinematic boundary conditions. One considers the diffusion problem with time-periodic unknown $\omega : \Omega \times]0, 2\pi] \mapsto \mathbb{R}^3$, still defined in a cylindrical domain Ω, and kinematic boundary conditions:

$$\begin{cases} \dfrac{\partial \omega}{\partial t} - \nu \Delta \omega = 0 \quad \text{in} \quad \Omega \times]0, 2\pi], \\ u(x,t) = \cos t \; e_\theta(x) \quad \text{on} \quad \partial\Omega \times]0, 2\pi], \end{cases} \tag{6.19}$$

where the velocity field $u(x,t)$ is built from ω by $u = \mathcal{A}\omega$, where operator $\mathcal{A}$ is based on formula (6.8), or by using a hybrid technique (see [11]).

Note that this problem is slightly more general than the one in the last section, because one has nonhomogeneous kinematic boundary conditions, whose main implication is that spurious velocity vanishes toward boundary value. Indeed, it has been shown (see [24]) that the parabolic problem

$$\begin{cases} \dfrac{\partial \omega}{\partial t} - \nu \Delta \omega = 0 \quad \text{in} \quad \Omega \times]0, 2\pi], \\ \nu \dfrac{\partial \omega_z}{\partial n}(x,t) = -\dfrac{\partial u_\theta}{\partial t}(x,t) = \sin t \quad \text{on} \quad \partial\Omega \times]0, 2\pi], \\ \nu \dfrac{\partial \omega_\theta}{\partial n}(x,t) = \dfrac{\partial u_z}{\partial t}(x,t) = 0 \quad \text{on} \quad \partial\Omega \times]0, 2\pi], \\ \omega_r = 0 \quad \text{on} \quad \partial\Omega \times]0, 2\pi] \end{cases} \tag{6.20}$$

approximates well problem (6.19) in this particular context (i.e., when residual velocities coming from (6.20) with homogeneous conditions vanish toward $\partial_t u$ as stated above). One can notice that the solution is then invariant in z and θ.

Then let the functions

$$\begin{cases} K^+(r) = \mathfrak{Re}\left(Ke_0(cr)\right) + \mathfrak{Im}\left(Ke_0(cr)\right), \\ K^-(r) = \mathfrak{Re}\left(Ke_0(cr)\right) - \mathfrak{Im}\left(Ke_0(cr)\right) \end{cases}$$

be based on Kelvin functions (sometimes also called Thompson functions), with $c = 1/\sqrt{\nu}$. The function

$$\omega^*(r,t) = \left(\alpha K^-(r) - \beta K^+(r)\right)\cos t - \left(\alpha K^+(r) + \beta K^-(r)\right)\sin t$$

is then a time-periodic solution of $\partial_t\omega - \nu\Delta\omega = 0$. Setting $A = c\sqrt{2}/2|\,Ke_1(c)\,|$ helps to check that $\beta - i\alpha = Ke_1(c)/A$ is the only pair of parameters that makes $\omega^* e_z$ satisfy (6.20).

Let us consider (6.20) over a time step $]t_n, t_{n+1}[$ with $t_n = n\delta t$ and its initial value denoted $\omega^n(x)$. The initial value for $n = 0$ is set with the exact solution $\omega_z^0(x) = \omega^*(|x|, t_n)$ and other components of ω^0 set to zero.

Playing with linearity of the heat equation (without approximation), (6.20) can be solved in two steps over $]t_n, t_{n+1}[$. The first step is the computation of the solution with arbitrary boundary conditions, in practice using a second order PSE scheme on a grid with a uniform cylindrical lattice (see [11]). The second step is the enforcement of the boundary condition, in the spirit of section 6.1.

Nevertheless, one can notice that truncated PSE schemes are consistent with a flux and can be tuned to provide homogeneous Neumann boundary conditions. The flux error denoted q in section 6.1 can consequently be set to 0 in equations (6.3)–(6.5). The two steps can be naturally parallelized, and the second step can be computed with the integral scheme presented herein in the pure Neumann boundary condition context, which, using formula (6.18), reads as

$$\omega_z(x, t_n + \delta t) \simeq \omega_z(x, t_n) - 2\int_0^{\delta t}\int_{\partial\Omega} \frac{G_{\xi,\tau}(x, \delta t)}{1 + \frac{1}{R}(\sqrt{\nu\tau/\pi})}\,\sin(t_n + \tau)\,\mathrm{d}\sigma(\xi)\mathrm{d}\tau,$$

where R is the cylinder radius, without any action on other components since 0 is solution. Note that Theorem 5.3 is valid only at $t_n = 0$; thus this density evaluation is first order (from Corollary 5.2). One can then use a time quadrature to compute $\omega(x, t_{n+1})$, such as using the midpoint rule (see (6.5) with $q = 0$).

This one-dimensional reducible example allows us to compare the three-dimensional algorithm presented herein with the two-dimensional algorithm from [24] given for Neumann boundary conditions, and compare both of them with the exact solution. These three quantities are plotted in Figure 6 at times $t = 0.75$, 1, 1.5, and 2. The curves show a good agreement qualitatively, but the main result is that the algorithm allows us to enforce very well the kinematic boundary conditions: the residual velocity (rebuilt from vorticity ω by the operator $\mathcal{A}$ defined in section 6.3) is close to 10^{-6}, and this code was run in simple precision.

7. Toroidal examples. This section aims to illustrate the optimality of the convergence ratio obtained by Theorem 5.1. In order to proceed, one has to consider an example for which Theorem 5.3 is not valid.

The differences in the hypotheses of Theorems 5.1 and 5.3 lead us to consider either a nondifferentiable 1-Hölder continuous function on a torsion-free surface (the case already studied in section 6.2), or a smooth function density defined on a surface presenting torsion (i.e., whose tensor of map third derivatives is not identically zero).

In order to build such a surface and analyze properties of the integrodifferential operator $\widetilde{\mathcal{H}}$ on it, one considers the heat equation with pure Neumann boundary conditions:

$$\left\{\begin{array}{l} \dfrac{\partial\omega}{\partial t} - \nu\Delta\omega = 0 \ \text{ in } \ \Omega\times]0, \delta t], \\ \nu\,\dfrac{\partial\omega}{\partial n}(x,t) = F(x,t) \ \text{ on } \ \partial\Omega\times]0,\delta t], \\ \omega(x,0) = 0 \ \text{ on } \ \partial\Omega, \end{array}\right. \tag{7.1}$$

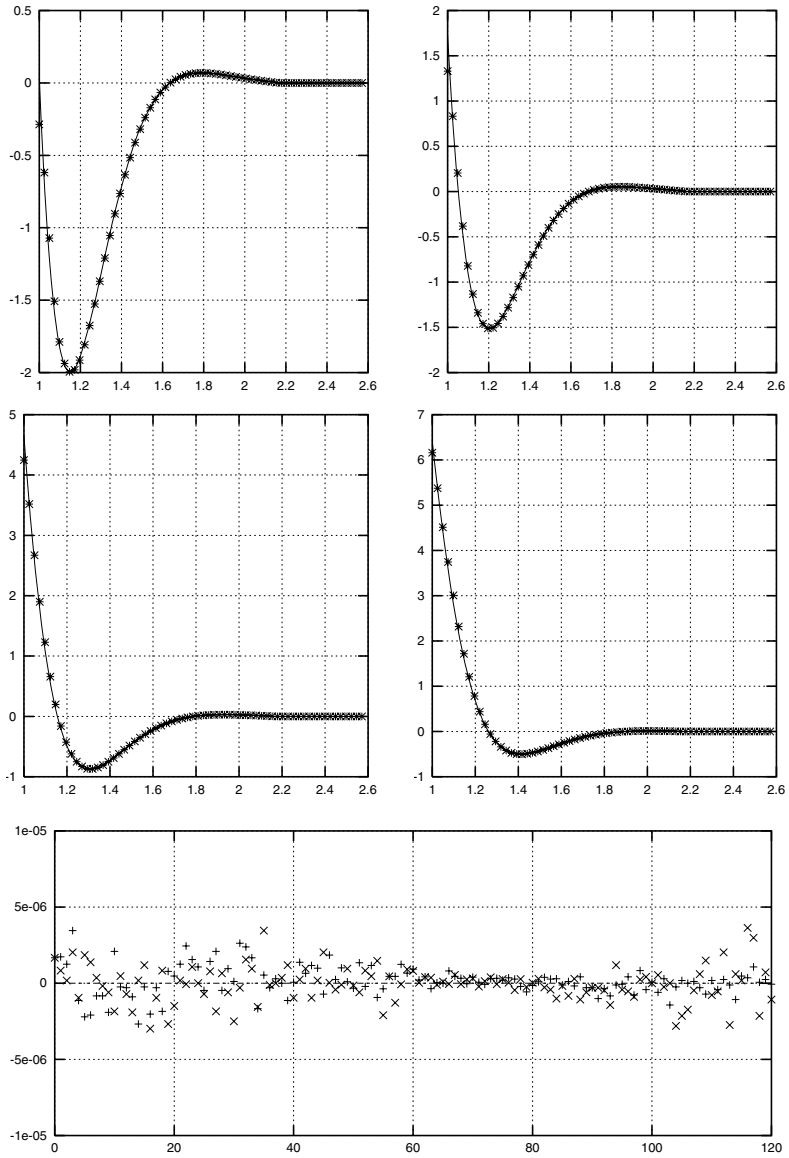


FIG. 6. *Numerical solutions compared to the exact axisymmetric and time-periodic solution of the heat equation at times* $t = 0.75, 1.0, 1.5, 2.0$*, with exact solution (–), three-dimensional scheme (+), and two-dimensional scheme (×) from* [24]*. Bottom: residual tangential velocity versus time step number* n *(time step* $\delta t = 0.025$*).*

where $\partial\Omega$ is successively the usual torus in section 7.1 and then modified with harmonic perturbations introducing torsion in sections 7.2 and 7.3. Large and small wavelength perturbations are involved, generating, respectively, a "twisted" and a "rippled" torus.

Instead of solving the heat equation, we will discuss properties of the related operator $\widetilde{\mathcal{H}}$ defined by formula (2.2), which reads as follows:

$$\widetilde{\mathcal{H}}(x,t)1 = -\frac{\nu}{16\pi^{3/2}} \int_0^t \int_{\partial\Omega} \frac{(x-\xi)\cdot n_x}{\big(\nu(t-\tau)\big)^{5/2}} e^{-|x-\xi|^2/4\nu(t-\tau)} \mathrm{d}\sigma(\xi)\mathrm{d}\tau. \tag{7.2}$$

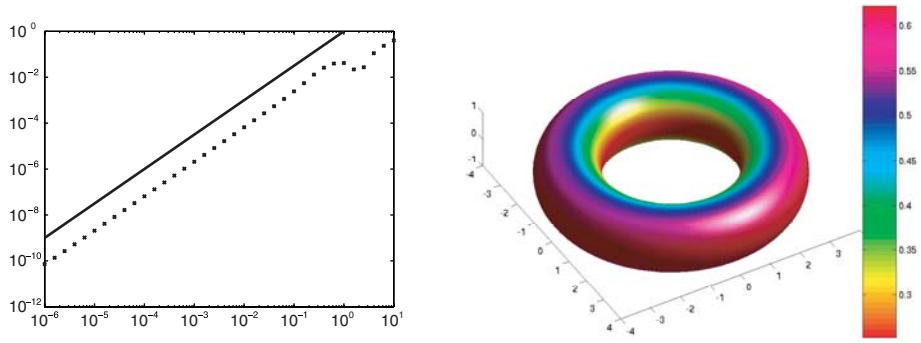


FIG. 7. *Right: plot of this surface, with color being the curvature coefficient given by formula* (7.12). *Left: pointwise error of the estimate of* $\widetilde{\mathcal{H}}(x,t)1$ *in* $\theta_0 = \pi/2$ *versus* νt ($\times$ *is formula* (7.13), *and the solid line is* $(\nu t)^{3/2}$).

7.1. The torsionless torus. One sets two radii r and $R > r$, and two angles $\theta \in [-\pi, \pi[$ and $\zeta \in [-\pi, \pi[$, both defined modulo 2π. The torus $\mathbb{T}$ is then the image of these domains by the function

$$f(\theta,\zeta) = \big((R + r\cos\theta)\cos\zeta, (R + r\cos\theta)\sin\zeta, r\sin\theta\big)\,, \tag{7.3}$$

which satisfies

$$\left\| \frac{\partial f}{\partial \theta} \wedge \frac{\partial f}{\partial \zeta} \right\|_2 = r\, I_\theta \quad \text{with} \quad I_\theta = (R + r\cos\theta) > 0. \tag{7.4}$$

The resulting surface is shown in Figure 7.

One can consider the points on the $\zeta = 0$ section, defined by

$$x = f(\theta_0, 0) = \big(I_{\theta_0}, 0, r\sin\theta_0\big)\,, \tag{7.5}$$

as arbitrary points of the surface without loss of generality, since the torus is globally ζ-invariant. The normal vector to $\mathbb{T}$ in x is then

$$n_x = (\cos\theta_0, 0, \sin\theta_0)\,. \tag{7.6}$$

Moreover, in order to describe a neighborhood of x, one also defines

$$\xi = f(\theta,\zeta) = \big(I_\theta\cos\zeta, I_\theta\sin\zeta, r\sin\theta\big)\,. \tag{7.7}$$

The integrodifferential operator defined by formula (7.2) reads as

$$\widetilde{\mathcal{H}}(x,t)1 = -\frac{1}{2\pi^{3/2}} \int_0^{4\nu t} \int_{\partial\Omega} \frac{(x-\xi)\cdot n_x}{u^{5/2}} e^{-|x-\xi|^2/u} \mathrm{d}\sigma(\xi)\mathrm{d}u \tag{7.8}$$

for $u = 4\nu(t-\tau)$. In order to obtain a two-dimensional integral calculus, one can set

$$U_t(\theta,\zeta) = \frac{(x - \xi(\theta,\zeta))^2}{4\nu t} \tag{7.9}$$

so that

$$\int_0^{4\nu t} e^{-(x-\xi(\theta,\zeta))^2/u} u^{-5/2}\,\mathrm{d}u = \frac{\mathrm{E}\left(U_t(\theta,\zeta)^{1/2}\right) - U_t(\theta,\zeta)^{1/2} e^{-U_t(\theta,\zeta)}}{U_t(\theta,\zeta)^{3/2}}\,(4\nu t)^{3/2}, \tag{7.10}$$

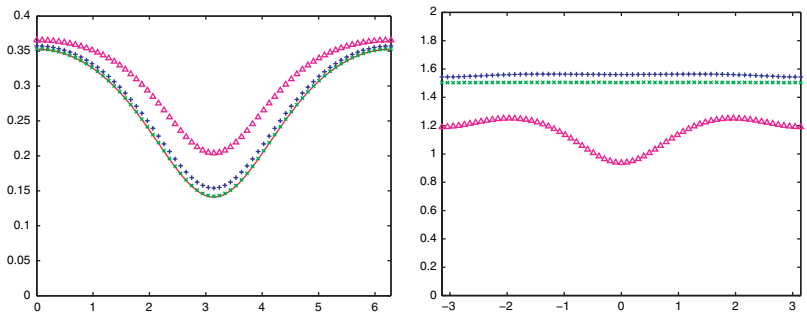


FIG. 8. *Convergence of estimates for the torsionless torus. Left: uniform convergence of* $(\nu\delta t)^{-1/2}\widetilde{\mathcal{H}}(x,\delta t)1$ *toward* $\bar\kappa(x)/\sqrt{\pi}$ *versus angle* θ_0. *Right: resulting numerical order* $N(\nu\delta t,\theta_0)$ *obtained by formula* (7.14) *versus angle* θ_0. *Legend:* $\triangle$ *is* $\nu\delta t = 1$, $+$ *is* $\nu\delta t = 0.1$, *and* $\times$ *is* $\nu\delta t = 0.01$.

where E is the scaled erf complementary function defined by

$$\mathrm{E}(x) = -\int_x^{+\infty} e^{-z^2}\mathrm{d}z = \frac{\sqrt{\pi}}{2}(\mathrm{erf}(x) - 1).$$

This leads to a new expression for $\widetilde{\mathcal{H}}(x,t)1$:

(7.11)
$$-\frac{r}{(4\pi\nu t)^{3/2}}\int_0^{\pi}\int_{\theta_0-\pi}^{\theta_0+\pi}\frac{\mathrm{E}\left(U_t(\theta,\zeta)^{1/2}\right) - U_t(\theta,\zeta)^{1/2}e^{-U_t(\theta,\zeta)}}{U_t(\theta,\zeta)^{3/2}}\,(x-\xi(\theta,\zeta))\cdot n_x\, I_\theta\,\mathrm{d}\theta\mathrm{d}\zeta.$$

This two-dimensional integral is then computed by a fifth order Gauss–Legendre quadrature formula with 2000^3 elements, once the singularity in $(\theta_0, 0)$ has been smoothed by setting $\theta = \theta_0 \pm \hat\theta^s$ and $\zeta = \hat\zeta^s$ with $s = 3$.

It can be shown (but is not developed herein) that the mean curvature in $x = f(\theta,\zeta)$ is given by

$$\bar\kappa(x) = \frac{1}{2}\left(\frac{1}{r} + \frac{\cos\theta}{R + r\cos\theta}\right). \tag{7.12}$$

Theorem 5.3 then predicts, since the torus is a $\mathcal{C}^\infty$ torsionless two-dimensional submanifold of $\mathbb{R}^3$, that

$$\widetilde{\mathcal{H}}(x,t)1 - \bar\kappa(x)\sqrt{\frac{\nu t}{\pi}} = \mathcal{O}(t^{3/2}). \tag{7.13}$$

One verifies that this 3/2 order is reached with the computation of the difference of the two expressions above for $x = \pi/2$. The left-hand picture in Figure 7 shows that indeed the difference scales as $t^{3/2}$. In order to measure the convergence more uniformly, one introduces the numerical order of convergence $N(t)$ defined by

$$N(\nu\delta t,\theta_0) = \log_{10}\left(\frac{\widetilde{\mathcal{H}}(x,\delta t)1 - \bar\kappa(x)\sqrt{\frac{\nu\delta t}{\pi}}}{\widetilde{\mathcal{H}}(x,\delta t/10)1 - \bar\kappa(x)\sqrt{\frac{\nu\delta t/10}{\pi}}}\right) \tag{7.14}$$

with $x = f(\theta_0, 0)$ chosen on the $\zeta = 0$ section (which is the generality since this torus is ζ-invariant). This function is plotted in the right-hand graph of Figure 8 and shows a convergence toward the 3/2 order everywhere. The convergence order suggested by Theorem 5.3 is consequently optimal.

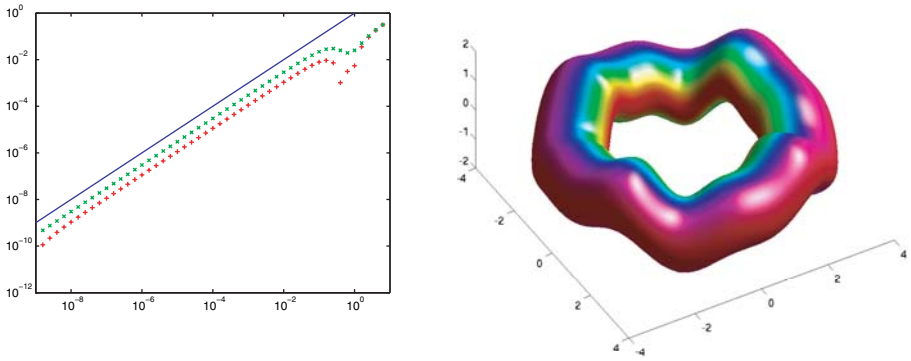


FIG. 9. *Right: plot of the twisted torus (color is the same as for Figure* 7*). Left: pointwise error of the estimate of* $\widetilde{\mathcal{H}}(x,t)1$ *versus* νt *(*$\times$ *is formula* (7.13) *for* $\theta_0 = \pi/4$, $+$ *is formula* (7.13) *for* $\theta_0 = \pi/2$, *and the solid line is* νt*).*

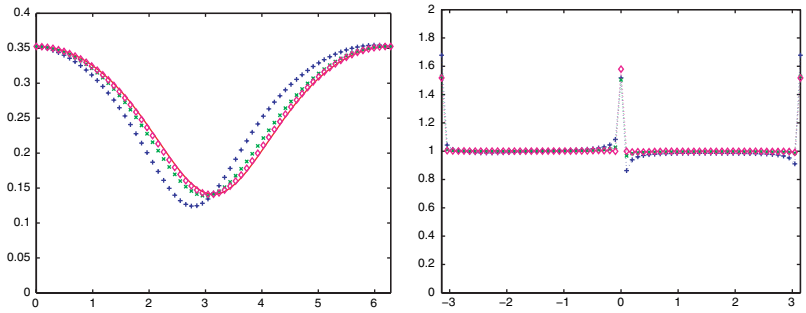


FIG. 10. *Convergence of estimate for the twisted torus. Left: uniform convergence of* $(\nu\delta t)^{-1/2}\widetilde{\mathcal{H}}(x,\delta t)1$ *toward* $\bar{\kappa}(x)/\sqrt{\pi}$ *versus angle* θ_0*. Right: resulting numerical order* $N(\nu\delta t, \theta_0)$ *obtained by formula* (7.14) *versus angle* θ_0*. Legend:* $+$ *is* $\nu\delta t = 10^{-2}$, $\times$ *is* $\nu\delta t = 10^{-3}$, *and* $\triangle$ *is* $\nu\delta t = 10^{-4}$.

7.2. The "twisted" torus. In order to introduce torsion effects to the geometry, one considers two strictly positive numbers $\bar{A}$ and $\bar{m}$ and the function

$$g(\zeta) = \bar{A}\,(2\sin(\bar{m}\zeta) - \sin(2\bar{m}\zeta))\,, \tag{7.15}$$

which satisfies $g(0) = g'(0) = g''(0) = 0$ and $g'''(0) = 6\bar{A}\bar{m}^3 \neq 0$.

The surface defined by

$$f(\theta,\zeta) + g(\zeta)e_z \tag{7.16}$$

with $\bar{A} = 0.2$ and $\bar{m} = 4$ (where e_z denotes the third vector of the canonical basis of $\mathbb{R}^3$) is called herein the "twisted" torus and is plotted in Figures 9 and 10.

It presents nonzero torsion everywhere on the section $\zeta = 0$, except for $\theta_0 = 0$ and $\theta_0 = \pi$ for which the mapping is tangential (thus introduces no torsion), without changing curvature, slope, and location of this section when compared to the torsionless torus discussed in the last section (note that in this case the Jacobian is not as obvious as before and is thus not explicitly given herein).

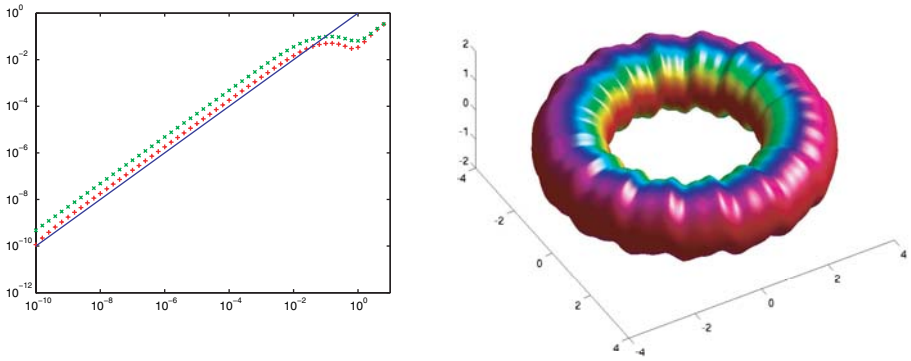


FIG. 11. *Right: plot of the rippled torus (color is the same as for Figure 7). Left: pointwise error of the estimate of* $\widetilde{\mathcal{H}}(x,t)1$ *versus* νt *(*$\times$ *is formula* (7.13) *for* $\theta_0 = \pi/4$, $+$ *is formula* (7.13) *for* $\theta_0 = \pi/2$, *and the solid line is* νt*).*

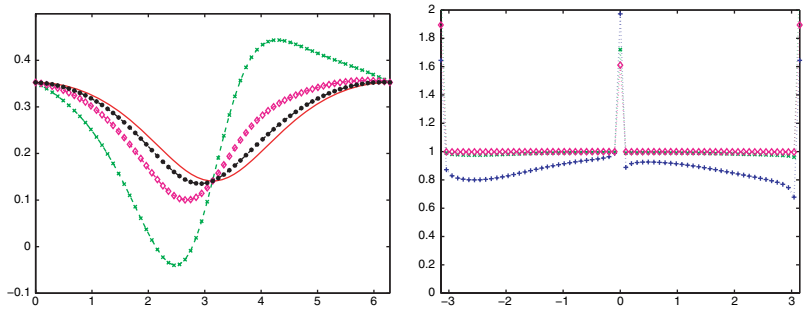


FIG. 12. *Convergence of estimate for the rippled torus. Left: uniform convergence of* $(\nu\delta t)^{-1/2}\widetilde{\mathcal{H}}(x,\delta t)1$ *toward* $\bar{\kappa}(x)/\sqrt{\pi}$ *versus angle* θ_0*. Right: resulting numerical order* $N(\nu\delta t, \theta_0)$ *obtained by formula* (7.14) *versus angle* θ_0*. Legend:* $+$ *is* $\nu\delta t = 10^{-2}$, $\times$ *is* $\nu\delta t = 10^{-3}$, $\triangle$ *is* $\nu\delta t = 10^{-4}$, *and* $*$ *is* $\nu\delta t = 10^{-5}$.

According to Theorem 5.1 and the strong need in Theorem 5.3 of the torsion-free surface to reach order 3/2, one expects in the present case to observe a first order convergence rate, except for the singular value of θ_0 mentioned above.

7.3. The "rippled" torus. In this section one considers the same kind of torus as in section 7.2, with perturbation parameters chosen as $\bar{A} = 0.05$ and $\bar{m} = 16$. This example provides a torsion 16 times stronger than the previous one, which makes torsion effects even clearer. Indeed, one can see in Figure 11 that convergence is globally first order, as observed for the twisted torus, but with a shift of accuracy due to stronger torsion effects.

Figure 12 shows that the first order is induced by a broken symmetry led by the torsion tensor, but also that the order is still constant over the section, except for $\theta_0 = 0$ for which the torsion effect is tangential; thus it does not act on body torsion and allows a 3/2 convergence order at this special point.

8. Conclusion. In this article, we have proved that the solution of the heat equation whose sources are provided only at boundaries can be explicited analytically up to order 3/2, exhibiting a square root in time depending on the boundary curvature

and the Dirichlet part of the Robin–Fourier coefficients of the boundary conditions. The solution is expressed in its integral formulation, involving a Gaussian kernel and a surface density. Most of the present study focuses on properties of this density.

The main result obtained herein is that since the density is analytically provided, one gets a very fast estimation of the solution of the heat equation for early times. This leads to a fast numerical scheme for kinematic boundary conditions or in addition to a scheme not satisfying algebraically the Robin–Fourier boundary condition.

The order depends on whether the manifold defining the domain boundary is torsionless or not, and on the manifold and the density regularities. Since small times are considered, the Gaussian kernel of the heat equation has small standard deviation, and thus its effect is localized (if not compactly supported). Therefore classical results have been extended to a class of noncompact manifolds, satisfying a few properties denoted (C1)–(C5).

We first discussed the error estimation due to restriction and then the error resulting from the substitution of the manifold by its best quadratic approximant. The error coming from the flattening process was also discussed, obtaining finally an integral expression which can be symbolically carried out.

Several applications illustrate that the limit convergence rates given by Theorems 5.1 and 5.3 are optimal. As examples, we investigate numerically the two- and three-dimensional cylinders, whose different eigenvectors of the curvature matrix induce density anisotropy. These cylindrical examples allow us to show the effect of density regularity on the double heat layer. The effect of manifold torsion is finally investigated for smoothly perturbated toroidal manifolds.

Acknowledgments. The author thanks Georges-Henri Cottet and Petros Koumoutsakos for their helpful contribution in the early work leading to the three-dimensional results.

REFERENCES

[1] S. Alinhac and P. Gérard, *Opérateurs pseudo-différentiels et théorème de Nash-Moser*, Editions du CNRS, Meudon, France, 1991.

[2] C. Anderson and C. Greengard, *On vortex methods*, SIAM J. Numer. Anal., 22 (1985), pp. 413–440.

[3] J. T. Beale, *A convergent boundary integral method for three-dimensional water waves*, Math. Comp., 70 (2000), pp. 977–1029.

[4] J. T. Beale, T. Y. Hou, and J. S. Lowengrub, *Convergence of a boundary integral method for water waves*, SIAM J. Numer. Anal., 33 (1996), pp. 1797–1843.

[5] J. Chazarin and A. Piriou, *Introduction à la théorie des équations aux dérivées partielles linéaires*, Gauthier-Villars, Paris, 1981.

[6] A. K. Chaniotis, D. Poulikakos, and P. Koumoutsakos, *Remeshed smoothed particle hydrodynamics for the simulation of viscous and heat conducting flows*, J. Comput. Phys., 182 (2002), pp. 67–90.

[7] A. J. Chorin, *Numerical study of slightly viscous flow*, J. Fluid Mech., 57 (1973), pp. 785–796.

[8] G.-H. Cottet, *A vorticity creation algorithm for the Navier–Stokes equations in arbitrary domain*, in Navier–Stokes Equations and Related Nonlinear Problems, A. Sequeira, ed., Plenum, New York, 1995.

[9] G.-H. Cottet and P. D. Koumoutsakos, *Vortex Methods, Theory and Practice*, Cambridge University Press, Cambridge, UK, 2000.

[10] G.-H. Cottet and P. Poncet, *Particle methods for direct numerical simulations of three-dimensional wakes*, J. Turbul., 3 (2002), pp. 1–9.

[11] G.-H. Cottet and P. Poncet, *Advances in direct numerical simulations of 3D wall-bounded flows by vortex-in-cell methods*, J. Comput. Phys., 193 (2004), pp. 136–158.

[12] P. Degond and S. Mas-Gallic, *The weighted particle method for convection-diffusion equations*, Math. Comp., 53 (1989), pp. 485–526.

[13] L. Dragoş and A. Dinu, *A direct boundary integral method for the three-dimensional lifting flow*, Comput. Methods Appl. Mech. Engrg., 127 (1995), pp. 357–370.
[14] A. Friedman, *Partial Differential Equations of Parabolic Type*, Prentice-Hall, Englewood Cliffs, NJ, 1964.
[15] L. Greengard and V. Rokhlin, *A fast algorithm for particle simulation*, J. Comput. Phys., 73 (1987), pp. 325–348.
[16] Ph. Guillaume, A. Huard, and C. Le Calvez, *A block constant approximate inverse for preconditioning large linear systems*, SIAM J. Matrix Anal. Appl., 24 (2003), pp. 822–851.
[17] D. J. Haroldsen and D. Meiron, *Numerical calculation of three-dimensional interfacial potential flows using the point vortex method*, SIAM J. Sci. Comput., 20 (1998), pp. 648–683.
[18] J. L. Hess, *Review of integral-equation techniques for solving potential flow problems with emphasis on the surface-source method*, Comput. Methods Appl. Mech. Engrg., 5 (1975), pp. 145–196.
[19] J. L. Hess, *Panel methods in computational fluid dynamics*, Ann. Rev. Fluid Mech., 22 (1990), pp. 255–274.
[20] S. Itô, *Fundamental solutions of parabolic differential equations and boundary value problems*, Japan J. Math., 27 (1957), pp. 55–102.
[21] J. Katz and A. Plotkin, *Low-Speed Aerodynamics*, McGraw-Hill, New York, 1991.
[22] J. Kim and P. Moin, *Application of a fractional-step method to incompressible Navier–Stokes equations*, J. Comput. Phys., 59 (1985), pp. 308–323.
[23] R. B. Kinney and Z. M. Cielak, *Analysis of unsteady viscous flow past an airfoil. Part* I: *Theoretical development*, AIAA J., 15 (1977), pp. 1712–1717.
[24] P. D. Koumoutsakos, A. Leonard, and F. Pepin, *Boundary conditions for viscous vortex methods*, J. Comput. Phys., 113 (1994), p. 52.
[25] P. D. Koumoutsakos and A. Leonard, *High-resolution simulations of the flow around an impulsively started cylinder using vortex methods*, J. Fluid Mech., 296 (1995), pp. 1–38.
[26] S. V. Kozlov, I. K. Lifanov, and A. A. Mikhailov, *A new approach to mathematical modelling of flow of ideal fluid around bodies*, Soviet J. Numer. Anal. Math. Modelling, 6 (1991), pp. 209–222.
[27] E. E. Levi, *Sulle equazioni lineari totalmente ellittiche alle derivate parziali*, Rend. Circ. Mat. Palermo, 24 (1907), pp. 275–317.
[28] O. A. Ladyzhenskaya, V. A. Solonnikov, and N. N. Ural'ceva, *Linear and Quasilinear Equations of Parabolic Type*, AMS, Providence, RI, 1968.
[29] M. J. Lighthill, *Boundary Layer Theory*, Oxford University Press, Oxford, UK, 1963.
[30] M. L. Ould-Sahili, G.-H. Cottet, and M. El Hamraoui, *Blending finite-differencies and vortex methods for incompressible flow computations*, SIAM J. Sci. Comput., 22 (2000), pp. 1655–1674.
[31] W. Pogorzelski, *Propriétés des intégrales de l'équation parabolique normale*, Ann. Polon. Math., 4 (1957), pp. 61–92.
[32] P. Poncet, *Topological aspects of the three-dimensional wake behind rotary oscillating circular cylinders*, J. Fluid Mech., 517 (2004), pp. 27–53.
[33] P. Ramachandran, S. C. Rajan, and M. Ramakrishna, *A fast, two-dimensional panel method*, SIAM J. Sci. Comput., 24 (2003), pp. 1864–1878.
[34] A. Rathsfeld, *The invertibility of the double layer potential operator in the space of continuous functions defined on a polyhedron: The panel method*, Appl. Anal., 45 (1992), pp. 135–177.
[35] J. E. Romate, *Local error analysis of three-dimensional panel methods in terms of curvilinear surface coordinates*, SIAM J. Numer. Anal., 27 (1990), pp. 529–542.
[36] M. Taylor, *Pseudodifferential Operators*, Princeton University Press, Princeton, NJ, 1981.
[37] J. C. Wu, *Numerical boundary conditions for viscous flow problems*, AIAA J., 14 (1976), pp. 1042–1049.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2298–2320

THE FINITE ELEMENT APPROXIMATION OF THE NONLINEAR POISSON–BOLTZMANN EQUATION*

LONG CHEN†, MICHAEL J. HOLST‡, AND JINCHAO XU§

Abstract. A widely used electrostatics model in the biomolecular modeling community, the nonlinear Poisson–Boltzmann equation, along with its finite element approximation, are analyzed in this paper. A regularized Poisson–Boltzmann equation is introduced as an auxiliary problem, making it possible to study the original nonlinear equation with delta distribution sources. A priori error estimates for the finite element approximation are obtained for the regularized Poisson–Boltzmann equation based on certain quasi-uniform grids in two and three dimensions. Adaptive finite element approximation through local refinement driven by an a posteriori error estimate is shown to converge. The Poisson–Boltzmann equation does not appear to have been previously studied in detail theoretically, and it is hoped that this paper will help provide molecular modelers with a better foundation for their analytical and computational work with the Poisson–Boltzmann equation. Note that this article apparently gives the first rigorous convergence result for a numerical discretization technique for the nonlinear Poisson–Boltzmann equation with delta distribution sources, and it also introduces the first provably convergent adaptive method for the equation. This last result is currently one of only a handful of existing convergence results of this type for nonlinear problems.

Key words. nonlinear Poisson–Boltzmann equation, finite element methods, a priori and a posteriori error estimate, convergence of adaptive methods

AMS subject classifications. 65N12, 65N30, 65N50, 65Y20, 92B05

DOI. 10.1137/060675514

1. Introduction.

In this paper, we shall design and analyze finite element approximations of a widely used electrostatics model in the biomolecular modeling community, the nonlinear Poisson–Boltzmann equation (PBE):

$$-\nabla \cdot (\varepsilon \nabla \tilde{u}) + \bar{\kappa}^2 \sinh(\tilde{u}) = \sum_{i=1}^{N_m} q_i \delta_i \quad \text{in } \mathbb{R}^d,\ d = 2, 3, \tag{1.1}$$

where the dielectric ε and the modified Debye–Hückel parameter $\bar{\kappa}$ are piecewise constants in domains Ω_m (the domain for the biomolecule of interest) and Ω_s (the domain for a solvent surrounding the biomolecule), and $\delta_i := \delta(x - x_i)$ is a Dirac distribution at point x_i. The importance of (1.1) in biomolecular modeling is well-established; cf. [14, 44] for thorough discussions. Some analytical solutions are known,

*Received by the editors November 20, 2006; accepted for publication (in revised form) June 19, 2007; published electronically October 24, 2007.
http://www.siam.org/journals/sinum/45-6/67551.html

†Department of Mathematics, University of California at San Diego, La Jolla, CA 92093. Current address: Department of Mathematics, University of California at Irvine, CA 92697 (chenlong@math.uci.edu). This author's research was supported in part by NSF awards 0411723 and 022560, in part by DOE awards DE-FG02-04ER25620 and DE-FG02-05ER25707, and in part by NIH award P41RR08605.

‡Department of Mathematics, University of California at San Diego, La Jolla, CA 92093 (mholst@math.ucsd.edu). This author's research was supported in part by NSF awards 0411723 and 022560, in part by DOE awards DE-FG02-04ER25620 and DE-FG02-05ER25707, and in part by NIH award P41RR08605.

§The School of Mathematical Science, Peking University, Beijing, 100871 China and Department of Mathematics, Pennsylvania State University, University Park, PA 16801 (xu@math.psu.edu). This author's research was supported in part by NSF DMS0308946, DMS-0619587, DMS-0609727, and NSFC-10528102.

but only for unrealistic structure geometries, and usually only for linearizations of the equation; cf. [30] for a collection of these solutions and for references to the large amount of literature on analytical solutions to the PBE and similar equations. The current technological advances are more demanding and require the solution of highly nonlinear problems in complicated geometries. To this end, numerical methods, including the finite element method, are widely used to solve the nonlinear PBE [30, 31, 5, 6, 45, 19, 57].

The main difficulties for the rigorous analysis and provably good numerical approximation of solutions to the nonlinear Poisson–Boltzmann equation include: (1) Dirac distribution sources, (2) exponential rapid nonlinearities, and (3) discontinuous coefficients. We shall address these difficulties in this paper. To deal with the δ distribution sources, we decompose $\tilde{u}$ as an unknown function in H^1 and a known singular function, namely,

$$\tilde{u} = u + G, \qquad \text{with } G = \sum_{i=1}^{N_m} G_i,$$

where G_i is the fundamental solution of $-\varepsilon_m \Delta G_i = q_i \delta_i$ in $\mathbb{R}^d$. Substituting this decomposition into the PBE, we then obtain the so-called regularized Poisson–Boltzmann equation (RPBE):

$$-\nabla \cdot (\varepsilon \nabla u) + \bar{\kappa}^2 \sinh(u + G) = \nabla \cdot ((\varepsilon - \varepsilon_m)\nabla G) \quad \text{in } \mathbb{R}^d,\ d = 2, 3.$$

The singularities of the δ distributions are transferred to G, which then exhibits degenerate behavior at each $\{x_i\} \subset \Omega_m$. At those points, both $\sinh G(x_i)$ and $\nabla G(x_i)$ exhibit blowup. However, since G is known analytically, one avoids having to build numerical approximations to G. Moreover, both of the coefficients $\bar{\kappa}$ and $\varepsilon - \varepsilon_m$ are zero inside Ω_m where the blowup behavior arises. Due to this cutoff nature of coefficients, we obtain a well-defined nonlinear second-order elliptic equation for the regularized solution u with a source term in H^{-1}. We will show that it also admits a unique solution $u \in H^1$, even though the original solution $\tilde{u} \notin H^1$ due to the singularities present in G.

Singular function expansions are a common technique in applied and computational mathematics for this type of singularity; this type of expansion has been previously proposed for the Poisson–Boltzmann equation in [59] and was shown (empirically) to allow for more accurate finite difference approximations. In their work, the motivation for the technique was the poor discrete approximation of arbitrarily placed delta distributions using only the fixed corners of uniform finite difference meshes. In the present work, our interest is in developing finite element methods using completely unstructured meshes, so we are able to place the delta distributions precisely where they should be and do not have this problem with approximate delta function placement. Our motivation here for considering a singular function expansion is rather that the solution to the Poisson–Boltzmann equation is simply not smooth enough to either analyze or approximate using standard methods without using some sort of two-scale or multiscale expansion that represents the nonsmooth part of the solution analytically. In fact, it will turn out that expanding the solution into the sum of three functions, namely, a known singular function, an unknown solution to a linear auxiliary problem, and an unknown solution to a second nonlinear auxiliary problem, is the key to establishing some fundamental results and estimates for the continuous

problem and is also the key to developing a complete approximation theory for the discrete problem as well as provably convergent nonadaptive and adaptive numerical methods.

Starting with some basic results on existence, uniqueness, and a priori estimates for the continuous problem, we analyze the finite element discretization and derive discrete analogues of the continuous results to show that discretization leads to a well-posed discrete problem. Using maximum principles for the continuous and discrete problems, we derive a priori L^∞-estimates for the continuous and discrete solutions to control the nonlinearity, allowing us to obtain a priori error estimates for our finite element approximation of the form

$$\|u - u_h\|_1 \lesssim \inf_{v_h \in V_D^h} \|u - v_h\|_1,$$

where V_D^h is the linear finite element subspace defined over quasi-uniform triangulations with a certain boundary condition, and u_h is the finite element approximation of u in V_D^h. The result is *quasi-optimal* in the sense it implies that the finite element approximation to the RPBE is within a constant of being the best approximation from the subspace V_D^h. After establishing these results for finite element approximations, we describe an adaptive approximation algorithm that uses mesh adaptation through local refinement driven by a posteriori error estimates. The adaptive algorithm can be viewed as a mechanism for dealing with the primary remaining difficulty in the RPBE, namely, the discontinuities of the coefficients across the interface between the solvent and the molecular regions. Finally, we shall prove that our adaptive finite element method will produce a sequence of approximations that converges to the solution of the continuous nonlinear PBE. This last result is one of only a handful of existing results of this type for nonlinear elliptic equations (the others being [24, 49, 15]).

The outline of this paper is as follows. In section 2, we give a brief derivation and overview of the Poisson–Boltzmann equation. In section 3, we derive a regularized form of the Poisson–Boltzmann equation by using a singular function expansion. In section 4, we give some basic existence and uniqueness results for the RPBE. In section 5, we derive an a priori L^∞-estimate for the continuous problem. After introducing finite element methods for the RPBE, in section 6 we derive an analogous a priori L^∞-estimate for the discrete problem, and based on this we obtain a quasi-optimal a priori error estimate for the finite element approximation. In section 7, we describe the adaptive algorithm, present an a posteriori error estimate, and prove a general convergence result for the algorithm. In the last section, we summarize our work and give further remarks on the practical aspects using results in the present paper.

2. The Poisson–Boltzmann equation. In this section we shall give a brief introduction to the nonlinear Poisson–Boltzmann equation. A detailed derivation can be found in [48, 30].

The nonlinear PBE, a second-order nonlinear partial differential equation, is fundamental to Debye–Hückel continuum electrostatic theory [22]. It determines a dimensionless potential around a charged biological structure immersed in a salt solution. The PBE arises from the Gauss law, represented mathematically by the Poisson equation, which relates the electrostatic potential Φ in a dielectric to the charge density ρ:

$$-\nabla \cdot (\varepsilon \nabla \Phi) = \rho,$$

where ε is the dielectric constant of the medium and here is typically piecewise constant. Usually it jumps by one or two orders of magnitude at the interface between

the charged structure (a biological molecular or membrane) and the solvent (a salt solution). The charge density ρ consist of two components: $\rho = \rho_{\text{macro}} + \rho_{\text{ion}}$. For the macromolecule, the charge density is a summation of δ distributions at N_m point charges in the point charge behavior, i.e.,

$$\rho_{\text{macro}}(x) = \sum_{i=1}^{N_m} q_i \delta(x - x_i), \qquad q_i = \frac{4\pi e_c^2}{\kappa_B T} z_i,$$

where $\kappa_B > 0$ is the Boltzmann constant, T is the temperature, e_c is the unit of charge, and z_i is the amount of charge.

For the mobile ions in the solvent, the charge density ρ_{ion} cannot be given in a deterministic way. Instead it will be given by the Boltzmann distribution. If the solvent contains N types of ions, of valence Z_i and of bulk concentration c_i, then a Boltzmann assumption about the equilibrium distribution of the ions leads to

$$\rho_{\text{ion}} = \sum_{i=1}^{N} c_i Z_i e_c \exp\left(-Z_i \frac{e_c \Phi}{\kappa_B T}\right).$$

For a symmetric $1:1$ electrolyte, $N = 2$, $c_i = c_0$, and $Z_i = (-1)^i$, which yields

$$\rho_{\text{ion}} = -2c_0 e_c \sinh\left(\frac{e_c \Phi}{\kappa_B T}\right).$$

We can now write the PBE for modeling the electrostatic potential of a solvated biological structure. Let us denote the molecule region by $\Omega_m \subset \mathbb{R}^d$ and consider the solvent region $\Omega_s = \mathbb{R}^d \backslash \bar{\Omega}_m$. We use $\tilde{u}$ to denote the dimensionless potential and $\bar{\kappa}^2$ to denote the modified Debye–Hückel parameter (which is a function of the ionic strength of the solvent). The nonlinear Poisson–Boltzmann equation is then

$$-\nabla \cdot (\varepsilon \nabla \tilde{u}) + \bar{\kappa}^2 \sinh(\tilde{u}) = \sum_{i=1}^{N_m} q_i \delta_i \quad \text{in } \mathbb{R}^d, \tag{2.1}$$

$$\tilde{u}(\infty) = 0, \tag{2.2}$$

where

$$\varepsilon = \begin{cases} \varepsilon_m & \text{if} \quad x \in \Omega_m, \\ \varepsilon_s & \text{if} \quad x \in \Omega_s, \end{cases} \qquad \text{and} \qquad \bar{\kappa} = \begin{cases} 0 & \text{if} \quad x \in \Omega_m, \\ \sqrt{\varepsilon_s}\kappa > 0 & \text{if} \quad x \in \Omega_s. \end{cases}$$

It has been determined empirically that $\varepsilon_m \approx 2$ and $\varepsilon_s \approx 80$. The structure itself (e.g., a biological molecule or a membrane) is represented implicitly by ε and $\bar{\kappa}$, as well as explicitly by the N_m point charges $q_i = z_i e_c$ at the positions x_i. The charge positions are located in the strict interior of the molecular region Ω_m. A physically reasonable mathematical assumption is that all charge locations obey the following lower bound on their distance to the solvent region Ω_s for some $\sigma > 0$:

$$|x - x_i| \geq \sigma \quad \forall x \in \Omega_s, \quad i = 1, \ldots, N_m. \tag{2.3}$$

In some models employing the PBE, there is a third region Ω_l (the Stern layer [11]), a layer between Ω_m and Ω_s. In the presence of a Stern layer, the parameter σ in (2.3)

increases in value. Our analysis and results can be easily generalized to this case as well.

Some analytical solutions of the nonlinear PBE are known, but only for unrealistic structure geometries and usually only for linearizations of the equation; cf. [30] for a collection of these solutions and for references to the large amount of literature on analytical solutions to the PBE and similar equations. However, the problem is highly nonlinear. Surface potentials of the linear and the nonlinear PBE differ by over an order of magnitude [45]. Hence, using the nonlinear version of the PBE model is fundamentally important to accurately describe physical effects, and access to reliable and accurate numerical approximation techniques for the nonlinear PBE is critically important in this research area.

We finish this section by making some remarks about an alternative equivalent formulation of the PBE. It is well known (cf. [48, 30]) that the PBE is formally equivalent to a coupling of two equations for the electrostatic potential in different regions Ω_m and Ω_s through the boundary interface. In fact, this equivalence can be rigorously justified; some results of this type will appear in [29]. Inside Ω_m, there are no ions. Thus the equation is simply the Poisson equation

$$-\nabla \cdot (\varepsilon_m \nabla \tilde{u}) = \sum_{i=1}^{N_m} q_i \delta_i \quad \text{in } \Omega_m.$$

In the solvent region Ω_s, there are no atoms. Thus the density is given purely by the Boltzmann distribution

$$-\nabla \cdot (\varepsilon_s \nabla \tilde{u}) + \bar{\kappa}^2 \sinh(\tilde{u}) = 0 \quad \text{in } \Omega_s.$$

These two equations are coupled together through the boundary conditions on the interface $\Gamma := \partial\Omega_m = \partial\Omega_s \cap \Omega_m$:

$$[\tilde{u}]_\Gamma = 0, \quad \text{and} \quad \left[\varepsilon \frac{\partial \tilde{u}}{\partial n_\Gamma}\right]_\Gamma = 0,$$

where $[f]|_\Gamma = \lim_{t\to 0} f(x + t n_\Gamma) - f(x - t n_\Gamma)$, with n_Γ being the unit outward normal direction of interface Γ. We will assume Γ to be sufficiently smooth, say, of class C^2.

Solving the individual subdomain systems and coupling them through the boundary, in the spirit of a nonoverlapping domain decomposition method, is nontrivial due to the complicated boundary conditions and subdomain shapes. Approaches such as mortar-based finite element methods to solve the coupled equations for linear or nonlinear PBE can be found in [19, 52].

3. Regularization of the continuous problem. In this section, we shall introduce a regularized version of the nonlinear PBE for both analysis and discretization purposes. We first transfer the original equation posed on the whole space to a truncated domain using an artificial boundary condition taken from an approximate analytical solution. Then we use the fundamental solution in the whole space to get rid of the singularities caused by δ distributions. We shall mainly focus on more difficult problems in three dimensions. Formulation and results in two dimensions are similar and relatively easy.

Let $\Omega \subset \mathbb{R}^3$ with a convex and Lipschitz-continuous boundary $\partial\Omega$, and $\Omega_m \subset \Omega$. In the numerical simulation, for simplicity, we usually choose Ω to be a ball or cube

containing a molecule region. The solvent region is chosen as $\Omega_s \cap \Omega$ and will be still denoted by Ω_s. On $\partial\Omega$ we choose the boundary condition $\tilde{u} = g$, with

$$g = \left(\frac{e_c^2}{k_B T}\right) \sum_{i=1}^{N_i} \frac{e^{-\kappa|x-x_i|}}{\varepsilon_s |x - x_i|}. \tag{3.1}$$

The boundary condition is usually taken to be induced by a known analytical solution to one of several possible simplifications of the linearized PBE. Far from the molecule, such analytical solutions provide a highly accurate boundary condition approximation for the general nonlinear PBE on a truncation of $\mathbb{R}^3$. For example, (3.1) arises from the use of the Green's function for the Helmholtz operator arising from linearizations of the Poisson–Boltzmann operator, where a single constant global dielectric value of ε_s is used to generate the approximate boundary condition. (This is the case of a rod-like molecule approximation; cf. [30].) Another approach to handling the boundary condition more accurately is to solve the PBE with boundary conditions such as (3.1) on a large Ω (with a coarse mesh) and then solve it in a smaller Ω (with a fine mesh) with the boundary condition provided by the earlier coarse mesh solution. The theoretical justification of this approach can be found at [28] using the two-grid theory [54]. We are not going to discuss more on the choice of the boundary condition in this paper.

Employing (3.1) we obtain the nonlinear PBE on a truncated domain:

$$-\nabla \cdot (\varepsilon \nabla \tilde{u}) + \bar{\kappa}^2 \sinh(\tilde{u}) = \sum_{i=1}^{N_m} q_i \delta_i \qquad \text{in } \Omega, \tag{3.2}$$

$$\tilde{u} = g \qquad \text{on } \partial\Omega. \tag{3.3}$$

This is, in most respects, a standard boundary-value problem for a nonlinear second-order elliptic partial differential equation. However, the right side contains a linear combination of δ distributions, which individually and together are not in $H^{-1}(\Omega)$; thus we cannot apply standard techniques such as classical potential theory. This has at times been the source of some confusion in the molecular modeling community, especially with respect to the design of convergent numerical methods. More precisely, we will see shortly that the solution to the nonlinear Poisson–Boltzmann equation is simply not globally smooth enough to expect standard numerical methods (currently used by most PBE simulators) to produce approximations that converge to the solution to the PBE in the limit of mesh refinement.

In order to gain a better understanding of the properties of solutions to the nonlinear PBE, primarily so that we can design new provably convergent numerical methods, we shall propose a decomposition of the solution to separate out the singularity caused by the δ distributions. This decomposition will turn out to be the key idea that will allow us to design discretization techniques for the nonlinear PBE which have provably good approximation properties and, based on this, also design a new type of adaptive algorithm which is provably convergent for the nonlinear PBE.

We now give this decomposition. It is well known that the function

$$G_i = \frac{q_i}{\varepsilon_m} \frac{1}{|x - x_i|}$$

solves the equation

$$-\nabla \cdot (\varepsilon_m \nabla G_i) = q_i \delta_i \quad \text{in } \mathbb{R}^3.$$

We thus decompose the unknown $\tilde{u}$ as an unknown smooth function u and a known singular function G:

$$\tilde{u} = u + G,$$

with

$$G = \sum_{i=1}^{N_m} G_i. \tag{3.4}$$

Substituting the decomposition into (3.2), we then obtain

$$-\nabla \cdot (\varepsilon \nabla u) + \bar{\kappa}^2 \sinh(u+G) = \nabla \cdot ((\varepsilon - \varepsilon_m)\nabla G) \qquad \text{in } \Omega, \tag{3.5}$$

$$u = g - G \qquad \text{on } \partial\Omega, \tag{3.6}$$

and call it the RPBE. The singularities of the δ distribtuions are transferred to G, which then exhibits degenerate behavior at each $\{x_i\} \subset \Omega_m$. At those points, both $\sinh G(x_i)$ and $\nabla G(x_i)$ exhibit blowup. However, since G is known analytically, one avoids having to build numerical approximations to G. Moreover, both of the coefficients $\bar{\kappa}$ and $\varepsilon - \varepsilon_m$ are zero inside Ω_m, where the blowup behavior arises. Due to this cutoff nature of coefficients, the RPBE is a mathematically well defined nonlinear second-order elliptic equation for the regularized solution u with the source term in H^{-1}. We give a fairly standard argument in the next section to show that it also admits a unique solution $u \in H^1$, even though the original solution $\tilde{u} \notin H^1$ due to the singularities present in G. In the remainder of the paper we shift our focus to establishing additional estimates and developing an approximation theory to guide the design of convergent methods, both nonadaptive and adaptive.

Before moving on, it is useful to note that, away from $\{x_i\}$, the function G is smooth. In particular, we shall make use of the fact that $G \in C^\infty(\Omega_s) \cap C^\infty(\Gamma) \cap C^\infty(\partial\Omega)$ in the later analysis. Also, a key technical tool will be a further decomposition of the regularized solution u into linear and nonlinear parts, $u = u^l + u^n$, where u^l satisfies

$$-\nabla \cdot (\varepsilon \nabla u^l) = \nabla \cdot ((\varepsilon - \varepsilon_m)\nabla G) \qquad \text{in } \Omega, \tag{3.7}$$

$$u^l = 0 \qquad \text{on } \partial\Omega, \tag{3.8}$$

and where u^n satisfies

$$-\nabla \cdot (\varepsilon \nabla u^n) + \bar{\kappa}^2 \sinh(u^n + u^l + G) = 0 \qquad \text{in } \Omega, \tag{3.9}$$

$$u^n = g - G \qquad \text{on } \partial\Omega. \tag{3.10}$$

4. Existence and uniqueness. In this section we shall discuss the existence and uniqueness of the solution of the continuous RPBE. The arguments we use in this section appear essentially in [30], except there the PBE was artificially regularized by replacing the delta distributions with H^{-1}-approximations directly rather than being regularized through a singular function expansion. A different analysis from that appearing below, giving a more precise characterization of the particular function spaces involved and containing various auxiliary results such as the rigorous equivalence of different PBE formulations, will appear in [29].

We first write out the weak formulation. Since $\Delta G = 0$ away from $\{x_i\}$, through integration by parts we get the weak formulation of RPBE: Find

$$u \in M := \{v \in H^1(\Omega) \,|\, e^v, e^{-v} \in L^2(\Omega_s), \text{ and } v = g - G \text{ on } \partial\Omega\}$$

such that

$$A(u,v) + (B(u),v) + \langle f_G, v\rangle = 0 \quad \forall v \in H_0^1(\Omega), \tag{4.1}$$

where

- $A(u,v) = (\varepsilon\nabla u, \nabla u)$,
- $(B(u),v) = (\bar{\kappa}^2 \sinh(u+G), v)$, and
- $\langle f_G, v\rangle = \int_\Omega (\varepsilon - \varepsilon_m)\nabla G \cdot \nabla v$.

Let us define the energy on M:

$$E(w) = \int_\Omega \frac{\varepsilon}{2}|\nabla w|^2 + \bar{\kappa}^2 \cosh(w+G) + \langle f_G, w\rangle.$$

It is easy to characterize the solution of (4.1) as the minimizer of the energy.

LEMMA 4.1. *If u is the solution of the optimization problem, i.e.,*

$$E(u) = \inf_{w\in M} E(w),$$

then u is the solution of (4.1).

Proof. For any $v \in H_0^1(\Omega)$ and any $t \in \mathbb{R}$, the function $F(t) = E(u+tv)$ attains the minimal point at $t = 0$, and thus $F'(0) = 0$, which gives the desired result. ∎

We now recall some standard variational analysis on the existences of the minimizer. In what follows we suppose S is a set in some Banach space V with norm $\|\cdot\|$, and $J(u)$ is a functional defined on S. S is called *weakly sequential compact* if, for any sequence $\{u_k\} \subset S$, there exists a subsequence $\{u_{k_i}\}$ such that $u_{k_i} \rightharpoonup u \in S$, where $\rightharpoonup$ stands for the convergence in the weak topology. For any $u_k \rightharpoonup u$, if $J(u_k) \to J(u)$, we say J is *weakly continuous* at u; if

$$J(u) \le \liminf_{k\to\infty} J(u_k),$$

we say J is *weakly lower semicontinuous* (w.l.s.c.) at u. The following theorem can be proved by the definition easily.

THEOREM 4.2. *If*

1. *S is weakly sequential compact, and*
2. *J is weakly lower semicontinuous on S,*

then there exists $u \in S$ such that

$$J(u) = \inf_{w\in S} J(w).$$

We shall give conditions for the weakly sequential compactness and weakly lower semicontinuity. First we use the fact that a bounded set in a reflexive Banach space is weakly sequential compact.

LEMMA 4.3. *One has the following results:*

1. *The closed unit ball in a reflexive Banach space V is weakly sequential compact.*
2. *If $\lim_{\|v\|\to\infty} J(v) = \infty$, then*

$$\inf_{w\in V} J(w) = \inf_{w\in S} J(w).$$

The next lemma concerns when the functional is w.l.s.c. The proof can be found at [58].

LEMMA 4.4. *If J is a convex functional on a convex set S and J is Gâteaux differentiable, then J is w.l.s.c. on S.*

Now we are in the position to establish the existence and uniqueness of solutions to the RPBE.

THEOREM 4.5. *There exists a unique $u \in M \subset H^1(\Omega)$ such that*

$$E(u) = \inf_{w \in M} E(w).$$

Proof. It is easy to see $E(w)$ is differentiable in M with

$$\langle DE(u), v\rangle = A(u,v) + (B(u), v) + \langle f_G, v\rangle.$$

To prove the existence of the minimizer, we need only to verify that

1. M is a convex set,
2. E is convex on M, and
3. $\lim_{\|v\|_1 \to \infty} E(v) = \infty$.

The verification of (1) is easy and thus skipped here. (2) comes from the convexity of functions x^2 and $\cosh(x)$. Indeed E is *strictly* convex. (3) is a consequence of the inequality

$$E(v) \geq C(\varepsilon, \bar{\kappa})\|v\|_1^2 + C(G, g), \tag{4.2}$$

which can be proved as following. First, by Young's inequality we have for any $\delta > 0$

$$\langle f_G, v\rangle \leq \varepsilon_s \|\nabla G\|_{\Omega_s} \|\nabla v\|_{\Omega_s} \leq \frac{1}{\delta}\|\nabla G\|_{\Omega_s}^2 + \delta \varepsilon_s^2 \|\nabla_v\|_{\Omega_s}^2.$$

Since $\cosh(x) \geq 0$, we have then $E(v) \geq C(\varepsilon, \bar{\kappa})\|\nabla v\|^2 - (1/\delta)\|\nabla G\|_{\Omega_s}^2$, where we can ensure $C(\varepsilon, \bar{\kappa}) > 0$ if δ is chosen to be sufficiently small. Then using norm equivalence on M, we get (4.2). The uniqueness of the minimizer comes from the strict convexity of E. ∎

5. Continuous a priori L^∞-estimates. In this section, we shall derive a priori L^∞-estimates of the solution of the RPBE. The main result of this section is the following theorem.

THEOREM 5.1. *Let u be the weak solution of RPBE in $H^1(\Omega)$. Then u is also in $L^\infty(\Omega)$.*

Note that we cannot apply the analysis of [32, 33] directly to the RPBE, since the right side $f_G \in H^{-1}(\Omega)$ and does not lie in $L^\infty(\Omega)$ as required for use of these results. We shall overcome this difficulty through further decomposition of u into linear and nonlinear parts.

Let $u = u^l + u^n$, where $u^l \in H_0^1(\Omega)$ satisfies the linear elliptic equation (the weak form of (3.7)–(3.8))

$$A(u^l, v) + \langle f_G, v\rangle = 0 \quad \forall v \in H_0^1(\Omega) \tag{5.1}$$

and where $u^n \in M$ satisfies the nonlinear elliptic equation (the weak form of (3.9)–(3.10))

$$A(u^n, v) + (B(u^n + u^l), v) = 0 \quad \forall v \in H_0^1(\Omega). \tag{5.2}$$

Theorem 5.1 then follows from the estimates of u^l and u^n in Lemmas 5.2 and 5.3; cf. (5.3) and (5.4).

LEMMA 5.2. *Let u^l be the weak solution of* (5.1). *Then*

$$u^l \in L^\infty(\Omega). \tag{5.3}$$

Proof. Since $\Delta G = 0$ in Ω_s, using integral by parts we can rewrite the functional f_G as

$$\langle f_G, v\rangle = ((\varepsilon - \varepsilon_m)\nabla G, \nabla v) = \left([\varepsilon]\frac{\partial G}{\partial n_\Gamma}, v\right)_\Gamma,$$

where $[\varepsilon] = \varepsilon_s - \varepsilon_m$ is the jump of ε at the interface. We shall still use f_G to denote the smooth function $[\varepsilon]\frac{\partial G}{\partial n_\Gamma}$ on Γ.

It is easy to see that the linear equation (5.1) is the weak formulation of the elliptic interface problem

$$-\nabla\cdot(\varepsilon\nabla u^l) = 0 \text{ in } \Omega \quad [u^l] = 0,\ \left[\varepsilon\frac{\partial u^l}{\partial n}\right] = f_G \text{ on } \Gamma, \quad \text{and } u = 0 \text{ on } \partial\Omega.$$

Since $f_G \in C^\infty(\Gamma)$ and $\Gamma \in C^2$, by the regularity result of the elliptic interface problem [4, 12, 20, 42], we have $u^l \in H^2(\Omega_m)\cap H^2(\Omega_s)\cap H_0^1(\Omega)$. In particular by the embedding theorem we conclude that $u^l \in L^\infty(\Omega)$. ∎

To derive a similar estimate for the nonlinear part u^n, we define

$$\alpha' = \arg\max_c \left(\bar\kappa^2\sinh(c + \sup_{x\in\Omega_s}(u^l + G)) \le 0\right), \qquad \alpha = \min\left(\alpha', \inf_{\partial\Omega}(g - G)\right),$$
$$\beta' = \arg\min_c \left(\bar\kappa^2\sinh(c + \inf_{x\in\Omega_s}(u^l + G)) \ge 0\right), \qquad \beta = \max\left(\beta', \sup_{\partial\Omega}(g - G)\right).$$

The next lemma gives the a priori L^∞-estimate of u^n.

LEMMA 5.3. *Let u^n be the weak solution of* (5.2). *Then $\alpha \le u^n \le \beta$, and thus*

$$u^n \in L^\infty(\Omega). \tag{5.4}$$

Proof. We use a cutoff-function argument similar to that used in [32]. Since the boundary condition $g - G \in C^\infty(\partial\Omega)$, we can find a $u_D \in H^1(\Omega)$ such that $u_D = g - G$ on $\partial\Omega$ in the trace sense, or more precisely

$$Tu_D = g - G,$$

where $T : \Omega \mapsto \partial\Omega$ is the trace operator. Then the solution can be written $u^n = u_D + u_0$, with $u_0 \in H_0^1(\Omega)$. Let $\overline{\phi} = (u^n - \beta)^+ = \max(u^n - \beta, 0)$ and $\underline{\phi} = (u^n - \alpha)^- = \min(u^n - \alpha, 0)$. Then from

$$0 \le \overline{\phi} = (u^n - \beta)^+ = (u_D + u_0 - \beta)^+ \le (u_D - \beta)^+ + u_0^+,$$
$$0 \ge \underline{\phi} = (u^n - \alpha)^- = (u_D + u_0 - \alpha)^- \ge (u_D - \alpha)^- + u_0^-,$$

and

$$0 \le T\overline{\phi} \le T(u_D - \beta)^+ + Tu_0^+ = 0,$$
$$0 \ge T\underline{\phi} \ge T(u_D - \alpha)^- + Tu_0^- = 0,$$

we conclude that both $\overline{\phi}, \underline{\phi} \in H_0^1(\Omega)$. Thus for either $\phi = \overline{\phi}$ or $\phi = \underline{\phi}$, we have

$$(\varepsilon\nabla u^n, \nabla\phi) + (\bar\kappa^2\sinh(u^n + u^l + G), \phi) = 0.$$

Note that $\overline{\phi} \geq 0$ in Ω and its support is the set $\overline{\mathcal{Y}} = \{x \in \bar{\Omega} \,|\, u^n(x) \geq \beta\}$. On $\overline{\mathcal{Y}}$, we have

$$\bar{\kappa}^2 \sinh(u^n + u^l + G) \geq \bar{\kappa}^2 \sinh\left(\beta' + \inf_{x \in \Omega_s} (u^l + G)\right) \geq 0.$$

Similarly, $\underline{\phi} \leq 0$ in Ω with support set $\underline{\mathcal{Y}} = \{x \in \bar{\Omega} \,|\, u^n(x) \leq \alpha\}$. On $\underline{\mathcal{Y}}$, we now have

$$\bar{\kappa}^2 \sinh(u^n + u^l + G) \leq \bar{\kappa}^2 \sinh\left(\alpha' + \inf_{x \in \Omega_s} (u^l + G)\right) \leq 0.$$

Together this implies

$$0 \geq (\varepsilon \nabla u^n, \nabla \phi) = (\varepsilon \nabla (u^n - \beta), \nabla \phi) = \varepsilon \|\nabla \phi\|^2 \geq 0$$

for either $\phi = \overline{\phi}$ or $\phi = \underline{\phi}$. Using the Poincare inequality we have finally

$$0 \leq \|\phi\| \lesssim \|\nabla \phi\| \leq 0,$$

giving $\phi = 0$, again for either $\phi = \overline{\phi}$ or $\phi = \underline{\phi}$. Thus $\alpha \leq u^n \leq \beta$ in Ω. ∎

6. Finite element methods for the regularized Poisson–Boltzmann equation. In this section we shall discuss the finite element discretization of RPBE using linear finite element spaces V_D^h and prove the existence and uniqueness of the finite element approximation u_h. Furthermore, under some assumptions on the grids we shall derive a priori L^∞-estimates for u_h and use these to prove that u_h is a quasi-optimal approximation of u in the H^1 norm in the sense that

$$\|u - u_h\|_1 \lesssim \inf_{v_h \in V_D^h} \|u - v_h\|_1. \tag{6.1}$$

While the term on the left in (6.1) is in general difficult to analyze, the term on the right represents the fundamental question addressed by classical approximation theory in normed spaces, of which much is known. To bound the term on the right from above, one picks a function in V_D^h which is particularly easy to work with, namely, a nodal or generalized interpolant of u, and then one employs standard techniques in interpolation theory. Therefore, it is clear that the importance of approximation results such as (6.1) are that they completely separate the details of the Poisson–Boltzmann equation from the approximation theory, making available all known results on finite element interpolation of functions in Sobolev spaces (cf. [21]).

Now we assume Ω can be triangulated exactly (e.g., Ω is a cube) with a shape regular and conforming (in the sense of [21]) triangulation $\mathcal{T}_h$. Here $h = h_{\max}$ represents the mesh size which is the maximum diameter of elements in the triangulation. We further assume in the triangulation that the discrete interface Γ_h approximates the known interface Γ to the second order, i.e., $d(\Gamma, \Gamma_h) \leq Ch^2$.

Given such a triangulation $\mathcal{T}_h$ of Ω, we construct the linear finite element space $V^h := \{v \in H^1(\Omega), v|_\tau \in \mathcal{P}_1(\tau)\ \forall \tau \in \mathcal{T}_h\}$. Since the boundary condition $g - G \in C^\infty(\partial\Omega)$, we can find a $u_D \in H^1(\Omega)$ such that $u_D = g - G$ on $\partial\Omega$ in the trace sense. Then the solution can be uniquely written as $u = u_D + u_0$, with $u_0 \in H_0^1$. Thus we will use $H_D^1(\Omega) := H_0^1(\Omega) + u_D$ to denote the affine space with a specified boundary condition and $V_D^h = V^h \cap H_D^1(\Omega)$ to denote the finite element affine space of $H_D^1(\Omega)$. Similarly $V_0^h = V^h \cap H_0^1(\Omega)$. Here to simplify the analysis the boundary condition is assumed to be represented exactly.

Recall that the weak form of RPBE is

$$\text{Find } u \in H_D^1(\Omega) \text{ such that (s.t.) } A(u,v)+(B(u),v)+\langle f_G, v\rangle = 0 \ \ \forall v \in H_0^1(\Omega). \tag{6.2}$$

We are interested in the quality of the finite element approximation:

$$\text{Find } u_h \in V_D^h \quad \text{s.t. } A(u_h, v_h) + (B(u_h), v_h) + \langle f_G, v\rangle = 0 \ \ \forall v_h \in V_0^h. \tag{6.3}$$

It is easy to show that the finite element approximation u_h is the minimizer of E in V_D^h, i.e., $E(u_h) = \inf_{v_h \in V_D^h} E(v_h)$. Then the existence and uniqueness follows from section 3 since V_D^h is convex. As in the continuous setting, it will be convenient to split the discrete solution to the RPBE into linear and nonlinear parts $u_h = u_h^l + u_h^n$, where u_h^l and u_h^n satisfy, respectively,

$$\text{Find } u_h^l \in V_0^h \text{ s.t. } A(u_h^l, v_h) + \langle f_G, v\rangle = 0 \ \ \forall v_h \in V_0^h, \tag{6.4}$$

$$\text{Find } u_h^n \in V_D^h \text{ s.t. } A(u_h^n, v_h) + \langle B(u_h^n + u_h^l), v_h\rangle = 0 \ \ \forall v_h \in V_0^h. \tag{6.5}$$

6.1. Quasi-optimal a priori error estimate. We begin with the following properties of the bilinear form A and and operator B.

LEMMA 6.1. 1. *The bilinear form $A(u,v)$ satisfies the coercivity and continuity conditions. That is, for $u, v \in H^1(\Omega)$*

$$\|u\|_1^2 \lesssim A(u,u), \quad \textit{and} \quad A(u,v) \lesssim \|u\|_1 \|v\|_1.$$

2. *The operator B is monotone in the sense that*

$$(B(u) - B(v), u - v) \geq \bar{\kappa}^2 \|u - v\|^2 \geq 0.$$

3. *The operator B is bounded in the sense that for $u, v \in L^\infty(\Omega), w \in L^2(\Omega)$,*

$$(B(u) - B(v), w) \leq C\|u - v\|\|w\|.$$

Proof. The proof of (1) and (2) is straightforward. We now prove (3). By the mean value theorem, there exists $\theta \in (0,1)$ such that

$$B(u) - B(v) = \bar{\kappa}^2 \cosh(\theta u + (1-\theta)v + G)(u - v).$$

Then by the convexity of cosh and the fact that $u, v \in L^\infty(\Omega), G \in C^\infty(\Omega_s)$, we get

$$\|\cosh(\theta u + (1-\theta)v + G)\|_{\infty,\Omega_s} \leq \|\cosh(u + G)\|_{\infty,\Omega_s} + \|\cosh(v + G)\|_{\infty,\Omega_s} \leq C.$$

The desired result then follows since $B(\cdot)$ is nonzero only in Ω_s. ∎

THEOREM 6.2. *Let u and u_h be the solution of RPBE and its finite element approximation, respectively. When u_h is uniformly bounded, we have*

$$\|u - u_h\|_1 \lesssim \inf_{v_h \in V^h} \|u - v_h\|_1.$$

Proof. By the definition, the error $u - u_h$ satisfies

$$A(u - u_h, w_h) + (B(u) - B(u_h), w_h) = 0 \quad \forall w_h \in V_0^h.$$

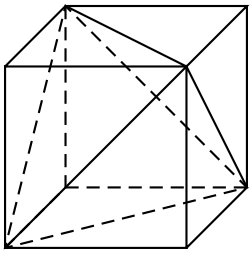


FIG. 6.1. *Divide a cube into* 5 *tetrahedra.*

We then have, for any $v_h \in V_D^h$,

$$\begin{aligned}\|u-u_h\|_1^2 &\lesssim A(u-u_h,u-u_h) = A(u-u_h,u-v_h)+A(u-u_h,v_h-u_h)\\ &\lesssim \|u-u_h\|_1\|u-v_h\|_1 - (B(u)-B(u_h),v_h-u_h).\end{aligned}$$

The second term on the right side is estimated by

$$\begin{aligned}-(B(u)-B(u_h),v_h-u_h) &= -(B(u)-B(u_h),u-u_h)+(B(u)-B(u_h),u-v_h)\\ &\le (B(u)-B(u_h),u-v_h)\\ &\lesssim \|u-u_h\|_1\|u-v_h\|_1.\end{aligned}$$

Here we make use of the monotonicity of B in the second step and the boundness of B in the third step. In summary we obtain for any $v_h \in V_D^h$

$$\|u-u_h\|_1 \lesssim \|u-v_h\|_1,$$

which leads to the desired result by taking the infimum. ∎

6.2. Discrete a priori L^∞-estimates. We now derive L^∞-estimates of the finite element approximation u_h. To this end, we have to put assumptions on the grid. Let (a_{ij}) denote the matrix of the elliptic operator $(\varepsilon\nabla u,\nabla v)$, i.e., $a_{i,j} = A(\varphi_i,\varphi_j)$. Two nodes i and j are adjacent if there is an edge connecting them.

(A1) The off-diagonal term $a_{i,j}$, i,j are adjacent, satisfies

$$a_{i,j} \le -\frac{\rho}{h^2}\sum_{e_{i,j}\subset T}|T|, \quad \text{with } \rho > 0.$$

We now give example grids satisfying (A1). In three dimensions, to simplify the generation of the grid, we choose Ω as a cube and divide into small cubes with length h. For each small cube, we divide it into 5 tetrahedra; see Figure 6.1 for a prototype of the triangulation of one cube. Neighbor cubes are triangulated in the same fashion (with different reflection to make the triangulation conforming). By the formula of the local stiffness matrix in [33, 55], it is easy to verify that the grids will satisfy assumption (A1). We comment that the uniform grid obtained by dividing each cube into 6 tetrahedra will not satisfy the assumption (A1), since in this case if i,j are vertices of diagonal of some cube, then $a_{ij} = 0$.

THEOREM 6.3. *In general dimension $\mathbb{R}^d, d \ge 2$, with assumption* (A1) *and h sufficiently small, the finite element approximation u_h of RPBE satisfies*

$$\|u_h\|_\infty \le C,$$

where C is independent of h.

Proof. We shall use the decomposition $u_h = u_h^n + u_h^l$. By the regularity result [42], we know $u^l \in B^{3/2}_{2,\infty}(\Omega)$ and thus obtain a priori estimate on quasi-uniform grids

$$\|u^l - u_h^l\|_\infty \le C h_{\max}^s \le C \mathrm{diam}(\Omega)^s \quad \text{for some } s \in (0, 3/2).$$

This implies that $\|u_h^l\|_\infty \le \|u^l\|_\infty + \|u^l - u_h^l\|_\infty \le C$ is uniformly bounded with respect to $h_{\max}$. The estimate of u_h^n follows from Theorem 3.3 in [33], where the grid assumption (A1) is used. ∎

In two dimensions, we can relax the assumption on the grid and obtain a similar result. Later we will see that, due to this relaxation, the local refinement in two dimensions is pretty simple.

(A1′) The off-diagonal terms $a_{i,j} \le 0, j \ne i$; i.e., the stiffness matrix corresponding to $A(\cdot,\cdot)$ is an M-matrix.

THEOREM 6.4. *For a two-dimensional triangulation satisfying* (A1′)*, the finite element approximation u_h of RPBE is bounded, i.e.,*

$$\|u_h\|_\infty \le C.$$

Proof. Similarly $\|u_h^l\|_\infty \le C$ is uniformly bounded. In two dimensions the estimate of u_h^n follows from Theorem 3.1 in [33], where the grid assumption (A1) is used. ∎

7. Convergence of adaptive finite element approximation. In this section, we shall follow the framework presented in [50, 51] to derive an a posteriori error estimate. Furthermore we shall present an adaptive method through local refinement based on this error estimator and prove that it will converge. The a priori L^∞-estimates of the continuous and discrete problems derived in the previous sections play an important role here.

7.1. A posteriori error estimate. There are several approaches to adaptive error control, among which the one based on a posteriori error estimation is usually the most effective and most general. Although most existing work on a posteriori estimates has been for linear problems, extensions to the nonlinear case can be made through linearization. For example, consider the nonlinear problem

$$F(u) = 0, \quad F \in C^1(\mathcal{B}_1, B_2^*), \tag{7.1}$$

where the Banach spaces $\mathcal{B}_1$ and $\mathcal{B}_2$ are, e.g., Sobolev spaces and where $\mathcal{B}^*$ denotes the dual space of $\mathcal{B}$. Consider now also a discretization of (7.1)

$$F_h(u_h) = 0, \quad F_h \in C^0(U_h, V_h^*), \tag{7.2}$$

where $U_h \subset \mathcal{B}_1$ and $V_h \subset \mathcal{B}_2$. For the RPBE and a finite element discretization, the function spaces would be taken to be $\mathcal{B}_1 = \mathcal{B}_2 = H_0^1(\Omega)$. The nonlinear residual $F(u_h)$ can be used to estimate the error through the use of a linearization inequality

$$C_1 \|F(u_h)\|_{\mathcal{B}_2^*} \le \|u - u_h\|_{\mathcal{B}_1} \le C_2 \|F(u_h)\|_{\mathcal{B}_2^*}. \tag{7.3}$$

See, for example, [50] for a proof of this linearization result under weak assumptions on F. The estimator is then based on an upper bound on the dual norm of the nonlinear residual on the right in (7.3).

In this section, to show the main idea, we will assume $F_h(u_h) = F(u_h)$ by making the following assumption on the grid.

(A2) The smooth interface Γ is replaced by its discrete approximation Γ_h such that ε and $\bar{\kappa}$ are piecewise constants on each element of the triangulation $\mathcal{T}_h$.

In our setting of the weak formulation, we need to estimate $\|F(u_h)\|_{-1,\Omega}$. To this end, we first introduce quite a bit of notation. We assume that the d-dimensional domain Ω has been exactly triangulated with a set $\mathcal{T}_h$ of shape-regular d-simplices (the finite dimension d is arbitrary, not restricted to $d \leq 3$, throughout this discussion). A family of simplices will be referred to here as shape-regular in the sense of [21].

It will be convenient to introduce the following notation:

$$
\begin{array}{lcl}
\mathcal{T}_h & = & \text{the set of shape-regular simplices triangulating the domain } \Omega. \\
\mathcal{N}(\tau) & = & \text{the union of faces contained in simplex set } \tau \text{ lying on } \partial\Omega. \\
\mathcal{I}(\tau) & = & \text{the union of faces contained in simplex set } \tau \text{ not in } \mathcal{N}(\tau). \\
\mathcal{F}(\tau) & = & \mathcal{N}(\tau) \cup \mathcal{I}(\tau). \\
\mathcal{F} & = & \cup_{\tau \in \mathcal{T}_h} \mathcal{F}(\tau). \\
\omega_\tau & = & \bigcup \{ \tilde{\tau} \in \mathcal{T}_h \mid \tau \bigcap \tilde{\tau} \neq \varnothing, \text{ where } \tau \in \mathcal{T}_h \}. \\
\omega_S & = & \bigcup \{ \tilde{\tau} \in \mathcal{T}_h \mid S \bigcap \tilde{\tau} \neq \varnothing, \text{ where } S \in \mathcal{F} \}. \\
h_\tau & = & \text{the diameter of the simplex } \tau. \\
h_S & = & \text{the diameter of the face } S.
\end{array}
$$

When the argument to one of the face set functions $\mathcal{N}$, $\mathcal{I}$, or $\mathcal{F}$ is in fact the entire set of simplices, we will leave off the explicit dependence on $\mathcal{S}$ without danger of confusion. Finally, we will also need some notation to represent discontinuous jumps in function values across faces interior to the triangulation. For any face $S \in \mathcal{N}$, let n_S denote the unit outward normal; for any face $S \in \mathcal{I}$, take n_S to be an arbitrary (but fixed) choice of one of the two possible face normal orientations. Now, for any $v \in L^2(\Omega)$ such that $v \in C^0(\tau)\ \forall \tau \in \mathcal{T}_h$, define the *jump function*:

$$
[v]_S(x) = \lim_{t \to 0^+} v(x + t n_S) - \lim_{t \to 0^-} v(x - t n_S).
$$

We now define our a posteriori error estimator

$$
\eta_\tau^2(u_h) = h_\tau^2 \|B(u_h)\|_{0,\tau}^2 + \frac{1}{2} \sum_{S \in \mathcal{I}(\tau)} h_S \| \left[n_S \cdot (\varepsilon \nabla u_h + (\varepsilon - \varepsilon_m) \nabla G) \right]_S \|_{0,S}^2, \tag{7.4}
$$

and the oscillation

$$
\operatorname{osc}_\tau^2(u_h) = h_\tau^4 \left(\|\nabla u_h\|_{0,\tau}^2 + \|\nabla G\|_{0,\tau}^2 \right). \tag{7.5}
$$

THEOREM 7.1. *Let $u \in H^1(\Omega)$ be a weak solution of the RPBE and u_h be the finite element approximation with a grid satisfying assumptions* (A1) *and* (A2)*. There exist two constants depending only on the shape regularity of $\mathcal{T}_h$ such that*

$$
\|u - u_h\|_1^2 \leq C_1\, \eta_h^2 + C_2 \operatorname{osc}_h^2, \tag{7.6}
$$

where

$$
\eta_h^2 := \sum_{\tau \in \mathcal{T}_h} \eta_\tau^2(u_h), \quad \textit{and} \quad \operatorname{osc}_h^2 := \sum_{\tau \in \mathcal{T}_h \cap \Omega_s} \operatorname{osc}_\tau^2(u_h).
$$

Proof. We shall apply the general estimate in [51, Chapter 2] (see also [50]) to

$$
\underline{a}(x, u, \nabla u) = \varepsilon \nabla u + (\varepsilon - \varepsilon_m) \nabla G, \quad \text{and} \quad b(x, u, \nabla u) = -\bar{\kappa}^2 \sinh(u + G).
$$

We then use the following facts to get the desired result:

- $\nabla \cdot (\varepsilon \nabla u_h) \mid_\tau = 0 \ \forall \tau \in \mathcal{T}_h$ by the assumption (A2) of the grid;
- $\nabla \cdot ((\varepsilon - \varepsilon_m)\nabla G) \mid_\tau = 0 \ \forall \tau \in \mathcal{T}_h$ since $\Delta G(x) = 0$ if $x \notin \{x_i\}$.
- For $\tau \in \mathcal{T}_h \cap \Omega_s$, let $\bar{u}_h$ and $\bar{G}$ denote the average of u_h and G over τ, respectively. We then have

$$\begin{aligned} \| \sinh(u_h + G) - \sinh(\bar{u}_h + \bar{G})\|_{0,\tau} &\le |\cosh(\xi)| \|u_h - \bar{u}_h + G - \bar{G}\|_{0,\tau} \\ &\le C h_\tau^2 (\|\nabla u_h\|_{0,\tau} + \|\nabla G\|_{0,\tau}). \end{aligned}$$

  Here we use the L^∞-estimates of u and u_h to conclude that $|\cosh(\xi)| \le C$ and the standard error estimate for $\|u_h - \bar{u}_h\|_{0,\tau}$ and $\|G - \bar{G}\|_{0,\tau}$. ∎

We give some remarks on our error estimator and the oscillation term. First, using (4.2) one can easily show that $\|\nabla u_h\|_{0,\Omega} \le C$ uniformly with respect to h and thus $\mathrm{osc}_\tau = O(h_\tau^2)$. Comparing to the order of $\eta_\tau = O(h_\tau)$, the error estimator η_τ will dominate in the upper bound. Second, in (7.4) the jump of $[n_S \cdot (\varepsilon - \varepsilon_m)\nabla G]_S \neq 0$ only if $S \in \Gamma_h$. This additional term with order $O([\varepsilon])$ will emphasize the elements around the interface where the refinement most occurs.

Although it is clear that the upper bound is the key to bounding the error, the lower bound can also be quite useful; it can help to ensure that the adaptive procedure does not do too much work by overrefining an area where it is unnecessary. Again using the general framework for the a posteriori error estimate in [50, 51], we have the following lower bound result.

THEOREM 7.2. *There exists two constants C_3, C_4 depending only on the shape regularity of $\mathcal{T}_h$ such that*

$$\eta_\tau^2(u_h) \le C_3 \|u - u_h\|_{1,\omega_\tau}^2 + C_4 \sum_{\tilde{\tau} \in \omega_\tau \cap \Omega_s} \mathrm{osc}_{\tilde{\tau}}^2(u_h) \quad \forall \tau \in \mathcal{T}_h.$$

7.2. Marking and refinement strategy. Given an initial triangulation $\mathcal{T}_0$, we shall generate a sequence of nested conforming triangulations $\mathcal{T}_k$ using the following loop:

$$\textbf{SOLVE} \ \rightarrow \ \textbf{ESTIMATE} \rightarrow \ \textbf{MARK} \ \rightarrow \ \textbf{REFINE}. \tag{7.7}$$

More precisely to get $\mathcal{T}_{k+1}$ from $\mathcal{T}_k$ we first solve the discrete equation to get u_k on $\mathcal{T}_k$. The error is estimated using u_k and used to mark a set of triangles that are to be refined. Triangles are refined in such a way that the triangulation is still shape-regular and conforming.

We have discussed the step **ESTIMATE** in detail, and we shall not discuss the step **SOLVE**, which deserves a separate investigation. We assume that the solutions of the finite-dimensional problems can be solved to any accuracy efficiently. Examples of such optimal solvers are the multigrid method or the multigrid-based preconditioned conjugate gradient method [53, 13, 27, 56]. In particular we refer to [1, 2] for recent work on adaptive grids in three dimensions and [31, 30] for solving the PBE with inexact Newton methods.

We now present the marking strategy which is crucial for our adaptive methods. We shall focus on one iteration of loop (7.7) and thus use $\mathcal{T}_H$ for the coarse mesh and $\mathcal{T}_h$ for the refined mesh. Quantities related to those meshes will be distinguished by a subscript H or h, respectively.

Let $\theta_i, i = 1, 2$ be two numbers in $(0, 1)$.

1. Mark $\mathcal{M}_{1,H}$ such that

$$\sum_{\tau \in \mathcal{M}_{1,H}} \eta_\tau^2(u_H) \ge \theta_1 \sum_{\tau \in \mathcal{T}_H} \eta_\tau^2(u_H).$$

2. If

$$\operatorname{osc}_H \geq \eta_H \tag{7.8}$$

or

$$C_4 \sum_{\tilde{\tau}\in\cup_{\tau\in M_H}\omega_\tau} \operatorname{osc}_\tau^2(u_H) \geq \frac{1}{2}\sum_{\tau\in\mathcal{M}_H}\eta_\tau^2(u_H), \tag{7.9}$$

then extend $\mathcal{M}_{1,H}$ to $\mathcal{M}_{2,H}$ such that

$$\sum_{\tau\in\mathcal{M}_{2,H}} \operatorname{osc}_\tau^2(u_H) \geq \theta_2 \sum_{\tau\in\mathcal{T}_H} \operatorname{osc}_\tau^2(u_H).$$

Unlike the marking strategy for reducing oscillation in the adaptive finite element methods in [37, 38], in the second step, we put a switch (7.8)–(7.9). In our setting, the oscillation $\operatorname{osc}_H = O(H^2)$ is in general a high-order term. The marking step (2) is seldom applied.

In the **REFINE** step, we need to carefully choose the rule for dividing the marked triangles such that the mesh obtained by this dividing rule is still conforming and shape-regular. Such refinement rules include red and green refinement [7], longest refinement [41, 40], and newest vertex bisection [43, 35, 36]. For the **REFINE** step, we are going to impose the following assumptions.

(A3) Each $\tau \in \mathcal{M}_H$, as well as each of its faces, contains a node of $\mathcal{T}_h$ in its interior.

(A4) Let $\mathcal{T}_h$ be a refinement of $\mathcal{T}_H$ such that the corresponding finite element spaces are nested, i.e., $V^H \subset V^h$.

With those assumptions, we can have the discrete lower bound between two nested grids. Let $\mathcal{T}_H$ be a shape-regular triangulation, and let $\mathcal{T}_h$ be a refinement of $\mathcal{T}_H$ obtained by local refinement of marked elements set $\mathcal{M}_H$. The assumption (A3) is known as the *interior nodes property* in [38]. Such a requirement ensures that the refined finite element space V^h is fine enough to capture the difference of solutions.

THEOREM 7.3. *Let $\mathcal{T}_H$ be a shape-regular triangulation, and let $\mathcal{T}_h$ be a refinement of $\mathcal{T}_H$ obtained by some local refinement methods of marked elements set $\mathcal{M}_H$, such that it satisfies assumptions* (A3) *and* (A4). *Then there exist two constants, depending only on the shape regularity of $\mathcal{T}_H$, such that*

$$\eta_\tau^2(u_H) \leq C_3\|u_h - u_H\|_{1,\omega_\tau}^2 + C_4 \sum_{\tilde{\tau}\in\omega_\tau} \operatorname{osc}_{\tilde{\tau}}^2(u_H) \quad \forall \tau \in \mathcal{M}_H. \tag{7.10}$$

Proof. The proof is standard using the discrete bubble functions on τ and each face $S \in \partial\tau$. ∎

7.3. Convergence analysis. We shall prove that the repeating of loop (7.7) will produce a convergent solution u_k to u. The convergent analysis of the adaptive finite element method is an active topic. In the literature it is mainly restricted to the linear equations [17, 47, 16, 37, 25, 9, 38, 34, 26, 8]. The convergence analysis for the nonlinear equation is relatively rare [24, 49, 15].

LEMMA 7.4. *Let $\mathcal{T}_H$ and $\mathcal{T}_h$ satisfy assumptions* (A1)–(A4). *Then there exist two constants depending only on the shape regularity of $\mathcal{T}_H$ such that*

$$\|u - u_H\|_1^2 \leq C_5\|u_h - u_H\|_1^2 + C_6 \operatorname{osc}_H^2.$$

When (7.8) *and* (7.9) *do not hold, we have a stronger inequality*

$$\|u - u_H\|_1^2 \le C_7 \|u_h - u_H\|_1^2,$$

where C_7 *depends only on the shape regularity of* $\mathcal{T}_H$.

Proof. By the upper bound and marking strategy

$$\begin{aligned}\|u - u_H\|_1^2 &\le C_1\eta_H^2 + C_2\mathrm{osc}_H^2 \\ &\le C_1\theta_1^{-1}\sum_{\tau\in\mathcal{M}_{1,H}}\eta_\tau^2(u_H) + C_2\,\mathrm{osc}_H^2 \\ &\le C_5\|u_h - u_H\|_1^2 + C_6\,\mathrm{osc}_H^2,\end{aligned}$$

with

$$C_5 = C_1\theta_1^{-1}C_3^{-1}, \quad \text{and} \quad C_6 = (C_2 + 2C_3^{-1}C_4).$$

If (7.8) does not hold, i.e., $\mathrm{osc}_H \le \eta_H$, the first inequality becomes

$$\|u - u_H\|_1^2 \le (C_1 + C_2)\eta_H^2.$$

If (7.9) does not hold, we can easily modify the lower bound (7.10) as

$$\sum_{\tau\in\mathcal{M}_{1,H}}\eta_\tau^2(u_H) \le 2C_3\|u_h - u_H\|_1^2.$$

Then the inequality follows similarly. ∎

For $\tau_h \subset \tau_H$, let $h_{\tau_h} = \gamma H_{\tau_H}$, with $\gamma \in (0,1)$. The next lemma shows that even the oscillation is not small; there is also a reduction result. For the marked set $\mathcal{M}_H \subset \mathcal{T}_H$, we shall use $\overline{\mathcal{M}_H}$ to denoted the refined elements in $\mathcal{T}_h$.

LEMMA 7.5. *If* $\mathcal{M}_{2,H}\backslash\mathcal{M}_{1,H} \notin \varnothing$, *there exist* ρ_1, ρ_2 *such that*

$$\mathrm{osc}_h^2 \le \rho_1\,\mathrm{osc}_H^2 + \rho_2\|u_h - u_H\|_1^2.$$

Proof.

$$\begin{aligned}\mathrm{osc}_h^2 &\le \sum_{\tau\in\mathcal{T}_h}\mathrm{osc}_\tau^2(u_H) + C\sum_{\tau\in\mathcal{T}_h}(h_\tau^4\|\nabla(u_h - u_H)\|_\tau^2) \\ &\le \sum_{\tau_h\in\overline{\mathcal{M}}_{2,H}}\mathrm{osc}_\tau^2(u_H) + \sum_{\tau_h\in\mathcal{T}_h\backslash\overline{\mathcal{M}}_{2,H}}\mathrm{osc}_\tau^2(u_H) + Ch^2\|\nabla(u_h - u_H)\|^2 \\ &\le \gamma^2\sum_{\tau_H\in\mathcal{M}_{2,H}}\mathrm{osc}_\tau^2(u_H) + \sum_{\tau_H\in\mathcal{T}_H\backslash\mathcal{M}_{2,H}}\mathrm{osc}_\tau^2(u_H) + Ch^2\|\nabla(u_h - u_H)\|^2 \\ &\le \mathrm{osc}_H^2 + (\gamma^2 - 1)\sum_{\tau_H\in\mathcal{M}_{2,H}}\mathrm{osc}_\tau^2(u_H) + Ch^2\|\nabla(u_h - u_H)\|^2 \\ &\le \rho_1\,\mathrm{osc}_H^2 + \rho_2\|u_h - u_H\|_1^2,\end{aligned}$$

with $\rho_1 = 1 - (1-\gamma^2)/\theta_2 \in (0,1)$, and $\rho_2 = Ch^2$. ∎

We shall choose θ_2 sufficiently close to 1 and $h_{\max} < 1/c$ to ensure $\rho_i \in (0,1), i = 1,2$.

For the nonlinear problem, we do not have the orthogonality in H^1 norms. But we shall use the trivial identity

$$E(u_H) - E(u) = E(u_H) - E(u_h) + E(u_h) - E(u). \tag{7.11}$$

The following lemma proves the equivalence of energy error and error in H^1 norm. Again the L^∞ norm estimate of u and u_h is crucial.

LEMMA 7.6. *If both $\mathcal{T}_h$ and $\mathcal{T}_H$ satisfy the assumption* (A1)*, then*

- $E(u_h) - E(u) \simeq \|u_h - u\|_1^2$;
- $E(u_H) - E(u) \simeq \|u_H - u\|_1^2$;
- $E(u_H) - E(u_h) \simeq \|u_H - u_h\|_1^2$.

Proof. By the Taylor expansion

$$E(u_H) - E(u_h) = \langle DE(u_h), u_H - u_h\rangle + (D^2E(\xi)(u_H - u_h), u_H - u_h).$$

The first term is zero since u_h is the minimizer. The desired result follows from the bound

$$\bar{\kappa}^2 \le \|D^2E(\xi)\|_\infty = \bar{\kappa}^2 \|\cosh(\xi + G)\|_{\infty,\Omega_s} \le C.$$

Other inequalities follow from the same line. ∎

Our adaptive finite element methods (AFEMs) consist of the iteration of loop (7.7) with the estimate, marking, and refinement parts discussed before. Also the grids generated by the algorithm will satisfy assumptions (A1)–(A4). Hereafter we replace the subscript h by an iteration counter called k and introduce some notation to simplify the proof. Let u_k be the solution in the kth iteration, $\delta_k := E(u_k) - E(u)$, $d_k = E(u_k) - E(u_{k+1})$, and $o_k = \mathrm{osc}^2(u_k)$

THEOREM 7.7. *The adaptive method using loop* (7.7) *will produce a convergent approximation in the sense that*

$$\lim_{k\to 0} \|u - u_k\|_1 = 0.$$

Proof. By Lemma 7.6, we need only to show $\delta^k \to 0$ as $k \to 0$. We first discuss the easier case: When osc_H is the high-order term in the sense that the inequalities (7.8) and (7.9) do not hold, we have the error reduction

$$\|u - u_H\|_1^2 \le C\|u_h - u_H\|^2.$$

Using Lemma 7.5 and (7.11), we have

$$E(u_H) - E(u) \le C(E(u_H) - E(u_h)),$$

which is equivalent to $\delta_H \le C\delta_H - C\delta_h$. Then $\delta_h \le (1 - 1/C)\delta_H$, and thus

$$\delta^k \le \alpha^k \delta^0, \quad \text{with } \alpha = (1 - 1/C) \in (0, 1).$$

When the oscillation is not small, i.e., (7.8) or (7.9) holds, we can get only

$$\Lambda_1\delta_k \le d_k + \Lambda_2 o_k, \quad \text{with } \Lambda_1 \in (0, 1). \tag{7.12}$$

We shall use techniques from [34] to prove the convergence. Recall that we have

$$\delta_{k+1} = \delta_k - d_k. \tag{7.13}$$

For any $\beta \in (0, 1)$, $\beta \times (7.12) + (7.13)$ gives

$$\delta_{k+1} \le \alpha\delta_k + \beta\Lambda_2 o_k - (1 - \beta)d_k, \quad \text{with } \alpha = (1 - \beta\Lambda_1) \in (0, 1). \tag{7.14}$$

Recall that we have

$$o_{k+1} \le \rho_1 o_k + \rho_2 d_k. \tag{7.15}$$

Let $\gamma = (1-\beta)/\rho_2$; (7.15) $\times\, \gamma$ + (7.14) gives

$$\delta_{k+1} + \gamma o_{k+1} \le \alpha \delta_k + (\beta \Lambda_2 + \rho_1 \gamma) o_k.$$

Let $1 > \mu > \rho_1$. We choose

$$\beta = \frac{\frac{\mu-\rho_1}{\rho_2}}{\Lambda_2 + \frac{\mu-\rho_1}{\rho_2}} \in (0,1)$$

to get

$$\delta_{k+1} + \gamma o_{k+1} \le \max(\alpha, \mu)(\delta_k + \gamma o_k),$$

which also implies the convergence of our AFEM. ∎

8. Summary and concluding remarks. In this article we have established a number of basic theoretical results for the nonlinear Poisson–Boltzmann equation and for its approximation using finite element methods. We began by showing that the problem is well-posed through the use of an auxiliary or *regularized* version of the equation and then established a number of basic estimates for the solution to the regularized problem. The Poisson–Boltzmann equation does not appear to have been previously studied in detail theoretically, and it is hoped that this paper, together with a more complete analysis of the continuous PDE solution theory to appear in [29], will help provide molecular modelers with a better theoretical foundation for their analytical and computational work with the Poisson–Boltzmann equation. The bulk of this article then focused on designing a numerical discretization procedure based on the regularized problem and on establishing rigorously that the discretization procedure converged to the solution to the original (nonregularized) nonlinear Poisson–Boltzmann equation. Based on these results, we also designed an adaptive finite element approximation procedure and then gave a fairly involved technical argument showing that this adaptive procedure also converges in the limit of mesh refinement. This article apparently gives the first convergence result for a numerical discretization technique for the nonlinear Poisson–Boltzmann equation with delta distribution sources, and it also introduces the first provably convergent adaptive method for the equation. This last result is one of only a handful of convergence results of this type for nonlinear elliptic equations (the others being [24, 49, 15]).

Several of the theoretical results in the paper rest on some basic assumptions on the underlying simplex mesh partitioning of the domain, namely, assumptions (A1)–(A4); we now make a few comments on these assumptions. To begin, we required a refinement procedure that would preserve the L^∞ norm estimate of u_h. Meeting this requirement in the two-dimensional setting is relatively easy; one can choose Ω as a square and start with a uniform mesh of a square. For the refinement methods, one can use longest edge or newest vertex bisection. Subdivisions obtained by these two methods contain only one type of triangle: isosceles right triangles. Thus the assumption (A1′) always holds. In the three-dimensional setting, this is more tricky. Bisection will introduce some obtuse angles in the refined elements. One needs to use a three-dimensional analogue of red-green refinement [10]. However, this will not produce nested subspaces; i.e., assumption (A4) is invalid. For convergence analysis

based on red-green refinement, we could use the technique in [46] to relax the assumption (A4). Since this will only add technical difficulties but does not exhibit principally new phenomena, we omit them here. Another approach to relax the assumption (A1) is to use pointwise a posteriori error estimates developed in [39] for monotone semilinear equations. We can start with a quasi-uniform triangulation and refine the triangulation according to the pointwise a posteriori error estimator to make sure $\|u - u_h\|_\infty \leq C$. Then together with the L^∞ norm estimate of u, by the triangulation inequality $\|u_h\|_\infty \leq \|u\|_\infty + \|u - u_h\|_\infty \leq C$, we have the control of $\|u_h\|_\infty$. Note that the pointwise a posteriori error estimates developed in [39] are for elliptic-type equations with continuous coefficients. To use this approach we need to adapt the estimate for the jump coefficients case which will be a further research topic.

Assumption (A2) is needed to approximate the interface well in an a priori manner. Of course, one can include this approximation effect into the a posteriori error estimate (namely, the term $\|F(u_h) - F_h(u_h)\|$) and use this to drive local refinement to improve the approximation to the desired level for the assumption or use the strategy for the oscillation to include it in the refinement loop. However, we note that, since the interface is known a priori from, e.g., x-ray crystallography information, we do not need to solve the equation (which is generally the more expensive route) to solve this problem; we view this as primarily a mesh generation problem. Robust algorithms to produce well-shaped tetrahedral meshes which are constrained to exactly match some interior embedded two-manifold are available in the literature; for example, see [18, 3]. A simple algorithm can be based entirely on local refinement with the marking and refinement strategy, but without having to solve the PBE to produce error indicators: If the element cross the interface, then it gets refined. This strategy was employed in [5].

After this work was done, we learned that the assumption (A3) is not needed for the convergence of adaptive finite element methods for a linear elliptic equation. As an ongoing project, we are extending it to the nonlinear Poisson–Boltzmann equation.

Finally, we make some remarks on the practical realization of a convergent discretization procedure based on the two-way (or three-way) expansion into a known singular function and solution(s) of an associated regularized version of the problem. Methods for building high-quality approximate solutions of the regularized nonlinear PBE, either by solving (3.5)–(3.6) at once or by solving for the linear and nonlinear pieces separately by solving (5.1)–(5.2) and then adding the solutions together, are well-understood. The techniques described in [28], taken together with the approximation framework and the adaptive algorithm proposed in the present article, moves us a step closer to the goal of a complete optimal solution to this problem, in terms of approximation quality for a given number of degrees of freedom, computational complexity of solving the corresponding discrete equations, and the storage requirements of the resulting algorithms. What remains is simply the cost of evaluating the singular function G in forming the source terms in (3.5) or (5.1). The source terms are evaluated using numerical quadrature schemes: sampling the integrand at specially chosen discrete points in each element and then summing the results up using an appropriate weighting. This is equivalent to computing all pairwise interactions between the collection of quadrature points (a fixed constant number of points per simplex) and the number of charges forming G. Given that G is typically formed from at most a few thousand charges, the algorithm evaluating G at the quadrature points should scale linearly with the number of quadrature points, which is a (small) constant multiple of the number of simplices. This can be accomplished using techniques such distance-classing and fast multiple-type methods.

REFERENCES

[1] B. Aksoylu, S. Bond, and M. Holst, *An odyssey into local refinement and multilevel preconditioning* III: *Implementation and numerical experiments*, SIAM J. Sci. Comput., 25 (2003), pp. 478–498.

[2] B. Aksoylu and M. Holst, *Optimality of multilevel preconditioners for local mesh refinement in three dimensions*, SIAM J. Numer. Anal., 44 (2006), pp. 1005–1025.

[3] P. Alliez, D. Cohen-Steiner, M. Yvinec, and M. Desbrun, *Variational tetrahedral meshing*, ACM Trans. Graphics, 24 (2005), pp. 617–625.

[4] I. Babuška, *The finite element method for elliptic equations with discontinuous coefficients*, Computing, 5 (1970), pp. 207–213.

[5] N. Baker, M. Holst, and F. Wang, *Adaptive multilevel finite element solution of the Poisson–Boltzmann equations* II: *Refinement at solvent-accessible surfaces in biomolecular systems*, J. Comput. Chem., 21 (2000), pp. 1343–1352.

[6] N. Baker, D. Sept, M. Holst, and J. A. McCammon, *The adaptive multilevel finite element solution of the Poisson–Boltzmann equations on massively parallel computers*, IBM Journal of Research and Development, 45 (2001), pp. 427–438.

[7] R. E. Bank, A. H. Sherman, and A. Weiser, *Refinement algorithms and data structures for regular local mesh refinement*, in Scientific Computing, IMACS/North-Holland, Amsterdam, 1983, pp. 3–17.

[8] E. Bänsch, P. Morin, and R. H. Nochetto, *An adaptive Uzawa FEM for the Stokes problem: Convergence without the inf-sup condition*, SIAM J. Numer. Anal., 40 (2002), pp. 1207–1229.

[9] P. Binev, W. Dahmen, and R. DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.

[10] F. Bornemann, B. Erdmann, and R. Kornhuber, *Adaptive multilevel methods in three space dimensions*, Internat. J. Numer. Methods Engrg., 36 (1993), pp. 3187–3203.

[11] I. Borukhov, D. Andelman, and H. Orland, *Steric effects in electrolytes: A modified Poisson–Boltzmann equation*, Phys. Rev. Lett., 79 (1997), pp. 435–438.

[12] J. Bramble and J. King, *A finite element method for interface problems in domains with smooth boundaries and interfaces*, Adv. Comput. Math., 6 (1996), pp. 109–138.

[13] J. H. Bramble and X. Zhang, *The analysis of multigrid methods*, in Handbook of Numerical Analysis VII, North-Holland, Amsterdam, 2000, pp. 173–415.

[14] J. M. Briggs and J. A. McCammon, *Computation unravels mysteries of molecular biophysics*, Comput. Phys., 6 (1990), pp. 238–243.

[15] C. Carstensen, *Convergence of adaptive FEM for a class of degenerate convex minimization problem*, IMA J. Numer. Anal., to appear.

[16] C. Carstensen and R. H. W. Hoppe, *Convergence analysis of an adaptive nonconforming finite element methods*, Numer. Math., 103 (2006), pp. 251–266.

[17] L. Chen, M. Holst, and J. Xu, *Convergence and Optimality of Adaptive Mixed Finite Element Methods*, manuscript, 2006.

[18] L. Chen and M. J. Holst, *Mesh Adaptation Based on Optimal Delaunay Triangulations*, preprint, 2006.

[19] W. Chen, Y. Shen, and Q. Xia, *A mortar finite element approximation for the linear Poisson–Boltzmann equation*, Appl. Math. Comput., 164 (2005), pp. 11–23.

[20] Z. Chen and J. Zou, *Finite element methods and their convergence for elliptic and parabolic interface problems*, Numer. Math., 79 (1998), pp. 175–202.

[21] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, in Stud. Math. Appl. 4, North-Holland, Amsterdam, 1978.

[22] P. Debye and E. Hückel, Physik. Z., 185 (1923).

[23] P. Debye and E. Hückel, *Zur theorie der elektrolyte.* I. *gefrierpunktserniedrigung und verwandte erscheinungen*, Physikalische Zeitschrift, 24 (1923), pp. 185–206.

[24] W. Dörfler, *A robust adaptive strategy for the non-linear Poisson's equation*, Computing, 55 (1995), pp. 289–304.

[25] W. Dörfler, *A convergent adaptive algorithm for Poisson's equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.

[26] W. Dörfler and O. Wilderotter, *An adaptive finite element method for a linear elliptic equation with variable coefficients*, ZAMM Z. Angew. Math. Mech., 80 (2000), pp. 481–491.

[27] W. Hackbusch, *Multigrid Methods and Applications*, in Computational Mathematics 4, Springer–Verlag, Berlin, 1985.

[28] M. Holst, N. Baker, and F. Wang, *Adaptive multilevel finite element solution of the Poisson–Boltzmann equations* I: *Algorithms and examples*, J. Comput. Chem., 21 (2000), pp. 1319–1342.

[29] M. Holst and B. Li, *Boundary-Value Problems of the Poisson–Boltzmann Equation*, preprint, 2006.
[30] M. J. Holst, *The Poisson–Boltzmann Equation: Analysis and Multilevel Numerical Solution*, Ph.D. thesis, University of Illinois at Urbana-Champaign, 1994.
[31] M. J. Holst and F. Said, *Numerical solution of the nonlinear Poisson–Boltzmann equation: Developing more robust and efficient methods*, J. Comput. Chem., 16 (1995), pp. 337–364.
[32] J. W. Jerome, *Consistency of semiconductor modeling: An existence/stability analysis for the stationary van Roosbroeck system*, SIAM J. Appl. Math., 45 (1985), pp. 565–590.
[33] T. Kerkhoven and J. W. Jerome, l_∞ *stability of finite element approximations of elliptic gradient equations*, Numer. Math., 57 (1990), pp. 561–575.
[34] K. Mekchay and R. H. Nochetto, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827.
[35] W. F. Mitchell, *Unified Multilevel Adaptive Finite Element Methods for Elliptic Problems*, Ph.D. thesis, University of Illinois at Urbana-Champsign, 1988.
[36] W. F. Mitchell, *A comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math. Software (TOMS) archive, 15 (1989), pp. 326–347.
[37] P. Morin, R. H. Nochetto, and K. G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
[38] P. Morin, R. H. Nochetto, and K. G. Siebert, *Convergence of adaptive finite element methods*, SIAM Rev., 44 (2002), pp. 631–658.
[39] R. H. Nochetto, A. Schmidt, K. G. Siebert, and A. Veeser, *Pointwise a posteriori error estimates for monotone semi-linear equations*, Numer. Math., 104 (2006), pp. 515–538.
[40] M. C. Rivara, *Design and data structure for fully adaptive, multigrid finite element software*, ACM Trans. Math. Software, 10 (1984), pp. 242–264.
[41] M.-C. Rivara, *Mesh refinement processes based on the generalized bisection of simplices*, SIAM J. Numer. Anal., 21 (1984), pp. 604–613.
[42] G. Savare, *Regularity results for elliptic equations in Lipschitz domains*, J. Funct. Anal., 152 (1998), pp. 176–201.
[43] E. G. Sewell, *Automatic Generation of Triangulations for Piecewise Polynomial Approximation*, Ph.D. dissertation, Purdue University, West Lafayette, IN, 1972.
[44] K. Sharp and B. Honig, *Electrostatic interactions in macromolecules: Theory and applications*, Annu. Kev. Biophys. Biophys. Chem., 19 (1990), pp. 301–332.
[45] A. I. Shestakov, J. L. Milovich, and A. Noy, *Solution of the nonlinear Poisson–Boltzmann equation using pseudo-transient continuation and the finite element method*, Journal of Colloid and Interface Science, 247 (2002), pp. 62–79.
[46] R. Stevenson, *An optimal adaptive finite element method*, SIAM J. Numer. Anal., 42 (2005), pp. 2188–2217.
[47] R. Stevenson, *Optimality of a standard adaptive finite element method*, Found. Comput. Math., 7 (2007), pp. 245–269
[48] C. Tanford, *Physical Chemistry of Macromolecules*, John Wiley & Sons, New York, 1961.
[49] A. Veeser, *Convergent adaptive finite elements for the nonlinear Laplacian*, Numer. Math., 92 (2002), pp. 743–770.
[50] R. Verfurth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp., 62 (1994), pp. 445–475.
[51] R. Verfürth, *A Review of a Posteriori Error Estimation and Adaptive Mesh Refinement Tecniques*, B. G. Teubner, Leipzig, 1996.
[52] D. Xie and S. Zhou, *A new minimization protocol for solving nonlinear Poisson–Boltzmann mortar finite element equation*, BIT Numerical Mathematics, to appear.
[53] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[54] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs*, SIAM J. Numer. Anal., 33 (1996), pp. 1759–1777.
[55] J. Xu and L. Zikatanov, *A monotone finite element scheme for convection diffusion equations*, Math. Comp., 68 (1999), pp. 1429–1446.
[56] J. Xu and L. Zikatanov, *The method of alternating projections and the method of subspace corrections in Hilbert space*, J. Amer. Math. Soc., 15 (2002), pp. 573–597.
[57] Y. Yang, *Some Studies on Finite Element Computing for the Poisson–Boltzmann Equation*, Ph.D. thesis, Institute of Computational Mathematics, Academy of Mathematics and System Sciences, Chinese Academy of Sciences, 2005.
[58] K. Yosida, *Functional Analysis*, Grundlehren Math. Wiss. 123, Springer–Verlag, Berlin, 1980.
[59] Z. Zhou, P. Payne, M. Vasquez, N. Kuhn, and M. Levitt, *Finite-difference solution of the Poisson–Boltzmann equation: Complete elimination of self-energy*, J. Comput. Chem., 11 (1996), pp. 1344–1351.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2321–2344

OVERLAPPING ADDITIVE SCHWARZ PRECONDITIONERS FOR ELLIPTIC PROBLEMS WITH DEGENERATE LOCALLY ANISOTROPIC COEFFICIENTS*

SVEN BEUCHLER† AND SERGEY V. NEPOMNYASCHIKH‡

Abstract. In this paper, we consider degenerate and locally anisotropic boundary value problems on the unit square. These problems are discretized by piecewise linear finite elements on a triangular mesh of isosceles right triangles. The system of linear algebraic equations is solved by a preconditioned gradient method using a domain decomposition preconditioner with overlap. We prove that the condition number of the preconditioned system is bounded by a constant independent of the discretization parameter. Moreover, the preconditioning operation requires $\mathcal{O}(N)$ operations, where N is the number of unknowns. Several numerical experiments show the performance of the proposed method.

Key words. solution of discretized equations, finite elements, domain decomposition

AMS subject classifications. 65N22, 65N30, 65N55, 65F50

DOI. 10.1137/060675678

1. Introduction.

In this paper, we investigate the degenerate and locally anisotropic boundary value problem

$$\begin{aligned} -\omega_1^2(y)u_{xx} - \omega_2^2(x)u_{yy} &= f \quad \text{in } \Omega = (0,1)^2, \\ u &= 0 \quad \text{on } \partial\Omega \end{aligned} \tag{1.1}$$

with some strongly monotonic increasing and bounded weight functions $\omega_i : [0,1] \mapsto \mathbb{R}$ satisfying $\omega_1(0)\omega_2(0) = 0$.

In the past, degenerate problems have been considered relatively rarely. One reason is the unphysical behavior of the partial differential equation (PDE), which is quite unusual in technical applications. One work focusing on this type of PDE is the book of Kufner and Sändig [13]. Nowadays, problems of this type are becoming more and more popular because there are stochastic PDEs of a similar structure. An example of an isotropic degenerate stochastic PDE is the elliptic part of the Black–Scholes PDE [18].

An example of a locally anisotropic degenerate elliptic problem is the solver related to the problem of the subdomains for the p-version of the finite element method (FEM) using quadrilateral elements. This solver can be interpreted as the h-version FEM-discretization matrix of problem (1.1) with $\omega_1(\xi) = \omega_2(\xi) = \xi$. We refer the reader to [1], [12] for more details.

The discretization of (1.1) using the h-version of the FEM leads to a linear system of algebraic equations

$$\mathcal{K}\underline{u} = \underline{f}. \tag{1.2}$$

*Received by the editors November 22, 2006; accepted for publication (in revised form) March 23, 2007; published electronically November 9, 2007.
http://www.siam.org/journals/sinum/45-6/67567.html
†Institute of Computational Mathematics, University of Linz, Altenbergerstraße 69, A-4040 Linz, Austria (sven.beuchler@jku.at).
‡Institute for Computational Mathematics and Computational Geophysics, SD Russian Academy of Sciences, Novosibirsk 630090, Russia (svnep@oapmg.sscc.ru).

It is well known from the literature that preconditioned conjugate gradient (PCG) methods with domain decomposition (DD) preconditioners are among the most efficient iterative solvers for systems of type (1.2); see, e.g., [7], [14], [10], [20], [15]. In this paper, we will propose and analyze overlapping DD preconditioners.

The type of overlapping DD preconditioners presented in this paper was originally developed in [17] for problems with jumping coefficients; see also the recent research for highly jumping coefficients in [19], [9]. In our recent paper [4], we analyzed these overlapping DD preconditioners for isotropic degenerate problems. In most cases, the optimality of this method has been shown. Here, we adapt these techniques to problem (1.1). For tensor product discretizations, we will prove the optimality of the method. Moreover, this method can easily be extended to more general h-version FEM discretizations, too.

Only a limited number of papers have investigated fast solvers for degenerate elliptic problems. The paper [6] deals with the Laplacian in two dimensions in polar coordinates. In the paper [8], multigrid methods for some other types of degenerate problems are proposed. Multigrid solvers for finite element discretizations of (1.1) have been investigated in [1]; see also [2]. However, the convergence of the V-cycle was not yet proved. The paper [12] develops nonoverlapping DD preconditioners for $\omega_1(\xi) = \omega_2(\xi) = \xi$. The paper [5] proposes wavelet methods for several classes of degenerate elliptic problems on the unit square. One of them is problem (1.1) under the restriction $\lim_{\xi\to 0+} \frac{\xi^3}{\omega_i^2(\xi)} = 0$, $i = 1, 2$, on the weight functions. Moreover, a fast direct solver based on eigenvalue computations combined with the fast Fourier transform and solving tridiagonal systems can be designed if at least one of the weight functions ω_i, $i = 1, 2$, is assumed to be constant on $(0, 1)$ and if a tensor product discretization is used.

The remaining part of this paper is organized as follows. In section 2, we introduce the reader to our problem and to our notation. The preconditioners are defined in section 3; moreover, the main theorems with the condition number estimates are stated. In section 4, we formulate some auxiliary results from the additive Schwarz method (ASM), which are required for the proofs of our main theorems given in section 5. In section 6, we present some numerical experiments which show the performance of the presented methods.

Throughout this paper, the integer k denotes the level number. For two real symmetric and positive definite $n \times n$ matrices A, B, the relation $A \preceq B$ means that $A - cB$ is negative definite, where $c > 0$ is a constant independent of n. The relation $A \sim B$ means $A \preceq B$ and $B \preceq A$, i.e., the matrices A and B are spectrally equivalent. The parameter c denotes a generic constant. The isomorphism between a function $u = \sum_i u_i \psi_i \in L^2$ and the corresponding vector of coefficients $\underline{u} = [u_i]_i$ in the basis $[\Psi] = [\psi_1, \psi_2, \ldots]$ is denoted by $u = [\Psi]\underline{u}$. The Greek letters μ and ν stand for pairs of integers (μ_1, μ_2) and (ν_1, ν_2), respectively. The closure of an open set M is denoted by $\overline{M}$.

2. Setting of the problem. In this paper, we investigate the following boundary value problem. Let $\Omega = (0, 1)^2$.

Find $u \in \mathbb{H}_{\mathcal{D},0} := \{u \in L_2(\Omega) : \int_\Omega (\nabla u)^T \mathcal{D} \nabla u < \infty, u\,|_{\partial\Omega} = 0\}$ such that

$$a(u, v) := \int_\Omega (\nabla v)^T D \nabla u = (f, v) \quad \forall v \in \mathbb{H}_{\mathcal{D},0} \tag{2.1}$$

with the coefficient matrix $\mathcal{D}(x, y) = \left[\begin{smallmatrix} \omega_1^2(y) & 0 \\ 0 & \omega_2^2(x) \end{smallmatrix}\right]$ and weight functions ω_i, $i = 1, 2$,

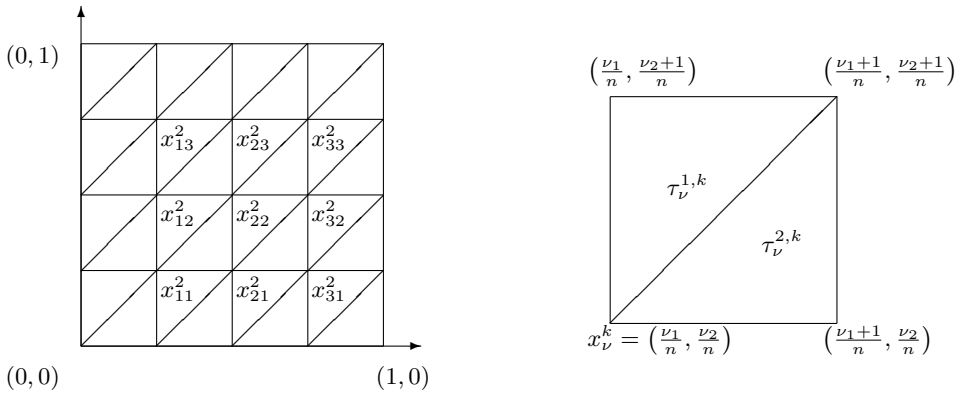


FIG. 2.1. *Mesh for the finite element method for $k = 2$ (left); notation of the triangles (right).*

which satisfy the following assumption.

ASSUMPTION 2.1. *The functions $\omega_i : [0,1] \mapsto \mathbb{R}$, $i = 1, 2$,*

- *are monotonically increasing;*
- *are continuous; and*
- *satisfy the estimate*

$$\omega_i(2\xi) \le c_\omega \omega_i(\xi) \quad \forall \xi \in \left(0, \frac{1}{2}\right] \tag{2.2}$$

with some constants $c_\omega > 0$.

We discretize problem (2.1) by piecewise linear finite elements on the regular Cartesian grid consisting of congruent, isosceles, and right triangles. For this purpose, some notation is introduced. Let k be the level of approximation and $n = 2^k$. Let $x^k_\nu = \left(\frac{\nu_1}{n}, \frac{\nu_2}{n}\right)$, where $\nu = (\nu_1, \nu_2) \in \{0, 1, \ldots, n\}^2$. The domain Ω is divided into congruent, isosceles, and right triangles $\tau^{s,k}_\nu$, where $\nu \in \{0, 1, \ldots, n-1\}^2$ and $s = 1, 2$; see Figure 2.1. The triangle $\tau^{1,k}_\nu$ has the three vertices $\left(\frac{\nu_1}{n}, \frac{\nu_2}{n}\right), \left(\frac{\nu_1+1}{n}, \frac{\nu_2+1}{n}\right)$, and $\left(\frac{\nu_1}{n}, \frac{\nu_2+1}{n}\right)$, and $\tau^{2,k}_\nu$ has the three vertices $\left(\frac{\nu_1}{n}, \frac{\nu_2}{n}\right), \left(\frac{\nu_1+1}{n}, \frac{\nu_2+1}{n}\right)$, and $\left(\frac{\nu_1+1}{n}, \frac{\nu_2}{n}\right)$; see Figure 2.1. Piecewise linear finite elements are used on the mesh $T_k = \{\tau^{s,k}_\nu\}_{\nu \in \{0,1,\ldots,n-1\}^2, s \in \{1,2\}}$. The subspace of piecewise linear functions ϕ^k_ν with

$$\phi^k_\nu \in H^1_0(\Omega),\ \phi^k_\nu\,|_{\tau^{s,k}_\mu} \in \mathbb{P}_1(\tau^{s,k}_\mu)$$

is denoted by $\mathbb{V}_k$, where $\mathbb{P}_1$ is the space of polynomials of degree ≤ 1. A basis of $\mathbb{V}_k$ is the system of the usual hat-functions $\Phi_k = \{\phi^k_\nu\}_{\nu \in \mathcal{I}_n}$ with $\mathcal{I}_n = \{(\nu_1, \nu_2) \in \mathbb{N}^2, \nu_1, \nu_2 \le n-1\}$ uniquely defined by

$$\phi^k_\nu(x^k_\mu) = \delta_{\nu\mu}$$

and $\phi^k_\nu \in \mathbb{V}_k$, where $\delta_{\nu\mu}$ is the Kronecker delta for multi-indices. Now, we can formulate the discretized problem.

Find $u^k \in \mathbb{V}_k$ such that

$$a(u^k, v^k) = (f, v^k) \quad \forall v^k \in \mathbb{V}_k \tag{2.3}$$

holds. Problem (2.3) is equivalent to solving the system of linear algebraic equations

$$K_k \underline{u}_k = \underline{f}_k, \tag{2.4}$$

where $K_k = \left[a(\phi_\nu^k, \phi_\mu^k)\right]_{\nu,\mu\in\mathcal{I}_n}$, $\underline{u}_k = [u_\nu]_{\nu\in\mathcal{I}_n}$, and $\underline{f}_k = \left[(f, \phi_\mu^k)\right]_{\mu\in\mathcal{I}_n}$. The size of the matrix K_k is $N \times N$ with $N = (n-1)^2$.

3. Definition of the preconditioners. In this section, we define the overlapping preconditioners for the matrix K_k (2.3). We distinguish between two cases:

- the weight function ω_1 is assumed to be constant;
- both weight functions satisfy $\omega_i(0) = 0$, $i = 1, 2$.

3.1. The case $\omega_1(\xi) = 1$. We introduce the following notation. Let

- $\Omega_{i,x} = (2^{-1-i}, 2^{-i}) \times (0,1)$, $i = 0, \dots, k-2$, be a strip of width 2^{-i-1};
- $\Omega_{k-1,x} = (0, 2^{-k+1}) \times (0,1)$ be a strip of width 2^{-k+1}, i.e., $2h$;
- $\Gamma_{i,x} = \{2^{-i}\} \times (0,1)$, $i = 1, \dots, k-1$, be the interface of two neighboring stripes;
- $\tilde{\Omega}_{j,x} = \operatorname{int}(\bigcup_{i=j}^{k-1} \overline{\Omega_{i,x}}) = (0, 2^{-i}) \times (0,1)$ be a strip of width 2^{-i}; and
- $n_j = 2^{k-j} - 1$ be the number of interior grid points in $\tilde{\Omega}_{j,x}$ in the x-direction, $N_j = (n-1)n_j$ the total number of interior grid points, and $n_k = -1$.

Moreover, let

$$\varepsilon_{2,j} = \left(\omega_2\left(2^{-j}\right)\right)^2.$$

Figure 3.1 displays a sketch with the notation for $k = 4$.

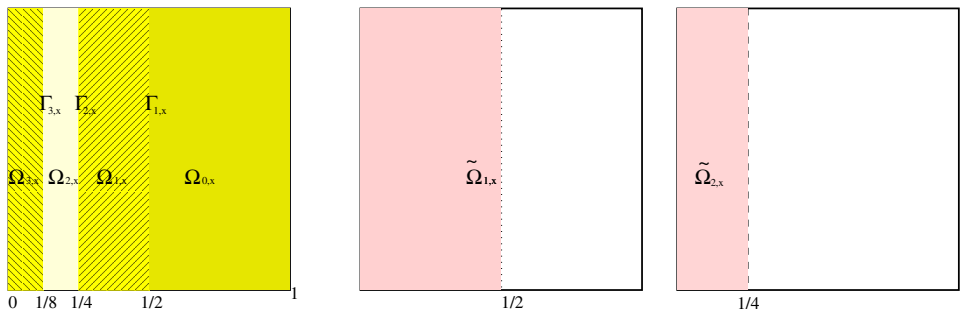


FIG. 3.1. *Notation for $k = 4$.*

We will develop two overlapping additive Schwarz preconditioners with inexact subproblem solvers. For the first preconditioner, we split the domain Ω into the substripes $\{\tilde{\Omega}_{i,x}\}_{i=0}^{k-1}$. On $\{\tilde{\Omega}_{i,x}\}_{i=0}^{k-1}$, we choose the constant diffusion matrix $\operatorname{argsup}_{(x,y)\in\tilde{\Omega}_{j,x}} \mathcal{D}(x,y)$, i.e., the componentwise supremum of the original diffusion matrix $\mathcal{D}$. For the second preconditioner, we consider the domain decomposition into $\{\overline{\Omega_{i,x}} \cup \overline{\Omega_{i+1,x}}\}_{i=0}^{k-2}$ and into $\Omega_{k-1,x}$. The construction of the diffusion matrix is similar to the first preconditioner.

For the correct mathematical definition of the preconditioners, we introduce the bilinear form

$$a_j(u,v) = \int_{\tilde{\Omega}_{j,x}} (\nabla u)^T \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon_{2,j} \end{bmatrix} \nabla v, \quad j = 0, \dots, k-1. \tag{3.1}$$

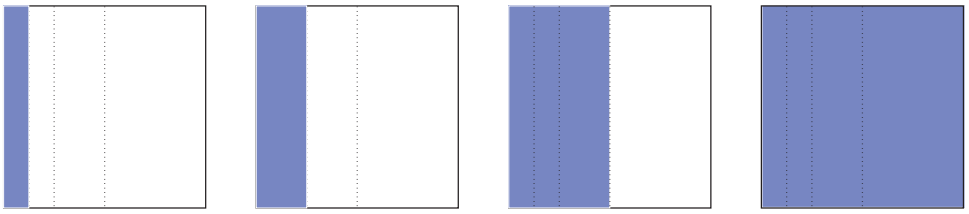


FIG. 3.2. *Computational domains for C (3.4), Θ_3, Θ_2, Θ_1, and Θ_0, and corresponding diffusion matrices for $\omega_2^2(\xi) = \xi^2$ (from left to right).*

This is a bilinear form with constant coefficients on $\tilde{\Omega}_{j,x}$. Let C_j be the stiffness matrix

$$C_j = \left[a_j(\phi_\mu^k, \phi_\nu^k)\right]_{\mu,\nu\in I_j}, \quad j = 0, \dots, k-1, \tag{3.2}$$

$$\text{with} \quad I_j = \left\{\nu \in \mathbb{N}^2 : \frac{1}{n}\nu \in \tilde{\Omega}_{j,x}\right\} = \{(\nu_1, \nu_2) \in \mathbb{N}^2 : \nu_1 \le n_j, \nu_2 \le n_0\}$$

corresponding to the bilinear form $a_j(\cdot,\cdot)$. Finally, let

$$\Theta_j = \begin{bmatrix} C_j & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{N-N_j} \end{bmatrix} \in \mathbb{R}^{N\times N} \tag{3.3}$$

be the corresponding global assembled stiffness matrix. The computational domains for Θ_j are displayed in Figure 3.2.

Then we define

$$C^{-1} = \sum_{j=0}^{k-1} \Theta_j^+ \tag{3.4}$$

as a first preconditioner for K_k, where B^+ denotes the pseudoinverse of a matrix B. Note that the locally anisotropic diffusion matrix $\mathcal{D}(x,y)$ is hidden in the matrix Θ_j.

This preconditioner turns out not to be optimal; see Theorem 3.2. To develop an optimal preconditioner, we have to modify C. Therefore, let

$$C_{j,mod} = \left[\int_{\overline{\Omega}_{j+1,x}\cup\overline{\Omega}_{j,x}} \nabla\phi_\mu^k \cdot \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon_{2,j} \end{bmatrix} \nabla\phi_\nu^k\right]_{\mu,\nu\in I_{j,mod}} \tag{3.5}$$

$$\begin{aligned}\text{with} \quad I_{j,mod} &= \left\{\nu \in \mathbb{N}^2 : \frac{1}{n}\nu \in \operatorname{int}(\overline{\Omega}_{j+1,x} \cup \overline{\Omega}_{j,x})\right\} \\ &= \{(\nu_1, \nu_2) \in \mathbb{N}^2 : n_{j+2} + 2 \le \nu_1 \le n_j, \nu_2 \le n_0\}.\end{aligned}$$

This is the discretized operator on $\overline{\Omega_{j+1,x}} \cup \overline{\Omega_{j,x}}$ with Dirichlet boundary conditions at all edges. Moreover, let

$$\Theta_{j,mod} = \begin{bmatrix} \mathbf{0}_{N_{j+2}+n_0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & C_{j,mod} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{N-N_j} \end{bmatrix} \in \mathbb{R}^{N\times N}, \quad j = 0, \dots, k-2,$$

be the corresponding assembled matrix; see Figure 3.3 for the computational domains.

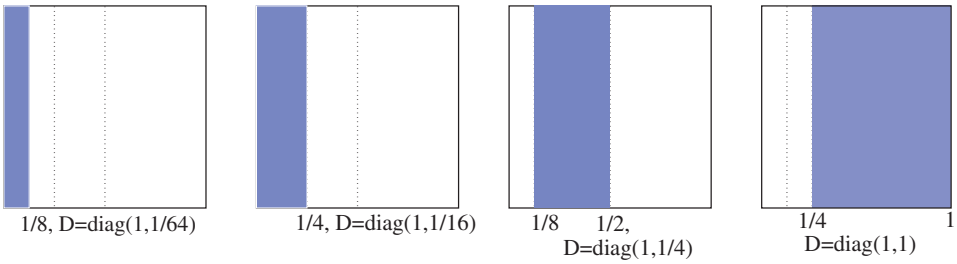


FIG. 3.3. *Computational domains for C_{mod} (3.6), Θ_3, $\Theta_{2,mod}$, $\Theta_{1,mod}$, and $\Theta_{0,mod}$, and corresponding diffusion matrices for $\omega_2^2(\xi) = \xi^2$ (from left to right).*

The second overlapping preconditioner for K_k is defined as

$$C_{mod}^{-1} = \sum_{j=0}^{k-2} \Theta_{j,mod}^{+} + \Theta_{k-1}^{+}. \tag{3.6}$$

Then we can formulate the following theorem.

THEOREM 3.1. *Let C_{mod} be defined via (3.6) and let $\mathcal{D}(x,y) = \left[\begin{smallmatrix} 1 & 0 \\ 0 & \omega_2^2(x) \end{smallmatrix}\right]$. Then the matrix C_{mod}^{-1} is symmetric positive definite and satisfies $K_k \sim C_{mod}$.*

Proof. A detailed proof is given in subsection 5.2. ∎

Concerning the first preconditioner (3.4), we can now prove the following result.

THEOREM 3.2. *Let C be defined via (3.4) and let $\mathcal{D}(x,y) = \left[\begin{smallmatrix} 1 & 0 \\ 0 & \omega_2^2(x) \end{smallmatrix}\right]$. Then the spectral equivalence relations $\frac{1}{k} K_k \leq C \preceq K_k$ hold.*

Proof. The proof is given in subsection 5.3. ∎

3.2. The general case. In addition to the notation of subsection 3.1, we define the following:

- $\Omega_{i,y} = (0,1) \times (2^{-1-i}, 2^{-i})$, $i = 0, \ldots, k-2$.
- $\Omega_{k-1,y} = (0,1) \times (0, 2^{-k+1})$.
- $\tilde{\Omega}_{j,y} = \operatorname{int}\left(\bigcup_{i=j}^{k-1} \overline{\Omega_{i,y}}\right) = (0,1) \times (0, 2^{-i})$.
- $\Omega_{jj'} = \operatorname{int}((\overline{\Omega_{j,x}} \cup \overline{\Omega_{j+1,x}}) \cap (\overline{\Omega}_{j',y} \cup \overline{\Omega}_{j'+1,y}))$ for $j, j' = 0, \ldots, k-2$. Note that $\Omega_{jj'} = (2^{-2-j}, 2^{-j}) \times (2^{-2-j'}, 2^{-j'})$ for $j, j' \leq k-3$.
- Moreover, let

$$\varepsilon_{1,j} = \left(\omega_1\left(2^{-j}\right)\right)^2.$$

Again, we will define two preconditioners. In comparison to the preconditioners of subsection 3.1, we use the domain decomposition in both directions. More precisely, we introduce the bilinear form

$$b_{j,j'}(u,v) = \int_{\tilde{\Omega}_{j,x} \cap \tilde{\Omega}_{j',y}} \nabla u \cdot \begin{bmatrix} \varepsilon_{1,j'} & 0 \\ 0 & \varepsilon_{2,j} \end{bmatrix} \nabla v, \quad j', j = 0, \ldots, k-1.$$

For $j = 0, \ldots, k-1$, let $B_{j,j'}$ be the stiffness matrix

$$B_{j,j'} = \left[b_{j,j'}(\phi_\mu^k, \phi_\nu^k)\right]_{\mu,\nu \in J_{j,j'}} \quad \text{with}$$

$$J_{j,j'} = \{\nu \in \mathbb{N}^2, \frac{1}{n}\nu \in \tilde{\Omega}_{j,x} \cap \tilde{\Omega}_{j',y}\} = \{(\nu_1, \nu_2) \in \mathbb{N}^2, \nu_1 \leq n_j, \nu_2 \leq n_{j'}\},$$

corresponding to the bilinear form $b_{j,j'}(\cdot,\cdot)$ with Dirichlet boundary conditions at all edges.

The corresponding global assembled stiffness matrices is denoted by the matrix $\Upsilon_{j,j'} \in \mathbb{R}^{N\times N}$. Then we define

$$B^{-1} = \sum_{j=0}^{k-1}\sum_{j'=0}^{k-1} \Upsilon^{+}_{j,j'} \tag{3.7}$$

as a first preconditioner for K_k. Note that the locally anisotropic diffusion matrix $\mathcal{D}(x,y)$ is hidden in the matrix $\Upsilon_{j,j'}$. This gives us a nonoptimal preconditioner; see Theorem 3.4. Moreover, we introduce an optimal preconditioner. For $0 \le j, j' \le k-2$, let

$$B_{j,j',mod} = \left[\int_{\Omega_{jj'}} \nabla\phi^k_\mu \cdot \begin{bmatrix} \varepsilon_{1,j'} & 0 \\ 0 & \varepsilon_{2,j} \end{bmatrix} \nabla\phi^k_\nu\right]_{\mu,\nu\in J_{j,j',mod}} \quad \text{with} \tag{3.8}$$

$$\begin{aligned} J_{j,j',mod} &= \left\{\nu \in \mathbb{N}^2, \frac{1}{n}\nu \in \Omega_{jj'}\right\} \\ &= \{(\nu_1,\nu_2) \in \mathbb{N}^2, n_{j+2}+2 \le \nu_1 \le n_j, n_{j'+2}+2 \le \nu_2 \le n_{j'}\}. \end{aligned}$$

This matrix is the finite element discretization matrix of an operator with piecewise constant coefficients on $\Omega_{jj'} = (\overline{\Omega_{j+1,x}} \cup \overline{\Omega_{j,x}}) \cap (\overline{\Omega_{j'+1,y}} \cup \overline{\Omega_{j',y}})$ and Dirichlet boundary conditions at all edges. For $j, j' \le k-2$, the corresponding global assembled stiffness matrices are denoted by the matrices $\Upsilon_{j,j',mod} \in \mathbb{R}^{N\times N}$. If $j = k-1$ or $j' = k-1$, we set

$$\Upsilon_{j,j',mod} = \Upsilon_{j,j'}.$$

The corresponding computational domains for the matrices $\Upsilon_{j,j'}$ are displayed in Figure 3.4.

The second overlapping preconditioner for K_k is defined as

$$B^{-1}_{mod} = \sum_{j=0}^{k-1}\sum_{j'=0}^{k-1} \Upsilon^{+}_{j,j',mod}. \tag{3.9}$$

THEOREM 3.3. *Let B_{mod} be defined via (3.9) and let $\mathcal{D}(x,y) = \left[\begin{smallmatrix} \omega_1^2(y) & 0 \\ 0 & \omega_2^2(x) \end{smallmatrix}\right]$. Then the matrix B^{-1}_{mod} is symmetric positive definite and satisfies $K_k \sim B_{mod}$.*

Proof. The proof is given in subsection 5.4. ∎

Concerning the first preconditioner (3.7), we can prove the following result.

THEOREM 3.4. *Let B be defined via (3.7) and let $\mathcal{D}(x,y) = \left[\begin{smallmatrix} \omega_1^2(y) & 0 \\ 0 & \omega_2^2(x) \end{smallmatrix}\right]$. Then the spectral equivalence relations $\frac{1}{k^2}K_k \le B \preceq K_k$ hold.*

Proof. The proof is similar to the proof of Theorem 3.2. ∎

Remark 3.5. We can replace the matrices $B_{j,j',mod}$ in (3.8) by

$$B_{j,j',var} = \left[\int_{\Omega_{jj'}} \nabla\phi^k_\mu \cdot \begin{bmatrix} \omega_1^2(y) & 0 \\ 0 & \omega_2^2(x) \end{bmatrix} \nabla\phi^k_\nu\right]_{\mu,\nu\in J_{j,j',mod}}.$$

Let $\Upsilon_{j,j',var}$ be the assembled matrices and

$$B^{-1}_{var} = \sum_{j=0}^{k-1}\sum_{j'=0}^{k-1} \Upsilon^{+}_{j,j',var}.$$

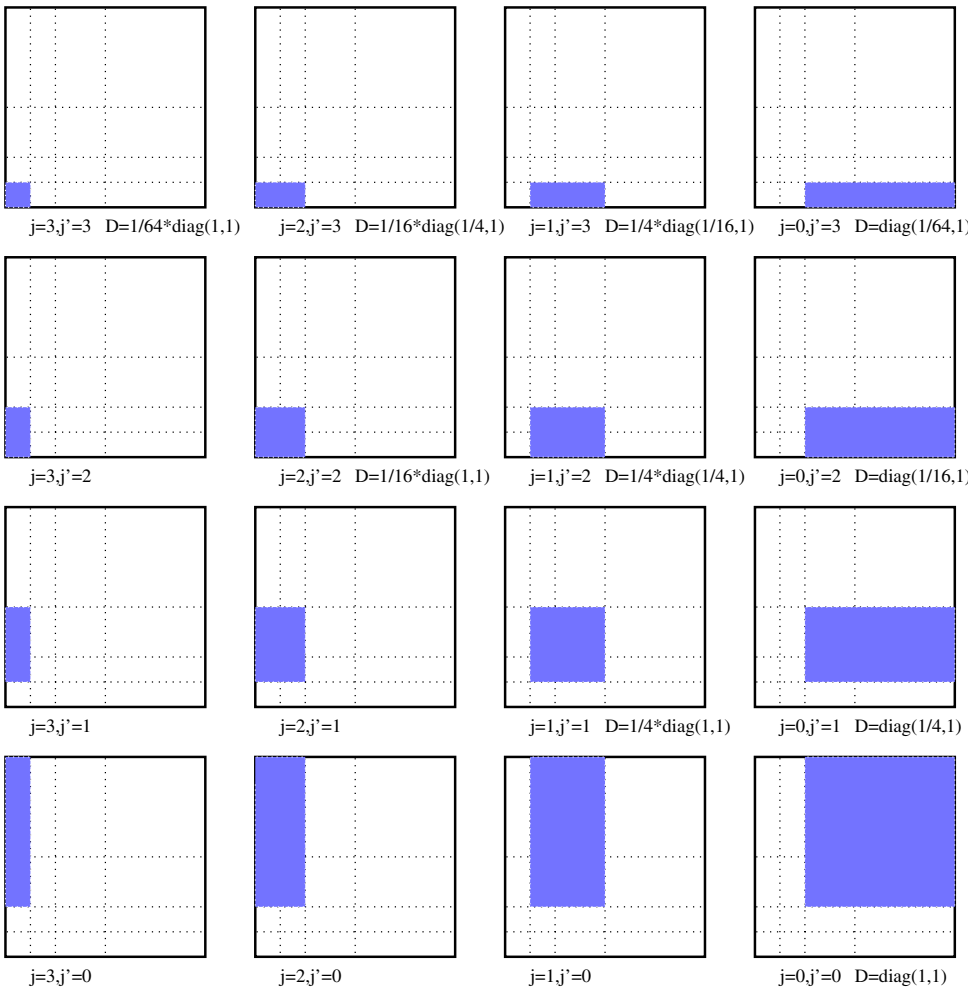


FIG. 3.4. *Corresponding domains $\Omega_{jj'}$ and diffusion matrices with weight functions $\omega_i^2(\xi) = \xi^2$ for $\Upsilon_{j,j',mod}$, $i = 1, 2$.*

Due to (2.2), we have $B_{j,j',var} \sim B_{j,j',mod}$, which gives $B_{mod} \sim B_{var}$. In the preconditioner B_{var}, we now have an operator with variable coefficients. However, the constants do not change too much since we have the estimate

$$\sup_{x_1,x_2 \in (\overline{\Omega}_{j,x} \cup \overline{\Omega}_{j+1,x})} \frac{\omega_2^2(x_1)}{\omega_2^2(x_2)} \leq c_\omega^4$$

from our assumption (2.2).

3.3. Computational aspects. In this subsection, we investigate the preconditioning operation $\mathcal{C}^{-1}\underline{w}$ for the preconditioners (3.4), (3.6), (3.7), and (3.9). We present algorithms to perform this preconditioning operation in optimal arithmetical complexity.

Let us start with the case $\omega_1(\xi) = 1$. Here, we have developed the preconditioners

$$C^{-1} = \sum_{j=0}^{k-1} \Theta_j^+$$

(see (3.4)) and

$$C_{mod}^{-1} = \sum_{j=0}^{k-2} \Theta_{j,mod}^+ + \Theta_{k-1}^+$$

(see (3.6)). In both cases, we have to solve systems of linear algebraic equations with the discretization of an operator with constant coefficients on a rectangle using triangular finite elements. The corresponding operators are now

$$-\varepsilon_j^2 u_{xx} - u_{yy} \tag{3.10}$$

with some numbers $0 < \varepsilon_j \leq 1$. The computational domains are displayed in Figures 3.2 and 3.3. These domains are the same as for the preconditioners C and C_{mod} of [4].

Using multigrid preconditioners combined with a line smoother, optimal solvers for Θ_j and $\Theta_{j,mod}$ can easily be designed; see [11]. The line smoother is necessary to remove the anisotropy of the operator. It can be shown that the multigrid preconditioner with line smoother and V-cycle is an optimal method independent of the parameter ε_j [11]. With the same arguments as in the isotropic case (see [4]), we can prove that the cost for the operations $\underline{w} = C^{-1}\underline{r}$ and $\underline{w} = C_{mod}^{-1}\underline{r}$ depends linearly on the number of unknowns.

In the general case, the application of the preconditioning operations $B^{-1}\underline{r}$ (3.7) and $B_{mod}^{-1}\underline{r}$ (3.9) implies again the solution of systems of linear algebraic equations with discretizations of operators of type (3.10). However, these systems have to be solved on the smaller subdomains $(\overline{\Omega}_{j,x} \cup \overline{\Omega}_{j+1,x}) \cap (\overline{\Omega}_{j',y} \cup \overline{\Omega}_{j'+1,y})$; see Figure 3.4 for $k = 4$. The structure of the diffusion matrices are displayed below each of the 16 panels for the weight functions $\omega_i^2(\xi) = \xi^2$, $i = 1, 2$. The diffusion matrices are isotropic for $j \approx j'$ and globally anisotropic elsewhere. Therefore, a multigrid algorithm or multigrid preconditioner with line smoother should be used as a solution method for $|j - j'| \gg 1$. Similarly as for the preconditioners C (3.4) and C_{mod} (3.6), the optimality of the preconditioning operations $B^{-1}\underline{r}$ and $B_{mod}^{-1}\underline{r}$ can be shown.

Summarizing, we now have to solve globally anisotropic problems with constant coefficients instead of locally anisotropic problems with changing directions of the anisotropy. This is much simpler than the original problem since solvers for the problem with constant coefficients are known in the literature. However, this method cannot remove the anisotropic behavior of the problem.

4. Some preliminaries. In this section, we formulate some auxiliary results from ASM which are necessary to prove our main results. The proofs can be found in the literature.

4.1. Preliminaries from ASM. The first result is a general result for preconditioned ASM.

LEMMA 4.1. *Let $\mathbb{H}$ be a Hilbert space with the scalar product $(\cdot, \cdot)$. Moreover, let $\mathbb{H}_i$, $i = 1, \ldots, m$, be subspaces of $\mathbb{H}$ such that*

$$\mathbb{H} = \mathbb{H}_1 + \mathbb{H}_2 + \cdots + \mathbb{H}_m.$$

Let $\mathcal{A} : \mathbb{H} \mapsto \mathbb{H}$ be a linear, self-adjoint, bounded, and positive definite operator and let

$$(u, v)_{\mathcal{A}} = (\mathcal{A}u, v) \quad \forall u, v \in \mathbb{H}.$$

We denote by P_i, $i = 1, \ldots, m$, the orthogonal projection operators from $\mathbb{H}$ onto $\mathbb{H}_i$ with respect to the scalar product $(\cdot, \cdot)_{\mathcal{A}}$. We assume that for any $u \in \mathbb{H}$ there exists a decomposition $u = u_1 + \cdots + u_m$ such that

$$c_1 \sum_{i=1}^{m} (u_i, u_i)_{\mathcal{A}} \leq (u, u)_{\mathcal{A}} \tag{4.1}$$

with a positive constant c_1. Moreover, let c_2 be some positive constant such that

$$\sum_{i=1}^{m} (P_i u, u)_{\mathcal{A}} \leq c_2 \, (u, u)_{\mathcal{A}} \quad \forall u \in \mathbb{H}. \tag{4.2}$$

Also, let $\mathcal{B}_i : \mathbb{H}_i \mapsto \mathbb{H}_i$, $i = 1, \ldots, m$, be some self-adjoint and surjective operators such that

$$c_3 \, (\mathcal{B}_i u_i, u_i) \leq (\mathcal{A} P_i u_i, P_i u_i) \leq c_4 \, (\mathcal{B}_i u_i, u_i) \quad \forall u_i \in \mathbb{H}_i, \; i = 1, \ldots, m. \tag{4.3}$$

Let $\mathcal{B}^{-1} = \mathcal{B}_1^{+} + \cdots + \mathcal{B}_m^{+}$, where $\mathcal{B}_i^{+}$ denotes the pseudoinverse operator of $\mathcal{B}_i$. Then

$$c_1 c_3 \left(\mathcal{A}^{-1} u, u\right) \leq \left(\mathcal{B}^{-1} u, u\right) \leq c_2 c_4 \left(\mathcal{A}^{-1} u, u\right) \quad \forall u \in \mathbb{H}.$$

Proof. The proof can be found in [16]. ∎

The second result is a technical result for some overlapping preconditioners, in which the domain is split into stripes as displayed in Figure 3.1. First, let us introduce some notation which is similar to the notation in Figure 3.1.

- Let

  $$\overline{\Omega} = \bigcup_{j=0}^{k-1} \overline{\Omega}_j$$

  be a domain Ω which is decomposed into stripes Ω_i, i.e.,

  $$\Omega_i \cap \Omega_j = \emptyset \quad \text{for } i \neq j, \quad \overline{\Omega}_i \cap \overline{\Omega}_j = \begin{cases} \Gamma_i, & i = j + 1, \\ \Gamma_j, & i = j - 1, \\ \overline{\Omega}_i, & i = j, \\ \emptyset, & |i - j| \geq 2, \end{cases}$$

  and let $\overline{\Omega}_{k-1} \cap \partial\Omega = \Gamma_k$.
- Let τ_k be a triangulation of Ω which is admissible to the decomposition of Ω into Ω_i.
- Let $\Phi_k = [\phi_i]_{i=1}^{N}$ be the basis of hat-functions according to the triangulation τ_k and let $\mathbb{V}_k = \text{span}\Phi_k$ be the corresponding finite element space.
- Let $a(\cdot, \cdot) : \mathbb{V}_k \times \mathbb{V}_k \mapsto \mathbb{R}$ be a symmetric and positive definite bilinear form and let

  $$\|u\|_{a,\Omega} = a(u, u)$$

  be the energetic norm. In the same way, let

  $$\|u\|_{a,\tilde{\Omega}} = a \mid_{\tilde{\Omega}} (u, u)$$

  be the restriction of the norm onto some subdomain $\tilde{\Omega} \subset \Omega$.

- For $j = 0, \ldots, k-2$, let $\mathbb{Y}_j = \{u \in \mathbb{V}_k : \text{supp } u \subset \overline{\Omega}_j \cup \overline{\Omega}_{j+1}\}$ be the restriction of the finite element space $\mathbb{V}_k$ onto $\overline{\Omega}_j \cup \overline{\Omega}_{j+1}$ with Dirichlet boundary conditions at the boundaries Γ_j and Γ_{j+2}. For $j = k-1$, we set $\mathbb{Y}_{k-1} = \{u \in \mathbb{V}_k : \text{supp } u \subset \overline{\Omega}_{k-1}\}$.
- Let

$$\|w\|^2_{\Gamma_j,\text{left}} = \min_{\substack{u\in\mathbb{V}_k \\ u|_{\Gamma_j}=w \\ u|_{\Gamma_{j+1}}=0}} \|u\|^2_{a,\Omega_j} \quad \text{and} \tag{4.4}$$

$$\|w\|^2_{\Gamma_j,\text{right}} = \min_{\substack{u\in\mathbb{V}_k \\ u|_{\Gamma_j}=w \\ u|_{\Gamma_{j-1}}=0}} \|u\|^2_{a,\Omega_{j-1}}$$

  be the left and right trace norms on Γ_j.
- Let $\mathcal{T}_{j,\text{left}} : \mathbb{V}_k\,|_{\Gamma_j} \mapsto \mathbb{V}_k\,|_{\Omega_j}$ and $\mathcal{T}_{j,\text{right}} : \mathbb{V}_k\,|_{\Gamma_j} \mapsto \mathbb{V}_k\,|_{\Omega_{j-1}}$ be the minimal energetic extension operators from Γ_j to Ω_j and Ω_{j-1}, i.e.,

$$\|w\|_{\Gamma_j,\text{left}} = \|\mathcal{T}_{j,\text{left}} w\|_{a,\Omega_j} \quad \text{and} \quad \|w\|_{\Gamma_j,\text{right}} = \|\mathcal{T}_{j,\text{right}} w\|_{a,\Omega_{j-1}}.$$

THEOREM 4.2. *In addition to the above assumptions, let us assume that there exists an integer j_0 such that the following hold:*

- *There exists a constant $\gamma < 1$, which is independent of the discretization parameter and j, such that*

$$a(\mathcal{T}_{j,\text{left}} u, \mathcal{T}_{j+1,\text{right}} v) \le \gamma \|u\|_{\Gamma_j,\text{left}} \|v\|_{u,\Gamma_{j+1},\text{right}} \quad \forall j = 0, \ldots, j_0,\ \ \forall u \in \mathbb{Y}_j\,|_{\Gamma_j},\ \forall v \in \mathbb{Y}_{j+1}\,|_{\Gamma_{j+1}}. \tag{4.5}$$

- *There exists a constant $q_0 < 1$ and a constant c_2, which are independent of j and the discretization parameter, such that*

$$q_0^{-1} \|w\|^2_{\Gamma_j,\text{left}} \le \|w\|^2_{\Gamma_j,\text{right}} \le c_2 \|w\|^2_{\Gamma_j,\text{left}} \quad \forall w, \quad j = j_0. \tag{4.6}$$

- *There exists a constant c_1, which is independent of the discretization parameter, such that*

$$c_1^{-1} \|w\|^2_{\Gamma_j,\text{left}} \le \|w\|^2_{\Gamma_j,\text{right}} \le c_2 \|w\|^2_{\Gamma_j,\text{left}} \quad \forall w, \quad \forall j = j_0+1, \ldots, k-1. \tag{4.7}$$

Then there exists a decomposition $u = \sum_{j=0}^{k-1} u_j$ with $u_j \in \mathbb{Y}_j$ such that

$$c_L^2 \sum_{j=0}^{k-1} a(u_j, u_j) \le a(u,u) \quad \forall u \in \mathbb{V}_k.$$

The constant $c_L > 0$ depends only on γ, c_1, c_2, and q_0. Moreover, for all decompositions of $u = \sum_{j=0}^{k-1} u_j$ with $u_j \in \mathbb{Y}_j$, the estimate

$$a(u,u) \le\ 2 \sum_{j=0}^{k-1} a(u_j, u_j) \quad \forall u \in \mathbb{V}_k$$

holds.

Proof. The proof can be found in [4]. ∎

Next, we construct a bilinear form $a_p(\cdot,\cdot)$ with piecewise constant coefficients which is spectrally equivalent to the original bilinear form $a(\cdot,\cdot)$; cf. (2.1). This idea has originally been developed in [12]. For $i = 1, 2$, let

$$\chi_i^2(\xi) = \varepsilon_{i,j}, \quad \xi \in (2^{-j-1}, 2^{-j}) \quad \text{with} \quad \varepsilon_{i,j} := \omega_i^2(2^{-j}) \tag{4.8}$$

be a piecewise constant coefficient function and let

$$a_p(u,v) := \int_\Omega (\nabla v)^T \begin{bmatrix} \chi_1^2(y) & 0 \\ 0 & \chi_2^2(x) \end{bmatrix} \nabla u \tag{4.9}$$

be the corresponding bilinear form. Moreover, we define the energetic norm

$$\|u\|_p^2 := a_p(u,u) \quad \forall u \in \mathbb{V}_k \tag{4.10}$$

with respect to the bilinear form $a_p(\cdot,\cdot)$. The stiffness matrix with respect to the basis Φ_k is denoted by $K_{k,p}$, i.e.,

$$K_{k,p} = [a_p(\phi_{il}, \phi_{i'l'})]_{(i,i'),(l,l')=(1,1)}^{n_0,n_0}. \tag{4.11}$$

LEMMA 4.3. *Let us assume that the weight functions ω_i, $i = 1, 2$, satisfy Assumption* 2.1. *Then we have*

$$a(u,u) \le a_p(u,u) \le 2c_\omega^2\, a(u,u) \quad \forall u \in \mathbb{V}_k.$$

The constant c_ω is from (2.2).

Proof. The proof is similar to the proof of Lemma 4.3 in [4]. ∎

Remark 4.4.

- In the case of the weight function $\omega_i^2(\xi) = \xi^\alpha$ with $\alpha > 0$, we have $c_\omega^2 = 2^\alpha$.
- A direct consequence of (2.4), (4.11), and Lemma 4.3 is the spectral equivalence estimate

$$\frac{1}{2} c_\omega^{-2} K_{k,p} \le K_k \le K_{k,p}.$$

4.2. Some estimates for tridiagonal matrices. Finally, some estimates for tridiagonal matrices with constant main and subdiagonals are required. For a fixed integer m and some positive parameter κ, we introduce

$$F_m = \begin{bmatrix} 2+\kappa & -1 & & & 0 \\ -1 & 2+\kappa & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2+\kappa & -1 \\ 0 & & & -1 & 2+\kappa \end{bmatrix} \in \mathbb{R}^{m-1\times m-1},$$

$$\tilde{F}_m = \begin{bmatrix} 1+\frac{\kappa}{2} & e_1^T \\ e_1 & F_m \end{bmatrix} \in \mathbb{R}^{m\times m}, e_1 = (1,0,\ldots,0)^T \in \mathbb{R}^{m-1\times 1}, \tag{4.12}$$

and the real number

$$q = 1 + \frac{\kappa}{2} + \frac{1}{2}\sqrt{\kappa(\kappa+4)}. \tag{4.13}$$

We prove the following lemma.

LEMMA 4.5. *Let F_m and $\tilde{F}_m$ be defined via* (4.12). *Then the following assertions are valid:*

- *The determinant of F_m satisfies the equation*

$$\det F_{m+1} = \frac{q^{m+2} - q^{-m}}{q^2 - 1} = q^{-m} \sum_{i=0}^{m} q^{2i}. \tag{4.14}$$

- *Let $s_m = 1 + \frac{\kappa}{2} - e_1^T F_m^{-1} e_1$ be the Schur complement of $\tilde{F}_m$ with respect to the first row and column. Then*

$$s_m = \frac{\kappa}{2} + \frac{1 + q^{2m-1}}{1 + q + \cdots + q^{2m-1}}. \tag{4.15}$$

Moreover, the estimate

$$\frac{\kappa}{2} + \frac{1}{m} \leq s_m \leq \frac{\kappa}{2} + \frac{1}{m} \frac{1 + q^{2m-1}}{2\sqrt{q^{2m-1}}} \tag{4.16}$$

holds.

- *Let $\hat{s}_m = e_1^T F_m^{-1} e_m$, $e_m = (0, \ldots, 0, 1)^T$. Then*

$$|\hat{s}_m| = \frac{1}{\det F_{m-1}}. \tag{4.17}$$

- *Let $\gamma_m = \frac{|\hat{s}_m|}{s_m}$. Then*

$$\gamma_m \leq \frac{2}{q^{m-1}}. \tag{4.18}$$

Proof. Relation (4.14) is a consequence of the following recursion:

$$\begin{aligned} \det F_m &= (2 + \kappa) \det F_{m-1} - \det F_{m-2}, \\ \det F_0 &= 1, \ \det F_1 = 2 + \kappa. \end{aligned}$$

The solution of this linear recursion of second order gives the first assertion. The second relation follows from the first one by the geometric series

$$1 + q^2 + \cdots + q^{2m} = \frac{q^{2m+2} - 1}{q^2 - 1}.$$

To prove relation (4.15), we compute the Schur complement by using Cramer's rule and (4.14) explicitly. Since

$$e_1^T F_M^{-1} e_1 = (F_m^{-1})_{(1,1)} = \frac{\det F_{m-1}}{\det F_m},$$

we conclude

$$s_m = \frac{\kappa}{2} + \frac{\det F_m - \det F_{m-1}}{\det F_m}.$$

We simplify the second summand with (4.14) and obtain

$$\det F_m - \det F_{m-1} = q^{-m+1} \sum_{i=0}^{2m-2} (-q)^i = q^{-m+1} \frac{q^{2m-1} + 1}{1 + q}. \tag{4.19}$$

Hence,

$$\frac{\det F_m - \det F_{m-1}}{\det F_m} = \frac{1+q^{2m-1}}{1+q+\cdots+q^{2m-1}}, \tag{4.20}$$

which proves (4.15). To prove (4.16), we start from (4.15) and use the convexity of the function $f:(1,\infty)\mapsto\mathbb{R}$ given by $f(x)=q^x$ for $q>1$. Then we have

$$\begin{aligned} 1+q^{2m-1} &\geq q+q^{2m-2},\\ 1+q^{2m-1} &\geq q^2+q^{2m-3},\\ &\vdots \quad \vdots\\ 1+q^{2m-1} &\geq q^{m-1}+q^m. \end{aligned}$$

Summing up over all inequalities yields

$$\frac{1+q^{2m-1}}{1+q+\cdots+q^{2m-1}} \geq \frac{1}{m},$$

which proves the lower estimate. For the upper estimate, the inequality of the mean values between arithmetical and geometrical means is used. Then we have

$$1+q+\cdots+q^{2m-1} \geq 2m\sqrt[2m]{q\cdot q^2\cdot\ldots\cdot q^{2m-1}} = 2m\sqrt{q^{2m-1}}.$$

This proves the lower estimate of (4.16).

The proof of (4.17) is similar to the proof of (4.15).

For the proof of (4.18), we use relations (4.15), (4.17) and equation (4.19). We obtain

$$\gamma_m = \frac{|\hat{s}_m|}{s_m} \leq \frac{1}{\det F_m - \det F_{m-1}} = \frac{q^{m-1}(1+q)}{q^{2m-1}+1} \leq \frac{1+q}{qq^{m-1}}.$$

Since $\frac{1+q}{q}\leq 2$ for $q\geq 1$, the assertion (4.18) follows, which proves the lemma. ∎

The next lemma gives some asymptotic estimates for the Schur complement s_m and the constant γ_m.

LEMMA 4.6.

- *Let $m\geq\max\{\frac{1}{\sqrt{\kappa}},2\}$. Then we have*

$$\gamma_m < \frac{20}{21}. \tag{4.21}$$

- *Let $2\leq m\leq\frac{1}{\sqrt{\kappa}}$ and $m\in\mathbb{N}$. Then the estimate*

$$\frac{1}{m} \leq s_m \leq \frac{9}{5}\frac{1}{m} \tag{4.22}$$

is valid.

Proof. To prove the first assertion, we use (4.13) and obtain

$$q^{m-1} = \left(1+\frac{\kappa}{2}+\frac{1}{2}\sqrt{\kappa(\kappa+4)}\right)^{m-1} \geq \left(1+\frac{\kappa}{2}+\sqrt{\kappa}\right)^{m-1}.$$

With $m \geq \frac{1}{\sqrt{\kappa}}$, we can conclude that

$$q^{m-1} \geq \left(1 + \frac{1}{2m^2} + \frac{1}{m}\right)^{m-1}.$$

The series $\{a_m\}_m$ given by $a_m = (1 + \frac{1}{2m^2} + \frac{1}{m})^{m-1}$ is monotonically increasing and satisfies $\lim_{m\to\infty} a_m = e$. Moreover, $a_m \geq \frac{21}{10}$ for $m \geq 4$. This gives

$$q^{m-1} \geq \frac{21}{10}.$$

Using (4.18), the assertion follows for $m \geq 4$. The case $m = 2$ implies that $\kappa \geq \frac{1}{4}$ and $q \geq \frac{13}{8}$. A direct computation shows

$$\gamma_2 \leq \frac{5}{6} < \frac{20}{21}.$$

A similar proof can be given for $m = 3$.

To prove the second assertion, we start with $\kappa < m^{-2}$. With the arguments we used above, we have

$$q^{2m-1} \leq \left(1 + \frac{1}{2m^2} + \frac{1}{m}\right)^{2m-1} \leq e^2.$$

Moreover, the function $f : [1, \infty) \mapsto \mathbb{R}$ given by

$$f(x) = \frac{1+x}{\sqrt{x}} = \left(\sqrt[4]{x} - \frac{1}{\sqrt[4]{x}}\right)^2 - 2$$

is monotonically increasing for $x \geq 1$. Hence, we can estimate

$$\frac{1 + q^{2m-1}}{\sqrt{q^{2m-1}}} \leq \frac{1 + e^2}{e}.$$

Now, we insert this estimate into (4.16) and can conclude that

$$\frac{1}{m} \leq s_m \leq \frac{\kappa}{2} + \frac{1}{m}\frac{1 + q^{2m-1}}{2\sqrt{q^{2m-1}}} \leq \frac{1}{m}\left(\frac{1}{2m} + \frac{1 + e^2}{2e}\right) \leq \frac{1}{m}\left(\frac{1}{4} + \frac{31}{20}\right) = \frac{9}{5m}.$$

This proves the lemma. ∎

5. Condition number estimates. In this section, we prove the central theorems of this paper. The proofs exploit the tensor product structure of the problems and use some auxiliary results from the one-dimensional case. The results for the one-dimensional case are presented in subsection 5.1. The proofs of Theorems 3.1, 3.2, and 3.3 are presented in subsections 5.2, 5.3, and 5.4, respectively.

5.1. Some one-dimensional auxiliary results. In this subsection, we prove some auxiliary results for the corresponding one-dimensional case. We start with the definition of a corresponding bilinear form and discretization matrices.

For $n = 2^k$ and $i = 1, 2$, let $I_{n-1} \in \mathbb{R}^{n-1\times n-1}$,

$$T_{n-1} = \begin{bmatrix} 2 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 2 \end{bmatrix} \in \mathbb{R}^{n-1\times n-1} \quad \text{and} \tag{5.1}$$

$$M_{\omega_i} = \operatorname{diag}(d_s)_{s=1}^{n-1} \quad \text{with} \quad d_s = \begin{cases} \varepsilon_{i,j} & \text{if} \quad 2^{k-j-1} < s < 2^{k-j}, \\ \frac{1}{2}(\varepsilon_{i,j} + \varepsilon_{i,j-1}) & \text{if} \quad 2^{k-j} = s, \end{cases} \tag{5.2}$$

be the identity, the unweighted Laplacian in one dimension, and a scaled weighted mass matrix with piecewise constant coefficients, respectively. The coefficients $\varepsilon_{i,j}$ are defined via (4.8).

Since T_{n_0} is the one-dimensional Laplacian, we introduce linear finite elements on the equidistant mesh $\mathcal{M}_n = \bigcup_{s=0}^{n-1} \tau_s^n$, where $\tau_s^n = \left(\frac{s}{n}, \frac{s+1}{n}\right)$. The one-dimensional hat-functions on this mesh,

$$\phi_s^n(x) = \begin{cases} nx - (s-1) & \text{on } \tau_{s-1}^n, \\ (s+1) - nx & \text{on } \tau_s^n, \\ 0 & \text{otherwise,} \end{cases} \qquad s = 1, \dots, n-1, \tag{5.3}$$

are a basis of the finite element space $\mathbb{X}_n = \operatorname{span}[\phi_s^n]_{s=1}^{n-1} = \operatorname{span}[\Phi_1]$. Let $\lambda > 0$. Then the matrix $T_{n-1} + \lambda M_{\omega_i}$ defines a coercive bilinear form $a_1(\cdot,\cdot)$ on $\mathbb{X}_n$, i.e.,

$$\underline{u}^T (T_{n-1} + \lambda M_{\omega_i}) \underline{v} = a_1([\Phi_1]\underline{u}, [\Phi_1]\underline{v}) := \frac{1}{n} \int_0^1 u'(x) v'(x)\, \mathrm{d}x + \sum_{s=1}^{n-1} \rho_{i,s} u\left(\frac{s}{n}\right) v\left(\frac{s}{n}\right) \tag{5.4}$$

with $\rho_{i,s} = \frac{1}{2}\lambda(\chi_i^2|_{\tau_{s-1}^n} + \chi_i^2|_{\tau_s^n})$, $i = 1, 2$. Due to the symmetry and positive definiteness of the matrix $T_{n-1} + \lambda M_{\omega_i}$, the bilinear form $a_1(\cdot,\cdot)$ is symmetric and coercive.

For $j = 0, \dots, k-2$, let $\Omega_j = \left(2^{-j-1}, 2^{-j}\right)$ and $\Omega_{k-1} = (0, 2^{k-1})$. Moreover, we introduce

$$\begin{aligned} \tilde{\mathbb{W}}_j &= \operatorname{span}\{\phi_i^n\}_{i=n_{j+1}+2}^{n_j}, \quad j = 0, \dots, k-1, \quad \text{and} \\ \mathbb{W}_j &= \operatorname{span}\{\phi_i^n\}_{i=n_{j+2}+2}^{n_j}, \quad j = 0, \dots, k-2, \quad \mathbb{W}_{k-1} = \tilde{\mathbb{W}}_{k-1}, \end{aligned}$$

where n_j is defined in subsection 3.1. Due to this definition, the spaces $\mathbb{W}_j$ and $\tilde{\mathbb{W}}_j$ are formed by those finite element functions of $\mathbb{X}_n$ which have support inside $\overline{\Omega_{j+1,x}} \cup \overline{\Omega_{j,x}}$ and $\overline{\Omega_{j,x}}$, respectively.

LEMMA 5.1. *There exists a decomposition $u = \sum_{j=0}^{k-1} u_j$ with $u_j \in \mathbb{W}_j$ such that*

$$a_1(u,u) \geq c^2 \sum_{j=0}^{k-1} a_1(u_j, u_j) \quad \forall u \in \mathbb{X}_n.$$

The constant $c^2 > 0$ does not depend on n and ρ_i.

Proof. We adapt the notation of Theorem 4.2, i.e., let $\Gamma_{j+1} = \overline{\Omega}_{j+1} \cup \overline{\Omega}_j$ and

$$\|w\|^2_{\Gamma_j,\text{left}} = \min_{\substack{u\in\mathbb{X}_n \\ u|_{\Gamma_j}=w \\ u|_{\Gamma_{j+1}}=0}} \|u\|^2_{a_1,\Omega_j} \quad \text{and} \quad \|w\|^2_{\Gamma_j,\text{right}} = \min_{\substack{u\in\mathbb{X}_n \\ u|_{\Gamma_j}=w \\ u|_{\Gamma_{j-1}}=0}} \|u\|^2_{a_1,\Omega_{j-1}} \tag{5.5}$$

be the left and right trace norms on Γ_j. Moreover, the minimal energy extension operators with respect to $a_1(\cdot,\cdot)$ from Γ_j to Ω_j and Ω_{j-1} are denoted by $\mathcal{T}_{j,\text{left}}$ and $\mathcal{T}_{j,\text{right}}$, respectively. Since the coefficients of the bilinear form $a_1(\cdot,\cdot)$ are constant on Ω_j and the discretization is symmetric with respect to the left and right boundaries, we have

$$\|w\|^2_{\Gamma_j,\text{left}} = \|w\|^2_{\Gamma_{j+1},\text{right}} \quad \forall w \in \mathbb{R} \tag{5.6}$$

for which the minima in (5.5) are achieved.

We now fix a stripe Ω_j. A simple computation shows

$$\left[a_1\left|_{\Omega_j}(\phi^n_l, \phi^n_{l'})\right]^{n_j}_{l,l'=n_{j+1}+2} = T_{m_j-1} + \kappa_{i,j} I_{m_j-1}, \quad i = 1,2, \tag{5.7}$$

with $m_j = 2^{k-j-1}$ and $\kappa_{i,j} = \varepsilon_{i,j}\lambda$; i.e., this matrix has the structure of the matrix F_m (4.12) with $m = m_j$ and $\kappa = \kappa_{i,j}$, $i = 1,2$. So, it is possible to apply the results about the matrix F_m. Due to the properties of the Schur complement, we have

$$\|w\|^2_{\Gamma_j,\text{left}} = \|w\|^2_{\Gamma_{j+1},\text{right}} = w^2 s_{m_j} \quad \forall w \in \mathbb{R} \tag{5.8}$$

with s_{m_j} defined via (4.15). A simple computation shows

$$a_1(\mathcal{T}_{j,\text{left}} u, \mathcal{T}_{j+1,\text{right}} v) = u \hat{s}_{m_j} v \quad \forall u, v \in \mathbb{R} \tag{5.9}$$

with $\hat{s}_{m_j}$ of (4.17). Hence, we can conclude that

$$\gamma^2_{m_j} = \max_{\substack{u,v\in\mathbb{R}\\ u,v\neq 0}} \frac{a_1(\mathcal{T}_{j,\text{left}} u, \mathcal{T}_{j+1,\text{right}} v)}{\|u\|_{\Gamma_j,\text{left}} \|v\|_{\Gamma_{j+1},\text{right}}} = \frac{\hat{s}_{m_j}}{s_{m_j}}. \tag{5.10}$$

The series $\{m_j\}_{j=0}^{k-1}$ is monotonically decreasing by definition. The series $\{\kappa_{i,j}\}_{j=0}^{k-1}$ is monotonically decreasing by Assumption 2.1. Hence, the series $\{m_j \kappa^2_{i,j}\}_{j=0}^{k-1}$ is monotonically decreasing, too. Consequently, there exists a number $j_0 \in \{-1, 0, \ldots, k\}$ such that

$$m_j \geq \kappa^{-2}_{i,j} \quad \forall j \leq j_0 \quad \text{and} \quad m_j \leq \kappa^{-2}_{i,j} \quad \forall j > j_0, i = 1,2. \tag{5.11}$$

Now, we verify the assumptions of Theorem 4.2. Using (5.10) and (4.21), we have

$$\gamma^2_{m_j} \leq \frac{20}{21} \quad \text{for} \quad j < j_0.$$

This gives (4.5). Using (5.8) and (4.22), the lower estimate in (4.7) follows with

$$q = \frac{9}{10} < 1 \quad \text{for} \quad j > j_0.$$

The estimates (4.6) and the upper estimate in (4.7) are a consequence of Assumption 2.1; i.e., the weight function before the mass matrix does not vary too much on two neighboring stripes. Using Theorem 4.2, the assertion follows. ∎

We now define an overlapping preconditioner of the type (3.6), (3.9) for the matrix

$$A_\lambda = \lambda M_{\omega_i} + T_{n_0}, \quad \lambda > 0. \tag{5.12}$$

First, we have to introduce some auxiliary matrices. For $j = 0, \dots, k-2$, let

$$M_{i,j} = \begin{bmatrix} \mathbf{0}_{n_{j+2}+1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \varepsilon_{i,j} I_{n_j - n_{j+2}-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{n_0 - n_j} \end{bmatrix} \in \mathbb{R}^{n_0 \times n_0},\ i = 1, 2,$$

$$T_{j,n_0} = \begin{bmatrix} \mathbf{0}_{n_{j+2}+1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & T_{n_j - n_{j+2}-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0}_{n_0 - n_j} \end{bmatrix} \in \mathbb{R}^{n_0 \times n_0},$$

where $\varepsilon_{i,j}$ is defined via (4.8). For $j = k-1$, we set

$$M_{i,k-1} = \begin{bmatrix} \varepsilon_{i,k-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_0-1} \end{bmatrix} \in \mathbb{R}^{n_0 \times n_0},\ i = 1, 2, \quad T_{j,n_0} = \begin{bmatrix} 2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0}_{n_0-1} \end{bmatrix} \in \mathbb{R}^{n_0 \times n_0}.$$

Now, we can define

$$C_1^{-1} = \sum_{j=0}^{k-1} (\lambda M_{i,j} + T_{j,n-1})^+ \tag{5.13}$$

as a preconditioner for A_λ. Now, we are able to formulate the final lemma.

LEMMA 5.2. *For $i = 1, 2$ and $\lambda > 0$, let A_λ and C_1 be defined via* (5.12) *and* (5.13)*, respectively. Then $c_1 C_1 \le A_\lambda \le c_2 C_1$. The constants do not depend on the structure of the weight functions ω_i or the parameter λ.*

Proof. We apply Lemma 4.1 with the bilinear form $(\cdot,\cdot)_{\mathcal{A}} = a_1(\cdot,\cdot)$ and verify the assumptions (4.1), (4.2), and (4.3). The space splitting implies $\beta = 2$ (cf. Theorem 4.2), which proves (4.2). Relation (4.1) follows from Lemma 5.1.

The bilinear form $a_1(\cdot,\cdot)$ (5.4) is the sum of two terms—a stiffness term and a mass term. The coefficient before the stiffness term is constant. The coefficient before the mass term is piecewise constant, i.e., $\varepsilon_{i,j}$ on Ω_j, $i = 1, 2$. Hence, the maximum of the coefficients on $\overline{\Omega}_j \cup \overline{\Omega}_{j+1}$ is $\varepsilon_{i,j}$, and the minimum is $\varepsilon_{i,j+1}$. In the preconditioner C_1 (5.13), the coefficient on $\overline{\Omega}_j \cup \overline{\Omega}_{j+1}$ is replaced by $\varepsilon_{i,j}$. Assumption 2.1 implies that the ratio of coefficients $\varepsilon_{i,j+1}^{-1}\varepsilon_{i,j}$ is bounded. This gives (4.3) and proves the lemma for the matrix C_1. ∎

5.2. The proof of Theorem 3.1. Now, we prove Theorem 3.1.

Proof. Due to Lemma 4.3, it suffices to show the result for the matrix $K_{k,p}$ (4.11). A simple computation shows that

$$K_{k,p} = T_{n_0} \otimes M_{\omega_2} + I_{n_0} \otimes T_{n_0},$$

where the matrices T_n and M_{ω_2} are defined via (5.1) and (5.2). Since the matrix T_{n_0} is symmetric and positive definite, we have

$$T_{n_0} = Q^T \Lambda Q \quad \text{with} \quad Q^T Q = I_{n_0},\ \Lambda = \operatorname{diag}[\lambda_i]_i, \quad \lambda_i > 0.$$

Hence,

$$\begin{aligned} K_{k,p} &= (Q^T \otimes I_{n_0})(\Lambda \otimes M_{\omega_2} + I_{n_0} \otimes T_{n_0})(Q \otimes I_{n_0}) \\ &= (Q^T \otimes I_{n_0})\operatorname{blockdiag}\left[\lambda_i M_{\omega_2} + T_{n_0}\right]_i (Q \otimes I_{n_0}). \end{aligned}$$

We now apply Lemma 5.2 and obtain

$$\begin{aligned}
K_{k,p}^{-1} &= (Q^T \otimes I_{n_0})\text{blockdiag}\left[(\lambda_i M_{\omega_2} + T_{n_0})^{-1}\right]_i (Q \otimes I_{n_0}) \\
&\sim (Q^T \otimes I_{n_0})\text{blockdiag}\left[\sum_{j=0}^{k-1}(\lambda_i M_j + T_{j,n_0})^+\right]_i (Q \otimes I_{n_0}) \\
&= (Q^T \otimes I_{n_0})\sum_{j=0}^{k-1}(\Lambda \otimes M_j + I_{n_0} \otimes T_{j,n_0})^+ (Q \otimes I_{n_0}) \\
&= \sum_{j=0}^{k-1}\left((Q^T \otimes I_{n_0})(\Lambda \otimes M_j + I_{n_0} \otimes T_{j,n_0})(Q \otimes I_{n_0})\right)^+ \\
&= \sum_{j=0}^{k-1}(T_{n_0} \otimes M_j + I_{n_0} \otimes T_{j,n_0})^+ = C_{mod}^{-1},
\end{aligned}$$

which proves the result. ∎

5.3. The proof of Theorem 3.2. The proof of Theorem 3.2 requires the following auxiliary result about block matrices.

LEMMA 5.3. *Let* $A = \left[\begin{smallmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{smallmatrix}\right]$ *be a symmetric positive definite matrix and* $C_0 = \left[\begin{smallmatrix} A_{11}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{smallmatrix}\right]$. *Then we have* $C_0 \le A^{-1}$.

Proof. Since A is positive definite, A is nonsingular. The inverse of A can be expressed by

$$A^{-1} = \begin{bmatrix} A_{11}^{-1} + A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix},$$

where $S = A_{22} - A_{21}A_{11}^{-1}A_{12}$ denotes the Schur complement. The matrix

$$A^{-1} - C_0 = \begin{bmatrix} A_{11}^{-1}A_{12}S^{-1}A_{21}A_{11}^{-1} & -A_{11}^{-1}A_{12}S^{-1} \\ -S^{-1}A_{21}A_{11}^{-1} & S^{-1} \end{bmatrix} = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

is positive semidefinite since $B_{22} = S^{-1}$ is positive definite and $B_{11} - B_{12}B_{22}^{-1}B_{21} = \mathbf{0}$. This proves the lemma. ∎

Now, we are able to prove Theorem 3.2.

Proof. We start with the proof of the lower inequality. We apply Lemma 4.1 for the bilinear forms $(\cdot,\cdot)_{\mathcal{A}} = a(\cdot,\cdot)$ and $(\mathcal{B}_i\cdot,\cdot) = a_i(\cdot,\cdot)$, $i = 0,\dots,k-1$; see (3.1). We verify the assumptions $(\mathcal{A}P_iu_i, P_iu_i) \le c_4\,(\mathcal{B}_iu_i, u_i)$ (see (4.3)) and $\sum_{i=0}^{k-1}(P_iu,u)_{\mathcal{A}} \le c_2\,(u,u)_{\mathcal{A}}$ (see (4.2)) of this lemma. The monotonicity of the weight function gives $c_4 = 1$ in (4.3). The space splitting into k subspaces implies $c_2 = k$ in (4.2). Using Lemma 4.1, we have $C^{-1} \le kK_k^{-1}$, which proves the lower estimate.

To prove the upper estimate, we note that $C_j = \left[\begin{smallmatrix} * & \\ * & C_{j,mod}^{*} \end{smallmatrix}\right] > 0$; cf. the definition of the matrices C_j and $C_{j,mod}$ in (3.2) and (3.5), respectively. Lemma 5.3 implies $\Theta_{j,mod}^{+} \le \Theta_j^{+}$. Using (3.4) and (3.6), we conclude that $C_{mod}^{-1} \le C^{-1}$, or, equivalently, $C \le C_{mod}$. By Theorem 3.1, we obtain $C \le C_{mod} \preceq K_k$. ∎

5.4. The proof of Theorem 3.3.

Proof. As in the previous case, we use the tensor product structure of the stiffness matrices K_k and $K_{k,p}$ (4.11). Due to Lemma 4.3, it suffices to prove $B_{mod}^{-1} \sim K_{k,p}$. A simple computation shows that

$$K_{k,p} = T_{n_0} \otimes M_{\omega_2} + M_{\omega_1} \otimes T_{n_0};$$

see (5.1) and (5.2) for the definition of the involved matrices. Since M_{ω_1} and T_{n_0} are symmetric and positive definite matrices, we can conclude

$$M_{\omega_1}^{-1/2} T_{n_0} M_{\omega_1}^{-1/2} = \tilde{Q}^T \tilde{\Lambda} \tilde{Q} \quad \text{with} \quad \tilde{Q}^T \tilde{Q} = I_{n_0},\ \tilde{\Lambda} = \operatorname{diag}[\tilde{\lambda}_i]_i, \quad \tilde{\lambda}_i > 0.$$

This gives

$$\begin{aligned} K_{k,p} &= (M_{\omega_1}^{1/2} \tilde{Q} \otimes I_{n_0})(\tilde{\Lambda} \otimes M_{\omega_2} + I_{n_0} \otimes T_{n_0})(\tilde{Q}^T M_{\omega_1}^{1/2} \otimes I_{n_0}) \\ &= (M_{\omega_1}^{1/2} \tilde{Q} \otimes I_{n_0}) \operatorname{blockdiag} \left[\tilde{\lambda}_i M_{\omega_2} + T_{n_0} \right]_i (\tilde{Q}^T M_{\omega_1}^{1/2} \otimes I_{n_0}). \end{aligned}$$

Using Lemma 5.2, we can conclude

$$(\tilde{\lambda}_i M_{\omega_2} + T_{n_0})^{-1} \sim \sum_{j=0}^{k-1} (\tilde{\lambda}_i M_{2,j} + T_{j,n_0})^+.$$

Hence, we can proceed with the estimates

$$\begin{aligned} K_{k,p}^{-1} &\sim (M_{\omega_1}^{-1/2} \tilde{Q} \otimes I_{n_0}) \operatorname{blockdiag} \left[\sum_{j=0}^{k-1} (\tilde{\lambda}_i M_{2,j} + T_{j,n_0})^+ \right]_i (\tilde{Q}^T M_{\omega_1}^{-1/2} \otimes I_{n_0}) \\ &= (M_{\omega_1}^{-1/2} \tilde{Q} \otimes I_{n_0}) \left[\sum_{j=0}^{k-1} \tilde{\Lambda} \otimes M_{2,j} + I_{n_0} \otimes T_{j,n_0} \right] (\tilde{Q}^T M_{\omega_1}^{-1/2} \otimes I_{n_0}) \\ &= \sum_{j=0}^{k-1} (T_{n_0} \otimes M_{2,j} + M_{\omega_1} \otimes T_{j,n_0})^+ := \sum_{j=0}^{k-1} C_{3,j}^+. \end{aligned} \tag{5.14}$$

In a second step, we derive a preconditioner for $C_{3,j}$. With the same tensor product arguments as above, we obtain

$$C_{3,j}^+ \sim \sum_{j'=0}^{k-1} (T_{j',n_0} \otimes M_{2,j} + M_{1,j'} \otimes T_{j,n_0})^+ \tag{5.15}$$

uniformly for all $j = 0, \ldots, k-1$. Combining (5.14) and (5.15), we have

$$K_{k,p}^{-1} \sim \sum_{j=0}^{k-1} \sum_{j'=0}^{k-1} (T_{j',n_0} \otimes M_{2,j} + M_{1,j'} \otimes T_{j,n_0})^+ = B_{mod}^{-1},$$

which proves the result. ∎

6. Numerical experiments. In this section, we present some numerical experiments.

6.1. The case $\omega_1(\xi) = 1$. In a first experiment, we investigate the preconditioner C (3.4). Figure 6.1 displays the maximal and minimal eigenvalues of the matrix $C^{-1}K_{k,p}$ for different weight functions. The minimal eigenvalue of the matrix $C^{-1}K_{k,p}$ is bounded from below by a positive constant for all types of investigated weight functions. The constants are very close to 1. The maximal eigenvalue is about k on level k.

In a second experiment, we investigate the preconditioner C_{mod} (3.6). In comparison to the first preconditioner, this preconditioner is optimal. Figure 6.2 displays

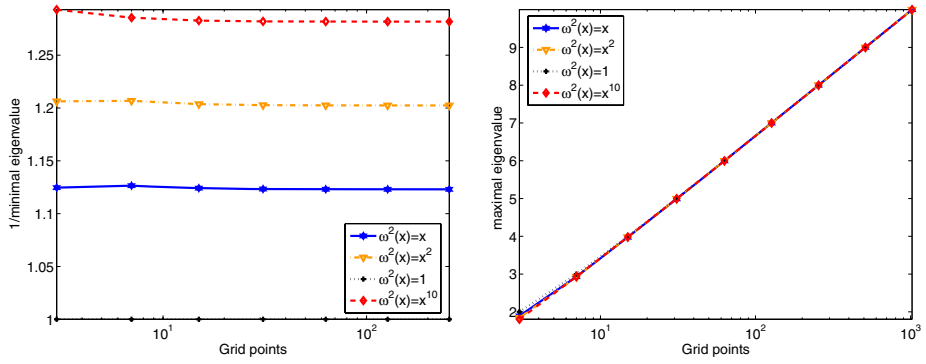


FIG. 6.1. *Eigenvalue bounds with the preconditioner* (3.4): *minimal eigenvalue (left), maximal eigenvalue (right).*

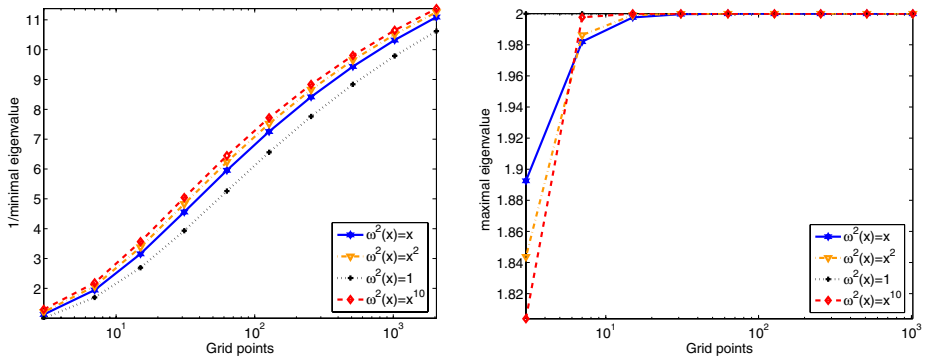


FIG. 6.2. *Eigenvalue bounds with the modified preconditioner* (3.6): *minimal eigenvalue (left), maximal eigenvalue (right).*

the maximal and minimal eigenvalues for the matrix $C_{mod}^{-1}K_{k,p}$ with the modified preconditioner (3.6) for different weight functions. The minimal eigenvalue of the matrix $C_{mod}^{-1}K_{k,p}$ is bounded from below by a positive constant for all types of investigated weight functions. The maximal eigenvalue is bounded from above by a constant of 2. The asymptotically optimal behavior can be seen only for relatively high level numbers. Thus, the condition number of $C^{-1}K_{k,p}$ is lower for $k \leq 10$ than the condition number of $C_{mod}^{-1}K_{k,p}$, although the condition number of $C_{mod}^{-1}K_{k,p}$ grows logarithmically, whereas the condition number of $C^{-1}K_{k,p}$ is bounded.

Finally, we investigated the preconditioners for the matrix K_k. The results for the minimal eigenvalue of the matrices $C^{-1}K_k$ and $C_{mod}^{-1}K_k$ are displayed in Figure 6.3, left and right panels, respectively. For the maximal eigenvalues, the results are the same as for the matrix K_k; i.e., the maximal eigenvalue of the matrix $C^{-1}K_k$ is about k, whereas the maximal eigenvalue of the matrix $C_{mod}^{-1}K_k$ is about 2.

Considering the minimal eigenvalue, the results are different. The results for the matrix $C_{mod}^{-1}K_k$ are comparable with the results for the matrix $C_{mod}^{-1}K_{k,p}$ if the weight function is not $\omega_2^2(\xi) = \xi^{10}$. Then an additional factor of about 2.5 can be seen. The minimal eigenvalue $C^{-1}K_k$ has the expected (pessimistic) additional factor of $2 \cdot 2^{\alpha}$ of Lemma 4.3 compared with the minimal eigenvalue $C^{-1}K_{k,p}$. Summarizing, the preconditioner C_{mod} (3.6) should be preferred for the matrix K_k with a weight function $\omega_2^2(\xi) = \xi^{\alpha}$, $\alpha > 1$.

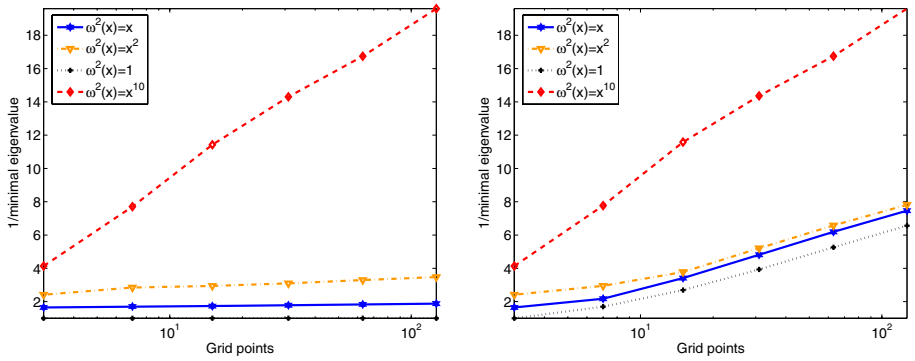


FIG. 6.3. *Minimal eigenvalue of the matrix $C^{-1}K_k$ (left) and minimal eigenvalue of the matrix $C_{mod}^{-1}K_k$ (right).*

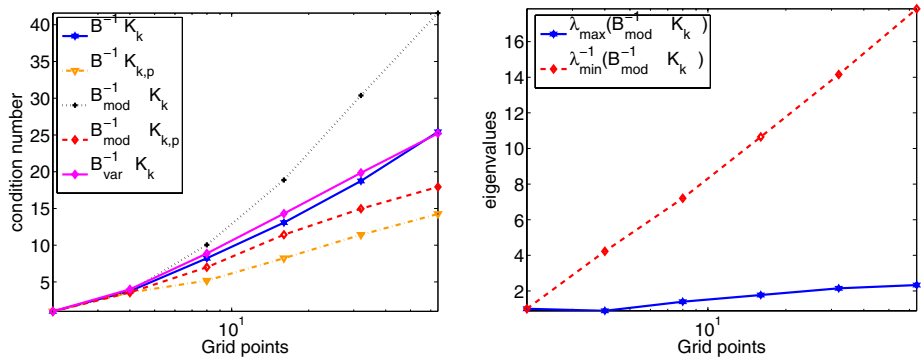


FIG. 6.4. *Condition numbers for the general case (left), eigenvalue bounds for $B_{mod}^{-1}K_k$ (right).*

6.2. The general case. In this subsection, we consider the systems

$$K_k\underline{u} = \underline{f} \quad \text{and} \quad K_{k,p}\underline{u} = \underline{f}.$$

In all experiments, we choose the weight functions $\omega_i^2(\xi) = \xi^2$, $i = 1, 2$. Figure 6.4 displays the condition number of $\mathcal{C}^{-1}K$ for five combinations of $\mathcal{C} = \{B, B_{mod}, B_{var}\}$ and $K = \{K_k, K_{k,p}\}$. The best results are obtained for the matrix $K_{k,p}$ with a piecewise constant coefficient function. Then the condition number is moderately increasing for both preconditioners. For the matrix K_k with the smooth coefficient function, the results are not as good as in the previous case if we take a preconditioner with constant coefficients. The preconditioner with variable but smooth coefficients B_{var} of Remark 3.5 behaves better.

In particular for $B_{mod}^{-1}K_k$ (3.9), an additional factor of about 4 can be seen, which arises from the estimates of Lemma 4.3. The eigenvalue bounds of $B_{mod}^{-1}K_k$ and $B_{var}^{-1}K_k$ are similar to those for the preconditioner C_{mod} (3.6) (cf. Figures 6.3 and 6.2), where the asymptotics can be seen only for relatively high level numbers. Hence, the nonoptimal preconditioner B (3.7) should be preferred for moderate level numbers of $k = 5, 6, 7$.

Finally, we investigate the iterations of the PCG method with the preconditioner B (3.7). In all experiments, we have taken a randomly chosen right-hand side and a

TABLE 6.1
PCG iterations for the systems $K_k\underline{u} = \underline{f}$ and $K_{k,p}\underline{u} = \underline{f}$ with the preconditioner B (3.7).

Level	3	4	5	6	7	8	9	10
K_k	16	21	25	29	33	37	40	43
$K_{k,p}$	16	20	25	30	34	38	41	43

relative accuracy of 10^{-5}. The results for both systems are displayed in Table 6.1. From the results, a slight increase of the iteration numbers can be seen. Since this preconditioner is not optimal (cf. Theorem 3.4), the growth of the PCG numbers is not surprising. The PCG iterations are about the same for the matrix K_k with continuous weight function and the matrix $K_{k,p}$ with piecewise constant weight function.

7. Concluding remarks. We will conclude the paper with a remark about an application for the p-version of the FEM in three dimensions. Using the basis of the integrated Legendre polynomials $\{\hat{L}_i\}_{i=2}^{p}$, it has been proved in [3] that the element stiffness matrix for odd polynomial degree p with respect to the interior bubbles is spectrally equivalent to the matrix

$$\begin{aligned} K_{pv} &= P^T \text{blockdiag}\,[T_n \otimes T_n \otimes M_\omega + T_n \otimes M_\omega \otimes T_n + M_\omega \otimes T_n \otimes T_n]_{i=1}^{8}\, P \\ &=: P^T \text{blockdiag}\,[K_3]_{i=1}^{8}\, P, \end{aligned}$$

where T_n is the matrix (5.1) of dimension $\frac{p-1}{2}$, M_ω is the matrix (5.2) with the weight function $\omega^2(\xi) = \xi^2$, and P is a permutation matrix. In [5], an optimal solver for K_{pv} based on wavelets has been derived.

Another preconditioner C_3 for K_3 can be developed in the same way as for K_k in (3.9) and (3.7). With the same tensor product arguments as in the proof of Theorem 3.3 presented in subsection 5.4, the optimality of the estimate $C_3 \sim K_{3,k}$ is proved. Using a fast Fourier transform for the remaining problem, we obtain a second fast solver for the block of the interior bubbles in the p-version of the FEM using hexahedral elements.

Acknowledgments. The work reported in this paper was initiated during the Special Semester on Computational Mechanics in Linz 2005. The second author thanks the RICAM for the hospitality during his stay in Linz.

REFERENCES

[1] S. BEUCHLER, *Multigrid solver for the inner problem in domain decomposition methods for p-FEM*, SIAM J. Numer. Anal., 40 (2002), pp. 928–944.
[2] S. BEUCHLER, *Multilevel solvers for a finite element discretization of a degenerate problem*, SIAM J. Numer. Anal., 42 (2004), pp. 1342–1356.
[3] S. BEUCHLER AND D. BRAESS, *Improvements for some condition number estimates in p-fem*, Numer. Linear Algebra Appl., 13 (2006), pp. 573–588.
[4] S. BEUCHLER AND S. NEPOMNYASCHIKH, *Overlapping additive Schwarz preconditioners for degenerated elliptic problems: Part* I. *Isotropic problems*, J. Numer. Math., to appear.
[5] S. BEUCHLER, R. SCHNEIDER, AND C. SCHWAB, *Multiresolution weighted norm equivalences and applications*, Numer. Math., 98 (2004), pp. 67–97.
[6] S. BÖRM AND R. HIPTMAIR, *Analysis of tensor product multigrid*, Numer. Algorithms, 26 (2001), pp. 219–234.
[7] J. BRAMBLE, J. PASCIAK, AND A. SCHATZ, *The construction of preconditioners for elliptic problems by substructuring* I, Math. Comp., 47 (1986), pp. 103–134.
[8] J. BRAMBLE AND X. ZHANG, *Uniform convergence of the multigrid V-cycle for an anisotropic problem*, Math. Comp., 70 (2001), pp. 453–470.

[9] I. G. Graham, P. Lechner, and R. Scheichl, *Domain decomposition for multiscale PDEs*, Numer. Math., 106 (2007), pp. 589–626.
[10] G. Haase, U. Langer, and A. Meyer, *The approximate Dirichlet domain decomposition method. Part* I: *An algebraic approach*, Computing, 47 (1991), pp. 137–151.
[11] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Heidelberg, 1985.
[12] V. G. Korneev, *An almost optimal method for Dirichlet problems on decomposition subdomains of the hierarchical hp-version*, Differ. Uravn., 37 (2001), pp. 1008–1018 (in Russian).
[13] A. Kufner and A.M. Sändig, *Some applications of weighted Sobolev spaces*, B.G. Teubner Verlagsgesellschaft, Leipzig, Germany, 1987.
[14] P.-L. Lions, *On the Schwarz alternating method* I, in Proceedings of the 1st Annual International Symposium on Domain Decomposition Methods for Partial Differential Equations, R. Glowinski, G. H. Golub, G. A. Meurant, and J. Périaux, eds., SIAM, Philadelphia, 1988, pp. 1–42.
[15] A. M. Matsokin and S. V. Nepomnyaschikh, *The Schwarz alternation method in a subspace*, Iz. VUZ Mat., 29 (1985), pp. 61–66.
[16] S. V. Nepomnyaschikh, *Method of splitting into subspaces for solving elliptic boundary value problems in complex-form domains*, Soviet J. Numer. Anal. Math. Modelling, 6 (1991), pp. 151–168.
[17] S. V. Nepomnyaschikh, *Preconditioning operators for elliptic problems with bad parameters*, in Proceedings of the 11th International Conference on Domain Decomposition Methods (London, 1998), DDM.org, Augsburg, 1999, pp. 82–88.
[18] O. Pironneau and F. Hecht, *Mesh adaption for the Black and Scholes equations*, East-West J. Numer. Math., 8 (2000), pp. 25–36.
[19] R. Scheichl and E. Vainikko, *Additive Schwarz and aggregation-based coarsening for elliptic problems with highly variable coefficients*, Computing, to appear.
[20] A. Toselli and O. Widlund, *Domain Decomposition Methods—Algorithms and Theory*, Springer-Verlag, Berlin, 2005.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2345–2367

SPECTRAL ANALYSIS OF NONSYMMETRIC QUASI-TOEPLITZ MATRICES WITH APPLICATIONS TO PRECONDITIONED MULTISTEP FORMULAS*

DANIELE BERTACCINI† AND FABIO DI BENEDETTO‡

Abstract. The eigenvalue spectrum of a class of nonsymmetric preconditioned matrices arising in time-dependent partial differential equations is analyzed and discussed. The matrices generated by the underlying numerical integrators are small rank perturbations of block Toeplitz matrices; circulant-like preconditioners based on the former are considered. The eigenvalue distribution of the preconditioned matrix influences often crucially the convergence of Krylov iterative accelerators. Due to several reasons (lack of symmetry, band structure, and coefficients depending on the size) the classical approach based on smooth generating functions gives very little insight here. Therefore, to characterize the eigenvalues, a *difference equation* approach exploiting the band Toeplitz and circulant patterns generalizing the well-known results of Trench is proposed.

Key words. circulant preconditioners, nonsymmetric Toeplitz matrices, eigenvalues, difference equations, linear systems of time-step integrators, linear multistep formulas, boundary value problems

AMS subject classifications. 15A18, 65F10, 65F15, 65N22, 65Q05

DOI. 10.1137/060650349

1. Introduction.

In this paper we focus on small rank perturbations of block nonsymmetric Toeplitz matrices preconditioned by circulant approximations introduced in [3, 4, 7].

An $n \times n$ matrix $A_n = (a_{j,k})$ is said to be *Toeplitz* if $a_{j,k} = a_{j-k}$, $j, k = 1, \ldots, n$, i.e., A_n is constant along its diagonals, and quasi-Toeplitz if it is a small rank perturbation of a Toeplitz matrix. $\breve{A}_n$ is *circulant* if it is Toeplitz and its diagonals satisfy $\breve{a}_{n-j} = \breve{a}_{-j}$. The circulant matrices $\breve{A}_n$ are diagonalized by the Fourier matrix $F = (F_{j,k})$, $F_{j,k} = e^{2\pi \mathbf{i} jk/n}/\sqrt{n}$, $j, k = 0, \ldots, n-1$, and $\mathbf{i}$ is the imaginary unit; see [19]. Circulant matrices are easily and efficiently invertible using the fast Fourier transform (FFT), as in [16].

Perturbations of block nonsymmetric Toeplitz matrices arise in the numerical approximation of time-dependent partial differential equations (PDEs) by generalizations of implicit multistep formulas used in boundary value form [20, 1, 14]. The techniques considered here could be adapted to other discretization schemes based on finite differences for PDEs.

Other circulant-like matrices used in the PDE context can be found in [8].

As explained in section 2.1, the matrices of the underlying linear systems can be written as follows:

$$M = A \otimes I - h\, B \otimes J, \tag{1.1}$$

*Received by the editors January 18, 2006; accepted for publication (in revised form) July 3, 2007; published electronically November 9, 2007. This work was partially supported by MIUR, grants 2002014121, 2004015437, 2006017542, and by grant 8111317 from Università di Roma "La Sapienza."

http://www.siam.org/journals/sinum/45-6/65034.html

†Dipartimento di Matematica, Università di Roma "La Sapienza," P. le A. Moro 2, Roma, Italy. Current address: Università di Roma "Tor Vergata," Dipartimento di Matematica, viale della Ricerca Scientifica 1, 00133, Roma, Italy (bertaccini@mat.uniroma2.it).

‡Dipartimento di Matematica, Università di Genova, via Dodecaneso 35, 16146 Genova, Italy (dibenede@dima.unige.it).

where A and B are $n \times n$ small rank perturbations of band Toeplitz matrices whose entries are given by the coefficients of the scheme involved, I is the identity, and J is an $m \times m$ matrix which can be large and sparse. More precisely, J is the Jacobian matrix of a system of equations discretized in space by finite differences; see [4] for details.

Unfortunately, when m and/or n are (even moderately) large, iterative solvers for (1.1), used without preconditioners or with general purpose preconditioners, such as those based on incomplete factorizations, often converge very slowly or not at all; see [4, section 5]. In general, direct methods cannot exploit the block structure of (1.1).

Preconditioners introduced in [4] take into account this structure. They are block-circulant and, in compact form, can be written as

$$P = \breve{A} \otimes I - h\, \breve{B} \otimes \tilde{J}, \tag{1.2}$$

where $\breve{A}$ and $\breve{B}$ are circulant-like approximations for A, B, respectively, and $\tilde{J}$ is a suitable approximation for J. Their performance has been tested in several papers [4, 5, 17].

The distribution of the eigenvalues of the matrices M and $P^{-1}M$ can influence the convergence of iterations of Krylov subspace methods. This is the case, e.g., if the condition number of the eigenvector matrix is moderate; see [22].

Tables of the condition number $\kappa_2(X)$ of the eigenvector matrix X for the (left) preconditioned matrix $P^{-1}M$ and related discussions can be found in [9], showing that eigenvalues can give reasonable information in our setting. Similar conclusions hold true for the nonpreconditioned case for most of the methods considered here. More details are reported in section 6.

A theoretical investigation of the eigenvalues is hard because, in general, P and M are nonsymmetric and nonsymmetrizable. Moreover, as explained in section 3 an analysis of the eigenvalues based on the generating function of the underlying Toeplitz matrices is not feasible here, although very meaningful for Hermitian matrices [16, 23].

These difficulties motivate us to a "direct" analysis, based on the generalized eigenvalue problem

$$M\, u = \lambda\, P\, u.$$

The tools used here are completely different from those in previous works such as [10], [7], or [25]. In particular, we cannot write $\breve{A}$, $\breve{B}$ as small rank perturbations of A, B.

By using instead linear difference equation theory and generalizing Trench's approach [33, 34], we derive closed formulas and first-order expansions for λ as a function of the time step h and of the eigenvalues of the Jacobian matrix J. This characterization involves the roots of a sparse polynomial whose degree is related to the size of A and B.

Our estimates are explicitly computed for some well-known 2-step integrators and compared with the "true" eigenvalues approximated by Matlab. The approach seems very useful for spectrum localization and is not too expensive provided that A and B have moderate size or an efficient rootfinder is associated with our technique.

The paper is organized as follows. Section 2 introduces the problem and the main circulant preconditioning techniques. In section 3 we discuss the relevant literature for spectral analysis, and we explain in more detail the motivation of our work. Section 4 is devoted to the spectral analysis, from the general case to the 2-step case study. In section 5 we describe two classical PDE examples, representing the test problems for our experiments of section 6.

2. Preliminaries. Let us consider a model problem based on a first-order initial-boundary value time-dependent partial differential equation

$$\begin{cases} \dfrac{\partial u}{\partial t} & = \mathcal{L}(u) + f, \quad x \in \mathbf{D}, \\ \mathcal{G}(u) & = g, \qquad\quad x \in \partial\mathbf{D}, \\ u & = u_0, \qquad\quad t = t_0, x \in \mathbf{D}, \end{cases} \tag{2.1}$$

where $\mathbf{D}$ is an open domain in $\mathbb{R}^N$, $N \geq 1$, and $\mathcal{L}$ is a differential operator, nonlinear in u in general. Equations (2.1) are *evolutionary* because they describe evolving phenomena and combine differentiation with respect to both space and time. For simplicity, we will focus on linear operators $\mathcal{L}$ and $\mathcal{G}$. However, most of the techniques considered here can be applied to a more general nonlinear framework by recalling that often numerical codes linearize the nonlinear algebraic equations by using a quasi-Newton step; see [21].

2.1. Linear multistep formulas in boundary value form. In the following, a brief description of a generalization of linear multistep formulas is given.

If the partial differential equation (2.1) is first discretized in space, we obtain a system of ordinary differential equations (ODEs). Such a system can be very large and is treated by means of a numerical method for ODEs.

Here we focus on linear multistep formulas applied in boundary value form (see [1, 14]), which generalize classical implicit linear multistep formulas by using both initial and boundary conditions even in the presence of an initial value problem. Such schemes have a relatively long history (see, e.g., [20, 1]) and can be very useful in some communities where "time" has no special orientation (see an example of these problems in the work by Shirley referred to in [26]).

More precisely, we suppose that (2.1), with solution $u(x,t)$, has been discretized in space on a certain grid Ω_τ, with mesh width $\tau > 0$, to yield a *semidiscrete* system

$$y'(t) = F(t, y(t)), \quad t_0 \leq t \leq t_s, \quad y(t_0) \text{ given}, \tag{2.2}$$

with $y(t) = (u_j(t))_{j=1}^m$, m being related to the number of grid points in space, and, for unidimensional spatial domain $\mathbf{D}$, i.e., $N = 1$ in (2.1), $u_j(t)$ approximates $u(x_j, t)$ at some x_j, $j = 1, \ldots, m$. The contribution of the discretized boundary conditions is enclosed in F. In order to approximate $u(x,t)$ on Ω_τ for $t = t_0, t_1, \ldots, t_s$, an appropriate temporal mesh, we apply an ODE method with step size $h > 0$.

Using the shortened notation $F_{n+i} = F(t_{n+i}, y_{n+i})$, $i = 0, \ldots, k$, if y_n approximates $y(t_n)$, linear multistep formulas in boundary value form are given by

$$\sum_{i=0}^{k} \alpha_i y_{n+i} = h \sum_{i=0}^{k} \beta_i F_{n+i}, \quad n = 0, \ldots, s-k, \tag{2.3}$$

where $y_0 = y(t_0)$ is provided by initial conditions of (2.1), while $y_1, \ldots, y_{\nu-1}$ and $y_{s-k+\nu+1}, \ldots, y_s$, computed at the mesh points $t_0, \ldots, t_{\nu-1}$, $t_{s-k+\nu+1}, \ldots, t_s$, are determined by using other difference formulas, usually of the same order of (2.3). In practical use, we couple three sets of formulas: $\nu - 1$ for $y_1, \ldots, y_{\nu-1}$, $s - k + 1$ with the coefficients as (2.3) and $k - \nu$ for $y_{s-k+\nu+1}, \ldots, y_s$. We note that the formulas of the first and third sets are still based on linear multistep finite differences expressions as (2.3), but each one has different coefficients (and is independent from those in (2.3)), while all formulas in the second set, based on (2.3), share the same coefficients $\alpha_0, \ldots, \alpha_k$, $\beta_0, \ldots, \beta_k$.

As announced, we simplify the notation; i.e., we suppose F linear, $F(t, y(t)) = J\,y(t) + g(t)$, where $J \in \mathbb{R}^{m\times m}$ and g is a vector-valued function of t. The underlying discrete boundary value problem can be solved by forming the following linear system:

$$\begin{aligned} &Mz = b, \quad M = A \otimes I_m - B \otimes (h\,J), \quad z^T = \left(y_0^T, y_1^T, \ldots, y_s^T\right), \\ &b = e_1 \otimes y_0 + h(B \otimes I_m)g, \quad g^T = (g(t_0)^T \ldots g(t_s)^T), \end{aligned} \tag{2.4}$$

where A, B are $(s+1)\times(s+1)$ real-valued quasi-Toeplitz nonsymmetric matrices and e_1 is the first column of the identity matrix. In practice, we accommodate the coefficients α_j, $\alpha_j^{(r)}$, $\quad j = 0, \ldots, k, \quad r = 1, \ldots, \nu-1, s-k+\nu+1, \ldots, s$ in A and β_j, $\beta_j^{(r)}$, $\quad j = 0, \ldots, k, \quad r = 1, \ldots, \nu-1, s-k+\nu+1, \ldots, s$ in B such that we can look at A as $\hat{A} + R_A$ and B as $\hat{B} + R_B$, $\hat{A}$ and $\hat{B}$ Toeplitz matrices with stencil

$$\left(0 \ldots 0\ \alpha_0 \ldots \underline{\alpha_\nu} \ldots \alpha_k 0 \ldots 0\right)$$

and

$$\left(0 \ldots 0\ \beta_0 \ldots \underline{\beta_\nu} \ldots \beta_k 0 \ldots 0\right),$$

respectively. The underlined element is the one on the main diagonal, and R_A and R_B have nonzero elements at most in their $\nu \times (k+1)$ upper left and $(k-\nu)\times(k+1)$ lower right corners.

The additional work needed for the solution of the discrete problems (2.4) with respect to those for the solution of implicit standard linear multistep formulas (i.e., used with only initial values) is justified by better stability and order properties; see [14, 4] for details and discussions.

More on the matrices A, B, and M generated by the schemes above can be found in [7, 6]. Examples of matrices A, B, and M for 2-step formulas will be given in what follows.

2.2. A review of block-circulant preconditioners. We noted in [4] that, in d dimensions, $d > 1$, when a fine enough spatial discretization is used in (2.1), direct methods are often not feasible to solve linear systems (2.4). Iterative methods are mandatory when the discrete problem is generated by a three-dimensional or even two-dimensional differential model (2.1). In [3, 4] Krylov subspace methods were proposed to solve (2.4). However, without preconditioning, the convergence can be very slow or iterations do not converge at all. Therefore, in [3, 4] a preconditioning strategy based on circulant matrices was introduced (see also [16]). Thus, other approximations were introduced in [5, 8]; see [8] for a more comprehensive bibliography. By left preconditioning we mean solving the equivalent nonsymmetric linear system

$$P^{-1}Mx = P^{-1}b \tag{2.5}$$

instead of $Mx = b$. Right preconditioning is obtained by considering

$$MP^{-1}y = b, \quad x = P^{-1}y.$$

Note that matrices MP^{-1} and $P^{-1}M$ are similar and hence share the same eigenvalues. Since we are interested in the eigenvalues of (2.5), our analysis is based entirely on left preconditioning.

In what follows, some block-circulant and block-circulant-like preconditioners for (2.4) are briefly reviewed.

Let us consider the following approximation of the matrix M:

$$P = \breve{A} \otimes I_m - h\breve{B} \otimes \tilde{J}, \tag{2.6}$$

where $\tilde{J}$ is a suitable approximation of the Jacobian matrix or the Jacobian itself. $\breve{A}$, $\breve{B}$ are circulant matrices whose entries are derived from the coefficients of the main method (2.3) as follows:

$$\begin{aligned} &\breve{A} = circ(\breve{a}),\ \breve{a}_j = c_{j,1}(s)\alpha_{j+\nu} + c_{j,2}(s)\alpha_{j+\nu-(s+1)}, \\ &\breve{B} = circ(\breve{b}),\ \breve{b}_j = c_{j,3}(s)\beta_{j+\nu} + c_{j,4}(s)\beta_{j+\nu-(s+1)}, j = 0, \ldots, s, \end{aligned} \tag{2.7}$$

where $circ(\cdot)$ denotes the circulant matrix having the first column specified in the argument, and the $c_{j,i}(s)$, $i = 1, \ldots, 4$, $j = 0, \ldots, s$, are linear in j. It is understood that α_j (β_j) is zero for $j < 0$ or $j > k$ in (2.7), so that the sparsity of A, B implies that of $\breve{A}$, $\breve{B}$. The coefficients $c_{i,j}(s)$ in (2.7) are chosen in such a way that $\breve{A}$, $\breve{B}$ are suitable approximations of A, B in (2.4), respectively.

The approximation of A, B with T. Chan's *optimal circulant preconditioner* (see [18]) requires that

$$c_{j,1}(s) = c_{j,3}(s) = 1 - \frac{j}{s+1},\ \text{and}\ c_{j,2}(s) = c_{j,4}(s) = \frac{j}{s+1}, j = 0, \ldots, s, \tag{2.8}$$

while for Strang's *natural (or simple) circulant preconditioner* (see [29])

$$\begin{aligned} &c_{j,1}(s) = c_{j,3}(s) = 1, \quad j = 0, \ldots, \left\lfloor \frac{s+1}{2} \right\rfloor, \\ &c_{j,2}(s) = c_{j,4}(s) = 1, \quad j = \left\lfloor \frac{s+1}{2} \right\rfloor + 1, \ldots, s, \quad c_{j,i}(s) = 0 \text{ otherwise.} \end{aligned}$$

On the other hand, if we consider, instead of (2.8), the following definition of the coefficients $c_{j,i}(s)$ for $\breve{A}$ and $\breve{B}$:

$$c_{j,1}(s) = c_{j,3}(s) = 1 + \frac{j}{s+1}, \quad c_{j,2}(s) = c_{j,4}(s) = \frac{j}{s+1}, \quad j = 0, \ldots, s, \tag{2.9}$$

we get the so-called P-circulant approximations which, used in (2.6), gives the P-circulant (block) preconditioner, introduced in [3, 4]. The latter definition avoids singularity problems which are sometimes typical of the former choices.

In [3, 4] and in [17] it was shown that both the P-circulant and generalized Strang preconditioned systems can be effective to accelerate the convergence. Unfortunately, when the Jacobian matrix J has some small (or zero) eigenvalues, the simple circulant or Strang preconditioner can be severely ill-conditioned or even singular (see [3, 4, 5]). An analysis of the spectrum for the preconditioned matrix based on simple circulant approximations can be found in [10]. However, we stress that the tools used here are completely different from those in the former. In particular, we cannot write anymore $\breve{A}$, $\breve{B}$ as small rank perturbations of A, B, respectively.

Therefore, we will focus on preconditioners (2.6) based on T. Chan's and the P-circulant approximations in the following discussions. Practical examples for the matrices A, B, $\breve{A}$, $\breve{B}$, M, and P can be found below.

Another approximation which was found effective (but is not considered here) is based on $\{\omega\}$-circulant approximations for matrices A and B in (2.4); see [8]. In

particular, $\breve{A}$, $\breve{B}$ are $\{\omega\}$-circulant matrices approximating A and B, respectively. The $\{\omega\}$-circulant matrices are Toeplitz matrices whose first entry of a row is given by multiplying the last entry of the preceding row by $\omega = \exp(\mathbf{i}\theta)$; see [19] for more details. Notice that the $\{1\}$-circulant matrices ($\theta = 0$) are just circulant matrices (and therefore generate simple or Strang's approximations for a given Toeplitz matrix), while $\{-1\}$-circulant matrices ($\theta = \pi$) are skew-circulant matrices.

We observe that various trigonometric approximations can be combined. For example, $\{\omega\}$-P-circulant preconditioners can be defined by using (2.9) to give the first row of the related $\{\omega\}$-circulant approximation. A similar combination can be made by using T. Chan's optimal circulant matrices. Moreover, it is straightforward to observe that P-circulant approximations can be seen as $\{\omega\}$-circulant preconditioners with $\theta = 0$, whose entries are defined as in (2.9). More comments on these generalizations can be found in [8].

2.3. Dahlquist's hypothesis. In the proposed eigenvalue analysis for the preconditioned linear systems (2.5), unless otherwise specified, we choose $J = \mu$ (i.e., a scalar) in (2.2), where $\mu \in \mathbb{C}^- := \{\lambda \in \mathbb{C} : \mathrm{Re}\lambda \leq 0\}$ ("Dahlquist's hypothesis"). It is customary to consider this scalar problem in the linear stability theory for ODEs. The parameter μ can be any eigenvalue of the Jacobian matrix J of the given PDE, supposed diagonalizable. Indeed, notice that, supposing J diagonalizable, we have

$$J = VDV^{-1}, \quad D = \mathrm{diag}(\mu_1, \ldots, \mu_m).$$

This framework is not restrictive for our analysis since

$$M = A \otimes I_m - hB \otimes \left(VDV^{-1}\right) = (I_{s+1} \otimes V)(A \otimes I_m - hB \otimes D)(I_{s+1} \otimes V^{-1}).$$

Assuming that the preconditioner is based on the exact Jacobian, a similar expression can be derived for P. Let us write

$$M(q) = A - qB, \quad P(q) = \breve{A} - q\breve{B}, \quad q = h\mu.$$

It is straightforward to observe that the eigenvalues of the preconditioned/transformed linear system (2.5) are given by the union of the eigenvalues of the finite sequence of matrices

$$\{P(q)^{-1}M(q)\}_q, \quad q = h\mu_i, \quad i = 1, \ldots, m.$$

Just to have an idea of the matrix structures, we sketch below the explicit expression of $M(q)$ and $P(q)$ for the particular example of a 2-step generalized Adams–Moulton method in connection to a P-circulant preconditioner, with an additional final condition given by the implicit Euler method:

$$M(q) = \begin{pmatrix} 1 & 0 & & & \\ -1 & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 1 & 0 \\ & & & -1 & 1 \end{pmatrix} - q \cdot \begin{pmatrix} \frac{2}{3} & -\frac{1}{12} & & & \\ \frac{5}{12} & \frac{2}{3} & -\frac{1}{12} & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{5}{12} & \frac{2}{3} & -\frac{1}{12} \\ & & & 0 & 1 \end{pmatrix},$$

$$P(q) = \begin{pmatrix} 1 & 0 & & & -\frac{s}{s+1} \\ -\frac{s}{s+1} & 1 & 0 & & \\ & \ddots & \ddots & \ddots & \\ & & -\frac{s}{s+1} & 1 & 0 \\ 0 & & & -\frac{s}{s+1} & 1 \end{pmatrix} - q \cdot \begin{pmatrix} \frac{2}{3} & \breve{b}_1 & & & \breve{b}_s \\ \breve{b}_s & \frac{2}{3} & \breve{b}_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \breve{b}_s & \frac{2}{3} & \breve{b}_1 \\ \breve{b}_1 & & & \breve{b}_s & \frac{2}{3} \end{pmatrix},$$

where $\breve{b}_1 = -\frac{s+2}{12(s+1)}, \breve{b}_s = \frac{5s}{12(s+1)}$.

We recall that the underlying generalizations of Adams–Moulton formulas, which should be used only as implicit methods with one initial (given) and one final condition, are (i) A-stable not only for $k = 2$ but for arbitrarily high-order $k+1$ and (ii) all formulas preserve important properties such as the time reversal symmetry and the Hamiltonian function; see [14]. Note that if we use the usual 2-step Adams–Moulton formula, we should supply another starting value y_1.

3. Motivation of the work. From now on, we will assume Dahlquist's hypothesis (see section 2.3) in order to simplify the theoretical analysis.

Understanding the behavior of iterative solvers for (2.4) requires the knowledge of the following features:

1. How does the spectrum of M depend on the discretization parameters? For instance, for which values of q (both involving the time step and the Jacobian of the PDE) can we ensure that the spectrum lies in $\mathbb{C}^+ := \{\lambda \in \mathbb{C} : \mathrm{Re}\lambda > 0\}$? Can we exclude the pathological situation where M is singular?
2. When a suitable preconditioner is applied, we know that the spectrum of $P^{-1}M$ is clustered; see, e.g., [23]. But which localization of the cluster (and of the outliers, if present) should be expected? Again, how does that localization depend on q?

Concerning the first issue, the literature contains plenty of spectral results involving Toeplitz matrices (see, e.g., [12, 13]), even though the nonsymmetric case is more difficult to treat (surprisingly, smoothness of the generating function can be a disadvantage: see [32]). In particular, this difficulty arises in our setting, where generating functions are trigonometric polynomials, and the accurate localization results typical for Hermitian matrices are no longer applicable. Moreover, such results are of the asymptotic type and require a critical assumption: *the entries of M must not depend on the size*. In other words, as the size varies we obtain a finite section of a *fixed* infinite matrix. This is not our case, since varying s (the size of matrices A and B and $s = O(h^{-1})$) gives a different value of q in M.

The only known results we can apply concern mainly algorithms for computation of a few eigenvalues (in [2]) or a theoretical analysis of the "pencil" $A - qB$ (in the sense of generalized eigenvalues) in [14]: in the latter book we can find conditions on q for which M is nonsingular, that is, a partial answer to our questions raised above.

In summary, to the best of our knowledge, *a general theoretical characterization of the eigenvalues of $M = A - qB$ as functions of q is still lacking*. The underlying algebraic setting is the (standard) eigenvalue problem for nonsymmetric Toeplitz matrices with small rank corrections.

Concerning the second issue, some mathematical tools for the spectral analysis of $P^{-1}M$ have been proposed in the literature (see, e.g., [4, 10, 17, 25]), but they all assume that $M - P$ has small rank. This is true for some choices of P (such as Strang's preconditioner and a few extensions), but several other important instances (such as T. Chan's or P-circulant approximations) give rise to matrices $M - P$ whose rank is usually full.

Therefore, *appropriate tools for the analysis of the case where P differs from M by more than a matrix whose rank is small*[1] *are still unknown.* An exception is provided in the Hermitian case, not of interest in this context. The underlying algebraic setting is the *generalized* eigenvalue problem for nonsymmetric Toeplitz matrices with small rank corrections.

The following sections will attempt to give some answers to the open questions discussed so far.

4. Spectral analysis. From now on we focus on the generalized eigenvalue problem for nonsymmetric quasi-Toeplitz matrices:

$$M(q)u = \lambda P(q)u, \qquad u \neq 0. \tag{4.1}$$

The standard eigenproblem falls into this notation by making the formal assumption $P(q) = I$ (in this section we are interested just in the *structure* of the matrices involved).

The lack of symmetry and the band structure imply that the classical approach based on generating functions gives very little insight here (see the results presented in [32]). Therefore, the best way to characterize eigenvalues (and potentially eigenvectors) by exploiting the band Toeplitz pattern seems to be the *difference equation* approach, proposed by Trench [33] for the standard, pure Toeplitz case.

Let the $s+1$ equations of (4.1), as well as the entries of u, be indexed from 0 to s; the indices from ν to $s-k+\nu$ correspond to the rows of $M(q)$ and $P(q)$ not affected by the low rank correction and containing all of the coefficients of the main method. The resulting relations

$$\sum_{i=0}^{k}(\alpha_i - q\beta_i)u_{i+j} = \lambda\sum_{i=0}^{k}(\breve{a}_{i-\nu} - q\breve{b}_{i-\nu})u_{i+j}, \quad j = 0, \ldots, s-k$$

(where we assume a periodic pattern for $\breve{a}_i$ and $\breve{b}_i$, whenever a subscript is out of range), can be treated as linear k-order homogeneous difference equations with constant coefficients. The first and last rows of (4.1) will provide us with initial and final conditions.

The eigenvector u is a nonzero solution of the difference problem and therefore can be characterized in terms of the algebraic *characteristic equation* of degree k:

$$\pi(z) - \lambda\breve{\pi}(z) = 0, \quad \pi(z) := \sum_{i=0}^{k}(\alpha_i - q\beta_i)z^i, \ \breve{\pi}(z) := \sum_{i=0}^{k}(\breve{a}_{i-\nu} - q\breve{b}_{i-\nu})z^i \tag{4.2}$$

(notice that $\breve{\pi}(z)$ simplifies into z^ν in the standard problem).

From now on we assume that, for each eigenvalue λ, all of the roots $z_1(\lambda), \ldots, z_k(\lambda)$ of the characteristic equation are distinct (otherwise, λ is called *defective* [34], but this pathological situation occurs just in isolated cases and for specific values of s). In this case, each component of the solution of the difference equation has the form

$$u_j = \sum_{l=1}^{k} c_l z_l(\lambda)^j, \quad j = 0, \ldots, s, \tag{4.3}$$

for suitable coefficients $c_1, \ldots, c_k$ determined by the boundary conditions.

[1] In the sense that s is supposed large with respect to the band of the Toeplitz matrices involved, and the rank is not depending on s.

More specifically, the first ν and the last $k-\nu$ rows of (4.1) represent additional conditions on the sequence u_j. In the standard problem, we have ν *initial* and $k-\nu$ *final* conditions since just the first and last entries of u are involved, respectively. In the generalized problem, the circulant structure of $P(q)$ determines a mixing of initial and final entries in all of these k equations, but for simplicity we keep the same terminology.

Substituting (4.3) into the mentioned equations, we obtain k homogeneous relations involving the unknown coefficients $c_1, \dots, c_k$, which can be put in matrix form as follows:

$$\begin{aligned} K_{\text{in}}(z_1(\lambda), \dots, z_k(\lambda))c = 0, \\ K_{\text{fin}}(z_1(\lambda), \dots, z_k(\lambda))c = 0, \end{aligned} \tag{4.4}$$

where $K_{\text{in}} \in \mathbb{C}^{\nu \times k}$, $K_{\text{fin}} \in \mathbb{C}^{(k-\nu)\times k}$ and we have emphasized the dependence of these Vandermonde-like matrices on the roots of the characteristic equation. The trivial solution $c = 0$ would imply $u = 0$ and therefore must be discarded; hence the square matrix

$$K(z_1(\lambda), \dots, z_k(\lambda)) := \begin{pmatrix} K_{\text{in}} \\ K_{\text{fin}} \end{pmatrix} \tag{4.5}$$

must be singular. Its (vanishing) determinant can be regarded as a function of λ having the same zeros of the characteristic polynomial of (4.1).

An alternative parameterization with respect to the roots $z_j(\lambda)$ can be useful for a different characterization of λ.

Let ζ be one of the roots, say, $z_1(\lambda)$. From one point of view, ζ is a function of λ, but it is understood that λ can be retrieved as well from ζ by means of the characteristic equation

$$\lambda(\zeta) = \frac{\pi(\zeta)}{\breve{\pi}(\zeta)} \qquad (\lambda(\zeta) = \zeta^{-\nu}\pi(\zeta) \text{ in the standard case}); \tag{4.6}$$

we remark that *any* root gives the same value of λ. The other roots can be expressed in terms of λ by inverting some elementary symmetric functions. For example, in the generalized problem with $k = 2$ and $\nu = 1$, the easiest way is to consider the ratio between the constant term and the leading coefficient in (4.2)

$$\frac{\alpha_0 - q\beta_0 - \lambda(\breve{a}_s - q\breve{b}_s)}{\alpha_2 - q\beta_2 - \lambda(\breve{a}_1 - q\breve{b}_1)} = z_1(\lambda)z_2(\lambda),$$

whence, after the substitution $\lambda = \lambda(\zeta)$ given in (4.6),

$$z_2(\lambda) = \zeta^{-1}\frac{(\alpha_0 - q\beta_0)\breve{\pi}(\zeta) - (\breve{a}_s - q\breve{b}_s)\pi(\zeta)}{(\alpha_2 - q\beta_2)\breve{\pi}(\zeta) - (\breve{a}_1 - q\breve{b}_1)\pi(\zeta)} =: \zeta_2(\zeta). \tag{4.7}$$

In general, we can assume that we have explicit functions $\zeta_2(\zeta), \dots, \zeta_k(\zeta)$ that replace $z_2(\lambda), \dots, z_k(\lambda)$ in the matrix K of (4.5). Thus

$$\det K(\zeta, \zeta_2(\zeta), \dots, \zeta_k(\zeta)) =: \det(\zeta; q)$$

is a function of the single complex variable ζ, containing q as a parameter.

TABLE 4.1
Coefficients for some 2-step formulas.

Type	α_0	α_1	α_2	β_0	β_1	β_2
Midpoint (MP)	-1	0	1	0	2	0
Simpson (S)	-1	0	1	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$
Adams–Moulton (AM)	-1	1	0	$\frac{5}{12}$	$\frac{2}{3}$	$-\frac{1}{12}$

As we will see in specific examples, the analysis of the function $\det(\zeta; q)$ can be sometimes reduced to the study of a sparse polynomial, which makes feasible a first-order analysis (perhaps a direct computation) of its roots $\zeta(q)$. Finally, the relation (4.6) allows us to obtain a knowledge of λ from that of $\zeta(q)$.

Remark 4.1. $\det(\zeta; q)$ has a number of redundant roots that should be discarded in order to simplify the analysis. Some of them are "spurious" values for which $\zeta = \zeta_j(\zeta)$ or $\zeta_j(\zeta) = \zeta_l(\zeta)$, with $j \neq l$ (the matrix K turns out to have two equal columns), violating the assumption of distinct roots. Furthermore, if ζ is a root of $\det(\zeta; q)$, then $\zeta_2(\zeta), \ldots, \zeta_k(\zeta)$ are roots as well, and they all give the same eigenvalue λ. In summary, since there are $s+1$ eigenvalues, we expect to find $k(s+1)$ roots of $\det(\zeta; q)$, plus the spurious roots (whose number cannot be estimated a priori, in general).

Remark 4.2. Once the behavior of $\zeta(q)$ has been obtained, in principle this can be used also for the study of eigenvectors: the key relation is (4.3), and the main issue would be the behavior (in terms of q) of the coefficients $c_1, \ldots, c_k$. This problem is not treated in the present paper, where we are interested only in the eigenvalues $\lambda(q)$.

4.1. A case study: 2-step formulas. In this paper, we focus on 2-step methods as the principal (or main) scheme (2.3) for a linear multistep formula in boundary value form, with one initial condition and one final condition provided by an implicit Euler scheme. For those methods, we have

$$k = 2, \nu = 1,$$

and Dahlquist's hypothesis of section 2.3 allows us to assume that

$$A = \begin{pmatrix} \alpha_1 & \alpha_2 & & & 0 \\ \alpha_0 & \alpha_1 & \alpha_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \alpha_0 & \alpha_1 & \alpha_2 \\ & & & -1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} \beta_1 & \beta_2 & & & \\ \beta_0 & \beta_1 & \beta_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_0 & \beta_1 & \beta_2 \\ & & & 0 & 1 \end{pmatrix}, \tag{4.8}$$

where parameters are given in Table 4.1 for the most common cases.

Circulant approximations for A and B are given by

$$\breve{A} = \begin{pmatrix} \breve{a}_0 & \breve{a}_1 & & & \breve{a}_s \\ \breve{a}_s & \breve{a}_0 & \breve{a}_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \breve{a}_s & \breve{a}_0 & \breve{a}_1 \\ \breve{a}_1 & & & \breve{a}_s & \breve{a}_0 \end{pmatrix}, \quad \breve{B} = \begin{pmatrix} \breve{b}_0 & \breve{b}_1 & & & \breve{b}_s \\ \breve{b}_s & \breve{b}_0 & \breve{b}_1 & & \\ & \ddots & \ddots & \ddots & \\ & & \breve{b}_s & \breve{b}_0 & \breve{b}_1 \\ \breve{b}_1 & & & \breve{b}_s & \breve{b}_0 \end{pmatrix}, \tag{4.9}$$

where the examples for its entries considered here are shown in Table 4.2.

TABLE 4.2
Entries for the preconditioner $P = P(q)$.

Type	$\breve{a}_0$	$\breve{a}_1$	$\breve{a}_s$	$\breve{b}_0$	$\breve{b}_1$	$\breve{b}_s$
MP, P-circ	0	$\frac{s+2}{s+1}$	$-\frac{s}{s+1}$	2	0	0
S, P-circ	0	$\frac{s+2}{s+1}$	$-\frac{s}{s+1}$	$\frac{4}{3}$	$\frac{s+2}{3(s+1)}$	$\frac{s}{3(s+1)}$
AM, Chan	1	0	$-\frac{s}{s+1}$	$\frac{2}{3}$	$-\frac{s}{12(s+1)}$	$\frac{5s}{12(s+1)}$
AM, P-circ	1	0	$-\frac{s}{s+1}$	$\frac{2}{3}$	$-\frac{s+2}{12(s+1)}$	$\frac{5s}{12(s+1)}$

We are not interested in T. Chan's approximation for the midpoint and Simpson methods, since it becomes singular in such cases [7, p. 1819].

The characteristic equation (4.2) has a quadratic form, with

$$\pi(z) = \gamma_0 + \gamma_1 z + \gamma_2 z^2 \qquad (\gamma_i := \alpha_i - q\beta_i),$$

$$\breve{\pi}(z) = \breve{g}_s + \breve{g}_0 z + \breve{g}_1 z^2 \qquad (\breve{g}_i := \breve{a}_i - q\breve{b}_i);$$

its roots ζ and ζ_2 are then related through (4.7), where we observe that the numerator vanishes for $\zeta = 0$, whereas the denominator loses its quadratic term. Hence we define

$$N(\zeta) := \frac{\gamma_0\breve{\pi}(\zeta) - \breve{g}_s\pi(\zeta)}{\zeta}, \quad D(\zeta) := \gamma_2\breve{\pi}(\zeta) - \breve{g}_1\pi(\zeta),$$

which are both linear polynomials such that

$$\zeta_2(\zeta) = \frac{N(\zeta)}{D(\zeta)}. \tag{4.10}$$

In light of Remark 4.1, we know that the function $\det(\zeta; q)$ has two spurious roots for which $\zeta = \zeta_2$, satisfying the quadratic equation $N(\zeta) - \zeta D(\zeta) = 0$; hence we know in advance that $N(\zeta) - \zeta D(\zeta)$ exactly divides $\det(\zeta; q)$.

In order to form the explicit expression of $\det(\zeta; q)$, first we must compute the 2×2 matrix

$$K(\zeta, \zeta_2) = \begin{pmatrix} \gamma_{\text{in}}(\zeta) & \gamma_{\text{in}}(\zeta_2) \\ \gamma_{\text{fin}}(\zeta) & \gamma_{\text{fin}}(\zeta_2) \end{pmatrix},$$

where $\gamma_{\text{in}}(\cdot)$ and $\gamma_{\text{fin}}(\cdot)$ are suitable polynomials obtained by imposing boundary conditions on the main difference equation.

More precisely, since $k = 2$ and $\nu = 1$, we have just one initial condition (the first equation in (4.1))

$$\gamma_1 u_0 + \gamma_2 u_1 = \lambda(\breve{g}_0 u_0 + \breve{g}_1 u_1 + \breve{g}_s u_s)$$

and one final condition (the last of (4.1))

$$-u_{s-1} + (1-q)u_s = \lambda(\breve{g}_1 u_0 + \breve{g}_s u_{s-1} + \breve{g}_0 u_s),$$

where u_j given by (4.3),

$$u_j = c_1\zeta^j + c_2\zeta_2^j,$$

should be substituted.

This way we obtain two homogeneous equations in the unknowns c_1, c_2, whose coefficients contribute to the matrix K. For instance, the coefficient of c_1 in the initial condition is

$$\gamma_{\text{in}}(\zeta) = \gamma_1 + \gamma_2\zeta - \lambda(\breve{g}_0 + \breve{g}_1\zeta + \breve{g}_s\zeta^s),$$

whereas that in the final condition is

$$\gamma_{\text{fin}}(\zeta) = -\zeta^{s-1} + (1-q)\zeta^s - \lambda(\breve{g}_1 + \breve{g}_s\zeta^{s-1} + \breve{g}_0\zeta^s).$$

The same holds for c_2 with ζ_2 in place of ζ; it must be remembered that

$$\lambda(\zeta) = \frac{\pi(\zeta)}{\breve{\pi}(\zeta)} = \frac{\pi(\zeta_2)}{\breve{\pi}(\zeta_2)}. \tag{4.11}$$

Some further algebraic manipulations give the following compact formulas:

$$\gamma_{\text{in}}(z) = \pi_0(z) - \frac{\pi(z)}{\breve{\pi}(z)}\mathcal{R}[z^s\breve{\pi}], \quad \gamma_{\text{fin}}(z) = z^{s-2}\pi_s(z) - \frac{\pi(z)}{\breve{\pi}(z)}\mathcal{R}[z^{s-1}\breve{\pi}],$$

where the notation $\mathcal{R}[P]$ means the s-degree remainder of P modulo $z^{s+1}-1$, and

$$\pi_0(z) := \frac{\pi(z) - \pi(0)}{z}, \quad \pi_s(z) := (1-q)z^2 - z.$$

A useful simplification arises by observing that in light of (4.11) λ needs not to be evaluated in ζ_2 when we form the second column of K. Therefore, the determinant is given by

$$\det(\zeta; q) = \gamma_{\text{in}}(\zeta)\gamma_{\text{fin}}(\zeta_2) - \gamma_{\text{in}}(\zeta_2)\gamma_{\text{fin}}(\zeta),$$

where

$$\gamma_{\text{in}}(\zeta_2) = \pi_0(\zeta_2) - \frac{\pi(\zeta)}{\breve{\pi}(\zeta)}Q_s(\zeta_2)$$

and

$$\gamma_{\text{fin}}(\zeta_2) = \zeta_2^{s-2}\pi_s(\zeta_2) - \frac{\pi(\zeta)}{\breve{\pi}(\zeta)}\tilde{Q}_s(\zeta_2),$$

with suitable s-degree polynomials Q_s and $\tilde{Q}_s$; the substitution (4.10) shows that $\det(\zeta; q)$ is a rational function whose denominator is $\breve{\pi}(\zeta)^2 D(\zeta)^s$. From the linearity of N and D, the function

$$d(\zeta; q) := \breve{\pi}(\zeta)^2 D(\zeta)^s \det(\zeta; q) \tag{4.12}$$

is a $(2s+4)$-degree polynomial in ζ, for which $N(\zeta) - \zeta D(\zeta)$ is a known exact divisor. Its significant roots occur in pairs $(\zeta(q), \zeta_2(q))$, each of them providing a unique value of $\lambda(q)$.

The formulas derived so far simplify very much if we are concerned with the nonpreconditioned case: it suffices to put formally $\breve{\pi}(z) := z$, so that

$$\zeta_2 = \frac{\gamma_0}{\gamma_2 \zeta}, \quad \gamma_{\text{in}}(z) := \pi_0(z) - \frac{\pi(z)}{z} = -\frac{\gamma_0}{z}, \quad \gamma_{\text{fin}}(z) := z^{s-2}(\pi_s(z) - z\pi(z)),$$

whence

$$\det(\zeta; q) = -\gamma_2 \zeta_2^{s-1}(\pi_s(\zeta_2) - \zeta_2 \pi(\zeta_2)) + \gamma_2 \zeta^{s-1}(\pi_s(\zeta) - \zeta\pi(\zeta));$$

here the denominator is just ζ^{s+2}, and the spurious roots of $d(\zeta; q) := \zeta^{s+2} \det(\zeta; q)$ are $\pm\sqrt{\gamma_0/\gamma_2}$.

It is important to observe that $d(\zeta; q)$ is a *sparse* polynomial, which makes a first-order analysis feasible.

We sketch below the essential formulas arising for the specific examples under consideration, which represent the individual instances of (4.10) for ζ_2, (4.12) for $d(\zeta; q)$, and (4.6) for $\lambda(q) := \lambda(\zeta(q))$. In the preconditioned cases, polynomials N and D have been scaled by a constant common factor σ which has been explicitly reported; hence the true expression of $d(\zeta; q)$ should be multiplied by σ^s, but obviously this correction has no influence on the roots and will not be considered in the subsequent analysis.

Nonpreconditioned matrices $M(q)$.

Midpoint (MP).

$$\zeta_2 = -\frac{1}{\zeta},$$
$$d(\zeta; q) = (-1)^s((1+q)\zeta + 1) + \zeta^{2s+3}(1 + q - \zeta),$$
$$\lambda(\zeta(q)) = \zeta(q) - 2q - \frac{1}{\zeta(q)}.$$

Simpson (S).

$$\zeta_2 = \frac{\gamma}{\zeta}, \quad \gamma := \frac{q/3 + 1}{q/3 - 1},$$
$$d(\zeta; q) = \gamma^{s+1}\left(\frac{q}{3} - 1\right)\left(1 + \frac{q}{3} + \left(1 + \frac{q}{3}\right)\zeta + \frac{q}{3\gamma}\zeta^2\right),$$
$$- \zeta^{2s+2}\left(\frac{q}{3} - 1\right)\left(\frac{q}{3} + \left(1 + \frac{q}{3}\right)\zeta + \left(\frac{q}{3} - 1\right)\zeta^2\right),$$
$$\lambda(\zeta(q)) = \left(1 - \frac{q}{3}\right)\zeta(q) - \frac{4}{3}q - \left(1 + \frac{q}{3}\right)/\zeta(q).$$

Adams–Moulton (AM).

$$\zeta_2 = \frac{\gamma}{\zeta}, \quad \gamma := -5 - \frac{12}{q},$$
$$d(\zeta; q) = \gamma^{s-1}\left(1 + \frac{5}{12}q\right)\left(\gamma\left(1 + \frac{5}{12}q\right) - \frac{1}{3}\gamma q\zeta + \frac{5}{12}q\zeta^2\right)$$
$$+ \zeta^{2s+2}\frac{q^2}{36}\left(\frac{5}{4} - \zeta - \frac{1}{4}\zeta^2\right),$$
$$\lambda(\zeta(q)) = \frac{q}{12}\zeta(q) + 1 - \frac{2}{3}q - \left(1 + \frac{5}{12}q\right)/\zeta(q).$$

Preconditioned matrices $P(q)^{-1}M(q)$.

MP, P-circulant.

$$\sigma = -\frac{s+1}{2}, \quad N(\zeta) = q - \zeta, \quad D(\zeta) = 1 + q\zeta, \quad \zeta_2 = \frac{N(\zeta)}{D(\zeta)},$$

$$\begin{aligned} d(\zeta;q) = {} & \frac{2\pi(\zeta)}{s+1}(N-\zeta D)\left[-\left(1-\frac{1}{s+1}\right)\zeta^s N^s + \left(1+\frac{1}{s+1}\right)D^s\right] \\ & + \left(1-\frac{1}{(s+1)^2}\right)\pi(\zeta)^2[N^s - (\zeta D)^s] \\ & - \frac{2(1+q)\breve{\pi}(\zeta)}{s+1}[N^{s+1} - (\zeta D)^{s+1}] + \frac{4}{(s+1)^2}[N^{s+2} - (\zeta D)^{s+2}], \end{aligned}$$

$$\pi(\zeta) = \zeta^2 - 2q\zeta - 1, \quad \breve{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}(\zeta^2+1), \quad \lambda(\zeta(q)) = \frac{\pi}{\breve{\pi}}.$$

S, P-circulant.

$$\sigma = -\frac{9(s+1)}{2}, \quad N(\zeta) = (3+q)(2q+(q-3)\zeta), \quad D(\zeta) = (3-q)(q+3+2q\zeta),$$

$$\begin{aligned} d(\zeta;q) = {} & \pi(\zeta)(N-\zeta D)(\breve{g}_1 D^{s-1}\phi_1 + \breve{g}_s(\zeta N)^{s-1}\psi_0) \\ & - \left(\breve{g}_1\breve{g}_s\pi(\zeta)^2 + \frac{2D\psi_0}{9(s+1)}\right)[N^s - (\zeta D)^s] - \frac{2\psi_1}{9(s+1)}[N^{s+1} - (\zeta D)^{s+1}] \end{aligned}$$

($\breve{g}_i := \breve{a}_i - q\breve{b}_i$, where $\breve{a}_i$ and $\breve{b}_i$ are given by Table 4.2),

$$\phi_1 := \frac{2D}{9(s+1)}, \quad \psi_0 := -\breve{\pi}(\zeta) + \frac{s}{s+1}\left(1+\frac{q}{3}\right)\pi(\zeta), \quad \psi_1 := (1-q)\breve{\pi}(\zeta) + \frac{4}{3}q\pi(\zeta),$$

$$\pi(\zeta) = \zeta^2 - 1 - \frac{q}{3}(\zeta^2 + 4\zeta + 1), \quad \breve{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}\left(1 + \zeta^2 + \frac{q}{3}(1-\zeta^2)\right).$$

AM, Chan.

$$\sigma = -\frac{12(s+1)}{2q/3 - 1}, \quad N(\zeta) = 5q + 12, \quad D(\zeta) = -q\zeta,$$

$$\begin{aligned} d(\zeta;q) = {} & \left(1-\frac{1}{s+1}\right)\pi(\zeta)(N-\zeta D)\left[\frac{q}{12}D^{s-1}\phi_1 - \left(1+\frac{5}{12}q\right)(\zeta N)^{s-1}\psi_0\right] \\ & + \zeta N D\phi_1\psi_0[N^{s-2} - (\zeta D)^{s-2}] + (\zeta N\phi_1\psi_1 + D\phi_0\psi_0)[N^{s-1} - (\zeta D)^{s-1}] \\ & + \left[\phi_0\psi_1 + \frac{q}{12}\left(1-\frac{1}{s+1}\right)^2\left(1+\frac{5}{12}q\right)\pi(\zeta)^2\right][N^s - (\zeta D)^s], \end{aligned}$$

$$\phi_0 := \left(1-\frac{2}{3}q\right)(\breve{\pi}(\zeta) - \pi(\zeta)), \quad \phi_1 := \frac{q}{12}\left[\breve{\pi}(\zeta) - \left(1-\frac{1}{s+1}\right)\pi(\zeta)\right],$$

$$\psi_0 := -\breve{\pi}(\zeta) + \left(1-\frac{1}{s+1}\right)\left(1+\frac{5}{12}q\right)\pi(\zeta), \quad \psi_1 := (1-q)\breve{\pi}(\zeta) - \left(1-\frac{2}{3}q\right)\pi(\zeta),$$

$$\pi(\zeta) = \zeta - 1 - \frac{q}{12}(\zeta^2 - 8\zeta - 5), \quad \breve{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}\left(-\frac{q}{12}\zeta^2 + 1 + \frac{5}{12}q\right).$$

AM, P-circulant.

$$\sigma = -(s{+}1),\ N(\zeta) = \left(1+\frac{5q}{12}\right)\left(1-\frac{2}{3}q+\frac{q\zeta}{6}\right),\ D(\zeta) = \frac{q}{12}\left[\left(1-\frac{2}{3}q\right)\zeta - 2 - \frac{5}{6}q\right],$$

$$d(\zeta;q) = \pi(\zeta)(N-\zeta D)(\breve{g}_1 D^{s-1}\phi_1 + \breve{g}_s(\zeta N)^{s-1}\psi_0) + \zeta N D\phi_1\psi_0[N^{s-2}-(\zeta D)^{s-2}]$$
$$+ (\zeta N\phi_1\psi_1 + D\phi_0\psi_0)[N^{s-1}-(\zeta D)^{s-1}] + (\phi_0\psi_1 - \breve{g}_1\breve{g}_s\pi(\zeta)^2)[N^s-(\zeta D)^s]$$

($\breve{g}_i := \breve{a}_i - q\breve{b}_i$, where $\breve{a}_i$ and $\breve{b}_i$ are given by Table 4.2),

$$\phi_0 := \left(1-\frac{2}{3}q\right)(\breve{\pi}(\zeta)-\pi(\zeta)), \quad \phi_1 := \frac{q}{12}\left[\breve{\pi}(\zeta) - \left(1+\frac{1}{s+1}\right)\pi(\zeta)\right],$$

$$\psi_0 := -\breve{\pi}(\zeta) + \left(1-\frac{1}{s+1}\right)\left(1+\frac{5}{12}q\right)\pi(\zeta), \quad \psi_1 := (1-q)\breve{\pi}(\zeta) - \left(1-\frac{2}{3}q\right)\pi(\zeta),$$

$$\pi(\zeta) = \zeta - 1 - \frac{q}{12}(\zeta^2 - 8\zeta - 5), \quad \breve{\pi}(\zeta) = \pi(\zeta) + \frac{1}{s+1}\left(\frac{q}{12}\zeta^2 + 1 + \frac{5}{12}q\right).$$

4.2. A first-order analysis. The parameterization of λ as a function of q obtained so far allows us to investigate the behavior of the eigenvalues for q small.

We recall that $q = h\mu$, where h is the time discretization step and μ represents any eigenvalue of the Jacobian matrix J in (2.4) related to the space discretization. Thus, a small value of q is a physically meaningful situation, occurring whenever, e.g., the Jacobian matrix has eigenvalues with a small modulus (as in the examples sketched in section 5) and/or a small time step is used. A particular care is required in the latter instance: we stress that $s \to \infty$ as $h \to 0$, so that the polynomial $d(\zeta;q)$ raises its degree, increasing the number of the roots $\zeta(q)$. However, the insights given by the first-order analysis are generally in good agreement with the localization of λ, as we will see in the numerical experiments of section 6.

In what follows, we present a first-order expansion of $\lambda(q)$ centered in zero for all of the three nonpreconditioned methods (MP, S, AM) and for two preconditioners (P-circulant approximations for MP and AM).

The starting point is the continuity of polynomial roots with respect to coefficients (provided that the degree remains constant). Hence $\zeta(q)$ is very close to $\zeta(0)$ for small q, and its first-order dependence on q can be made explicit.

In the MP method, $\zeta = \zeta(0)$ is a root of $d(\zeta;0) = (-1)^s(1+\zeta) + \zeta^{2s+3}(1-\zeta)$, and therefore

$$|\zeta|^{2s+3} = \left|\frac{1+\zeta}{1-\zeta}\right|.$$

Squaring both sides of the previous equation and letting $\zeta = \rho e^{\mathbf{i}\theta}$, after some algebraic manipulations, we get

$$\cos\theta = \frac{1+\rho^2}{2\rho}\cdot\frac{\rho^{4s+6}-1}{\rho^{4s+6}+1}. \tag{4.13}$$

(4.13) is the equation, in polar coordinates, of a curve containing all of the roots $\zeta(0)$ and lying in the following region of the complex plane:

$$\Omega = \left\{\theta \in \left(\frac{\pi}{2}, \frac{3}{2}\pi\right), \rho < 1\right\} \cup \left\{|\theta| < \frac{\pi}{2}, \rho > 1\right\} \cup \{\pm\mathbf{i}\},$$

where $\pm\mathbf{i}$ are exactly the spurious roots for which $\zeta = \zeta_2$.

Therefore, $\lambda(0)$ can be localized through the transformation $\lambda = \zeta - 1/\zeta$ of the previous curve. In particular, since $\mathrm{Re}\lambda = (\rho - 1/\rho)\cos\theta$, it is straightforward to observe that $\mathrm{Re}\lambda > 0$ whenever $\zeta \in \Omega$ (except for the spurious roots). By continuity, we have the useful result that the eigenvalues of M lie on $\mathbb{C}^+$ for q small enough. We recall that projection methods such as GMRES or BiCGstab show often a faster convergence behavior whenever the matrix of the linear systems we have to solve has all eigenvalues in one half-plane; see [22].

If we are interested in a deeper analysis, we can check that the roots $\zeta(0)$ are distinct and therefore

$$\zeta(q) \doteq \zeta(0) + \zeta'(0)q,$$

where $\doteq$ denotes a first-order approximation of the function on the left-hand side. Therefore,

$$\lambda(q) \doteq \zeta(0) - \frac{1}{\zeta(0)} + \left[\zeta'(0) + \frac{\zeta'(0)}{\zeta(0)^2} - 2\right] q;$$

the explicit expression of $\zeta'(0)$, if desired, can be retrieved from the classical theory on the conditioning of zeros of polynomials (see, e.g., [30, section 5.8]).

The Simpson method has a quite similar analysis. In addition, since the matrix A is the same as the previous case, the zero-order terms of $\zeta(q)$ and $\lambda(q)$ are exactly equal to the corresponding ones for MP. On the other hand, the first-order expansion for S has a different expression, which is reported below:

$$\lambda(q) \doteq \zeta(0) - \frac{1}{\zeta(0)} + \left[\zeta'(0) + \frac{\zeta'(0)}{\zeta(0)^2} - \frac{4}{3} - \frac{\zeta(0)}{3} - \frac{1}{3\zeta(0)}\right] q,$$

where $\zeta'(0)$ is different from the MP method.

The analysis of the AM method shows a further complication with respect to the previous cases. It is evident that $d(\zeta; q)$ loses several degrees when q goes to zero, so that many roots $\zeta(q)$ become infinite. Hence we are not able to predict the behavior of $\lambda(q)$, unless we apply an appropriate change of variable. For this purpose, let

$$\xi := q^{1/2}\zeta, \quad \beta := \gamma q,$$

and rewrite the polynomial d in terms of the new variable ξ. We obtain

$$\begin{aligned} q^s d(\xi; q) = \beta^{s-1}\left(1 + \frac{5}{12}q\right)\left(\beta\left(1 + \frac{5}{12}q\right) - \frac{1}{3}\beta q^{1/2}\xi + \frac{5}{12}q\xi^2\right) \\ + \xi^{2s+2}/36\left(\frac{5}{4}q - q^{1/2}\xi - \frac{1}{4}\xi^2\right), \end{aligned}$$

whence $\xi(q) \doteq \xi(0) + \xi'(0)q^{1/2}$, where $\xi(0)$ solves the equation

$$(-12)^s - \xi^{2s+4}/144 = 0,$$

that is, $\xi(0) = 2\sqrt{3}\exp(\mathbf{i}(\frac{l\pi}{s+2} + \frac{\pi}{2}))$, $l = 1, \ldots, s+1$. Other values of the index l would give spurious roots or already obtained values of λ. Taking into account the change of variable, the behaviors of ζ and λ are, respectively,

$$\zeta(q) \doteq \xi(0)q^{-1/2} + \xi'(0),$$

$$\lambda(q) \doteq 1 + \frac{\mathbf{i}}{\sqrt{3}}\cos\frac{l\pi}{s+2}q^{1/2} + \left[\frac{\xi'(0)}{12} + \frac{\xi'(0)}{\xi(0)^2} - \frac{2}{3}\right] q.$$

Hence, for small values of q, the eigenvalues of M are close to a vertical segment on $\mathbb{C}^+$ with the midpoint placed at 1.

The main difficulty arising in the preconditioned case is given by the presence of the spurious divisor $N - \zeta D$ in all instances of $d(\zeta; q)$. In order to perform the analysis, it is worth considering the quotient $\hat{d}(\zeta; q) := d(\zeta; q)/(N - \zeta D)$, studied for $q \approx 0$. Let

$$F_m(\zeta; q) := \frac{N^m - (\zeta D)^m}{N - \zeta D} = \sum_{j=0}^{m-1} N^j (\zeta D)^{m-j-1};$$

this expression appears in almost all of the terms of $\hat{d}(\zeta; q)$ and will determine the first-order behavior of the significant roots.

Concerning the P-circulant preconditioner for the MP method, for $q = 0$ we have $N(\zeta) = -\zeta$, $D(\zeta) = 1$. Therefore,

$$F_m(\zeta; 0) = \zeta^{m-1} \sum_{j=0}^{m-1} (-1)^j = \begin{cases} \zeta^{m-1} & \text{if } m \text{ is odd,} \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the zero-order localization of the roots $\zeta(q)$ strictly depends on the parity of s: more specifically, when s is odd they solve the equation

$$\begin{aligned} &(\zeta^2 - 1)\zeta^{s+1} + \frac{2}{s+1}(\zeta^2 - 1)(\zeta^{2s} + 1) \\ &+ \frac{1}{(s+1)^2}[4\zeta^{s+1} - (\zeta^2 - 1)\zeta^{s-1} + 2(\zeta^2 - 1)(1 - \zeta^{2s})] = 0, \end{aligned}$$

and when s is even the equation becomes

$$(\zeta^2 - 1)(1 - \zeta^s - \zeta^{2s}) + \frac{1}{s+1}[(\zeta^2 - 1)(\zeta^{2s} + 1) - (\zeta^2 + 1)\zeta^s] = 0.$$

For q very small, the eigenvalues of $P^{-1}M$ can be estimated from the roots $\zeta = \zeta(0)$ through the relation

$$\lambda(0) = \frac{\zeta^2 - 1}{\zeta^2 - 1 + \dfrac{1}{s+1}(\zeta^2 + 1)}.$$

After more heavy computations we are able to obtain the first-order terms in the expansions of $\zeta(q)$ and $\lambda(q)$. In section 6 we will present explicit estimates based on the formulas derived so far and compare them with values obtained numerically.

The difficulties found in the analysis of the AM method arise in the P-circulant preconditioned case as well. Many ingredients of $d(\zeta; q)$ degenerate for $q = 0$: among others, polynomials $N(\zeta), D(\zeta), \pi(\zeta)$, and $\breve{\pi}(\zeta)$ drop their degree. This causes several roots $\zeta(q)$ to go to infinity: also here we need a suitable change of variable.

Let $\xi := q\zeta$, and rewrite all of the polynomials $N, D, \pi, \breve{\pi}, \phi_i, \psi_i (i = 0, 1)$ in terms of the new variable, in particular,

$$N(\xi) = \left(1 + \frac{5}{12}q\right)\left(1 - \frac{2}{3}q + \frac{1}{6}\xi\right), \quad D(\xi) = \frac{1}{12}\left[\left(1 - \frac{2}{3}q\right)\xi - q\left(2 + \frac{5}{6}q\right)\right].$$

The "clean" polynomial $\hat{d}$ takes the following expression after some algebra:

$$\hat{d}(\xi; q) = q^{-s-1}\breve{d}(\xi; q),$$

where $\check{d}$ has constant degree $2s+2$ (independently on q) and its zero-order form is

$$\check{d}(\xi;0) = \left(-\frac{1}{s+1}\right)^s \xi^{s+1}\left\{-\frac{1}{12^{s+1}(s+1)}\left[\xi^{s-3}(12+2\xi)^2+\xi^{s+1}\right] - \left(1-\frac{1}{s+1}\right)\left(1+\frac{1}{12}\xi\right)\left(1+\frac{1}{6}\xi\right)^s\right\}.$$

Notice that $s+1$ roots of $\check{d}$ are distinct and behave as $\xi(q) \doteq \xi(0)+\xi'(0)q$, whence

$$\zeta(q) \doteq \xi(0)q^{-1}+\xi'(0).$$

These are the roots going to infinity, associated with the values of $\lambda(q)$ with

$$\lambda(0) = \frac{12+\xi(0)}{12+\xi(0)\left(1-\dfrac{1}{s+1}\right)};$$

the same eigenvalues are associated with the corresponding "dual" roots given by $\zeta_2 = N(\xi)/D(\xi)$ which are finite for $q=0$, as a direct look at N and D shows. Through the transformation $\xi = q\zeta_2$ we find the remaining $s+1$ roots of $\check{d}$, which collapse at the origin.

We will compare these results with numerical estimates for this setting as well; see section 6.

5. Model problems. As a first benchmark of our analysis, we consider two simple model problems which encompass two important types of spectra for their Jacobian matrices: real and negative and pure imaginary eigenvalues, respectively. Only one-dimensional (1D) problems are considered, but extensions to 2D and 3D cases are straightforward and not necessary in our setting.

Diffusion equation. As a typical example of a problem whose Jacobian matrix has negative (real) eigenvalues, we report the variable coefficient 1D diffusion equation with homogeneous Dirichlet boundary conditions at both ends. Let $a = a(x) \geq 0$ be a suitably smooth function.

$$\begin{cases} u_t - c(a\,u_x)_x = 0, & x \in [0, x_{\max}], t \in (0,T], \\ u(0,t) = u(x_{\max},t) = 0 & t \in (0,T], \\ u(x,0) = g(x), & x \in [0, x_{\max}]. \end{cases} \tag{5.1}$$

Discretizing the operator $\partial/\partial x$ in (5.1) with centered differences and step size $\Delta x = x_{\max}/(m+1)$ gives a sequence of systems of ODEs parameterized by Δx whose mth element is given by

$$\begin{cases} y'(t) = T_m y(t), & t \in [0,T], \\ y(0) = \eta, & \eta = (g(x_1)\dots g(x_m))^T, \end{cases} \tag{5.2}$$

where $x_j = j\Delta x$ and

$$T_m = \frac{c}{(\Delta x)^2}\begin{pmatrix} a_1 & b_1 & & \\ b_1 & \ddots & \ddots & \\ & \ddots & \ddots & b_{m-1} \\ & & b_{m-1} & a_m \end{pmatrix}, \tag{5.3}$$

where

$$a_j = -(a(x_{j-1/2}) + a(x_{j+1/2})), \quad b_j = a(x_{j+1/2}).$$

The Jacobian matrix T_m is $m \times m$ symmetric, tridiagonal, and weakly diagonally dominant with negative diagonal entries. From an extensive study performed in [15, 28] we get for each eigenvalue μ_j of T_m the bounds

$$-\frac{4c}{(\Delta x)^2} \max_x\{a(x)\} \le \mu_j \le -\frac{c\pi^2}{(x_{\max})^2} \min_x\{a(x)\}.$$

Note that, as Δx tends to zero, the systems of differential equations (5.2) become increasingly stiff, spreading the eigenvalues of the Jacobian matrix T_m along an interval in $(-4c \max_x\{a\}/(\Delta x)^2, 0)$ whose left boundary tends to $-\infty$ with $O((\Delta x)^{-2})$.

More precisely, the spectrum is equally distributed [31] as the values of the bivariate function $a(x)f(\theta)$, where

$$f(\theta) = \frac{2c}{(\Delta x)^2}(\cos(\theta) - 1), \quad \theta \in (-\pi, \pi),$$

is the so-called "generating function" related to the constant-coefficient version of the problem. As stated in [27, 24], if $a(x)$ has a zero at the origin of order α, the smallest eigenvalue shows an asymptotic behavior like $(\Delta x)^2/m^{\max(2,\alpha)}$.

Transport equation. The linear 1D transport equation with periodic boundary conditions and constant coefficient $c > 0$ in its simplest form reads:

$$\begin{cases} u_t + c\, u_x = 0, \\ u(x,0) = g(x), & x \in [0,\pi], \\ u(\pi,t) = u(0,t), & t \in [0,2\pi]. \end{cases} \tag{5.4}$$

Discretizing the partial derivative $\partial/\partial x$ with central differences and step size $\Delta x = \pi/m$, $x_j = j\Delta x$ gives a sequence of systems of ODEs parameterized by Δx whose mth element is given by

$$\begin{cases} y'(t) = C_m y(t), & t \in [0,2\pi], \\ y(0) = \eta, & \eta = (g(x_0) \dots\ g(x_{m-1}))^T, \end{cases} \tag{5.5}$$

with

$$C_m = \frac{c}{2\Delta x} \begin{pmatrix} 0 & -1 & & & 1 \\ 1 & \ddots & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & \ddots & -1 \\ -1 & & & 1 & 0 \end{pmatrix}. \tag{5.6}$$

The matrix C_m is circulant $m \times m$ with generating function

$$\tilde{f}(\theta) = \frac{c}{2\Delta x}\left(e^{-\mathbf{i}\theta} - e^{\mathbf{i}\theta} - e^{-\mathbf{i}(m-1)\theta} + e^{\mathbf{i}(m-1)\theta}\right) = \frac{-\mathbf{i}c\,(\sin\theta - \sin(m-1)\theta)}{\Delta x}, \tag{5.7}$$

where $\theta \in (-\pi, \pi)$. Therefore, the eigenvalues of C_m are distributed as $\tilde{f}(\theta)$ in (5.7) and lie in the purely imaginary (closed) interval

$$[-2\mathbf{i}c/\Delta x, 2\mathbf{i}c/\Delta x],$$

which becomes wider as $O(1/\Delta x)$ as we refine the discretization. This implies that a finer mesh for the time-step integrator is required to resolve the (oscillatory) solution as Δx (the step for the discretization in space) decreases to zero.

An explicit expression of the spectrum of C_m can be obtained by observing that

$$C_m = F\Lambda F^*, \tag{5.8}$$

where Λ is a diagonal matrix containing the eigenvalues μ_j of C_m and F is the Fourier matrix; see, e.g., [19]. Thus, from the expression of the eigenvalues of a circulant matrix, we have

$$\mu_j = -\frac{2\mathbf{i}c}{\Delta x}\left(\sin\frac{2\pi j}{m}\right), \quad j = 0, \ldots, m-1,$$

i.e., the generating function computed in the points $\theta_j = 2\pi j/m$, $j = 0, \ldots, m-1$, as usual.

It is worth noting that the Jacobian matrices for both of the proposed model problems are normal and therefore can be diagonalized by unitary matrices. This feature is useful in order to use the bounds for the convergence of a Krylov accelerator which uses the preconditioners analyzed here; see [9, Theorems 3.1 and 3.2]. In particular, by applying the cited results, for the underlying problems we can predict convergence in at most $O(\log s)$ (preconditioned) iterations.

6. Numerical estimates and comparisons. We compare the results of zero- and of some first-order approximations presented in section 4.2 with the eigenvalues computed by Matlab's QR method for the model problems in section 5. We do not report plots generated by Simpson's formula because they are very similar to those related to the midpoint formula.

In all tests, unless specified otherwise, we consider $s = m = 100$, $c = 1$, $T = 2\pi$, $x_{\max} = \pi$, $t_0 = 0$. The Jacobian matrix J is taken, in light of Dahlquist's hypothesis, as the smallest eigenvalue (in modulus) for each one of the model problems considered in the previous section. In the variable diffusion model problem, the diffusion function is of the form $a(x) = x^k$, $k > 0$ integer; i.e., it has a zero in the origin of multiplicity k. However, a similar eigenvalue distribution of the preconditioned and nonpreconditioned problems has been observed even in the absence of zeros on the real axis for various functions such as $a(x) = x^k + \epsilon$, where $\epsilon > 0$ is a small constant, varying with $O(m^{-1})$. We stress that, in both cases, eigenvalues of the Jacobian matrix (5.3), are negative, but some of them go to zero as the space discretization gets refined. On the other hand, the same asymptotic behavior holds for some nonzero eigenvalues of the Jacobian matrix (5.6), although the transport equation has constant coefficients.

Results of some tests are reported in Figures 6.1 (nonpreconditioned case), 6.2, and 6.3 (preconditioned case). In all three cases, the condition number $\kappa_2(X)$ of the eigenvector matrix X is modest. Therefore GMRES' convergence is well described by the eigenvalues.

Note that in all tests we get that even just zero-order approximations can give reasonable information on the qualitative behavior of the eigenvalues related to the smallest eigenvalues (in modulus) of the Jacobian matrix of the differential problem both in the nonpreconditioned and in the preconditioned cases, for variable and constant coefficient equations, provided that the mesh for the discretization in space is fine enough.

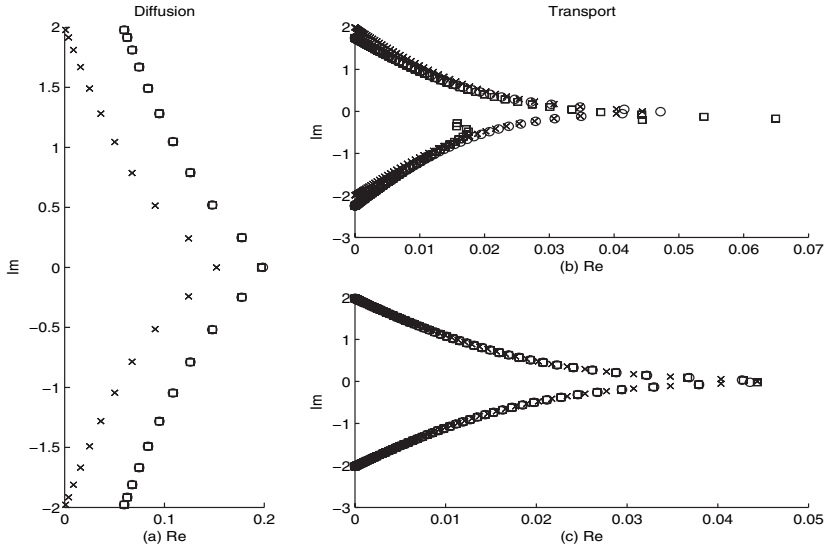


FIG. 6.1. *MP method, smallest eigenvalue (in modulus) for* (a) *the diffusion equation with* $a(x) = x^4$, $s = 20$, $m = 20$ *giving* $\kappa_2(X) \approx 7$, (b) *the transport equation,* $s = 100$, $m = 100$, $c = 1$ *giving* $\kappa_2(X) \approx 28$, *and* (c) *the same equation with* $c = 0.1$ *giving* $\kappa_2(X) \approx 28$; *x=order* 0, *o=eig*(M), $\square$*=order* 1 *approximations.*

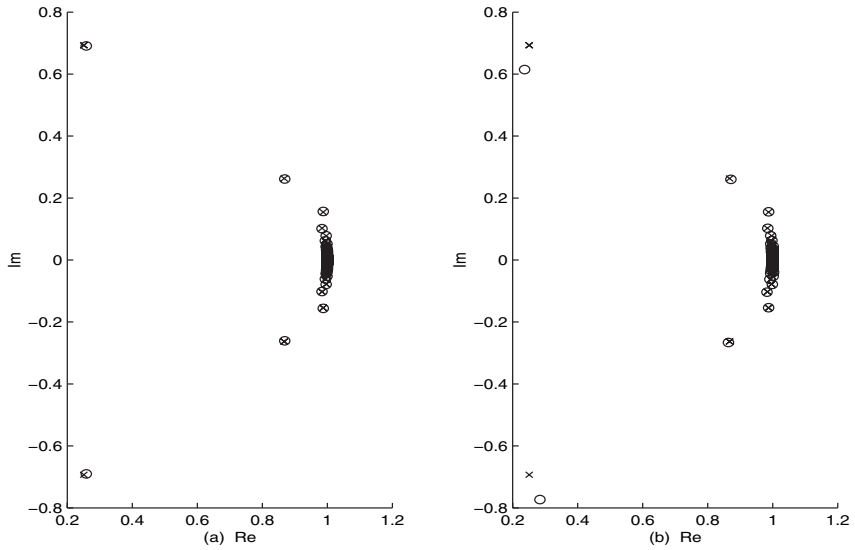


FIG. 6.2. *MP method with* P*-circulant preconditioning, smallest eigenvalue (in modulus) for* (a) *the diffusion equation with* $a(x) = x^4$, $s = 100$, $m = 100$ *and* (b) *the transport equation,* $s = 100$, $m = 100$; *+=order 0, o=eig*$(P^{-1}M)$ *approximations.*

In order to emphasize the effect of the first-order approximations with respect to zero order, just in Figure 6.1 (left) we use a rougher mesh with $s = m = 20$ for the midpoint formula without using preconditioning.

It is surprising that, for the transport equation (upper right plot in Figure 6.1), the order 1 approximation gives worse approximations than order 0 for some eigen-

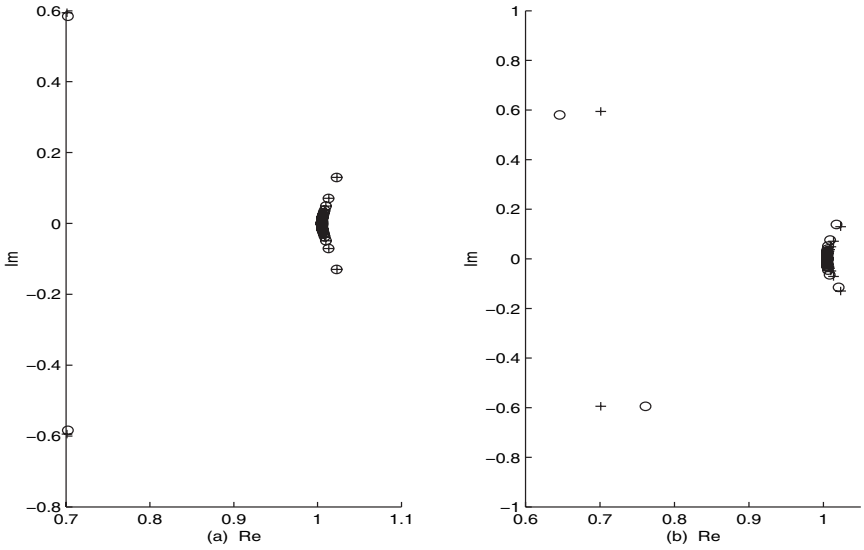


FIG. 6.3. *Adams–Moulton with* P*-circulant preconditioning, smallest eigenvalue (in modulus) for* (a) *the diffusion equation with* $a(x) = x^4$, $s = 100$, $m = 100$ *and* (b) *the transport equation,* $s = 100$, $m = 100$*; +=order 0, o=eig*$(P^{-1}M)$ *approximations.*

values: the "wrong" values come from roots $\zeta(0)$ very close to the real axis (the same occurs for the derivatives $\zeta'(0)$, $\zeta''(0)$, etc.). This phenomenon is probably explained by observing that q is pure imaginary in this setting, so that in the power series $\sum_{j=0}^{+\infty} \zeta^{(j)}(0)q^j$ just the even terms contribute to refine the real part, as well as the odd terms are related only to the imaginary part; this way the convergence radius of the series could be reduced, and the actual value of q could fall outside the region of analyticity. On the other hand, continuity still holds so that order 0 is always meaningful.

Our conjecture is confirmed by the lower right plot in Figure 6.1, where we have simply set $c = 0.1$: q has been divided by a factor of 10, and order 1 estimates become again more accurate than order 0.

For these moderate dimensions, every $\zeta(0)$ has been computed through the Matlab function `roots`. If one is interested in locating the spectrum of much larger matrices, we suggest the use of more efficient rootfinders specifically designed for sparse polynomials, such as MPSolve proposed in [11].

REFERENCES

[1] A. O. H. AXELSSON AND J. G. VERWER, *Boundary value techniques for initial value problems in ordinary differential equations*, Math. Comp., 45 (1985), pp. 153–171.

[2] R. M. BEAM AND R. F. WARMING, *The asymptotic spectra of banded Toeplitz and quasi-Toeplitz matrices*, SIAM J. Sci. Comput., 14 (1993), pp. 971–1006.

[3] D. BERTACCINI, *P-circulant preconditioners and the systems of ODE codes*, in Iterative Methods in Scientific Computation IV, IMACS Series in Computational and Applied Mathematics, D. R. Kincaid et al., eds., IMACS, NJ, 1999, pp. 179–193.

[4] D. BERTACCINI, *A circulant preconditioner for the systems of LMF-based ODE codes*, SIAM J. Sci. Comput., 22 (2000), pp. 767–786.

[5] D. BERTACCINI, *Reliable preconditioned iterative linear solvers for some numerical integrators*, Numer. Linear Algebra Appl., 8 (2001), pp. 111–125.

[6] D. Bertaccini and M. K. Ng, *The convergence rate of block preconditioned systems arising from LMF-based ODE codes*, BIT, 41 (2001), pp. 433–450.
[7] D. Bertaccini, *The spectrum of circulant-like preconditioners for some general linear multistep formulas for linear boundary value problems*, SIAM J. Numer. Anal., 40 (2002), pp. 1798–1822.
[8] D. Bertaccini and M. K. Ng, *Block $\{\omega\}$-circulant preconditioners for the systems of differential equations*, Calcolo, 40 (2003), pp. 71–90.
[9] D. Bertaccini and M. K. Ng, *Band-Toeplitz preconditioned GMRES iterations for time-dependent PDEs*, BIT, 43 (2003), pp. 901–914.
[10] D. Bertaccini, Y. Wen, and M. K. Ng, *The eigenvalues of preconditioned matrices for linear multistep formulas in boundary value form*, Numer. Linear Algebra Appl., 12 (2005), pp. 315–325.
[11] D. Bini and G. Fiorentino, *Design, analysis and implementation of a multiprecision polynomial rootfinder*, Numer. Algorithms, 23 (2000), pp. 127–173.
[12] A. Böttcher and B. Silbermann, *Introduction to Large Truncated Toeplitz Matrices*, Springer, New York, 1998.
[13] A. Böttcher and S. M. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM, Philadelphia, 2005.
[14] L. Brugnano and D. Trigiante, *Solving ODE by Linear Multistep Methods: Initial and Boundary Value Methods*, Gordon & Breach, Reading, U.K., 1998.
[15] R. H. Chan and T. Chan, *Circulant preconditioners for elliptic problems*, Numer. Linear Algebra Appl., 1 (1992), pp. 77–101.
[16] R. H. Chan and M. K. Ng, *Conjugate gradient methods for Toeplitz systems*, SIAM Rev., 38 (1996), pp. 427–482.
[17] R. H. Chan, M. K. Ng, and X. Q. Jin, *Circulant preconditioners for solving ordinary differential equations*, in Structured Matrices, Bini D. et al. eds., Nova Science, Hauppauge, NY, 2001.
[18] T. F. Chan, *An optimal circulant preconditioner for Toeplitz systems*, SIAM J. Sci. Comput., 9 (1988), pp. 766–771.
[19] P. J. Davis, *Circulant Matrices*, John Wiley & Sons, New York, 1979.
[20] L. Fox, *A note on the numerical integration of first-order differential equations*, Quart. J. Mech. Appl. Math., 7 (1954), pp. 367–378.
[21] E. Hairer and G. Wanner, *Solving Ordinary Differential Equations* II. *Stiff and Differential-Algebraic Problems*, Springer, Berlin, 1991.
[22] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM, Philadelphia, 1997.
[23] M. K. Ng, *Iterative Methods for Toeplitz Systems*, Oxford University Press, New York, 2004.
[24] D. Noutsos, S. Serra-Capizzano, and P. Vassalos, *The Conditioning of FD Matrix Sequences Coming from Semi-Elliptic Differential Equations*, manuscript.
[25] K. Otto, *Analysis of preconditioners for hyperbolic partial differential equations*, SIAM J. Numer. Anal., 33 (1996), pp. 2131–2165.
[26] U. Peskin and N. Moiseyev, *Time-independent scattering theory for time-periodic Hamiltonians: Formulation and complex scaling calculations of above threshold ionization spectra*, Phys. Rev. A, 49 (1994), pp. 3712–3728.
[27] S. Serra, *The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems*, Numer. Math., 81 (1999), pp. 461–495.
[28] S. Serra-Capizzano and C. T. Possio, *Spectral and structural analysis of high precision finite difference matrices for elliptic operators*, Linear Algebra Appl., 293 (1999), pp. 85–131.
[29] G. Strang, *A proposal for Toeplitz matrix calculations*, Stud. Appl. Math., 74 (1986), pp. 171–176.
[30] J. Stoer and R. Bulirsch, *Introduction to Numerical Analysis*, Springer, New York, 1980.
[31] P. Tilli, *Locally Toeplitz sequences: Spectral properties and applications*, Linear Algebra Appl., 278 (1998), pp. 91–120.
[32] P. Tilli, *Some results on complex Toeplitz eigenvalues*, J. Math. Anal. Appl., 239 (1999), pp. 390–401.
[33] W. F. Trench, *On the eigenvalue problem for Toeplitz band matrices*, Linear Algebra Appl., 64 (1985), pp. 199–214.
[34] W. F. Trench, *Numerical solution of the eigenvalue problem for Hermitian Toeplitz matrices*, SIAM J. Matrix Anal. Appl., 10 (1989), pp. 135–146.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2368–2391

AN INTERPOLATION ERROR ESTIMATE ON ANISOTROPIC MESHES IN $\mathcal{R}^n$ AND OPTIMAL METRICS FOR MESH REFINEMENT*

WEIMING CAO†

Abstract. In this paper, we extend the work in [W. Cao, *Math. Comp.*, to appear] to functions of n dimensions. We measure the anisotropic behavior of higher-order derivative tensors by the "largest" (in certain sense) ellipse/ellipsoid contained in the level curve/surface of the polynomial for directional derivatives. Given the anisotropic measure for the interpolated functions, we derive an error estimate for piecewise polynomial interpolations on meshes that are quasi-uniform under a given metric. By using the inertia properties for matrix eigenvalues [R. C. Thompson, *J. Math. Anal. Appl.*, 58 (1977), pp. 572–577] and Hölder's inequality, we can identify the optimal mesh metrics leading to the smallest error bound in various norms. Furthermore, we develop a dimensional reduction method to find the anisotropic measure approximately. We present two numerical examples for linear and quadratic interpolation on various anisotropic meshes generated with the optimal mesh metrics developed in this paper. Numerical results show that the smallest interpolation error is attained exactly on meshes optimal for the corresponding error norm as predicted.

Key words. interpolation error, anisotropic mesh, anisotropic measure, aspect ratio, mesh alignment, mesh metric

AMS subject classifications. 65D05, 65L50, 65N15, 65N50

DOI. 10.1137/060667992

1. Introduction.

It is well known in mesh generation and finite element analysis communities that in order to minimize linear interpolation errors, the eigenvalues and eigenvectors of Hessian matrices can be used to determine the element aspect ratio and mesh alignment direction for anisotropic mesh generation or refinement; see, e.g., [3, 4, 11, 13, 15, 19, 21, 23, 24]. In the case of quadratic or higher-order interpolations, the error is determined by the third- or higher-order partial derivatives of the interpolated functions. Measuring their anisotropic behavior is the key for anisotropic mesh design and refinement [2]. In particular, in order to determine an ideal element orientation and aspect ratio, one needs to define the "principal direction" and the "strength" to characterize the anisotropic behavior of the derivative tensors. In a previous paper [6], the author developed a method to measure the orientation and anisotropic ratio of the higher-order derivative tensors for two-dimensional functions. The technique is based on decomposing the homogeneous polynomials for directional derivatives into the product of linear and nonnegative quadratic polynomials. Then the anisotropic measure is defined by the directions of the lines and ellipses corresponding to those factors. An interpolation error estimate is further derived on anisotropic meshes that are quasi-uniform under given metrics. Optimal mesh metrics can be identified to minimize the error bound in various norms.

In this paper, we extend the work in [6] to functions of n dimensions. The difficulty in making such an extension is that it is generally impossible to factor a homogeneous polynomial in three or higher dimensions into the product of linear and nonnegative

*Received by the editors August 22, 2006; accepted for publication (in revised form) April 6, 2007; published electronically November 21, 2007. This work was supported in part by NSF grant DMS-0209313.

http://www.siam.org/journals/sinum/45-6/66799.html

†Department of Mathematics, University of Texas at San Antonio, San Antonio, TX 78249 (wcao@math.utsa.edu).

quadratic functions. However, the idea in [6], which characterizes the anisotropic behavior of derivative tensors by the largest ellipse contained in the level curve of the polynomial for directional derivatives (see Remark 2.1 in [6]), is still valid in higher dimensions. Algebraically, for any positive integer k, let $\nabla^{k+1}u(\boldsymbol{x}) = \nabla(\nabla^k u)(\boldsymbol{x})$ be the $(k+1)$th-order tensor for the partial derivatives of function u at a point $\boldsymbol{x}$. We may characterize its anisotropic behavior by a suitable $n \times n$ positive definite matrix Q satisfying the following constraint:

$$|(\boldsymbol{\xi} \cdot \nabla)^{k+1} u(\boldsymbol{x})| \leq |\boldsymbol{\xi} \cdot Q\boldsymbol{\xi}|^{\frac{k+1}{2}} \qquad \forall \boldsymbol{\xi} \in \mathcal{R}^n.$$

Based on the anisotropic measure of the interpolated functions, we derive an error estimate for the higher-order polynomial interpolation on meshes that are quasi-uniform under a given mesh metric. The error bound involves the same convergence rate with respect to the number of elements as in classical results on quasi-uniform meshes under Euclidean metrics. However, the constant in the error bound depends on an interplay between the anisotropic mesh (through the mesh metric) and the anisotropic behavior of $\nabla^{k+1}u$, which can be much smaller than that in the classical error bound. Furthermore, by using the inertia properties of matrix eigenvalues proved by Thompson [26] and Hölder's inequality, we show that the smallest bound for the $W^{m,p}$-error of kth-order interpolations is attained when the metric is defined as

$$M_{k+1,m,p} = c(\lambda_{max}(Q))^{\frac{mp}{(k+1-m)p+n}} |\det(Q)|^{-\frac{1}{(k+1-m)p+n}} \cdot Q,$$

where $\lambda_{max}(Q)$ is the largest eigenvalue of the matrix Q measuring the anisotropic behavior of $\nabla^{k+1}u$. In the case of linear interpolation ($k = 1$), Q can be chosen as $Q = [(\nabla^2 u)^2]^{1/2} + \delta \cdot I_n$, where δ is a user-specified small parameter and I_n is the identity matrix. Then the above optimal metric for minimizing the L^p-error (i.e., with $m = 0$) is identical to that presented in Chen, Sun, and Xu [7]. In the case of higher-order interpolations, the optimal choice of Q relies on a constrained minimization. We develop a dimensional reduction method to find an approximate Q to measure the anisotropic behavior of $\nabla^{k+1}u$. To test the optimality of the proposed mesh metrics, we generated various anisotropic meshes using the black-box anisotropic mesh generator `bamg` [3, 16] supplied with the metric $M_{k+1,m,p}$. We compare various error norms for linear and quadratic interpolations in two dimensions. It is found in all the cases that the smallest error norm is attained exactly with meshes based on the corresponding optimal mesh metric. We also present a three-dimensional example to demonstrate that the smallest error norms are attained at the best aspect ratios, as expected.

An outline of this paper is as follows. In section 2 we present the error estimate for interpolations on anisotropic meshes that are quasi-uniform under a given metric. The metrics leading to the smallest error bounds (in various norms) are identified. In section 3 we introduce the notion of anisotropic measure of higher-order derivative tensors, and present a dimension reduction algorithm to determine the measure approximately. In section 4 we present two numerical examples demonstrating the optimality of the proposed mesh metrics. We conclude the paper with some discussion.

2. Error estimates on anisotropic meshes.

2.1. Basic assumptions and lemmas. We introduce in this subsection some assumptions regarding the anisotropic behaviors of the meshes and the higher-order derivative tensors of the interpolated functions, and list several lemmas needed for

deriving the interpolation error estimates. Denote by $\{\mathcal{T}_N\}$ a family of triangulations of a polygonal domain $\Omega \in \mathcal{R}^n$ into simplicial elements (see [9]). Here N represents the total number of elements. We study the error estimates for piecewise polynomial interpolations over a class of anisotropic meshes $\{\mathcal{T}_N\}$ that are quasi-uniform under a given metric. More specifically, let M be a Riemannian metric on Ω. For each element $\tau \in \mathcal{T}_N$, define M_τ to be the average of M over τ, i.e.,

$$M_\tau = \frac{1}{|\tau|}\int_\tau M(\boldsymbol{x})d\boldsymbol{x}.$$

Since M_τ is an $n \times n$ symmetric positive definite matrix, we may express it in its eigen-decomposition form,

$$M_\tau = \ T_\tau \cdot D_\tau \cdot T_\tau', \tag{1}$$

where $D_\tau = \mathrm{diag}(d_1, d_2, \ldots, d_n)$ is composed of all the eigenvalues of M_τ, and T_τ is the orthogonal matrix composed of all the eigenvectors. Define

$$F_\tau = T_\tau D_\tau^{-\frac{1}{2}}. \tag{2}$$

Clearly $(M_\tau)^{-1} = F_\tau \cdot F_\tau'$. Let $\boldsymbol{x}_c$ be the center of element τ. Define an affine mapping $\tilde{\boldsymbol{x}} = F_\tau^{-1}(\boldsymbol{x} - \boldsymbol{x}_c)$. Then τ is transformed into a simplex element $\tilde{\tau}$ with its center at the origin. We call a family of triangulations $\{\mathcal{T}_N\}$ quasi-uniform under metric M if there exist positive constants c_1 and c_2 independent of N such that

(i) for all $\tau \in \mathcal{T}_N$, the smallest internal angle of $\tilde{\tau} = F_\tau^{-1}(\tau - \boldsymbol{x}_c)$ is bounded from below by c_1, i.e., $\tilde{\tau}$ is shape regular; and
(ii) $\max_{\tau\in\mathcal{T}_N} |\tilde{\tau}| \le c_2 \min_{\tau\in\mathcal{T}_N} |\tilde{\tau}|$.

Note that if $\{\mathcal{T}_N\}$ is quasi-uniform under metric M, then the geometric features of the mesh are determined by M. Specifically, the element size (area/volume) is proportional to $|\det(M)|^{-1/n}$, the element aspect ratio is proportional to $1/\sqrt{d_1} : 1/\sqrt{d_2} : \cdots : 1/\sqrt{d_n}$, and the element orientation (or mesh alignment) is determined by the directions of the eigenvectors of M. Indeed, some anisotropic mesh generators, such as `bamg` developed by Borouchaki et al. [3] and Hecht [16], take user-specified metrics to control directly the anisotropic behaviors of the meshes.

We also would like to point out that not all anisotropic meshes can be considered as quasi-uniform under suitable metrics. For instance, the mesh in Figure 1 is anisotropic. But it cannot be quasi-uniform under any metric, since for a quasi-uniform mesh there is at most a fixed number of neighboring elements next to each vertex; otherwise the metric would be singular at certain points. For general meshes in n dimensions, a necessary condition for a family of triangulations being quasi-uniform under a metric is that there is an upper limit for the number of neighboring elements at each element interface (i.e., vertex, edge, and face). We call it the *limited connectivity condition* for anisotropic meshes.

In order to derive the continuous form of the interpolation error estimate, we make an assumption on the smoothness of the mesh metric M.

Assumption A. Let M be a given Riemannian metric on Ω. There exists $\delta > 0$ such that for any neighborhood $\mathcal{N}_{\boldsymbol{z}}$ of any $\boldsymbol{z} \in \Omega$ with radius (under metric M) less than δ, the following is true for all $\boldsymbol{x} \in \mathcal{N}_{\boldsymbol{z}}$:

$$c_1\ \boldsymbol{\xi} \cdot \bar{M}\boldsymbol{\xi}\ \le\ \boldsymbol{\xi} \cdot M(\boldsymbol{x})\boldsymbol{\xi}\ \le\ c_2\ \boldsymbol{\xi} \cdot \bar{M}\boldsymbol{\xi} \qquad \forall \boldsymbol{\xi} \in \mathcal{R}^n, \tag{3}$$

where $\bar{M}$ is the average of M over $\mathcal{N}_{\boldsymbol{z}}$.

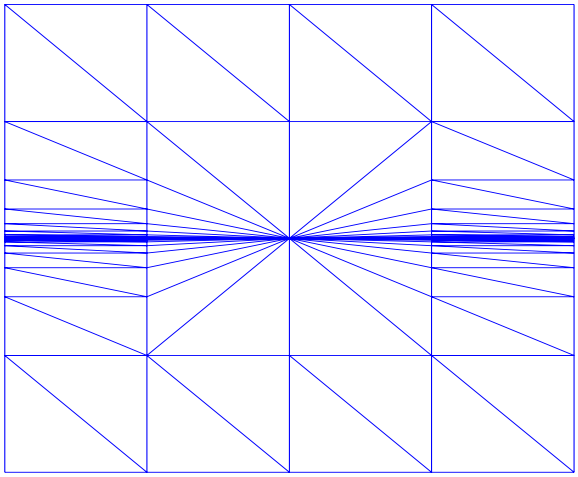


FIG. 1. *An example of anisotropic meshes that are not quasi-uniform under any metric.*

This assumption means that the metric M is equivalent to its local average over a sufficiently small neighborhood of each point. It basically requires the continuity of M. Indeed, if M is uniformly continuous over Ω, the above assumption holds obviously.

LEMMA 2.1. *Let M be a metric on Ω satisfying Assumption* A. *Let $F(\boldsymbol{x})F'(\boldsymbol{x}) = [M(\boldsymbol{x})]^{-1}$ be the decomposition of its inverse as given in* (2). *For any neighborhood $\mathcal{N}_{\boldsymbol{z}}$ of $\boldsymbol{z} \in \Omega$ with radius less than δ, we have for all $\boldsymbol{x} \in \mathcal{N}_{\boldsymbol{z}}$ that*

$$\|F^{-1}(\boldsymbol{x})\bar{F}\| \le \sqrt{c_2}, \qquad \|\bar{F}^{-1}F(\boldsymbol{x})\| \le 1/\sqrt{c_1}, \tag{4}$$

and that

$$(c_1)^n \,|\det(\bar{M})| \;\le\; \det(M(\boldsymbol{x})) \;\le\; (c_2)^n \,|\det(\bar{M})|, \tag{5}$$

where $\bar{M}$ is the average of M over $N_{\boldsymbol{z}}$, $\bar{F}\bar{F}' = \bar{M}^{-1}$ is the decomposition of $\bar{M}^{-1}$, and $\|\cdot\|$ stands for the 2-norm of matrices.

Proof. It is easy to see that

$$\|F^{-1}(\boldsymbol{x})\bar{F}\| = \max_{\boldsymbol{\xi}\in\mathcal{R}^n} \frac{\|F^{-1}(\boldsymbol{x})\bar{F}\boldsymbol{\xi}\|}{\|\boldsymbol{\xi}\|} \;=\; \left[\max_{\boldsymbol{\xi}\in\mathcal{R}^n} \frac{(\bar{F}\boldsymbol{\xi})\cdot M(\boldsymbol{x})(\bar{F}\boldsymbol{\xi})}{\boldsymbol{\xi}\cdot\boldsymbol{\xi}}\right]^{1/2}$$

$$= \left[\max_{\boldsymbol{\xi}\in\mathcal{R}^n} \frac{\boldsymbol{\xi}\cdot M(\boldsymbol{x})\boldsymbol{\xi}}{(\bar{F}^{-1}\boldsymbol{\xi})\cdot(\bar{F}^{-1}\boldsymbol{\xi})}\right]^{1/2} \;=\; \left[\max_{\boldsymbol{\xi}\in\mathcal{R}^n} \frac{\boldsymbol{\xi}\cdot M(\boldsymbol{x})\boldsymbol{\xi}}{\boldsymbol{\xi}\cdot\bar{M}\boldsymbol{\xi}}\right]^{1/2} \;\le\; \sqrt{c_2},$$

and similarly for the second inequality in (4). To show (5), we note that $c_i\boldsymbol{\xi}\cdot\bar{M}\boldsymbol{\xi} = 1$ $(i = 1, 2)$ and $\boldsymbol{\xi}\cdot M(\boldsymbol{x})\boldsymbol{\xi} = 1$ are three ellipsoids in $\mathcal{R}^n$, with their volumes being $\frac{\pi^{[\frac{n}{2}]}}{\Gamma(1+\frac{n}{2})}\cdot(c_i)^{-\frac{n}{2}}\cdot|\det(\bar{M})|^{-1/2}$, $i = 1, 2$, and $\frac{\pi^{[\frac{n}{2}]}}{\Gamma(1+\frac{n}{2})}|\det(M(\boldsymbol{x}))|^{-1/2}$, respectively. Here $[\frac{n}{2}]$ represents the integer part of $\frac{n}{2}$, and Γ is the Gamma function. Assumption A means that ellipsoid $\boldsymbol{\xi}\cdot M(\boldsymbol{x})\boldsymbol{\xi} = 1$ lies in between ellipsoids $c_i\boldsymbol{\xi}\cdot\bar{M}\boldsymbol{\xi} = 1$ $(i = 1, 2)$. Thus (5) follows easily from the ordering of their volumes. ∎

Because the error for polynomial interpolations of degree k is determined by the $(k+1)$th derivative tensor $\nabla^{k+1}u$ of the interpolated functions u, we make an assumption on its anisotropic behavior.

Assumption B. For each $\boldsymbol{x} \in \Omega$, there exists a positive definite matrix $Q(\boldsymbol{x})$ such that

$$|(\boldsymbol{\xi} \cdot \nabla)^{k+1} u(\boldsymbol{x})| \le |\boldsymbol{\xi} \cdot Q(\boldsymbol{x}) \boldsymbol{\xi}|^{\frac{k+1}{2}} \qquad \forall \boldsymbol{\xi} \in \mathcal{R}^n. \tag{6}$$

Note that $p_{k+1}(\boldsymbol{\xi}) \equiv (\boldsymbol{\xi} \cdot \nabla)^{k+1} u(\boldsymbol{x})$ is a homogeneous polynomial of $\boldsymbol{\xi}$ of degree $k+1$. For $\|\boldsymbol{\xi}\| = 1$, $p_{k+1}(\boldsymbol{\xi})$ is simply the $(k+1)$th-order directional derivative at $\boldsymbol{x}$ along $\boldsymbol{\xi}$. The right-hand side is also a homogeneous function of $\boldsymbol{\xi}$. Thus, geometrically Assumption B is equivalent to saying that at each $\boldsymbol{x}$, the ellipse/ellipsoid $\boldsymbol{\xi} \cdot Q\boldsymbol{\xi} = 1$ is contained in the level curve/surface $|p_{k+1}(\boldsymbol{\xi})| = 1$ for directional derivatives.

Some choices of Q can be made readily. For instance, if one is only interested in isotropic mesh refinement, then we may choose $Q(\boldsymbol{x}) = |||\nabla^{k+1} u(\boldsymbol{x})||| \cdot I_n$, where

$$|||\nabla^{k+1} u(\boldsymbol{x})||| = \max_{\|\boldsymbol{\xi}\|=1} |(\boldsymbol{\xi} \cdot \nabla)^{k+1} u|$$

is the largest $(k+1)$th-order directional derivative at $\boldsymbol{x}$, and I_n is the $n \times n$ identity matrix. This choice corresponds to defining Q by the largest circle/sphere contained in the level curve/surface. In the case of $k = 1$ (i.e., linear interpolation), Q can be chosen as

$$Q = abs(\nabla^2 u) + \delta \cdot I_n,$$

where $abs(A) = [A^T A]^{1/2}$ is the symmetric matrix of the same eigenvectors as matrix A, but of eigenvalues equal to the absolute eigenvalues of A. δ is a small positive constant to avoid Q from being degenerate in case $\nabla^2 u(\boldsymbol{x})$ becomes singular. Such a Q is called the majorization matrix for the Hessian $\nabla^2 u$ in Chen, Sun, and Xu [7].

In order to reflect accurately the anisotropic behavior of $\nabla^{k+1} u$ at $\boldsymbol{x}$, the matrix $Q(\boldsymbol{x})$ in (6) should be chosen as "small" (in certain sense) as possible. Also, its choice should be invariant under translation and rotation transforms of the coordinates, since the anisotropic behavior is so. The best choice will depend on how Assumption B is used. For minimizing the interpolation error in various norms, we present in section 3.1 an ideal choice of Q. We also present in section 3.2 an algorithm to find an approximate Q that characterizes roughly the anisotropic behavior of $\nabla^{k+1} u$.

Next, we list two lemmas regarding the anisotropic behavior of $\nabla^{k+1} u$ under affine transforms.

LEMMA 2.2. *Let* $|\alpha| = \sum \alpha_i$ *for any multi-index* $\alpha = (\alpha_1, \ldots, \alpha_n)$ *of nonnegative integers. For any integer* $k \ge 0$, $\sum_{|\alpha|=k+1} |\partial^\alpha u|$ *is equivalent to* $|||\nabla^{k+1} u|||$.

Proof. Define

$$\bar{\mathcal{P}}_{k+1} = \left\{ v(\boldsymbol{\xi}) = \sum_{|\alpha|=k+1} C_\alpha (\xi_1)^{\alpha_1} \cdots (\xi_n)^{\alpha_n} \;\middle|\; \forall C_\alpha \in R \right\}.$$

The $\bar{\mathcal{P}}_{k+1}$ is a finite-dimensional linear space. It is easy to verify that

$$\|v\|_{\bar{\mathcal{P}}_{k+1}} = \max_{\|\boldsymbol{\xi}\|=1} |v(\boldsymbol{\xi})|$$

is a norm on $\bar{\mathcal{P}}_{k+1}$. Indeed, if $\|v\|_{\bar{\mathcal{P}}_{k+1}} = 0$, then for any $\boldsymbol{\xi} \ne \mathbf{0}$, $|v(\boldsymbol{\xi})| = \|\boldsymbol{\xi}\|^{k+1} \cdot |v(\boldsymbol{\xi}/\|\boldsymbol{\xi}\|)| = 0$, thus $v \equiv 0$ and $\|\cdot\|_{\bar{\mathcal{P}}_{k+1}}$ is a norm.

On the other hand, $\sum_{|\alpha|=k+1} |C_\alpha|$ is clearly also a norm on $\bar{\mathcal{P}}_{k+1}$. Hence it is equivalent to $\|v\|_{\bar{\mathcal{P}}_{k+1}}$. Now, for $p_{k+1}(\boldsymbol{\xi}) = (\boldsymbol{\xi} \cdot \nabla)^{k+1} u \in \bar{\mathcal{P}}_{k+1}$, since its coefficients

C_α are multiples of $\partial^\alpha u$ with fixed positive constants, therefore $\sum_{|\alpha|=k+1} |\partial^\alpha u|$ is equivalent to $\sum_{|\alpha|=k+1} |C_\alpha|$, and to $\max_{\|\boldsymbol{\xi}\|=1} |p_{k+1}(\boldsymbol{\xi})| = |||\nabla^{k+1} u|||$, too. ∎

LEMMA 2.3. *Let $\boldsymbol{x} = F_\tau(\tilde{\boldsymbol{x}}) + \boldsymbol{x}_c$ be the affine mapping from $\tilde\tau$ to τ, and define $\tilde u(\tilde{\boldsymbol{x}}) = u(F_\tau(\tilde{\boldsymbol{x}}) + \boldsymbol{x}_c)$. Denote by $\tilde\nabla$ the gradient operator with respect to $\tilde{\boldsymbol{x}}$. Then under Assumption* B *we have*

$$|||\tilde\nabla^{k+1} \tilde u(\tilde{\boldsymbol{x}})||| \le \|F_\tau' Q F_\tau\|^{\frac{k+1}{2}}.$$

Proof. Let $\tilde p_{k+1}(\boldsymbol{\xi}) = (\boldsymbol{\xi}\cdot\tilde\nabla)^{k+1}\tilde u(\tilde{\boldsymbol{x}})$. By the fact that F_τ is constant, it follows from the chain rule for derivatives that $\tilde\nabla = F_\tau'\nabla$ and $\tilde p_{k+1}(\boldsymbol{\xi}) = [\boldsymbol{\xi}\cdot(F_\tau'\nabla)]^{k+1} u(\boldsymbol{x}) = p_{k+1}(F_\tau\boldsymbol{\xi})$. Hence Assumption B implies that

$$|||\tilde\nabla^{k+1}\tilde u(\tilde{\boldsymbol{x}})||| = \max_{\|\boldsymbol{\xi}\|=1} |p_{k+1}(F_\tau\boldsymbol{\xi})| \le \max_{\|\boldsymbol{\xi}\|=1} |(F_\tau\boldsymbol{\xi})' Q (F_\tau\boldsymbol{\xi})|^{\frac{k+1}{2}} = \|F_\tau' Q F_\tau\|^{\frac{k+1}{2}}. \qquad ∎$$

Finally, in order to minimize the interpolation errors and to select the optimal mesh metrics, we need the following inertial properties for matrix eigenvalues established in [26] by Thompson and an elementary inequality.

LEMMA 2.4. *Let A be an $n\times n$ symmetric matrix with eigenvalues $\alpha_1 \ge \cdots \ge \alpha_n \ge 0$, and let S be a nonsingular matrix with singular values $s_1 \ge \cdots \ge s_n$. Denote the eigenvalues of $B = S^* A S$ by $\beta_1 \ge \cdots \ge \beta_n$. Then $\beta_{i+j-n} \ge s_i^2 \alpha_j$ for all $1 \le i, j \le n$ with $i + j > n$. In particular, $\lambda_{max}(B) = \beta_1 \ge \max_{1\le i\le n}(s_i)^2 \alpha_{n+1-i}$.*

LEMMA 2.5. *Let $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ and $0 \le \alpha < 1$. For $\boldsymbol{t} = (t_1, t_2, \ldots, t_n)$, let*

$$f(\boldsymbol{t}) = |t_n|^\alpha \cdot \max_{1\le i\le n} \frac{\lambda_i}{t_i}.$$

Then f attains its infimum on the set

$$\mathcal{K} \;=\; \{\; \boldsymbol{t} \in \mathcal{R}^n \mid 0 < t_1 \le t_2 \le \cdots \le t_n, \;\; \textit{and } \Pi_{i=1}^n t_i = 1 \;\}$$

if and only if $\frac{\lambda_i}{t_i} = \text{const.}$ for all i.

Proof. First we show that f attains its infimum on $\mathcal{K}$. Let $\{\boldsymbol{t}^{(k)}\}$ be a sequence in $\mathcal{K}$ such that $\lim f(\boldsymbol{t}^{(k)}) = \inf_{\boldsymbol{t}\in\mathcal{K}} f(\boldsymbol{t}) < \infty$. Clearly $\{\boldsymbol{t}^{(k)}\}$ must be bounded; otherwise there exists a subsequence $\{\boldsymbol{t}^{(k')}\}$ whose nth components $t_n^{(k')} \to \infty$, while its ith components $t_i^{(k')} \to 0$ for some i, which would imply $f(\boldsymbol{t}^{(k)}) \to \infty$. Therefore, $\{\boldsymbol{t}^{(k)}\}$ is bounded and has a cluster point $\boldsymbol{t}^* \in \mathcal{K}$ with $f(\boldsymbol{t}^*) = \inf_{\boldsymbol{t}\in\mathcal{K}} f(\boldsymbol{t})$.

Next we show that

$$\frac{\lambda_i}{t_i^*} \le \frac{\lambda_n}{t_n^*} \qquad \forall i. \tag{7}$$

Suppose otherwise; then there exists $m \le n-1$ such that

$$\max_{1\le i\le n} \frac{\lambda_i}{t_i^*} = \frac{\lambda_m}{t_m^*} > \frac{\lambda_j}{t_j^*} \qquad \forall j \ge m+1.$$

For each $j \ge m+1$, since $\lambda_m \le \lambda_j$, we must have $t_m^* < t_j^*$. Let $\beta > 1$ and define $\tilde{\boldsymbol{t}} = (\tilde t_1, \tilde t_2, \ldots, \tilde t_n)$ with

$$\tilde t_i \;=\; \begin{cases} \beta^{1/m} t_i^* & \text{if } i \le m, \\ \beta^{-1/(n-m)} t_i^* & \text{if } i \ge m+1. \end{cases}$$

For β close enough to 1, we have $\tilde{\boldsymbol{t}} \in \mathcal{K}$ and

$$\begin{aligned} f(\tilde{\boldsymbol{t}}) &= \beta^{-\alpha/(n-m)} \\ &\quad \cdot \max\left[\beta^{-1/m} \cdot (t_n^*)^\alpha \max_{1\le i\le m} \frac{\lambda_i}{t_i^*}, \beta^{-1/(n-m)} \cdot (t_n^*)^\alpha \max_{m+1\le i\le n} \frac{\lambda_i}{t_i^*} \right] \\ &< f(\boldsymbol{t}^*). \end{aligned}$$

Finally we prove that

$$\frac{\lambda_i}{t_i^*} = \text{const.} \qquad \forall i. \tag{8}$$

Note that $\Pi_{i=1}^n t_i^* = 1$. It follows from (7) that

$$\Pi_{i=1}^n \lambda_i \le \left(\frac{\lambda_n}{t_n^*}\right)^n,$$

which implies

$$t_n^* \le \left(\Pi_{i=1}^n \frac{\lambda_n}{\lambda_i}\right)^{1/n}.$$

In particular, "=" in the above inequality holds if and only if (8) is satisfied. Thus

$$\inf_{\boldsymbol{t}\in\mathcal{K}} f(\boldsymbol{t}) = f(\boldsymbol{t}^*) = \lambda_n (t_n^*)^{\alpha-1} \ge \lambda_n \left(\Pi_{i=1}^n \frac{\lambda_n}{\lambda_i}\right)^{\frac{\alpha-1}{n}},$$

and "=" holds if and only if (8) is satisfied. □

2.2. Error estimate and optimal mesh metrics. We first recall some classical results for the interpolation error estimates under Euclidean metrics. Let k be a positive integer. Denote by P_k the set of all the polynomials of $\boldsymbol{x} \in \mathcal{R}^n$ of total degree less than or equal to k. Let Π_k be an interpolation operator whose restriction on each element preserves P_k. It is well known that on any shape regular element τ, and for any $0 \le m \le k$ and $p, q \in [1, \infty]$

$$|u - \Pi_k u|_{m,p,\tau} \le c|\tau|^{(k+1-m)/n+1/p-1/q} |u|_{k+1,q,\tau}, \tag{9}$$

provided that

$$W^{k+1,q}(\tau) \hookrightarrow C^s(\tau) \qquad \text{and} \qquad W^{k+1,q}(\tau) \hookrightarrow W^{m,p}(\tau), \tag{10}$$

where s is the highest degree of derivatives used in defining the interpolation Π_k. See, e.g., Theorem 3.1.5 of [9].

If we further assume that $\{\mathcal{T}_N\}$ is a family of quasi-uniform triangulations, i.e., all $\tau \in \mathcal{T}_N$, for all N, are shape regular and

$$\max_{\tau\in\mathcal{T}_N} |\tau| \le c \min_{\tau\in\mathcal{T}_N} |\tau|,$$

then we have globally

$$\left(\sum_{\forall \tau\in\mathcal{T}_N} |u - \Pi_k u|_{m,p,\tau}^p\right)^{1/p} \le cN^{-(k+1-m)/n-1/p+1/q} |u|_{k+1,q,\Omega}. \tag{11}$$

Now we present the main theorem of this paper on interpolation error estimates for anisotropic meshes.

THEOREM 2.1. *Let M be a Riemannian metric on Ω satisfying Assumption* A, *and let $F(\boldsymbol{x})\ F'(\boldsymbol{x}) = (M(\boldsymbol{x}))^{-1}$ be the decomposition of its inverse at each $\boldsymbol{x} \in \Omega$. Let $\{\mathcal{T}_N\}$ be a family of triangulations of Ω that is quasi-uniform under metric M. Let k be a positive integer, and let Π_k be an interpolation operator whose restriction on each element preserves P_k. For any function u satisfying Assumption* B, *any $0 \le m \le k$, and $p \in [1, \infty]$ satisfying* (10), *we have*

(12)

$$\left(\sum_{\tau \in \mathcal{T}_N} |u - \Pi_k u|^p_{m,p,\tau} \right)^{1/p}$$

$$\le\ cN^{-(k+1-m)/n} \left\{ \int_\Omega |\det(M)|^{\frac{1}{2}} \right\}^{(k+1-m)/n} \left\{ \int_\Omega \|F^{-1}\|^{mp} \|F' Q F\|^{\frac{(k+1)p}{2}} \right\}^{1/p}.$$

Furthermore, among all the Riemannian metrics, the optimal bound of the above estimate is attained when M is defined to be

$$M_{k+1,m,p} \ \equiv\ c(\lambda_{max}(Q))^{\frac{mp}{(k+1-m)p+n}} |\det(Q)|^{-\frac{1}{(k+1-m)p+n}} \cdot Q. \tag{13}$$

If $\{\mathcal{T}_N\}$ is quasi-uniform under $M_{k+1,m,p}$, we have

$$\left(\sum_{\tau \in \mathcal{T}_N} |u - \Pi_k u|^p_{m,p,\tau} \right)^{1/p} \tag{14}$$

$$\le\ cN^{-(k+1-m)/n} \parallel |\lambda_{max}(Q)|^{\frac{m}{2}} |\det(Q)|^{\frac{k+1-m}{2n}} \|_{L^{np/[(k+1-m)p+n]}(\Omega)}.$$

Proof. Consider an element $\tau \in \mathcal{T}_N$. Denote by M_τ the average of M over τ, and let F_τ be defined as in (2). By the fact that (see [9])

$$\sum_{|\alpha|=m} |\partial^\alpha v(\boldsymbol{x})| \le c \|F_\tau^{-1}\|^m \cdot \sum_{|\alpha|=m} |\partial^\alpha \tilde{v}(\tilde{\boldsymbol{x}})| \qquad \forall v,$$

we have

$$\begin{aligned} |u - \Pi_k u|^p_{m,p,\tau} &= \int_{\tilde{\tau}} \sum_{|\alpha|=m} |\partial^\alpha (u - \Pi_k u)(\boldsymbol{x})|^p \, d\tilde{\boldsymbol{x}} \\ &\le c\, \frac{|\tau|}{|\tilde{\tau}|}\ \|F_\tau^{-1}\|^{mp}\ |\tilde{u} - \tilde{\Pi}_k \tilde{u}|^p_{m,p,\tilde{\tau}}. \end{aligned}$$

Because $\tilde{\tau}$ is shape regular, we have from the classical error estimate (9) that

$$|\tilde{u} - \tilde{\Pi}_k \tilde{u}|_{m,p,\tilde{\tau}} \le c |\tilde{\tau}|^{(k+1-m)/n}\ |\tilde{u}|_{k+1,p,\tilde{\tau}},$$

which implies that

$$|u - \Pi_k u|^p_{m,p,\tau} \le c|\tau|\, |\tilde{\tau}|^{(k+1-m)p/n-1}\, \|F_\tau^{-1}\|^{mp} |\tilde{u}|^p_{k+1,p,\tilde{\tau}}.$$

It follows from Lemmas 2.2 and 2.3 that

$$\sum_{|\alpha|=k+1} |\partial^\alpha \tilde{u}(\tilde{\boldsymbol{x}})| \;\le\; c\,|\|\tilde{\nabla}^{k+1}\tilde{u}(\tilde{\boldsymbol{x}})|\| \;\le\; c\,\|F_\tau' Q(\boldsymbol{x}) F_\tau\|^{\frac{k+1}{2}}.$$

Hence

$$\begin{aligned}|u-\Pi_k u|^p_{m,p,\tau} &\le c|\tau|\,|\tilde{\tau}|^{(k+1-m)p/n-1}\,\|F_\tau^{-1}\|^{mp}\int_\tau \|F_\tau' Q(\boldsymbol{x})F_\tau\|^{\frac{(k+1)p}{2}}\cdot\det\left(\frac{\partial\tilde{\boldsymbol{x}}}{\partial\boldsymbol{x}}\right)\,d\boldsymbol{x}\\ &\le c|\tilde{\tau}|^{(k+1-m)p/n}\,\|F_\tau^{-1}\|^{mp}\int_\tau \|F_\tau' Q(\boldsymbol{x})F_\tau\|^{\frac{(k+1)p}{2}}\,d\boldsymbol{x}.\end{aligned}$$

Now we write F_τ on the right-hand side of the above inequality in terms of $M(\boldsymbol{x})$ directly. For each $\boldsymbol{x}\in\tau$, decompose $M(\boldsymbol{x})$ into its eigenvalues and eigenvectors as

$$M(\boldsymbol{x}) = T(\boldsymbol{x})\cdot D(\boldsymbol{x})\cdot T'(\boldsymbol{x}), \tag{15}$$

where $D(\boldsymbol{x})$ is the diagonal matrix composed of all the eigenvalues $d_1(\boldsymbol{x})\le d_2(\boldsymbol{x})\le\cdots\le d_n(\boldsymbol{x})$ and T is the orthogonal matrix composed of all the eigenvectors. Define also $F(\boldsymbol{x}) = T(\boldsymbol{x})D(\boldsymbol{x})^{-\frac{1}{2}}$ as in (2). Then by Assumption A about the smoothness of M, we have from Lemma 2.1 that

$$\begin{aligned}\|F_\tau' Q(\boldsymbol{x})F_\tau\| &\le \|(F^{-1}(\boldsymbol{x})F_\tau)'\cdot(F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x}))\cdot(F^{-1}(\boldsymbol{x})F_\tau)\|\\ &\le \|F^{-1}(\boldsymbol{x})F_\tau\|^2\cdot\|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\|\\ &\le c\|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\|\end{aligned}$$

and

$$\|F_\tau^{-1}\| \le \|F_\tau^{-1}F(\boldsymbol{x})\|\cdot\|F^{-1}(\boldsymbol{x})\| \le c\|F^{-1}(\boldsymbol{x})\|.$$

Therefore,

$$|u-\Pi_k u|^p_{m,p,\tau} \le c|\tilde{\tau}|^{(k+1-m)p/n}\int_\tau \|F^{-1}(\boldsymbol{x})\|^{mp}\|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\|^{\frac{(k+1)p}{2}}\,d\boldsymbol{x}. \tag{16}$$

Summing up the above inequality for all $\tau\in\mathcal{T}_N$, we find that

$$\begin{aligned}&\sum_{\tau\in\mathcal{T}_N}|u-\Pi_k u|^p_{m,p,\tau}\\ &\quad\le c\left(\max_{\tau\in\mathcal{T}_N}|\tilde{\tau}|\right)^{(k+1-m)p/n}\int_\Omega \|F^{-1}(\boldsymbol{x})\|^{mp}\|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\|^{\frac{(k+1)p}{2}}\,d\boldsymbol{x}.\end{aligned} \tag{17}$$

Now we estimate $|\tilde{\tau}|$. Note that $\tilde{\tau} = F_\tau^{-1}(\tau-\boldsymbol{x}_c)$ and $\det(F_\tau^{-1}) = |\det(M_\tau)|^{1/2}$. By the assumption that $\{\mathcal{T}_N\}$ is quasi-uniform under metric M, the sizes of all $\tilde{\tau}$'s are of the same order. Hence,

$$\begin{aligned}\max_{\tau\in\mathcal{T}_N}|\tilde{\tau}| &\le cN^{-1}\sum_{\tau\in\mathcal{T}_N}|\tilde{\tau}| = cN^{-1}\textstyle\sum_{\tau\in\mathcal{T}_N}\displaystyle\int_\tau|\det(F_\tau^{-1})|\,d\boldsymbol{x}\\ &= cN^{-1}\textstyle\sum_{\tau\in\mathcal{T}_N}\displaystyle\int_\tau|\det(M_\tau)|^{1/2}\,d\boldsymbol{x}\\ &\le cN^{-1}\textstyle\sum_{\tau\in\mathcal{T}_N}\displaystyle\int_\tau|\det(M(\boldsymbol{x}))|^{1/2}\,d\boldsymbol{x}\\ &= cN^{-1}\int_\Omega|\det(M(\boldsymbol{x}))|^{1/2}\,d\boldsymbol{x},\end{aligned}$$

where in the last inequality we used (5) in Lemma 2.1. Putting the above inequality into (17), we have the error estimate (12).

Next, we consider for what metric M the error bound on the right-hand side of (12) is the smallest. We determine M through its eigenvalues and eigenvectors. Since they are independent of each other, we proceed in three steps as follows. First, for any $\boldsymbol{x} \in \Omega$ and a given set of eigenvalues of $M(\boldsymbol{x})$, we determine the orthogonal matrix $T(\boldsymbol{x})$ so that the integrands on the right-hand side of (12) are the smallest possible. Then we determine the ratios among the eigenvalues to further minimize the integrands. Finally, the optimal distribution of $\det(M)$ on Ω is determined such that the error bound in (12) is minimized.

First, for fixed $\boldsymbol{x}$ we write $Q(\boldsymbol{x})$ in its eigen-decomposition form,

$$Q(\boldsymbol{x}) = S(\boldsymbol{x}) \cdot \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \cdot S'(\boldsymbol{x}),$$

where $0 < \lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ are eigenvalues of $Q(\boldsymbol{x})$ and S is the orthogonal matrix composed of all its eigenvectors. Applying Lemma 2.4 on the inertia properties of matrix eigenvalues (with $A = Q(x)$, $S = F(x)$, $\alpha_i = \lambda_{n+1-i}$, and $s_i = (d_i)^{-1/2}$), we have

$$\|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\| \;=\; \lambda_{\max}(F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})) \;\ge\; \max_{1\le i\le n}\{\lambda_i/d_i\}.$$

Moreover, the equality in the above relation is attained when $T(\boldsymbol{x}) = S(\boldsymbol{x})$. Thus, we conclude that

$$\min_{\forall T} \|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\| = \max_{1\le i\le n}\{\lambda_i/d_i\},$$

and the minimum value is attained at $T(\boldsymbol{x}) = S(\boldsymbol{x})$. Since $\|F^{-1}(\boldsymbol{x})\| = (d_n)^{1/2}$ is independent of T, then for the integrand $\|F^{-1}(\boldsymbol{x})\|^{mp}\|F'(\boldsymbol{x})Q(\boldsymbol{x})F(\boldsymbol{x})\|^{\frac{(k+1)p}{2}}$, this choice of $T(\boldsymbol{x})$ is also optimal, with the minimum value

$$(d_n)^{\frac{mp}{2}} \left[\max_{1\le i\le n}\{\lambda_i/d_i\}\right]^{\frac{(k+1)p}{2}}. \tag{18}$$

Now we consider minimizing (18) with respect to the ratios among eigenvalues of $M(\boldsymbol{x})$. For fixed $\det(M) = \Pi_{i=1}^n d_i$, it follows from Lemma 2.5 that the expression in (18) achieves its minimum if and only if

$$\lambda_i/d_i = \frac{1}{\mu} \qquad \forall i = 1, 2, \ldots, n,$$

where μ depends on $\boldsymbol{x}$ and will be determined later. This implies that $d_i = \mu \cdot \lambda_i$ for all i, and

$$M(\boldsymbol{x}) \;=\; \mu(\boldsymbol{x}) \cdot Q(\boldsymbol{x}). \tag{19}$$

The minimum value of (18) is

$$\mu^{-\frac{(k+1-m)p}{2}}(\lambda_n)^{\frac{mp}{2}}. \tag{20}$$

This is the smallest possible value of the integrand in (17) at $\boldsymbol{x}$ for all possible $T(\boldsymbol{x})$ and different ratios among $d_i(\boldsymbol{x})$, $i = 1, 2, \ldots, n$. With this optimal choice, the error

estimate (12) becomes

$$\sum_{\tau\in\mathcal{T}_N} |u-\Pi_k u|^p_{m,p,\tau} \tag{21}$$
$$\leq\ cN^{-(k+1-m)p/n}\left\{\int_\Omega \mu^{\frac{n}{2}}\,|\det(Q)|^{\frac{1}{2}}\right\}^{(k+1-m)p/n}\cdot\int_\Omega \mu^{-\frac{(k+1-m)p}{2}}(\lambda_n)^{\frac{mp}{2}}.$$

Finally, we determine the distribution of μ on Ω to minimize the right-hand side of the above error estimate. Let $\alpha=(k+1-m)p/n$, and let

$$f=\mu^{\frac{n}{2}}\cdot|\det(Q)|^{\frac{1}{2}},$$
$$g=\mu^{-\frac{(k+1-m)p}{2}}\cdot(\lambda_n)^{\frac{mp}{2}}.$$

Then the integrals on the right-hand side of (21) are $[\int f]^\alpha[\int g]$. It follows from Hölder's inequality that

$$\begin{aligned}\left\{\left[\int f\right]^\alpha\cdot\left[\int g\right]\right\}^{1/(\alpha+1)} &= \left[\int f\right]^{\alpha/(\alpha+1)}\cdot\left[\int g\right]^{1/(\alpha+1)}\\ &= \|f^{\frac{\alpha}{\alpha+1}}\|_{L^{(\alpha+1)/\alpha}}\cdot\|g^{\frac{1}{\alpha+1}}\|_{L^{\alpha+1}}\\ &\geq \int f^{\frac{\alpha}{\alpha+1}}\cdot g^{\frac{1}{\alpha+1}},\end{aligned}$$

and the equality holds if and only if f is a constant multiple of g. In this case

$$\mu=c(\lambda_n)^{\frac{mp}{(k+1-m)p+n}}|\det(Q)|^{-\frac{1}{(k+1-m)p+n}}, \tag{22}$$

and the metric M becomes

$$M=\mu\cdot Q=c(\lambda_n)^{\frac{mp}{(k+1-m)p+n}}|\det(Q)|^{-\frac{1}{(k+1-m)p+n}}\cdot Q.$$

With this optimal choice of metric M, we have the smallest error bound

$$\sum_{\tau\in\mathcal{T}_N} |u-\Pi_k u|^p_{m,p,\tau} \leq cN^{-\frac{(k+1-m)p}{n}}\left[\int\mu^{\frac{n}{2}}|\det(Q)|^{\frac{1}{2}}\right]^{\alpha+1} \tag{23}$$
$$\leq cN^{-\frac{(k+1-m)p}{n}}\left\{\int_\Omega(\lambda_n)^{\frac{mp}{2(\alpha+1)}}\,|\det(Q)|^{\frac{\alpha}{2(\alpha+1)}}\right\}^{\alpha+1}.$$

Let $\beta=\frac{p}{\alpha+1}=\frac{np}{(k+1-m)p+n}$. Then we may write the above inequality in the following form:

$$\left(\sum_{\tau\in\mathcal{T}_N} |u-\Pi_k u|^p_{m,p,\tau}\right)^{1/p} \leq cN^{-\frac{k+1-m}{n}}\left\{\int_\Omega[\,(\lambda_n)^{\frac{m}{2}}|\det(Q)|^{\frac{\alpha}{2p}}\,]^\beta\right\}^{1/\beta} \tag{24}$$
$$= cN^{-\frac{k+1-m}{n}}\left\|(\lambda_n)^{\frac{m}{2}}|\det(Q)|^{\frac{k+1-m}{2n}}\right\|_{L^{np/[(k+1-m)p+n]}(\Omega)}.$$

This completes the proof of the theorem. □

Remark 2.1. Theorem 2.1 covers the error estimate on isotropic meshes as a special case. Indeed, if we choose $Q(\boldsymbol{x}) = c\,|\|\nabla^{k+1}u(\boldsymbol{x})\|\|^{\frac{2}{k+1}} \cdot I_n$, then the optimal metric $M_{k+1,m}$ based on Q becomes $c\,|\|\nabla^{k+1}u\|\|^{\frac{2p}{(k+1-m)p+n}} \cdot I_n$, and that $\{\mathcal{T}_N\}$ is quasi-uniform under $M_{k+1,m,p}$ implies it is isotropic. In this case, error estimate (14) is reduced to

$$\left(\sum_{\tau\in\mathcal{T}_N} |u-\Pi_k u|^p_{m,p,\tau}\right)^{\frac{1}{p}} \;\le\; c\,N^{-\frac{k+1-m}{n}}\;|u|_{k+1,np/[(k+1-m)p+n],\Omega}.$$

The above error bound is sharper than estimate (11) with $q = p$, since $L^p(\Omega) \subset L^{\frac{np}{(k+1-m)p+n}}$ due to $m \le k$.

Remark 2.2. In the case of $k = 1$, namely for linear interpolation, we may choose $Q = abs(\nabla^2 u)+\delta\cdot I_n$, where $\delta > 0$ is a small parameter to avoid Q from being singular. In this case, the optimal metrics and error estimate with $m = 0$ (i.e., for L^p-error) stated in Theorem 2.1 are identical to those in [7] by Chen, Sun, and Xu. They are also identical to those in [18] by Huang. For the case of $k \ge 2$, error estimates and mesh metrics are also derived in [17, 18] based on the sum of the Hessians of $(k-1)$th partial derivatives. It is shown in our previous study [5] that metrics based on the sum of the Hessians can be problematic for general anisotropic meshes; see Remark 3 in [5] for details.

Remark 2.3. It is easy to see that if $\{\mathcal{T}_N\}$ is quasi-uniform under the optimal metric $M_{k+1,m,p}$, then estimate (16) for the $W^{m,p}(\tau)$-seminorm of the error is of the same magnitude on each element. In other words, $M_{k+1,m,p}$ is a matrix which makes the $W^{m,p}$-error evenly distributed over all elements. Therefore, the optimal metrics and meshes follow also the so-called equidistribution principle. This principle has been used extensively to justify the selection of optimal or nearly optimal meshes; see, e.g., [12, 18, 22].

3. Measuring the anisotropic behavior of $\nabla^{k+1}u(\boldsymbol{x})$.

3.1. Definition of anisotropic measure of $\nabla^{k+1}u(\boldsymbol{x})$. It is seen from Theorem 2.1 that a good interpolation error estimate relies on a proper matrix Q in Assumption B that characterizes the anisotropic behavior of $\nabla^{k+1}u$ at each point. In order to produce as tight as possible an error estimate, we need to measure as accurately as possible the anisotropic behavior of $\nabla^{k+1}u$. Note that the optimal error bound (14) is determined by the function $|\lambda_{max}(Q(\boldsymbol{x}))|^{\frac{m}{2}}\,|\det(Q(\boldsymbol{x}))|^{\frac{k+1-m}{2n}}$, where $Q(\boldsymbol{x})$ is a positive definite matrix satisfying Assumption B. Therefore, to produce the tightest error bound, we choose matrix $Q(\boldsymbol{x})$ in (13) and (14) to be $Q_{k+1,m}$, the solution of the following minimization problem:

$$\min_{Q\in\mathcal{V}_{k+1}(\boldsymbol{x})} |\lambda_{max}(Q)|^{\frac{m}{2}}\,|\det(Q)|^{\frac{k+1-m}{2n}}, \tag{25}$$

where $\mathcal{V}_{k+1}(\boldsymbol{x})$ is the set of all $n\times n$ symmetric positive definite matrices Q that satisfy

$$|p_{k+1}(\boldsymbol{\xi})| = |(\boldsymbol{\xi}\cdot\nabla)^{k+1}u(\boldsymbol{x})| \le (\boldsymbol{\xi}\cdot Q\boldsymbol{\xi})^{\frac{k+1}{2}} \qquad \forall\boldsymbol{\xi}\in\mathcal{R}^n. \tag{26}$$

We call $Q_{k+1,m}$ an *anisotropic measure* of $\nabla^{k+1}u(\boldsymbol{x})$. It depends not only on k, the interpolation degree, but also on the index m, which is associated with the norm used to measure the error. In the special case of $m = 0$ (i.e., for L^p-error estimates), the

above problem is reduced to minimizing $\det(Q)$ for all $Q \in \mathcal{V}_{k+1}(\boldsymbol{x})$. Geometrically, for any symmetric positive definite matrix Q, the level curve/surface $\boldsymbol{\xi} \cdot Q\boldsymbol{\xi} = 1$ is an ellipse/ellipsoid of area/volume proportional to $|\det(Q)|^{-1/2}$, and $Q \in \mathcal{V}_{k+1}(\boldsymbol{x})$ implies that this level curve/surface is enclosed in that of $|p_{k+1}(\boldsymbol{\xi})| = 1$. Thus the solution $Q_{k+1,m}$ to (25) corresponds to the largest ellipse/ellipsoid (in area/volume) contained in $|p_{k+1}(\boldsymbol{\xi})| = 1$.

The above definition of $Q_{k+1,m}$ based on constrained minimization formulation is not quite convenient for theoretical study and practical use. We may change it into an unconstrained problem. For this purpose, we write the eigen-decomposition of each $Q \in \mathcal{V}_{k+1}(\boldsymbol{x})$ in the form

$$Q = \nu \cdot S\ \Lambda\ S'. \tag{27}$$

Without loss of generality, we assume $\nu = |\det(Q)|^{\frac{1}{n}}$ and

$$\Lambda = \mathrm{diag}(a_1, a_2, \ldots,\ a_{n-1}, [\Pi_{i=1}^{n-1} a_i]^{-1})$$

with $0 < a_1 \le a_2 \le \cdots \le a_{n-1} \le [\Pi_{i=1}^{n-1} a_i]^{-1}$. Clearly, minimization with respect to all Q is equivalent to minimization with respect to all a_i, $\nu > 0$, and all $n \times n$ orthogonal matrices S. Note that $\det(Q) = \nu^n$ and the largest eigenvalue of Q is $\lambda_{max}(Q) = \nu \cdot [\Pi_{i=1}^{n-1} a_i]^{-1}$. The objective function in (25) is indeed

$$|\lambda_{max}(Q)|^{\frac{m}{2}} \cdot |\det(Q)|^{\frac{k+1-m}{2n}} = \nu^{\frac{k+1}{2}} \cdot |\Pi_{i=1}^{n-1} a_i|^{-\frac{m}{2}},$$

and the constraint $Q \in \mathcal{V}_{k+1}(\boldsymbol{x})$ becomes

$$|p_{k+1}(\boldsymbol{\xi})| \ \le\ [\ \nu\ (S'\boldsymbol{\xi}) \cdot \Lambda(S'\boldsymbol{\xi})\ ]^{\frac{k+1}{2}} \qquad \forall \boldsymbol{\xi} \in \mathcal{R}^n.$$

This condition is equivalent to

$$|p_{k+1}(S\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})| \ \le\ \nu^{\frac{k+1}{2}} \cdot \|\boldsymbol{\eta}\|^{k+1} \qquad \forall \boldsymbol{\eta} \in \mathcal{R}^n.$$

Since p_{k+1} is a homogeneous polynomial of degree $k+1$, it is further equivalent to

$$|p_{k+1}(S\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})| \le \nu^{\frac{k+1}{2}} \qquad \forall \boldsymbol{\eta} \in \mathcal{R}^n \text{ with } \|\boldsymbol{\eta}\| = 1.$$

Hence we conclude that the following ν value is optimal:

$$\nu = \left[\ \max_{\|\boldsymbol{\eta}\|=1} |p_{k+1}(S\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})|\ \right]^{\frac{2}{k+1}}, \tag{28}$$

with which the constraint $Q \in \mathcal{V}_{k+1}(\boldsymbol{x})$ is automatically satisfied. Now the constrained minimization problem (25) is reduced to finding the minimum of

$$(\Pi_{i=1}^{n-1} a_i)^{-\frac{m}{2}} \cdot \max_{\|\boldsymbol{\eta}\|=1} |p_{k+1}(S\Lambda^{-\frac{1}{2}}\eta)| \tag{29}$$

with respect to all $0 < a_1 \le a_2 \le \cdots \le a_{n-1} \le [\Pi_{i=1}^{n-1} a_i]^{-1}$ and all orthogonal matrices S.

It is possible that the minimization problem (29) does not have a solution, or problem (25) has a solution with some $\lambda_i = 0$. In this case, Q becomes singular, and the optimal metrics $M_{k+1,m,p}$ based on it will lead to elements of infinitely large aspect

ratio. In order to avoid such a degenerate mesh metric in practice, we put a cap on the ratios of the eigenvalues of Q. More precisely, we may restrict $\lambda_{min}(Q)/\lambda_{max}(Q) \geq \delta$, where $\delta \in (0,1]$ is a user-specified parameter. This requirement is guaranteed when $a_i \in [\delta^{\frac{1}{n}}, \delta^{-\frac{n-1}{n}}]$ for all $1 \leq i \leq n-1$. Therefore, we may seek the minimizer to (29) over all $\delta^{\frac{1}{n}} \leq a_1 \leq a_2 \leq \cdots \leq a_{n-1} \leq [\Pi_{n=1}^{n-1}]^{-1} \leq \delta^{-\frac{n-1}{n}}$ and all orthogonal S. Since the set of all these a_i's and T's is compact, and the objective function in (29) is continuous with respect to its variables, a positive definite minimizer $Q_{k+1,m}$ is guaranteed to exist for any specified $0 < \delta \leq 1$.

Examples. We present two examples to explain the definition of $Q_{k+1,m}$. First we consider the simplest case, $n = 2, k = 1$, which corresponds to linear interpolation in $\mathcal{R}^2$. Without loss of generality, suppose

$$\nabla^2 u = R_\phi \cdot \begin{bmatrix} \mu_1 & \\ & \mu_2 \end{bmatrix} R_\phi^T$$

with $|\mu_1| \leq |\mu_2|$, where R_ϕ is the matrix of rotation by angle ϕ counterclockwise. Let

$$Q = \nu \cdot R_\psi \cdot \begin{bmatrix} a & \\ & a^{-1} \end{bmatrix} R_\psi^T$$

with $0 < a \leq a^{-1}$ or $a \in (0,1]$. Then $Q_{k+1,m}$ is defined by the solution (a_*, ψ_*) to the following problem:

$$\min_{a \in (0,1], \psi \in [0,2\pi]} \left\{ a^{-\frac{m}{2}} \cdot \max_{\|\boldsymbol{\eta}\|=1} |p_2(R_\psi \Lambda^{-\frac{1}{2}} \boldsymbol{\eta})| \right\}. \tag{30}$$

Note that

$$\max_{\|\boldsymbol{\eta}\|=1} |p_2(R_\psi \Lambda^{-\frac{1}{2}} \boldsymbol{\eta})| \;=\; \max_{\|\boldsymbol{\xi}\|=1} \frac{|p_2(R_\psi \boldsymbol{\xi})|}{\|\Lambda^{\frac{1}{2}} \boldsymbol{\xi}\|^2}.$$

Using the polar coordinates for $\boldsymbol{\xi} \in \mathcal{R}^2$, the objective function of (30) can be written as $a^{-\frac{m}{2}} \cdot \max_{t \in [0,2\pi]} J(a, \psi, t)$, where

$$J(a, \psi, t) \equiv \left| \frac{\mu_1 \cos^2(t - \phi + \psi) + \mu_2 \sin^2(t - \phi + \psi)}{a \cdot \cos^2 t + a^{-1} \cdot \sin^2 t} \right|. \tag{31}$$

When $\mu_1 \cdot \mu_2 \geq 0$, we can show that

$$\max_{t \in [0,2\pi]} J(a, \psi, t) \geq \max_{t \in [0,2\pi]} J(a, \phi, t) \qquad \forall \psi \in [0, 2\pi]. \tag{32}$$

Indeed, it is easy to see that

$$\max_{t \in [0,2\pi]} J(a, \phi, t) = \max_{t \in [0,2\pi]} \left| \frac{\mu_1 \cos^2 t + \mu_2 \sin^2 t}{a \cdot \cos^2 t + a^{-1} \cdot \sin^2 t} \right| = \max(|\mu_1 \cdot a^{-1}|,\ |\mu_2 \cdot a|)$$

since $J(a, \phi, t)$ is a rational function of $\cos^2 t \in [0,1]$. On the other hand,

$$\max_{t \in [0,2\pi]} J(a, \psi, t) \geq J(a, \psi, 0) = \left| \frac{\mu_1 \cos^2(\phi - \psi) + \mu_2 \sin^2(\phi - \psi)}{a} \right| \geq |\mu_1 \cdot a^{-1}|$$

since $|\mu_1| \le |\mu_2|$, and similarly

$$\max_{t\in[0,2\pi]} J(a,\psi,t) \ge J\left(a,\psi,\phi-\psi+\frac{\pi}{2}\right)$$
$$= \left|\frac{\mu_2}{a\cdot\sin^2(\phi-\psi)+a^{-1}\cdot\cos^2(\phi-\psi)}\right| \ge |\mu_2\cdot a|$$

since $a \le 1$. Therefore, we conclude that

$$\min_{a\in(0,1],\psi\in[0,2\pi]}\left\{a^{-\frac{m}{2}}\cdot\max_{\|\boldsymbol{\eta}\|=1}|p_2(R_\psi\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})|\right\}$$
$$= \min_{a\in(0,1]}\left\{a^{-\frac{m}{2}}\cdot\max(|\mu_1\cdot a^{-1}|,|\mu_2\cdot a|)\right\}, \tag{33}$$

where the minimum is attained at $\psi = \phi$. Furthermore, since $m \le k = 1$, the minimizer for the right-hand side of the above equation is

$$a = \sqrt{\left|\frac{\mu_1}{\mu_2}\right|},$$

which implies by (28) that

$$\nu = \sqrt{|\mu_1\cdot\mu_2|} \tag{34}$$

and

$$Q_{k+1,m} = R_\phi\cdot\begin{bmatrix}|\mu_1| & \\ & |\mu_2|\end{bmatrix}R_\phi^T = abs(\nabla^2 u). \tag{35}$$

In the case of $\mu_1\cdot\mu_2 < 0$, (32) is not true in general. However, our numerical calculation shows that in this case the minimum of (30) is still attained at $a = \sqrt{|\mu_1/\{\mu_2\}|}$ and $\psi = \phi$, which results in the same ν and Q as in (34) and (35).

This example confirms from another perspective that the conventional choice is optimal by using the eigenvalues and eigenvectors of the Hessian matrix to characterize its anisotropic behavior in mesh generation and refinement for linear interpolation or linear elements; see [3, 4, 7, 8, 11, 13, 17, 18].

Our second example is for $u = \frac{1}{6}xy^2$ in $\mathcal{R}^2$ and $k = 2$ (quadratic interpolation). In this example, $p_3(\boldsymbol{\xi}) = \xi\eta^2$. Matrix Q can be determined as in (27) with a and ψ being the minimizer to the following problem:

$$\min_{a\in(0,1],\psi\in[0,2\pi]}\left\{a^{-\frac{m}{2}}\cdot\max_{\|\boldsymbol{\eta}\|=1}|p_3(R_\psi\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})|\right\}. \tag{36}$$

For each $a > 0$, it can be shown numerically that

$$\max_{\|\boldsymbol{\eta}\|=1}|p_3(\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})| \le \max_{\|\boldsymbol{\eta}\|=1}|p_3(R_\psi\Lambda^{-\frac{1}{2}}\boldsymbol{\eta})| \qquad \forall\psi\in[0,2\pi].$$

Thus $\psi = 0$ is an optimal solution, with minimum value $a^{-(m-1)/2}\max_t|\cos t\cdot\sin^2 t| = \frac{2\sqrt{3}}{9}a^{-(m-1)/2}$. In this case, problem (36) is reduced to $\min_{a\in(0,1]}\frac{2\sqrt{3}}{9}a^{-(m-1)/2}$. For $m = 0$, which corresponds to L^p-error estimates, the optimal a is 0; while for $m = 1$

(corresponding to H^1-error estimates), any positive $a \leq 1$ is optimal. The optimal ν value is $\frac{\sqrt[3]{4}}{3}$.

Applying Theorem 2.1 to the quadratic interpolation of $u = \frac{1}{6}xy^2$, we see that the L^2-error bound can be driven to 0 if we increase the aspect ratio of the elements while keeping their area fixed and alignment direction fixed along the x-axis. The H^1-error bound does not change (since $|\lambda_{max}(Q)|^{\frac{m}{2}} |\det(Q)|^{\frac{k+1}{2}}$ is a constant independent of a), which implies that using highly anisotropic triangles does not help reduce the H^1-error. This conclusion coincides with the study in [5] based on the exact formula for the quadratic interpolation errors; see Remark 2 in [5].

3.2. Estimate of the anisotropic measure: A dimension reduction method. The definition of the anisotropic measure for $\nabla^{k+1}u$ in the previous subsection involves a nonlinear minimization (25) with respect to the matrix Q. Solving this problem could be expensive in practice. Here we describe a method to find a suboptimal solution to the minimization problem, and give an approximate Q for the anisotropic measure.

Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the eigenvalues (in ascending order) of Q, and $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ be the corresponding eigenvectors. Notice that the objective function in (25), $|\lambda_{max}(Q)|^{\frac{m}{2}} |\det(Q)|^{\frac{k+1-m}{2n}}$, is the product of the eigenvalues of Q. We determine an approximate Q by choosing successively $(\lambda_i, \boldsymbol{v}_i)$ for $i = n, n-1, \ldots, 1$, such that each λ_i is the smallest possible to have constraint (26) hold. More precisely, we first choose λ_n as

$$\lambda_n = |||\nabla^{k+1}u|||^{\frac{2}{k+1}} = \left[\max_{\|\boldsymbol{\xi}\|=1} |p_{k+1}(\boldsymbol{\xi})| \right]^{\frac{2}{k+1}};$$

i.e., $|\lambda_n|^{\frac{k+1}{2}}$ is the largest $(k+1)$th-order directional derivative of u. The corresponding eigenvector $\boldsymbol{v}_n$ is chosen as the unit vector along which the $(k+1)$th directional derivative is the largest, i.e.,

$$\boldsymbol{v}_n = \arg \max_{\|\boldsymbol{\xi}\|=1} |p_{k+1}(\boldsymbol{\xi})|.$$

To determine the rest $n-1$ eigenpairs, let $\tilde{\boldsymbol{v}}_1, \ldots, \tilde{\boldsymbol{v}}_{n-1}$ be a set of orthonormal bases for the orthogonal complement of $span\{\boldsymbol{v}_n\}$ in $\mathcal{R}^n$. Then any $\boldsymbol{\xi} \in \mathcal{R}^n$ can be expressed as

$$\boldsymbol{\xi} = \sum_{i=1}^{n-1} \zeta_i \tilde{\boldsymbol{v}}_i + z\boldsymbol{v}_n$$

for some $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_{n-1}]' \in \mathcal{R}^{n-1}$ and $z \in \mathcal{R}$. Furthermore, let

$$T_{n-1} = \begin{bmatrix} \tilde{\boldsymbol{v}}_1' \\ \vdots \\ \tilde{\boldsymbol{v}}_{n-1}' \end{bmatrix} \cdot [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{n-1}], \qquad D_{n-1} = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_{n-1} \end{bmatrix},$$

and

$$Q_{n-1} = T_{n-1} D_{n-1} T_{n-1}'.$$

Then

$$\boldsymbol{\xi} \cdot Q\boldsymbol{\xi} = \boldsymbol{\zeta} \cdot Q_{n-1}\boldsymbol{\zeta} + \lambda_n z^2,$$

and constraint (26) is reduced to

$$\boldsymbol{\zeta}\cdot Q_{n-1}\boldsymbol{\zeta} \;\geq\; h(\boldsymbol{\zeta},z) \;\equiv \left|p_{k+1}\left(\sum_{i=1}^{n-1}\zeta_i\tilde{\boldsymbol{v}}_i + z\boldsymbol{v}_n\right)\right|^{\frac{2}{k+1}} - \lambda_n z^2 \qquad \forall\boldsymbol{\zeta}\in\mathcal{R}^{n-1}, \forall z\in\mathcal{R}.$$

This condition can be expressed equivalently as

$$\boldsymbol{\zeta}\cdot Q_{n-1}\boldsymbol{\zeta} \;\geq\; g(\boldsymbol{\zeta}) \;\equiv\; \max_{z\in\mathcal{R}} h(\boldsymbol{\zeta},z) \qquad \forall\boldsymbol{\zeta}\in\mathcal{R}^{n-1}. \tag{37}$$

It is easy to verify that $g(\boldsymbol{\zeta})$ is a homogeneous function of $\boldsymbol{\zeta}$. Indeed, for any $t\in\mathcal{R}$,

$$g(t\boldsymbol{\zeta}) \;=\; \max_{z\in\mathcal{R}} h(t\boldsymbol{\zeta},z) \;=\; \max_{z\in\mathcal{R}} h(t\boldsymbol{\zeta},tz) \;=\; \max_{z\in\mathcal{R}} t^2 h(\boldsymbol{\zeta},z) \;=\; t^2 g(\boldsymbol{\zeta}).$$

Therefore, to determine Q_{n-1} under constraint (37) is similar to the original problem to determine Q under constraint (26), except the former is of one dimension less than the later. We may repeat this process $n-1$ times to arrive at a one-dimensional problem, whose solution is ready to work out.

In practice, the evaluation of $g(\boldsymbol{\zeta})$ in the constraint (37) can be carried out by checking the critical points of $h(\boldsymbol{\zeta},z)$ in the z direction. Indeed, for given $\boldsymbol{\zeta}$,

$$\frac{\partial}{\partial z}p_{k+1}(\boldsymbol{\xi}) = (k+1)(\boldsymbol{\xi}\cdot\nabla)^k(\tilde{\boldsymbol{v}}_n\cdot\nabla)u.$$

Thus the critical points (excluding those making $p_{k+1}=0$, which are clearly not among the maxima of $h(\boldsymbol{\zeta},z)$) satisfy the following equation:

$$|p_{k+1}(\boldsymbol{\xi})|^{\frac{1-k}{k+1}}(\boldsymbol{\xi}\cdot\nabla)^k(\tilde{\boldsymbol{v}}_n\cdot\nabla)u = \lambda_n z.$$

It is equivalent to

$$\left[\ (\boldsymbol{\xi}\cdot\nabla)^k(\tilde{\boldsymbol{v}}_n\cdot\nabla)u\ \right]^{k+1} = (\lambda_n z)^{k+1}\cdot\left[\ (\boldsymbol{\xi}\cdot\nabla)^{k+1}u\ \right]^{k-1},$$

which is a polynomial equation for z of degree $k(k+1)$.

Remark 3.1. For two-dimensional problems ($n=2$), the author developed in [5, 6] a method to define a matrix Q_{k+1} characterizing the anisotropic behavior $\nabla^{k+1}u$ by using the factors of polynomial $p_{k+1}(\boldsymbol{\xi})$ for directional derivatives. For the cases of $k=1,2$, it can be shown that Q_{k+1} given in [5, 6] is equivalent to the matrix Q produced by the dimension reduction algorithm described here. For $k\geq 3$, we believe they are still equivalent. However, it is yet to be verified.

4. Numerical results. In this section, we present some numerical results to compare the error in various norms for interpolations based on anisotropic meshes generated with the optimal metric $M_{k+1,m,p}$ developed in Theorem 2.1.

Two-dimensional example. We consider linear and quadratic interpolations of the following function on $\Omega=[0,1]^2$:

$$u(x,y) = x^2+y^2+x^3+y^3+\exp(-K\ (y-d_1(x))^2)+\exp(-K\ (y-d_2(x))^2), \tag{38}$$

where $K=10000$ and

$$d_1(x) = -x(x-a)/[2(1-a)]+1, \qquad \text{with } a=1.25;$$
$$d_2(x) = (x-b)(x-c)/[2(1-b)(1-c)], \qquad \text{with } b=0.2, c=1.25.$$

TABLE 1

The error in various norms for linear (in big font) and quadratic (in small font) interpolations of function (38) *based on meshes generated under metrics* $M_{2,m,p}$. N_e *and* N_v *represent the total number of elements and nodes, respectively.*

	N_e	N_v	Metric	$\|e\|_{L^1}$	$\|e\|_{L^2}$	$\|e\|_{L^\infty}$	$\|e\|_{H^1}$
$N_e \approx 1,000$	1030	550	$M_{2,0,1}$	**8.55865e-03**	2.25843e-02	5.65950e-01	1.00987e+01
				7.00047e-04	3.38558e-03	2.82878e-01	3.12289e+00
	1021	542	$M_{2,0,2}$	1.00819e-02	**1.91599e-02**	4.45866e-01	8.39336e+00
				5.39823e-04	2.22625e-03	1.02203e-01	2.17432e+00
	1039	546	$M_{2,0,\infty}$	2.01321e-02	2.80857e-02	**2.00845e-01**	8.41199e+00
				6.22584e-04	1.94362e-03	8.06520e-02	2.00037e+00
	1017	535	$M_{2,1,2}$	2.85871e-02	4.33668e-02	6.68077e-01	**7.76100e+00**
				8.23694e-04	2.50331e-03	1.87979e-01	1.99802e+00
$N_e \approx 4,000$	3992	2069	$M_{2,0,1}$	**1.70237e-03**	4.38095e-03	1.96297e-01	4.08116e+00
				7.02603e-05	3.69513e-04	4.32572e-02	6.95222e-01
	4000	2063	$M_{2,0,2}$	2.10463e-03	**3.44617e-03**	7.64404e-02	3.41493e+00
				5.62782e-05	2.20319e-04	1.82189e-02	5.33361e-01
	4040	2068	$M_{2,0,\infty}$	6.66663e-03	8.14892e-03	**2.83157e-02**	3.08884e+00
				8.89956e-05	2.19451e-04	6.35742e-03	4.53298e-01
	4002	2047	$M_{2,1,2}$	1.27443e-02	1.74685e-02	1.58691e-01	**2.83627e+00**
				1.43980e-04	2.67353e-04	1.60196e-02	4.16979e-01
$N_e \approx 16,000$	15971	8136	$M_{2,0,1}$	**4.09208e-04**	2.80185e-03	4.46053e-01	2.44949e+00
				1.91768e-05	1.04155e-03	2.30520e-01	8.87463e-01
	16088	8171	$M_{2,0,2}$	4.91963e-04	**8.02090e-04**	5.75916e-02	1.54053e+00
				8.81760e-06	1.64703e-04	6.68060e-02	2.68500e-01
	15994	8088	$M_{2,0,\infty}$	1.79767e-03	2.06070e-03	**5.39628e-03**	1.37961e+00
				1.54309e-05	4.24226e-05	9.44125e-04	1.43448e-01
	16004	8088	$M_{2,1,2}$	4.44656e-03	5.47114e-03	9.68989e-02	**1.25693e+00**
				3.17102e-05	5.80640e-05	9.69865e-03	1.36711e-01

This function u has steep layers around two parabolas $y = d_i(x)$, $i = 1, 2$.

We calculate exactly all the second and third partial derivatives of u and determine matrix Q for measuring their anisotropic behaviors by using the dimension reduction algorithm described in section 3.2. Then we form the mesh metric $M_{k+1,m,p}$ according to (13), which is optimal for minimizing an upper bound of the $W^{m,p}$-norm of the interpolation error $e = u - \Pi_k u$. The constant multiple c in (13) is used to control the total number of elements.

We use the two-dimensional mesh generator `bamg` (bidimensional anisotropic mesh generator) developed by Borouchaki et al. [3] and Hecht [16] to create the anisotropic meshes. This package accepts a user-defined metric to create and refine an anisotropic mesh that is quasi-uniform under the given metric. We choose the following parameter setting in all our experiments:

```
"-NoRescaling -NbSmooth 5 -hmax 0.02 -hmin 0.0000005 -ratio 0
    -nbv 100000 -v 9"
```

In order to make the anisotropic mesh as uniform as possible under metric $M_{k+1,m,p}$, we call `bamg` iteratively with the metrics recalculated over the updated mesh. The final mesh is the one after 20 iterations for all the cases, where there is little change of the mesh and the interpolation error.

We consider specifically the linear ($k = 1$) and quadratic ($k = 2$) interpolations, and measure the errors in (i) L^1-norm ($m = 0, p = 1$), (ii) L^2-norm ($m = 0, p = 2$), (iii) L^∞-norm ($m = 0, p = \infty$), and (iv) H^1-seminorm ($m = 1, p = 2$). These error norms are calculated using numerical quadratures based on 7 and 28 Fekete points for

TABLE 2

The error in various norms for linear (in small font) and quadratic (in big font) interpolations of function (38) *based on meshes generated under metrics* $M_{3,m,p}$. N_e *and* N_v *represent the total number of elements and nodes, respectively.*

	N_e	N_v	Metric	$\|e\|_{L^1}$	$\|e\|_{L^2}$	$\|e\|_{L^\infty}$	$\|e\|_{H^1}$
$N_e \approx 1,000$	1026	542	$M_{3,0,1}$	1.19200e-02 **5.47899e-04**	2.02953e-02 2.63100e-03	4.01216e-01 1.41338e-01	8.96489e+00 2.59358e+00
	1016	534	$M_{3,0,2}$	1.84082e-02 5.84794e-04	2.70593e-02 **2.15282e-03**	4.59322e-01 1.06775e-01	8.95022e+00 2.42518e+00
	1019	536	$M_{3,0,\infty}$	2.92068e-02 8.38699e-04	4.64891e-02 2.23870e-03	2.47095e-01 **3.84038e-02**	8.85002e+00 2.16885e+00
	1000	526	$M_{3,1,2}$	2.86835e-02 8.41909e-04	4.47000e-02 2.57482e-03	3.66967e-01 1.39962e-01	8.01008e+00 **2.16436e+00**
$N_e \approx 4,000$	4002	2060	$M_{3,0,1}$	2.78079e-03 **4.10310e-05**	4.75471e-03 1.86169e-04	1.63081e-01 1.42471e-02	4.31546e+00 4.26067e-01
	3963	2033	$M_{3,0,2}$	4.95528e-03 4.69086e-05	6.45569e-03 **1.34854e-04**	1.04982e-01 6.66449e-03	4.07441e+00 3.53092e-01
	3991	2041	$M_{3,0,\infty}$	1.50754e-02 1.46617e-04	2.12624e-02 2.32106e-04	6.14308e-02 **2.95223e-03**	3.97760e+00 3.50702e-01
	4013	2051	$M_{3,1,2}$	1.90713e-02 2.04870e-04	2.86518e-02 3.38047e-04	1.09707e-01 1.18358e-02	3.94508e+00 **3.02695e-01**
$N_e \approx 16,000$	16038	8140	$M_{3,0,1}$	7.17854e-04 **4.04626e-06**	1.33199e-03 1.64799e-05	4.08086e-02 1.14591e-03	2.46107e+00 8.76529e-02
	16070	8136	$M_{3,0,2}$	1.28305e-03 5.01463e-06	1.70431e-03 **1.28541e-05**	2.82953e-02 6.64558e-04	2.36372e+00 7.34430e-02
	16058	8113	$M_{3,0,\infty}$	4.86946e-03 2.18202e-05	6.23055e-03 2.97789e-05	2.06393e-02 **2.82230e-04**	2.24020e+00 7.30433e-02
	15995	8080	$M_{3,1,2}$	8.42677e-03 4.88933e-05	1.14039e-02 7.11500e-05	3.99892e-02 1.58626e-03	2.21819e+00 **6.24718e-02**

linear and quadratic interpolations, respectively; see [25]. We list in Tables 1 and 2 the linear and quadratic interpolation errors in four norms, and display in Figures 2 and 3 the anisotropic meshes (of about 4,000 elements). Several observations are clearly at hand. (a) In all the cases, the smallest $W^{m,p}$-norm of the interpolation error is obtained when the mesh is generated according to the optimal metric $M_{k+1,m,p}$. This indicates not only the optimality of the metric $M_{k+1,m,p}$ stated in Theorem 2.1, but also that the matrix Q produced by the dimension reduction algorithm characterizes fairly accurately the anisotropic behavior of $\nabla^{k+1}u$. (b) When the number of elements is quadrupled (i.e., the element length scale is halved), the interpolation error is reduced by a factor 2^{k+1-m} as predicted in the error estimate. (c) The mesh metrics and the anisotropic meshes ideal for linear interpolation are not necessarily good for quadratic interpolation, and vice versa.

Three-dimensional example. We present here an example for the quadratic interpolation in three dimensions. Consider the following function on $\Omega = [0,1]^3$:

$$u(x,y,z) = x^3 + (\mu_1 y)^3 + (\mu_2 z)^3, \tag{39}$$

where $\mu_1 = 10, \mu_2 = 30$. For this function, $\nabla^3 u$ is constant over all Ω. By using the dimension reduction algorithm, it is easy to find the matrix Q that characterizes the anisotropic behavior of $\nabla^3 u$ as follows:

$$Q = \begin{bmatrix} 1 & & \\ & \mu_1 & \\ & & \mu_2 \end{bmatrix}.$$

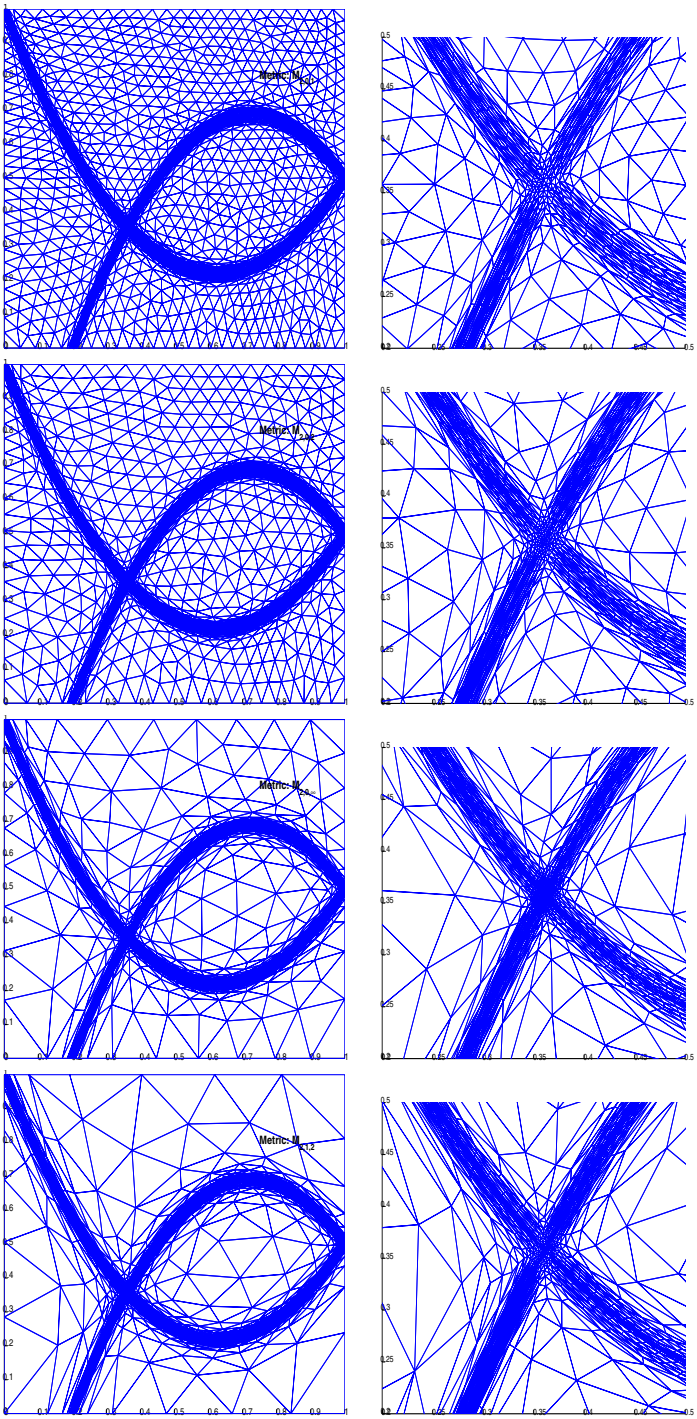


FIG. 2. *Anisotropic meshes that are quasi-uniform under respective metrics $M_{2,m,p}$ and their closeup views.*

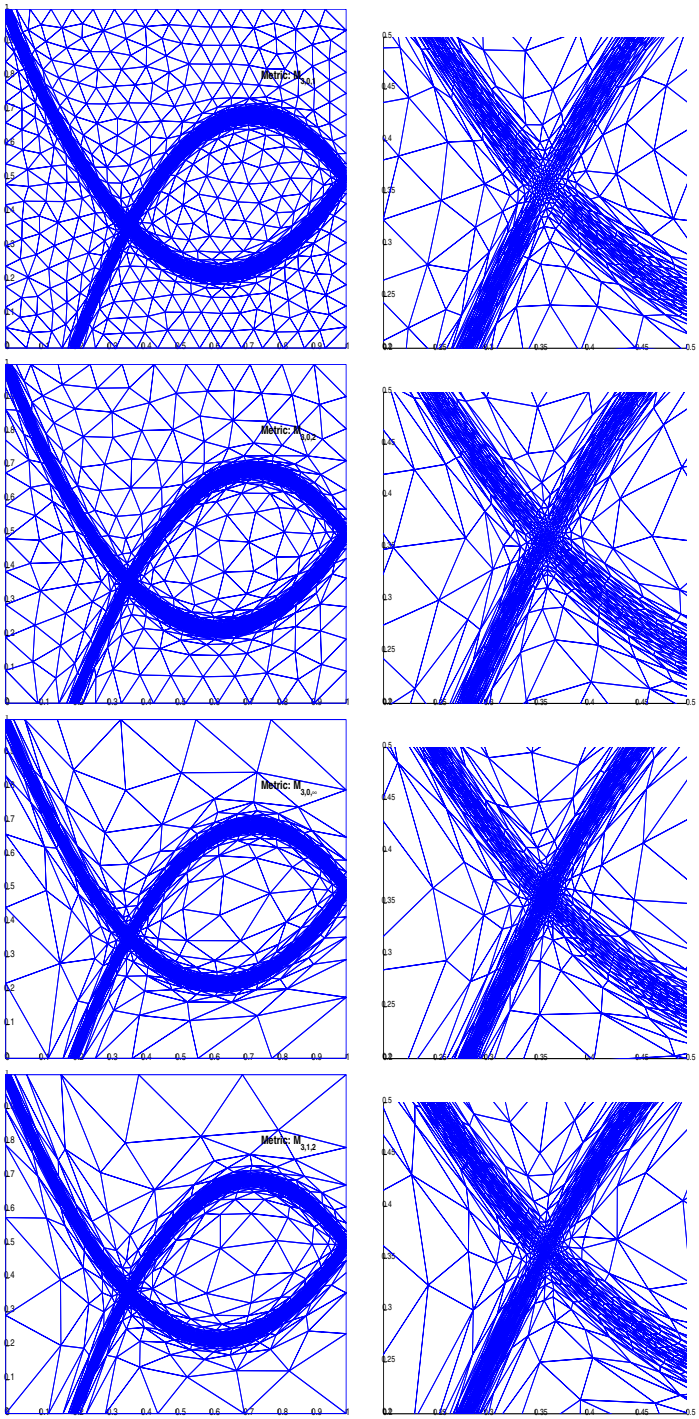


FIG. 3. *Anisotropic meshes that are quasi-uniform under respective metrics* $M_{3,m,p}$ *and their closeup views.*

The optimal mesh metric $M_{k+1,m,p}$ is a constant multiple of the above matrix. Thus in an anisotropic mesh that is quasi-uniform under $M_{k+1,m,p}$ all elements must be of about the same volume, and all elements must have approximately the same length scale $1 : \frac{1}{\mu_1} : \frac{1}{\mu_2}$ in x, y, and z directions. We create meshes of a specified length scale $1 : \frac{1}{L_y} : \frac{1}{L_z}$ in the following way. First we generate a quasi-uniform tetrahedral mesh using a three-dimensional Delaunay generator `gmsh` developed by Geuzaine and Remacle [14] over a rectangular box $[0,1] \times [0, L_y] \times [0, L_z]$. Then it is scaled in y and z directions by factors L_y and L_z to obtain the anisotropic mesh on Ω. When $L_x = \mu_1$ and $L_y = \mu_2$, the obtained mesh is quasi-uniform under $M_{k+1,m,p}$. This mesh should also be the minimizer for the $W^{m,p}$-error of interpolation Π_k (in this example, $M_{k+1,m,p}$ is identical for all m, k, and p). We calculate the error norms by using a quadrature formula with 24 points supplied by [10], which is exact for numerical integration of polynomials of degree less than or equal to 6. We vary L_y between $1 \sim 20$ and L_z between $10 \sim 40$ to obtain meshes of different aspect ratios, while keeping the total number of elements around $40,000$ (by setting the characteristic length `lc` in `gmsh` to be $0.05\sqrt[3]{L_y L_z}$).

We display in Figure 4 the error contour plots against L_y and L_z in the cases of

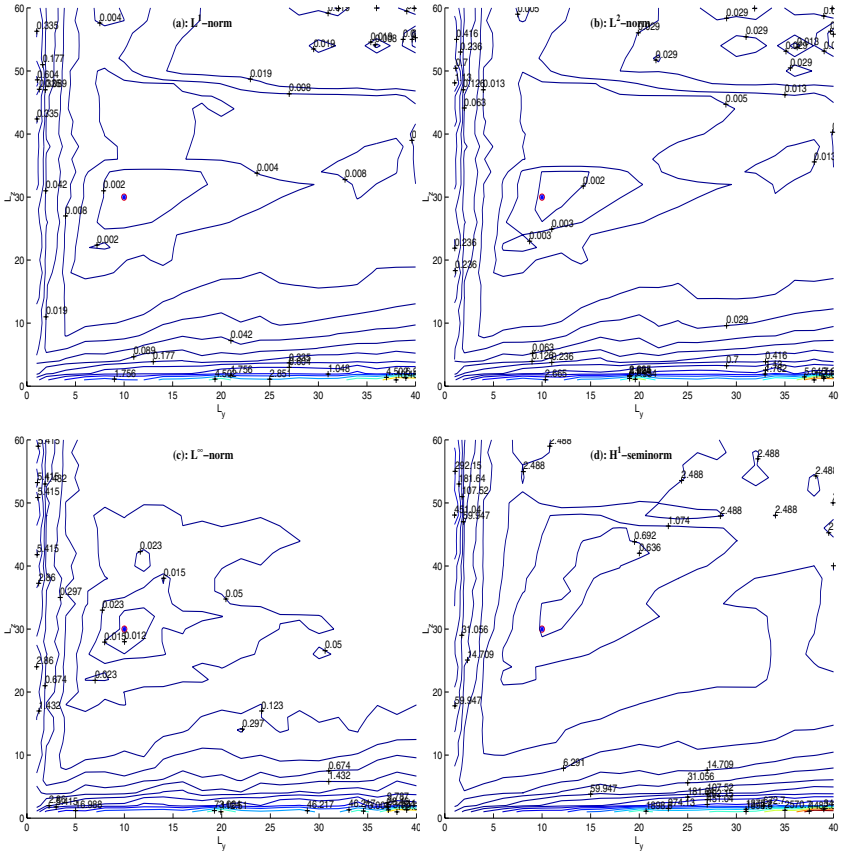


FIG. 4. *Contour plot of the* (a) L^1, (b) L^2, (c) L^∞, *and* (d) H^1 *errors with respect to various element aspect ratios* $1 : \frac{1}{L_y} : \frac{1}{L_z}$. *The star-circle mark at* $(10, 30)$ *indicates the optimal aspect ratio predicted by the error estimates.*

$m = 0, p = 1, 2, \infty$, and $m = 1, p = 2$. It is noted that the smallest interpolation error is achieved when the element aspect ratio is approximately equal to $1 : \frac{1}{\mu_1} : \frac{1}{\mu_2}$ in all four cases. This indicates the optimality of the metric $M_{k+1,m,p}$. It is also noted that the error is relatively insensitive to variation of the length scales in y and z directions when they are close to the optimal values. We believe this is because the anisotropic behavior of $\nabla^3 u$ in this example is relatively "mild." For interpolated functions of stronger anisotropic behaviors, the improvement by using the optimal mesh metrics can be more drastic. Due to the lack of a reliable anisotropic mesh generator in three dimensions, we are unable to test the optimality of $M_{k+1,m,p}$ for u with variable $\nabla^3 u$.

5. Conclusion and discussions. In the previous sections, we presented an error estimate for higher-order interpolations over anisotropic meshes in $\mathcal{R}^n$. It involves an interplay of the mesh features controlled by a given mesh metric and the anisotropic measures of the higher-order derivative tensors of interpolated functions. Based on the error estimate, we were able to identify the optimal mesh metrics leading to the smallest error bound for a subset of interpolated functions exhibiting similar anisotropic behaviors. Numerical results indicate that the meshes generated according to the optimal metrics produce the smallest interpolation error in the corresponding norms.

A critical component in applying the error estimate for anisotropic mesh generation or refinement is to measure the anisotropic behavior of higher-order derivative tensors. We define such a measure by the "largest" ellipse/ellipsoid contained in the level curve for directional derivatives. To avoid solving the minimization problem for defining the anisotropic measure, we developed a dimension reduction algorithm to produce the measure approximately.

The practical application of the results in this paper is always associated with the development of a reliable and efficient anisotropic mesh generator. While there have been many two-dimensional packages available (`bamg` is clearly among the best of them), general three-dimensional anisotropic meshing packages are yet to be developed and tested. Also, it is natural to apply the results in this paper to quadratic and higher-order finite element methods for solving PDEs. To this end, one needs to recover the higher-order derivatives of the PDEs solution from its numerical approximation. While there have been extensive studies along this direction for isotropic triangulations and linear elements [1, 20, 27, 28], the analysis and application for higher-order elements on anisotropic meshes are yet to be developed.

REFERENCES

[1] M. Ainsworth and J. T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*, Wiley Intersciences, New York, 2000.

[2] T. Apel, *Anisotropic Finite Elements: Local Estimates and Applications*, Adv. Numer. Math., Teubner, Stuttgart, 1999.

[3] H. Borouchaki, P. L. George, F. Hecht, P. Laug, and E. Saltel, *Delaunay mesh generation governed by metric specifications. I. Algorithms*, Finite Elem. Anal. Des., 25 (1997), pp. 61–83.

[4] W. Cao, *On the error of linear interpolation and the orientation, aspect ratio, and internal angles of a triangle*, SIAM J. Numer. Anal., 43 (2005), pp. 19–40.

[5] W. Cao, *Anisotropic measure of third order derivatives and the quadratic interpolation error on triangular elements*, SIAM J. Sci. Comput., 29 (2007), pp. 756–781.

[6] W. Cao, *An interpolation error estimate in $\mathcal{R}^2$ based on the anisotropic measures of higher order derivatives*, Math. Comp., to appear.

[7] L. Chen, P. Sun, and J. Xu, *Optimal anisotropic meshes for minimizing interpolation errors in L^p-norm*, Math. Comp., 79 (2007), pp. 179–204.

[8] L. Chen and J. Xu, *Optimal Delaunay triangulations*, J. Comput. Math., 22 (2004), pp. 299–308.
[9] P. G. Ciarlet, *The Finite Element Methods for Elliptic Problems*, Classics Appl. Math. 40, SIAM, Philadelphia, 2002.
[10] R. Cools, *An encyclopaedia of cubature formulas*, J. Complexity, 19 (2003), pp. 445–453.
[11] E. F. D'Azevedo and R. B. Simpson, *On optimal triangular meshes for minimizing the gradient error*, Numer. Math., 59 (1991), pp. 321–348.
[12] C. de Boor, *Good approximation by splines with variable knots* II, in Conference on the Numerical Solution of Differential Equations, Lecture Notes Math. 363, Springer-Verlag, Berlin, 1974, pp. 12–20.
[13] L. Formaggia and S. Perotto, *New anisotropic a priori error estimates*, Numer. Math., 89 (2001), pp. 641–667.
[14] C. Geuzaine and J.-F. Remacle, *Gmsh: A Three-Dimensional Finite Element Mesh Generator with Built-in Pre- and Post-Processing Facilities*, Technical report, Dept. of Electrical Engineering and Computer Sciences, University of Liège, Liège, Belgium, 2005.
[15] W. G. Habashi, J. Dompierre, Y. Bourgault, D. Ait-Ali-Yahia, M. Fortin, and M. Vallet, *Anisotropic mesh adaptation: Towards user-independent, mesh-independent and solver-independent CFD.* I. *General principles.* Internat. J. Numer. Methods Fluids, 32 (2000), pp. 725–744.
[16] F. Hecht, *Bidimensional Anisotropic Mesh Generator (User's Manual)*, INRIA, Rocquencourt, 1997, http://www-rocqq.inria.fr/gamma/cdrom/www/bamg/eng.htm
[17] W. Huang, *Measuring mesh qualities and application to variational mesh*, SIAM J. Sci. Comput., 26 (2005), pp. 1643–1666.
[18] W. Huang, *Mathematical principles of anisotropic mesh adaptation*, Commun. Comput. Phys., 1 (2006), pp. 276–310.
[19] E. J. Nadler, *Piecewise Linear Approximation on Triangulations of a Planar Region*, Ph.D. thesis, Division of Applied Mathematics, Brown University, Providence, RI, 1985.
[20] A. Naga and Z. Zhang, *A posteriori error estimates based on the polynomial preserving recovery*, SIAM J. Numer. Anal., 42 (2004), pp. 1780–1800.
[21] S. Rippa, *Long and thin triangles can be good for linear interpolation*, SIAM J. Numer. Anal., 29 (1992), pp. 257–270.
[22] R. D. Russell and J. Christiansen, *Adaptive mesh selection strategies for solving boundary value problems*, SIAM J. Numer. Anal., 15 (1978), pp. 59–80.
[23] J. R. Shewchuk, *What Is a Good Linear Finite Element? Interpolation, Conditioning, Anisotropy, and Quality Measure*, Preprint, Dept. of Electronic Engineering and Computer Sciences, University of California, Berkeley, CA, 2002.
[24] R. B. Simpson, *Anisotropic mesh transformations and optimal error control*, Appl. Numer. Math., 14 (1994), pp. 183–198.
[25] M. A. Taylor, B. A. Wingate, and R. E. Vincent, *An algorithm for computing Fekete points in the triangle*, SIAM J. Numer. Anal., 38 (2000), pp. 1707–1720.
[26] R. C. Thompson, *Inertial properties of eigenvalues.* II, J. Math. Anal. Appl., 58 (1977), pp. 572–577.
[27] Z. Zhang and A. Naga, *A new finite element gradient recovery method: Superconvergence property*, SIAM J. Sci. Comput., 26 (2005), pp. 1192–1213.
[28] O. C. Zienkiewicz and J. Z. Zhu, *The superconvergence patch recovery and a posteriori error estimates, Part* I: *The recovery technique*, Internat. J. Numer. Methods Engrg., 33 (1992), pp. 1331–1364.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2392–2407

AN INF-SUP STABLE AND RESIDUAL-FREE BUBBLE ELEMENT FOR THE OSEEN EQUATIONS*

LEOPOLDO P. FRANCA†, VOLKER JOHN‡, GUNAR MATTHIES§, AND LUTZ TOBISKA¶

Abstract. We investigate the residual-free bubble method for the linearized incompressible Navier–Stokes equations. Starting with a nonconforming inf-sup stable element pair for approximating the velocity and pressure, we enrich the velocity space by discretely divergence-free bubble functions to handle the influence of strong convection. An important feature of the method is that the stabilization does not generate an additional coupling between the mass equation and the momentum equation as is the case for the streamline upwind Petrov–Galerkin method applied to equal-order interpolation. Furthermore, the discrete solution is piecewise divergence-free, a property which is useful for the mass balance in transport equations coupled with the incompressible Navier–Stokes equations.

Key words. stabilized finite elements, Navier–Stokes equations, nonconforming finite elements

AMS subject classifications. 65N12, 65N30, 65N15

DOI. 10.1137/060661454

1. Introduction.

Finite element approximations of the Oseen equations need stability for advective-dominated flows and compatibility between the velocity and pressure spaces. The latter is also necessary for the Stokes flow.

Starting with the streamline upwind Petrov–Galerkin (SUPG) stabilization of Brooks and Hughes [9] for the advective term, this idea has been extended to the Stokes equations in [21], where a stabilized method is proposed accommodating low equal-order interpolation to be stable and convergent. This formulation circumvents the need to abide by inf-sup condition for many interpolations. In an attempt to get the stability features of these works, a method is proposed in [14] that at the same time is advective stable and overcomes the inf-sup restrictions of the standard Galerkin method. The analysis of these SUPG-type stabilizations, including the case of equal-order interpolations, can be found in [31]. The drawback of these methods is that various terms need to be added to the weak formulation. Residual-based stabilization methods which use inf-sup stable pairs of elements reduce the number of terms which have to be added to the Galerkin formulation [17, 25]. However, there is still a strong coupling of the form $(\nabla p, (\mathbf{b} \cdot \nabla)\mathbf{v}_h)$ which is difficult to handle, and an optimal L^2 error estimate for the pressure is missing in [17]. Several attempts have been made to relax the strong coupling of velocity and pressure and to introduce symmetric versions of the stabilizing terms; for an overview see [5]. Local projection-type methods have

*Received by the editors May 31, 2006; accepted for publication (in revised form) April 6, 2007; published electronically November 21, 2007. This work was supported by the NSF exchange program grant INT-0339107 and the DAAD exchange program D/03/36787.
http://www.siam.org/journals/sinum/45-6/66145.html

†University of Colorado at Denver, P.O. Box 173364, Campus Box 170, Denver, CO 80217-3364 (lfranca@math.cudenver.edu).

‡Fachbereich Mathematik, Universität des Saarlandes, Postfach 151150, D-66041 Saarbrücken, Germany (john@math.uni-sb.de).

§Fakultät für Mathematik, Ruhr-Universität Bochum, Universitätsstraße 150, D-44780 Bochum, Germany (gunar.matthies@ruhr-uni-bochum.de).

¶Institut für Analysis und Numerik, Otto-von-Guericke-Universität Magdeburg, Postfach 4120, D-39016 Magdeburg, Germany (tobiska@mathematik.uni-magdeburg.de).

been introduced for the Stokes problem in [2], extended to the transport equation in [3], and analyzed for low-order discretizations of the Oseen equations in [4]. They are designed for equal-order interpolation and allow a separation of the velocity and pressure in the stabilization terms. The disadvantage is that the finite element stencil is less compact than for the SUPG-type stabilization. They also suffer from the weak fulfillment of the incompressiblity constraint which is important for mass conservation in a transport equation coupled with the Navier–Stokes problem. In the edge-oriented stabilization technique, introduced in [10], we find the same problem of a much wider stencil which needs also some special data structure or an implicit defect correction.

Our method of enriching the velocity space of an inf-sup stable pair of finite elements by discretely divergence-free functions will always suppress additional coupling terms in the discrete formulation and lead to a separation of the velocity and pressure in the stabilization terms. Due to the use of inf-sup stable finite element pairs, the computed velocity field is always discretely divergence-free. As a first step in this paper, we analyze the simplest version of such an enrichment method, the Crouzeix–Raviart element of lowest order, i.e., piecewise linear nonconforming velocity and piecewise constant pressure approximations.

The plan of the paper is as follows. In section 2, the weak formulation of the Oseen equations and its Galerkin discretization is considered. Next, in section 3, we apply the residual-free bubble approach and highlight the advantages of using discretely divergence-free enrichments. The relation to the classical SUPG method is studied in section 4. Finally, an a priori error estimate for an approximate residual-free bubble method is derived in section 5. A numerical test example confirms the convergence rates.

Notations. We use the Sobolev spaces $W^{k,p}(D)$, $H^k(D) = W^{k,2}(D)$, $H_0^k(D)$, $L^2(D) = H^0(D)$, and write $\mathbf{W}^{k,p}(D)$, $\mathbf{H}^k(D)$, $\mathbf{H}_0^k(D)$, $\mathbf{L}^2(D)$ for their vector-valued versions. The norms and seminorms in the scalar and vector-valued versions of $W^{k,p}(D)$ are denoted by $\|\cdot\|_{k,p,D}$ and $|\cdot|_{k,p,D}$, respectively [12]. To simplify the notation, we drop D if $D = \Omega$ and p if $p = 2$. Moreover, we introduce the broken H^1 seminorm and norm for piecewise H^1 functions defined on a triangulation $\mathcal{T}_h$ by

$$|v|_{1,h} := \left(\sum_{K \in \mathcal{T}_h} |v|_{1,K}^2 \right)^{1/2}, \qquad \|v\|_{1,h} := \left(|v|_{1,h}^2 + \|v\|_0^2 \right)^{1/2}.$$

2. A linearized Navier–Stokes model. We consider the steady linearized Navier–Stokes model given by

$$-\nu \Delta \mathbf{u} + (\mathbf{b} \cdot \nabla)\mathbf{u} + \nabla p = \mathbf{f} \qquad \text{in } \Omega \subset \mathbb{R}^d, \tag{2.1}$$

$$\nabla \cdot \mathbf{u} = \mathbf{0} \qquad \text{in } \Omega, \tag{2.2}$$

$$\mathbf{u} = \mathbf{0} \qquad \text{on } \Gamma = \partial\Omega, \tag{2.3}$$

where $\mathbf{b} \in \mathbf{W}^{1,\infty}(\Omega)$ with $\nabla \cdot \mathbf{b} = 0$ in Ω, $\mathbf{f} \in \mathbf{L}^2(\Omega)$, and Ω denotes a bounded domain in $\mathbb{R}^d$ with $d = 2$ or $d = 3$. Homogeneous Dirichlet boundary conditions are considered for simplicity of presentation. The extension to nonhomogeneous Dirichlet boundary conditions is straightforward when the boundary data are interpolated in the space of restrictions of discretely divergence-free functions. For smooth boundary data, this is always possible and requires only additional technical details which do not lead to further insight into the method. The weak formulation of (2.1)–(2.3) reads:

Find $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that for all $(\mathbf{v}, q) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$,

$$a(\mathbf{u}, \mathbf{v}) + b(\mathbf{u}, \mathbf{v}) - (p, \nabla \cdot \mathbf{v}) + (q, \nabla \cdot \mathbf{u}) = (\mathbf{f}, \mathbf{v}), \tag{2.4}$$

where the bilinear forms a and b are defined by

$$a(\mathbf{u}, \mathbf{v}) := \nu(\nabla \mathbf{u}, \nabla \mathbf{v}), \quad b(\mathbf{u}, \mathbf{v}) := ((\mathbf{b} \cdot \nabla)\mathbf{u}, \mathbf{v}) \quad \forall \mathbf{u}, \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{2.5}$$

$(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$ or its vector-valued and tensor-valued versions, and

$$L_0^2(\Omega) = \{q \in L^2(\Omega) \ : \ (q, 1) = 0\}.$$

The property

$$b(\mathbf{v}, \mathbf{v}) = ((\mathbf{b} \cdot \nabla)\mathbf{v}, \mathbf{v}) = \frac{1}{2}(\mathbf{b} \cdot \nabla(\mathbf{v} \cdot \mathbf{v}), 1) = -\frac{1}{2}(\nabla \cdot \mathbf{b}, \mathbf{v} \cdot \mathbf{v}) = 0 \qquad \forall \mathbf{v} \in \mathbf{H}_0^1(\Omega)$$

of the bilinear form b guarantees that the Lax–Milgram lemma can be applied in the subspace of divergence-free functions. A unique pressure in $L_0^2(\Omega)$ follows from the Babuška–Brezzi condition for the pair $(\mathbf{H}_0^1(\Omega), L_0^2(\Omega))$ [18]. Therefore, there is a unique solution $(\mathbf{u}, p)$ of (2.4) for all positive ν.

For the finite element approximation, we use the nonconforming P_1^{nc}/P_0 element pair of Crouzeix–Raviart [13]. Let $\mathcal{T}_h$ be a regular decomposition of the domain $\Omega \subset \mathbb{R}^d$ into d-dimensional simplices $K \in \mathcal{T}_h$, where the mesh parameter h represents the maximum diameter of the elements $K \in \mathcal{T}_h$. We denote by $\mathcal{E}_h$ the set of all $(d-1)$-dimensional faces E of cells $K \in \mathcal{T}_h$. We choose for any face $E \in \mathcal{E}_h$ a unit normal $\mathbf{n}_E$ with an arbitrary but fixed orientation where $\mathbf{n}_E$ on boundary faces is the outer unit normal of Ω. We will write $\mathbf{n}_K$ for the outer unit normal with respect to the cell K. For a scalar piecewise continuous function ψ, the jump $[\psi]_E$ and the average $\{\psi\}_E$ on a face E are defined by

$$[\psi]_E := \begin{cases} (\psi|_K)|_E - (\psi|_{\widetilde{K}})|_E & \text{if} \quad E \not\subset \Gamma, \\ (\psi|_K)|_E & \text{if} \quad E \subset \Gamma, \end{cases}$$

$$\{\psi\}_E := \begin{cases} \frac{1}{2}\left((\psi|_K)|_E + (\psi|_{\widetilde{K}})|_E\right) & \text{if} \quad E \not\subset \Gamma, \\ \frac{1}{2}(\psi|_K)|_E & \text{if} \quad E \subset \Gamma, \end{cases}$$

where K and $\widetilde{K}$ are chosen such that $E = \partial K \cap \partial \widetilde{K}$ and $\mathbf{n}_K = \mathbf{n}_E$.

Note that the definition of the jump and the average on a boundary face corresponds to that on an inner face when extending the functions outside of Ω by zero. Furthermore, we have the relation

$$[\varphi\psi]_E = [\varphi]_E\{\psi\}_E + \{\varphi\}_E[\psi]_E$$

on both inner and boundary faces E. The jump and the average of vector-valued functions are defined in a componentwise manner.

Now our approximate spaces $\mathbf{V}_h \approx \mathbf{H}_0^1(\Omega)$ and $Q_h \approx L_0^2(\Omega)$ can be defined to be

$$\mathbf{V}_h := \left\{\mathbf{v}_h \in \mathbf{L}^2(\Omega) \ : \ \mathbf{v}_h\Big|_K \in P_1(K)^d \ \forall K \in \mathcal{T}_h, \ \int_E [\mathbf{v}_h]_E \, d\gamma = 0 \ \forall E \in \mathcal{E}_h\right\}, \tag{2.6}$$

$$Q_h := \left\{q_h \in L_0^2(\Omega) \ : \ q_h\Big|_K \in P_0(K) \ \forall K \in \mathcal{T}_h\right\}, \tag{2.7}$$

where $P_n(K)$ is the set of all polynomials on K of degree less than or equal to n. Note that a function $\mathbf{v}_h \in \mathbf{V}_h$—in general—is discontinuous across the inner faces E and does not vanish on the boundary.

Now, we introduce the discrete bilinear forms elementwise to be

$$a_h(\mathbf{u}_h, \mathbf{v}_h) := \nu \sum_{K\in\mathcal{T}_h} (\nabla_h \mathbf{u}_h, \nabla_h \mathbf{v}_h)_K, \tag{2.8}$$

$$b_h(\mathbf{u}_h, \mathbf{v}_h) := \sum_{K\in\mathcal{T}_h} ((\mathbf{b}\cdot\nabla_h)\mathbf{u}_h, \mathbf{v}_h)_K - \sum_{E\in\mathcal{E}_h} \langle \mathbf{b}\cdot\mathbf{n}_E[\mathbf{u}_h]_E, \{\mathbf{v}_h\}_E\rangle_E\,. \tag{2.9}$$

Here, the discrete versions of the gradient and the divergence operators, ∇ and $\nabla\cdot$, respectively, are understood in the following sense:

$$(\nabla_h \mathbf{v}_h)\big|_K := \nabla\left(\mathbf{v}_h\big|_K\right) \quad \forall \mathbf{v}_h \in \mathbf{V}_h,\ \forall K \in \mathcal{T}_h,$$

$$(\nabla_h \cdot \mathbf{v}_h)\big|_K := \nabla\cdot\left(\mathbf{v}_h\big|_K\right) \quad \forall \mathbf{v}_h \in \mathbf{V}_h,\ \forall K \in \mathcal{T}_h,$$

and $\langle\cdot,\cdot\rangle_E$ denotes the inner product in $L^2(E)$ and its vector-valued versions. To simplify the notation, we briefly write ∇ instead of ∇_h in expressions like (2.8) and (2.9). Clearly, we have

$$a_h(\mathbf{u}, \mathbf{v}) = a(\mathbf{u}, \mathbf{v}), \quad b_h(\mathbf{u}, \mathbf{v}) = b(\mathbf{u}, \mathbf{v}), \quad \mathbf{u}, \mathbf{v} \in \mathbf{H}^1(\Omega).$$

The additional term in the elementwise-defined bilinear form b_h (compare (2.9)) vanishes for $\mathbf{v}_h \in \mathbf{H}^1(\Omega)$. For functions $\mathbf{v}_h$ belonging to our nonconforming finite element space $\mathbf{V}_h$, it guarantees that we have

$$\begin{aligned} b_h(\mathbf{v}_h, \mathbf{v}_h) &= \frac{1}{2}\sum_{K\in\mathcal{T}_h} ((\mathbf{b}\cdot\nabla)(\mathbf{v}_h\cdot\mathbf{v}_h), 1)_K - \sum_{E\in\mathcal{E}_h} \langle \mathbf{b}\cdot\mathbf{n}_E[\mathbf{v}_h]_E, \{\mathbf{v}_h\}_E\rangle_E \\ &= \sum_{E\in\mathcal{E}_h} \left(\frac{1}{2}\langle \mathbf{b}\cdot\mathbf{n}_E[\mathbf{v}_h\cdot\mathbf{v}_h]_E, 1\rangle_E - \langle \mathbf{b}\cdot\mathbf{n}_E[\mathbf{v}_h]_E, \{\mathbf{v}_h\}_E\rangle_E\right) = 0, \end{aligned}$$

in analogy to $b(\mathbf{v}, \mathbf{v}) = 0$ for all $\mathbf{v} \in \mathbf{H}^1_0(\Omega)$.

The standard Galerkin finite element method reads:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$,

$$a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla_h\cdot\mathbf{v}_h) + (q_h, \nabla_h\cdot\mathbf{u}_h) = (\mathbf{f}, \mathbf{v}_h). \tag{2.10}$$

The finite element pair $(\mathbf{V}_h, Q_h)$ satisfies the discrete inf-sup stability condition

$$\exists \beta_0 > 0 \quad \forall q_h \in Q_h: \qquad \beta_0 \|q_h\|_0 \le \sup_{\mathbf{v}_h\in\mathbf{V}_h} \frac{(q_h, \nabla_h\cdot\mathbf{v}_h)}{|\mathbf{v}_h|_{1,h}}; \tag{2.11}$$

see [6, 13]. As a result, we have the unique solvability of (2.10). Error estimates which do not take into consideration the size of ν are standard, e.g., in the energy norm we have

$$\nu^{1/2}|\mathbf{u} - \mathbf{u}_h|_{1,h} + \|p - p_h\|_0 \le C(\nu)\ h\ (|u|_2 + |p|_1) \tag{2.12}$$

with a constant $C(\nu)$ depending on ν. We are interested in the case of small ν (high Reynolds numbers) in which numerical experiments show the need for stabilization [11, 29, 30]. In the next section, we will follow the concept of residual-free bubble stabilizations, which has been already successfully applied to scalar convection-diffusion equations [1, 7, 8, 16].

3. Residual-free bubble method. Let us enrich the velocity space $\mathbf{V}_h$ by the space of residual-free bubbles

$$\mathbf{B}_h := \bigoplus_{K \in \mathcal{T}_h} \mathbf{H}_0^1(K)$$

and denote the enriched space by $\mathbf{V}_{RFB}$. Since a piecewise linear function which vanishes at the boundary of each cell is identically zero, we conclude $\mathbf{V}_{RFB} = \mathbf{V}_h \oplus \mathbf{B}_h$. The pair $(\mathbf{V}_{RFB}, Q_h)$ satisfies the discrete inf-sup stability (2.11) as well. Note that a function from the bubble space $\mathbf{B}_h$ is discretely divergence-free since we have, for all $q_h \in Q_h,\ \mathbf{v}_B \in \mathbf{B}_h$,

$$(q_h, \nabla_h \cdot \mathbf{v}_B) = \sum_{K \in \mathcal{T}_h} q_h\big|_K (1, \nabla \cdot \mathbf{v}_B)_K = \sum_{K \in \mathcal{T}_h} q_h\big|_K \langle 1, \mathbf{v}_B \cdot \mathbf{n}_K \rangle_{\partial K} = 0.$$

In this sense the inf-sup stability will not be improved by enriching $\mathbf{V}_h$ by $\mathbf{B}_h$. Each element $\mathbf{u}_{RFB} \in \mathbf{V}_{RFB}$ can be uniquely represented in the form

$$\mathbf{u}_{RFB} = \mathbf{u}_h + \mathbf{u}_B \quad \text{with } \mathbf{u}_h \in \mathbf{V}_h,\ \mathbf{u}_B \in \mathbf{B}_h.$$

The Galerkin approximation of (2.4) with respect to the pair $(\mathbf{V}_{RFB}, Q_h)$ reads:

Find $(\mathbf{u}_h, \mathbf{u}_B, p_h) \in \mathbf{V}_h \times \mathbf{B}_h \times Q_h$ such that

$$a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{u}_B, \mathbf{v}_h) - (p_h, \nabla_h \cdot \mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) \quad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{3.1}$$

$$a_h(\mathbf{u}_B, \mathbf{v}_B) + b_h(\mathbf{u}_B, \mathbf{v}_B) + b_h(\mathbf{u}_h, \mathbf{v}_B) = (\mathbf{f}, \mathbf{v}_B) \quad \forall \mathbf{v}_B \in \mathbf{B}_h, \tag{3.2}$$

$$(q_h, \nabla_h \cdot \mathbf{u}_h) = 0 \qquad \forall q_h \in Q_h. \tag{3.3}$$

Note that in deriving (3.1)–(3.3) we have taken into consideration the orthogonality property

$$\begin{aligned} a_h(\mathbf{v}_B, \mathbf{w}_h) = a_h(\mathbf{w}_h, \mathbf{v}_B) &= \nu \sum_{K \in \mathcal{T}_h} (\nabla \mathbf{w}_h, \nabla \mathbf{v}_B)_K \\ &= \nu \sum_{K \in \mathcal{T}_h} \left(\left\langle \frac{\partial \mathbf{w}_h}{\partial \mathbf{n}_K}, \mathbf{v}_B \right\rangle_{\partial K} - (\Delta \mathbf{w}_h, \mathbf{v}_B)_K \right) = 0 \end{aligned}$$

and the property that $\mathbf{u}_B$ and $\mathbf{v}_B$ are discretely divergence-free. Equation (3.2) can be considered to define $\mathbf{u}_B$ as a functional of $\mathbf{u}_h$. In order to find a representation for $\mathbf{u}_B$, we define $M(\mathbf{u}_h)$, $F(\mathbf{f}) \in \mathbf{B}_h$ as the solutions of the problems:

Find $M(\mathbf{u}_h),\ F(\mathbf{f}) \in \mathbf{B}_h$ such that for all $\mathbf{v}_B \in \mathbf{B}_h$,

$$a_h(M(\mathbf{u}_h), \mathbf{v}_B) + b_h(M(\mathbf{u}_h), \mathbf{v}_B) = -b_h(\mathbf{u}_h, \mathbf{v}_B),$$

$$a_h(F(\mathbf{f}), \mathbf{v}_B) + b_h(F(\mathbf{f}), \mathbf{v}_B) = (\mathbf{f}, \mathbf{v}_B).$$

Then, the solution $\mathbf{u}_B$ of (3.2) can be represented in the form $\mathbf{u}_B = M(\mathbf{u}_h) + F(\mathbf{f})$. Elimination of $\mathbf{u}_B$ from (3.1) gives the residual-free bubble method for solving (2.4):

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that

$$a_{RFB}(\mathbf{u}_h, \mathbf{v}_h) - (p_h, \nabla_h \cdot \mathbf{v}_h) = l_{RFB}(\mathbf{v}_h) \qquad \forall \mathbf{v}_h \in \mathbf{V}_h, \tag{3.4}$$

$$(q_h, \nabla_h \cdot \mathbf{u}_h) = 0 \qquad \forall q_h \in Q_h, \tag{3.5}$$

where

$$a_{RFB}(\mathbf{u}_h, \mathbf{v}_h) = a_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(\mathbf{u}_h, \mathbf{v}_h) + b_h(M(\mathbf{u}_h), \mathbf{v}_h), \tag{3.6}$$

$$l_{RFB}(\mathbf{v}_h) = (\mathbf{f}, \mathbf{v}_h) - b_h(F(\mathbf{f}), \mathbf{v}_h). \tag{3.7}$$

The difficulty in realizing the *exact* residual-free method (3.4)–(3.5) is that we have to evaluate the terms $b_h(M(\mathbf{u}_h), \mathbf{v}_h)$ and $b_h(F(\mathbf{f}), \mathbf{v}_h)$, which essentially means solving an infinite-dimensional problem. Therefore, in practice some sort of approximation is used. We mention in particular the following approaches:

- stabilizing subgrid methods [8],
- pseudo-residual-free bubble method [7],
- two-level and three-level approaches [15, 16, 19, 20].

In the following we will reformulate the method (3.4)–(3.5) by looking at the constant coefficient case.

4. Relation to other stabilized methods. The case of continuous P_1 pressure and velocity approximations on triangles has been considered in [28]; for a systematic study on quadrilaterals with a continuous Q_1 pressure approximation and a sufficiently large velocity space see [24]. In that paper the fully nonlinear case of the Navier–Stokes equations has also been considered.

In the following we consider a discretization within the space $(\mathbf{V}_h \times Q_h)$, i.e., nonconforming piecewise linear velocity and piecewise constant pressure approximations. Let us assume that $\mathbf{b}$ and $\mathbf{f}$ are constants. Moreover, let $\varphi_K \in H_0^1(K)$ be the solution of the scalar convection-diffusion problem

$$-\nu\Delta\varphi_K + \mathbf{b}\cdot\nabla\varphi_K = 1 \text{ in } K, \quad \varphi_K = 0 \text{ on } \partial K.$$

Then, we obtain

$$M(\mathbf{u}_h)\Big|_K = -(\mathbf{b}\cdot\nabla)\mathbf{u}_h\Big|_K \varphi_K, \quad F(\mathbf{f})\Big|_K = \mathbf{f}\Big|_K \varphi_K.$$

The terms which appear in (3.6)–(3.7), in addition to the standard Galerkin approach, become

$$\begin{aligned} b_h(M(\mathbf{u}_h), \mathbf{v}_h) &= \sum_{K\in\mathcal{T}_h} ((\mathbf{b}\cdot\nabla)M(\mathbf{u}_h), \mathbf{v}_h)_K - \sum_{E\in\mathcal{E}_h} \langle \mathbf{b}\cdot\mathbf{n}_E[M(\mathbf{u}_h)]_E, \{\mathbf{v}_h\}_E\rangle_E \\ &= -\sum_{K\in\mathcal{T}_h} ((\mathbf{b}\cdot\nabla)\mathbf{v}_h, M(\mathbf{u}_h))_K \\ &= \sum_{K\in\mathcal{T}_h} ((\mathbf{b}\cdot\nabla)\mathbf{v}_h, (\mathbf{b}\cdot\nabla)\mathbf{u}_h\ \varphi_K)_K \\ &= \sum_{K\in\mathcal{T}_h} \tau_K((\mathbf{b}\cdot\nabla)\mathbf{v}_h, (\mathbf{b}\cdot\nabla)\mathbf{u}_h)_K, \end{aligned}$$

$$
\begin{aligned}
-b_h(F(\mathbf{f}),\mathbf{v}_h) &= -\sum_{K\in\mathcal{T}_h}((\mathbf{b}\cdot\nabla)F(\mathbf{f}),\mathbf{v}_h)_K + \sum_{E\in\mathcal{E}_h}\langle \mathbf{b}\cdot\mathbf{n}_E[F(\mathbf{f})]_E,\{\mathbf{v}_h\}_E\rangle_E \\
&= \sum_{K\in\mathcal{T}_h}((\mathbf{b}\cdot\nabla)\mathbf{v}_h,F(\mathbf{f}))_K \\
&= \sum_{K\in\mathcal{T}_h}((\mathbf{b}\cdot\nabla)\mathbf{v}_h,\mathbf{f}\,\varphi_K)_K \\
&= \sum_{K\in\mathcal{T}_h}\tau_K((\mathbf{b}\cdot\nabla)\mathbf{v}_h,\mathbf{f})_K,
\end{aligned}
$$

since $M(\mathbf{u}_h),\ F(\mathbf{f})\in\mathbf{B}_h$ where

$$\tau_K = \frac{1}{|K|}\int_K \varphi_K\,dx.$$

Thus, the exact residual-free bubble method for constant $\mathbf{b}$ and $\mathbf{f}$ is equal to:

Find $(\mathbf{u}_h,p_h)\in\mathbf{V}_h\times Q_h$ such that

$$\tilde{a}_{RFB}(\mathbf{u}_h,\mathbf{v}_h)-(p_h,\nabla_h\cdot\mathbf{v}_h)=\tilde{l}_{RFB}(\mathbf{v}_h)\qquad\forall\mathbf{v}_h\in\mathbf{V}_h, \tag{4.1}$$

$$(q_h,\nabla_h\cdot\mathbf{u}_h)=0\qquad\forall q_h\in Q_h, \tag{4.2}$$

where

$$\tilde{a}_{RFB}(\mathbf{u}_h,\mathbf{v}_h)=a_h(\mathbf{u}_h,\mathbf{v}_h)+b_h(\mathbf{u}_h,\mathbf{v}_h)+\sum_{K\in\mathcal{T}_h}\tau_K((\mathbf{b}\cdot\nabla)\mathbf{u}_h,(\mathbf{b}\cdot\nabla)\mathbf{v}_h)_K,$$

$$\tilde{l}_{RFB}(\mathbf{v}_h)=(\mathbf{f},\mathbf{v}_h)+\sum_{K\in\mathcal{T}_h}\tau_K(\mathbf{f},(\mathbf{b}\cdot\nabla)\mathbf{v}_h)_K.$$

Since on each $K\in\mathcal{T}_h$ it holds that $-\nu\Delta\mathbf{u}_h+\nabla p_h=0$, the method corresponds to the SUPG method analyzed in [26] for the fully nonlinear case of the Navier–Stokes equations. However, the influence of small ν on the error constants has not been investigated in that paper.

5. Error estimate for the generalized Oseen equations. We now turn to estimates with Reynolds-number-independent constants. It has been shown in a series of papers [22, 23, 27] that for nonconforming finite element discretizations applied to scalar convection-diffusion equations, one has to add certain jump terms to the discretization to recover the error estimates of the SUPG method known for conforming finite elements. Therefore, we expect to meet the same situation in the more complex problem of linearized Navier–Stokes equations and add

$$j_h(\mathbf{u}_h,\mathbf{v}_h):=\sum_{E\in\mathcal{E}_h}\gamma_E\langle[\mathbf{u}_h]_E,[\mathbf{v}_h]_E\rangle_E$$

with positive constants γ_E to the discrete formulation. In the case of a scalar convection-diffusion equation it turns out that it is enough to choose $\gamma_E\sim 1$ (see [22]), but due to the coupling with the pressure we have to choose γ_E differently; see Lemma 5.2. Note that the solution $(\mathbf{u},p)\in\mathbf{H}_0^1(\Omega)\times L_0^2(\Omega)$ satisfies $[\mathbf{u}]_E=0$ and consequently $j_h(\mathbf{u},\mathbf{v})=0$ for all $\mathbf{v}\in\mathbf{H}_0^1(\Omega)+\mathbf{V}_h$.

We shall consider and analyze the case of the generalized Oseen equations,

$$-\nu\Delta\mathbf{u} + (\mathbf{b}\cdot\nabla)\mathbf{u} + \sigma\mathbf{u} + \nabla p = \mathbf{f}, \quad \nabla\cdot\mathbf{u} = \mathbf{0} \quad \text{in } \Omega, \quad \mathbf{u} = \mathbf{0} \quad \text{on } \Gamma = \partial\Omega$$

which appears as a result of time discretizations of the nonstationary Navier–Stokes equations with $\sigma = (1/\Delta t)$. Its weak formulation reads:

Find $(\mathbf{u}, p) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$ such that for all $(\mathbf{v}, q) \in \mathbf{H}_0^1(\Omega) \times L_0^2(\Omega)$,

$$a^\sigma(\mathbf{u}, \mathbf{v}) + b(\mathbf{u}, \mathbf{v}) - (p, \nabla\cdot\mathbf{v}) + (q, \nabla\cdot\mathbf{u}) = (\mathbf{f}, \mathbf{v}), \tag{5.1}$$

where the bilinear form $a(\cdot,\cdot)$ in (2.4) has been replaced by the bilinear form

$$a^\sigma(\mathbf{u}, \mathbf{v}) := \nu(\nabla\mathbf{u}, \nabla\mathbf{v}) + \sigma(\mathbf{u}, \mathbf{v}).$$

Let us introduce the following notations:

$$\begin{aligned} A\left((\mathbf{u}, p), (\mathbf{v}, q)\right) &= a_h^\sigma(\mathbf{u}, \mathbf{v}) + b_h(\mathbf{u}, \mathbf{v}) + j_h(\mathbf{u}, \mathbf{v}) + \sum_{K\in\mathcal{T}_h} \tau_K((\mathbf{b}\cdot\nabla)\mathbf{u}, (\mathbf{b}\cdot\nabla)\mathbf{v})_K \\ &\quad - (p, \nabla_h\cdot\mathbf{v}) + (q, \nabla_h\cdot\mathbf{u}), \\ L\left((\mathbf{v}, q)\right) &= (\mathbf{f}, \mathbf{v}) + \sum_{K\in\mathcal{T}_h} \tau_K(\mathbf{f}, (\mathbf{b}\cdot\nabla)\mathbf{v})_K \end{aligned}$$

with $a_h^\sigma(\cdot,\cdot)$ being the discrete analogue of $a^\sigma(\cdot,\cdot)$, more precisely

$$a_h^\sigma(\mathbf{u}_h, \mathbf{v}_h) := \sum_{K\in\mathcal{T}_h} (\nu(\nabla\mathbf{u}_h, \nabla\mathbf{v}_h)_K + \sigma(\mathbf{u}_h, \mathbf{v}_h)_K) \quad \forall \mathbf{u}_h, \mathbf{v}_h \in \mathbf{V}_h + \mathbf{H}_0^1(\Omega).$$

The discrete problem to be studied now becomes:

Find $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ such that for all $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$,

$$A\left((\mathbf{u}_h, p_h), (\mathbf{v}_h, q_h)\right) = L\left((\mathbf{v}_h, q_h)\right). \tag{5.2}$$

The bilinear form $A(\cdot,\cdot)$ generates a norm on the product space $\mathbf{V}_h \times Q_h$

$$\begin{aligned} |||(\mathbf{v}, q)||| = \Bigg(&\nu|\mathbf{v}|_{1,h}^2 + \sigma\|\mathbf{v}\|_0^2 + (\nu+\sigma)\|q\|_0^2 \\ &+ j_h(\mathbf{v}, \mathbf{v}) + \sum_{K\in\mathcal{T}_h} \tau_K\|(\mathbf{b}\cdot\nabla)\mathbf{v}\|_{0,K}^2 \Bigg)^{1/2}. \end{aligned}$$

First we show an inf-sup condition for the bilinear form $A(\cdot,\cdot)$ on the product space $\mathbf{V}_h \times Q_h$.

LEMMA 5.1. *Assume that* $\max(\nu, \sigma, \tau_K, \gamma_E h_E) \le C$. *Then, there is a positive constant* β *independent of* $\nu > 0$ *such that for all* $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$,

$$|||(\mathbf{v}_h, q_h)||| \le \frac{1}{\beta} \sup_{(\mathbf{w}_h, r_h)\in\mathbf{V}_h\times Q_h} \frac{A\left((\mathbf{v}_h, q_h), (\mathbf{w}_h, r_h)\right)}{|||(\mathbf{w}_h, r_h)|||}. \tag{5.3}$$

Proof. Let us consider an arbitrary $(\mathbf{v}_h, q_h) \in \mathbf{V}_h \times Q_h$. Choosing $(\mathbf{w}_h, r_h) = (\mathbf{v}_h, q_h)$, we have

$$\begin{aligned} &A\big((\mathbf{v}_h, q_h), (\mathbf{v}_h, q_h)\big) \\ &\quad = \nu|\mathbf{v}_h|_{1,h}^2 + \sigma\|\mathbf{v}_h\|_0^2 + j_h(\mathbf{v}_h, \mathbf{v}_h) + \sum_{K\in\mathcal{T}_h} \tau_K\|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}^2 \end{aligned} \tag{5.4}$$

due to the property $b_h(\mathbf{v}_h, \mathbf{v}_h) = 0$ which has been shown in section 2.

Now let us consider another choice of $(\mathbf{w}_h, r_h)$. For any $q_h \in Q_h$ the discrete Babuška–Brezzi condition (2.11) guarantees the existence of a function $\mathbf{v}_{q_h} \in \mathbf{V}_h$ such that

$$(\nabla_h \cdot \mathbf{v}_{q_h}, q_h) = -(q_h, q_h), \qquad \|\mathbf{v}_{q_h}\|_{1,h} \leq C\|q_h\|_0.$$

Thus, by choosing $(\mathbf{w}_h, r_h) = (\mathbf{v}_{q_h}, 0)$ we obtain

$$\begin{aligned} A\left((\mathbf{v}_h, q_h), (\mathbf{v}_{q_h}, 0)\right) &= \|q_h\|_0^2 + a_h^\sigma(\mathbf{v}_h, \mathbf{v}_{q_h}) + b_h(\mathbf{v}_h, \mathbf{v}_{q_h}) + j_h(\mathbf{v}_h, \mathbf{v}_{q_h}) \\ &\quad + \sum_{K \in \mathcal{T}_h} \tau_K((\mathbf{b}\cdot\nabla)\mathbf{v}_h, (\mathbf{b}\cdot\nabla)\mathbf{v}_{q_h})_K. \end{aligned} \tag{5.5}$$

Now, the second term on the right-hand side of (5.5) can be bounded as follows:

$$\begin{aligned} |a_h^\sigma(\mathbf{v}_h, \mathbf{v}_{q_h})| &\leq C\ (\nu|\mathbf{v}_h|_{1,h} + \sigma\|\mathbf{v}_h\|_0)\ \|q_h\|_0 \\ &\leq C\ \left(\nu^2|\mathbf{v}_h|_{1,h}^2 + \sigma^2\|\mathbf{v}_h\|_0^2\right) + \frac{1}{8}\|q_h\|_0^2. \end{aligned}$$

Elementwise integration by parts of the third term on the right-hand side of (5.5) gives

$$\begin{aligned} b_h(\mathbf{v}_h, \mathbf{v}_{q_h}) &= \sum_{E \in \mathcal{E}_h} \langle \mathbf{b}\cdot\mathbf{n}_E, [\mathbf{v}_h \cdot \mathbf{v}_{q_h}]_E\rangle_E - \sum_{K \in \mathcal{T}_h} ((\mathbf{b}\cdot\nabla)\mathbf{v}_{q_h}, \mathbf{v}_h)_K \\ &\quad - \sum_{E \in \mathcal{E}_h} \langle \mathbf{b}\cdot\mathbf{n}_E[\mathbf{v}_h]_E, \{\mathbf{v}_{q_h}\}_E\rangle_E \\ &= \sum_{E \in \mathcal{E}_h} \langle \mathbf{b}\cdot\mathbf{n}_E[\mathbf{v}_{q_h}]_E, \{\mathbf{v}_h\}_E\rangle_E - \sum_{K \in \mathcal{T}_h} ((\mathbf{b}\cdot\nabla)\mathbf{v}_{q_h}, \mathbf{v}_h)_K. \end{aligned}$$

Let $\omega(E)$ denote the union of the cells K sharing a common face E. For any $\mathbf{v}_h \in V_h$ we have

$$\|[\mathbf{v}_h]_E\|_{0,E} \leq C\, h_E^{1/2}|\mathbf{v}_h|_{1,h,\omega(E)}, \qquad \|\{\mathbf{v}_h\}_E\|_{0,E} \leq C\, h_E^{-1/2}\|\mathbf{v}_h\|_{0,\omega(E)},$$

from which

$$|b_h(\mathbf{v}_h, \mathbf{v}_{q_h})| \leq C\,|\mathbf{v}_{q_h}|_{1,h}\,\|\mathbf{v}_h\|_0 \leq C\,\|q_h\|_0\,\|\mathbf{v}_h\|_0 \leq C\,\|\mathbf{v}_h\|_0^2 + \frac{1}{8}\|q_h\|_0^2$$

follows. Similarly, for the fourth term on the right-hand side of (5.5) we obtain

$$\begin{aligned} j_h(\mathbf{v}_h, \mathbf{v}_{q_h}) &\leq C \sum_{E \in \mathcal{E}_h} \gamma_E\|[\mathbf{v}_h]_E\|_{0,E}\, h_E^{1/2}|\mathbf{v}_{q_h}|_{1,h,\omega(E)} \\ &\leq C \sum_{E \in \mathcal{E}_h} \gamma_E^2 h_E\|[\mathbf{v}_h]_E\|_{0,E}^2 + \frac{1}{8}\|q_h\|_0^2. \end{aligned}$$

Finally, the fifth term on the right-hand side of (5.5) is estimated by

$$\begin{aligned} \left|\sum_{K \in \mathcal{T}_h} \tau_K((\mathbf{b}\cdot\nabla)\mathbf{v}_h, (\mathbf{b}\cdot\nabla)\mathbf{v}_{q_h})_K\right| &\leq \sum_{K \in \mathcal{T}_h} \tau_K\|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}\|(\mathbf{b}\cdot\nabla)\mathbf{v}_{q_h}\|_{0,K} \\ &\leq C \sum_{K \in \mathcal{T}_h} \tau_K^2\|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}^2 + \frac{1}{8}\|q_h\|_0^2. \end{aligned}$$

Combining the inequalities and taking into consideration that ν, τ_K, and $\gamma_E h_E$ are bounded from above, we get from (5.5)

$$A\big((\mathbf{v}_h, q_h), (\mathbf{v}_{q_h}, 0)\big) \geq \frac{1}{2}\|q_h\|_0^2$$

$$- C_1 \left[\nu|\mathbf{v}_h|_{1,h}^2 + \|\mathbf{v}_h\|_0^2 + j_h(\mathbf{v}_h, \mathbf{v}_h) + \sum_{K\in\mathcal{T}_h} \tau_K \|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}^2\right]. \tag{5.6}$$

Multiplying this inequality by $(\nu+\sigma)$, using the estimate $\nu+\sigma \leq C$ to bound

$$(\nu+\sigma)\nu|\mathbf{v}_h|_{1,h}^2 \leq C\nu|\mathbf{v}_h|_{1,h}^2,$$
$$(\nu+\sigma) j_h(\mathbf{v}_h, \mathbf{v}_h) \leq C j_h(\mathbf{v}_h, \mathbf{v}_h),$$
$$(\nu+\sigma)\sum_{K\in\mathcal{T}_h} \tau_K \|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}^2 \leq C \sum_{K\in\mathcal{T}_h} \tau_K \|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}^2,$$

and hiding the $\nu\|\mathbf{v}_h\|_0^2$ term by the discrete Poincaré's inequality

$$(\nu+\sigma)\|\mathbf{v}_h\|_0^2 = \nu\|\mathbf{v}_h\|_0^2 + \sigma\|\mathbf{v}_h\|_0^2 \leq C\nu|\mathbf{v}_h|_{1,h}^2 + \sigma\|\mathbf{v}_h\|_0^2,$$

we end up with

$$A\big((\mathbf{v}_h, q_h), ((\nu+\sigma)\mathbf{v}_{q_h}, 0)\big) \geq \frac{\nu+\sigma}{2}\|q_h\|_0^2$$

$$- C_2 \left[\nu|\mathbf{v}_h|_{1,h}^2 + \sigma\|\mathbf{v}_h\|_0^2 + j_h(\mathbf{v}_h, \mathbf{v}_h) + \sum_{K\in\mathcal{T}_h} \tau_K \|(\mathbf{b}\cdot\nabla)\mathbf{v}_h\|_{0,K}^2\right]. \tag{5.7}$$

From (5.4) and (5.7) we get for $(\mathbf{w}_h, r_h) := (1-\alpha)(\mathbf{v}_h, q_h) + \alpha((\nu+\sigma)\mathbf{v}_{q_h}, 0)$,

$$A\big((\mathbf{v}_h, q_h), (\mathbf{w}_h, r_h)\big) \geq \frac{\alpha}{2}|||(\mathbf{v}_h, q_h)|||^2 \tag{5.8}$$

with $\alpha = 2/(2C_2+3) \in (0,1)$. Moreover, analyzing each individual term in the triple norm, we can show that

$$|||(\mathbf{v}_{q_h}, 0)||| \leq C\|\mathbf{v}_{q_h}\|_{1,h} \leq C\|q_h\|_0,$$

and with $\nu+\sigma \leq C\sqrt{\nu+\sigma}$ we conclude that

$$\begin{aligned} |||(\mathbf{w}_h, r_h)||| &\leq (1-\alpha)|||(\mathbf{v}_h, q_h)||| + \alpha(\nu+\sigma)|||(\mathbf{v}_{q_h}, 0)||| \\ &\leq C_3|||(\mathbf{v}_h, q_h)||| \end{aligned}$$

follows. Thus, we obtain (5.3) with $\beta = \alpha/(2C_3)$. ∎

Remark. Note that for $\sigma > 0$ we have control over the L^2 norm of the velocity and the pressure *uniformly with respect to* ν. However, for $\sigma = 0$ we lose this uniform L^2 norm control. In this case, the pressure is only controlled by $\nu^{1/2}\|\cdot\|_0$. Taking into consideration Poincaré's inequality we see that the velocity is also controlled by $\nu^{1/2}\|\cdot\|_0$. This behavior, that the case $\sigma > 0$ leads to a uniform (with respect to ν) control of the L^2 norm of velocity and pressure, can be also observed in other stabilized methods; see, for example, [11].

Let the weak solution of the generalized Oseen equations belong additionally to $\mathbf{H}^2(\Omega) \times H^1(\Omega)$. Our formulation admits the following consistency property, where the parameter choice satisfies the assumption of Lemma 5.1.

LEMMA 5.2. *Let* $(\mathbf{u}, p) \in (\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)) \times (L_0^2(\Omega) \cap H^1(\Omega))$ *be the weak solution of* (5.1) *and let* $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ *be the discrete solution of* (5.2). *Then, the consistency error can be represented in the form*

$$\begin{aligned} R(\mathbf{u}, p; \mathbf{w}_h, r_h) :&= A\left((\mathbf{u} - \mathbf{u}_h, p - p_h), (\mathbf{w}_h, r_h)\right) \\ &= \sum_{E \in \mathcal{E}_h} \left\{ \left\langle \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}_E}, [\mathbf{w}_h]_E \right\rangle_E - \langle p, [\mathbf{w}_h]_E \cdot \mathbf{n}_E \rangle_E \right\} \\ &\quad + \sum_{K \in \mathcal{T}_h} \tau_K (\nu \Delta \mathbf{u} - \sigma \mathbf{u} - \nabla p, (\mathbf{b} \cdot \nabla) \mathbf{w}_h)_K. \end{aligned}$$

Furthermore, assume that $\tau_K \sim h_K^2$ *and* $\gamma_E \sim h_E^{-1}$. *Then, there is a positive constant* C *independent of* ν *such that*

$$|R(\mathbf{u}, p; \mathbf{w}_h, r_h)| \leq C\, h (\|\mathbf{u}\|_2 + \|p\|_1)\, |||(\mathbf{w}_h, r_h)||| \quad \forall (\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h.$$

Proof. The representation follows by testing the strong form of the problem with $\mathbf{w}_h$ and $(\mathbf{b} \cdot \nabla)\mathbf{w}_h$, respectively, elementwise integration by parts, and taking into consideration the definition of $A(\cdot, \cdot)$, (5.1), and (5.2). Following [13] we have

$$\left| \sum_{E \in \mathcal{E}_h} \left\langle \nu \frac{\partial \mathbf{u}}{\partial \mathbf{n}_E}, [\mathbf{w}_h]_E \right\rangle_E \right| \leq C\, h \nu \|\mathbf{u}\|_2\, |\mathbf{w}_h|_{1,h} \leq C\, h \|\mathbf{u}\|_2\, |||(\mathbf{w}_h, r_h)|||,$$

$$\left| \sum_{E \in \mathcal{E}_h} \langle p, [\mathbf{w}_h]_E \cdot \mathbf{n}_E \rangle_E \right| \leq Ch \|p\|_1\, |\mathbf{w}_h|_{1,h},$$

which shows that the second estimate does not lead to the desired estimate with a ν independent constant. Therefore, we bound the term in a different way as follows:

$$\begin{aligned} \left| \sum_{E \in \mathcal{E}_h} \langle p, [\mathbf{w}_h]_E \cdot \mathbf{n}_E \rangle_E \right| &\leq C \sum_{E \in \mathcal{E}_h} \gamma_E^{-1/2} h_E^{1/2} |p|_{1,h,\omega(E)}\ \gamma_E^{1/2} \|[\mathbf{w}_h]_E\|_{0,E} \\ &\leq Ch \|p\|_1 \sqrt{j_h(\mathbf{w}_h, \mathbf{w}_h)}. \end{aligned}$$

Concerning the last term of the consistency error, we get

$$\begin{aligned} &\left| \sum_{K \in \mathcal{T}_h} \tau_K (\nu \Delta \mathbf{u} - \sigma \mathbf{u} - \nabla p, (\mathbf{b} \cdot \nabla) \mathbf{w}_h)_K \right| \\ &\qquad \leq \sum_{K \in \mathcal{T}_h} \tau_K^{1/2} \left(\|\mathbf{u}\|_{2,K} + \|p\|_{1,K}\right)\ \tau_K^{1/2} \|(\mathbf{b} \cdot \nabla) \mathbf{w}_h\|_{0,K} \\ &\qquad \leq Ch \left(\|\mathbf{u}\|_2 + \|p\|_{1,K}\right) \left(\sum_{K \in \mathcal{T}_h} \tau_K \|(\mathbf{b} \cdot \nabla) \mathbf{w}_h\|_{0,K}^2 \right)^{1/2}. \end{aligned}$$

Summarizing the individual estimates we obtain the statement of the lemma. ∎

Next we shall investigate the interpolation error. First, note that a discretely divergence-free function is divergence-free on each cell K. Indeed, if χ_K denotes the characteristic function of K, $|K|$ and $|\Omega|$ denoting the measure of K and Ω, respectively, we conclude for a discretely divergence-free function $\mathbf{v}_h \in \mathbf{V}_h$ that the function $\nabla_h \cdot \mathbf{v}_h$ is piecewise constant and, thus, by setting $q_h = \chi_K - |K|/|\Omega| \in Q_h$,

$$\begin{aligned} 0 = (q_h, \nabla_h \cdot \mathbf{v}_h) &= (1, \nabla_h \cdot \mathbf{v}_h)_K - \frac{|K|}{|\Omega|}(1, \nabla_h \cdot \mathbf{v}_h)_\Omega \\ &= |K| \left(\nabla \cdot \mathbf{v}_h\big|_K\right) - \frac{|K|}{|\Omega|} \sum_{K \in \mathcal{T}_h} \langle 1, \mathbf{v}_h \cdot \mathbf{n}_K \rangle_{\partial K} \\ &= |K| \left(\nabla \cdot (\mathbf{v}_h\big|_K)\right). \end{aligned}$$

LEMMA 5.3. *The canonical interpolant $\mathbf{I}_h : \mathbf{H}_0^1(\Omega) \to \mathbf{V}_h$ defined by*

$$\frac{1}{|E|} \int_E (\mathbf{I}_h \mathbf{v} - \mathbf{v}) \, ds = \mathbf{0} \qquad \forall E \in \mathcal{E}_h$$

satisfies

$$(q_h, \nabla_h \cdot \mathbf{I}_h \mathbf{v}) = (q_h, \nabla \cdot \mathbf{v}) \qquad \forall\, q_h \in Q_h,\ \mathbf{v} \in \mathbf{H}_0^1(\Omega), \tag{5.9}$$

$$\|\mathbf{v} - \mathbf{I}_h \mathbf{v}\|_{0,K} + h_K |\mathbf{v} - \mathbf{I}_h \mathbf{v}|_{1,K} \le C\, h_K^2\, |\mathbf{v}|_{2,K} \qquad \forall\, \mathbf{v} \in \mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega). \tag{5.10}$$

Proof. For the proof see [13]. ∎

LEMMA 5.4. *Let $(\mathbf{u}, p) \in (\mathbf{H}_0^1(\Omega) \cap \mathbf{H}^2(\Omega)) \times (L_0^2(\Omega) \cap H^1(\Omega))$ be the weak solution of (5.1) and let $(\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h$ be the discrete solution of (5.2). Assume that $\tau_K \sim h_K^2$ and $\gamma_E \sim h_E^{-1}$. Then, for the canonical interpolant $\mathbf{I}_h : \mathbf{H}_0^1(\Omega) \to \mathbf{V}_h$ and the L^2 projection $J_h : L_0^2(\Omega) \to Q_h$ there is a constant C independent of ν such that*

$$|A\left((\mathbf{u} - \mathbf{I}_h \mathbf{u}, p - J_h p), (\mathbf{w}_h, r_h)\right)| \le C\, h (\|\mathbf{u}\|_2 + \|p\|_1)\, |||(\mathbf{w}_h, r_h)||| \tag{5.11}$$

for all $(\mathbf{w}_h, r_h) \in \mathbf{V}_h \times Q_h$.

Proof. Taking into consideration the definition of $|||\cdot|||$, we estimate each term in $A(\cdot,\cdot)$ separately. The estimate

$$|a_h^\sigma(\mathbf{u} - \mathbf{I}_h \mathbf{u}, \mathbf{w}_h)| \le C\, h\, \|\mathbf{u}\|_2\, |||(\mathbf{w}_h, r_h)|||$$

is standard. Using elementwise integration by parts, we obtain

$$b_h(\mathbf{u} - \mathbf{I}_h \mathbf{u}, \mathbf{w}_h) = \sum_{E \in \mathcal{E}_h} \langle \mathbf{b} \cdot \mathbf{n}_E [\mathbf{w}_h]_E, \{\mathbf{u} - \mathbf{I}_h \mathbf{u}\}_E \rangle_E - \sum_{K \in \mathcal{T}_h} (\mathbf{u} - \mathbf{I}_h \mathbf{u}, (\mathbf{b} \cdot \nabla) \mathbf{w}_h)_K$$

(see also the proof of Lemma 5.1). The first term on the right-hand side is estimated by

$$\begin{aligned} &\Big| \sum_{E \in \mathcal{E}_h} \langle \mathbf{b} \cdot \mathbf{n}_E [\mathbf{w}_h]_E, \{\mathbf{u} - \mathbf{I}_h \mathbf{u}\}_E \rangle_E \Big| \\ &\qquad \le C \sum_{E \in \mathcal{E}_h} \gamma_E^{-1/2} h_E^{3/2} \|\mathbf{u}\|_{2,\omega(E)}\, \gamma_E^{1/2} \|[\mathbf{w}_h]_E\|_{0,E} \\ &\qquad \le C\, h^2\, \|\mathbf{u}\|_2 \sqrt{j_h(\mathbf{w}_h, \mathbf{w}_h)} \end{aligned}$$

and the second one by

$$\left|\sum_{K\in\mathcal{T}_h}(\mathbf{u}-\mathbf{I}_h\mathbf{u},(\mathbf{b}\cdot\nabla)\mathbf{w}_h)_K\right| \le C\sum_{K\in\mathcal{T}_h}\tau_K^{-1/2}h_K^2\|\mathbf{u}\|_{2,K}\,\tau_K^{1/2}\|(\mathbf{b}\cdot\nabla)\mathbf{w}_h\|_{0,K}$$

$$\le C\,h\,\|\mathbf{u}\|_2\left(\sum_{K\in\mathcal{T}_h}\tau_K\|(\mathbf{b}\cdot\nabla)\mathbf{w}_h\|_{0,K}^2\right)^{1/2}.$$

The next expression is

$$|j_h(\mathbf{u}-\mathbf{I}_h\mathbf{u},\mathbf{w}_h)| \le C\sum_{E\in\mathcal{E}_h}\gamma_E^{1/2}h_E^{3/2}\|\mathbf{u}\|_{2,\omega(E)}\;\gamma_E^{1/2}\|[\mathbf{w}_h]_E\|_{0,E}$$

$$\le C\,h\,\|\mathbf{u}\|_2\,\sqrt{j_h(\mathbf{w}_h,\mathbf{w}_h)}$$

followed by

$$\sum_{K\in\mathcal{T}_h}\tau_K\big((\mathbf{b}\cdot\nabla)(\mathbf{u}-\mathbf{I}_h\mathbf{u}),(\mathbf{b}\cdot\nabla)\mathbf{w}_h\big)_K \le C\,h^2\,\|\mathbf{u}\|_2\left(\sum_{K\in\mathcal{T}_h}\tau_K\|(\mathbf{b}\cdot\nabla)\mathbf{w}_h\|_{0,K}^2\right)^{1/2}.$$

The orthogonality of the L^2 projection J_h and the property that any discretely divergence-free function is divergence-free on each cell yield that the last two terms become zero; i.e.,

$$(p-J_hp,\nabla_h\cdot\mathbf{w}_h)=0 \qquad \forall\mathbf{w}_h\in\mathbf{V}_h,$$

$$(r_h,\nabla_h\cdot(\mathbf{u}-\mathbf{I}_h\mathbf{u}))=0 \qquad \forall r_h\in Q_h.$$

Collecting all estimates, we get the statement of the lemma. ∎

THEOREM 5.5. *Let* $(\mathbf{u},p)\in(\mathbf{H}_0^1(\Omega)\cap\mathbf{H}^2(\Omega))\times(L_0^2(\Omega)\cap H^1(\Omega))$ *be the weak solution of* (5.1) *and let* $(\mathbf{u}_h,p_h)\in\mathbf{V}_h\times Q_h$ *be the discrete solution of* (5.2). *Assume that* $\tau_K\sim h_K^2$ *and* $\gamma_E\sim h_E^{-1}$. *Then, there is a positive constant* C *independent of* ν *such that*

$$|||(\mathbf{u}-\mathbf{u}_h,p-p_h)||| \le C\,h\,(\|\mathbf{u}\|_2+\|p\|_1). \tag{5.12}$$

Proof. Starting with Lemma 5.1 we have

$$|||(\mathbf{u}_h-\mathbf{I}_h\mathbf{u},p_h-J_hp)||| \le \frac{1}{\beta}\sup_{(\mathbf{w}_h,r_h)\in\mathbf{V}_h\times Q_h}\frac{A((\mathbf{u}_h-\mathbf{I}_h\mathbf{u},p_h-J_hp),(\mathbf{w}_h,r_h))}{|||(\mathbf{w}_h,r_h)|||}$$

$$\le \frac{1}{\beta}\sup_{(\mathbf{w}_h,r_h)\in\mathbf{V}_h\times Q_h}\frac{A((\mathbf{u}_h-\mathbf{u},p_h-p),(\mathbf{w}_h,r_h))}{|||(\mathbf{w}_h,r_h)|||}$$

$$+\frac{1}{\beta}\sup_{(\mathbf{w}_h,r_h)\in\mathbf{V}_h\times Q_h}\frac{A((\mathbf{u}-\mathbf{I}_h\mathbf{u},p-J_hp),(\mathbf{w}_h,r_h))}{|||(\mathbf{w}_h,r_h)|||}.$$

Now, the first term can be bounded by Lemma 5.2 and the second one by Lemma 5.4. It remains to apply the triangle inequality

$$|||(\mathbf{u}-\mathbf{u}_h,p-p_h)||| \le |||(\mathbf{u}-\mathbf{I}_h\mathbf{u},p-J_hp)||| + |||(\mathbf{u}_h-\mathbf{I}_h\mathbf{u},p_h-J_hp)|||$$

and the approximation properties of the interpolation operators $\mathbf{I}_h$ and J_h. ∎

Remark. According to the definition of the triple norm we have for $\sigma > 0$ an additional control *uniformly with respect to* ν over the L^2 norm of the velocity and the pressure. For $\sigma = 0$ we lose this control for $\nu \to 0$.

Remark. In the SUPG method the additional stabilizing term

$$\sum_{K \in \mathcal{T}_h} \gamma_K (\nabla \cdot \mathbf{u}_h, \nabla \cdot \mathbf{v}_h)_K$$

is often used [17, 31]. In our case of the Crouzeix–Raviart element, discretely divergence-free functions are piecewise divergence-free, therefore this term vanishes.

Remark. Often the SUPG parameter in the SUPG method is chosen in the advective regime as $\tau_K \sim h_K$, which is the correct choice for equal-order interpolation [9, 14, 31]. However, using inf-sup stable elements with different-order interpolation in the SUPG method, we have to take $\tau_K \sim h_K^2$ [5].

Numerical test. We consider the generalized Oseen equations (5.2) in $\Omega = (0,1)^2$ with the prescribed solution

$$\mathbf{u} = \begin{pmatrix} 2x^2(1-x)^2y(1-y)(1-2y) \\ -2y^2(1-y)^2x(1-x)(1-2x) \end{pmatrix}, \quad p = x^3 + y^3 - 0.5,$$

the convection field

$$\mathbf{b} = \begin{pmatrix} \sin(x)\sin(y) \\ \cos(x)\cos(y) \end{pmatrix},$$

and with the parameters $\nu = 10^{-3}$, $\sigma = 100$. The choice of σ corresponds to a length of the time step of 0.01 in the nonstationary Navier–Stokes equations.

The coarsest grid in the computations (level 0) consists of two triangles with the common edge from $(0,0)$ to $(1,1)$. On level 7, the system has 98 816 velocity degrees of freedom (including Dirichlet nodes) and 32 768 pressure degrees of freedom.

Results for different choices of the parameter γ_E in the jump term $j_h(\mathbf{u}_h, \mathbf{v}_h)$ are presented in Tables 5.1 and 5.2. In Table 5.1, computations without this jump term ($\gamma_E = 0$) and with the appropriate choice ($\gamma_E = 1$) known from scalar convection-diffusion equations (cf. [22]) are given. It can be observed that the order of convergence with respect to the natural norms for the Oseen equations is far below the optimal one in the convection-dominated regime; even an increase of errors occurs. However, optimal orders are obtained for the choice $\gamma_E = 1/h_E$, which is in agreement with our theoretical results presented in this section; see Table 5.2. In addition, the optimal order of convergence in the $|||\cdot|||$ norm, (5.12), can be seen.

TABLE 5.1
Results obtained with $\gamma_E = 0$ and $\gamma_E = 1$.

	$\gamma_E = 0$				$\gamma_E = 1$			
Level	$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_0$	Order	$\|p - p_h\|_0$	Order	$\|\nabla(\mathbf{u} - \mathbf{u}_h)\|_0$	Order	$\|p - p_h\|_0$	Order
3	3.057e-1	—	2.790e-1	—	2.211e-1	—	2.185e-1	—
4	5.899e-1	−0.949	2.625e-1	0.088	3.377e-1	−0.611	1.601e-1	0.449
5	1.083e+0	−0.876	2.487e-1	0.078	4.549e-1	−0.430	1.077e-1	0.572
6	1.748e+0	−0.691	2.166e-1	0.200	5.336e-1	−0.230	6.433e-2	0.744
7	2.205e+0	−0.335	1.474e-1	0.555	5.486e-1	−0.040	3.410e-2	0.916

TABLE 5.2
Results obtained with $\gamma_E = 1/h_E$.

Level	$\|\nabla(\mathbf{u}-\mathbf{u}_h)\|_0$	Order	$\|p-p_h\|_0$	Order	$\|\|\|(\mathbf{u}-\mathbf{u}_h, p-p_h)\|\|\|$	Order
3	8.610e-2	—	1.176e-1	—	1.179e+0	—
4	5.332e-2	0.691	4.389e-2	1.422	4.409e-1	1.418
5	2.775e-2	0.942	1.776e-2	1.306	1.789e-1	1.301
6	1.386e-2	1.002	8.196e-3	1.115	8.270e-2	1.113
7	6.895e-3	1.001	4.053e-3	1.021	4.090e-2	1.021

REFERENCES

[1] M. I. ASENSIO, A. RUSSO, AND G. SANGALLI, *The residual-free bubble numerical method with quadratic elements*, Math. Models Methods Appl. Sci., 14 (2004), pp. 641–661.

[2] R. BECKER AND M. BRAACK, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 173–199.

[3] R. BECKER AND M. BRAACK, *A two-level stabilization scheme for the Navier-Stokes equations*, in Numerical Mathematics and Advanced Applications, M. Feistauer, V. Dolejší, P. Knobloch, and K. Najzar, eds., Springer-Verlag, Berlin, 2004, pp. 123–130.

[4] M. BRAACK AND E. BURMAN, *Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method*, SIAM J. Numer. Anal., 43 (2006), pp. 2544–2566.

[5] M. BRAACK, E. BURMAN, V. JOHN, AND G. LUBE, *Stabilized finite element methods for the generalized Oseen problem*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 853–866.

[6] F. BREZZI AND M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.

[7] F. BREZZI, L. D. MARINI, AND A. RUSSO, *Applications of pseudo residual-free bubbles to the stabilization of convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 51–63.

[8] F. BREZZI, L. D. MARINI, AND A. RUSSO, *On the choice of stabilizing subgrid for convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 194 (2005), pp. 127–148.

[9] A. N. BROOKS AND T. J. R. HUGHES, *Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 32 (1982), pp. 199–259.

[10] E. BURMAN AND P. HANSBO, *The edge stabilization method for finite elements in cfd*, in Numerical Mathematics and Advanced Applications, M. Feistauer, V. Dolejší, P. Knobloch, and K. Najzar, eds., Springer-Verlag, Berlin, 2004, pp. 196–203.

[11] E. BURMAN AND P. HANSBO, *A stabilized non-conforming finite element method for incompressible flow*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 2881–2899.

[12] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis: Finite Element Methods, Vol. 2, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.

[13] M. CROUZEIX AND P.-A. RAVIART, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*. I, Rev. Française Automat. Informat. Recherche Opérationnelle Sér. Rouge, 7 (1973), pp. 33–76.

[14] L. P. FRANCA AND S. L. FREY, *Stabilized finite element methods:* II. *The incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 99 (1992), pp. 209–233.

[15] L. P. FRANCA AND A. NESLITURK, *On a two-level finite element method for the incompressible Navier-Stokes equations*, Internat. J. Numer. Methods Engrg., 52 (2001), pp. 433–453.

[16] L. P. FRANCA, A. NESLITURK, AND M. STYNES, *On the stability of residual-free bubbles for convection-diffusion problems and their approximation by a two-level finite element method*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 35–49.

[17] T. GELHARD, G. LUBE, M. A. OLSHANSKII, AND J.-H. STARCKE, *Stabilized finite element schemes with LBB-stable elements for incompressible flows*, J. Comput. Appl. Math., 177 (2005), pp. 243–267.

[18] V. GIRAULT AND P.-A. RAVIART, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986.

[19] V. GRAVEMEIER, W. A. WALL, AND E. RAMM, *A three-level finite element method for the instationary incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1323–1366.

[20] V. Gravemeier, W. A. Wall, and E. Ramm, *Large eddy simulation of turbulent incompressible flows by a three-level finite element method*, Internat. J. Numer. Methods Fluids, 48 (2005), pp. 1067–1099.

[21] T. J. R. Hughes, L. P. Franca, and M. Balestra, *A new finite element formulation for computational fluid dynamics:* V. *Circumventing the Babuška–Brezzi condition: A stable Petrov-Galerkin formulation of the Stokes problem accommodating equal-order interpolations*, Comput. Methods Appl. Mech. Engrg., 59 (1986), pp. 85–99.

[22] V. John, G. Matthies, F. Schieweck, and L. Tobiska, *A streamline-diffusion method for nonconforming finite element approximations applied to convection-diffusion problems*, Comput. Methods Appl. Mech. Engrg., 166 (1998), pp. 85–97.

[23] V. John, J. Maubach, and L. Tobiska, *Nonconforming streamline-diffusion-finite-element-methods for convection-diffusion problems*, Numer. Math., 78 (1997), pp. 165–188.

[24] P. Knobloch and L. Tobiska, *Stabilization methods of bubble type for the Q_1/Q_1-element applied to the incompressible Navier-Stokes equations*, M2AN Math. Model Numer. Anal., 34 (2000), pp. 85–107.

[25] G. Lube and G. Rapin, *Residual-based stabilized higher-order FEM for a generalized Oseen problem*, Math. Models Methods Appl. Sci., 16 (2006), pp. 949–966.

[26] G. Lube and L. Tobiska, *A nonconforming finite element method of streamline diffusion type for the incompressible Navier–Stokes equations*, J. Comput. Math., 8 (1990), pp. 147–158.

[27] G. Matthies and L. Tobiska, *The streamline diffusion method for conforming and nonconforming finite elements of lowest order applied to convection-diffusion problems*, Computing, 66 (2001), pp. 343–364.

[28] A. Russo, *Bubble stabilization of finite element methods for the linearized incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 132 (1996), pp. 335–343.

[29] F. Schieweck and L. Tobiska, *A nonconforming finite element method of upstream type applied to the stationary Navier-Stokes equation*, M2AN Math. Model Numer. Anal., 23 (1989), pp. 627–647.

[30] F. Schieweck and L. Tobiska, *An optimal order error estimate for an upwind discretization of the Navier-Stokes equation*, Numer. Methods Partial Differential Equations, 12 (1996), pp. 407–421.

[31] L. Tobiska and R. Verfürth, *Analysis of a streamline diffusion finite element method for the Stokes and Navier–Stokes equations*, SIAM J. Numer. Anal., 33 (1996), pp. 107–127.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2408–2441

A NEW STICKY PARTICLE METHOD FOR PRESSURELESS GAS DYNAMICS*

ALINA CHERTOCK†, ALEXANDER KURGANOV‡, AND YURII RYKOV§

Abstract. We first present a new sticky particle method for the system of pressureless gas dynamics. The method is based on the idea of sticky particles, which seems to work perfectly well for the models with point mass concentrations and strong singularity formations. In this method, the solution is sought in the form of a linear combination of δ-functions, whose positions and coefficients represent locations, masses, and momenta of the particles, respectively. The locations of the particles are then evolved in time according to a system of ODEs, obtained from a weak formulation of the system of PDEs. The particle velocities are approximated in a special way using global conservative piecewise polynomial reconstruction technique over an auxiliary Cartesian mesh. This velocities correction procedure leads to a desired interaction between the particles and hence to clustering of particles at the singularities followed by the merger of the clustered particles into a new particle located at their center of mass. The proposed sticky particle method is then analytically studied. We show that our particle approximation satisfies the original system of pressureless gas dynamics in a weak sense, but only within a certain residual, which is rigorously estimated. We also explain why the relevant errors should diminish as the total number of particles increases. Finally, we numerically test our new sticky particle method on a variety of one- and two-dimensional problems as well as compare the obtained results with those computed by a high-resolution finite-volume scheme. Our simulations demonstrate the superiority of the results obtained by the sticky particle method that accurately tracks the evolution of developing discontinuities and does not smear the developing δ-shocks.

Key words. nonstrictly hyperbolic systems of conservation laws, pressureless gas dynamics, mass concentration, strong singularities, δ-shock, sticky particle method

AMS subject classifications. 65M25, 65M12, 35L65, 35L67

DOI. 10.1137/050644124

1. Introduction.

We consider the Euler equations of pressureless gas dynamics:

$$\mathbf{w}_t + \nabla_{\mathbf{x}} \cdot (\mathbf{u} \otimes \mathbf{w}) = \mathbf{0}. \tag{1.1}$$

Here, $\mathbf{x} := (x, y, \ldots)$ is an n-dimensional vector of spatial variables, $\mathbf{u} := (u, v, \ldots)$ is the corresponding velocity vector, and $\mathbf{w} \equiv (w^1, w^2, w^3, \ldots)^T := (\rho, \rho u, \rho v, \ldots)^T$ is the $(n+1)$-dimensional vector of unknown function, where ρ is the density.

This system arises in the modeling of the formation of large-scale structures in the universe [24]. It can be formally obtained as the limit of the isotropic Euler equations of gas dynamics as pressure tends to zero or as the macroscopic limit of a Boltzmann equation when the Maxwellian has zero temperature.

*Received by the editors November 2, 2005; accepted for publication (in revised form) April 6, 2007; published electronically November 21, 2007.
http://www.siam.org/journals/sinum/45-6/64412.html

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695 (chertock@math.ncsu.edu). This author's work was supported in part by NSF grant DMS-0410023.

‡Mathematics Department, Tulane University, New Orleans, LA 70118, and Department of Mathematics, University of Michigan, Ann Arbor, MI 48109 (kurganov@math.tulane.edu). This author's work was supported in part by NSF grants DMS-0310585 and DMS-0610430.

§Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, 125047 Moscow, Russia (rykov@Keldysh.ru). This author's work was supported in part by Russian Foundation for Basic Research grant 030100189 and by Program 01 of the Division of Mathematical Sciences of the Russian Academy of Sciences.

Even in the simplest one-dimensional (1-D) case, the system (1.1), which can be rewritten as

$$\begin{cases} \rho_t + (\rho u)_x = 0, \\ (\rho u)_t + (\rho u^2)_x = 0, \end{cases} \tag{1.2}$$

is mathematically challenging since it is nonstrictly hyperbolic and its Jacobian is not diagonalizable. For smooth solutions, the system (1.2) is equivalent to

$$\rho_t + (\rho u)_x = 0, \tag{1.3}$$

$$u_t + u u_x = 0. \tag{1.4}$$

Notice that (1.4) is the inviscid Burgers equation, which is, in fact, decoupled from (1.3). It is well known that the solution of the initial-value problem associated with (1.4), as long as it stays smooth, can be easily obtained by the method of characteristics. The density ρ can then be determined from (1.3), which becomes a linear transport equation. However, if the initial velocity $u(x,0)$ is not monotone increasing, the characteristics will intersect within a finite time, and the solution will lose its initial smoothness, and thus it must be understood in a weak sense. As in the general theory of weak solutions of hyperbolic systems of conservation laws, one has to introduce discontinuity lines. Let $x = \xi(t)$ be such a line and assume that the solution accepts finite values $u^{\pm} := u(\xi(t) \pm 0, t)$ and $\rho^{\pm} := \rho(\xi(t) \pm 0, t)$ from both sides of discontinuity. The jumps then must satisfy the Rankine–Hugoniot conditions, namely,

$$\begin{cases} \xi'(t)\,(\rho^+ - \rho^-) = \rho^+ u^+ - \rho^- u^-, \\ \xi'(t)\,(\rho^+ u^+ - \rho^- u^-) = \rho^+ \left(u^+\right)^2 - \rho^- \left(u^-\right)^2. \end{cases}$$

After eliminating ξ' from this system, one obtains $\rho^+ \rho^- \left(u^+ - u^-\right)^2 = 0$, which implies $u^+ = u^-$. Therefore, in order to support the shock discontinuity in the velocity field, the density must have a stronger (than a shock) singularity at $x = \xi'(t)$. Since in the Burgers equation, the characteristic lines impinge each other and thus, as part of the system (1.2), cause a mass concentration at the velocity discontinuity line, resulting in the formation of a δ-type singularity in the density field there.

The two-dimensional (2-D) version of (1.1) reads as

$$\begin{cases} \rho_t + (\rho u)_x + (\rho v)_y = 0, \\ (\rho u)_t + (\rho u^2)_x + (\rho u v)_y = 0, \\ (\rho v)_t + (\rho u v)_x + (\rho v^2)_y = 0. \end{cases} \tag{1.5}$$

Compared to the 1-D case, solutions of the 2-D system have a similar but essentially more sophisticated mechanism of singularities formation due to the dimensionality factor: strong singularities may now be formed either along surfaces or at separate points (we expect that in the three space dimensions the situation is even more complex). The system (1.5) has been intensively studied at the theoretical level (see, e.g., [2, 4, 6, 7, 8, 20, 21]). However, no more or less complete analytical results concerning the existence and uniqueness of solutions in the 2-D case are available. This is primarily related to the difficulties in the theoretical description of the collision of 2-D shocks. (See section 3.2 for an extensive numerical study of this phenomenon.)

Formation and further evolution of singular shocks, their interactions as well as the emergence of vacuum states, make development of numerical methods for the

system (1.1) a challenging problem. A numerical method based on the movement of a system of particles was introduced in [19]. Several finite-volume [17], kinetic [2, 3, 5], and relaxation [1] methods have been recently proposed. These methods are able to reasonably accurately capture δ-shocks, but their applicability is rather limited; for example, most of these methods do not work well for problems where the velocities change sign in regions where the density varies smoothly [17].

We develop a simple, efficient, and low-dissipative sticky particle method for pressureless gas dynamics. The derivation of our method is based on a weak formulation of the system (1.5) and can be viewed as a practical implementation of the sticky particle method from [4]. We first approximate $\mathbf{w}$ by a collection of N particles, located at $(x_i^N(t), y_i^N(t))$, $i = 1, \dots, N$, at time t, and carrying fixed masses and momenta. The particle locations are then evolved according to the corresponding system of ODEs, derived by plugging the particle approximation into a weak form of (1.1). In order to prevent the situation, in which approaching particles simply pass by each other without any interaction (such an undesirable situation is obviously impossible in the 1-D case, but is almost unavoidable in the 2-D case), we divide the computational domain into a set of auxiliary cells and compute the total mass and momenta in each cell. The particle velocities are then approximated using the global conservative piecewise polynomial interpolants of ρ, ρu, and ρv, constructed over an auxiliary Cartesian mesh. This way an interaction of all particles located in the same cell is guaranteed. When the particles get even closer to each other, we unite them into a new particle, located at the center of mass of the original clustered particles. The mass (momentum) of the new particle is simply the sum of the masses (corresponding momenta) of the replaced particles, and the velocities of the new particle are uniquely determined from the conservation requirements. This particle merger procedure results in mass concentration, which is an essential property of pressureless gases.

We would like to note that our 2-D sticky particle method can be extended to any number of space dimensions in a rather straightforward manner. In this paper, we restrict our consideration to the 1-D and 2-D cases only, since, to the best of our knowledge, no analytical results on three-dimensional (3-D) pressureless gas dynamics system are available, and it is therefore hard to set up convincing 3-D numerical experiments.

We test our method on a number of 1-D and 2-D numerical examples, in which we compare the results obtained by the new (nondissipative) sticky particle method and by the (dissipative) second-order central-upwind scheme from [11]. The latter scheme is a high-resolution Godunov-type finite-volume method that belongs to a family of central schemes, which may serve as "black-box" solvers for multidimensional hyperbolic systems of conservation laws. The prototype of modern central schemes is the first-order Lax–Friedrichs scheme [9, 16], which is the most universal method for solving (multidimensional systems of) time-dependent PDEs. However, its excessive numerical dissipation prevents sharp resolution and therefore in practice one has to use higher-order schemes. The first high-resolution nonoscillatory central scheme—the second-order Nessyahu–Tadmor scheme—was proposed in [18]. The amount of numerical dissipation present in projection-evolution central schemes was further reduced by incorporating some more upwinding information on local speeds of propagation into the evolution step [12, 14] (the resulting schemes thus have been referred to as central-upwind schemes) and, more recently, by enhancing the accuracy of the projection step [11, 13]. We note that the only upwinding information required by the central-upwind schemes is the eigenvalues of the Jacobians, and therefore application of these schemes to nonstrictly hyperbolic systems like (1.5) is straightforward.

The paper is organized as follows. We start in section 2 by introducing the new sticky particle method for the system (1.5). We then describe, in section 2.1, the velocity correction procedure and, in section 2.2, an algorithm of the unification of clustering particles. The main analytical result in section 2.3 is Theorem 2.1, where we show that even though our particle solution fails to satisfy (1.1) in a weak sense defined in [21], the relevant errors can be rigorously estimated. We then provide a heuristic justification why these errors tend to zero as $N \to \infty$. We conclude in section 3 with several 1-D and 2-D numerical examples and demonstrate that the new method accurately tracks the evolution of developing discontinuities. We also compare solutions computed by the sticky particle method with the corresponding solutions computed using the second-order semidiscrete central-upwind scheme, developed in [11, 12, 14]. A brief description of the central-upwind scheme for the pressureless gas dynamics system (1.5) is provided in Appendix A.

2. Derivation of the sticky particle method. We consider the system (1.1) subject to the compactly supported (or periodic) initial data,

$$\mathbf{w}(\mathbf{x},0) \equiv (\rho(\mathbf{x},0), \rho u(\mathbf{x},0), \rho v(\mathbf{x},0))^T, \quad \mathbf{x} := (x,y)^T, \tag{2.1}$$

and look for the solution of the initial-value problem (1.1), (2.1) in the particle form,

$$\mathbf{w}^N(\mathbf{x},t) = \sum_{i=1}^{N} \boldsymbol{\alpha}_i(t)\delta(\mathbf{x} - \mathbf{x}_i^N(t)),\ \mathbf{x}_i^N := (x_i^N, y_i^N)^T,\ \boldsymbol{\alpha}_i = (m_i, m_i u_i, m_i v_i)^T. \tag{2.2}$$

Here, N is a total number of particles, $\mathbf{x}_i^N(t)$ is the location of the ith particle at time t, and $m_i, m_i u_i$, and $m_i v_i$ are its mass, the x-, and the y-momenta, respectively.

In order to study the particle time evolution, we plug (2.2) into the weak formulation of the system (1.1),

$$\int_0^\infty \iint_{\mathbb{R}^2} \left\{\mathbf{w}^N \cdot [\boldsymbol{\varphi}_t + u\boldsymbol{\varphi}_x + v\boldsymbol{\varphi}_y]\right\} d\mathbf{x}dt - \iint_{\mathbb{R}^2} \mathbf{w}^N(\mathbf{x},0)\cdot\boldsymbol{\varphi}(\mathbf{x},0)\,d\mathbf{x} = 0, \tag{2.3}$$

where $\boldsymbol{\varphi}$ is an arbitrary C_0^1 test function. As a result, (2.3) reduces to

$$\sum_{i=1}^{N} \int_0^\infty \boldsymbol{\alpha}_i(t)\cdot\left\{\boldsymbol{\varphi}_t + u\boldsymbol{\varphi}_x + v\boldsymbol{\varphi}_y\right\}\Big|_{(\mathbf{x},t)=(\mathbf{x}_i^N(t),t)} dt - \sum_{i=1}^{N} \boldsymbol{\alpha}_i(0)\cdot\boldsymbol{\varphi}(\mathbf{x}_i^N(0),0) = 0, \tag{2.4}$$

which should be satisfied for any $\boldsymbol{\varphi}$. Evolving particle locations according to the following system of ODEs:

$$\frac{dx_i^N(t)}{dt} = u(\mathbf{x}_i^N(t),t), \quad \frac{dy_i^N(t)}{dt} = v(\mathbf{x}_i^N(t),t), \qquad i = 1,\dots,N, \tag{2.5}$$

and integrating by parts, we rewrite (2.4) as

$$\sum_{i=1}^{N} \int_0^\infty \frac{d\boldsymbol{\alpha}_i(t)}{dt}\cdot\boldsymbol{\varphi}(\mathbf{x}_i^N(t),t)\,dt = 0.$$

The last equation implies

$$\frac{d\boldsymbol{\alpha}_i(t)}{dt} = 0, \quad i = 1,\dots,N, \tag{2.6}$$

that is, the particle weights remain constant in time. Thus, the weights can be determined from the initial conditions, for instance, in the following manner. We divide the computational domain Ω into N subdomains Ω_i, $i = 1, \dots, N$, and place an ith particle with

$$\boldsymbol{\alpha}_i := \iint\limits_{\Omega_i} \mathbf{w}(\mathbf{x}, 0)\, d\mathbf{x}$$

into the center of Ω_i, denoted by $\mathbf{x}_i^N(0) \equiv (x_i^N(0), y_i^N(0))$, which will serve as initial data for the ODE system (2.5).

2.1. Particle velocities. In order to be able to solve the system of ODEs (2.5), one would need to recover the particle velocities at any given time moment. The simplest (and the least dissipative) way to compute the velocities is to divide the corresponding particle momenta by its mass, that is, by taking

$$u_i \equiv u(\mathbf{x}_i^N(t), t) := \frac{m_i u_i}{m_i}, \quad v_i \equiv v(\mathbf{x}_i^N(t), t) := \frac{m_i v_i}{m_i}. \tag{2.7}$$

In fact, this means that every particle travels with constant velocity until it collides with another particle (see section 2.2). This approach can be rigorously justified through the weak formulation (2.3) and it seems to work perfectly in the 1-D case, in which collision of approaching particles is unavoidable. However, in the 2-D case, the probability of collision of two particles approaching the same singularity curve is zero unless a special symmetry in initial particle locations has been imposed (see Example 5 in section 3.2).

We propose an alternative way of particle velocities reconstruction, which is independent of an initial placement of particles. Our approach is based on a global piecewise polynomial reconstruction technique, which is widely used in finite-volume framework (see Appendix A and the references therein). To adopt this technique to a mesh-free particle method we introduce an auxiliary Cartesian grid (which may vary in time). The grid should be adapted to the particle distribution so that the number of particles in every cell is about the same. In our numerical experiments, we have used the simplest strategy: we have adapted the auxiliary grid to the initial (uniform) particle distribution only by taking the size of the cells to be twice larger than the distance between the particles. A more sophisticated adaptation strategy may lead to more accurate results, but its optimization may substantially increase the complexity of the proposed sticky particle method.

Taking a simple uniform auxiliary grid, $x_j \equiv j\Delta x$, $y_k \equiv k\Delta y$, we first compute the cell averages of the conserved quantities at time t,

$$\bar{\mathbf{w}}_{j,k}(t) = \frac{1}{\Delta x \Delta y} \sum_{i:\, \mathbf{x}_i^N(t) \in I_{j,k}} \boldsymbol{\alpha}_i, \qquad I_{j,k} = [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [y_{k-\frac{1}{2}}, y_{k+\frac{1}{2}}]. \tag{2.8}$$

Using these cell averages, we then reconstruct a nonoscillatory piecewise polynomial interpolant of an appropriate order of accuracy, denoted by

$$\widetilde{\mathbf{w}}(\mathbf{x}, t) := (w^1(\mathbf{x}, t), w^2(\mathbf{x}, t), w^3(\mathbf{x}, t))^T,$$

which is used to compute the particle velocities,

$$u_i := \frac{\widetilde{w}^2(\mathbf{x}_i^N(t), t)}{\widetilde{w}^1(\mathbf{x}_i^N(t), t)}, \quad v_i := \frac{\widetilde{w}^3(\mathbf{x}_i^N(t), t)}{\widetilde{w}^1(\mathbf{x}_i^N(t), t)}. \tag{2.9}$$

Notice that in order to ensure that no mass (momentum) is artificially lost (created) at this step, the reconstruction must be performed in a conservative manner, namely, one should guarantee that

$$\sum_{i:\,\mathbf{x}_i^N(t)\in I_{j,k}} \widetilde{\mathbf{w}}(\mathbf{x}_i^N(t),t)=\bar{\mathbf{w}}_{j,k}(t).$$

We achieve the conservation (in fact, while the mass is conserved exactly, only approximate momentum conservation is guaranteed; see the computation in section 2.3) by taking $\widetilde{\mathbf{w}}$ to be a second-order accurate piecewise linear reconstruction centered at the center of mass of the particles located in the $I_{j,k}$ cell,

$$\begin{aligned}\widetilde{\mathbf{w}}(x,y,t)=\bar{\mathbf{w}}_{j,k}&+(\mathbf{s}_x)_{j,k}(x-x_{j,k}^{\mathrm{CM}}(t))\\&+(\mathbf{s}_y)_{j,k}(y-y_{j,k}^{\mathrm{CM}}(t))\quad\text{for}\quad(x,y)\in I_{j,k},\end{aligned}\tag{2.10}$$

where the coordinates of the center of mass are

$$x_{j,k}^{\mathrm{CM}}(t):=\frac{\displaystyle\sum_{i:\,\mathbf{x}_i^N(t)\in I_{j,k}} m_i x_i^N(t)}{\displaystyle\sum_{i:\,\mathbf{x}_i^N(t)\in I_{j,k}} m_i},\quad y_{j,k}^{\mathrm{CM}}(t):=\frac{\displaystyle\sum_{i:\,\mathbf{x}_i^N(t)\in I_{j,k}} m_i y_i^N(t)}{\displaystyle\sum_{i:\,\mathbf{x}_i^N(t)\in I_{j,k}} m_i},\tag{2.11}$$

and the slopes $(\mathbf{s}_x)_{j,k}$ and $(\mathbf{s}_y)_{j,k}$ are (at least) first-order approximations of the derivatives $\mathbf{w}_x(x_{j,k}^{\mathrm{CM}}(t),y_{j,k}^{\mathrm{CM}}(t))$ and $\mathbf{w}_y(x_{j,k}^{\mathrm{CM}}(t),y_{j,k}^{\mathrm{CM}}(t))$, respectively.

Finally, in order to ensure a nonoscillatory nature of the reconstruction (2.10), the slopes $(\mathbf{s}_x)_{j,k}$ and $(\mathbf{s}_y)_{j,k}$ should be computed using a nonlinear limiter. In our numerical experiments, we have used the minmod limiter applied in the following way.

Let us take, for example, the first component of $\mathbf{w}$ (density) and consider the four planes, denoted by $\pi_{j,k}^{\mathrm{NE}},\pi_{j,k}^{\mathrm{NW}},\pi_{j,k}^{\mathrm{SE}},\pi_{j,k}^{\mathrm{SW}}$, that pass through the following four triplets of points:

$$\begin{aligned}\pi_{j,k}^{\mathrm{NE}}&:\left\{(x_{j,k}^{\mathrm{CM}},y_{j,k}^{\mathrm{CM}},\bar{w}_{j,k}^1),\,(x_{j,k+1}^{\mathrm{CM}},y_{j,k+1}^{\mathrm{CM}},\bar{w}_{j,k+1}^1),\,(x_{j+1,k}^{\mathrm{CM}},y_{j+1,k}^{\mathrm{CM}},\bar{w}_{j+1,k}^1)\right\},\\ \pi_{j,k}^{\mathrm{NW}}&:\left\{(x_{j,k}^{\mathrm{CM}},y_{j,k}^{\mathrm{CM}},\bar{w}_{j,k}^1),\,(x_{j,k+1}^{\mathrm{CM}},y_{j,k+1}^{\mathrm{CM}},\bar{w}_{j,k+1}^1),\,(x_{j-1,k}^{\mathrm{CM}},y_{j-1,k}^{\mathrm{CM}},\bar{w}_{j-1,k}^1)\right\},\\ \pi_{j,k}^{\mathrm{SE}}&:\left\{(x_{j,k}^{\mathrm{CM}},y_{j,k}^{\mathrm{CM}},\bar{w}_{j,k}^1),\,(x_{j,k-1}^{\mathrm{CM}},y_{j,k-1}^{\mathrm{CM}},\bar{w}_{j,k-1}^1),\,(x_{j+1,k}^{\mathrm{CM}},y_{j+1,k}^{\mathrm{CM}},\bar{w}_{j+1,k}^1)\right\},\\ \pi_{j,k}^{\mathrm{SW}}&:\left\{(x_{j,k}^{\mathrm{CM}},y_{j,k}^{\mathrm{CM}},\bar{w}_{j,k}^1),\,(x_{j,k-1}^{\mathrm{CM}},y_{j,k-1}^{\mathrm{CM}},\bar{w}_{j,k-1}^1),\,(x_{j-1,k}^{\mathrm{CM}},y_{j-1,k}^{\mathrm{CM}},\bar{w}_{j-1,k}^1)\right\}\end{aligned}\tag{2.12}$$

(the dependence of $\{x_{j,k}^{\mathrm{CM}}\}$ and $\{y_{j,k}^{\mathrm{CM}}\}$ on t has been omitted here for briefness). We then denote the gradients of these planes by $((\pi_x)_{j,k}^{\mathrm{NE}},(\pi_y)_{j,k}^{\mathrm{NE}})^T$, $((\pi_x)_{j,k}^{\mathrm{NW}},(\pi_y)_{j,k}^{\mathrm{NW}})^T$, etc., and take the first component of the slopes in (2.10) to be

$$\begin{aligned}(s_x^1)_{j,k}&=\mathrm{minmod}\left((\pi_x)_{j,k}^{\mathrm{NE}},(\pi_x)_{j,k}^{\mathrm{NW}},(\pi_x)_{j,k}^{\mathrm{SE}},(\pi_x)_{j,k}^{\mathrm{SW}}\right),\\ (s_y^1)_{j,k}&=\mathrm{minmod}\left((\pi_y)_{j,k}^{\mathrm{NE}},(\pi_y)_{j,k}^{\mathrm{NW}},(\pi_y)_{j,k}^{\mathrm{SE}},(\pi_y)_{j,k}^{\mathrm{SW}}\right),\end{aligned}\tag{2.13}$$

where the minmod function is defined by

$$\text{minmod}(c_1,c_2,\ldots) := \begin{cases} \min(c_1,c_2,\ldots) & \text{if } \ c_i>0 \ \forall i,\\ \max(c_1,c_2,\ldots) & \text{if } \ c_i<0 \ \forall i,\\ 0 & \text{otherwise.}\end{cases} \tag{2.14}$$

The reconstructions for the other two fields of $\mathbf{w}$ (momenta) are obtained in a similar way.

Remarks.

1. It may happen that one of the planes in (2.12) is perpendicular to the (x,y)-plane or is not uniquely determined. Then this plane is not taken into account, and its gradient is excluded from the formulae for the slopes in (2.13).
2. As was mentioned in section 1, our velocity recovery procedure ensures that there is an interaction between the particles, located inside the same auxiliary grid cell. As is illustrated in our numerical experiments (see section 3.2), this leads to the desired clustering of particles at the singularities.

2.2. Unification of clustering particles. A major drawback of particle methods is that, in general, their resolution and efficiency significantly deteriorate when too many particles cluster near the same point at the singularity. To prevent such an undesired situation, we unite clustering particles according to the following algorithm. We choose a certain critical distance d_{cr} and as soon as the distance between any two particles gets smaller than the critical distance, we unite them into a new "heavier" particle.

Let us assume that at some time t, the distance between the ith and the jth particles, $|\mathbf{x}_i(t)-\mathbf{x}_j(t)|$, is smaller than d_{cr}. We then replace these two particles with a new one of the following total mass and momenta:

$$\boldsymbol{\alpha}_{\text{new}} = \boldsymbol{\alpha}_i + \boldsymbol{\alpha}_j, \tag{2.15}$$

located at the center of mass of the replaced particles, namely,

$$\mathbf{x}^N_{\text{new}} = \frac{m_i\mathbf{x}^N_i + m_j\mathbf{x}^N_j}{m_i+m_j}. \tag{2.16}$$

The velocities of the new particle are determined according to the procedure in section 2.1. After completing the replacement process (2.15)–(2.16), we check whether any other two particles are to be united, and if not, the remaining set of particles is further evolved in time according to (2.2), (2.5), (2.9) until another particle clustering occurs. Then, the unification procedure is repeated, and so forth.

Remark. The critical distance d_{cr} should be chosen experimentally. In all our numerical examples, except for Example 4, this distance was taken a quarter of a minimal initial distance between the particles (note that the initial distribution of particles is rather uniform in every numerical example below). In Example 4, d_{cr} was made proportional to the size of the shrinking support of the solution.

We would also like to stress that our numerical experiments clearly indicate that the presented sticky particle method does not seem to be sensitive to the choice of d_{cr}.

2.3. On convergence of the sticky particle method. In previous sections, a sequence of approximate solutions $\{\mathbf{w}^N\}_{N=1}^{\infty}$ of the system (1.1) for a fixed time interval $[0,T]$ has been constructed based on the dynamics of moving particles. In this section, we show that the measures $\mathbf{w}^N$ do not satisfy (1.1) in a weak sense.

Nevertheless, in Theorem 2.1, we obtain rigorous estimates for relevant errors and further discuss the heuristic justification why these errors tend to zero as $N \to \infty$.

In order to exactly formulate the theorem, let us first describe the interactions of moving particles in detail. Consider a time interval $[t_1, t_2] \subset [0, T]$, some number p of moving particles, and a time moment t_0 such that the particles evolve according to the ODE system (2.5)–(2.6) for $t \in [t_1, t_0)$ and $t \in (t_0, t_2]$, while at time t_0 the particles either coalesce (Case I) or change the velocities according to (2.9) (Case II).

For the considered group of p particles, $\mathcal{P}$, with the total mass

$$M := \sum_{i \in \mathcal{P}} m_i,$$

we introduce the following notation.

- Prior to $t = t_0$ we denote by
  $\boldsymbol{\alpha}_i = (m_i, m_i u_i, m_i v_i)^T$: weights of the particles,
  $(x_i(t), y_i(t))$: their locations at time $t < t_0$,
  $(x_i^0, y_i^0) = (x_i(t_0), y_i(t_0))$: final locations of the particles at time $t = t_0$.
- At $t = t_0$ the considered p particles either
  − coalesce (Case I) and then we denote by
  $\boldsymbol{\alpha} = (M, MU, MV)$: weights of the newly formed particle of mass M,
  $U = \dfrac{\sum_{i \in \mathcal{P}} m_i u_i}{\sum_{i \in \mathcal{P}} m_i}$ and $V = \dfrac{\sum_{i \in \mathcal{P}} m_i v_i}{\sum_{i \in \mathcal{P}} m_i}$: its x- and y-velocities,
  $(X_0, Y_0) = (X(t_0), Y(t_0))$: its initial position at time $t = t_0$,
  $(X(t), Y(t))$: its location at time $t > t_0$;
  or
  − undergo the velocities correction (Case II) and then we denote by
  $\widetilde{\boldsymbol{\alpha}}_i = (m_i, m_i \widetilde{u}_i, m_i \widetilde{v}_i)^T$: new weights of the original p particles,
  $(x_i(t), y_i(t))$: their locations, which are not instantaneously affected by the velocities correction procedure and thus change continuously.

We also denote by $\left(x^{\mathrm{CM}}(t), y^{\mathrm{CM}}(t)\right)$ the location of the center of mass of the considered group of p particles,

$$x^{\mathrm{CM}}(t) = \frac{\sum\limits_{i \in \mathcal{P}} m_i x_i(t)}{\sum\limits_{i \in \mathcal{P}} m_i}, \quad y^{\mathrm{CM}}(t) = \frac{\sum\limits_{i \in \mathcal{P}} m_i y_i(t)}{\sum\limits_{i \in \mathcal{P}} m_i}. \tag{2.17}$$

Let us call by *the event with respect to Case* I the situation when some number of particles coalesce at some time moment and at some location. Suppose E_{C1} is the set of such events, and denote by N_{C1} the number of such events that take place in the computational domain within the specified time interval. It is obvious that, in general, N_{C1} is less than the initial number of particles N since each possible merging reduces the number of particles by at least one.

Let us call by *the event with respect to Case* II the situation when the velocities of a particle change according to (2.9) at some time moment. Suppose E_{C2} is the set of such events, and denote by N_{C2} the number of such events that take place in the computational domain within the specified time interval. Notice that all existing particles, whose total number is always less than or equal to N, can undergo the velocity correction procedure at every time step. The minimal distance between the particles is controlled by the particle unification procedure and is thus proportional to $1/\sqrt{N}$. Due to the CFL condition, the size of each time step is proportional to the

minimal distance between the particles. Therefore, the total number of time steps in our 2-D sticky particle method is proportional to $\sqrt{N}$, and hence $N_{\mathrm{C2}} \lesssim N^{3/2}$.

We are now ready to formulate the following theorem.

THEOREM 2.1. *Let* $\mathbf{R}$ *be the residual of the particle solution* $\mathbf{w}^N$*, that is, let* $\mathbf{w}^N$ *satisfy the equation*

$$\mathbf{w}_t + (u\mathbf{w})_x + (v\mathbf{w})_y = \mathbf{R}(x,y,z)$$

in the weak sense defined in [21, *Definition* 1] *for any time interval* $[t_1,t_2] \subset [0,T]$.

If the slopes $(\mathbf{s}_x)_{j,k}$ *and* $(\mathbf{s}_y)_{j,k}$ *in the piecewise linear reconstruction* (2.9) *are set to be* 0 *in all cells* $I_{j,k}$*, then the size of the residual can be estimated by*

$$|\mathbf{R}| \le C\varepsilon \sum_{E_{\mathrm{C1}}\cup E_{\mathrm{C2}}} \left(\frac{\sum\limits_{i<l} m_i m_l \left(|u_i-u_l|+|v_i-v_l|\right)}{\sum\limits_{l} m_l} + \varepsilon \sum_i m_i \left(1+|u_i|+|v_i|\right) \right), \tag{2.18}$$

where the summation is taken over the particles that participate in the specific event from E_{C1} *or* E_{C2}*, and* $\varepsilon := \sqrt{(\Delta x)^2+(\Delta y)^2}$ *is the diameter of the auxiliary grid cell, which is assumed to tend to* 0 *as* $N \to \infty$.

In the case where the slopes $(\mathbf{s}_x)_{j,k}$ *and* $(\mathbf{s}_y)_{j,k}$ *in* (2.9) *are defined according to formulae* (2.10)–(2.12)*, the estimate* (2.18) *is also true, provided the following bound*

$$\left|(s_x^r)_{j,k}\left(x_i - x_{j,k}^{\mathrm{CM}}\right) + (s_y^r)_{j,k}\left(y_i - y_{j,k}^{\mathrm{CM}}\right)\right| \le C\varepsilon \bar{w}_{j,k}^r, \quad r=1,2,3, \tag{2.19}$$

is true at each auxiliary $I_{j,k}$ *cell and for each particle such that* $(x_i,y_i) \in I_{j,k}$*. Here,* $x_{j,k}^{\mathrm{CM}}$ *and* $y_{j,k}^{\mathrm{CM}}$*, given by* (2.11)*, are the coordinates of the center of mass of the particles, located in* $I_{j,k}$ *at the time moment when the velocity correction procedure is performed.*

Remark. The conditions (2.19) are rather technical. It is clear that for the reconstruction (2.10)–(2.14) they hold in smooth parts of the solution (away from vacuum), where all the slopes are bounded. In the nonsmooth parts of the solution and near vacuum, the conditions (2.19) represent a certain nonoscillatory requirement, which may or may not be satisfied by the reconstruction (2.10)–(2.14).

Proof. We start by observing that there is a finite number (which may be proportional to N) of time moments in the interval $[0,T]$ at which some particle velocities change according to either Case I or Case II. Therefore, it is enough to consider such time intervals $[t_1,t_2]$ that contain only a single moment $t=t_0$ of the velocities change.

Let us next fix a test function, $\boldsymbol{\psi} \equiv (\psi^1,\psi^2,\psi^3)^T \in C_0^1(\mathbb{R}^2)$ and consider the following two sets of time moments:

$$\mathcal{T}_1 := \{t_{1_i} \in [t_1,t_0], i=1,\dots,q_1\} \quad \text{and} \quad \mathcal{T}_2 := \{t_{2_i} \in [t_0,t_2], i=1,\dots,q_2\},$$

such that some particle either enters or leaves the domain

$$\Phi := \operatorname{supp}\psi^1 \cup \operatorname{supp}\psi^2 \cup \operatorname{supp}\psi^3 \tag{2.20}$$

at these times.

Notice that it suffices to consider the sets $\mathcal{T}_1$ and $\mathcal{T}_2$ to be finite. If not, then the supports of functions ψ^i, $i=1,2,3$, can be placed into larger convex sets Λ_i and the functions ψ^i can be extended to Λ_i by zero. As has been mentioned above, there is only a finite number of time moments in the interval $[0,T]$ at which some particle velocities change according to either Case I or Case II. Between these time moments

all the particles freely move along straight lines and, due to the convexity of Λ_i, they can intersect the boundaries of Λ_i at most twice. Therefore, replacing $\operatorname{supp}\psi^i$ with Λ_i in (2.20) will make $\mathcal{T}_1$ and $\mathcal{T}_2$ finite.

The conservation laws are thus satisfied in any time interval $[t_{1_i}, t_{1_{i+1}}]$ or $[t_{2_k}, t_{2_{k+1}}]$ since no velocities correction procedures are performed and since the test function $\boldsymbol{\psi}$ vanishes at the points where particles enter or leave the domain Φ. Hence, it is enough to consider only the particles dynamics in the time interval $[\max_i t_{1_i}, \min_k t_{2_k}]$, such that at time $t = \max_i t_{1_i}$ there are p particles (from $\mathcal{P}$) inside the domain Φ and no particles enter or leave Φ until $t = \min_k t_{2_k}$. In order to simplify the notation, we again denote such interval by $[t_1, t_2]$.

Case I. First, we suppose that the particle formed at the time moment $t = t_0$ stays inside the domain Φ. We then plug the particle solution (2.2) into the weak formulation (in the sense of [21, Definition 1]) of (1.1) over the time interval $[t_1, t_2]$ to compute the residuals for the equations of mass and momenta conservation.

- From the *mass conservation* equation we obtain

$$\begin{aligned}
&\int_{t_1}^{t_0} \left\{ \sum_{i\in\mathcal{P}} \left[\psi_x^1(x_i(\tau), y_i(\tau)) m_i u_i + \psi_y^1(x_i(\tau), y_i(\tau)) m_i v_i \right] \right\} d\tau \\
&+ \int_{t_0}^{t_2} \left\{ \psi_x^1(X(\tau), Y(\tau)) MU + \psi_y^1(X(\tau), Y(\tau)) MV \right\} d\tau \\
&= \int_{t_1}^{t_0} \frac{d}{d\tau} \sum_{i\in\mathcal{P}} m_i \psi^1(x_i(\tau), y_i(\tau))\, d\tau + \int_{t_0}^{t_2} \frac{d}{d\tau} M\psi^1(X(\tau), Y(\tau))\, d\tau \\
&= M\psi^1(X(t_2), Y(t_2)) - \sum_{i\in\mathcal{P}} m_i \psi^1(x_i(t_1), y_i(t_1)) + R^1,
\end{aligned}$$

where

$$R^1 = \sum_{i\in\mathcal{P}} m_i \psi^1(x_i^0, y_i^0) - M\psi^1(X_0, Y_0). \tag{2.21}$$

Rewriting (2.21), using the Taylor expansion about (X_0, Y_0) and taking into account (2.17) for $t = t_0$, we arrive at

$$\begin{aligned}
R^1 &= \sum_{i\in\mathcal{P}} m_i \left[\psi^1(x_i^0, y_i^0) - \psi^1(X_0, Y_0) \right] \\
&= \sum_{i\in\mathcal{P}} m_i \left[\psi_x^1(X_0, Y_0) \left(x_i^0 - X_0 \right) + \psi_y^1(X_0, Y_0) \left(y_i^0 - Y_0 \right) + \mathcal{O}(\varepsilon^2) \right] \\
&= \psi_x^1(X_0, Y_0) \left[\sum_{i\in\mathcal{P}} m_i x_i^0 - MX_0 \right] + \psi_y^1(X_0, Y_0) \left[\sum_{i\in\mathcal{P}} m_i y_i^0 - MY_0 \right] + M \cdot \mathcal{O}(\varepsilon^2) \\
&= M \cdot \mathcal{O}(\varepsilon^2). \qquad (2.22)
\end{aligned}$$

- From the *x-momentum conservation* equation we obtain

$$\begin{aligned}
&\int_{t_1}^{t_0} \left\{ \sum_{i\in\mathcal{P}} \left[\psi_x^2(x_i(\tau), y_i(\tau)) u_i \cdot m_i u_i + \psi_y^2(x_i(\tau), y_i(\tau)) v_i \cdot m_i u_i \right] \right\} d\tau \\
&+ \int_{t_0}^{t_2} \left\{ \psi_x^2(X(\tau), Y(\tau)) U \cdot MU + \psi_y^2(X(\tau), Y(\tau)) V \cdot MU \right\} d\tau
\end{aligned}$$

$$= \int_{t_1}^{t_0} \frac{d}{d\tau} \sum_{i\in\mathcal{P}} m_i u_i \psi^2(x_i(\tau), y_i(\tau))\, d\tau + \int_{t_0}^{t_2} \frac{d}{d\tau} MU\psi^2(X(\tau), Y(\tau))\, d\tau$$
$$= MU\psi^2(X(t_2), Y(t_2)) - \sum_{i\in\mathcal{P}} m_i u_i \psi^2(x_i(t_1), y_i(t_1)) + R^2,$$

where

$$R^2 = \sum_{i\in\mathcal{P}} m_i u_i \psi^2(x_i^0, y_i^0) - MU\psi^2(X_0, Y_0). \tag{2.23}$$

We now rewrite (2.23) and use the Taylor expansion about (X_0, Y_0) and (2.17) to obtain

$$R^2 = \sum_{i\in\mathcal{P}} m_i u_i \left[\psi^2(x_i^0, y_i^0) - \psi^2(X_0, Y_0)\right]$$
$$= \sum_{i\in\mathcal{P}} m_i u_i \left[\psi_x^2(X_0, Y_0)\left(x_i^0 - X_0\right) + \psi_y^2(X_0, Y_0)\left(y_i^0 - Y_0\right) + \mathcal{O}(\varepsilon^2)\right]$$
$$= \psi_x^2(X_0, Y_0)\left[\sum_{i\in\mathcal{P}} m_i u_i x_i^0 - MUX_0\right] + \psi_y^2(X_0, Y_0)\left[\sum_{i\in\mathcal{P}} m_i u_i y_i^0 - MUY_0\right] + MU\cdot\mathcal{O}(\varepsilon^2)$$
$$= \psi_x^2(X_0, Y_0)\left[\sum_{i\in\mathcal{P}} m_i u_i x_i^0 - \frac{\sum\limits_{i\in\mathcal{P}} m_i u_i \cdot \sum\limits_{l\in\mathcal{P}} m_l x_l^0}{\sum\limits_{l\in\mathcal{P}} m_l}\right]$$
$$+ \psi_y^2(X_0, Y_0)\left[\sum_{i\in\mathcal{P}} m_i u_i y_i^0 - \frac{\sum\limits_{i\in\mathcal{P}} m_i u_i \cdot \sum\limits_{l\in\mathcal{P}} m_l y_l^0}{\sum\limits_{l\in\mathcal{P}} m_l}\right] + MU\cdot\mathcal{O}(\varepsilon^2).$$

Then, taking into account that

$$\sum_{i\in\mathcal{P}} m_i u_i x_i^0 - \frac{\sum\limits_{i\in\mathcal{P}} m_i u_i \cdot \sum\limits_{l\in\mathcal{P}} m_l x_l^0}{\sum\limits_{l\in\mathcal{P}} m_l} = \frac{\sum\limits_{i,l\in\mathcal{P}} \left(m_l m_i u_i x_i^0 - m_i m_l u_i x_l^0\right)}{\sum\limits_{l\in\mathcal{P}} m_l}$$
$$= \frac{\sum\limits_{i,l\in\mathcal{P}} m_i m_l u_i \left(x_i^0 - x_l^0\right)}{\sum\limits_{l\in\mathcal{P}} m_l} = \frac{\sum\limits_{i,l\in\mathcal{P}:i<l} \left[m_i m_l u_i \left(x_i^0 - x_l^0\right) + m_l m_i u_l \left(x_l^0 - x_i^0\right)\right]}{\sum\limits_{l\in\mathcal{P}} m_l}$$
$$= \frac{\sum\limits_{i,l\in\mathcal{P}:i<l} m_i m_l \left(x_i^0 - x_l^0\right)(u_i - u_l)}{\sum\limits_{l\in\mathcal{P}} m_l},$$

we end up with

$$R^2 = \psi_x^2(X_0, Y_0)\frac{\sum\limits_{i,l\in\mathcal{P}:i<l} m_i m_l (x_i^0 - x_l^0)(u_i - u_l)}{\sum\limits_{l\in\mathcal{P}} m_l}$$
$$+ \psi_y^2(X_0, Y_0)\frac{\sum\limits_{i,l\in\mathcal{P}:i<l} m_i m_l (y_i^0 - y_l^0)(u_i - u_l)}{\sum\limits_{l\in\mathcal{P}} m_l} + MU\cdot\mathcal{O}(\varepsilon^2). \tag{2.24}$$

• In a similar manner, we consider the third component of the residual,

$$R^3 = \sum_{i\in\mathcal{P}} m_i v_i \psi^3(x_i^0, y_i^0) - MV\psi^3(X_0, Y_0), \tag{2.25}$$

and then from the *y-momentum conservation* equation derive

$$\begin{aligned} R^3 = {} & \psi_x^3(X_0, Y_0) \frac{\sum\limits_{i,l\in\mathcal{P}:i<l} m_i m_l (x_i^0 - x_l^0)(v_i - v_l)}{\sum\limits_{l\in\mathcal{P}} m_l} \\ & + \psi_y^3(X_0, Y_0) \frac{\sum\limits_{i,l\in\mathcal{P}:i<l} m_i m_l (y_i^0 - y_l^0)(v_i - v_l)}{\sum\limits_{l\in\mathcal{P}} m_l} + MV\cdot\mathcal{O}(\varepsilon^2). \end{aligned} \tag{2.26}$$

Finally, combining formulae (2.22), (2.24), and (2.26) and using the fact that the distance between (x_i^0, y_i^0) and (x_l^0, y_l^0) is less than $d_{cr} < \varepsilon$, we immediately conclude with the desired estimate (2.18).

Remark. Recall that formulae (2.22), (2.24), and (2.26) were derived under the assumption that the particle formed at $t = t_0$ stays inside the domain Φ. If not, then $\boldsymbol{\psi}(X_0, Y_0) = \mathbf{0}$ and all the particles from $\mathcal{P}$ lie within the distance $d_{cr} < \varepsilon$ from the boundary of $\boldsymbol{\psi}$. Thus the estimate (2.18) follows (as can be seen from formulae (2.21), (2.23), and (2.25)) since $\boldsymbol{\psi} \in C_0^1(\mathbb{R}^2)$ and $|\boldsymbol{\psi}(x_i^0, y_i^0)| < C\varepsilon^2$ for all $(x_i^0, y_i^0) \in \mathcal{P}$.

Case II. As in Case I, we plug the particle solution (2.2) into the weak formulation (in the sense of [21, Definition 1]) of (1.1) over the time interval $[t_1, t_2]$ to compute the corresponding residuals. However, since the set of particles participating in the velocities correction procedure at time $t = t_0$, described in section 2.1, coincides (in general) with the set of all existing particles (including the ones lying outside the domain Φ), the residuals computation is carried out as follows.

• From the *mass conservation* equation we obtain

$$\begin{aligned} & \int_{t_1}^{t_0} \left\{ \sum_{i\in\mathcal{P}} \left[\psi_x^1(x_i(\tau), y_i(\tau)) m_i u_i + \psi_y^1(x_i(\tau), y_i(\tau)) m_i v_i \right] \right\} d\tau \\ & + \int_{t_0}^{t_2} \left\{ \sum_{i\in\mathcal{P}} \left[\psi_x^1(x_i(\tau), y_i(\tau)) m_i \widetilde{u}_i + \psi_y^1(x_i(\tau), y_i(\tau)) m_i \widetilde{v}_i \right] \right\} d\tau \\ & = \int_{t_1}^{t_0} \frac{d}{d\tau} \sum_{i\in\mathcal{P}} m_i \psi^1(x_i(\tau), y_i(\tau))\, d\tau + \int_{t_0}^{t_2} \frac{d}{d\tau} \sum_{i\in\mathcal{P}} m_i \psi^1(x_i(\tau), y_i(\tau))\, d\tau \\ & = \sum_{i\in\mathcal{P}} m_i \psi^1(x_i(t_2), y_i(t_2)) - \sum_{i\in\mathcal{P}} m_i \psi^1(x_i(t_1), y_i(t_1)). \end{aligned}$$

Unlike Case I, the particle trajectories are now continuous within the time interval $[t_1, t_2]$ because only particle velocities may change at $t = t_0$. Therefore, the first component of the residual is

$$R^1 = \sum_{i\in\mathcal{P}} m_i \psi^1(x_i^0, y_i^0) - \sum_{i\in\mathcal{P}} m_i \psi^1(x_i^0, y_i^0) = 0. \tag{2.27}$$

• From the *x-momentum conservation* equation we obtain

$$\int_{t_1}^{t_0} \left\{ \sum_{i\in\mathcal{P}} \left[\psi_x^2(x_i(\tau), y_i(\tau)) u_i \cdot m_i u_i + \psi_y^2(x_i(\tau), y_i(\tau)) v_i \cdot m_i u_i \right] \right\} d\tau$$

$$
\begin{aligned}
&+\int_{t_0}^{t_2}\left\{\sum_{i\in\mathcal{P}}\left[\psi_x^2(x_i(\tau),y_i(\tau))\widetilde{u}_i\cdot m_i\widetilde{u}_i+\psi_y^2(x_i(\tau),y_i(\tau))\widetilde{v}_i\cdot m_i\widetilde{u}_i\right]\right\}d\tau\\
&=\int_{t_1}^{t_0}\frac{d}{d\tau}\sum_{i\in\mathcal{P}}m_iu_i\psi^2(x_i(\tau),y_i(\tau))\,d\tau+\int_{t_0}^{t_2}\frac{d}{d\tau}\sum_{i\in\mathcal{P}}m_i\widetilde{u}_i\psi^2(x_i(\tau),y_i(\tau))\,d\tau\\
&=\sum_{i\in\mathcal{P}}m_i\widetilde{u}_i\psi^2(x_i(t_2),y_i(t_2))-\sum_{i\in\mathcal{P}}m_iu_i\psi^2(x_i(t_1),y_i(t_1))+R^2,
\end{aligned}
$$

where

$$
R^2=\sum_{i\in\mathcal{P}}\psi^2(x_i^0,y_i^0)(m_iu_i-m_i\widetilde{u}_i). \tag{2.28}
$$

Note that the summation in (2.28) is taken over the particles located in the domain Φ, which consists of a certain number of auxiliary cells (or their parts) $I_{j,k}$. Thus, the residual R^2 can be written as the sum of residuals in each cell $I_{j,k}$ that contains (at least) one particle and has a nonempty intersection with Φ. Let us now consider such a cell, denote the set of particles, located in it at time moment $t=t_0$, by $\mathcal{P}_{j,k}$, and the residual in this cell by $R^2_{j,k}$.

Applying the Taylor expansion about the center of mass $(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k})$ given by (2.11) yields

$$
\begin{aligned}
R^2_{j,k}=&\sum_{i\in\mathcal{P}_{j,k}}\Big[\psi^2\left(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k}\right)+\psi^2_x\left(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k}\right)\left(x_i^0-x^{\rm CM}_{j,k}\right)\\
&\qquad+\psi^2_y\left(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k}\right)\left(y_i^0-y^{\rm CM}_{j,k}\right)+\mathcal{O}(\varepsilon^2)\Big]m_i(u_i-\widetilde{u}_i)\\
=&\left[\psi^2\left(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k}\right)+\mathcal{O}(\varepsilon^2)\right]\sum_{i\in\mathcal{P}_{j,k}}m_i(u_i-\widetilde{u}_i)\\
&+\psi^2_x\left(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k}\right)\sum_{i\in\mathcal{P}_{j,k}}m_i(u_i-\widetilde{u}_i)\left(x_i^0-x^{\rm CM}_{j,k}\right)\\
&+\psi^2_y\left(x^{\rm CM}_{j,k},y^{\rm CM}_{j,k}\right)\sum_{i\in\mathcal{P}_{j,k}}m_i(u_i-\widetilde{u}_i)\left(y_i^0-y^{\rm CM}_{j,k}\right).
\end{aligned}
\tag{2.29}
$$

Next, we consider each sum on the right-hand side (RHS) of (2.29) separately. For the first sum, use formulae (2.8)–(2.10) to rewrite

$$
\begin{aligned}
&\sum_{i\in\mathcal{P}_{j,k}}m_i(u_i-\widetilde{u}_i)=\sum_{i\in\mathcal{P}_{j,k}}m_i\left[u_i-\frac{\bar{w}^2_{j,k}+\left(s^2_x\right)_{j,k}\left(x_i^0-x^{\rm CM}_{j,k}\right)+\left(s^2_y\right)_{j,k}\left(y_i^0-y^{\rm CM}_{j,k}\right)}{\bar{w}^1_{j,k}+\left(s^1_x\right)_{j,k}\left(x_i^0-x^{\rm CM}_{j,k}\right)+\left(s^1_y\right)_{j,k}\left(y_i^0-y^{\rm CM}_{j,k}\right)}\right]\\
&=\sum_{i\in\mathcal{P}_{j,k}}m_i\left\{u_i-\frac{1}{\bar{w}^1_{j,k}}\left[\bar{w}^2_{j,k}+\left(s^2_x\right)_{j,k}\left(x_i^0-x^{\rm CM}_{j,k}\right)+\left(s^2_y\right)_{j,k}\left(y_i^0-y^{\rm CM}_{j,k}\right)\right]\right.\\
&\left.\times\left[1+\frac{\left(s^1_x\right)_{j,k}\left(x_i^0-x^{\rm CM}_{j,k}\right)+\left(s^1_y\right)_{j,k}\left(y_i^0-y^{\rm CM}_{j,k}\right)}{\bar{w}^1_{j,k}}\right]^{-1}\right\}.
\end{aligned}
\tag{2.30}
$$

Taking into account (2.19), the last term in (2.30) is equal to

$$
1-\frac{\left(s^1_x\right)_{j,k}\left(x_i^0-x^{\rm CM}_{j,k}\right)+\left(s^1_y\right)_{j,k}\left(y_i^0-y^{\rm CM}_{j,k}\right)}{\bar{w}^1_{j,k}}+\mathcal{O}(\varepsilon^2),
$$

and thus

$$\sum_{i\in\mathcal{P}_{j,k}} m_i(u_i-\widetilde{u}_i)=\frac{1}{\bar{w}^1_{j,k}}\Bigg\{\bar{w}^1_{j,k}\sum_{i\in\mathcal{P}_{j,k}} m_iu_i$$
$$-\sum_{i\in\mathcal{P}_{j,k}} m_i\Bigg[\bar{w}^2_{j,k}-\frac{\bar{w}^2_{j,k}}{\bar{w}^1_{j,k}}\left(\left(s^1_x\right)_{j,k}\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)+\left(s^1_y\right)_{j,k}\left(y^0_i-y^{\mathrm{CM}}_{j,k}\right)\right)$$
$$+\left(s^2_x\right)_{j,k}\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)+\left(s^2_y\right)_{j,k}\left(y^0_i-y^{\mathrm{CM}}_{j,k}\right)+\bar{w}^2_{j,k}\mathcal{O}(\varepsilon^2)\Bigg]\Bigg\}.$$

Finally, using the definition of cell averages (2.8) and the fact that the center of mass $(x^{\mathrm{CM}}_{j,k},y^{\mathrm{CM}}_{j,k})$ satisfies (2.11), we obtain

$$\sum_{i\in\mathcal{P}_{j,k}} m_i(u_i-\widetilde{u}_i)=\frac{1}{\bar{w}^1_{j,k}}\Bigg\{\sum_{i\in\mathcal{P}_{j,k}} m_i\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)\left[\frac{\bar{w}^2_{j,k}}{\bar{w}^1_{j,k}}\left(s^1_x\right)_{j,k}-\left(s^2_x\right)_{j,k}\right]$$
$$+\sum_{i\in\mathcal{P}_{j,k}} m_i\left(y^0_i-y^{\mathrm{CM}}_{j,k}\right)\left[\frac{\bar{w}^2_{j,k}}{\bar{w}^1_{j,k}}\left(s^1_y\right)_{j,k}-\left(s^2_y\right)_{j,k}\right]+\bar{w}^2_{j,k}\mathcal{O}(\varepsilon^2)\sum_{i\in\mathcal{P}_{j,k}} m_i\Bigg\}$$
$$=\Delta x\Delta y\bar{w}^2_{j,k}\mathcal{O}(\varepsilon^2)=\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\cdot\mathcal{O}(\varepsilon^2). \tag{2.31}$$

Remark. Note that if $\left(\mathbf{s}_x\right)_{j,k}=\left(\mathbf{s}_y\right)_{j,k}=\mathbf{0}$, then the RHS of (2.31) vanishes and $\sum_{i\in\mathcal{P}_{j,k}} m_i(u_i-\widetilde{u}_i)=0$. Otherwise, one has an approximate x-momentum conservation only.

We next consider the second sum on the RHS of (2.29) and in a similar manner obtain

$$\sum_{i\in\mathcal{P}_{j,k}} m_i(u_i-\widetilde{u}_i)\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)=\frac{1}{\bar{w}^1_{j,k}}\Bigg\{\bar{w}^1_{j,k}\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)$$
$$-\sum_{i\in\mathcal{P}_{j,k}} m_i\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)\Bigg(\bar{w}^2_{j,k}+\left[\left(s^2_x\right)_{j,k}-\frac{\bar{w}^2_{j,k}}{\bar{w}^1_{j,k}}\left(s^1_x\right)_{j,k}\right]\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)$$
$$+\left[\left(s^2_y\right)_{j,k}-\frac{\bar{w}^2_{j,k}}{\bar{w}^1_{j,k}}\left(s^1_y\right)_{j,k}\right]\left(y^0_i-y^{\mathrm{CM}}_{j,k}\right)+\bar{w}^2_{j,k}\mathcal{O}(\varepsilon^2)\Bigg)\Bigg\}$$
$$=\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)+\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\cdot\mathcal{O}(\varepsilon^2). \tag{2.32}$$

Then, using the definition of the center of mass (2.11) we rewrite the first term on the RHS of (2.32) as

$$\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\left(x^0_i-x^{\mathrm{CM}}_{j,k}\right)=\sum_{i\in\mathcal{P}_{j,k}} m_iu_ix^0_i-\frac{\sum\limits_{i\in\mathcal{P}_{j,k}} m_iu_i\sum\limits_{l\in\mathcal{P}_{j,k}} m_lx^0_l}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l}$$
$$=\frac{\sum\limits_{i,l\in\mathcal{P}_{j,k}}\left(m_lm_iu_ix^0_i-m_im_lu_ix^0_l\right)}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l}=\frac{\sum\limits_{i,l\in\mathcal{P}_{j,k}:i<l} m_im_l\left(x^0_i-x^0_l\right)(u_i-u_l)}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l},$$

and conclude with

(2.33)
$$\sum_{i\in\mathcal{P}_{j,k}} m_i(u_i-\widetilde{u}_i)\left(x_i^0-x_{j,k}^{\rm CM}\right)=\frac{\sum\limits_{i,l\in\mathcal{P}_{j,k}:i<l} m_im_l\left(x_i^0-x_l^0\right)(u_i-u_l)}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l}+\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\cdot\mathcal{O}(\varepsilon^2).$$

The estimate on the third sum on the RHS of (2.29) is completely analogous:

(2.34)
$$\sum_{i\in\mathcal{P}_{j,k}} m_i(u_i-\widetilde{u}_i)\left(y_i^0-y_{j,k}^{\rm CM}\right)=\frac{\sum\limits_{i,l\in\mathcal{P}_{j,k}:i<l} m_im_l\left(y_i^0-y_l^0\right)(u_i-u_l)}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l}+\sum_{i\in\mathcal{P}_{j,k}} m_iu_i\cdot\mathcal{O}(\varepsilon^2).$$

We now substitute (2.31), (2.33), and (2.34) into (2.29) and sum up all $R^2_{j,k}$ to end up with the following estimate:

$$|R^2|\le C\varepsilon\sum_{j,k}\left(\frac{\sum\limits_{i,l\in\mathcal{P}_{j,k}:i<l} m_im_l|u_i-u_l|}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l}+\varepsilon\sum_{i\in\mathcal{P}_{j,k}} m_i|u_i|\right). \tag{2.35}$$

• A similar estimate on R^3 is obtained from the *y-momentum conservation* equation,

$$|R^3|\le C\varepsilon\sum_{j,k}\left(\frac{\sum\limits_{i,l\in\mathcal{P}_{j,k}:i<l} m_im_l|v_i-v_l|}{\sum\limits_{l\in\mathcal{P}_{j,k}} m_l}+\varepsilon\sum_{i\in\mathcal{P}_{j,k}} m_i|v_i|\right). \tag{2.36}$$

Finally, adding up (2.27), (2.35), and (2.36), and considering velocity corrections occurring in different auxiliary cells to be different events from the set $E_{\rm C2}$, we obtain the estimate (2.18) in Case II.

This completes the proof of the theorem since in our 2-D sticky particle method the only contributions to the residual $\mathbf{R}$ come from the particle interactions enforced by the merger (Case I) and velocity correction (Case II) procedures. □

Remark. Note that as has been shown in the proof (see the estimate (2.31) and the remark after it), the use of the second-order reconstruction (2.10) results in additional errors in momenta conservation equations compared with the first-order ($(\mathbf{s}_x)_{j,k}=(\mathbf{s}_y)_{j,k}=0$ for all j,k) approach. However, a more accurate velocity reconstruction typically leads to a more accurate particle dynamics, while the momenta conservation errors and their contributions to the corresponding residuals (the second terms on the RHS of (2.35) and (2.36)) are relatively small.

We conclude this section with a brief discussion of the result established in Theorem 2.1, which provides us with an estimate on the size of the residual. We view this result as a step toward the convergence proof of the proposed 2-D sticky particle method. Completing the proof would require obtaining more precise estimates on the residual, which, in general, may be rather difficult. However, according to the conjecture in [21], the following scenario of mass concentration occurs. Let us first mention that the system (1.5) has straight bicharacteristic lines, which usually intersect at some time moment (analogously to the 1-D case) and form curves in the (x,y)-plane (their representation in the (t,x,y)-space is not a characteristic surface, but a surface defined by a generalization of the Hugoniot relations) with a finite mass distributed

along the curves as a δ-function. These curves then start to impinge each other and form the singularities with finite masses at separate points. The collisions of the curves take place, in general, transversally, but no rigorous theoretical description of such a solution behavior is available. Assuming now that solutions of (1.5) have such a structure (this assumption has also been supported by the numerical experiments reported in section 3.2, Examples 7, 8a, and 8b), it is possible to show that $|\mathbf{R}| \to 0$ as $N \to \infty$.

Indeed, following the above scenario when particles coalesce (Case I) they form curves with finite masses. In this case, the differences $|u_i - u_l|$ and $|v_i - v_l|$ are finite, the considered cluster of particles $\mathcal{P}$ merges into a particle with mass $\sim \varepsilon$, while other, nonclustered, particles have masses $\sim \varepsilon^2$. Also, $m_l \sim \varepsilon^2$ and $|\mathcal{P}| \sim 1/\varepsilon$. Thus, $\sum_{l\in\mathcal{P}} m_l \sim \varepsilon$, $\sum_{i,l\in\mathcal{P}:i<l} m_i m_l \sim \varepsilon^3$, and hence one gets $|\mathbf{R}| \sim \varepsilon^3 \cdot N_{\mathrm{C1}}$. Finally, $N_{\mathrm{C1}} < C/\varepsilon^2$ since it is bounded by the total number of particles N, and thus we obtain that $|\mathbf{R}| \sim \varepsilon$, which tends to zero as N tends to infinity.

We now consider the situation of "pure" Case 2, when only the velocities correction procedure is performed and no particles coalesce. In this case, taking into account that the corrected velocities are also close and masses of particles are of order ε^2, one has $|\mathbf{R}| \sim \varepsilon^4 \cdot N_{\mathrm{C2}}$. But, as has been mentioned before, $N_{\mathrm{C2}} \sim N^{3/2} \sim 1/\varepsilon^3$, and thus we obtain that $|\mathbf{R}| \sim \varepsilon$ in Case II as well.

We hope that the presented heuristic convergence arguments can be "upgraded" to a rigorous convergence proof and we plan to do so in forthcoming papers.

3. Numerical examples. In this section, we test the new sticky particle (SP) method presented in section 2 on a number of 1-D and 2-D numerical examples. We also compare solutions computed by the particle method with the corresponding solutions computed using the second-order semidiscrete central-upwind (CU) scheme, developed in [11, 12, 14]. A brief description of the CU scheme for the pressureless gas dynamics system (1.5) is provided in Appendix A. Numerical time integration has been performed using the strong stability preserving Runge–Kutta method [10].

Note that in all the examples below, we do not reconstruct point values of the computed density from its particle distribution at the final time but rather plot the total mass m of each particle. For the purpose of fair comparison, the solutions computed by the finite-volume CU scheme are always presented in a similar way, that is, we plot the total mass in each cell rather than the corresponding cell averages.

3.1. One-dimensional examples. The following four examples are devoted to the 1-D system (1.2). A 1-D version of our SP method can be easily deduced from its 2-D formulation in section 2.

Example 1. In the first numerical test, taken from [5], we solve the system (1.2) subject to the following Riemann initial data:

$$(\rho(x,0), u(x,0)) = \begin{cases} (1.00, \quad 0.5) & \text{if } \ x < 0, \\ (0.25, -0.4) & \text{if } \ x > 0. \end{cases} \tag{3.1}$$

In this example, a δ-shock develops immediately and propagates with speed 0.2.

We take $\Delta x = 0.005$ for the CU scheme and the initial uniform distribution of particles, placed Δx away from each other, for the SP method. In Figure 1, the particle/cell masses and the corresponding velocities, computed by both the SP method and the CU scheme, are plotted at time $t = 0.5$. Note that because of the point mass concentration occurring at the δ-shock, the masses are presented in the logarithmic scale so that a more detailed structure of the solution can be seen.

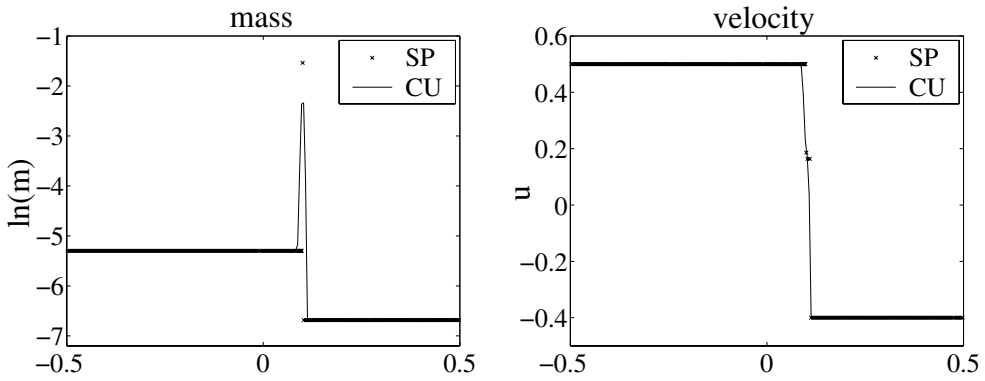


FIG. 1. *Solution of* (1.2), (3.1) *computed by the SP method and the CU scheme.*

Figure 1 demonstrates that both schemes are able to capture the δ-shock with the correct propagation speed, but one can clearly see the superiority of the results obtained by the SP method, which does not smear the δ-shock.

Example 2. We consider a test problem of the collision of two compactly supported clouds. The initial data, prescribed at $t = -1$, are taken from [17],

$$(\rho(x,-1),u(x,-1)) = \begin{cases} (2,\ \ 1) & \text{if} \quad -2 < x < -1, \\ (1,-1) & \text{if} \quad\ \ 1 < x < 5, \\ (0,\ \ 0) & \text{otherwise.} \end{cases} \tag{3.2}$$

The two clouds collide at time $t = 0$. The left cloud is fully accelerated into the δ-wave at about $t \approx 1.21$ and the right cloud is fully accelerated at about $t \approx 4.25$. We use a uniform spatial grid with $\Delta x = 0.0125$ for the CU scheme. The SP method is started with 400 particles, placed only in the intervals $[-2,-1]$ and $[1,5]$, where the dust is initially present. Figures 2 and 3 show the particle/cell masses (in the logarithmic scale) at times $t = -1$, -0.5, 0, 0.5, 1, 1.5, 3.5, and 6. As one can observe, both methods give the same correct location of the δ-wave. However, both the δ-wave and the contact discontinuities computed by the CU scheme are smeared over a number of cells, while the resolution achieved by the SP method is almost perfect. We note that the mass computed by the SP method is concentrated in a single point by time $t = 6$.

Example 3. In this example, we demonstrate an interaction of two singular shocks by numerically solving the system (1.2) subject to the following initial data:

$$(\rho(x,0),u(x,0)) = \begin{cases} (0.25,1.00) & \text{if} \ -2.75 < x < -0.75, \\ (0.25,0.50) & \text{if} \ -0.75 < x < 0.5, \\ (1.00,-1.00) & \text{if} \quad\ \ 0.5 < x < 1.5, \\ (0.00,0.00) & \text{otherwise.} \end{cases} \tag{3.3}$$

In Figure 4, we plot the particle/cell masses (in the logarithmic scale) computed by both the SP method and the CU scheme at times $t = 0$, 0.5, 1, 1.5, 2, and 2.5. We start the SP method with $N = 425$ particles, which are uniformly distributed in the interval $[-2.75, 1.5]$. For the CU scheme, we use a uniform spatial grid with $\Delta x = 0.01$. Again, one can clearly see that the SP method outperforms the finite-volume CU scheme by far.

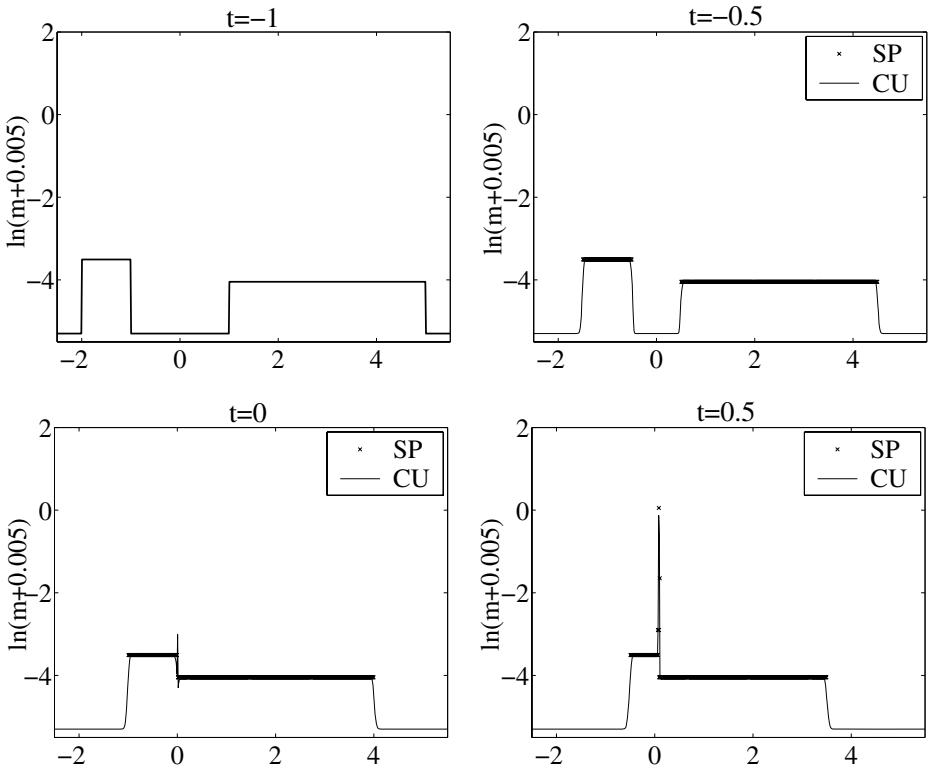


FIG. 2. *Solution (masses) of* (1.2), (3.2) *computed by the SP method and the CU scheme.*

Example 4. The last 1-D example is devoted to a problem where the velocity u changes its sign in the region with varying density. This significantly increases the level of complexity of the problem due to a special way the singularity forms, as demonstrated below.

We consider the system (1.2) subject to the smooth initial data:

(3.4)
$$\rho(x,0) = \begin{cases} 2-\sin x & \text{if } -\pi \le x \le \pi, \\ 0 & \text{otherwise,} \end{cases} \qquad u(x,0) = \begin{cases} 1-x & \text{if } -\pi \le x \le \pi, \\ 0 & \text{otherwise,} \end{cases}$$

for which the exact solution can be found analytically as follows. A continuous part of the solution is obtained by the method of characteristics:

$$u(X(t),t) = 1 - x_0, \qquad \rho(X(t),t) = \frac{2-\sin x_0}{1-t}, \tag{3.5}$$

where

$$X(t) = x_0 + t(1-x_0) \tag{3.6}$$

is the characteristic line starting at $x = x_0$. Obviously, the solution (3.5)–(3.6) is valid in the domain bounded by the characteristics $X_-(t) = -\pi + t(1+\pi)$ and $X_+(t) = \pi + t(1-\pi)$ and thus exists until $t = 1$ only; see Figure 5.

As t approaches 1, the density tends to infinity, more and more mass is concentrated near the point $x = 1$, and therefore one can anticipate a massive particle

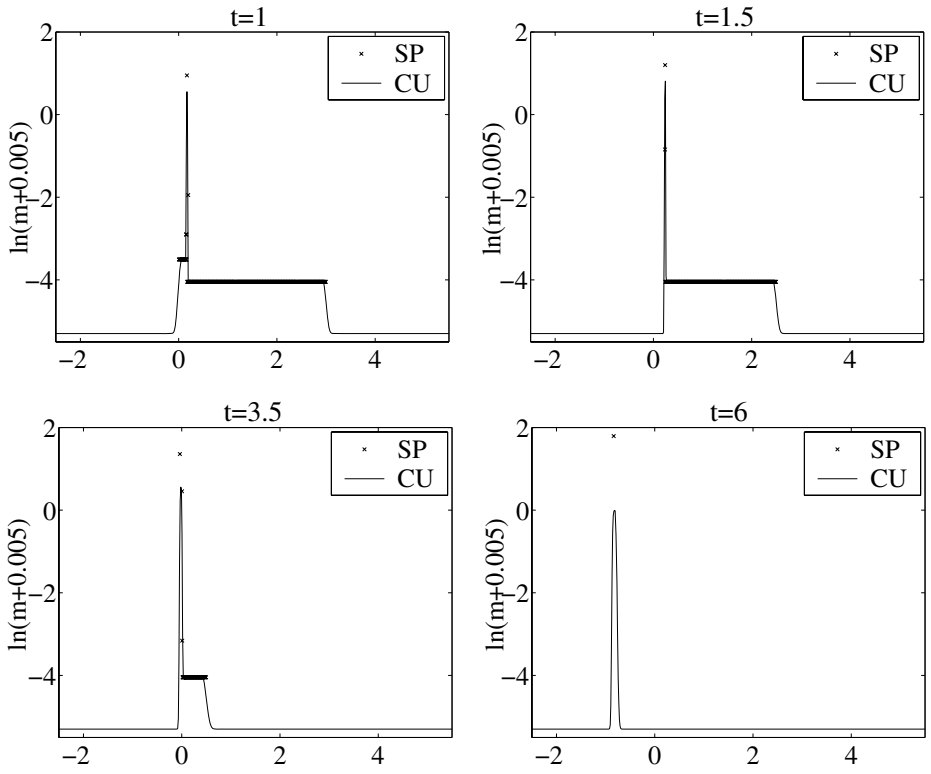


FIG. 3. *Solution (masses) of* (1.2), (3.2) *computed by the SP method and the CU scheme.*

formation at this point. In order to determine a singular part of the solution of (1.2), (3.4) we use the variational representation of the generalized solution of pressureless gas dynamics equations introduced in [8]. To this end, we consider the function

$$F(x_0) \equiv F(x_0;x,t) := \int_0^{x_0} [s - x + t\,u(s,0)]\,\rho(s,0)\,ds = \int_0^{x_0} [s - x + t\,(1-s)]\,(2-\sin s)\,ds, \tag{3.7}$$

and according to [8], the smoothness of the solution depends on a number of points at which the global minimum of F is attained. If F has only one global minimum point, then the solution is continuous at (x,t); otherwise the solution develops a shock discontinuity in velocity and a δ-shock in density there. In the latter case, suppose that there exists a set of points $\{x_0^1, x_0^2, \ldots\}$ at which F assumes its global minimum, and denote $x_0^- := \min\{x_0^1, x_0^2, \ldots\}$ and $x_0^+ := \max\{x_0^1, x_0^2, \ldots\}$. Then the left and right values of ρ and u at (x,t) are computed from (3.5) with $x_0 = x_0^-$ and $x_0 = x_0^+$, respectively. In addition, the δ-function singularity at this point ("a massive particle") has the following mass and momentum:

$$M = \int_{x_0^-}^{x_0^+} \rho(s,0)\,ds, \qquad I = \int_{x_0^-}^{x_0^+} u(s,0)\rho(s,0)\,ds, \tag{3.8}$$

and according to mass and momentum conservation, the speed of the massive particle

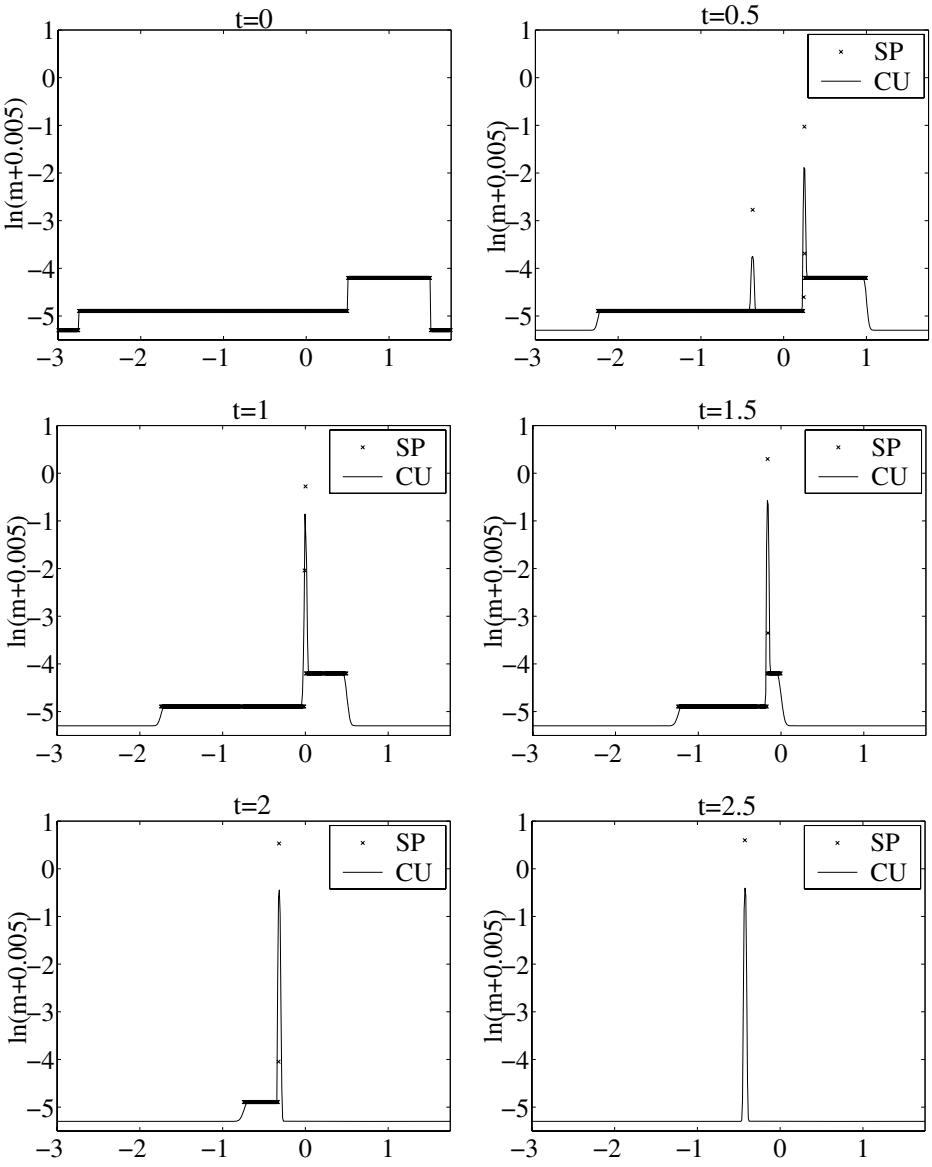


FIG. 4. *Solution (masses) of* (1.2), (3.3) *computed by the SP method and the CU scheme.*

is

$$\frac{dX}{dt} = \frac{I}{M}. \tag{3.9}$$

For the problem under consideration, the singularity is first formed at the point $(x,t) = (1,1)$, and for $t \geq 1$, the global minimum of F is attained at two points only: $x_0^- = -\pi$ and $x_0^+ = \pi$. Therefore, by $t = 1$ all the mass is concentrated in one massive particle with the mass $M = 4\pi$ and the momentum $I = 6\pi$ (according to (3.8)), and the movement of this particle is described, according to (3.9), by the formula $X(t) = (3t-1)/2$, $t \geq 1$; see Figure 5.

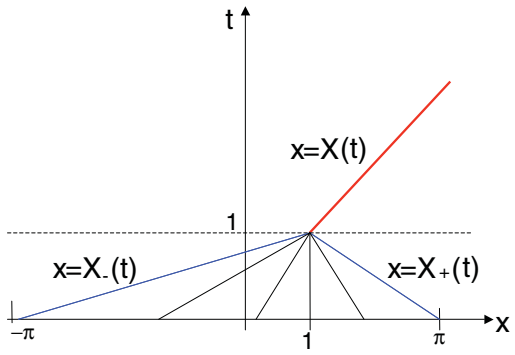


FIG. 5. *Characteristics diagram for the initial-value problem* (1.2), (3.4).

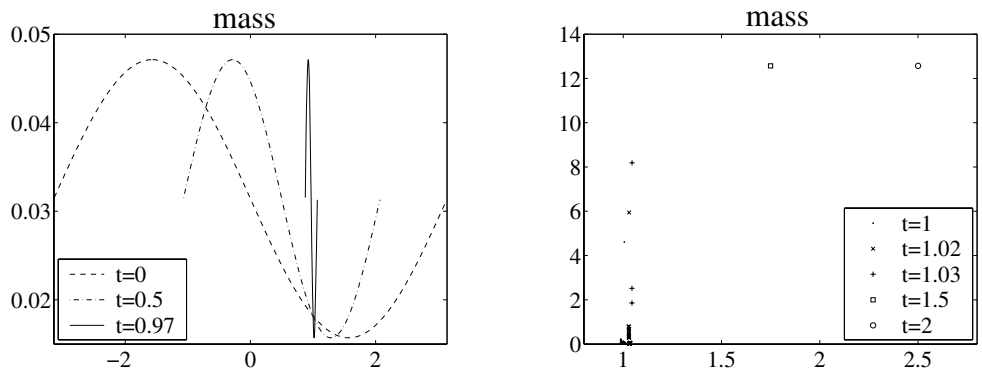


FIG. 6. *Solution (masses) of* (1.2), (3.4) *computed by the SP method for* $t < 1$ *(left), when the solution is smooth inside its shrinking support, and for* $t \geq 1$ *(right), when the total mass* $M = 4\pi$ *is concentrated in one particle, propagating with the constant speed* $I/M = 3/2$. *Note that due to a certain arbitrariness in the selection of the unification parameter* $d_{\rm cr}$, *the final collapse of the numerical solution occurs at a slightly later time* $t \approx 1.03$.

We now turn to the presentation of our numerical results. We start the SP simulations with 400 particles uniformly distributed over the interval $[-\pi, \pi]$. In Figure 6, we plot the particle masses computed by the SP method only, since the CU scheme could not be applied to this problem at large times ($t \sim 1$ and larger). We note that, as indicated in [17], other finite-volume methods are likely to fail to capture the solution of the initial-value problem (1.2), (3.4) as well.

3.2. Two-dimensional examples.

Example 5. We start by numerically solving the 2-D analogue of the 1-D problem considered in Example 1, namely, we solve the system (1.5) in the square domain $[-1, 1] \times [-1, 1]$ subject to the 1-D Riemann initial data, artificially extended to two space dimensions:

$$(\rho(x, y, 0), u(x, y, 0), v(x, y, 0)) = \begin{cases} (1.00, \ \ 0.5, \ 0) & \text{if } \ x < 0, \\ (0.25, -0.4, \ 0) & \text{if } \ x > 0. \end{cases} \tag{3.10}$$

The purpose of this simple example is to demonstrate the failure of the "standard" velocity recovery procedure (2.7) and the ability of the alternative procedure (2.9), developed in section 2.1, to force the desired interaction of nearby particles.

Recall that in this example a δ-shock develops immediately at the line $x = 0$ and then propagates to the right with speed 0.2. As has been already mentioned, the probability of collision of two particles approaching the same singularity curve in two dimensions is, in general, zero, and therefore using formula (2.7) for computing velocities requires a special symmetric setting of the initial locations of particles; see Figure 7 (left). Obviously, if at time $t = 0$ the particles are placed as shown in Figure 7 (right) and if the unification parameter is reasonably small ($d_{\rm cr} < \Delta y/2$), the particles moving from the left and from the right will never interact and the δ-shock will not be captured numerically. We note that for a more complicated, truly 2-D initial data it may be impossible to impose any kind of symmetry, so the situation with the data as in Figure 7 (right) is generic.

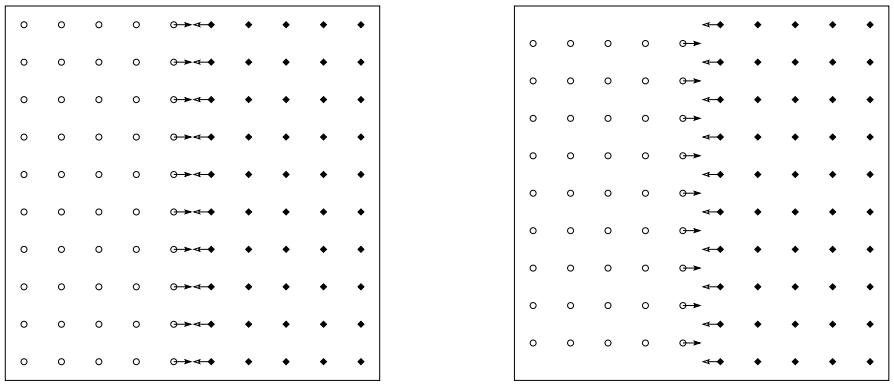


FIG. 7. *Initial locations of particles in Example* 5*: symmetric (left) and asymmetric (right) cases.*

On the other hand, the velocity recovery procedure (2.9) ensures an interaction between particles independently of their initial placement. In Figures 8 and 9, we show the masses and x-velocities of the particles at time $t = 0.5$. They are computed by the SP method with the initial locations of particles as in Figure 7 (left) and Figure 7 (right), respectively. As one can see, in both cases the SP method combined with the velocity recovery procedure (2.9) leads to the desired clustering of particles at the singularity. Moreover, the resolution achieved in the case of asymmetric initial particle distribution is almost as good as in the symmetric case.

Example 6. Next, we turn to genuinely 2-D problems. First, consider the system (1.5) subject to the following initial data:

$$(\rho(x,y,0), u(x,y,0), v(x,y,0)) = \begin{cases} (2,2,1) & \text{if } (x,y) \in \Omega, \\ (0,0,0) & \text{if } (x,y) \in \partial\Omega, \\ (1,0,0) & \text{otherwise,} \end{cases} \tag{3.11}$$

where $\Omega = \{x < 0,\, y < 1\} \cup \{x > 0,\, y > 0,\, x^2 + y^2 < 1\} \cup \{y < 0,\, 0 < x < 1\}$. The initial location of the discontinuity $\partial\Omega$ is shown in Figure 10. According to [21], the exact solution of the initial-value problem (1.5), (3.11) develops a δ-shock in density,

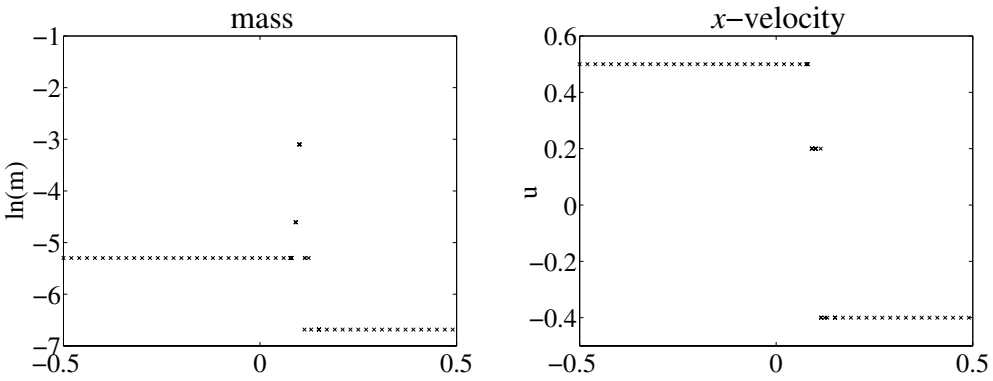


FIG. 8. *Side view on the solution of* (1.5), (3.10) *computed by the SP method. The initial location of particles is shown in Figure* 7 *(left).*

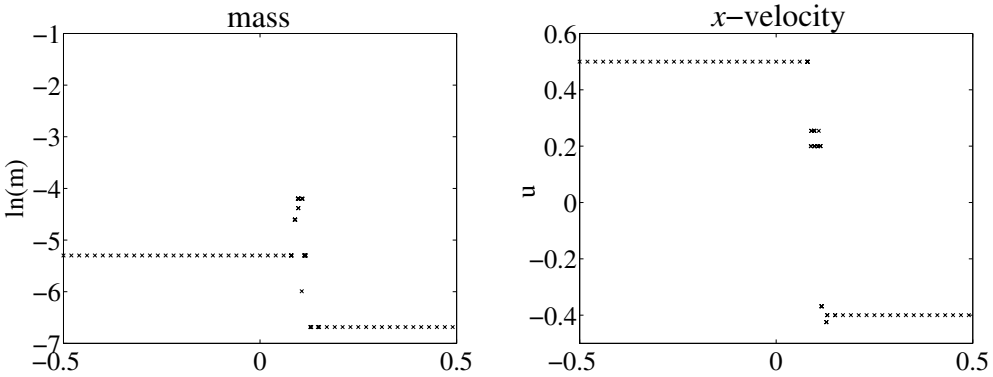


FIG. 9. *Side view on the solution of* (1.5), (3.10) *computed by the SP method. The initial location of particles is shown in Figure* 7 *(right).*

and the evolution of the shock curve is described by the following system of ODEs:

$$\begin{cases} \dfrac{dX}{dt} = u_- \dfrac{\sqrt{\rho_-}}{\sqrt{\rho_-}+\sqrt{\rho_+}} = \dfrac{2\sqrt{2}}{\sqrt{2}+1}, \\ \dfrac{dY}{dt} = v_- \dfrac{\sqrt{\rho_-}}{\sqrt{\rho_-}+\sqrt{\rho_+}} = \dfrac{\sqrt{2}}{\sqrt{2}+1}, \end{cases} \tag{3.12}$$

where $(\rho_-, u_-, v_-) := (2, 2, 1)$ are the initial values inside the domain Ω and $\rho_+ := 1$ is the initial value of the density on the other side of the initial shock curve.

Numerically, we restrict the initial data (3.11) to the finite domain $[-4, 4]\times[-4, 4]$ and consider the following initial-boundary value problem: (1.5), (3.11) together with the solid wall boundary conditions. The numerical solutions, computed by the SP method at time $t = 2$ with 50×50 and 100×100 initially uniformly distributed particles, are plotted in Figure 10. The size of each point in the figure is proportional to the mass accumulated in the particle located there. The exact solution of the initial-boundary value problem is not known, but in the domain $[0, 4] \times [-2, 4]$ it coincides with the solution of the original initial-value problem (1.5), (3.11), and as can be clearly seen in Figure 10, the SP method accurately tracks the evolution of

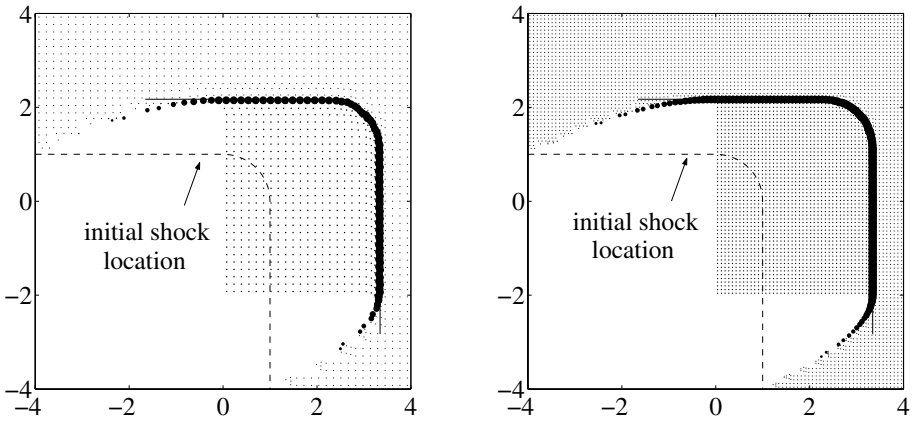


FIG. 10. *Top view on the solution (masses) of* (1.5), (3.11) *computed by the SP method with* 50×50 *(left) and* 100×100 *(right) particles. The solid line is obtained from the initial shock curve (the dashed line) by the evolution according to* (3.12).

the corresponding part of the shock curve described by (3.12). Outside the domain $[0,4] \times [-2,4]$, the solution is obviously affected by the boundedness of the cloud, but the obtained numerical solution looks reasonable, as supported by the performed mesh refinement study.

Example 7. Next, we consider an example with nonzero mass and momenta at the initial shock curve. We numerically solve the system (1.5) subject to the following initial data:

$$(\rho(\mathbf{x},0), u(\mathbf{x},0), v(\mathbf{x},0)) = \begin{cases} (2,2,2) & \text{if } \mathbf{x} \in \Omega, \\ (10\,\delta(\mathrm{dist}(\mathbf{x},\partial\Omega)),2,1) & \text{if } \mathbf{x} \in \partial\Omega, \\ (2,0,0) & \text{otherwise,} \end{cases} \tag{3.13}$$

where $\mathbf{x} \equiv (x,y)$ and the domain Ω is the same as in Example 6: $\Omega = \{x<0,\, y<1\} \cup \{x>0,\, y>0,\, x^2+y^2<1\} \cup \{y<0,\, 0<x<1\}$. In the practical implementation, we replace the δ-function along the curve $\partial\Omega$ with its approximation by a step function; namely, we take

$$\rho(\mathbf{x},0) = \begin{cases} \dfrac{10\sqrt{2}}{\sqrt{(\Delta x)^2+(\Delta y)^2}} & \text{if } \mathrm{dist}(\mathbf{x},\partial\Omega) \le \dfrac{\sqrt{(\Delta x)^2+(\Delta y)^2}}{2\sqrt{2}}, \\ 2 & \text{otherwise.} \end{cases}$$

The numerical solutions at time $t = 1.5$ obtained using the SP method with 101×101 particles (initially uniformly distributed) and the CU scheme with $\Delta x = \Delta y = 0.08$ are plotted in Figures 11 and 12. Note that the maximal mass value of the solution obtained by the CU scheme is 0.6299 while the maximal mass obtained by the SP method is 2.7009. As before, the size of each point in the figures is proportional to the mass accumulated in the particle located there.

Even though a complete structure of the exact solution of the initial-value problem (1.5), (3.13) is not available, the obtained solution behavior has been expected (see the discussion at the end of section 2). It is instructive to compare the computed numerical solution with theoretical results presented in [21]. According to [21], if

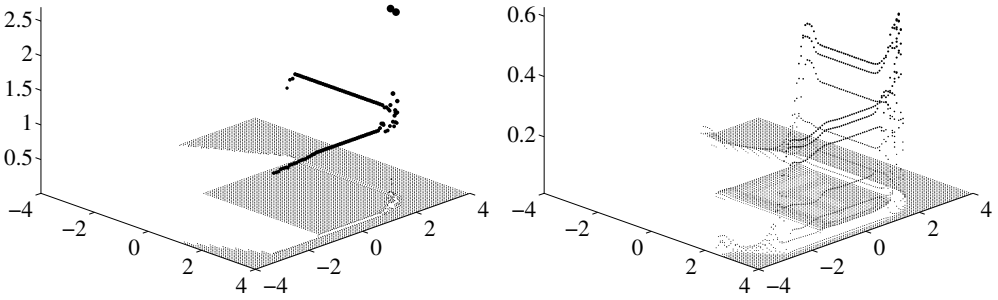


FIG. 11. *Solution (masses) of* (1.5), (3.13) *computed by the SP method (left) and the CU scheme (right).*

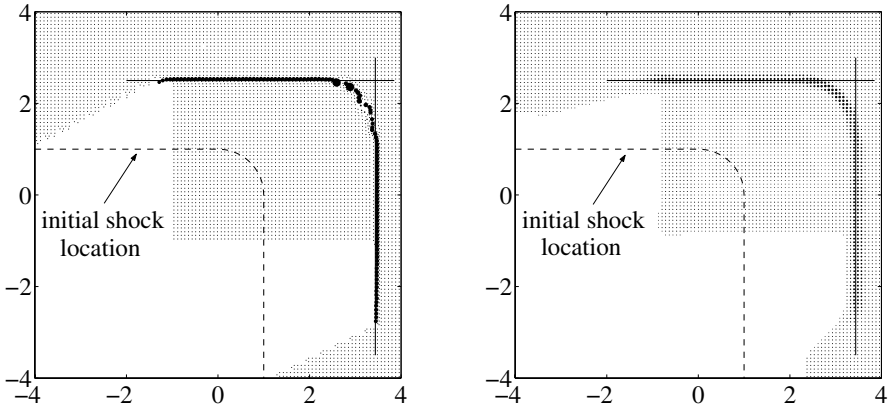


FIG. 12. *Top view on the solution (masses) of* (1.5), (3.13) *computed by the SP method (left) and the CU scheme (right). The solid lines are obtained from the initial shock curve (the dashed line) by the evolution according to* (3.14).

initially a shock curve with mass distribution P_0 and velocities (U_0, V_0) is located along the line $x_0(l) = C \equiv \text{const}$, $y_0(l) = l$, then its location at a later time is given by

$$\begin{cases} x = C + \left[\dfrac{u}{2} - \dfrac{\left(\frac{1}{2}u - U_0\right) P_0}{\rho u t + P_0}\right] t, \\ uy - vx = \dfrac{(uV_0 - vU_0)P_0}{\rho u} \ln\left(1 + \dfrac{\rho u t}{P_0}\right) + ul - vC, \end{cases} \tag{3.14}$$

where ρ, u, and v are the density and the corresponding velocities inside the domain. If we now consider a part of the initial shock curve $\partial\Omega$, namely, $x_0 = 1$, $y_0 = l$, $0 \le l \le 1$, and substitute the corresponding values of $P_0 = 10$, $U_0 = 2$, $V_0 = 1$, and $\rho = u = v = 2$ into the first formula in (3.14), we obtain that at time $t = 1.5$ the shock line should be located at $x = 3.4375$. Similarly, it can be shown that the initial shock line $x_0 = l$, $y_0 = 1$, $0 \le l \le 1$ should move to $y = 2.5$ by the time $t = 1.5$. As one can see from Figure 12, both methods accurately track the evolution of the corresponding

parts of the shock curve: in this figure, the horizontal solid line ($y = 2.5$) and the vertical solid line ($x = 3.4375$) represent the exact shock locations, while the dots are used to plot the numerical solution obtained by the SP method (left) and the CU scheme (right). One can clearly observe a much better resolution of the discontinuity achieved by the SP method.

Example 8. We now consider the system (1.5) subject to the following initial data:

$$(\rho(\mathbf{x},0),u(\mathbf{x},0),v(\mathbf{x},0)) = \begin{cases} (2,2,1) & \text{if } \mathbf{x}\in\Omega_1,\\ (2,1,2) & \text{if } \mathbf{x}\in\Omega_2,\\ (2,\eta,\eta) & \text{if } \mathbf{x}\in\Omega_3,\\ (0,0,0) & \text{if } \mathbf{x}\in\partial\Omega_1\cup\partial\Omega_2\cup\partial\Omega_3,\\ (1,0,0) & \text{otherwise,}\end{cases} \tag{3.15}$$

where $\Omega_1 = \{x < 0,\ x/2+1 < y < 1\}$, $\Omega_2 = \{y < 0,\ y/2+1 < x < 1\}$, and $\Omega_3 = \{x < 0,\ y < x/2+1\}\cup\{x > 0,\ y > 0,\ x^2+y^2 < 1\}\cup\{y < 0,\ x < y/2+1\}$. The initial locations of the discontinuities are shown in Figures 13 and 16. As in the previous two examples, we restrict the initial data (3.15) to the finite domain $[-4,4]\times[-4,4]$ and supplement the initial-value problem (1.5), (3.15) with the solid wall boundary conditions.

Example 8a. We first take $\eta = 1$ in (3.15). In this case, δ-shocks are immediately formed along the initial shock curves. Then, they propagate and develop stronger δ-type singularities at two points, which later merge into a single one in the upper right corner of the computational domain (as in the previous numerical example, the exact solution of the initial-value problem (1.5), (3.15) is not available, but the obtained solution behavior is in line with our expectations; see the discussion at the end of section 2).

We apply the SP method with initially uniformly distributed 100×100 particles and present the solutions, computed at times $t = 2$ and $t = 4$, in Figures 13 and 14. Once again, the size of each point in the figures is proportional to the mass accumulated in the particle located there. For comparison purposes, we also apply the CU scheme with $\Delta x = \Delta y = 0.08$ to the same initial-boundary value problem. The obtained solution, presented in Figure 15 (left), clearly demonstrates that the resolution achieved by the SP method is by far superior. However, since the exact solution of this test problem is unavailable and since there is a very big discrepancy between the solutions computed by the SP and CU methods, we also apply the CU scheme on a much finer grid with $\Delta x = \Delta y = 0.02$. The obtained solution, shown in Figure 15 (right), looks more like the SP solution in Figure 14 (right), but the resolution is still not as high as the one achieved by our SP method; compare, for instance, the maximal mass values—7.2927, 3.9223, and 1.0205—of the solutions, computed by the SP method, the CU scheme on the fine grid, and the CU scheme on the coarse grid, respectively.

Example 8b. Next, we take $\eta = 2-1/\sqrt{2}$ in (3.15). In this case, we observe a more clear structure of the formed δ-shocks, which then interact with two contact waves. This interaction, as in Example 8a, leads to formation of strong singularities. Such a structure—strong singularities emerging from δ-shock curves—is anticipated as a typical one for 2-D pressureless gases; see the discussion at the end of section 2. See also [21] and the references therein.

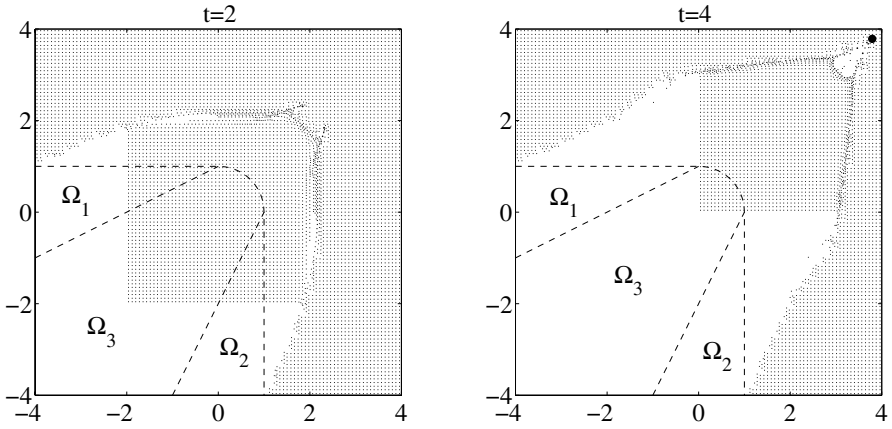


FIG. 13. *Top view on the solution (masses) of* (1.5), (3.15) *with* $\eta = 1$ *at* $t = 2$ *(left) and* $t = 4$ *(right) computed by the SP method. The dashed lines represent the initial location of discontinuities.*

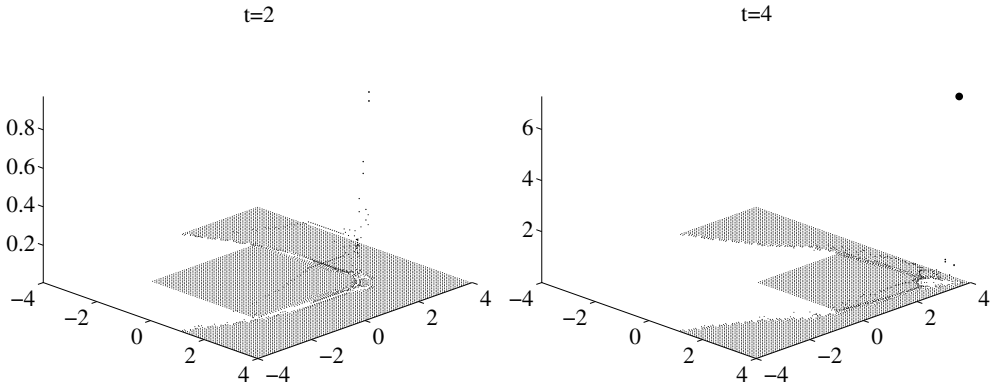


FIG. 14. *Solution (masses) of* (1.5), (3.15) *with* $\eta = 1$ *at* $t = 2$ *(left) and* $t = 4$ *(right) computed by the SP method.*

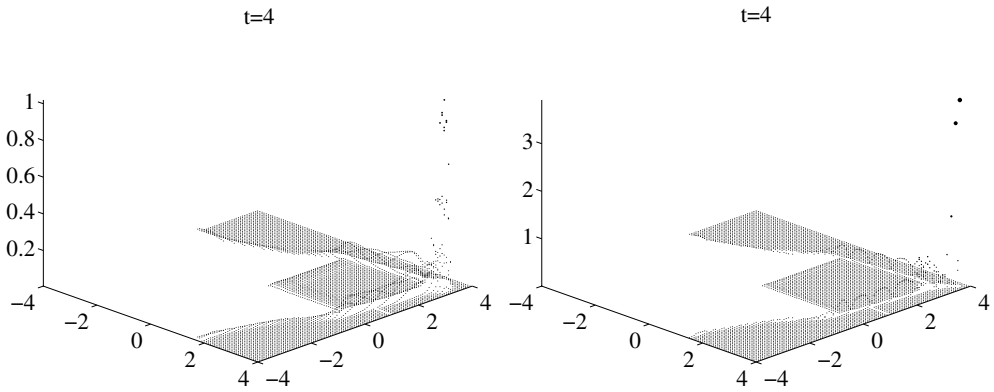


FIG. 15. *Solution (masses) of* (1.5), (3.15) *with* $\eta = 1$ *at* $t = 4$ *computed by the CU scheme with* $\Delta x = \Delta y = 0.08$ *(left) and* $\Delta x = \Delta y = 0.02$ *(right).*

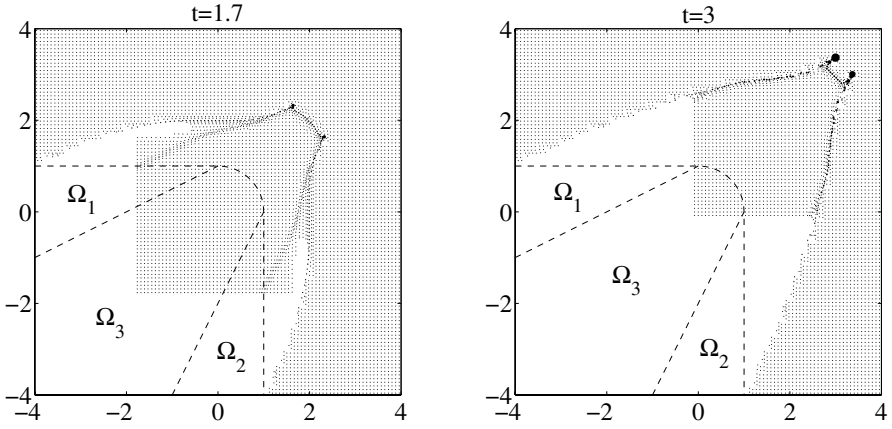


FIG. 16. *Top view on the solution (masses) of* (1.5), (3.15) *with* $\eta = 2 - 1/\sqrt{2}$ *at* $t = 1.7$ *(left) and* $t = 3$ *(right) computed by the SP method. The dashed lines represent the initial shock location.*

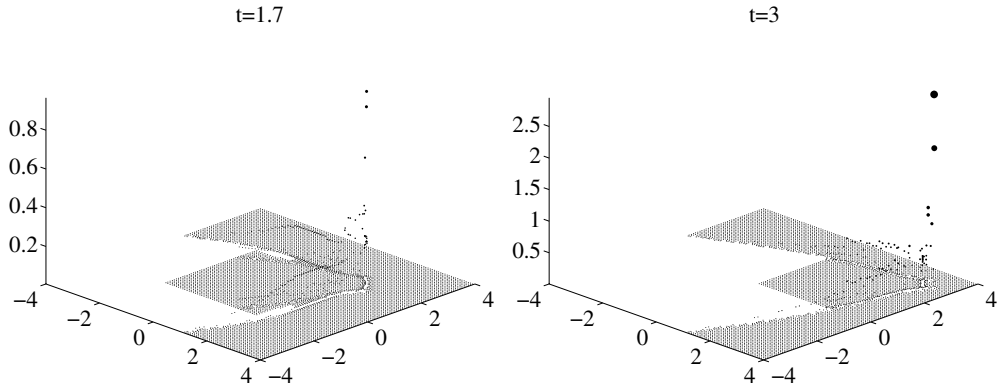


FIG. 17. *Solution (masses) of* (1.5), (3.15) *with* $\eta = 2 - 1/\sqrt{2}$ *at* $t = 1.7$ *(left) and* $t = 3$ *(right) computed by the SP method.*

We apply the SP method with initially uniformly distributed 100×100 particles and present the solutions, computed at times $t = 1.7$ and $t = 3$, in Figures 16 and 17. As before, the size of each point in these figures is proportional to the mass accumulated in the particle located there. We compare the SP solution, presented in Figure 17 (right), with the solution computed by the CU scheme with $\Delta x = \Delta y = 0.08$, which is plotted in Figure 18 (left). One can clearly see the superiority of the results achieved by the SP method. We also apply the CU scheme on a much finer grid with $\Delta x = \Delta y = 0.02$. The obtained solution, shown in Figure 18 (right), looks more like the SP solution in Figure 17 (right), but the resolution is still not as high as the one achieved by the SP method; compare, as before, the maximal mass values—2.9529, 1.7261, and 0.4724—of the solutions, computed by the SP method, the CU scheme on the fine grid, and the CU scheme on the coarse grid, respectively.

Example 9. Finally, we consider the system (1.5) subject to the initial data taken from [1, 23]. In this example, $\rho(x, 0)$ is a Gaussian field, shown in Figures 19 and 20 (a detailed description of its generation can be found in [23, section 5.1]) and

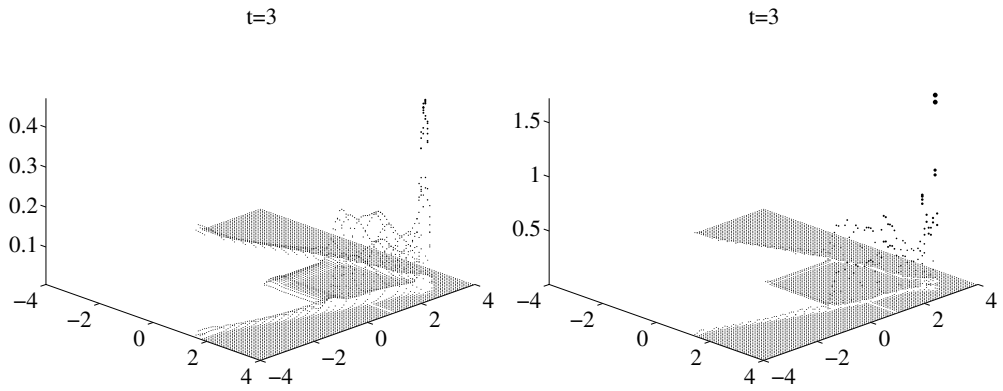


FIG. 18. *Solution (masses) of* (1.5), (3.15) *with* $\eta = 2 - 1/\sqrt{2}$ *at* $t = 3$ *computed by the CU scheme with* $\Delta x = \Delta y = 0.08$ *(left) and* $\Delta x = \Delta y = 0.02$ *(right).*

the initial velocity vector is the solution of the following elliptic problem:

$$\text{(3.16)} \qquad \begin{cases} u = -\phi_x, \\ v = -\phi_y, \\ \Delta\phi = 4\pi G(\rho - \bar{\rho}) \quad \text{in } \Omega = [0, 251] \times [0, 251], \end{cases}$$

where G is the gravitational constant and $\bar{\rho} = \frac{1}{|\Omega|} \int_\Omega \rho \, dx \, dy$. All the boundary conditions are assumed to be periodic. These 2-D physical data are derived from large-scale structure simulations related to the cosmological model of Zeldovich [24]. In Figures 19 and 20, we observe the formation of the large-scale structures computed by the SP method and the CU scheme, respectively. We use a uniform spatial grid with $\Delta x = \Delta y = 1$ for the CU scheme and the uniform initial distribution of 251×251 particles. In order to compare the results, the total mass of each cell computed by the CU scheme has been recalculated at the location of particles. Again, the size of each point in Figures 19 and 20 is proportional to the total mass at this point, that is, bigger points correspond to larger masses. As one can see, for small times both schemes produce very similar results, while for larger times a numerical diffusion present in the CU scheme "takes over" (compare the corresponding results at times $t = 4000$ and $t = 15000$ in Figures 19 and 20). In fact, the maximum mass accumulated at one point by the SP method is about 15 times larger than the one accumulated by the CU scheme. We also would like to point out that, as a result of unification of clustering particles, the number of particles is decreasing in time and therefore the efficiency of the SP method is increasing. For instance, the number of particles at times $t = 1000$, 4000, and 15000 is 2767, 1068, and 454, respectively, while computations using the CU scheme are being performed on a 251×251 grid for all times. This also results in much smaller runtime for the SP method compared to the CU scheme.

4. Concluding remarks. We have presented a new sticky particle (SP) method for the system of Euler equations of pressureless gas dynamics that arises in the modeling of the formation of large-scale structures in the universe. The main feature of interest in this problem is the formation of strong singularities (δ-functions along the surfaces as well as at separate points) and the emergence of vacuum states, and therefore particle methods seem to be a natural choice for numerical simulations of such models. The proposed SP method has been studied both analytically and numerically.

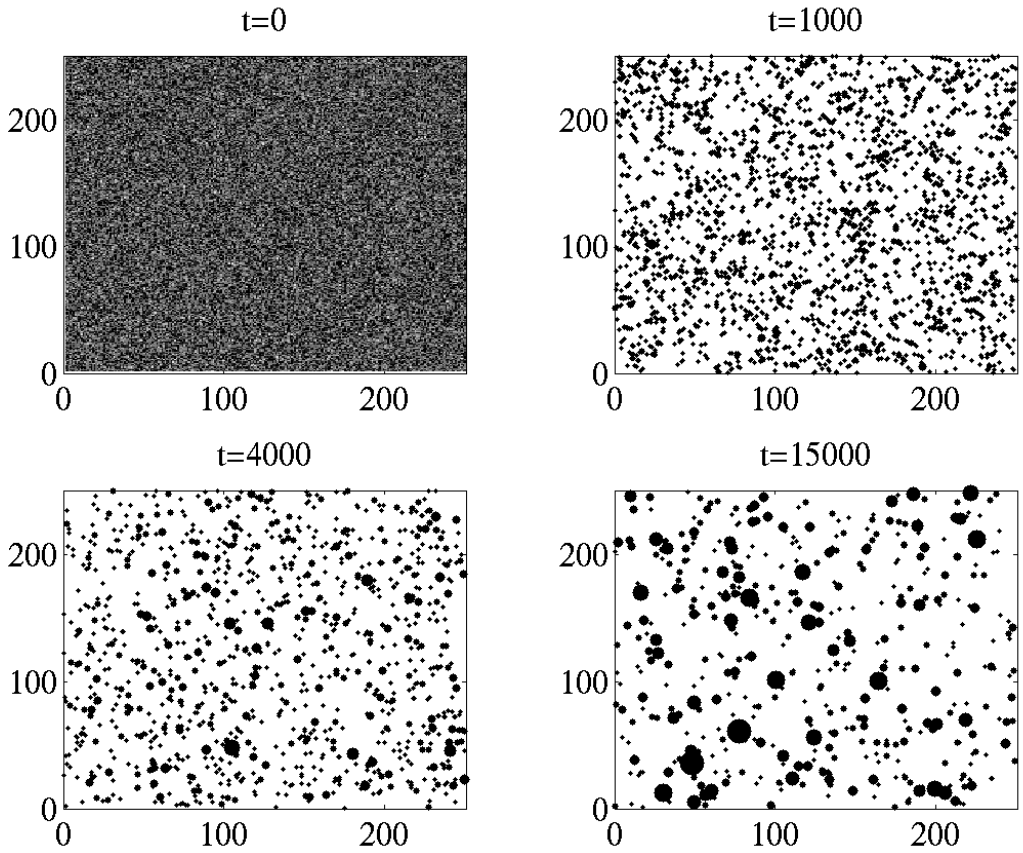


FIG. 19. *Top view on the solution (masses) of* (1.5), (3.16) *computed by the SP method at different times.*

It has been shown that the particle approximation satisfies the original system of pressureless gas dynamics in a weak sense, but only within a certain residual, which has been rigorously estimated. It has also been explained why the relevant errors should diminish as the total number of particles increases. Numerical experiments in one and two space dimensions have been performed (3-D extension of the SP method is out of scope of this paper, but it can be carried out rather straightforwardly). The SP method has been compared to the second-order CU scheme. Our numerical experiments clearly demonstrate the superiority of results obtained by the SP method, which seems to be a robust, accurate, and efficient alternative to existing numerical methods for pressureless gas dynamics.

Appendix A. Semidiscrete central-upwind schemes for pressureless gas dynamics. Here, we briefly describe semidiscrete CU schemes for the 2-D system of pressureless gas dynamics (1.5), which can be written in the following flux-vector form:

$$\mathbf{w}_t + \mathbf{f}(\mathbf{w})_x + \mathbf{g}(\mathbf{w})_y = 0,$$

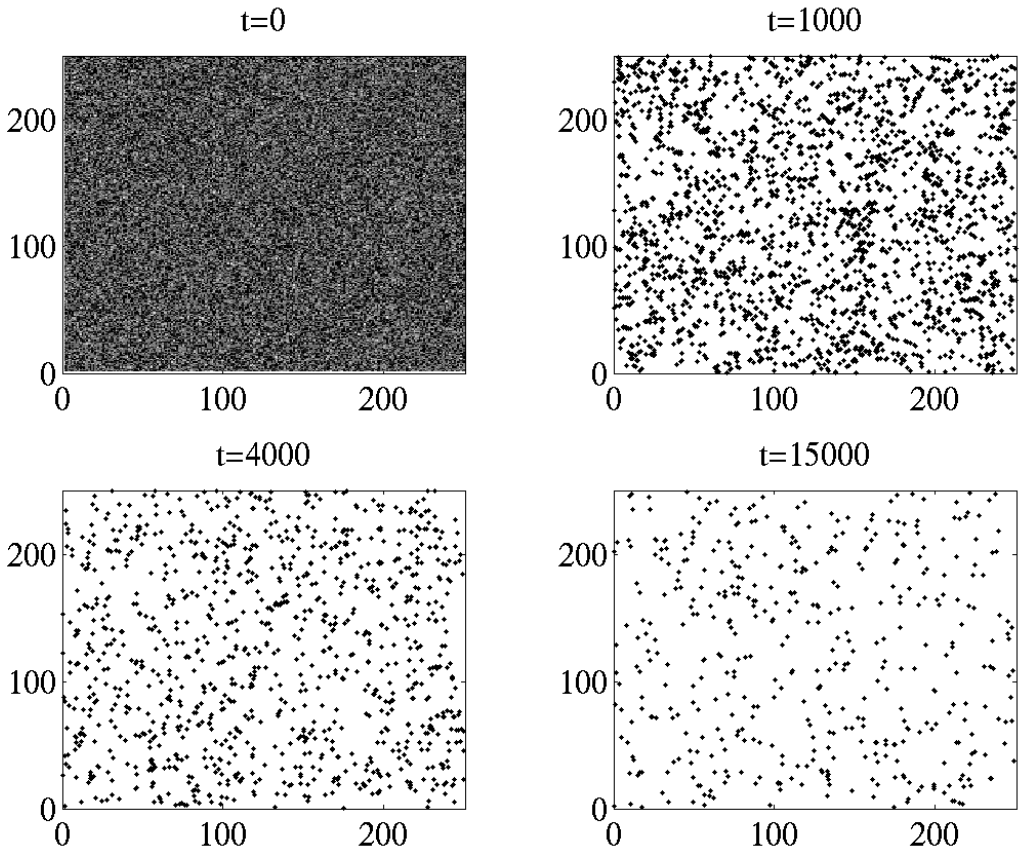


FIG. 20. *Top view on the solution (masses) of* (1.5), (3.16) *computed by the CU scheme at different times.*

where

$$\mathbf{w} := \begin{pmatrix} \rho \\ \rho u \\ \rho v \end{pmatrix}, \quad \mathbf{f}(\mathbf{w}) := \begin{pmatrix} \rho u \\ \rho u^2 \\ \rho uv \end{pmatrix}, \quad \mathbf{g}(\mathbf{w}) := \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 \end{pmatrix}.$$

We consider a uniform spatial grid $x_\mu := \mu\Delta x$, $y_\nu := \nu\Delta y$, and denote the computed quantities, the cell averages, by

$$\bar{\mathbf{w}}_{j,k}(t) := \frac{1}{\Delta x \Delta y} \iint\limits_{I_{j,k}} \mathbf{w}(\xi,\eta,t)\,d\eta\,d\xi, \quad I_{j,k} := [x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}}] \times [y_{k-\frac{1}{2}}, y_{k+\frac{1}{2}}].$$

The cell averages are evolved in time according to the semidiscrete CU scheme

$$\frac{d}{dt}\bar{\mathbf{w}}_{j,k}(t) = -\frac{H^x_{j+\frac{1}{2},k}(t) - H^x_{j-\frac{1}{2},k}(t)}{\Delta x} - \frac{H^y_{j,k+\frac{1}{2}}(t) - H^y_{j,k-\frac{1}{2}}(t)}{\Delta y},$$

where the numerical fluxes $H^x_{j+\frac12,k}$ and $H^y_{j,k+\frac12}$ are given by (see [11] for the derivation)

(A.1)
$$H^x_{j+\frac12,k}=\frac{a^+_{j+\frac12,k}\mathbf{f}(\mathbf{w}^{\rm E}_{j,k})-a^-_{j+\frac12,k}\mathbf{f}(\mathbf{w}^{\rm W}_{j+1,k})}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}}+a^+_{j+\frac12,k}a^-_{j+\frac12,k}\left[\frac{\mathbf{w}^{\rm W}_{j+1,k}-\mathbf{w}^{\rm E}_{j,k}}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}}-\mathbf{q}^x_{j+\frac12,k}\right]$$

and

(A.2)
$$H^y_{j,k+\frac12}=\frac{b^+_{j,k+\frac12}\mathbf{g}(\mathbf{w}^{\rm N}_{j,k})-b^-_{j,k+\frac12}\mathbf{g}(\mathbf{w}^{\rm S}_{j,k+1})}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}}+b^+_{j,k+\frac12}b^-_{j,k+\frac12}\left[\frac{\mathbf{w}^{\rm S}_{j,k+1}-\mathbf{w}^{\rm N}_{j,k}}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}}-\mathbf{q}^y_{j,k+\frac12}\right].$$

Note that all the quantities in (A.1) and (A.2) depend on t, but we will omit this dependence in order to simplify the notation.

In (A.1)–(A.2), the point values $\mathbf{w}^{E(W,S,N)}$ are to be computed from a conservative, nonoscillatory piecewise polynomial reconstruction of an appropriate order. For example, the second-order CU scheme would employ a piecewise linear reconstruction

$$\widetilde{\mathbf{w}}(x,y,t)=\bar{\mathbf{w}}_{j,k}(t)+(\mathbf{w}_x)_{j,k}(x-x_j)+(\mathbf{w}_y)_{j,k}(y-y_k)\quad\text{for}\quad(x,y)\in I_{j,k},\tag{A.3}$$

and the corresponding point values will be

$$\mathbf{w}^{\rm E(W)}_{j,k}:=\bar{\mathbf{w}}_{j,k}(t)\pm\frac{\Delta x}{2}(\mathbf{w}_x)_{j,k},\quad \mathbf{w}^{\rm N(S)}_{j,k}:=\bar{\mathbf{w}}_{j,k}(t)\pm\frac{\Delta y}{2}(\mathbf{w}_y)_{j,k}.$$

To ensure a nonoscillatory property of this reconstruction and thus of the second-order CU scheme, the slopes in (A.3) should be computed with the help of a nonlinear limiter. In our numerical experiments, we have used a one-parameter family of the *minmod* limiters [15, 18, 22]:

$$(\mathbf{w}_x)_{j,k}=\text{minmod}\left(\theta\frac{\bar{\mathbf{w}}_{j+1,k}-\bar{\mathbf{w}}_{j,k}}{\Delta x},\ \frac{\bar{\mathbf{w}}_{j+1,k}-\bar{\mathbf{w}}_{j-1,k}}{2\Delta x},\ \theta\frac{\bar{\mathbf{w}}_{j,k}-\bar{\mathbf{w}}_{j-1,k}}{\Delta x}\right),$$
$$(\mathbf{w}_y)_{j,k}=\text{minmod}\left(\theta\frac{\bar{\mathbf{w}}_{j,k+1}-\bar{\mathbf{w}}_{j,k}}{\Delta y},\ \frac{\bar{\mathbf{w}}_{j,k+1}-\bar{\mathbf{w}}_{j,k-1}}{2\Delta y},\ \theta\frac{\bar{\mathbf{w}}_{j,k}-\bar{\mathbf{w}}_{j,k-1}}{\Delta y}\right),$$

where $\theta\in[1,2]$, and the multivariate minmod function is defined by (2.14). Notice that larger θ's correspond to less dissipative but, in general, more oscillatory limiters (we have used $\theta=1.5$ in all the reported numerical experiments).

Since all the eigenvalues of the Jacobians $\frac{\partial\mathbf{f}}{\partial\mathbf{w}}$ and $\frac{\partial\mathbf{g}}{\partial\mathbf{w}}$ are of multiplicity 3 and are equal to u and v, respectively, the one-sided local speeds in (A.1)–(A.2) are easy to estimate:

$$a^+_{j+\frac12,k}:=\max\left\{u^{\rm W}_{j+1,k},u^{\rm E}_{j,k},0\right\},\quad a^-_{j+\frac12,k}:=\min\left\{u^{\rm W}_{j+1,k},u^{\rm E}_{j,k},0\right\},$$
$$b^+_{j,k+\frac12}:=\max\left\{v^{\rm S}_{j,k+1},v^{\rm N}_{j,k},0\right\},\quad b^-_{j,k+\frac12}:=\min\left\{v^{\rm S}_{j,k+1},v^{\rm N}_{j,k},0\right\}.$$

Finally, $\mathbf{q}^x_{j+\frac12,k}$ and $\mathbf{q}^y_{j,k+\frac12}$ are the "antidiffusion" terms that help to reduce numerical dissipation present at nonoscillatory central schemes [11]:

$$\mathbf{q}^x_{j+\frac12,k}=\text{minmod}\left(\frac{\mathbf{w}^{\rm NW}_{j+1,k}-\mathbf{w}^{\rm int}_{j+\frac12,k}}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}},\frac{\mathbf{w}^{\rm int}_{j+\frac12,k}-\mathbf{w}^{\rm NE}_{j,k}}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}},\right.$$
$$\left.\frac{\mathbf{w}^{\rm SW}_{j+1,k}-\mathbf{w}^{\rm int}_{j+\frac12,k}}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}},\frac{\mathbf{w}^{\rm int}_{j+\frac12,k}-\mathbf{w}^{\rm SE}_{j,k}}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}}\right),$$

$$\mathbf{q}^y_{j,k+\frac12}=\text{minmod}\left(\frac{\mathbf{w}^{\text{SW}}_{j,k+1}-\mathbf{w}^{\text{int}}_{j,k+\frac12}}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}},\frac{\mathbf{w}^{\text{int}}_{j,k+\frac12}-\mathbf{w}^{\text{NW}}_{j,k}}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}},\right.$$
$$\left.\frac{\mathbf{w}^{\text{SE}}_{j,k+1}-\mathbf{w}^{\text{int}}_{j,k+\frac12}}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}},\frac{\mathbf{w}^{\text{int}}_{j,k+\frac12}-\mathbf{w}^{\text{NE}}_{j,k}}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}}\right),$$

where

$$\mathbf{w}^{\text{int}}_{j+\frac12,k}=\frac{a^+_{j+\frac12,k}\mathbf{w}^{\text{W}}_{j+1,k}-a^-_{j+\frac12,k}\mathbf{w}^{\text{E}}_{j,k}-\left\{\mathbf{f}(\mathbf{w}^{\text{W}}_{j+1,k})-\mathbf{f}(\mathbf{w}^{\text{E}}_{j,k})\right\}}{a^+_{j+\frac12,k}-a^-_{j+\frac12,k}},$$
$$\mathbf{w}^{\text{int}}_{j,k+\frac12}=\frac{b^+_{j,k+\frac12}\mathbf{w}^{\text{S}}_{j,k+1}-b^-_{j,k+\frac12}\mathbf{w}^{\text{N}}_{j,k}-\left\{\mathbf{g}(\mathbf{w}^{\text{S}}_{j,k+1})-\mathbf{g}(\mathbf{w}^{\text{N}}_{j,k})\right\}}{b^+_{j,k+\frac12}-b^-_{j,k+\frac12}},$$

and the point values at the cell corners are

$$\mathbf{w}^{\text{NE(NW)}}_{j,k}:=\bar{\mathbf{w}}_{j,k}(t)\pm\frac{\Delta x}{2}(\mathbf{w}_x)_{j,k}+\frac{\Delta y}{2}(\mathbf{w}_y)_{j,k},$$
$$\mathbf{w}^{\text{SE(SW)}}_{j,k}:=\bar{\mathbf{w}}_{j,k}(t)\pm\frac{\Delta x}{2}(\mathbf{w}_x)_{j,k}-\frac{\Delta y}{2}(\mathbf{w}_y)_{j,k}.$$

REFERENCES

[1] C. Berthon, M. Breuss, and M.-O. Titeux, *A relaxation scheme for the approximation of the pressureless Euler equations*, Numer. Methods Partial Differential Equations, 22 (2006), pp. 484–505.

[2] F. Bouchut, *On zero pressure gas dynamics*, in Advances in Kinetic Theory and Computing, World Scientific, River Edge, NJ, 1994, pp. 171–190.

[3] F. Bouchut and G. Bonnaud, *Numerical simulation of relativistic plasmas in hydrodynamic regime*, Z. Angew. Math. Mech., 76 (1996), pp. 287–290.

[4] F. Bouchut and F. James, *Duality solutions for pressureless gases, monotone scalar conservation laws, and uniqueness*, Comm. Partial Differential Equations, 24 (1999), pp. 2173–2189.

[5] F. Bouchut, S. Jin, and X. Li, *Numerical approximations of pressureless and isothermal gas dynamics*, SIAM J. Numer. Anal., 41 (2003), pp. 135–158.

[6] Y. Brenier and E. Grenier, *Sticky particles and scalar conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2317–2328.

[7] G.-Q. Chen and H. Liu, *Formation of δ-shocks and vacuum states in the vanishing pressure limit of solutions to the Euler equations for isentropic fluids*, SIAM J. Math. Anal., 34 (2003), pp. 925–938.

[8] W. E, Yu. G. Rykov, and Ya. G. Sinai, *Generalized variational principles, global weak solutions and behavior with random initial data for systems of conservation laws arising in adhesion particle dynamics*, Comm. Math. Phys., 177 (1996), pp. 349–380.

[9] K. O. Friedrichs, *Symmetric hyperbolic linear differential equations*, Comm. Pure Appl. Math., 7 (1954), pp. 345–392.

[10] S. Gottlieb, C.-W. Shu, and E. Tadmor, *High order time discretization methods with the strong stability property*, SIAM Rev., 43 (2001), pp. 89–112.

[11] A. Kurganov and C.-T. Lin, *On the reduction of numerical dissipation in central-upwind schemes*, Commun. Comput. Phys., 2 (2007), pp. 141–163.

[12] A. Kurganov, S. Noelle, and G. Petrova, *Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 21 (2001), pp. 707–740.

[13] A. Kurganov and G. Petrova, *Central schemes and contact discontinuities*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 1259–1275.

[14] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.

[15] B. van Leer, *Towards the ultimate conservative difference scheme,* V. *A second order sequel to Godunov's method*, J. Comput. Phys., 32 (1979), pp. 101–136.

[16] P. D. Lax, *Weak solutions of nonlinear hyperbolic equations and their numerical computation*, Comm. Pure Appl. Math., 7 (1954), pp. 159–193.
[17] R. J. LeVeque, *The dynamics of pressureless dust clouds and delta waves*, J. Hyperbolic Differ. Equ., 1 (2004), pp. 315–327.
[18] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.
[19] Yu. G. Rykov, *The Numerical Method for Solving the 2-D System of Gas Dynamics without Pressure*, KIAM Preprint 76, Moscow, 1996.
[20] Yu. G. Rykov, *Propagation of singularities of shock wave type in a system of equations of two-dimensional pressureless gas dynamics*, Mat. Zametki, 66 (1999), pp. 760–769 (in Russian); translation in Math. Notes, 66 (1999), pp. 628–635 (2000).
[21] Yu. G. Rykov, *On the nonhamiltonian character of shocks in 2-D pressureless gas*, Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8), 5 (2002), pp. 55–78.
[22] P. K. Sweby, *High resolution schemes using flux limiters for hyperbolic conservation laws*, SIAM J. Numer. Anal., 21 (1984), pp. 995–1011.
[23] M. Vergassola, B. Dubrulle, U. Frish, and A. Noullez, *Burgers equations, Dervil's staircases and the mass distribution for large scale structures*, Astron. Astophys., 289 (1994), pp. 325–356.
[24] Ya. B. Zeldovich, *Gravitational instability: An approximate theory for large density perturbations*, Astron. Astrophys., 5 (1970), pp. 84–89.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2442–2467

CENTRAL DISCONTINUOUS GALERKIN METHODS ON OVERLAPPING CELLS WITH A NONOSCILLATORY HIERARCHICAL RECONSTRUCTION*

YINGJIE LIU†, CHI-WANG SHU‡, EITAN TADMOR§, AND MENGPING ZHANG¶

Abstract. The central scheme of Nessyahu and Tadmor [*J. Comput. Phys.*, 87 (1990), pp. 408–463] solves hyperbolic conservation laws on a staggered mesh and avoids solving Riemann problems across cell boundaries. To overcome the difficulty of excessive numerical dissipation for small time steps, the recent work of Kurganov and Tadmor [*J. Comput. Phys.*, 160 (2000), pp. 241–282] employs a variable control volume, which in turn yields a semidiscrete nonstaggered central scheme. Another approach, which we advocate here, is to view the staggered meshes as a collection of overlapping cells and to realize the computed solution by its overlapping cell averages. This leads to a simple technique to avoid the excessive numerical dissipation for small time steps [Y. Liu, *J. Comput. Phys.*, 209 (2005), pp. 82–104]. At the heart of the proposed approach is the evolution of *two* pieces of information per cell, instead of one cell average which characterizes all central and upwind Godunov-type finite volume schemes. Overlapping cells lend themselves to the development of a central-type discontinuous Galerkin (DG) method, following the series of works by Cockburn and Shu [*J. Comput. Phys.*, 141 (1998), pp. 199–224] and the references therein. In this paper we develop a central DG technique for hyperbolic conservation laws, where we take advantage of the redundant representation of the solution on overlapping cells. The use of redundant overlapping cells opens new possibilities beyond those of Godunov-type schemes. In particular, the central DG is coupled with a novel reconstruction procedure which removes spurious oscillations in the presence of shocks. This reconstruction is motivated by the moments limiter of Biswas, Devine, and Flaherty [*Appl. Numer. Math.*, 14 (1994), pp. 255–283] but is otherwise different in its hierarchical approach. The new hierarchical reconstruction involves a MUSCL or a second order ENO reconstruction in each stage of a multilayer reconstruction process without characteristic decomposition. It is compact, easy to implement over arbitrary meshes, and retains the overall preprocessed order of accuracy while effectively removing spurious oscillations around shocks.

Key words. central scheme, discontinuous Galerkin method, ENO scheme, MUSCL scheme, TVD scheme

AMS subject classifications. 65M60, 65M12

DOI. 10.1137/060666974

*Received by the editors August 7, 2006; accepted for publication (in revised form) April 26, 2007; published electronically November 21, 2007.

http://www.siam.org/journals/sinum/45-6/66697.html

†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (yingjie@math.gatech.edu, http://www.math.gatech.edu/~yingjie/). The research of this author was supported in part by NSF grant DMS-0511815.

‡Division of Applied Mathematics, Brown University, Providence, RI 02912 (shu@dam.brown.edu, http://www.dam.brown.edu/people/shu/). The research of this author was supported in part by the Chinese Academy of Sciences while this author was visiting the University of Science and Technology of China (grant 2004-1-8) and the Institute of Computational Mathematics and Scientific/Engineering Computing. Additional support was provided by ARO grant W911NF-04-1-0291 and NSF grant DMS-0510345.

§Department of Mathematics, Institute for Physical Science and Technology and Center of Scientific Computation and Mathematical Modeling (CSCAMM), University of Maryland, College Park, MD 20742 (tadmor@cscamm.umd.edu, http://www.cscamm.umd.edu/people/faculty/tadmor). The research of this author was supported in part by NSF grant 04-07704 and ONR grant N00014-91-J-1076.

¶Department of Mathematics, University of Science and Technology of China, Hefei, Anhui 230026, China (mpzhang@ustc.edu.cn). The research of this author was supported in part by Chinese Academy of Sciences grant 2004-1-8.

1. Introduction. The first order Godunov and Lax–Friedrichs (LxF) schemes are, respectively, the forerunners for the large classes of upwind and central high-resolution schemes for nonlinear conservation laws and related equations. The Godunov scheme captures shock waves monotonically in narrow transition layers. It is based on evolving a piecewise cell average representation of the solution by evaluating the fluxes at the boundaries of each cell which are obtained from the solution of (approximate) Riemann problems along the boundary interfaces. Various higher order generalizations of Godunov scheme have been developed since the mid 1970s. They employ higher order piecewise polynomials which are reconstructed from the evolving cell averages "in the direction of smoothness." We mention here the notable examples of the high-resolution upwind FCT, MUSCL, TVD, PPM, ENO, and WENO schemes [8, 49, 19, 16, 20, 35], and this list is far from complete. The use of intricate Riemann solvers can be avoided at the expense of using the more diffusive LxF scheme. The excessive numerical dissipation can be reduced significantly, however, when higher order piecewise polynomial reconstructions are used in conjunction with the staggered formulation of the LxF scheme. The central scheme of Nessyahu and Tadmor (NT) [40] provides such a second order generalization of the staggered LxF scheme. It is based on the same piecewise linear reconstructions of cell averages used with upwind schemes, yet the solution of (approximate) Riemann problems is avoided. High-resolution generalizations of the NT scheme were developed since the 1990s as the class of central schemes in, e.g., [43, 3, 22, 21, 36, 6, 25, 2, 27, 28, 32], and here too the list is far from complete. The relaxation scheme of Jin and Xin [23] provides another approach which leads to a staggered central stencil for solving nonlinear conservation laws.

Being free of the (eigenstructure of) the underlying Riemann problems, central schemes provide black-box–type methods for the approximate solution of nonlinear hyperbolic conservation laws and other closely related equations [5]. Essentially, one only needs to supply the flux functions. But the staggered high order central schemes of order, say, $r > 1$ still share one disadvantage with the original LxF scheme, namely, the amplitude of their numerical viscosity of order $\mathcal{O}((\Delta x)^{r+1}/\Delta t)$. It excludes the use of small time steps, Δt, which are too small relative to the spatial grid size Δx. The problem lies with the space-time control volumes which are staggered "$\Delta x/2$-away" from each other. (Similar difficulties occur with the two-dimensional (2D) conservative front tracking method which was overcome by Glimm et al. in [17] using space-time cells instead of rezoning.) This problem was addressed by Kurganov and Tadmor who introduced, in [28], a new type of central scheme whose numerical viscosity is independent of $\mathcal{O}(1/\Delta t)$). This was achieved by using variable control volumes so that cells are staggered only "$\mathcal{O}(\Delta t)$-away" from each other. The latest version of the central-upwind scheme has been recently derived in [26]. It allows implementation of central schemes with arbitrarily small time step, and, in particular, it yields a semidiscrete formulation which can be conveniently integrated by ODE solvers, e.g., the strong stability preserving (SSP) Runge–Kutta methods of [45]; consult, e.g., [18]. Similar advantages of a semidiscrete formulation can be achieved when a local LxF building block is used over nonstaggered meshes; see, e.g., Shu and Osher [45, 46] and Liu and Osher [34]. The upwind and central schemes mentioned so far share one thing in common—they evolve *one* piece of information per cell, that is, the cell average. Upwind schemes use Riemann solvers, while central schemes use simpler quadrature rules. For higher accuracy, they both employ piecewise polynomial representation of the solution which is reconstructed from these cell averages.

In [38], Y. Liu introduced an alternative technique for controlling the numerical

dissipation of central schemes. The main idea is to evolve the solution over *overlapping cells*. That is, two sets of cell averages are realized over interlacing grids. The solution is then represented as a convex combination—an "$\mathcal{O}(\Delta t)$-weighted" combination of these overlapping cell averages. The resulting scheme has a numerical viscosity which is independent of $\mathcal{O}(1/\Delta t)$, and as such it admits a semidiscrete formulation which can be integrated using SSP methods. The use of overlapping cells, however, is fundamentally different in that it evolves *two* independent quantities for each given cell, that is, the two overlapping subcell averages. The use of overlapping cells opens many new possibilities. For example, instead of the usual reconstructions such as MUSCL and (W)ENO, overlapping cells offer a more efficient approach for high-resolution: by adding the two subcell averages, we recover the evolution of a full cell average, where by taking their difference, we independently evolve an approximate slope, rather than reconstructing it from neighboring averages. This makes feasible the use of the central discontinuous Galerkin (DG) approach over overlapping cells, following the series of works by Cockburn and Shu [13, 14, 15]. Thus, in particular, the use of overlapping cells yields the versatility of finite element (Galerkin) methods which can be easily formulated on general unstructured meshes with any formal order, since no reconstruction is involved. In this paper, we further develop the staggered central DG method introduced in [37] for solving hyperbolic conservation laws. Numerical tests are performed up to third order accuracy on uniform staggered meshes in one and two dimensions. Stability analysis and error estimates, and extension of the method for time-dependent and steady state convection-diffusion equations, constitute ongoing work and will be reported in the future. Here, one does not reconstruct a piecewise-polynomial representation of the solution; rather it is part of the evolution of higher moments. Still, a nonlinear limiting procedure is necessary to reduce spurious oscillations for high order methods. We introduce here such a general nonoscillatory procedure, the so-called hierarchical reconstruction, interesting in its own right, which is closely related to the moment limiters of Biswas, Devine, and Flaherty [7] and to the earlier work of Cockburn and Shu [13]. Since this limiting procedure requires only linear reconstructions using information from adjacent cells without characteristic decomposition, it can be easily implemented for any shapes of the cells and hence is practical also for unstructured meshes or even dynamically moving meshes (e.g., Tang and Tang [47]), although we do not pursue it in this paper. We refer the reader to [3] and the references therein for a systematic study of central schemes on unstructured grids using the framework of discontinuous finite elements.

This paper is organized as follows. In section 2, we briefly describe the central finite volume scheme on overlapping cells as the background. The natural extension to the central DG scheme on overlapping cells is discussed in section 3. In subsection 3.1 we study the numerical convergence rate for a number of linear and nonlinear equations having smooth solutions. In section 4, we introduce a general nonoscillatory hierarchical reconstruction procedure and use it as a limiter for the central DG scheme on overlapping cells to control spurious oscillations in the presence of shocks. Numerical results testing the accuracy of the proposed schemes are included in sections 3 and 4. Additional numerical results are presented in section 5. Concluding remarks and a plan for future work are included in section 6.

2. Central schemes on overlapping cells. Consider the scalar one-dimensional (1D) conservation law

$$\frac{\partial u}{\partial t} + \frac{\partial f(u)}{\partial x} = 0, \qquad (x,t) \in \mathcal{R} \times (0,T). \tag{2.1}$$

Set $\{x_i := x_0 + i\Delta x\}$, let $C_{i+1/2} := [x_i, x_{i+1})$ be a uniform partition of $\mathcal{R}$, and let $\{\overline{U}^n_{i+1/2}\}$ denote the set of approximate cell averages $\overline{U}^n_{i+1/2} \approx (1/\Delta x)\int_{C_{i+1/2}} u(x,t^n)dx$. Similarly, we set $D_i := [x_{i-1/2}, x_{i+1/2})$ as the dual partition and let $\{\overline{V}^n_i\}$ denote the corresponding set of approximate cell averages $\overline{V}^n_i \approx (1/\Delta x)\int_{D_i} u(x,t^n)dx$. Starting with these two piecewise-constant approximations,[1]

$$\sum_i \overline{U}^n_{i+1/2}\mathbf{1}_{C_{i+1/2}}(x) \quad \text{and} \quad \sum_i \overline{V}^n_i \mathbf{1}_{D_i}(x),$$

we proceed to compute our approximate solution at the next time level, $t^{n+1} := t^n + \Delta t^n$. To this end, we reconstruct two higher order nonoscillatory piecewise-polynomial approximations,

$$U^n(x) = \sum_i U_{i+1/2}(x)\mathbf{1}_{C_{i+1/2}}(x) \quad \text{and} \quad V^n(x) = \sum_i V_i(x)\mathbf{1}_{D_i}(x),$$

with breakpoints at x_i, $i = 0, \pm1, \pm2, \dots$, and, respectively, at $x_{i+1/2}$, $i = 0, \pm1, \pm2, \dots$. These piecewise polynomials should be conservative in the sense that $\int_{C_{j+1/2}} U^n(x)dx = \Delta x\overline{U}^n_{j+1/2}$ and $\int_{D_j} V^n(x)dx = \Delta x\overline{V}^n_j$ for all j's. There are large libraries for such conservative, accurate, and nonoscillatory reconstructions; we refer, for example, to the second order example of MUSCL [48], the third order example of [36], the well-known class of high order (W)ENO reconstructions [20, 44], etc. Following Nessyahu and Tadmor [40], the central scheme associated with these piecewise polynomials reads

$$\overline{V}^{n+1}_i = \frac{1}{\Delta x}\int_{D_i} U^n(x)dx - \frac{\Delta t^n}{\Delta x}\Big[f(U^{n+\frac12}(x_{i+1/2})) - f(U^{n+\frac12}(x_{i-1/2}))\Big], \tag{2.2a}$$

$$\overline{U}^{n+1}_{i+1/2} = \frac{1}{\Delta x}\int_{C_{i+1/2}} V^n(x)dx - \frac{\Delta t^n}{\Delta x}\Big[f(V^{n+\frac12}(x_{i+1})) - f(V^{n+\frac12}(x_i))\Big]. \tag{2.2b}$$

To guarantee second order accuracy, the right-hand sides of (2.2a), (2.2b) require the approximate values of $U^{n+\frac12}(x_{j+1/2}) \approx u(x_{j+1/2}, t^{n+\frac12})$ and $V^{n+\frac12}(x_j) \approx u(x_j, t^{n+\frac12})$ to be evaluated at the midpoint $t + \Delta t^n/2$. Replacing the midpoint rule with higher order quadratures yields higher order accuracy; see, e.g., [36, 6].

The central NT scheme (2.2) and its higher order generalizations provide effective high-resolution "black-box" solvers to a wide variety of nonlinear conservation laws. However, when Δt is very small, e.g., with $\Delta t = \mathcal{O}\big((\Delta x)^2\big)$ as required by the CFL condition for convection-diffusion equations, the numerical dissipation of the NT schemes becomes excessively large. The excessive dissipation is due to the staggered grids where, at each time step, cell averages are shifted $\Delta x/2$-away from each other: indeed, at the extreme of $f(u) \equiv 0$, the central scheme (2.2) is reduced to reaveraging at every time step. To address this difficulty, Kurganov and Tadmor [28] suggested removing this excessive dissipation by using staggered grids which are shifted only $\mathcal{O}(\Delta t)$-away from each other. This amounts to using control volumes of width $\mathcal{O}(\Delta t)$ so that the resulting schemes admit a semidiscrete limit as $\Delta t \to 0$, the so-called "central-upwind" schemes introduced in [28] and further generalized in [27]. Liu [38] introduced another modification of the NT scheme which removes its $\mathcal{O}(1/\Delta t)$ dependency of numerical dissipation. In this approach, one takes advantage of the *redundant*

[1] Here and below, $\mathbf{1}_\Omega(x)$ denotes the characteristic function of Ω.

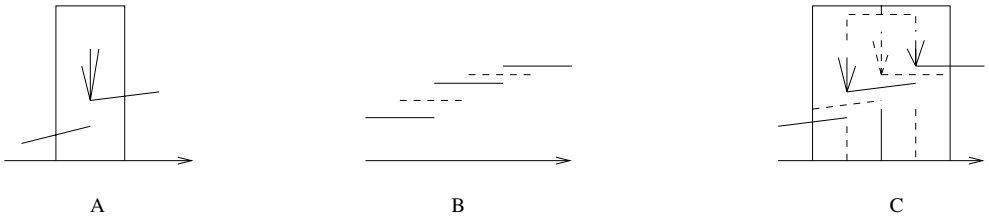


FIG. 1. (A) *NT scheme;* (B) 1*D overlapping cells;* (C) *overlapping cells create self-similarity for the grid over time and allow a convex combination of the overlapping cell averages to control the numerical dissipation.*

representation of the solution over overlapping cells, $\overline{V}_i^n$ and $\overline{U}_{i+1/2}^n$. The idea is to use an $\mathcal{O}(\Delta t)$-dependent weighted average of $\overline{U}_{i+1/2}^n$ and $\overline{V}_i^n$. In fact the difference between them is the local dissipation error. To simplify our discussion, we momentarily give up on second order accuracy in time, setting $U^{n+\frac{1}{2}} = U^n$ and $V^{n+\frac{1}{2}} = V^n$ in (2.2a) and (2.2b). The resulting first order forward-Euler formulation of the new central scheme (consult Figure 1) reads

$$\overline{V}_i^{n+1} = \theta\Big(\frac{1}{\Delta x}\int_{D_i} U^n(x)dx\Big) + (1-\theta)\overline{V}_i^n - \frac{\Delta t^n}{\Delta x}\Big[f(U^n(x_{i+1/2})) - f(U^n(x_{i-1/2}))\Big], \tag{2.3a}$$

$$\overline{U}_{i+1/2}^{n+1} = \theta\Big(\frac{1}{\Delta x}\int_{C_{i+1/2}} V^n(x)dx\Big) + (1-\theta)\overline{U}_{i+1/2}^n - \frac{\Delta t^n}{\Delta x}\Big[f(V^n(x_{i+1})) - f(V^n(x_i))\Big]. \tag{2.3b}$$

Here $\theta := \Delta t^n/\Delta\tau^n$, where $\Delta\tau^n$ is an upper bound for the time step, dictated by the CFL condition. We refer the readers to [40] and [38] for more details to facilitate the full understanding of the sketches in Figure 1. Note that when $\theta = 1$, (2.3a), (2.3b) is reduced to the first order, forward-Euler–based version of the NT scheme (2.2a), (2.2b). Moreover, writing

$$\theta\Big(\frac{1}{\Delta x}\int_{D_i} U^n(x)dx\Big) + (1-\theta)\overline{V}_i^n = \overline{V}_i^n + \frac{\Delta t^n}{\Delta\tau^n}\Big(\frac{1}{\Delta x}\int_{D_i} U^n(x)dx - \overline{V}_i^n\Big),$$

and recalling that $\Delta\tau^n = \mathcal{O}(\Delta x)$ due to the CFL restriction, it follows that the local dissipative error now has a prefactor of order Δt^n, and hence the cumulative error will be independent of $\mathcal{O}(\Delta t)$. The reduced dissipation allows us to pass to a semidiscrete formulation: subtracting $\overline{V}_i^n$ and $\overline{U}_{i+1/2}^n$ from both sides, multiplying by $\frac{1}{\Delta t^n}$, and then passing to the limit as $\Delta t^n \to 0$, we end up with

$$\frac{d}{dt}\overline{V}_i(t^n) = \frac{1}{\Delta\tau^n}\left(\frac{1}{\Delta x}\int_{D_i} U^n(x)dx - \overline{V}_i^n\right) - \frac{1}{\Delta x}\Big[f(U^n(x_{i+1/2})) - f(U^n(x_{i-1/2}))\Big], \tag{2.4a}$$

$$\frac{d}{dt}\overline{U}_{i+1/2}(t^n) = \frac{1}{\Delta\tau^n}\left(\frac{1}{\Delta x}\int_{C_{i+1/2}} V^n(x)dx - \overline{U}_{i+1/2}^n\right) - \frac{1}{\Delta x}\Big[f(V^n(x_{i+1})) - f(V^n(x_i))\Big]. \tag{2.4b}$$

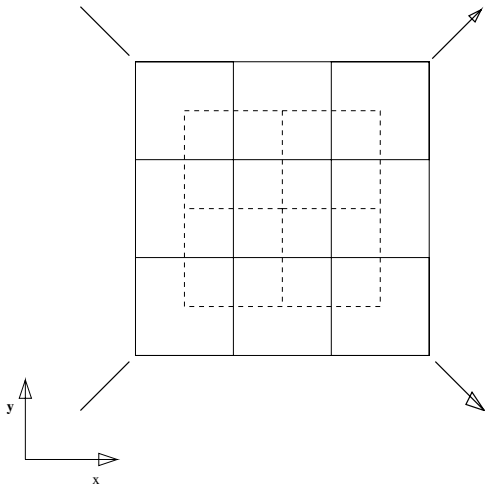


FIG. 2. *2D overlapping cells formed by collapsing the staggered dual cells on two adjacent time levels to one time level.*

The spatial accuracy of the semidiscrete central scheme (2.4) is dictated by the order of the reconstruction $U^n(x)$ and $V^n(x)$. The SSP Runge–Kutta methods yield the matching high order discretization in time.

We conclude this section by quoting [38] regarding the nonoscillatory behavior of the central scheme (2.4), which is quantified here in terms of the total-variation-diminishing (TVD) property; see, e.g., [19].

THEOREM 1. *Consider the central schemes* (2.2) *and* (2.3) *which are set with the same initial values* $\overline{V}_i^n$ *and* $\overline{U}_{i+1/2}^n$ *at* $t = t^n$. *If the NT scheme is TVD, then so is the central scheme* (2.3).

There are two reconstruction procedures for overlapping cells: one is the standard procedure to reconstruct the two classes of cell averages $\{\overline{V}_i^n : i = 0, \pm 1, \pm 2, \ldots\}$ and $\{\overline{U}_{i+1/2}^n : i = 0, \pm 1, \pm 2, \ldots\}$; the other couples these two classes for reconstruction of the final representation of the solution. Thus, this approach is redundant. At the same time, numerical examples in [38] have shown that by coupling the reconstructions, redundancy does provide improved resolution when compared with the one-cell average evolution approach of Godunov-type schemes.

3. A central discontinuous Galerkin method on overlapping cells for conservation laws. Following the general strategy of the DG methods (see, e.g., Lesaint and Raviart [31], Cockburn [10], and Cockburn and Shu [13, 15]), the central-type DG method on overlapping cells can be derived [37]. Consider the system of conservation laws

$$\frac{\partial u_k}{\partial t} + \nabla_{\mathbf{x}} \cdot \mathbf{f}_k(\mathbf{u}) = 0, \qquad (\mathbf{x}, t) \in \mathcal{R}^d \times (0, T), \quad k = 1, \ldots, m, \tag{3.1}$$

where $\mathbf{u} = (u_1, \ldots, u_m)^\top$. For simplicity, assume a uniform staggered rectangular mesh, depicted in Figure 2, for the 2D case, and we note that a similar formulation is used for irregular staggered meshes, e.g., the Voronoi mesh consisting of a triangular mesh and its dual.

Let $\{C_{I+1/2}\}$, $I = (i_1, i_2, \ldots, i_d)$, be a partition of R^d into uniform square cells, depicted by solid lines in Figure 2, and tagged by their cell centroids at the *half*

integers, $\mathbf{x}_{I+1/2} := (I+1/2)\Delta x$. Let $\mathcal{M}$ denote the set of piecewise polynomials of degree r on the cells $\{C_{I+1/2}\}$; no continuity is assumed across cell boundaries. Let $\{D_I\}$ be the dual mesh which consists of a $\Delta x/2$ shift of the $C_{I+1/2}$'s, depicted by dashed lines in Figure 2. Let $\mathbf{x}_I$ be the cell centroid of the cell D_I and let $\mathcal{N}$ denote the set of piecewise polynomials of degree r over the cells $\{D_I\}$; again, no continuity is assumed across the cell boundary. The weak formulation of (3.1) over these cells reads

$$\frac{d}{dt}\int_{C_{I+1/2}} u_k\phi d\mathbf{x} = \int_{C_{I+1/2}} \mathbf{f}_k\cdot\nabla_{\mathbf{x}}\phi d\mathbf{x} - \int_{\partial C_{I+1/2}} (\mathbf{f}_k\cdot\mathbf{n})\phi ds \quad \forall\phi\in\mathcal{M},\ k=1,\dots,m, \tag{3.2a}$$

$$\frac{d}{dt}\int_{D_I} u_k\psi d\mathbf{x} = \int_{D_I} \mathbf{f}_k\cdot\nabla_{\mathbf{x}}\psi d\mathbf{x} - \int_{\partial D_I} (\mathbf{f}_k\cdot\mathbf{n})\psi ds \quad \forall\psi\in\mathcal{N},\ k=1,\dots,m, \tag{3.2b}$$

where $\mathbf{n}$ is the unit outer normal of the corresponding cell and ϕ and ψ are test functions. As in the 1D setup, we let

$$\mathbf{U}^n(\mathbf{x}) = \sum_{I+1/2} \mathbf{U}^n_{I+1/2}(\mathbf{x})\mathbf{1}_{C_{I+1/2}}(\mathbf{x}) \in \mathcal{M} \quad \text{and} \quad \mathbf{V}^n(\mathbf{x}) = \sum_I \mathbf{V}^n_I(\mathbf{x})\mathbf{1}_{D_I}(\mathbf{x}) \in \mathcal{N}$$

denote two representations of the numerical solution, approximating $\mathbf{u}(\cdot,t^n)$ over the two overlapping grids, $\{C_{I+1/2}\}$ and $\{D_I\}$. Observe that each of the two vector functions, $\mathbf{U}^n$ with smooth pieces supported on the $C_{I+1/2}$'s and $\mathbf{V}^n$ with smooth pieces supported on the D_I's, consists of m components

$$\mathbf{U}^n(\mathbf{x}) = (U^n_1(\mathbf{x}),\dots,U^n_m(\mathbf{x}))^\top \quad \text{and} \quad \mathbf{V}^n(\mathbf{x}) = (V^n_1(\mathbf{x}),\dots,V^n_m(\mathbf{x}))^\top.$$

Given these conservative, accurate, and nonoscillatory approximations at t^n we proceed to compute the approximate solution at the next time level, $t^{n+1} = t^n + \Delta t^n$. To this end, the exact solution $\mathbf{u}(\mathbf{x},t^n)$ of (3.1) in the right-hand side of (3.2a) is replaced by $\mathbf{V}^n(\mathbf{x}) = (V^n_1,\dots,V^n_m)^\top$; similarly, for the right-hand side of (3.2b) we use the approximate solution $\mathbf{U}^n(\mathbf{x}) = (U^n_1,\dots,U^n_m)^\top$. Further, time derivatives on the left are replaced by forward-Euler time differencing where we use the same type of θ-weighting of the $\mathbf{U}^n$'s and the $\mathbf{V}^n$'s as in (2.3a), (2.3b). In the resulting central DG method one seeks piecewise polynomials $\{\mathbf{U}^{n+1}_{I+1/2}\}\in\mathcal{M}$ and $\{\mathbf{V}^{n+1}_I\}\in\mathcal{N}$ such that for all I's,

$$\begin{aligned}\int_{C_{I+1/2}} U^{n+1}_k(\mathbf{x})\phi(\mathbf{x})d\mathbf{x} &= \theta\int_{C_{I+1/2}} V^n_k(\mathbf{x})\phi(\mathbf{x})d\mathbf{x} + (1-\theta)\int_{C_{I+1/2}} U^n_k(\mathbf{x})\phi(\mathbf{x})d\mathbf{x}\\ &+ \Delta t^n\int_{C_{I+1/2}} \mathbf{f}_k(\mathbf{V}^n(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\phi d\mathbf{x}\\ &- \Delta t^n\int_{\partial C_{I+1/2}} (\mathbf{f}_k(\mathbf{V}^n(\mathbf{x}))\cdot\mathbf{n})\phi(\mathbf{x})ds \quad \forall\phi\in\mathcal{M},\ k=1,\dots,m,\end{aligned} \tag{3.3a}$$

$$\int_{D_I} V_k^{n+1}(\mathbf{x})\psi(\mathbf{x})d\mathbf{x} = \theta \int_{D_I} U_k^n(\mathbf{x})\psi(\mathbf{x})d\mathbf{x} + (1-\theta)\int_{D_I} V_k^n(\mathbf{x})\psi(\mathbf{x})d\mathbf{x}$$
$$+ \Delta t^n \int_{D_I} \mathbf{f}_k(\mathbf{U}^n(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\psi(\mathbf{x})d\mathbf{x} \tag{3.3b}$$
$$- \Delta t^n \int_{\partial D_I} (\mathbf{f}_k(\mathbf{U}^n(\mathbf{x}))\cdot\mathbf{n})\psi(\mathbf{x})ds \quad \forall\psi\in\mathcal{N},\ k=1,\dots,m.$$

Here $\theta = \Delta t^n/\Delta\tau^n \le 1$, $\Delta\tau^n$ is the maximum time step size determined by the CFL restriction, and $\Delta t^n = t^{n+1} - t^n$ is the current time step size. $\Delta\tau^n$ = (CFL factor) × Δx/(maximum characteristic speed), where the CFL factor should be less than 1/2. At time t^n, $\Delta\tau^n$ is first chosen with a certain CFL factor, then Δt^n has the freedom to take any value in $(0, \Delta\tau^n]$ without introducing excessive dissipation. The smaller $\Delta\tau^n$ is chosen, the larger the numerical dissipation is. We find in numerical experiments that setting $\Delta\tau^n$ with CFL factor 0.4 is robust. In some numerical tests with fewer interactions of discontinuities, we can choose larger $\Delta\tau^n$. This forward-Euler step is going to be used in an SSP Runge–Kutta method of desired order. For the pure hyperbolic problem, Δt^n can be chosen as large as possible, i.e., $\Delta t^n = \Delta\tau^n$ for efficiency.

The resulting central DG is the 2D analogue of the 1D central scheme (2.4). And as in the 1D case, the semidiscrete version of (3.3) can be obtained; higher order Runge–Kutta time discretization can be used to match the high order accuracy of the spatial reconstructions. We conclude with the semidiscrete central DG approximation of (3.1) such that for all admissible test functions ϕ and ψ and all I's,

$$\frac{d}{dt}\int_{C_{I+1/2}} U_k\phi d\mathbf{x} = \frac{1}{\Delta\tau}\int_{C_{I+1/2}} (V_k(\mathbf{x}) - U_k(\mathbf{x}))\phi(\mathbf{x})d\mathbf{x}$$
$$+ \int_{C_{I+1/2}} \mathbf{f}_k(\mathbf{V}(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\phi d\mathbf{x} \tag{3.4a}$$
$$- \int_{\partial C_{I+1/2}} (\mathbf{f}_k(\mathbf{V}(\mathbf{x}))\cdot\mathbf{n})\phi(\mathbf{x})ds \quad \forall\phi\in\mathcal{M},\ k=1,\dots,m,$$
$$\frac{d}{dt}\int_{D_I} V_k\psi d\mathbf{x} = \frac{1}{\Delta\tau}\int_{D_I} (U_k(\mathbf{x}) - V_k(\mathbf{x}))\psi(\mathbf{x})d\mathbf{x}$$
$$+ \int_{D_I} \mathbf{f}_k(\mathbf{U}(\mathbf{x}))\cdot\nabla_{\mathbf{x}}\psi(\mathbf{x})d\mathbf{x} \tag{3.4b}$$
$$- \int_{\partial D_I} (\mathbf{f}_k(\mathbf{U}(\mathbf{x}))\cdot\mathbf{n})\psi(\mathbf{x})ds \quad \forall\psi\in\mathcal{N},\ k=1,\dots,m.$$

For example, consider the piecewise quadratic element in two dimensions; see, e.g., Figure 2. We use the third order SSP Runge–Kutta method [45] to discretize (3.4) in time, which ends up with calling the forward-Euler step (3.3) three times. Let cell $C_{I+1/2}$ as in (3.3a) be the cell bounded by solid lines in the center of Figure 2, and let

$$\begin{aligned} U_{I+1/2}(x - x_{I+1/2}, y - y_{I+1/2}) = {} & U_{I+1/2}(0,0) + \partial_x U_{I+1/2}(0,0)(x - x_{I+1/2}) \\ & + \partial_y U_{I+1/2}(0,0)(y - y_{I+1/2}) \\ & + \frac{1}{2}\partial_{xx} U_{I+1/2}(0,0)(x - x_{I+1/2})^2 \\ & + \partial_{xy} U_{I+1/2}(0,0)(x - x_{I+1/2})(y - y_{I+1/2}) \\ & + \frac{1}{2}\partial_{yy} U_{I+1/2}(0,0)(y - y_{I+1/2})^2 \end{aligned}$$

TABLE 1
P^1 *version of the central DG scheme* (3.4) *for the linear convection equation* (3.5).

Δx	1/20	1/40	1/80	1/160	1/320
L_1 *error*	8.91E-3	2.25E-3	5.66E-4	1.42E-4	3.54E-5
order	-	1.99	1.99	1.99	2.00
L_∞ *error*	5.92E-3	1.55E-3	3.96E-4	1.00E-4	2.51E-5
order	-	1.93	1.97	1.99	1.99

TABLE 2
P^1 *version of the central DG scheme* (3.4) *for the* 2*D Burgers equation.*

Δx	1/2	1/4	1/8	1/16	1/32
L_1 *error*	6.69E-2	3.29E-2	5.04E-3	1.66E-3	3.88E-4
order	-	1.02	2.70	1.60	2.10
L_∞ *error*	3.85E-2	2.05E-2	7.69E-3	1.19E-3	2.75E-4
order	-	0.91	1.41	2.69	2.11

be $U_k^{n+1}|_{C_{I+1/2}}$, i.e., $U_k^{n+1}(\mathbf{x})$ restricted in cell $C_{I+1/2}$, where $(x_{I+1/2}, y_{I+1/2})$ is the cell centroid of cell $C_{I+1/2}$. There are six coefficients to be determined in this polynomial in cell $C_{I+1/2}$, namely,

$$U_{I+1/2}(0,0),\ \partial_x U_{I+1/2}(0,0),\ \partial_y U_{I+1/2}(0,0),$$
$$\frac{1}{2}\partial_{xx} U_{I+1/2}(0,0),\ \partial_{xy} U_{I+1/2}(0,0),\ \frac{1}{2}\partial_{yy} U_{I+1/2}(0,0).$$

By letting

$$\phi(\mathbf{x}) = 1,\ x - x_{I+1/2},\ y - y_{I+1/2},\ (x - x_{I+1/2})^2,$$
$$(x - x_{I+1/2})(y - y_{I+1/2}),\ \text{or}\ (y - y_{I+1/2})^2,$$

we obtain six linear equations in (3.3a) to solve for $U_{I+1/2}(x - x_{I+1/2},\, y - y_{I+1/2})$. The last two integrals in (3.3a) can be approximated by Gaussian quadratures, such as the three-point Gaussian quadrature for line integrals. The other integrals on the right-hand side of (3.3a) can be evaluated exactly.

3.1. Numerical errors for smooth solutions. In this subsection we study the convergence rate for a number of equations having smooth solutions. The examples are computed by linear schemes described previously without using any limiter.

Example 1. Let us start with the following linear transport equation with periodic boundary conditions:

$$u_t + a u_x = 0, \qquad (x,t) \in (0,2) \times (0,2), \tag{3.5}$$
$$u(x,0) = 1 + \sin(\pi x), \qquad x \in (0,2),$$

where $a = 1$ by default.

The test results at $T = 2$ for the P^1 (piecewise linear) version of the central DG scheme on overlapping cells (3.4) are listed in Table 1, with second order Runge–Kutta time discretization. The CFL factor is 0.4 for choosing $\Delta\tau$ and the actual time step size Δt is chosen with $\theta = 0.9$. It can be seen that the expected second order accuracy is achieved. Similar results for the 2D Burgers equation can be found in Table 2. The results for the P^2 (piecewise quadratic) version of the scheme (3.4) for the linear convection equation (3.5) are listed in Table 3, with a third order TVD

TABLE 3
P^2 *version of the central DG scheme* (3.4) *for the linear convection equation* (3.5).

Δx	1/20	1/40	1/80	1/160	1/320
L_1 *error*	6.50E-5	8.12E-6	1.02E-6	1.27E-7	1.59E-8
order	-	3.00	2.99	3.01	3.00
L_∞ *error*	4.68E-5	5.90E-6	7.40E-7	9.27E-8	1.16E-8
order	-	2.99	3.00	3.00	3.00

TABLE 4
P^2 *version of the central DG scheme* (3.4) *for* (3.5) *with* $a = 0$.

Δx	1/20	1/40	1/80	1/160
L_∞ *error*	9.32E-7	5.89E-8	3.70E-9	2.32E-10
order	-	3.98	3.99	4.00
L_∞ *error,* $\Delta t = \Delta x^2$	9.32E-7	5.89E-8	3.70E-9	2.32E-10
order	-	3.98	3.99	4.00

TABLE 5
P^2 *version of the central DG scheme* (3.4) *for the* 1*D Burgers equation.*

Δx	1/10	1/20	1/40	1/80	1/160
L_1 *error*	2.72E-5	3.41E-6	4.29E-7	5.37E-8	6.78E-9
order	-	3.00	2.99	3.00	2.99
L_∞ *error*	4.00E-5	7.06E-6	8.27E-7	1.04E-7	1.31E-8
order	-	2.50	3.09	2.99	2.99

Runge–Kutta time discretization [45]. The results for the same equation with $a = 0$ are listed in Table 4, in which the first row is computed with the previously chosen Δt and the second row is computed with $\Delta t = \Delta x^2$. We observe that the staggered dissipation error does not increase with a diminishing time step size. We remark that for this special case with $a = 0$, the $\Delta\tau$ can be chosen as $+\infty$, since there is no CFL restriction on the stability time step. With this choice of $\Delta\tau$, our scheme will maintain exactly the initial condition for this degenerated PDE. If we choose a finite $\Delta\tau^n$ anyway, then the initial solution may not be maintained exactly. As to the order of accuracy, we can see that the expected third order accuracy is achieved in Table 3, and fourth order accuracy, which is one order higher than expected, is achieved in Table 4.

Example 2. We test the scheme for the Burgers equation $u_t + (\frac{1}{2}u^2)_x = 0$, $u(x, 0) = \frac{1}{4} + \frac{1}{2}\sin(\pi x)$. The errors are shown in Table 5 at the final time $T = 0.1$ when the solution is still smooth.

Example 3. We conduct a convergence test for the P^1 version of the scheme (3.4) on a 2D problem [11] which is the Burgers equation with periodic initial data: $u_t + (\frac{1}{2}u^2)_x + (\frac{1}{2}u^2)_y = 0$ on $[-1, 1] \times [-1, 1]$, $u(x, y, 0) = \frac{1}{4} + \frac{1}{2}\sin(\pi(x + y))$. The numerical solutions are computed at the final time $T = 0.1$ when the exact solution is still smooth. The CFL factor is 0.4 for choosing $\Delta\tau$ and the actual time step size Δt is chosen with $\theta = 0.9$. The errors are shown in Table 2. Again we observe the expected second order accuracy. Further we test the P^2 version of scheme for the 2D Burgers equation. The errors are shown in Table 6 at the final time $T = 0.1$.

Example 4. The solution of the 2D Burgers equation may contain linear waves; hence we also test the scheme on another 2D equation $u_t + (\frac{1}{2}u^2)_x + (\frac{1}{4}u^4)_y = 0$, $u(x, 0) = \frac{1}{4} + \frac{1}{2}\sin(\pi(x + y))$. The accuracy of the numerical solution is shown at $T = 0.1$ in Table 7.

TABLE 6
P^2 *version of the central DG scheme* (3.4) *for the* 2*D Burgers equation.*

Δx	1/4	1/8	1/16	1/32	1/64
L_1 *error*	8.33E-3	9.58E-4	1.36E-4	1.65E-5	2.14E-6
order	-	3.12	2.82	3.04	2.95
L_∞ *error*	4.56E-3	8.20E-4	1.48E-4	1.95E-5	2.58E-6
order	-	2.48	2.47	2.92	2.92

TABLE 7
P^2 *version of the central DG scheme* (3.4) *for the* 2*D nonlinear equation.*

Δx	1/4	1/8	1/16	1/32	1/64
L_1 *error*	5.35E-3	5.75E-4	6.80E-5	7.81E-6	9.77E-7
order	-	3.22	3.08	3.12	3.00
L_∞ *error*	2.57E-3	3.16E-4	8.00E-5	1.10E-5	1.53E-6
order	-	3.02	1.98	2.86	2.85

It seems that for all these cases the expected third order accuracy is achieved for the P^2 version of scheme, at least for the L_1 errors.

4. A general nonoscillatory hierarchical reconstruction procedure. Compared to finite volume schemes which evolve only cell averages over time, DG methods compute and evolve a high order polynomial in each cell. The challenge lies in determining how to take advantage of the extra information provided by the DG method in each cell and use it in the limiting process where the solution is nonsmooth. The first idea is given by Cockburn and Shu [13] for the DG method which limits the variation between a cell edge value and its cell average by the differences between the cell averages of the current and neighboring cells. The higher Legendre moments are truncated in a cell if nonsmoothness is detected. This process is shown to be total variation bounded in the means. A generalization is introduced in Biswas, Devine, and Flaherty [7], which detects the nonsmoothness in higher degree moments and applies the limiting when necessary from higher to lower moments. In Qiu and Shu [42, 41], a high order WENO reconstruction is used as a limiter for the so-called troubled cells, where the polynomial defined at Gaussian points is reconstructed from a WENO procedure and is projected back to the finite element space to replace the one computed by the DG method. In [41], the Hermite WENO reconstruction takes not only cell averages of a function, but also cell averages of its first order derivatives in order to obtain a compact reconstruction. A similar strategy is used in our nonoscillatory hierarchical reconstruction, where cell averages of various orders of derivatives of a function are to be calculated and used in the reconstruction of linear polynomials at each stage. Our limiting procedure is closely related to that of [7]. Our departure from [7] begins with a different point of view, where the approximation in each cell is viewed as a high degree polynomial, instead of the combination of orthogonal Legendre polynomials advocated in [7]. Instead of a limiting procedure which is trying to set an acceptable range for the coefficient of the Legendre moments (by using the coefficients of lower degree moments), we reconstruct the *complete set* of coefficients of the m-degree polynomial terms, using a nonoscillatory conservative reconstruction which involves previous reconstructed terms of degrees above m. The resulting, so-called *hierarchical reconstruction* algorithm is easy to implement in a multidimensional setting, and there is no need to transform an irregular mesh cell into a rectangular one or use a dimension-by-dimension extension of a 1D limiter. It

is essentially independent of the shapes of mesh cells and is compact because of the conservative nonoscillatory linear reconstruction (such as the MUSCL or second order ENO reconstruction; see [1] for an implementation for unstructured meshes) used at each stage. We now give the details of this reconstruction procedure. For simplicity we discuss only the scalar case. For systems a component-by-component extension is applied without characteristic decomposition.

From scheme (3.4) with the SSP Runge–Kutta methods, we obtain numerical solutions $U^n(\mathbf{x})$ and $V^n(\mathbf{x})$ at time t^n. To simplify the notation we drop the superscript n and write

$$U(\mathbf{x}) = \sum_{I+1/2} U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})\mathbf{1}_{C_{I+1/2}}(\mathbf{x}) \in \mathcal{M},$$
$$V(\mathbf{x}) = \sum_{I} V_I(\mathbf{x}-\mathbf{x}_I)\mathbf{1}_{D_I}(\mathbf{x}) \in \mathcal{N},$$

recalling that $\mathbf{x}_{I+1/2}$ and $\mathbf{x}_I$ are centroids of cells $C_{I+1/2}$ and D_I, respectively; $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ and $V_I(\mathbf{x}-\mathbf{x}_I)$ are the polynomials (of degree r) in cells $C_{I+1/2}$ and D_I, respectively.[2] The task is to reconstruct a "limited" version of these polynomials, retaining high-resolution and removing spurious oscillations. In the following, we discuss a hierarchical reconstruction procedure for recomputing the polynomial $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ by using polynomials in cells adjacent to cell $C_{I+1/2}$. For convenience these adjacent cells are renamed as the set $\{C_J\}$ (which contain cells $C_{I+1/2}$, D_I, etc.), and the polynomials (of degree r) supported on them are thus renamed as $\{U_J(\mathbf{x}-\mathbf{x}_J)\}$, respectively, where $\mathbf{x}_J$ is the cell centroid of cell C_J. We write $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ in terms of its Taylor expansion,

$$U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2}) = \sum_{m=0}^{r}\sum_{|\mathbf{m}|=m}\frac{1}{\mathbf{m}!}U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})(\mathbf{x}-\mathbf{x}_{I+1/2})^{\mathbf{m}},$$

where $\frac{1}{\mathbf{m}!}U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$ are the coefficients which participate in its typical m-degree terms,

$$\sum_{|\mathbf{m}|=m}\frac{1}{\mathbf{m}!}U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})(\mathbf{x}-\mathbf{x}_{I+1/2})^{\mathbf{m}}, \qquad |\mathbf{m}| = 0,\dots,r,$$

$\mathbf{m} = (m_1, m_2, \dots, m_d)$ is the multi-index,

$$|\mathbf{m}| = \sum_{i=1}^{d} m_i, \quad \mathbf{m}! = \prod_{i=1}^{d} m_i!, \quad U^{(\mathbf{m})}_{I+1/2}(\mathbf{x}) = \partial^{m_d}_{x_d}\cdots\partial^{m_1}_{x_1}U_{I+1/2}(\mathbf{x}),$$

and $\mathbf{x} = (x_1, \dots, x_d)$. The following hierarchical reconstruction describes a procedure to compute the new coefficients,

$$\frac{1}{\mathbf{m}!}\widetilde{U}^{(\mathbf{m})}_{I+1/2}(\mathbf{0}), \qquad m = r, r-1, \dots, 0,$$

in $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$, iterating from the highest to the lowest degree terms.

[2] These polynomials could be oscillatory. There could be other methods to compute these polynomials such as a finite volume reconstruction from cell averages.

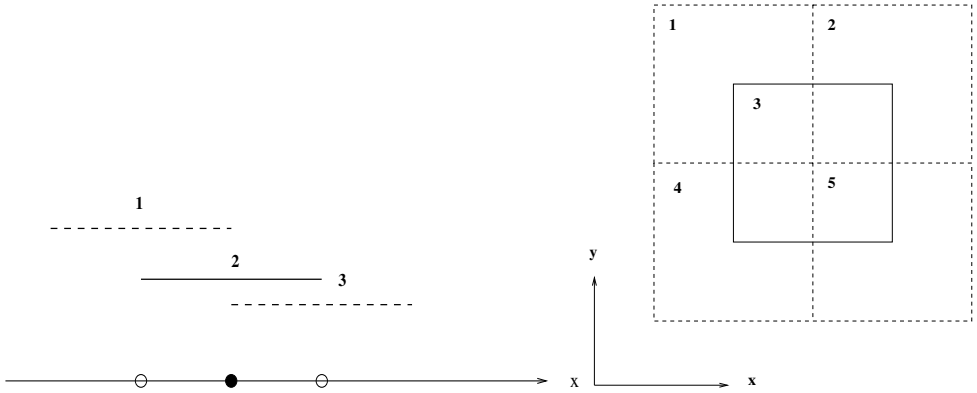


FIG. 3. *Left:* 1*D nonoscillatory hierarchical reconstruction for cell* 2 *involves only overlapping cells* 1, 2, *and* 3. *Right:* 2*D nonoscillatory hierarchical reconstruction for cell* 3 *involves only overlapping cells* 1, 2, 3, 4, *and* 5.

4.1. An example for piecewise quadratic finite element space in one dimension. Suppose $U_j(x-x_j) = U_j(0)+U_j'(0)(x-x_j)+\frac{1}{2}U_j''(0)(x-x_j)^2$, $j = 1, 2, 3$, are given at cells C_1, C_2, and C_3, respectively (see Figure 3, left), where x_j is the center of cell C_j. These polynomials could be oscillatory if located near a discontinuity of the weak solution. The following algorithm computes a new value for each coefficient in the polynomial defined on cell C_2 in order to reduce the oscillation while keeping the accuracy (in the smooth area) and resolution.

Step 1. (1) Take the first derivative for them to obtain $L_j(x - x_j) = U_j'(0) + U_j''(0)(x - x_j)$, $j = 1, 2, 3$.

(2) Calculate the cell average of $L_j(x - x_j)$ on cell C_j to obtain $\overline{L}_j = U_j'(0)$, $j = 1, 2, 3$.

(3) With the three cell averages one can apply a MUSCL or second order ENO procedure to reconstruct a nonoscillatory linear polynomial in cell C_2. The slope of this new linear polynomial corresponds to the slope $U_2''(0)$ of the original linear polynomial $L_2(x-x_2)$ in cell C_2 and is denoted by $\widetilde{U}_2''(0)$. The details can be explained as follows.

Using the technique of [1], let the new linear polynomial $\widetilde{L}_2(x - x_2)$ in cell C_2 be determined by solving

$$\frac{1}{|C_j|}\int_{C_j} \widetilde{L}_2(x - x_2)dx = \overline{L}_j, \quad j = 1, 2. \tag{4.1}$$

We now obtain the slope of $\widetilde{L}_2(x - x_2)$, which is only a candidate for the new value of $U_2''(0)$. The set of cells $\{C_1, C_2\}$ chosen by C_j in (4.1) is called a stencil for cell C_2. We can similarly determine another candidate for the new value of $U_2''(0)$ by solving (4.1) with C_j chosen from another stencil $\{C_2, C_3\}$ of cells. Finally we let

$$\widetilde{U}_2''(0) = \text{minmod}\big(\text{ candidates of } U_2''(0)\ \big),$$

where

$$\text{minmod}\{c_1, c_2, \ldots, c_m\} = \begin{cases} \min\{c_1, c_2, \ldots, c_m\} & \text{if } c_1, c_2, \ldots, c_m > 0, \\ \max\{c_1, c_2, \ldots, c_m\} & \text{if } c_1, c_2, \ldots, c_m < 0, \\ 0, & \text{otherwise.} \end{cases}$$

TABLE 8
P^2 version of the central DG scheme (3.4) *with the hierarchical reconstruction Algorithm* 1 *for the Burgers equation. MUSCL is used in Algorithm* 1.

Δx	1/10	1/20	1/40	1/80	1/160
L_1 *error*	4.24E-4	5.33E-5	6.71E-6	8.44E-7	1.07E-7
order	-	2.99	2.99	2.99	2.98
L_∞ *error*	5.13E-4	6.20E-5	7.38E-6	1.29E-6	2.66E-7
order	-	3.05	3.07	2.52	2.28

TABLE 9
P^2 version of the central DG scheme (3.4) *with the hierarchical reconstruction Algorithm* 1 *for the Burgers equation. Second order ENO is used in Algorithm* 1.

Δx	1/10	1/20	1/40	1/80	1/160
L_1 *error*	4.51E-4	5.36E-5	6.85E-6	8.54E-7	1.08E-7
order	-	3.07	2.97	3.00	2.98
L_∞ *error*	5.24E-4	6.17E-5	1.03E-5	1.81E-6	3.27E-7
order	-	3.09	2.58	2.51	2.47

This is a MUSCL reconstruction. To use the second order ENO reconstruction, we replace the minmod function by the following minmod_2 function:

$$\text{minmod}_2\{c_1, c_2, \ldots, c_m\} = c_j \quad \text{if } |c_j| = \min\{|c_1|, |c_2|, \ldots, |c_m|\}.$$

In order to find the new value $\widetilde{U}_2'(0)$ for $U_2'(0)$ by using the above MUSCL or second order ENO reconstruction, we need to find the cell averages of the linear part $U_2(0) + U_2'(0)(x - x_2)$ on cell C_2 and its neighbors C_1 and C_3.

Step 2. (1) Compute the cell average of $U_j(x - x_j)$ on cell C_j to obtain $\overline{U_j}$, $j = 1, 2, 3$.

(2) Compute the cell average of the polynomial $\widetilde{R}_2(x - x_2) = \frac{1}{2}\widetilde{U}_2''(0)(x - x_2)^2$ on cell C_j to obtain $\overline{R}_j$, $j = 1, 2, 3$.

(3) Redefine $\overline{L}_j = \overline{U_j} - \overline{R}_j$, $j = 1, 2, 3$. *These are the approximate cell averages of the linear part $U_2(0) + U_2'(0)(x - x_2)$ on cells C_1, C_2, and C_3.*

(4) Similar to Step 1, we solve (4.1) to obtain a linear polynomial in cell C_2. The slope of this linear polynomial corresponds to the slope $U_2'(0)$ of the linear polynomial $U_2(0)+U_2'(0)(x-x_2)$ and is only a candidate for the new value of $U_2'(0)$. Another candidate can be obtained by solving (4.1) with C_j chosen from another stencil $\{C_2, C_3\}$ of cells. Finally let

$$\widetilde{U}_2'(0) = \text{minmod}\big(\text{ candidates of } U_2'(0)\ \big).$$

For the second order ENO reconstruction, the minmod function can be replaced by the minmod_2 function. To keep the cell average invariant, we let the new value for $U_2(0)$ be $\widetilde{U}_2(0) = \overline{L}_2$.

The convergence test results with Algorithm 1 for Example 2 can be found in Tables 8 and 9. We observe that the order of accuracy is maintained, although (as expected for any limiter) the magnitude of the error is increased for the same mesh (see Table 5 for a comparison).

4.2. Hierarchical reconstruction—General description. In the following, we discuss a hierarchical reconstruction procedure for recomputing the polynomial $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ by using polynomials in cells adjacent to cell $C_{I+1/2}$. Recall that these adjacent cells are renamed as the set $\{C_J\}$ and the polynomials (of degree r)

supported on them are thus renamed as $\{U_J(\mathbf{x}-\mathbf{x}_J)\}$, respectively. The following hierarchical reconstruction describes a procedure to compute the new coefficients,

$$\frac{1}{\mathbf{m}!}\widetilde{U}^{(\mathbf{m})}_{I+1/2}(\mathbf{0}), \qquad m=r,r-1,\dots,0,$$

in $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$, iterating from the highest to the lowest degree terms.

To reconstruct $\widetilde{U}^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$, we first compute many *candidates* of $U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$ (sometimes still denoted as $\widetilde{U}^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$ with specification), and we then let the new coefficient for $U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$ be

$$\widetilde{U}^{(\mathbf{m})}_{I+1/2}(\mathbf{0}) = F\big(\text{ candidates of } U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})\ \big),$$

where F is a convex limiter of its arguments, e.g., the minmod function.

In order to find these candidates of $U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$, $|\mathbf{m}|=m$, we take an $(m-1)$th order partial derivative of $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ and denote it by

$$\partial^{m-1}U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2}) = L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2}) + R_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2}),$$

where $L_{I+1/2}$ is the linear part and $R_{I+1/2}$ contains all second and higher degree terms of $\partial^{m-1}U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$. Clearly, every coefficient in the first degree terms of $L_{I+1/2}$ is in the set $\{U^{(\mathbf{m})}_{I+1/2}(\mathbf{0}) : |\mathbf{m}|=m\}$. And for every $\mathbf{m}$ subject to $|\mathbf{m}|=m$, one can always take some $(m-1)$th order partial derivatives of $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ so that $U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$ is a coefficient in the first degree terms of $L_{I+1/2}$. Thus, a "candidate" for a coefficient in the first degree terms of $L_{I+1/2}$ is the candidate for the corresponding $U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$.

In order to find the candidates for all the coefficients in the first degree terms of $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$, we only need to know the cell averages of $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ on $d+1$ distinct mesh cells adjacent to cell $C_{I+1/2}$, recalling that d is the spatial dimension. Assume $C_{J_0}, C_{J_1}, \dots, C_{J_d} \in \{C_J\}$ are these cells and $\overline{L}_{J_0}, \overline{L}_{J_1}, \dots, \overline{L}_{J_d}$ are the corresponding cell averages. The set of these $d+1$ cells with the associated cell averages is called a *stencil*. Let a linear polynomial $\widetilde{L}_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ be determined by

$$\frac{1}{|C_{J_l}|}\int_{C_{J_l}} \widetilde{L}_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})d\mathbf{x} = \overline{L}_{J_l}, \quad l=0,1,\dots,d. \tag{4.2}$$

Then the coefficients in the first degree terms of $\widetilde{L}_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ will be the candidates for the corresponding coefficients of $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$. Therefore, a stencil located near cell $C_{I+1/2}$ will determine a set of candidates for all coefficients in the first degree terms of $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$. The key is to determine the new approximate cell averages of $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ on the cells of $\{C_J\}$, which is outlined by the following algorithm.

ALGORITHM 1.

Step 1. *Suppose* $r \geq 2$. *For* $m=r,r-1,\dots,2$, *do the following:*

(a) *Take an* $(m-1)$*th order partial derivative for each of* $\{U_J(\mathbf{x}-\mathbf{x}_J)\}$ *to obtain polynomials* $\{\partial^{m-1}U_J(\mathbf{x}-\mathbf{x}_J)\}$, *respectively. In particular, denote* $\partial^{m-1}U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2}) = L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2}) + R_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$, *where* $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *is the linear part of* $\partial^{m-1}U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *and* $R_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *is the remainder.*

(b) *Calculate the cell averages of* $\{\partial^{m-1}U_J(\mathbf{x}-\mathbf{x}_J)\}$ *on cells* $\{C_J\}$ *to obtain* $\{\overline{\partial^{m-1}U_J}\}$, *respectively.*

(c) *Let* $\widetilde{R}_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *be the* $R_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *with its coefficients replaced by the corresponding new coefficients.*[3] *Calculate the cell averages of* $\widetilde{R}_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *on cells* $\{C_J\}$ *to obtain* $\{\overline{R}_J\}$, *respectively.*

(d) *Let* $\overline{L}_J=\overline{\partial^{m-1}U_J}-\overline{R}_J$ *for all* J.

(e) *Form stencils out of the new approximate cell averages* $\{\overline{L}_J\}$ *by using a nonoscillatory finite volume MUSCL or second order ENO strategy. Each stencil will determine a set of candidates for the coefficients in the first degree terms of* $L_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$, *which are also candidates for the corresponding* $U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$*'s,* $|\mathbf{m}|=m$.

(f) *Repeat from* (a) *to* (e) *until all possible combinations of the* $(m-1)$*th order partial derivatives are taken. Then the candidates for all coefficients in the* m*th degree terms of* $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *have been computed. For each of these coefficients, say* $\frac{1}{\mathbf{m}!}U^{(\mathbf{m})}_{I+1/2}(\mathbf{0})$, $|\mathbf{m}|=m$, *let the new coefficient* $\widetilde{U}^{(\mathbf{m})}_{I+1/2}(\mathbf{0})=F(\text{candidates of}\,U^{(\mathbf{m})}_{I+1/2}(\mathbf{0}))$.

Step 2. *In order to find the new coefficients in the zeroth and first degree terms of* $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$, *we perform the procedure of Step* 1(a)–(f) *with* $m=1$, *and make sure that the new approximate cell average* $\overline{L}_{I+1/2}$ *is in each of the stencils, which ensures that the cell average of* $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *on cell* $C_{I+1/2}$ *is not changed with the new coefficients. The new coefficient in the zeroth degree term of* $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *is* $\overline{L}_{I+1/2}$, *which ensures that the cell average of* $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *in cell* $C_{I+1/2}$ *is invariant with the new coefficients. At this stage all new coefficients of* $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ *have been found.*

Remarks. 1. The coefficients of the polynomials can be updated after Algorithm 1 has been applied to all mesh cells, or at the mth stage when all new coefficients for those in the mth degree terms of all polynomials have been computed (in this case, $\{\overline{\partial^0 U_J}\}$ used in Step 2 should be the cell averages of the original polynomials to ensure that they are invariant). The latter case is supposed to be more diffusive. In numerical experiments we find their results are about the same. All numerical results presented in this paper are performed with the former implementation.

2. One motivation for us to develop this hierarchical reconstruction is that the limiting for the DG scheme on nonstaggered meshes is different from that for scheme (3.4). For the usual DG scheme the time evolution of the cell averages is completely determined by the fluxes; however, in (3.4), cell interior values are also involved. We find in numerical experiments that the moment limiter [7] does not work as robustly for scheme (3.4) as it does for the DG scheme on nonstaggered meshes. The proposed hierarchical reconstruction process is quite general and could be useful for conventional DG or even finite volume schemes. These will be explored in the future.

3. Scheme (3.4) with Algorithm 1 and with piecewise linear elements is identical to the second order central scheme on overlapping cells [38].

4. It is more efficient to apply the hierarchical reconstruction process only in places where it is needed by using nonsmoothness detectors (see, e.g., [42, 9]). This will be explored in the future.

[3] At this stage, we have already found new values for all coefficients in the terms of $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ of degree above m. These coefficients remain in $R_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ (after taking an $(m-1)$th order partial of $U_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$). When they are replaced by their corresponding new values, $R_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$ becomes $\widetilde{R}_{I+1/2}(\mathbf{x}-\mathbf{x}_{I+1/2})$. See Step 2(2) in section 4.1 as an example.

The most important point is that even though the linear reconstruction used in Algorithm 1 is only second order accurate, the approximation order of accuracy of a polynomial in a cell is unaffected by the algorithm, and we have the following condition.

CONDITION 1. *Let* $\{\mathbf{x}_{J_0}, \mathbf{x}_{J_1}, \ldots, \mathbf{x}_{J_d}\}$ *be the* $d+1$ *cell centroids of a stencil. Then there is a point among them, say* $\mathbf{x}_{J_0}$*, such that the matrix* $A = \frac{1}{\Delta x}[\mathbf{x}_{J_1} - \mathbf{x}_{J_0}, \mathbf{x}_{J_2} - \mathbf{x}_{J_0}, \ldots, \mathbf{x}_{J_d} - \mathbf{x}_{J_0}]$ *is nonsingular. Further, there is a constant* $\alpha > 0$ *independent of* Δx *such that* $||A^{-1}|| \le \alpha$.

In two dimensions, this condition means that $\mathbf{x}_{J_0}, \mathbf{x}_{J_1}, \mathbf{x}_{J_2}$ are not along a straight line. Further, the angle between the line passing $\mathbf{x}_{J_0}, \mathbf{x}_{J_1}$ and the line passing $\mathbf{x}_{J_0}, \mathbf{x}_{J_2}$ has a positive lower bound independent of Δx. This condition is satisfied for stencils such as $\{C_3, C_1, C_2\}$, $\{C_3, C_2, C_5\}$, $\{C_3, C_5, C_4\}$, and $\{C_3, C_4, C_1\}$ in Figure 3 (right), and is not satisfied for $\{C_1, C_3, C_5\}$.

THEOREM 2. *Suppose* $\{U_J(\mathbf{x} - \mathbf{x}_J)\}$ *in Algorithm* 1 *approximate a* C^{r+1} *function* $u(\mathbf{x})$ *with pointwise error* $\mathcal{O}\big((\Delta x)^{r+1}\big)$ *within their respective cell* $\{C_J\}$*, and all cells in* $\{C_J\}$ *are contained in a circle centered at* $\mathbf{x}_{I+1/2}$ *with radius* $\mathcal{O}(\Delta x)$*. Let the* $d+1$ *cell centroids in every stencil used in Algorithm* 1 *satisfy Condition* 1*. Then after the application of Algorithm* 1*, the polynomial* $\widetilde{U}_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$*, i.e.,* $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ *with its coefficients replaced by the corresponding new values, also approximates the function* $u(\mathbf{x})$ *with pointwise error* $\mathcal{O}\big((\Delta x)^{r+1}\big)$ *within cell* $C_{I+1/2}$*. The cell average of* $\widetilde{U}_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ *on cell* $C_{I+1/2}$ *is the same as that of* $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$.

Proof. From the assumption we know that the coefficients in the mth degree terms of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$, $0 \le m \le r$, are the $(r-m+1)$th order approximation to the corresponding coefficients of the Taylor expansion of $u(\mathbf{x})$ at $\mathbf{x}_{I+1/2}$.

Assume that when starting to compute new values for the coefficients of the mth degree terms of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$, $1 \le m \le r$, all the computed new values (if there are any) for the coefficients of the lth degree terms ($m < l \le r$, if they exist) of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ are their $(r-l+1)$th order approximations. In fact, when $m = r$, there are no new coefficients which have been computed at Step 1(a). However, the following argument will show that the new coefficients computed at Step 1(f) for coefficients of the rth degree terms of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ are their first order approximations.

Let $L_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2}) = c_0 + \mathbf{c}_1 \cdot (\mathbf{x} - \mathbf{x}_{I+1/2})$ in Step 1(a) and let $\widehat{L}(\mathbf{x}) = \widehat{c}_0 + \widehat{\mathbf{c}}_1 \cdot (\mathbf{x} - \mathbf{x}_{I+1/2})$ be the corresponding linear part in the Taylor expansion of the same (as for U_J) $(m-1)$th partial derivative of $u(x)$ at $\mathbf{x}_{I+1/2}$. Therefore c_0 and $\mathbf{c}_1$ approximate $\widehat{c}_0$ and $\widehat{\mathbf{c}}_1$ to the order of $\mathcal{O}\big((\Delta x)^{r-m+2}\big)$ and $\mathcal{O}\big((\Delta x)^{r-m+1}\big)$, respectively. Also from the above assumptions it is easy to see that $\overline{L}_J = \overline{\partial^{m-1} U_J} - \overline{R}_J$ in Step 1(d) approximates the cell average of $\widehat{L}(\mathbf{x})$ on cell C_J to the order of $O(\Delta x^{r-m+2})$ for all cells C_J adjacent to cell $C_{I+1/2}$.

Reconstructing $\widetilde{L}_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2}) = \widetilde{c}_0 + \widetilde{\mathbf{c}}_1 \cdot (\mathbf{x} - \mathbf{x}_{I+1/2})$ from a stencil $C_{J_0}, C_{J_1}, \ldots, C_{J_d} \in \{C_J\}$ entails finding $\widetilde{c}_0$ and $\widetilde{\mathbf{c}}_1$ that satisfy the equations (see (4.2))

$$\begin{aligned} \frac{1}{|C_{J_l}|} \int_{C_{J_l}} (\widetilde{c}_0 + \widetilde{\mathbf{c}}_1 \cdot (\mathbf{x} - \mathbf{x}_{I+1/2})) d\mathbf{x} &= \widetilde{c}_0 + \widetilde{\mathbf{c}}_1 \cdot (\mathbf{x}_{J_l} - \mathbf{x}_{I+1/2}) \\ &= \overline{L}_{J_l} = \widehat{c}_0 + \widehat{\mathbf{c}}_1 \cdot (\mathbf{x}_{J_l} - \mathbf{x}_{I+1/2}) \\ &\quad + \mathcal{O}\big((\Delta x)^{r-m+2}\big), \end{aligned} \tag{4.3}$$

where $\mathbf{x}_{J_l}$ is the cell centroid of cell C_{J_l}, $l = 0, \ldots, d$. The solutions are candidates for c_0 and $\mathbf{c}_1$, respectively. Subtracting the first equation ($l = 0$) from the rest of the

equations in (4.3), we can obtain

$$A^T(\widetilde{\mathbf{c}}_1 - \widehat{\mathbf{c}}_1) = \mathcal{O}\big((\Delta x)^{r-m+1}\big),$$

where $A = \frac{1}{\Delta x}[\mathbf{x}_{J_1} - \mathbf{x}_{J_0}, \mathbf{x}_{J_2} - \mathbf{x}_{J_0}, \ldots, \mathbf{x}_{J_d} - \mathbf{x}_{J_0}]$. From Condition 1, $||A^{-1}||$ is bounded independently of Δx. We conclude that the candidate

$$\widetilde{\mathbf{c}}_1 = \widehat{\mathbf{c}}_1 + \mathcal{O}\big((\Delta x)^{r-m+1}\big). \tag{4.4}$$

Also since $||\mathbf{x}_{J_l} - \mathbf{x}_{I+1/2}|| = \mathcal{O}(\Delta x)$, $l = 0, 1, \ldots, d$, by substituting the estimate of the candidate $\widetilde{\mathbf{c}}_1$ back into one of the equations of (4.3), we obtain that the candidate

$$\widetilde{c}_0 = \widehat{c}_0 + \mathcal{O}\big((\Delta x)^{r-m+2}\big). \tag{4.5}$$

Since the function F used in Step 1(f) is a convex combination of its arguments, it does not change the approximation order of its arguments. Therefore estimate (4.4) implies that the new values for coefficients of the mth degree terms of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ are their $(r-m+1)$th order approximations. Estimate (4.4) moves the induction till $m = 1$ and estimate (4.5) implies that in Step 2 the new value for the coefficient of the zeroth degree term of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ is its $O(\Delta x^{r+1})$ approximation. Step 2 clearly ensures that the cell average of $U_{I+1/2}(\mathbf{x} - \mathbf{x}_{I+1/2})$ on cell $C_{I+1/2}$ is unchanged with the new coefficients. The proof is now complete.

4.3. Implementation for piecewise quadratic finite element space in two dimensions. Suppose on cell C_j (see Figure 3, right) a quadratic polynomial is given as

$$\begin{aligned} U_j(x - x_j, y - y_j) = {} & U_j(0,0) + \partial_x U_j(0,0)(x - x_j) + \partial_y U_j(0,0)(y - y_j) \\ & + \frac{1}{2}\partial_{xx} U_j(0,0)(x - x_j)^2 \\ & + \partial_{xy} U_j(0,0)(x - x_j)(y - y_j) + \frac{1}{2}\partial_{yy} U_j(0,0)(y - y_j)^2, \end{aligned}$$

where (x_j, y_j) is the cell centroid of cell C_j, $j = 1, 2, \ldots, 5$.

According to Step 1 of Algorithm 1, take the first partial derivative with respect to x for them to obtain $L_j(x - x_j,\ y - y_j) = \partial_x U_j(0,0) + \partial_{xx} U_j(0,0)(x - x_j) + \partial_{xy} U_j(0,0)(y - y_j)$, $j = 1, 2, \ldots, 5$. Calculate the cell average of $L_j(x - x_j,\ y - y_j)$ on cell C_j to obtain $\overline{L}_j = \partial_x U_j(0,0)$, $j = 1, 2, \ldots, 5$ (note that $R_3(x - x_3,\ y - y_3) \equiv 0$). With the five new approximate cell averages $\{\overline{L}_j : j = 1, 2, \ldots, 5\}$, one can apply a MUSCL or a second order ENO procedure to reconstruct a nonoscillatory linear polynomial

$$\widetilde{L}_3(x - x_3, y - y_3) = \partial_x \widetilde{U}_3(0,0) + \partial_{xx} \widetilde{U}_3(0,0)(x - x_3) + \partial_{xy} \widetilde{U}_3(0,0)(y - y_3)$$

in cell C_3. For example, one can form the four stencils $\{C_3, C_1, C_2\}$, $\{C_3, C_2, C_5\}$, $\{C_3, C_5, C_4\}$, and $\{C_3, C_4, C_1\}$. For the first stencil, solve the following equations for $\partial_{xx} \widetilde{U}_3(0,0)$ and $\partial_{xy} \widetilde{U}_3(0,0)$:

$$\begin{aligned} \frac{1}{|C_j|}\int_{C_j} \widetilde{L}_3(x - x_3, y - y_3)\,dx\,dy &= \overline{L}_3 + \partial_{xx} \widetilde{U}_3(0,0)(x_j - x_3) + \partial_{xy} \widetilde{U}_3(0,0)(y_j - y_3) \\ &= \overline{L}_j, \end{aligned}$$

TABLE 10
P^2 version of the central DG scheme (3.4) with the hierarchical reconstruction Algorithm 1 for the 2D Burgers equation. Second order ENO is used in Algorithm 1.

Δx	1/4	1/8	1/16	1/32	1/64
L_1 *error*	8.00E-2	1.24E-2	1.58E-3	1.92E-4	2.40E-5
order	-	2.69	2.97	3.04	3.00
L_∞ *error*	4.90E-2	9.85E-3	1.68E-3	2.01E-4	2.68E-5
order	-	2.31	2.55	3.06	2.91

$j = 1, 2$; similarly for other stencils. We obtain two sets of candidates for $\partial_{xx}U_3(0,0)$ and $\partial_{xy}U_3(0,0)$, respectively. By taking the first partial derivative with respect to y for $U_j(x - x_j,\ y - y_j)$, $j = 1, 2, \ldots, 5$, we similarly obtain a set of candidates for $\partial_{yy}U_3(0,0)$ and enlarge the set of candidates for $\partial_{xy}U_3(0,0)$. Putting all candidates for $\partial_{xx}U_3(0,0)$ into the arguments of the minmod (or minmod_2) function, we obtain the new coefficient $\partial_{xx}\widetilde{U}_3(0,0)$ for $\partial_{xx}U_3(0,0)$. Applying the same procedure, we obtain new coefficients $\partial_{xy}\widetilde{U}_3(0,0)$ and $\partial_{yy}\widetilde{U}_3(0,0)$.

According to Step 2 of Algorithm 1, we compute the cell average of $U_j(x - x_j,\ y - y_j)$ on cell C_j to obtain $\overline{U_j}$, $j = 1, 2, \ldots, 5$, and compute cell averages of the polynomial

$$\widetilde{R}_3(x - x_3, y - y_3) = \frac{1}{2}\partial_{xx}\widetilde{U}_3(0,0)(x - x_3)^2 + \partial_{xy}\widetilde{U}_3(0,0)(x - x_3)(y - y_3) + \frac{1}{2}\partial_{yy}\widetilde{U}_3(0,0)(y - y_3)^2$$

on cell $C_1, C_2, \ldots, C_5$ to obtain $\overline{R}_1, \overline{R}_2, \ldots, \overline{R}_5$, respectively. Redefine $\overline{L}_j = \overline{U_j} - \overline{R}_j$, $j = 1, 2, \ldots, 5$. The same MUSCL or second order ENO procedure as described previously can be applied to the five cell averages $\{\overline{L}_j : j = 1, 2, \ldots, 5\}$ to obtain the new coefficients $\partial_x\widetilde{U}_3(0,0)$ and $\partial_y\widetilde{U}_3(0,0)$. Finally let the new coefficient $\widetilde{U}_3(0,0) = \overline{L}_3$.

The convergence test results with Algorithm 1 for Example 3 can be found in Table 10. We again observe that the order of accuracy is maintained, although (as expected for any limiter) the magnitude of the error is increased for the same mesh (see Table 6 for a comparison).

5. Additional numerical examples. Scheme (3.4) with the piecewise rth degree polynomial space is referred to as CO-DG-$(r+1)$, where "C" stands for "central" and "O" stands for "overlapping cells." When the hierarchical reconstruction Algorithm 1 is applied, it is referred to as CO-DG-hr1-$(r+1)$. To specify whether a linear MUSCL (with the minmod limiter) or ENO (with the minmod_2 limiter) reconstruction is used in Algorithm 1, we refer it as CO-DG-hr1m-$(r+1)$ or CO-DG-hr1e-$(r+1)$, respectively.

The corresponding (up to third order) TVD Runge–Kutta time discretization methods [45] are applied to the above schemes. Only the solution in one class of the overlapping cells is shown in the graphs throughout this section. For systems of equations, the componentwise extensions of the scalar schemes (without characteristic decomposition) have been used in all the computations.

Example 5. We compute the Euler equation with Lax's initial data. $u_t + f(u)_x = 0$ with $u = (\rho, \rho v, E)^T$, $f(u) = (\rho v, \rho v^2 + p, v(E + p))^T$, $p = (\gamma - 1)(E - \frac{1}{2}\rho v^2)$, $\gamma = 1.4$. Initially, the density ρ, momentum ρv, and total energy E are 0.445, 0.311, and 8.928 in $(0, 0.5)$, and are 0.5, 0, and 1.4275 in $(0.5, 1)$. The computed results by CO-DG-hr1e-3 and CO-DG-hr1m-3 are shown at $T = 0.16$ in Figure 4, with $\Delta x = 1/200$,

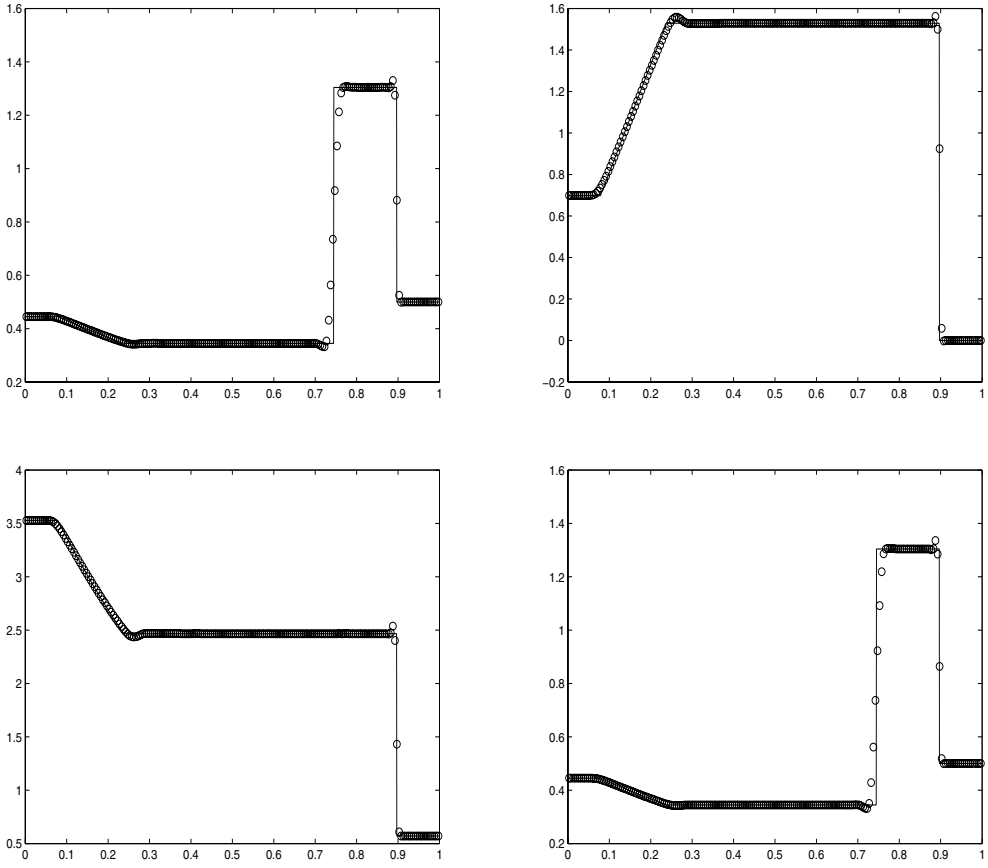


FIG. 4. *Lax's problem,* $\Delta x = 1/200$. *From left to right, top to bottom,* (1) *density (CO-DG-hr*1*e*-3*);* (2) *velocity (CO-DG-hr*1*e*-3*);* (3) *pressure (CO-DG-hr*1*e*-3*);* (4) *density (CO-DG-hr*1*m*-3*).*

$\Delta\tau^n$ chosen with a CFL factor 0.4, $\Delta t^n = 0.5\Delta\tau^n$. The solid line reference solutions are analytic solutions to the Riemann problem. We observe that the resolution is quite good with very small over/undershoots. The only concern is that the contact discontinuity is much more smeared than that of the regular third order DG scheme with a total variation bounded limiter (Figure 20 in [12]). We hope to improve this performance by reducing the usage of the reconstruction limiter through a troubled-cell indicator in future work.

Example 6. The Woodward and Colella blast wave problem [50] for the Euler equation computed by CO-DG-hr1e-3. Initially, the density, momentum, and total energy are $1, 0, 2500$ in $(0, 0.1)$; $1, 0, 0.025$ in $(0.1, 0.9)$; and $1, 0, 250$ in $(0.9, 1)$. The density, velocity, and pressure profiles are plotted in Figure 5 for $T = 0.01$ and $T = 0.038$. The solid line reference solutions are computed by a third order central scheme on overlapping cells [38] on a much refined mesh ($\Delta x = 1/2000$). $\Delta\tau^n$ is chosen with a CFL factor 0.4, $\Delta t^n = \frac{1}{2}\Delta\tau^n$. We observe stable results with good resolution for this very demanding problem in terms of numerical stability.

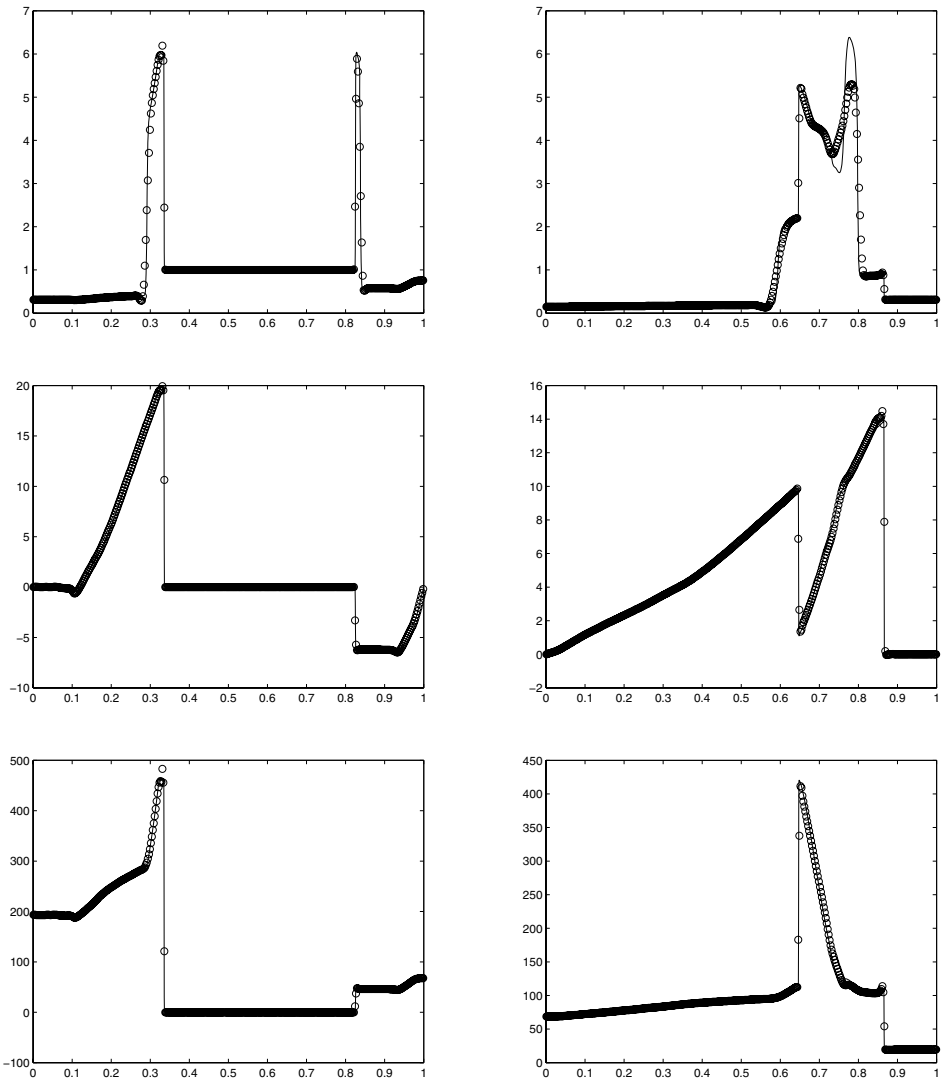


FIG. 5. *Woodward and Colella blast wave problem computed by CO-DG-hr*1*e*-3, $\Delta x = 1/400$. *Top: density; middle: velocity; bottom: pressure. Left:* $T = 0.01$. *Right:* $T = 0.038$.

Example 7. Shu–Osher problem [46]. It is the Euler equation with an initial data

$$(\rho, v, p) = (3.857143,\ 2.629369,\ 10.333333) \quad \text{for } x < -4,$$
$$(\rho, v, p) = (1 + 0.2\sin(5x),\ 0,\ 1) \quad \text{for } x \geq -4.$$

The density profiles are plotted at $T = 1.8$, with $\Delta x = 1/40$; see Figure 6. $\Delta\tau^n$ is chosen with a CFL factor 0.5, $\Delta t^n = 0.5\Delta\tau^n$. The solid line is the numerical solution on a fine mesh ($\Delta x = 1/200$) computed by a central scheme on overlapping cells [38]. We observe very good resolution for this example. In order to see the resolution of the 2D nonoscillatory hierarchical reconstruction algorithm, we put the Shu–Osher problem to a 2D domain $[-5, 5] \times [0, 0.25]$ and solve the 2D Euler equation. Initially

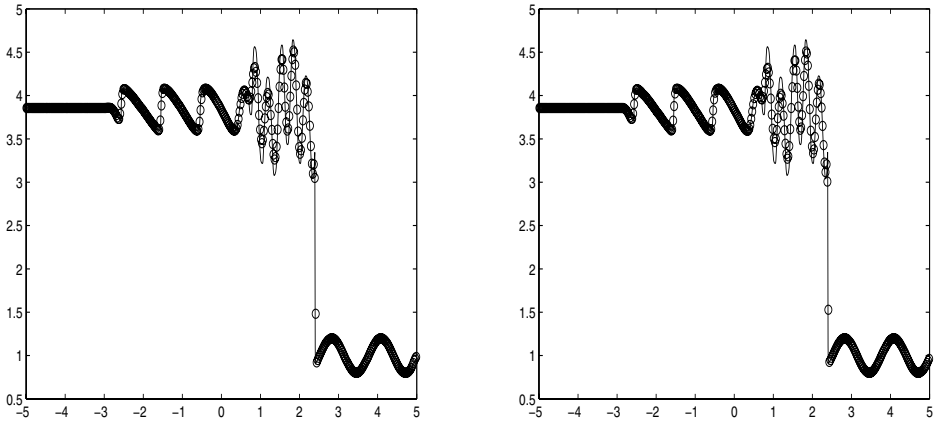


FIG. 6. *Shu–Osher problem,* $\Delta x = 1/40$. *Left: CO-DG-dr*1*m*-3. *Right: CO-DG-hr*1*e*-3.

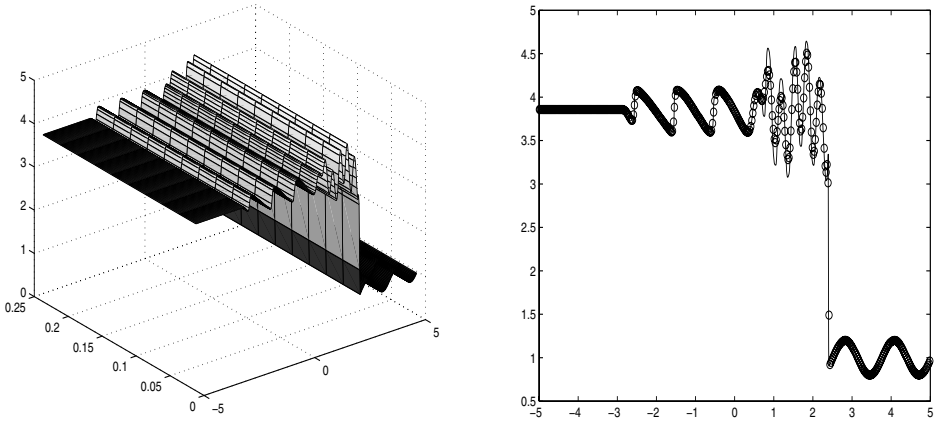


FIG. 7. *Shu–Osher problem in two dimensions,* $\Delta x = \Delta y = 1/40$. *CO-DG-hr*1*e*-3. *Left: density in the xy plane. Right: density along the line* $y = 0.25/3$.

the density variation is only along the x direction. The density profiles at $T = 1.8$ are plotted in Figure 7.

Example 8. 2D Riemann problem [29] for the Euler equation computed by CO-DG-hr1e-3. The 2D Euler equation can be written as

$$\mathbf{u}_t + \mathbf{f}(\mathbf{u})_x + \mathbf{g}(\mathbf{u})_y = 0, \quad \mathbf{u} = (\rho, \rho u, \rho v, E)^T, \quad p = (\gamma - 1)(E - \tfrac{1}{2}\rho(u^2 + v^2)),$$
$$\mathbf{f}(\mathbf{u}) = (\rho u, \rho u^2 + p, \rho uv, u(E + p))^T, \quad \mathbf{g}(\mathbf{u}) = (\rho v, \rho uv, \rho v^2 + p, v(E + p))^T,$$

where $\gamma = 1.4$. The computational domain is $[0, 1] \times [0, 1]$. The initial states are constants within each of the 4 quadrants. Counterclockwise from the upper right quadrant, they are labeled (ρ_i, u_i, v_i, p_i), $i = 1, 2, 3, 4$. Initially, $\rho_1 = 1.1$, $u_1 = 0$, $v_1 = 0$, $p_1 = 1.1$; $\rho_2 = 0.5065$, $u_2 = 0.8939$, $v_2 = 0$, $p_2 = 0.35$; $\rho_3 = 1.1$, $u_3 = 0.8939$, $v_3 = 0.8939$, $p_3 = 1.1$; and $\rho_4 = 0.5065$, $u_4 = 0$, $v_4 = 0.8939$, $p_4 = 0.35$. The density

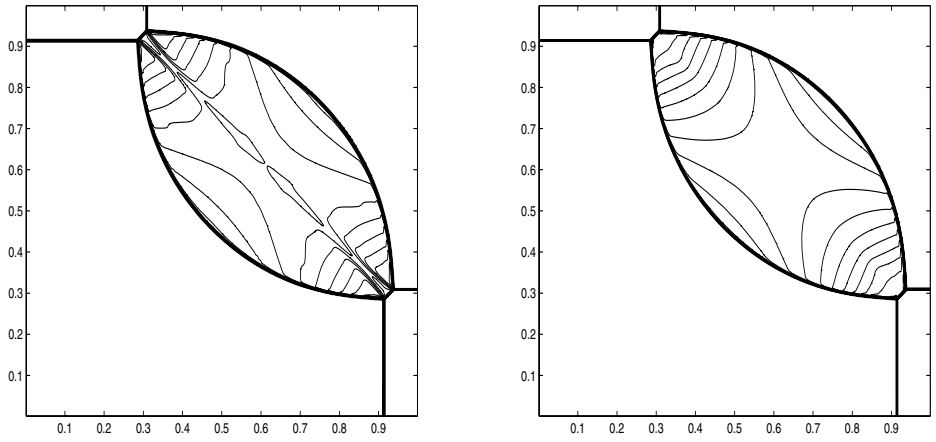


FIG. 8. *A 2D Riemann problem* [29] *computed by CO-DG-hr1e-3.* $\Delta x = \Delta y = 1/400$*, Left: density. Right: pressure.*

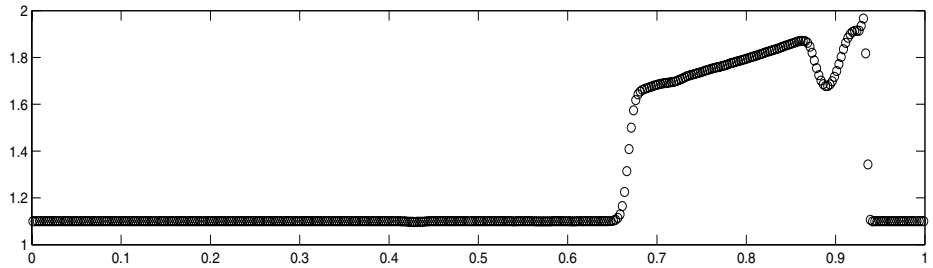


FIG. 9. *A 2D Riemann problem* [29]. *Density profile along* $y = 1/3$.

and pressure profiles are plotted at $T = 0.25$ in Figure 8, with 30 equally spaced contours. The numerical resolution is quite good for this problem. The density profile along $y = 1/3$ is plotted in Figure 9. There is no oscillation near the discontinuities.

Example 9. Double Mach reflection [50] computed by CO-DG-hr1e-3. A planar Mach 10 shock is incident on an oblique wedge at a $\pi/3$ angle. The air in front of the shock has density 1.4, pressure 1, and velocity 0. The boundary condition is described in [50]. The density and pressure profiles are plotted at $T = 0.2$ in Figure 10, with 30 equally spaced contours. $\Delta x = \Delta y = 1/120$, $\Delta\tau^n$ chosen with a CFL factor 0.4, $\Delta t^n = 0.99\Delta\tau^n$. We can see in the lower graph (the cross section density profile along $y = 1/3$) that the computed result is nonoscillatory.

6. Concluding remarks and a plan for future work. In this paper we have developed a central DG method based on staggered overlapping cells, with a numerical viscosity which stays bounded when the time step size goes to zero. Time discretization is via the standard TVD Runge–Kutta method. We have also developed a multilayer hierarchical reconstruction procedure and used it as a limiter for our central DG scheme. The limiter is able to maintain the original order of accuracy and can effectively control spurious oscillations for discontinuous solutions. In future

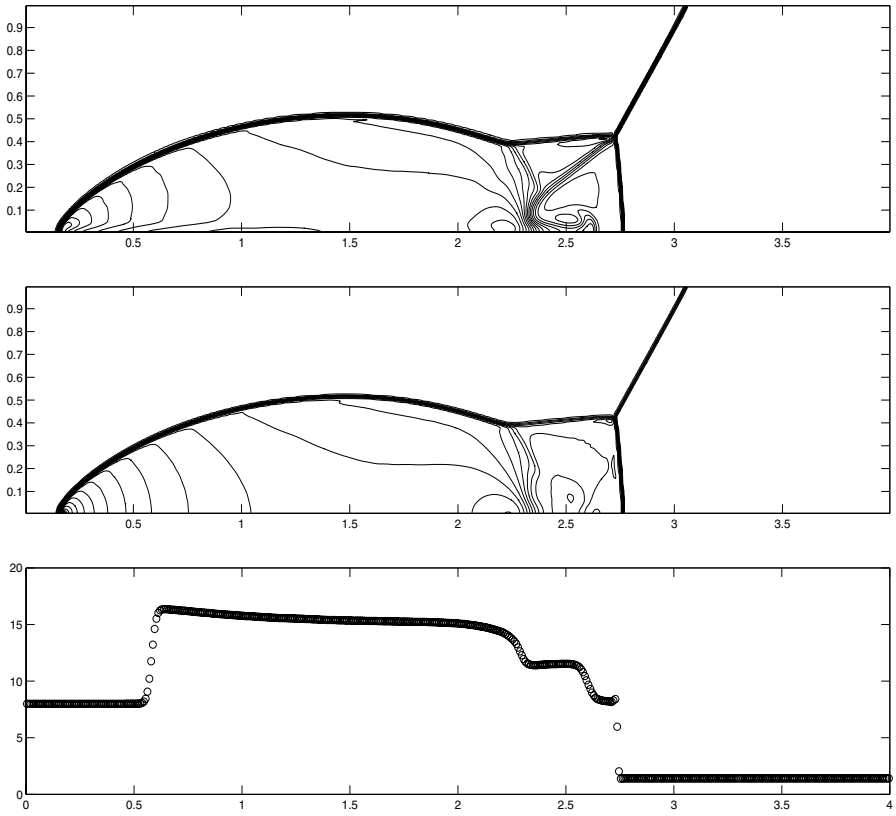


FIG. 10. *Double Mach reflection computed by CO-DG-hr*1*e*-3, $\Delta x = \Delta y = 1/120$. *Top: density contours. Middle: pressure contours. Bottom: density cut along the line* $y = 1/3$.

work we will generalize the method to convection-diffusion equations, improve the limiter by applying troubled-cell indicators, and also study further the hierarchical reconstruction procedure as a limiter for the regular DG methods and finite volume schemes. A stability analysis and error estimates for the central DG scheme as well as a comparison between the regular DG and central DG schemes will also be performed.

The examples reported in the paper are aimed to show the flexibility of the new approach to use with a Runge–Kutta method, and its capability to handle small time steps, without introducing excessive numerical dissipation. The more efficient way to overcome the small time step restriction with the presence of a diffusion term is to use implicit-explicit time discretization, e.g., Ascher, Ruuth, and Spiteri [4], Kennedy and Carpenter [24], and Liotta, Romano, and Russo [33], which treats the advection part explicitly and the diffusion part implicitly, thus avoiding the $O(\Delta x^2)$ stability restriction on the time step due to the diffusion term; another way would be to use a fast explicit Runge–Kutta solver, e.g., Lebedev [30] or Medovikov [39].

Even though in all the numerical examples the reconstruction is performed componentwisely, we have also performed some preliminary tests on the nonoscillatory hierarchical reconstruction with local characteristic decomposition and have not found any significant difference. We plan to conduct more careful study on this subject in the future.

REFERENCES

[1] R. Abgrall, *On essentially non-oscillatory schemes on unstructured meshes: Analysis and implementation*, J. Comput. Phys., 114 (1994), pp. 45–58.

[2] P. Arminjon and A. St-Cyr, *Nessyahu-Tadmor-type central finite volume methods without predictor for 3D Cartesian and unstructured tetrahedral grids*, Appl. Numer. Math., 46 (2003), pp. 135–155.

[3] P. Arminjon, M. C. Viallon, A. Madrane, and L. Kaddouri, *Discontinuous finite elements and 2-dimensional finite volume versions of the Lax-Friedrichs and Nessyahu-Tadmor difference schemes for compressible flows on unstructured grids*, in CFD Review, M. Hafez and K. Oshima, eds., John Wiley, New York, 1997, pp. 241–261.

[4] U. Ascher, S. Ruuth, and R. J. Spiteri, *Implicit-explicit Runge-Kutta methods for time dependent partial differential equations*, Appl. Numer. Math., 25 (1997), pp. 151–167.

[5] J. Balbas and E. Tadmor, *CentPack: A Package for High-Resolution Central Schemes for Nonlinear Conservation Laws and Related Problems*, http://www.cscamm.umd.edu/centpack.

[6] F. Bianco, G. Puppo, and G. Russo, *High-order central schemes for hyperbolic systems of conservation laws*, SIAM J. Sci. Comput., 21 (1999), pp. 294–322.

[7] R. Biswas, K. Devine, and J. Flaherty, *Parallel, adaptive finite element methods for conservation laws*, Appl. Numer. Math., 14 (1994), pp. 255–283.

[8] J. Boris and D. Book, *Flux corrected transport.* I. *SHASTA, a fluid transport algorithm that works*, J. Comput. Phys., 11 (1973), pp. 38–69.

[9] N. Chevaugeon, J. Xin, P. Hu, X. Li, D. Cler, J. Flaherty, and M. Shephard, *Discontinuous Galerkin methods applied to shock and blast problems*, J. Sci. Comput., 22/23 (2005), pp. 227–243.

[10] B. Cockburn, *An introduction to the discontinuous Galerkin method for convection-dominated problems*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations (Cetraro, 1997), Lecture Notes in Math. 1697, Springer, Berlin, 1998, pp. 151–268.

[11] B. Cockburn, S.-C. Hou, and C.-W. Shu, *The Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws*, Math. Comp., 54 (1990), pp. 545–581.

[12] B. Cockburn, S.-Y. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws.* III. *One dimensional systems*, J. Comput. Phys., 84 (1989), pp. 90–113.

[13] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws.* II. *General framework*, Math. Comp., 52 (1989), pp. 411–435.

[14] B. Cockburn and C.-W. Shu, *The Runge-Kutta local projection p^1-discontinuous-Galerkin finite element method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér., 25 (1991), pp. 337–361.

[15] B. Cockburn and C.-W. Shu, *The Runge-Kutta discontinuous Galerkin method for conservation laws* V: *Multidimensional systems*, J. Comput. Phys., 141 (1998), pp. 199–224.

[16] P. Colella and P. Woodward, *The piecewise parabolic method (PPM) for gas-dynamical simulation*, J. Comput. Phys., 54 (1984), pp. 174–201.

[17] J. Glimm, X. Li, Y. Liu, Z. Xu, and N. Zhao, *Conservative front tracking with improved accuracy*, SIAM J. Numer. Anal., 41 (2003), pp. 1926–1947.

[18] S. Gottlieb, C.-W. Shu, and E. Tadmor, *Strong stability-preserving high-order time discretization methods*, SIAM Rev., 43 (2001), pp. 89–112.

[19] A. Harten, *High resolution schemes for hyperbolic conservation laws*, J. Comput. Phys., 49 (1983), pp. 357–393.

[20] A. Harten, B. Engquist, S. Osher, and S. R. Chakravarthy, *Uniformly high order accuracy essentially non-oscillatory schemes* III, J. Comput. Phys., 71 (1987), pp. 231–303.

[21] G.-S. Jiang, D. Levy, C.-T. Lin, S. Osher, and E. Tadmor, *High-resolution nonoscillatory central schemes with nonstaggered grids for hyperbolic conservation laws*, SIAM J. Numer. Anal., 35 (1998), pp. 2147–2168.

[22] G.-S. Jiang and E. Tadmor, *Nonoscillatory central schemes for multidimensional hyperbolic conservation laws*, SIAM J. Sci. Comput., 19 (1998), pp. 1892–1917.

[23] S. Jin and Z. Xin, *The relaxation schemes for systems of conservation laws in arbitrary space dimensions*, Comm. Pure Appl. Math., 48 (1995), pp. 235–276.

[24] C. A. Kennedy and M. H. Carpenter, *Additive Runge-Kutta schemes for convection-diffusion-reaction equations*, Appl. Numer. Math., 44 (2003), pp. 139–181.

[25] A. Kurganov and D. Levy, *A third-order semidiscrete central scheme for conservation laws and convection-diffusion equations*, SIAM J. Sci. Comput., 22 (2000), pp. 1461–1488.

[26] A. Kurganov and C.-T. Lin, *On the reduction of numerical dissipation in central-upwind schemes*, Commun. Comput. Phys., 2 (2007), pp. 141–163.
[27] A. Kurganov, S. Noelle, and G. Petrova, *Semidiscrete central-upwind schemes for hyperbolic conservation laws and Hamilton–Jacobi equations*, SIAM J. Sci. Comput., 23 (2001), pp. 707–740.
[28] A. Kurganov and E. Tadmor, *New high-resolution central schemes for nonlinear conservation laws and convection-diffusion equations*, J. Comput. Phys., 160 (2000), pp. 241–282.
[29] P. D. Lax and X.-D. Liu, *Solution of two-dimensional Riemann problems of gas dynamics by positive schemes*, SIAM J. Sci. Comput., 19 (1998), pp. 319–340.
[30] V. I. Lebedev, *Explicit difference schemes for solving stiff systems of ODEs and PDEs with complex spectrum*, Russian J. Numer. Anal. Math. Modelling, 13 (1998), pp. 107–116.
[31] P. Lesaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, in Mathematical Aspects of Finite Elements in Partial Differential Equations, C. de Boor, ed., Academic Press, New York, 1974, pp. 89–123.
[32] D. Levy, G. Puppo, and G. Russo, *A fourth-order central WENO scheme for multidimensional hyperbolic systems of conservation laws*, SIAM J. Sci. Comput., 24 (2002), pp. 480–506.
[33] S. F. Liotta, V. Romano, and G. Russo, *Central schemes for balance laws of relaxation type*, SIAM J. Numer. Anal., 38 (2000), pp. 1337–1356.
[34] X. D. Liu and S. Osher, *Convex ENO high order multi-dimensional schemes without field by field decomposition or staggered grids*, J. Comput. Phys., 142 (1998), pp. 304–330.
[35] X. D. Liu, S. Osher, and T. Chan, *Weighted essentially non-oscillatory schemes*, J. Comput. Phys., 115 (1994), pp. 408–463.
[36] X.-D. Liu and E. Tadmor, *Third order nonoscillatory central scheme for hyperbolic conservation laws*, Numer. Math., 79 (1998), pp. 397–425.
[37] Y. Liu, *Central schemes and central discontinuous Galerkin methods on overlapping cells*, Conference on Analysis, Modeling and Computation of PDE and Multiphase Flow, Stony Brook, NY, 2004.
[38] Y. Liu, *Central schemes on overlapping cells*, J. Comput. Phys., 209 (2005), pp. 82–104.
[39] A. A. Medovikov, *High order explicit methods for parabolic equations*, BIT, 38 (1998), pp. 372–390.
[40] H. Nessyahu and E. Tadmor, *Non-oscillatory central differencing for hyperbolic conservation laws*, J. Comput. Phys., 87 (1990), pp. 408–463.
[41] J. Qiu and C.-W. Shu, *Hermite WENO schemes and their application as limiters for Runge-Kutta discontinuous Galerkin method.* II. *Two dimensional case*, Comput. & Fluids, 34 (2005), pp. 642–663.
[42] J. Qiu and C.-W. Shu, *Runge–Kutta discontinuous Galerkin method using WENO limiters*, SIAM J. Sci. Comput., 26 (2005), pp. 907–929.
[43] R. Sanders and A. Weiser, *High resolution staggered mesh approach for nonlinear hyperbolic systems of conservation laws*, J. Comput. Phys., 101 (1992), pp. 314–329.
[44] C.-W. Shu, *Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws*, in Advanced Numerical Approximation of Nonlinear Hyperbolic Equations, A. Quarteroni, ed., Lecture Notes in Math. 1697, Springer, Berlin, 1998, pp. 325–432.
[45] C.-W. Shu and S. Osher, *Efficient implementation of essentially non-oscillatory shock-capturing schemes*, J. Comput. Phys., 77 (1988), pp. 439–471.
[46] C.-W. Shu and S. Osher, *Efficient implementation of essentially nonoscillatory shock-capturing schemes,* II, J. Comput. Phys., 83 (1989), pp. 32–78.
[47] H. Tang and T. Tang, *Adaptive mesh methods for one- and two-dimensional hyperbolic conservation laws*, SIAM J. Numer. Anal., 41 (2003), pp. 487–515.
[48] B. van Leer, *Towards the ultimate conservative difference scheme* I, in Proceedings of the Third International Conference on Numerical Methods in Fluid Mechanics, Vol. I, Lecture Notes in Phys. 18, Springer-Verlag, Berlin, 1973, pp. 163–168.
[49] B. van Leer, *Towards the ultimate conservative difference scheme* II. *Monotonicity and conservation combined in a second order scheme*, J. Comput. Phys., 14 (1974), pp. 361–370.
[50] P. Woodward and P. Colella, *The numerical simulation of two-dimensional fluid flow with strong shocks*, J. Comput. Phys., 54 (1984), pp. 115–173.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2468–2482

MONOTONIC PARAREAL CONTROL FOR QUANTUM SYSTEMS*

YVON MADAY†, JULIEN SALOMON‡, AND GABRIEL TURINICI§

Abstract. Following encouraging experimental results in quantum control, numerical simulations have known significant improvements through the introduction of efficient optimization algorithms. Yet, the computational cost still prevents using these procedures for high-dimensional systems often present in quantum chemistry. Using parareal framework, we present here a time parallelization of these schemes which allows us to reduce significantly their computational cost while still finding convenient controls.

Key words. quantum control, monotonic schemes, optimal control, parareal scheme, time parallelization

AMS subject classifications. 49J20, 68W10

DOI. 10.1137/050647086

1. Introduction.

In the last decade, quantum control has witnessed significant developments including encouraging experimental results [5, 6, 9, 15, 16, 23, 24, 29, 36]. At the computational level [7, 25], the introduction of the monotonic algorithms of Krotov (by Tannor [31]), Zhu and Rabitz [37], or the unified form of Maday and Turinici [21] allows us to design efficient methods to obtain laser fields controlling the molecular dynamics. On the other hand, parareal scheme (that stands for parallelization in real time) has shown a convenient efficiency in the case of the Schrödinger equation; see, e.g., [2, 33]. In what follows, we combine these two approaches by using monotonic algorithms as the inner loop of a time-parallelization procedure.

Let us first briefly present the model and the corresponding optimal control framework used in this paper. Consider a quantum system described by its wavefunction $\psi(x,t)$, also called *state* in what follows. Here "x" $\in \Omega$ denotes the relevant spatial coordinates (the symbol x will often be omitted in what follows for reason of simplicity). The operator $V(x)$ is the potential part. The dynamics of this system is characterized by its internal Hamiltonian:

$$H(x) = H_0 + V(x).$$

In this equation H_0, the kinetic part, could be

$$H_0 = -\frac{1}{2}\sum_{n=1}^{p}\frac{1}{m_n}\Delta_{r_n},$$

where p is the number of particles considered, m_n their masses, and Δ_{r_n} the Laplace operator with respect to their coordinates.

*Received by the editors December 8, 2005; accepted for publication (in revised form) April 15, 2007; published electronically November 28, 2007.
http://www.siam.org/journals/sinum/45-6/64708.html

†Laboratoire J.-L. Lions, Université P. and M. Curie, 76252 Paris Cedex 05, France (maday@ann.jussieu.fr) and Division of Applied Mathematics, Brown University, Providence, RI 02912.

‡Laboratoire J.-L. Lions, Université P. and M. Curie, 76252 Paris Cedex 05, France (salomon@ann.jussieu.fr) and CEREMADE, Université Paris Dauphine, Pl. du Maréchal Lattre de Tassigny, 75775 Paris Cedex 16, France (salomon@ceremade.dauphine.fr).

§CEREMADE, Université Paris Dauphine, Pl. du Maréchal Lattre de Tassigny, 75775 Paris Cedex 16, France (gabriel.turinici@dauphine.fr).

A way to control such a system is to light it with a laser pulse. We denote by $\varepsilon(t)$ the intensity of this control field. The contribution of this parameter is taken into account by introducing a perturbative term in the Hamiltonian which then reads $H(x) - \mu(x)\varepsilon(t)$. The evolution of $\psi^\varepsilon(x,t)$ is governed by the Schrödinger equation (we work in atomic units, i.e., $\hbar = 1$):

$$\left\{ \begin{array}{rcl} i\frac{\partial}{\partial t}\psi^\varepsilon(x,t) & = & (H(x) - \mu(x)\varepsilon(t))\psi^\varepsilon(x,t), \\ \psi^\varepsilon(x,0) & = & \psi_{init}(x), \end{array} \right. \tag{1.1}$$

where ψ_{init} is the initial condition for ψ^ε subject to the constraint:

$$\|\psi_{init}\|_{L^2(\Omega)} = 1.$$

Since H is self-adjoint, from (1.1) the norm of the state is constant with respect to the time. In the numerical simulations, the ground state, i.e., a unitary eigenvector of H associated with the lowest eigenvalue, is generally taken as initial condition.

The optimal control framework can then be applied to this bilinear control system to design relevant control fields. The quality of the pulse is evaluated via a cost functional. A general example of such a function is

$$J(\varepsilon) = \|\psi^\varepsilon(T) - \psi_{target}\|^2_{L^2(\Omega)} + \int_0^T \alpha(t)\varepsilon^2(t)dt, \tag{1.2}$$

where T is the total time of control, α a positive function, and ψ_{target} a target state which has to be reached.

At the minimum of the cost functional J, the Euler–Lagrange critical point equations are satisfied; a standard way to write these equations is to use a Lagrange multiplier $\chi^\varepsilon(x,t)$ called *adjoint state*. The following critical point equations are thus obtained [37]:

$$\left\{ \begin{array}{l} i\frac{\partial}{\partial t}\psi^\varepsilon(x,t) = (H(x) - \varepsilon(t)\mu(x))\psi^\varepsilon(x,t), \\ \psi^\varepsilon(x,0) = \psi_{init}(x), \end{array} \right. \tag{1.3}$$

$$\left\{ \begin{array}{l} i\frac{\partial}{\partial t}\chi^\varepsilon(x,t) = (H(x) - \varepsilon(t)\mu(x))\chi^\varepsilon(x,t), \\ \chi^\varepsilon(x,T) = \psi_{target}(x), \end{array} \right. \tag{1.4}$$

$$\alpha(t)\varepsilon(t) = -Im\langle\chi^\varepsilon(t)|\mu|\psi^\varepsilon(t)\rangle, \tag{1.5}$$

where A is an operator on $L^2(\Omega)$ and $\langle f|A|g\rangle = \int_\Omega \overline{f(x)}A(g)(x)dx$.

Numerous optimization procedures exist to compute iteratively sequences $(\varepsilon^k)_{k\in\mathbf{N}}$ that approximate the solution of (1.3)–(1.5). The common feature of these algorithms is that they involve repeated resolutions of Schrödinger equations (1.3) and (1.4), which induce a heavy computational time cost. Depending on their order, the mere computation of ε^k can also be time consuming due to the high nonlinearity of the cost functional. In order to reduce the computation time, some time parallelization of the resolution of (1.3)–(1.5) can be done. A standard method consists of subdividing the interval $[0,T]$ into subintervals and to compute iteratively the corresponding adequate initial conditions for parallel resolution on each subinterval. This approach is also the base of the multiple shooting methods (see [22, sect. 17.1] and, e.g., [8]). In [10] a comparison between these methods shows that the parareal algorithms can be recast as a multiple shooting algorithm where the Jacobian matrix is approached by a finite difference method on a coarse grid.

Another body of related literature was introduced in the pioneering works of Hackbush; see [11, 12] and also [14, 19, 35]. The parareal method corresponds to a two-level multigrid with a larger coarsening rate and an unusual smoother which corresponds to a single phase of a bicolor relaxation scheme.

Such time parallelization procedures have already been associated to optimization procedures to tackle control problems, e.g., in the case of ordinary differential equations [13], or linear control of hyperbolic [17], or parabolic evolution equations [4, 34]. On the contrary, we present here a new method to treat the bilinear optimal control of the Schrödinger equation (1.1), and consider a particular decomposition of J in sub-cost functionals corresponding to the time subdivision. In this framework, an iterative optimization procedure is designed that converges to a critical point of J.

The paper is organized as follows: parareal optimal control settings corresponding to our quantum control problem is presented in section 2. In section 3, we give an algorithm achieving a parallel in time optimization. We prove the convergence of this procedure towards a critical point of a discrete version of the cost functional J in section 4 and we finally give some numerical results in section 5.

2. Parareal control setting. Throughout this section the control field ε is either a function of $L^2([0,T])$ or its corresponding time discretization.

2.1. Main features of the parallelization. Following Lions, Maday, and Turinici [18], we now introduce the necessary concepts and tools involved in the time parallelization proposed by the parareal approach.

2.1.1. Subdivision of $[0,T]$ and virtual controls. Let $N \geq 1$ be an integer. In order to design the time parallelized optimization procedure, we introduce a subdivision of $[0,T]$:

$$[0,T] = \bigcup_{\ell=0}^{N-1} [T_\ell, T_{\ell+1}],$$

with $T_0 = 0$ and $T_N = T$. Consider also a set $\Lambda = (\lambda_\ell)_{\ell=1,\dots,N-1} \in \left(L^2(\Omega)\right)^{N-1}$. In what follows, Λ will be called the set of *virtual controls*. For notational simplicity, we will also denote by λ_0 the initial state ψ_{init} and by λ_N the target state ψ_{target}. The resolution of (1.1) is now substituted by the resolution of the N problems:

$$\left\{ \begin{array}{rcl} i\frac{\partial}{\partial t}\psi_\ell^{\varepsilon_\ell}(x,t) & = & (H(x) - \varepsilon_\ell(t)\mu(x))\psi_\ell^{\varepsilon_\ell}(x,t), \\ \psi_\ell^{\varepsilon_\ell}(x,T_\ell) & = & \lambda_\ell(x), \quad \ell = 0,\dots,N-1, \end{array} \right. \tag{2.1}$$

where ε_ℓ is the restriction of ε to $[T_\ell, T_{\ell+1}]$ (with $\ell = 0,\dots,N-1$). By (2.1), $\psi_\ell^{\varepsilon_\ell}$ also depends on λ_ℓ. In order to simplify the notations, we omit this dependence. The solution of (2.1) coincides to that of (1.1) if and only if

$$\forall \ell = 0,\dots,N-1, \qquad \psi_\ell^{\varepsilon_\ell}(x,T_{\ell+1}) = \lambda_{\ell+1}(x). \tag{2.2}$$

The parareal framework provides different methods to iteratively compute a solution Λ^* of (2.1)–(2.2).

2.1.2. Coarse and fine propagators. Suppose that the numerical simulation of (1.1) can be realized both by a coarse propagator and a fine propagator. Because the use of the coarse propagator is considered to be cheap, it can be used for the

resolution of (1.1) over the whole interval of control $[0, T]$. On the contrary, the fine propagator will only be used for parallel resolutions on $[T_\ell, T_{\ell+1}]$.

The analysis of the algorithm presented below requires that the L^2-norm of the finely approximated solution be constant with respect to the time (as is the case for the exact solution of (1.1)). Because of its numerical accuracy, we choose to consider the Strang-second-order split-operator solver [3, 30], that fulfills this property. Let us present it in some detail.

Consider two parameters of the time discretization n and $\delta t = \frac{T}{n}$, and define $n_0 = 0 < n_1 < \cdots < n_\ell < \cdots < n_N = n$, the time indexes associated with $(T_\ell)_{\ell=0,\ldots,N}$. Let us also introduce, for $\ell = 0, \ldots, N-1$, $j = n_\ell, \ldots, n_{\ell+1}-1$, the discretized control field $\varepsilon_{\ell,j}$ and for $\ell = 0, \ldots, N-1$, $j = n_\ell, \ldots, n_{\ell+1}$, the discretized state $\psi^{\varepsilon_\ell}_{\ell,j}$ that stand, respectively, for approximations of $\varepsilon_\ell(j\delta t)$ and $\psi^{\varepsilon_\ell}_\ell(j\delta t)$. The time discretization of (2.1) is given by

$$\left\{ \begin{array}{rcl} \psi^{\varepsilon_\ell}_{\ell,j+1} & = & e^{\frac{H_0\delta t}{2i}} e^{\frac{V-\mu\varepsilon_{\ell,j}}{i}\delta t} e^{\frac{H_0\delta t}{2i}} \psi^{\varepsilon_\ell}_{\ell,j}, \quad j = n_\ell, \ldots, n_{\ell+1}-1, \\ \psi^{\varepsilon_\ell}_{\ell,n_\ell} & = & \lambda_\ell, \quad \ell = 0, \ldots, N-1. \end{array} \right. \tag{2.3}$$

We refer the reader to [20] for additional details about the corresponding full discretization.

Remark 1. Though we focus on the Strang-second-order split-operator scheme, the methodology presented in this paper can be adapted to other time discretizations. Indeed, the analysis done below requires only that the norm of the wavefunction be preserved during the propagation.

2.1.3. Parareal strategy. Parareal algorithms aim at computing in parallel on each subinterval the solution of evolution equations such as (1.1). To do this, they propose various iterative methods to update the virtual controls after each parallel propagation. The purpose of what follows is to define an updated algorithm that allows one to couple the parareal approach with an optimization procedure of quantum control.

2.2. Parareal cost functionals. Let $\Lambda = (\lambda_\ell)_{\ell=1,\ldots,N-1}$ be a set of virtual controls and $(\psi^{\varepsilon_\ell}_\ell)_{\ell=0,\ldots,N-1}$ the corresponding time discretized solutions of the parallel propagations (2.3), with $\psi^{\varepsilon_\ell}_\ell = (\psi^{\varepsilon_\ell}_{\ell,j})_{j=n_\ell,\ldots,n_{\ell+1}}$. In order to design a formulation combining optimal control and parareal framework, let us also consider the following cost functional:

$$J_\parallel(\varepsilon, \Lambda) = \sum_{\ell=0}^{N-1} \beta_\ell \|\psi^{\varepsilon_\ell}_{\ell,n_{\ell+1}} - \lambda_{\ell+1}\|^2_{L^2(\Omega)} + \delta t \sum_{\ell=0}^{N-1} \sum_{j=n_\ell}^{n_{\ell+1}-1} \alpha_j \varepsilon^2_{\ell,j}, \tag{2.4}$$

where $\beta_\ell = \frac{n}{n_{\ell+1}-n_\ell}$ and α_j is an approximation $\alpha(j\delta t)$. This cost functional can be decomposed as follows:

$$J_\parallel(\varepsilon, \Lambda) = \sum_{\ell=0}^{N-1} \beta_\ell J_\ell(\varepsilon_\ell, \lambda_\ell, \lambda_{\ell+1}),$$

where J_ℓ are the *parareal cost functionals*:

$$J_\ell(\varepsilon_\ell, \lambda_\ell, \lambda_{\ell+1}) = \|\psi^{\varepsilon_\ell}_{\ell,n_{\ell+1}} - \lambda_{\ell+1}\|^2_{L^2(\Omega)} + \delta t \sum_{j=n_\ell}^{n_{\ell+1}-1} \alpha'_{\ell,j} \varepsilon^2_{\ell,j},$$

with:

$$\alpha'_{\ell,j} = \frac{\alpha_j}{\beta_\ell}. \tag{2.5}$$

Note that the optimization problems defined on $[T_\ell, T_{\ell+1}]$ via J_ℓ are similar to the initial control problem on $[0,T]$ corresponding to (1.2).

3. Monotonic parareal algorithm. Our aim is to optimize $J_\parallel(\varepsilon,\Lambda)$ with respect to its two variables. We first present the main features of our algorithm and then give further details on each step.

3.1. Structure of the algorithm. To couple parareal framework and the control optimization, we propose the following methodology: given $\nu > 0$, consider the termination criterion $c(\varepsilon) = J(\varepsilon) + \nu \sum_{\ell=0}^{N-2} |\varepsilon_{\ell,n_{\ell+1}-1} - \varepsilon_{\ell+1,n_{\ell+1}}|^2$. Given an initial control field ε^k and a tolerance $tol \geq 0$, the computation of ε^{k+1} is done as follows:

1. If $c(\varepsilon^k) \leq tol$, then stop.
2. Compute a coarse solution $\psi^k = \psi^{\varepsilon^k}$ of (1.3).
3. Compute a coarse solution $\chi^k = \chi^{\varepsilon^k}$ of (1.4).
4. Using ψ^k and χ^k, compute Λ^k, which optimizes $J_\parallel(\varepsilon^k,\Lambda)$ with respect to Λ.
5. On each interval $[T_\ell, T_{\ell+1}]$, compute in parallel some control field ε_ℓ^{k+1}, which optimizes the cost functionals $J_\ell(\varepsilon_\ell, \lambda_\ell^k, \lambda_{\ell+1}^k)$ with respect to ε_ℓ.
6. Define ε^{k+1} as the concatenation of the control fields ε_ℓ^{k+1}.
7. Assign $k \leftarrow k+1$. Return to step 1.

This algorithm is similar to an alternate direction descent algorithm, in the sense that it alternatively optimizes $J_\parallel(\varepsilon,\Lambda)$ with respect to Λ and to ε.

Of course, to take advantage of the time parallelization, steps 2 and 3 of the previous algorithm are to be computed using the coarse propagator, whereas step 5 can be done simultaneously and by fine propagations.

Remark 2. Solving (1.4) exploits in an essential manner the time-reversibility of the Schrödinger equation. Further work is required to extend this algorithm to nonreversible cases.

3.2. Virtual controls definition. We present in this section some results which will enable us to achieve efficiently step 4 of the monotonic parareal algorithm. As we do not intend to deal with the coarse propagator properties, we will consider in this section that steps 2 and 3 are done with the split-operator method presented in section 2.1.2, with the (small) time step δt. Even though this is not what we want to do ultimately, the results below keep a practical interest since the most expensive part of the algorithm is the update of the control field which will be done in parallel. Thus, given a control field $\varepsilon = (\varepsilon_j)_{j=0,\dots,n-1}$, the states $\psi^\varepsilon = (\psi_j^\varepsilon)_{j=0,\dots,n}$ and the adjoint states $\chi^\varepsilon = (\chi_j^\varepsilon)_{j=0,\dots,n}$ are computed by

$$\left\{ \begin{array}{rcl} \psi_{j+1}^\varepsilon & = & e^{\frac{H_0\delta t}{2i}} e^{\frac{V-\mu\varepsilon_j}{i}\delta t} e^{\frac{H_0\delta t}{2i}} \psi_j^\varepsilon, \\ \psi_0^\varepsilon & = & \psi_{init}, \end{array} \right. \tag{3.1}$$

$$\left\{ \begin{array}{rcl} \chi_{j-1}^\varepsilon & = & e^{-\frac{H_0\delta t}{2i}} e^{-\frac{V-\mu\varepsilon_{j-1}}{i}\delta t} e^{-\frac{H_0\delta t}{2i}} \chi_j^\varepsilon, \\ \chi_n^\varepsilon & = & \psi_{target}. \end{array} \right. \tag{3.2}$$

For reasons of simplicity, we use in what follows the following notation:

$$\mathcal{F}_{j,j'}^\varepsilon \psi_j^\varepsilon = \psi_{j'}^\varepsilon,$$

where $0 \le j, j' \le n$ are two (time) indices, and $\psi^\varepsilon = (\psi_j^\varepsilon)_{j=0,\dots,n}$ is the solution of (3.1). We still denote by J the time discretized cost functional corresponding to (1.2):

$$J(\varepsilon) = \|\psi_n^\varepsilon - \psi_{target}\|^2_{L^2(\Omega)} + \delta t \sum_{j=0}^{n-1} \alpha_j \varepsilon_j^2. \tag{3.3}$$

The following theorem provides the optimal choice of virtual controls Λ.

THEOREM 3.1. *With the previous notations, let us define* $\Lambda^\varepsilon = (\lambda_\ell^\varepsilon)_{\ell=1,\dots,N-1}$ *by*

$$\lambda_\ell^\varepsilon = (1-\gamma_\ell)\psi_{n_\ell}^\varepsilon + \gamma_\ell \chi_{n_\ell}^\varepsilon, \tag{3.4}$$

where $\gamma_\ell = \frac{n_\ell}{n}$. *Then*

$$\Lambda^\varepsilon = \operatorname{argmin}_\Lambda (J_\parallel(\varepsilon, \Lambda)). \tag{3.5}$$

Moreover, we have

$$J_\parallel(\varepsilon, \Lambda^\varepsilon) = J(\varepsilon).$$

Proof. For a given $\Lambda = (\lambda_\ell)_{\ell=1,\dots,N-1}$, let us first prove that $J(\varepsilon)$ is a lower bound for $J_\parallel(\varepsilon, \Lambda)$. The Cauchy–Schwarz inequality gives

$$\sum_{\ell=0}^{N-1} \beta_\ell \|\psi_{\ell,n_{\ell+1}}^{\varepsilon_\ell} - \lambda_{\ell+1}\|^2_{L^2(\Omega)} = \left(\sum_{\ell=0}^{N-1} \frac{1}{\beta_\ell}\right) \sum_{\ell=0}^{N-1} \beta_\ell \|\psi_{\ell,n_{\ell+1}}^{\varepsilon_\ell} - \lambda_{\ell+1}\|^2_{L^2(\Omega)} \tag{3.6}$$

$$\ge \left(\sum_{\ell=0}^{N-1} \|\psi_{\ell,n_{\ell+1}}^{\varepsilon_\ell} - \lambda_{\ell+1}\|_{L^2(\Omega)}\right)^2. \tag{3.7}$$

Recalling that $\mathcal{F}^\varepsilon$ is a unitary operator, we have

$\forall \ell = 1, \dots, N-1,$

$$\begin{aligned} \|\psi_{\ell,n_{\ell+1}}^{\varepsilon_\ell} - \lambda_{\ell+1}\|_{L^2(\Omega)} &= \|\mathcal{F}^\varepsilon_{n_\ell,n_{\ell+1}}(\lambda_\ell) - \lambda_{\ell+1}\|_{L^2(\Omega)} \\ &= \big\|\mathcal{F}^\varepsilon_{n_{\ell+1},n}\big(\mathcal{F}^\varepsilon_{n_\ell,n_{\ell+1}}(\lambda_\ell) - \lambda_{\ell+1}\big)\big\|_{L^2(\Omega)} \\ &= \|\mathcal{F}^\varepsilon_{n_\ell,n}(\lambda_\ell) - \mathcal{F}^\varepsilon_{n_{\ell+1},n}(\lambda_{\ell+1})\|_{L^2(\Omega)}. \end{aligned} \tag{3.8}$$

Hence

$$\begin{aligned} \sum_{\ell=0}^{N-1} \|\psi_{\ell,n_{\ell+1}}^{\varepsilon_\ell} - \lambda_{\ell+1}\|_{L^2(\Omega)} &= \sum_{\ell=0}^{N-1} \|\mathcal{F}^\varepsilon_{n_\ell,n}(\lambda_\ell) - \mathcal{F}^\varepsilon_{n_{\ell+1},n}(\lambda_{\ell+1})\|_{L^2(\Omega)} \\ &\ge \|\mathcal{F}^\varepsilon_{0,n}(\psi_{init}) - \psi_{target}\|_{L^2(\Omega)}. \end{aligned} \tag{3.9}$$

Combining (3.9) and (3.6) we obtain, since $\mathcal{F}^\varepsilon_{0,n}(\psi_{init}) = \psi_n^\varepsilon$,

$$J_\parallel(\varepsilon, \Lambda) \ge J(\varepsilon).$$

By (3.4), we have

$$\begin{aligned} \psi_{\ell,n_{\ell+1}}^{\varepsilon_\ell} - \lambda_{\ell+1}^\varepsilon &= \mathcal{F}^\varepsilon_{n_\ell,n_{\ell+1}}(\lambda_\ell^\varepsilon) - \lambda_{\ell+1}^\varepsilon \\ &= \mathcal{F}^\varepsilon_{n_\ell,n_{\ell+1}}\big((1-\gamma_\ell)\psi_{n_\ell}^\varepsilon + \gamma_\ell \chi_{n_\ell}^\varepsilon\big) - \big((1-\gamma_{l+1})\psi_{n_{\ell+1}}^\varepsilon + \gamma_{\ell+1}\chi_{n_{\ell+1}}^\varepsilon\big) \\ &= (\gamma_{\ell+1} - \gamma_\ell)\big(\psi_{n_{\ell+1}}^\varepsilon - \chi_{n_{\ell+1}}^\varepsilon\big). \end{aligned} \tag{3.10}$$

Hence,

$$\|\psi^{\varepsilon_\ell}_{\ell,n_{\ell+1}} - \lambda^{\varepsilon}_{\ell+1}\|_{L^2(\Omega)} = \frac{1}{\beta_\ell}\|\psi^{\varepsilon}_n - \psi_{target}\|_{L^2(\Omega)}.$$

Combining this equality with (2.4), we obtain the following:

$$\begin{aligned} J_{\|}(\varepsilon,\Lambda^\varepsilon) &= \sum_{\ell=0}^{N-1}\frac{1}{\beta_\ell}\|\psi^\varepsilon_n - \psi_{target}\|^2_{L^2(\Omega)} + \delta t\sum_{\ell=0}^{N-1}\sum_{j=n_\ell}^{n_{\ell+1}-1}\alpha_j\varepsilon^2_{\ell,j}, \\ &= \|\psi^\varepsilon_n - \psi_{target}\|^2_{L^2(\Omega)} + \delta t\sum_{\ell=0}^{N-1}\sum_{j=n_\ell}^{n_{\ell+1}-1}\alpha_j\varepsilon^2_{\ell,j}, \end{aligned} \tag{3.11}$$

and the theorem follows. ∎

Remark 3. The trajectory $\left((1-\frac{j}{N})\psi^\varepsilon_j + \frac{j}{N}\chi^\varepsilon_j\right)_{j=0,\dots,n}$ is an ideal trajectory that reaches exactly the target ψ_{target}. The choice $\Lambda = \Lambda^\varepsilon$ is equivalent to define the virtual controls on this trajectory. This interpretation is closely related to the concept of reference trajectory tracking: through the introduction of the parareal cost functionals, the initial problem is transformed into $N-1$ optimization problems that aim to minimize on each subinterval the distance between the current trajectory and this ideal (unknown) trajectory.

Remark 4. An alternative definition for Λ can be

$$\lambda_\ell = \frac{(1-\gamma_\ell)\psi^\varepsilon_{n_\ell} + \gamma_\ell\chi^\varepsilon_{n_\ell}}{\|(1-\gamma_\ell)\psi^\varepsilon_{n_\ell} + \gamma_\ell\chi^\varepsilon_{n_\ell}\|_{L^2(\Omega)}}, \tag{3.12}$$

which has the advantage that the norms of the virtual controls are preserved. Furthermore, it can be proved that λ corresponds to a critical point of $J_{\|}(\varepsilon,\Lambda)$ under the constraint

$$\forall \ell = 1,\dots,N-1, \quad \|\lambda_\ell\|_{L^2(\Omega)} = 1.$$

3.3. Monotonic schemes. Let us now describe a practical implementation of step 5 of the monotonic parareal algorithm. Given a set of virtual controls $\Lambda = (\lambda_\ell)_{\ell=1,\dots,N-1}$ (recall that $\lambda_0 = \psi_0$ and $\lambda_N = \psi_{target}$), the parareal cost functional J_ℓ reads

$$\begin{aligned} J_\ell(\varepsilon_\ell,\lambda_\ell,\lambda_{\ell+1}) &= \|\lambda_\ell\|^2_{L^2(\Omega)} + \|\lambda_{\ell+1}\|^2_{L^2(\Omega)} - 2Re\langle\psi_{\ell,n_{\ell+1}},\lambda_{\ell+1}\rangle \\ &\quad + \delta t\sum_{j=n_\ell}^{n_{\ell+1}-1}\alpha'_{\ell,j}\varepsilon^2_{\ell,j}, \quad \ell = 0,\dots,N-1. \end{aligned} \tag{3.13}$$

The first two terms of J_ℓ will not change during the optimization of ε. An efficient way to minimize cost functionals of the form $\widetilde{J}(\varepsilon) = -2Re\langle\psi(T),\psi_{target}\rangle + \int_0^T\alpha(t)\varepsilon^2(t)dt$ associated with the Schrödinger equation is to use monotonic schemes [21, 37].

In our case, the time discretized monotonic scheme corresponding to (3.13) can be defined by the following procedure.

Let us consider $(\delta,\eta) \in [0,2[\times[0,2[$ and introduce the notations

$$\widetilde{\chi}_{\ell,j} = e^{\frac{H_0\delta t}{2i}}\chi_{\ell,j}, \quad \breve{\psi}_{\ell,j} = e^{\frac{H_0\delta t}{2i}}\psi_{\ell,j}, \quad \breve{\chi}_{\ell,j} = e^{-\frac{H_0\delta t}{2i}}\chi_{\ell,j}, \quad \widetilde{\psi}_{\ell,j} = e^{-\frac{H_0\delta t}{2i}}\psi_{\ell,j},$$

and $\mu^*(h)$, an approximation of μ, defined by

$$\mu^*(h) = \frac{e^{-i\mu h\delta t} - Id}{-ih\delta t}, \tag{3.14}$$

where Id is the identity operator.

Given a control field ε^k, its restriction ε^k_ℓ to the interval $[T_\ell, T_{\ell+1}]$ and the corresponding $\psi^k_\ell = (\psi^k_{\ell,j})_{j=n_\ell,\dots,n_{\ell+1}}$ are defined iteratively by

$$\begin{cases} \psi^k_{\ell,j+1} &= e^{\frac{H_0\delta t}{2i}} e^{\frac{V-\mu\varepsilon^k_{\ell,j}}{i}\delta t} e^{\frac{H_0\delta t}{2i}} \psi^k_{\ell,j}, \quad j = n_\ell, \dots, n_{\ell+1}-1, \\ \psi^k_{\ell,n_\ell} &= \lambda_\ell, \quad \ell = 0, \dots, N-1. \end{cases} \tag{3.15}$$

The computation of ε^{k+1}_ℓ is performed as follows:

1. For $\ell = 0, \dots, N-1$, compute iteratively an intermediate control field $\widetilde{\varepsilon}^k_\ell$ and its corresponding adjoint state χ^k_ℓ by

$$\chi^k_{\ell,j} = e^{-\frac{H_0\delta t}{2i}} e^{\frac{-V+\mu\widetilde{\varepsilon}^k_{\ell,j}}{i}\delta t} e^{-\frac{H_0\delta t}{2i}} \chi^k_{\ell,j+1}, \quad j = n_\ell, \dots, n_{\ell+1}-1, \tag{3.16}$$

where $\widetilde{\varepsilon}^k_{\ell,j}$ is such that

$$\widetilde{\varepsilon}^k_{\ell,j} = (1-\eta)\varepsilon^k_{\ell,j} - \frac{\eta}{\alpha'_{\ell,j}} Im\langle \breve{\chi}^k_{\ell,j+1} | \mu^*(\varepsilon^k_{\ell,j} - \widetilde{\varepsilon}^k_{\ell,j}) | \widetilde{\psi}^k_{\ell,j+1} \rangle, \tag{3.17}$$

with the final condition

$$\chi^k_{\ell,n_{\ell+1}} = \lambda_{\ell+1}.$$

2. For $\ell = 0, \dots, N-1$, compute iteratively the control field ε^{k+1}_ℓ and its corresponding state ψ^{k+1}_ℓ by

$$\psi^{k+1}_{\ell,j+1} = e^{\frac{H_0\delta t}{2i}} e^{\frac{V-\mu\varepsilon^{k+1}_{\ell,j}}{i}\delta t} e^{\frac{H_0\delta t}{2i}} \psi^{k+1}_{\ell,j}, \quad j = n_\ell, \dots, n_{\ell+1}-1,$$

where $\varepsilon^{k+1}_{\ell,j}$ is such that

$$\varepsilon^{k+1}_{\ell,j} = (1-\delta)\widetilde{\varepsilon}^k_{\ell,j} - \frac{\delta}{\alpha'_{\ell,j}} Im\langle \widetilde{\chi}^k_j | \mu^*(\varepsilon^{k+1}_{\ell,j} - \widetilde{\varepsilon}^k_{\ell,j}) | \breve{\psi}^{k+1}_{\ell,j} \rangle, \tag{3.18}$$

with the initial condition

$$\psi^{k+1}_{\ell,n_\ell} = \lambda_\ell. \tag{3.19}$$

The implicit equations (3.17) and (3.18) are solved independently for $\widetilde{\varepsilon}^k_{\ell,j}$ and $\varepsilon^{k+1}_{\ell,j}$, respectively, at each time step by a Newton method (all other parameters involved are known). We refer the reader to [20] for a proof of the existence of solutions and further details on this scheme.

In what follows, the initial value ε^0 of the monotonic scheme is considered fixed. An important property of this algorithm is given in the following theorem [21].

THEOREM 3.2. *For any $\eta, \delta \in [0,2]$ the algorithm given in* (3.16)–(3.19) *is well defined and converges monotonically in the sense that*

$$\forall \ell = 0, \dots, N-1,$$

$$J_\ell(\varepsilon_\ell^{k+1}, \lambda_\ell, \lambda_{\ell+1}) - J_\ell(\varepsilon_\ell^{k}, \lambda_\ell, \lambda_{\ell+1}) = -\delta t \sum_{j=n_\ell}^{n_{\ell+1}-1} \alpha'_{\ell,j} (\varepsilon_{j,\ell}^{k+1} - \varepsilon_{j,\ell}^{k})^2 \leq 0. \tag{3.20}$$

This optimization procedure can be done in parallel on each interval $[T_\ell, T_{\ell+1}]$. Consequently, the computations can be carried out with fine propagators $\mathcal{F}^{\varepsilon_\ell^{k+1}}$.

Remark 5. As was the case for step 4 (see Remark 3), this step of the algorithm is also linked to the concept of reference trajectory tracking: in the monotonic schemes, the control field $\widetilde{\varepsilon}_{\ell,j1}^{k}$ (resp., $\varepsilon_{\ell,j}^{k+1}$) is chosen such that the distance between the current states and adjoint state $\|\psi_{\ell,j}^{k} - \chi_{\ell,j}^{k}\|_{L^2(\Omega)}$ (resp., $\|\psi_{\ell,j+1}^{k+1} - \chi_{\ell,j+1}^{k}\|_{L^2(\Omega)}$) decreases at each time step. We refer the reader to [28, 32] for details about the relationship between the monotonic schemes and local tracking algorithms.

Remark 6. Note that several iterations of this scheme can be done during step 5 of the monotonic parareal algorithm, especially in case of slow convergence (see Table 5.1 in section 5.4 for numerical results about it).

The algorithm is schematically depicted in Figure 3.1.

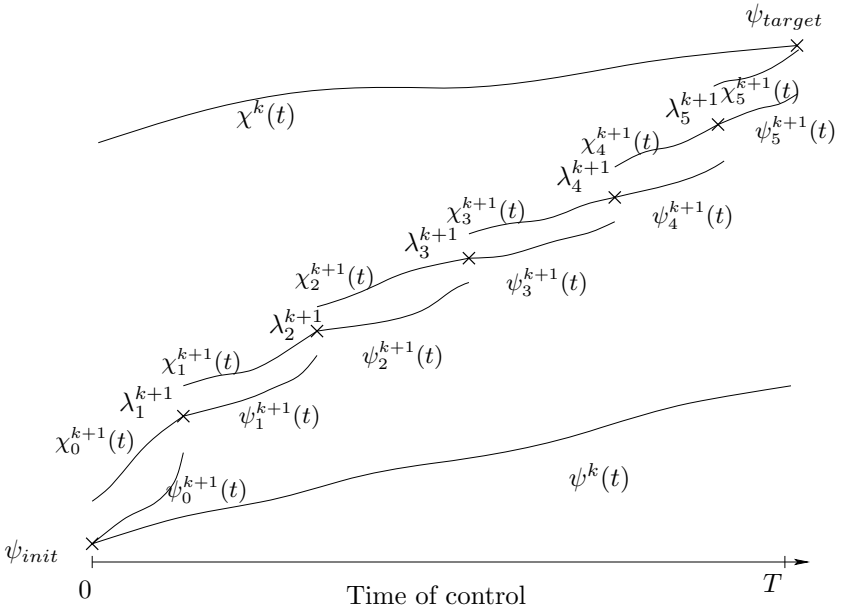


FIG. 3.1. *Symbolic representation of one iteration of the monotonic parareal algorithm. The optimization is achieved in parallel on each subinterval $[T_\ell, T_{\ell+1}]$. The virtual controls λ_ℓ^k are updated at each iteration.*

3.4. Monotonicity of the algorithm. The combination of the two previous strategies allows us to define

$$\Lambda^k = \Lambda^{\varepsilon^k}$$

and ε^{k+1} as the concatenation of the sequence $(\varepsilon_\ell^{k+1})_{\ell=0,\dots,N-1}$. We have thus obtained a global monotonic algorithm since

$$J(\varepsilon^{k+1}) = J_\parallel(\varepsilon^{k+1}, \Lambda^{k+1}) \leq J_\parallel(\varepsilon^{k+1}, \Lambda^k) \leq J_\parallel(\varepsilon^k, \Lambda^k) = J(\varepsilon^k).$$

4. Convergence of the algorithm. The convergence of the sequence $(\varepsilon^k)_{k\in\mathbf{N}}$ defined by the previous algorithm is described in the next theorem.

THEOREM 4.1. *Given an initial control field ε^0, consider the sequence $(\varepsilon^k)_{k\in\mathbf{N}}$ obtained by the algorithm (3.16)–(3.19), where the coarse propagator in steps 2–3 is taken to be the same as the fine propagator. The sequence $(\varepsilon^k)_{k\in\mathbf{N}}$ converges towards a critical point of J.*

Proof. Since the proof is very similar to that presented in [26], we give only a brief overview.

Denote by C_J the set of critical points of J. Using the previous notations, this set reads

$$C_J = \left\{\varepsilon /\ \forall j = 0,\dots,N-1,\ Im\langle\widetilde{\chi}_j^\varepsilon|\mu|\breve{\psi}_j^\varepsilon\rangle + \alpha_j\varepsilon_j = 0\right\}. \tag{4.1}$$

Let C_{ε^0} be the set of limit points of $(\varepsilon^k)_{k\in\mathbf{N}}$.

Since $\|\lambda_0^k\|_{L^2(\Omega)} = \|\lambda_N^k\|_{L^2(\Omega)} = 1$, (3.4) implies that

$$\forall \ell = 0,\dots,N, \quad \|\lambda_\ell^k\|_{L^2(\Omega)} \leq 1.$$

It can then be proved by induction that (see [20, Theorem 3])

$$\forall k \in \mathbf{N}, \quad \forall j = 0,\dots,n-1, \quad |\varepsilon_j^k| \leq M, \tag{4.2}$$

with

$$M = \max\left(\|\varepsilon^0\|_\infty, \max\left(1, \frac{\delta}{2-\delta}, \frac{\eta}{2-\eta}\right)\frac{\|\mu\|_*}{\min_{j=1,\dots,n-1}(\alpha_j)}\right), \tag{4.3}$$

where $\|\mu\|_*$ denotes the operator norm of μ and $\|\varepsilon^0\|_\infty = \max_{j=0,\dots,n-1}(|\varepsilon_j^0|)$. Consider now a converging subsequence $(\varepsilon^{k_p})_{p\in\mathbf{N}}$ and its limit $\varepsilon^\infty \in C_{\varepsilon^0}$. The corresponding state $\psi^{\varepsilon^{k_p}} = \psi^{k_p}$ and adjoint state $\chi^{\varepsilon^{k_p}} = \chi^{k_p}$ defined by (3.1) and (3.2) converge, respectively, towards $\psi^{\varepsilon^\infty}$ and $\chi^{\varepsilon^\infty}$. By (3.4), the sequence $(\Lambda_p^k)_{p\in\mathbf{N}}$ converges towards $\Lambda^{\varepsilon^\infty}$. Then, a similar proof indicates that $(\psi_\ell^{k_p})_{p\in\mathbf{N}}$ and $(\chi_\ell^{k_p})_{p\in\mathbf{N}}$ converge towards $\psi_\ell^{\varepsilon^\infty}$ and $\chi_\ell^{\varepsilon^\infty}$. Thanks to the similar structures of J_ℓ and J, one can adapt the proof of Lemma 3.3 in [26] which shows that

$$\forall \ell = 0,\dots,N-1, \quad \forall j = n_\ell,\dots,n_{\ell+1}-1, \quad Im\langle\widetilde{\chi}_{\ell,j}^{\varepsilon^\infty}|\mu|\breve{\psi}_{\ell,j}^{\varepsilon^\infty}\rangle + \alpha'_{\ell,j}\varepsilon_{\ell,j}^\infty = 0. \tag{4.4}$$

Another use of (3.4) proves that

$$\forall \ell = 0,\dots,N-1, \quad \forall j = n_\ell,\dots,n_{\ell+1}-1, \quad \frac{1}{\beta_\ell} Im\langle\widetilde{\chi}_{\ell,j}^{\varepsilon^\infty}|\mu|\breve{\psi}_{\ell,j}^{\varepsilon^\infty}\rangle = Im\langle\widetilde{\chi}_j^{\varepsilon^\infty}|\mu|\breve{\psi}_j^{\varepsilon^\infty}\rangle. \tag{4.5}$$

Combining (4.4) and (4.5) with (2.5), we obtain that

$$C_{\varepsilon^0} \subset C_J. \tag{4.6}$$

Since $(J(\varepsilon^k))_{k\in\mathbf{N}}$ is a bounded (by $2 + \max_{j=0,\dots,n-1}(\alpha_j)TM$) monotonic sequence, it converges towards a limit denoted by l_{ε^0}.

We are now in the position to reproduce the analysis in Theorem 4.5 of [26] which shows that

$$\exists \theta \in \left]0, \frac{1}{2}\right], \exists \kappa > 0, \exists k_0 \in \mathbf{N} / \ \forall k \geq k_0,$$
$$(l_{\varepsilon^0} - J(\varepsilon^k))^\theta - (l_{\varepsilon^0} - J(\varepsilon^{k+1}))^\theta \geq \kappa \|\varepsilon^{k+1} - \varepsilon^k\|_2, \tag{4.7}$$

where $\|.\|_2$ denotes the usual Euclidean norm on $\mathbf{R}^n$. Hence, the sequence $(\varepsilon^k)_{k\in\mathbf{N}}$ is a Cauchy sequence and by (4.6) the theorem follows. □

5. Numerical results.

5.1. Model. In order to test the performance of the algorithm, a case already treated in the literature has been considered. The system is a molecule of HCN modelled as a rigid rotator. We refer the reader to [1, 27] for numerical details concerning this system. The goal is to control the orientation of the system; this is expressed through the requirement that $\|\psi(T) - \psi_{target}\|_{L^2(\Omega)}$ is minimized, where the target function ψ_{target} corresponds to an orientated state.

5.2. Propagators. All the propagations are done through a Strang-second-order split-operator type as, e.g., in (3.15). The coarse propagator, corresponding to (3.1) and (3.2), is used with a large time step, whereas the fine propagator, as it appears in (3.18) and (3.17), is used with a small time step. The ratio of these two time steps is 10. We have observed that a renormalization (3.12) slightly reduces the speed of convergence, but has no effect on the converged control fields.

5.3. Numerical convergence. Let us present some results concerning the convergence of the monotonic parareal algorithm.

5.3.1. Evolution of $(\varepsilon^k)_{k\in\mathbf{N}}$. In a first test, $N = 10$ identical subintervals are considered to parallelize the algorithm. Figures 5.1–5.3 show a typical evolution of the sequence $(\varepsilon^k)_{k\in\mathbf{N}}$ computed by our algorithm. Figure 5.4 represents the control field obtained without parallelization by a monotonic algorithm applied to J. Note that the discontinuities that are visible on Figure 5.1 and even on Figure 5.2 disappear as the number of iteration increases.

5.3.2. Variation of the number of subintervals. The control field obtained at the numerical convergence appears to be independent of the number N of subintervals. This is coherent with Theorem 4.1 which claims that the limit depends only on the initial cost functional J. In order to evaluate the effect of N on the convergence of the algorithm, we have run the algorithm for different values of this parameter. Figure 5.5 shows the evolution of the cost functional values for $N = 1, 2, 5, 10, 50, 100$. The convergence speed decreases when N becomes larger. One has thus to find an optimum between the acceleration obtained by parallelization and the reduction of the convergence speed induced by it. This compromise depends on the choice of the coarse and fine propagators. In our case, the parallel optimizations and the coarse propagations allow us to reach a satisfactory cost functional value with an actual gain in "wall-clock" time roughly equal to 7. We have not sought to optimize the coarse and the fine propagators for these numerical tests. In particular, a coarser propagator should improve this result.

5.4. About the full efficiency. Consider again $N = 10$ identical subintervals. An ideal time parallelization should divide the computational time by 10. In order to

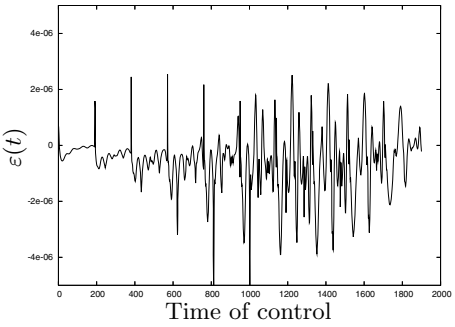


FIG. 5.1. *Field of control obtained after one iteration.*

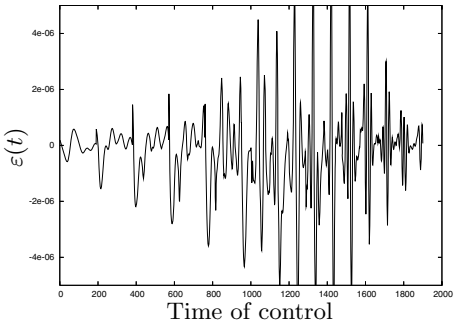


FIG. 5.2. *Field of control obtained after* 10 *iterations.*

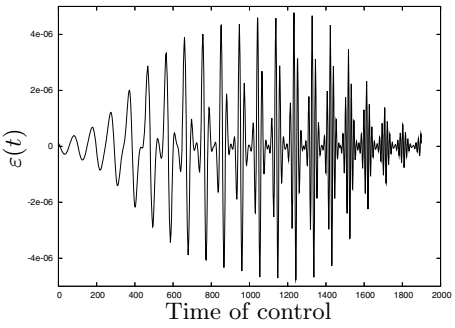


FIG. 5.3. *Field of control obtained after* 250 *iterations. In such nonlinear equations the typical number of iterations is in the range* 10^2–10^4 [1, 27].

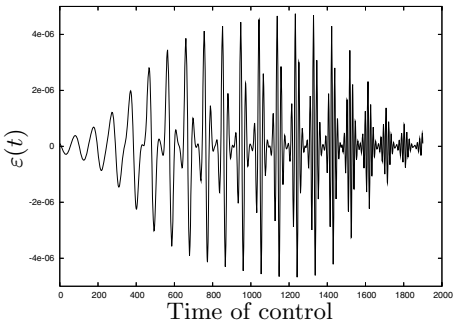


FIG. 5.4. *Field of control obtained by a nonparallelized monotonic algorithm (i.e., with* $N = 1$*) after* 250 *iterations.*

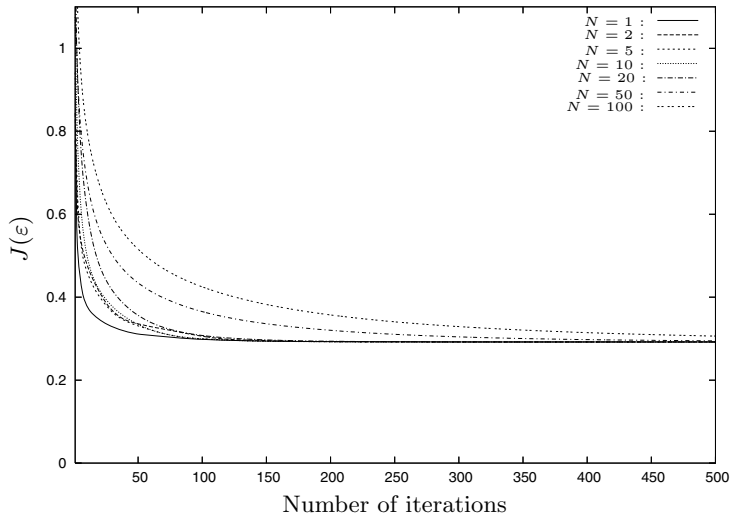


FIG. 5.5. *Evolution of the cost functional value during the first* 500 *iterations.*

TABLE 5.1

Results obtained for several subiterations of the monotonic schemes (3.16)–(3.19) *during step* 5 *of the monotonic parareal algorithm.*

	k	m	Comp. time	$J(\varepsilon^k)$
Case 0	100	1	$100.10.T_f$	0.2983
Case 1	100	1	$100.(T_f + T_C)$	0.2986
Case 2	50	2	$50.(2T_f + T_C)$	0.3062
Case 3	25	4	$25.(4T_f + T_C)$	0.3295

approach such a full efficiency, a strategy would be to increase the parallel computations (step 5) with respect to the coarse propagations (steps 2 and 3 of the algorithm). In this perspective, the influence on the convergence of the number of iterations (denoted by m) of the monotonic algorithm (3.17)–(3.18) done during step 5 has been tested (see Remark 6). Let us denote by k the number of iterations of the monotonic parareal algorithm, by T_f the computational time of one iteration of the monotonic algorithm in a subinterval, and by T_C the computational time of both step 2 and step 3. Table 5.1 summarizes the cases that have been tested.

Case 0 corresponds to the nonparallelized monotonic algorithm (i.e., with $N = 1$) computed with the fine propagator. Figure 5.6 shows that the best strategy corresponds to case 1. Further work is needed to reach the full efficiency (corresponding in our case to a computational time close to $100.T_f$).

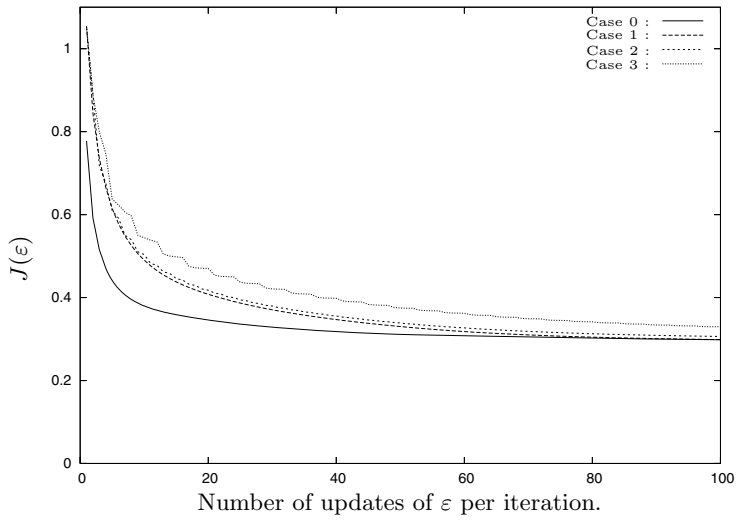


FIG. 5.6. *Evolutions of the cost functional values for different values of* m.

REFERENCES

[1] A. AUGER, A. BEN HAJ YEDDER, E. CANCES, C. LE BRIS, C. M. DION, A. KELLER, AND O. ATABEK, *Optimal laser control of molecular systems: Methodology and results*, Math. Models Methods Appl. Sci., 12 (2002), pp. 1281–1315.

[2] L. BAFFICO, S. BENARD, Y. MADAY, G. TURINICI, AND G. ZÉRAH, *Parallel in time molecular dynamics simulations*, Phys. Rev. E (3), 66 (2002), 057701.

[3] A. D. BANDRAUK AND H. SHEN, *Exponential split operator methods for solving coupled time-dependent Schrödinger equations*, J. Chem. Phys., 99 (1993), pp. 1185–1193.

[4] A Borzí, *Multigrid methods for parabolic distributed optimal control problems*, J. Comput. Appl. Math., 157 (2003), pp. 365–382.
[5] T. Brixner, N. H. Damrauer, P. Niklaus, and G. Gerber, *Photoselective adaptive femtosecond quantum control in the liquid phase*, Nature, 414 (2001), pp. 57–60.
[6] T. Brixner, G. Krampert, T. Pfeifer, R. Selle, G. Gerber, M. Wollenhaupt, O. Graefe, C. Horn, D. Liese, and T. Baumert, *Quantum control by ultrafast polarization shaping*, Phys. Rev. Lett., 92 (2004), 208301.
[7] E. Brown and H. Rabitz, *Some mathematical and algorithmic challenges in the control of quantum dynamics phenomena*, J. Math. Chem., 31 (2002), pp. 17–63.
[8] P. Chartier and B. Philippe, *A parallel shooting technique for solving dissipative ODEs*, Computing, 51 (1993), pp. 209–236.
[9] F. Courvoisier, V. Boutou, J. Kasparian, E. Salmon, G. Méjean, J. Yu, and J.-P. Wolf, *Ultraintense light filaments transmitted through clouds*, Appl. Phys. Lett., 83 (2003), pp. 213–215.
[10] M. J. Gander and S. Vandewalle, *Analysis of the parareal time-parallel time-integration method*, SIAM J. Sci. Comput., 29 (2007), pp. 556–578.
[11] W. Hackbusch, *Parabolic multi-grid methods*, in Computing Methods in Applied Sciences and Engineering VI, R. Glowinski and J.-L. Lions, eds., 1984, pp. 189–197.
[12] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, Berlin, 1985.
[13] M. Heinkenschloss, *A time-domain decomposition iterative method for the solution of distributed linear quadratic optimal control problems*, J. Comput. Appl. Math., 173 (2005), pp. 169–198.
[14] G. Horton, *The time-parallel multigrid method*, Comm. Appl. Numer. Methods, 8 (1992), pp. 585–595.
[15] R. S. Judson and H. Rabitz, *Teaching lasers to control molecules*, Phys. Rev. Lett., 68 (1992), pp. 1500–1503.
[16] J. Kasparian, M. Rodriguez, G. Méjean, J. Yu, E. Salmon, H. Wille, R. Bourayou, S. Frey, Y.-B. André, A. Mysyrowicz, R. Sauerbrey, J.-P. Wolf, and L. Wöste, *White-light filaments for atmospheric analysis*, Science, 301 (2003), pp. 61–64.
[17] J. E. Lagnese and G. Leugering, *Time-domain decomposition of optimal control problems for the wave equation*, Systems Control Lett., 48 (2003), pp. 229–242.
[18] J.-L. Lions, Y. Maday, and G. Turinici, *A parareal in time discretization of PDE*, C. R. Acad. Sci. Paris Sér. I Math., 332 (2001), pp. 661–668.
[19] C. Lubich and A. Ostermann, *Multi-grid dynamic iteration for parabolic equations*, BIT, 27 (1987), pp. 216–234.
[20] Y. Maday, J. Salomon, and G. Turinici, *Monotonic time-discretized schemes in quantum control*, Numer. Math., 103 (2006), pp. 323–338.
[21] Y. Maday and G. Turinici, *New formulations of monotonically convergent quantum control algorithms*, J. Chem. Phys., 118 (2003), pp. 8191–8196.
[22] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C++: The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, UK, 2002.
[23] H. Rabitz, R. J. Levis, and G. M. Menkir, *Selective bond dissociation and rearrangement with optimally tailored, strong-field laser pulses*, Science, 292 (2001), pp. 709–713.
[24] H. Rabitz, R. de Vivie-Riedle, M. Motzkus, and K. Kompa, *Whither the future of controlling quantum phenomena?*, Science, 288 (2000), pp. 824–828.
[25] H. Rabitz, G. Turinici, and E. Brown, *Control of quantum dynamics: Concepts, procedures and future prospects*, in Computational Chemistry, Special Volume, Handbook of Numerical Analysis, Vol. X, C. Le Bris, ed., Elsevier Science B. V., Amsterdam, The Netherlands, 2003.
[26] J. Salomon, *Convergence of the time-discretized monotonic schemes*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 77–93.
[27] J. Salomon, C. Dion, and G. Turinici, *Optimal molecular alignment and orientation through rotational ladder climbing*, J. Chem. Phys., 123 (2005), p. 144310.
[28] J. Salomon and G. Turinici, *On the relationship between the local tracking procedures and monotonic schemes in quantum optimal control*, J. Chem. Phys., 124 (2006), p. 074102.
[29] S. Shi, A. Woody, and H. Rabitz, *Optimal control of selective vibrational excitation in harmonic linear chain molecules*, J. Chem. Phys., 88 (1988), pp. 6870–6883.
[30] G. Strang, *Accurate partial difference methods.* I. *Linear Cauchy problems*, Arch. Rational Mech. Anal., 12 (1963), pp. 392–402.
[31] D. Tannor, V. Kazakov, and V. Orlov, *Control of photochemical branching: Novel procedures for finding optimal pulses and global upper bounds*, in Time Dependent Quantum

Molecular Dynamics, J. Broeckhove and L. Lathouwers, eds., Plenum, New York, 1992, pp. 347–360.

[32] G. Turinici, *Equivalence between local tracking procedures and monotonic algorithms in quantum control*, in Proceedings of the 44th IEEE Conference on Decision and Control, Sevilla, Spain, 2005, pp. 8203–8208.

[33] G. Turinici and Y. Maday, *A parallel in time approach for quantum control: The parareal algorithm*, in Proceedings of the 41st IEEE Conference on Decision and Control, Las Vegas, NV, 2002, pp. 62–66.

[34] S. Ulbrich, *Generalized SQP-Methods with "Parareal" Time-Domain Decomposition for Time-Dependent PDE-Constrained Optimization*, Technical report, Fachbereich Mathematik, TU Darmstadt, Darmstadt, Germany, 2005. To appear in Real-Time PDE-Constrained Optimization, L. T. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. van Bloemen Waanders, eds., SIAM, Philadelphia, 2007, pp. 145–168.

[35] S. Vandewalle and R. Piessens, *Efficient parallel algorithms for solving initial-boundary value and time-periodic parabolic partial differential equations*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1330–1346.

[36] G. Vogt, G. Krampert, P. Niklaus, P. Nuernberger, and G. Gerber, *Optimal control of photoisomerization*, Phys. Rev. Lett., 94 (2005), 68305.

[37] W. Zhu and H. Rabitz, *A rapid monotonically convergent iteration algorithm for quantum optimal control over the expectation value of a positive definite operator*, J. Chem. Phys., 109 (1998), pp. 385–391.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2483–2509

NODAL AUXILIARY SPACE PRECONDITIONING IN H(curl) AND H(div) SPACES*

RALF HIPTMAIR† AND JINCHAO XU‡

Abstract. In this paper, we develop and analyze a general approach to preconditioning linear systems of equations arising from conforming finite element discretizations of $\mathbf{H}(\mathbf{curl},\Omega)$- and $\mathbf{H}(\mathrm{div},\Omega)$-elliptic variational problems. The preconditioners exclusively rely on solvers for discrete Poisson problems. We prove mesh-independent effectivity of the preconditioners by using the abstract theory of auxiliary space preconditioning. The main tools are discrete analogues of so-called regular decomposition results in the function spaces $\mathbf{H}(\mathbf{curl},\Omega)$ and $\mathbf{H}(\mathrm{div},\Omega)$. Our preconditioner for $\mathbf{H}(\mathbf{curl},\Omega)$ is similar to an algorithm proposed in [R. Beck, *Algebraic Multigrid by Component Splitting for Edge Elements on Simplicial Triangulations*, Tech. rep. SC 99-40, ZIB, Berlin, Germany, 1999].

Key words. auxiliary space preconditioning, fictitious space preconditioning, $\mathbf{H}(\mathbf{curl})$ and $\mathbf{H}(\mathrm{div})$, edge and face finite elements, algebraic multigrid

AMS subject classifications. 65N22, 65F10, 65N30, 65N55

DOI. 10.1137/060660588

1. Introduction.

On a polyhedron Ω, scaled such that $\mathrm{diam}(\Omega) = 1$, we consider the variational problems

$$\mathbf{u} \in \mathcal{H}(\mathbf{curl}): \quad (\mathbf{curl}\,\mathbf{u}, \mathbf{curl}\,\mathbf{v})_0 + \tau\,(\mathbf{u},\mathbf{v})_0 = (\mathbf{f},\mathbf{v})_0 \quad \forall \mathbf{v} \in \mathcal{H}(\mathbf{curl},\Omega), \tag{1.1}$$

$$\mathbf{u} \in \mathcal{H}(\mathrm{div}): \quad (\mathrm{div}\,\mathbf{u}, \mathrm{div}\,\mathbf{v})_0 + \tau\,(\mathbf{u},\mathbf{v})_0 = (\mathbf{f},\mathbf{v})_0 \quad \forall \mathbf{v} \in \mathcal{H}(\mathrm{div},\Omega), \tag{1.2}$$

where $\mathbf{f}$ is a vector field in $(L^2(\Omega))^3$ and $\tau \geq 0$. We admit both homogeneous natural and essential boundary conditions; that is, $\mathcal{H}(\mathrm{div},\Omega)$ and $\mathcal{H}(\mathbf{curl},\Omega)$ can stand for either $\mathbf{H}(\mathbf{curl},\Omega)$ and $\mathbf{H}(\mathrm{div},\Omega)$ or $\mathbf{H}_0(\mathbf{curl},\Omega)$ and $\mathbf{H}_0(\mathrm{div},\Omega)$, respectively. The parameter τ controls the relative weight of the second and zero order terms in the bilinear forms.

More generally, the bilinear forms of (1.1) and (1.2) could feature spatially varying coefficients. So far, our theoretical analysis can take into account variations in the coefficients only very crudely. Thus, for the sake of simplicity, we have decided to focus on the constant coefficient case. Variable coefficients will be covered in some numerical experiments.

Variational problems of the form (1.1) and (1.2) arise in different applications, for instance, in

- (1.1) as a variational formulation of the eddy current model in computational electromagnetics [9], and
- (1.2) in the context of equivalent operator preconditioning for mixed finite element and first order system least squares (FOSLS) schemes for second order elliptic problems [3].

*Received by the editors May 23, 2006; accepted for publication (in revised form) May 7, 2007; published electronically November 28, 2007.
http://www.siam.org/journals/sinum/45-6/66058.html
†Seminar for Applied Mathematics, ETH Zürich, CH-8092 Zürich, Switzerland (hiptmair@sam.math.ethz.ch).
‡Mathematics Department, The Pennsylvania State University, University Park, PA 16802 (xu@math.psu.edu).

Geometric multigrid approaches to discrete linear problems arising from the Galerkin finite element discretization of (1.1) and (1.2) are well known [2, 4, 19, 22]. They supply mesh-independent iterative solvers and preconditioners, provided a hierarchy of uniformly shape regular meshes is available. Algebraic multigrid (AMG) methods that dispense with this requirement have been proposed in [7, 33], but they noticeably deteriorate on fine meshes, let alone permit a comprehensive theoretical analysis. The auxiliary space approach [38] allows us to harness powerful and asymptotically optimal AMG methods developed for discrete second order elliptic boundary value problems in order to get fast iterative solvers for discretized $\mathbf{H}(\mathbf{curl},\Omega)$- and $\mathbf{H}(\mathrm{div},\Omega)$-elliptic problems. As these auxiliary discrete second order elliptic boundary value problems arise from the use of Lagrangian finite elements, which are known as nodal finite elements in computational electromagnetism [10], we have tagged this special auxiliary space technique as *nodal*.

There is a close relationship between the variational problems (1.1) and (1.2) (cf. [20, section 2]) which allows a fairly parallel treatment of both. Thus we opt for a unified presentation, starting from an abstract variational problem

$$\mathbf{u}\in\mathcal{H}(\mathsf{D},\Omega):\quad a(\mathbf{u},\mathbf{v}):=(\mathsf{D}\mathbf{u},\mathsf{D}\mathbf{v})_0+\tau\,(\mathbf{u},\mathbf{v})_0=f(\mathbf{v})\quad\forall\mathbf{u},\mathbf{v}\in\mathcal{H}(\mathsf{D},\Omega), \tag{1.3}$$

where f is a continuous linear functional on the Hilbert space $\mathcal{H}(\mathsf{D},\Omega)$. Relating (1.3) to (1.1) and (1.2) discloses the meaning of D, f, and $\mathcal{H}(\mathsf{D},\Omega)$; see also Table 3.1. The bilinear form $a(\cdot,\cdot)$ induces the *energy norm*

$$\|\mathbf{v}\|_A^2:=a(\mathbf{v},\mathbf{v}),\quad \mathbf{v}\in\mathcal{H}(\mathsf{D},\Omega), \tag{1.4}$$

which is merely a seminorm, if $\tau=0$. The energy norm is closely related to the Hilbert space norm on $\mathcal{H}(\mathsf{D},\Omega)$

$$\|\mathbf{v}\|_{\mathcal{H}(\mathsf{D},\Omega)}^2:=\|\mathsf{D}\mathbf{v}\|_{L^2(\Omega)}^2+\|\mathbf{v}\|_{L^2(\Omega)}^2,\qquad \mathbf{v}\in\mathcal{H}(\mathsf{D},\Omega). \tag{1.5}$$

The principal challenge confronted in the development of preconditioners for discretized versions of (1.1) and (1.2) is the presence of a large kernel of D: in contrast to the case $\mathsf{D}=\mathbf{grad}$, these kernels have infinite dimension for $\mathsf{D}=\mathbf{curl}$ (comprising, e.g., all gradients) and $\mathsf{D}=\mathrm{div}$ (comprising, e.g., all rotations). This entails a separate treatment of these kernels by the preconditioner, which can exploit the fact that in suitable **curl**- and div-conforming finite element spaces the kernels possess a convenient representation through potentials. On the complement of the kernel the variational problem should display strong ellipticity and be amenable to standard elliptic preconditioning techniques; cf. the reasoning in [19, section 3] and [5, section 5].

Roughly speaking, on the complement of the kernels, the differential operators underlying (1.1) and (1.2) can be expected to be spectrally equivalent to a second order differential operator applied to the components of the vector fields. However, using a discrete second order differential operator as preconditioner is not possible immediately, because it does not fit the **curl**- and div-conforming finite element space. This is why we need the auxiliary space preconditioning technology [38] to link the finite element spaces on which (1.3) is discretized and the vectorial $H^1(\Omega)$-conforming finite elements that underlie the preconditioning operator.

The main rationale for pursuing this method in [5] was that the evaluation of the preconditioner boils down to inverting discrete scalar second order elliptic operators approximately. Fast AMG methods for this purpose abound; see [36, Appendix A]. Thus, AMG codes can be harnessed for $\mathbf{H}(\mathbf{curl},\Omega)$- and $\mathbf{H}(\mathrm{div},\Omega)$-elliptic problems.

The principal idea underlying our approach can be gleaned from the understanding that stable space decompositions are at the heart of preconditioners for symmetric positive definite variational problems; see section 2 for further explanations. Results that we have dubbed regular decompositions provide fundamental stable splittings for the spaces $\mathcal{H}(\mathsf{D},\Omega)$; see section 3 for details. For instance, for the space $\mathbf{H}_0(\mathbf{curl},\Omega)$ we have a stable splitting

$$\mathbf{H}_0(\mathbf{curl},\Omega) = (H^1_0(\Omega))^3 + \mathbf{grad}\, H^1_0(\Omega).$$

This suggests that a preconditioner for $\mathbf{H}_0(\mathbf{curl},\Omega)$-elliptic variational problem can be based on solving $H^1_0(\Omega)$-elliptic variational problems. However, to make this idea work, the splittings have to be adapted to the discrete setting. Thus, in sections 5 and 6 we establish stable discrete regular decompositions and corresponding norm equivalences. This yields the desired preconditioners, whose implementation will be discussed in section 7. In the end, we supplement the asymptotic theoretical estimate with numerical studies of the performance of the preconditioners. We refer the reader to [27,28] for more numerical results.

2. Auxiliary space preconditioning: Abstract theory. In this section, we give a self-contained description of preconditioning techniques based on fictitious or auxiliary spaces as developed in [18,31,38].

Let V stand for a real Hilbert space with inner product $a(\cdot,\cdot)$ and (energy) norm $\|\cdot\|_A$. The fictitious space method targets linear variational problems

$$u \in V: \quad a(u,v) = f(v) \quad \forall v \in V. \tag{2.1}$$

Its main building blocks are

1. a *fictitious space* $\bar{V}$, that is, another real Hilbert space equipped with the inner product $\overline{a}(\cdot,\cdot)$, which induces the norm $\|\cdot\|_{\bar{\mathsf{A}}}$, and
2. a continuous and surjective linear transfer operator $\Pi : \bar{V} \mapsto V$.

We tag dual spaces by $'$ and adjoint operators by *, and we use angle brackets for duality pairings. Then, writing $\mathsf{A} : V \mapsto V'$ and $\bar{\mathsf{A}} : \bar{V} \mapsto \bar{V}'$ for the isomorphisms associated with $a(\cdot,\cdot)$ and $\overline{a}(\cdot,\cdot)$, respectively, the fictitious space preconditioner is given by

$$\mathsf{B} = \Pi \circ \bar{\mathsf{A}}^{-1} \circ \Pi^* : V' \mapsto V. \tag{2.2}$$

Obviously, the operator B is associated with a symmetric bilinear form on the dual space V'. The next lemma confirms that this form is actually positive definite, which renders B a valid preconditioner.

LEMMA 2.1. *If $\Pi : \bar{V} \mapsto V$ is surjective, then the operator B is an isomorphism.*

Proof. Π being surjective means that it is an open mapping and Π^* is injective. As $\overline{a}$ is positive definite, we infer

$$\langle \varphi, \mathsf{B}\varphi \rangle_{V'\times V} = \left\langle \Pi^*\varphi, \bar{\mathsf{A}}^{-1}\Pi^*\varphi \right\rangle_{\bar{V}'\times\bar{V}} > 0 \quad \forall \varphi \in V' \setminus \{0\}.$$

From this we conclude the assertion of the theorem. ∎

The next theorem is known as the *fictitious space lemma* [31], for which we provide the elementary proof for the sake of completeness.

THEOREM 2.2 (fictitious space lemma). *Assume that Π is surjective and*

$$\exists c_0 > 0: \quad \forall v \in V: \quad \exists \bar{v} \in \bar{V}: \quad v = \Pi\bar{v} \quad \wedge \quad \|\bar{v}\|_{\bar{\mathsf{A}}} \le c_0 \|v\|_A, \tag{2.3}$$

$$\exists c_1 > 0: \quad \|\Pi\bar{v}\|_A \le c_1 \|\bar{v}\|_{\bar{\mathsf{A}}} \quad \forall \bar{v} \in \bar{V}. \tag{2.4}$$

Then

$$c_0^{-2}\,\|v\|_A^2 \le a(\mathsf{BA}v,v) \le c_1^2\,\|v\|_A^2 \quad \forall v\in V. \tag{2.5}$$

Proof. The proof makes use of only the Cauchy–Schwarz inequality:

$$\begin{aligned} a(\mathsf{BA}v,v) &\le a(\mathsf{BA}v,\mathsf{BA}v)^{1/2}a(v,v)^{1/2}\\ &= a(\Pi\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v,\Pi\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v)^{1/2}\,\|v\|_A\\ &\le c_1\,\overline{a}(\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v,\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v)^{1/2}\,\|v\|_A\\ &= c_1\,\left\langle\Pi^*\mathsf{A}v,\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v\right\rangle_{\bar V'\times\bar V}^{1/2}\,\|v\|_A = c_1\,a(\mathsf{BA}v,v)^{1/2}\,\|v\|_A\,. \end{aligned}$$

Next, we rely on the assumption (2.3) and get

$$\begin{aligned} a(v,v) &= \langle\mathsf{A}v,\Pi\bar v\rangle_{V'\times V} = \langle\Pi^*\mathsf{A}v,\bar v\rangle_{\bar V'\times\bar V} = \overline{a}(\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v,\bar v)\\ &\le \overline{a}(\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v,\bar{\mathsf{A}}^{-1}\Pi^*\mathsf{A}v)^{1/2}\,\|\bar v\|_{\bar{\mathsf{A}}} \le c_0\,a(\mathsf{BA}v,v)^{1/2}\,\|v\|_A\,. \end{aligned}$$

∎

From (2.5) we immediately get an estimate for the spectral condition number of the preconditioned operator A:

$$\kappa(\mathsf{BA}) := \frac{\lambda_{\max}(\mathsf{BA})}{\lambda_{\min}(\mathsf{BA})} \le (c_0c_1)^2. \tag{2.6}$$

Corollary 2.3. *Under the assumptions of the previous theorem, we have the following estimate for the spectral condition number:*

$$\kappa((\Pi\bar{\mathsf{B}}\Pi^*)\mathsf{A}) \le \kappa(\bar{\mathsf{B}}\bar{\mathsf{A}})(c_0c_1)^2,$$

where $\bar{\mathsf{B}}:\bar V'\mapsto\bar V$ *is supposed to be a preconditioner for* $\bar{\mathsf{A}}$.

Proof. This result is a consequence of the obvious inequality

$$\kappa((\Pi\bar B\Pi^*)\mathsf{A}) \le \kappa(\bar B\bar{\mathsf{A}})\kappa(\mathsf{BA}).$$ ∎

The auxiliary space method as pioneered in [38] can be viewed as a fictitious space approach relying on the special choice

$$\bar V = V\times W_1\times\cdots\times W_J, \tag{2.7}$$

where $W_1,\dots,W_J$, $J\in\mathbb{N}$, are Hilbert spaces endowed with inner products $\overline{a}_j(\cdot,\cdot)$, $j=1,\dots,J$. They provide the so-called auxiliary spaces.

A distinctive feature of the auxiliary space method is the presence of V in (2.7), but as a component of $\bar V$ the space V will be equipped with an inner product $s(\cdot,\cdot)$ different from $a(\cdot,\cdot)$. The operator $\mathsf{S}:V\mapsto V'$ induced by $s(\cdot,\cdot)$ on V is usually called the *smoother*. In other words, the auxiliary space method adopts the fictitious space approach with the inner product

$$\overline{a}(\bar v,\bar v) := s(v_0,v_0)+\sum_{j=1}^J\overline{a}_j(w_j,w_j), \quad \begin{array}{l}\bar v=(v_0,w_1,\dots,w_J)\in\bar V,\\ v_0\in V,\,w_j\in W_j.\end{array} \tag{2.8}$$

Furthermore, for each W_j we need a linear transfer operator $\Pi_j : W_j\mapsto V$, from which we build the surjective operator

$$\Pi := \begin{pmatrix} Id & & & \\ & \Pi_1 & & \\ & & \ddots & \\ & & & \Pi_J \end{pmatrix} : \bar V\mapsto V. \tag{2.9}$$

Now, all components of the auxiliary space preconditioner are in place and the formula (2.2) becomes

$$\mathsf{B} = \mathsf{S}^{-1} + \sum_{j=1}^{J} \Pi_j \circ \bar{\mathsf{A}}_j^{-1} \circ \Pi_j^*. \tag{2.10}$$

The verification of the assumptions of Theorem 2.2 for the preconditioner boils down to three steps.

1. Find bounds $c_j > 0$ for norms of the transfer operators Π_j:

$$\|\Pi_j w_j\|_A \le c_j \overline{a}(w_j, w_j)^{1/2}, \quad w_j \in W_j. \tag{2.11}$$

2. Investigate the continuity of S^{-1}:

$$\exists c_s > 0 : \quad \|v\|_A \le c_s s(v,v)^{1/2} \quad \forall v \in V. \tag{2.12}$$

3. Establish that for every $v \in V$ there are $v_0 \in V$ and $w_j \in W_j$ such that $v = v_0 + \sum_{j=1}^{J} \Pi_j w_j$ and

$$s(v_0, v_0) + \sum_{j=1}^{J} \overline{a}_j(w_j, w_j) \le c_0^2 \|v\|_A^2, \tag{2.13}$$

where $c_0 > 0$ should be small and independent of v.

Then, the assertion of Theorem 2.2 translates to

$$\kappa(\mathsf{BA}) \le c_0^2 (c_s^2 + c_1^2 + \cdots + c_J^2). \tag{2.14}$$

It goes without saying that in the spirit of Corollary 2.3, the bilinear forms $\overline{a}_j$ on the auxiliary spaces W_j can be replaced with spectrally equivalent bilinear forms $\bar{b}_j$; i.e., we may use preconditioners $\bar{\mathsf{B}}_j$ for the operators $\bar{\mathsf{A}}_j$. The impact of this approximation can be gauged as in Corollary 2.3.

In the applications we have in mind all the spaces will be finite element spaces and will feature bases comprised of locally supported functions. Plugging basis functions into the bilinear forms $a(\cdot,\cdot)$ and $\overline{a}_j(\cdot,\cdot)$, we obtain the Galerkin matrices $\mathbf{A} \in \mathbb{R}^{N,N}$, $N := \dim V$, $\overline{\mathbf{A}}_j \in \mathbb{R}^{N_j,N_j}$, $N_j := \dim W_j$. The smoother is provided by local relaxation procedures: for instance, if Jacobi smoothing is used, an algebraic representation of the associated operator S is given by the diagonal part $\mathbf{D}_A$ of the matrix $\mathbf{A}$. Hence, the algebraic version of the preconditioner from (2.10) reads

$$\mathbf{B} = \mathbf{D}_A^{-1} + \sum_{j=1}^{J} \mathbf{P}_j \overline{\mathbf{A}}_j^{-1} \mathbf{P}_j^T, \tag{2.15}$$

where $\mathbf{P}_j \in \mathbb{R}^{N,N_j}$ is the matrix representation of Π_j. When using symmetric Gauss–Seidel smoothing $\mathbf{D}_A^{-1}$ has to be replaced with $\mathbf{L}_A^{-1} + \mathbf{L}_A^{-T} - \mathbf{L}_A^{-1}\mathbf{A}\mathbf{L}_A^{-T}$, where $\mathbf{L}_A$ stands for the lower triangular part of (the symmetric matrix) $\mathbf{A}$.

Remark 1. Naturally, we can also try to apply the successive subspace correction idea [37] to the multiple auxiliary spaces to obtain the following iterative method for the operator equation $\mathsf{A}u = f$, $f \in V'$:

$$u \leftarrow u + \mathsf{S}^{-1}(f - \mathsf{A}u), \qquad u \leftarrow u + \Pi_j \bar{\mathsf{B}}_j \Pi_j^*(f - \mathsf{A}u), \quad 1 \le j \le J. \tag{2.16}$$

It is easy to see that a sufficient condition for the convergence of this successive auxiliary space method is

$$\lambda_{\max}(\bar{\mathsf{B}}_j \bar{\mathsf{A}}_j) \leq \frac{\omega}{c_j}, \qquad 1 \leq j \leq J, \tag{2.17}$$

for some $0 < \omega < 2$. Under the above conditions, the convergence rate of the iteration (2.16) depends only on c_0, c_s, c_j $(1 \leq j \leq J)$, ω, and J.

3. Regular decompositions. The abstract theory of the previous section has identified the uniform stability of decompositions (cf. (2.13)) as a key prerequisite of successful auxiliary space preconditioning. This connects well with the pivotal role of certain stable decomposition in the analysis of variational problems linked with $\mathbf{H}(\mathbf{curl}, \Omega)$ and $\mathbf{H}(\mathrm{div}, \Omega)$ [6,12,14]. In the subsequent discussion, to avoid topological obstructions, we restrict ourselves to "simple" domains.

ASSUMPTION 3.1. *We assume that Ω is homotopy equivalent to a ball.*

This makes it possible to use *potential representations* for the kernels of the differential operators.

LEMMA 3.1 (exact sequence property).

$$\textit{Assumption } 3.1 \quad \Rightarrow \quad \begin{array}{l} \mathbf{H}(\mathbf{curl}\,0, \Omega) := \{\mathbf{v} \in \mathbf{H}(\mathbf{curl}, \Omega) : \mathbf{curl}\,\mathbf{v} = 0\} = \mathbf{grad}\, H^1(\Omega), \\ \mathbf{H}_0(\mathbf{curl}\,0, \Omega) := \{\mathbf{v} \in \mathbf{H}_0(\mathbf{curl}, \Omega) : \mathbf{curl}\,\mathbf{v} = 0\} = \mathbf{grad}\, H^1_0(\Omega), \\ \mathbf{H}(\mathrm{div}\,0, \Omega) := \{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \Omega) : \mathrm{div}\,\mathbf{v} = 0\} = \mathbf{curl}\,\mathbf{H}(\mathbf{curl}, \Omega), \\ \mathbf{H}_0(\mathrm{div}\,0, \Omega) := \{\mathbf{v} \in \mathbf{H}_0(\mathrm{div}, \Omega) : \mathrm{div}\,\mathbf{v} = 0\} = \mathbf{curl}\,\mathbf{H}_0(\mathbf{curl}, \Omega). \end{array}$$

In the unifying framework, this lemma can be recast into

$$\text{Assumption } 3.1 \quad \Rightarrow \quad \mathcal{H}(\mathsf{D}0, \Omega) := \{\mathbf{v} \in \mathcal{H}(\mathsf{D}, \Omega) : \mathsf{D}\mathbf{v} = 0\} = \mathsf{D}^- \mathcal{H}(\mathsf{D}^-, \Omega), \tag{3.1}$$

where D^- is the differential operator characterizing the *potential* space $\mathcal{H}(\mathsf{D}^-, \Omega)$; see the "translation table" Table 3.1.

TABLE 3.1
Translation table for unifying notational framework, generic case.

D	$\mathcal{H}(\mathsf{D}, \Omega)$	D^-	$\mathcal{H}(\mathsf{D}^-, \Omega)$	D^+	$\boldsymbol{\mathcal{H}}^1$
grad	$H^1(\Omega)$ $H^1_0(\Omega)$	Id	$\{\mathrm{const}\}$ $\{0\}$	**curl**	$H^1(\Omega)$ $H^1_0(\Omega)$
curl	$\mathbf{H}(\mathbf{curl}, \Omega)$ $\mathbf{H}_0(\mathbf{curl}, \Omega)$	**grad**	$H^1(\Omega)$ $H^1_0(\Omega)$	div	$(H^1(\Omega))^3$ $(H^1_0(\Omega))^3$
div	$\mathbf{H}(\mathrm{div}, \Omega)$ $\mathbf{H}_0(\mathrm{div}, \Omega)$	**curl**	$\mathbf{H}(\mathbf{curl}, \Omega)$ $\mathbf{H}_0(\mathbf{curl}, \Omega)$	0	$(H^1(\Omega))^3$ $(H^1_0(\Omega))^3$

Remark 2. If Assumption 3.1 fails to hold, De Rham cohomology theory teaches that the potential representations will be available only up to contributions from cohomology spaces of a small and finite dimension. They will not matter much for the overall performance of a preconditioner, and so we decided to forgo a discussion of general topologies.

The starting point for the development of the auxiliary space preconditioners presented in this paper is theoretical results that, roughly speaking, state that the gap between $(H^1(\Omega))^3$ and $\mathcal{H}(\mathsf{D}, \Omega)$ can be bridged by contributions from the kernel of D. A rigorous statement is made by the following so-called regular decomposition

results. In light of the unified treatment we aim for, operators with similar function will be denoted alike, though they are different in **curl**- and div-contexts, respectively.

LEMMA 3.2 (existence of regular vector potentials [21, Lemma 2.5]). *There is a continuous mapping* $\mathsf{L} : \{\mathbf{v} \in \mathbf{H}(\mathrm{div}, \mathbb{R}^3), \mathrm{div}\,\mathbf{v} = 0\} \mapsto (H^1(\mathbb{R}^3))^3$ *such that* $\mathbf{curl}\,\mathsf{L}v = v$ *and* $\mathrm{div}\,\mathsf{L}v = 0$.

LEMMA 3.3 (regular decomposition of $\mathbf{H}(\mathbf{curl}, \Omega)$ [21, Lemma 2.4]). *There are continuous maps* $\mathsf{R} : \mathbf{H}(\mathbf{curl}, \Omega) \mapsto (H^1(\Omega))^3$, $\mathsf{Z} : \mathbf{H}(\mathbf{curl}, \Omega) \mapsto H^1(\Omega)$ *such that* $\mathsf{R} + \mathbf{grad} \circ \mathsf{Z} = \mathsf{Id}$ *on* $\mathbf{H}(\mathbf{curl}, \Omega)$ *and* $\mathsf{R}u = 0 \Leftrightarrow \mathbf{curl}\,u = 0$.

LEMMA 3.4 (regular decomposition of $\mathbf{H}_0(\mathbf{curl}, \Omega)$ [32, section 2]). *There are continuous linear operators* $\mathsf{R} : \mathbf{H}_0(\mathbf{curl}, \Omega) \mapsto (H^1_0(\Omega))^3$, $\mathsf{Z} : \mathbf{H}_0(\mathbf{curl}, \Omega) \mapsto H^1_0(\Omega)$ *such that* $\mathsf{R} + \mathbf{grad} \circ \mathsf{Z} = \mathsf{Id}$ *on* $\mathbf{H}_0(\mathbf{curl}, \Omega)$ *and* $\mathsf{R}u = 0 \Leftrightarrow \mathbf{curl}\,u = 0$.

COROLLARY 3.5. *Both operators* R *introduced in Lemmas* 3.3 *and* 3.4 *satisfy*

$$\exists C = C(\Omega) > 0 : \quad \|\mathsf{R}\mathbf{v}\|_{H^1(\Omega)} \leq C \|\mathbf{curl}\,\mathbf{v}\|_{L^2(\Omega)} \quad \forall \mathbf{v} \in \mathcal{H}(\mathbf{curl}).$$

LEMMA 3.6 (existence of regular velocity fields, [17, Corollary 2.4]). *There is a continuous linear operator* $\mathsf{K} : L^2_0(\Omega) := \{\mathbf{v} \in L^2(\Omega), \int_\Omega \mathbf{v}\,\mathrm{d}\boldsymbol{x} = 0\} \mapsto (H^1_0(\Omega))^3$ *such that* $\mathrm{div} \circ \mathsf{K} = \mathsf{Id}$ *on* $L^2_0(\Omega)$.

LEMMA 3.7 (regular decomposition of $\mathbf{H}(\mathrm{div}, \Omega)$). *There are continuous linear operators* $\mathsf{R} : \mathbf{H}(\mathrm{div}, \Omega) \mapsto (H^1(\Omega))^3$, $\mathsf{Z} : \mathbf{H}(\mathrm{div}, \Omega) \mapsto (H^1(\Omega))^3$ *such that* $\mathsf{R} + \mathbf{curl} \circ \mathsf{Z} = \mathsf{Id}$ *on* $\mathbf{H}(\mathrm{div}, \Omega)$ *and* $\mathsf{R}u = 0 \Leftrightarrow \mathrm{div}\,u = 0$.

Proof. For $u \in \mathbf{H}(\mathrm{div}, \Omega)$ perform a trivial extension by zero of $\mathrm{div}\,u$ to an element of $L^2(\mathbb{R}^3)$. By elementary Fourier transform techniques (see [17, section 3.3]) we establish the existence of $w \in \mathbf{H}(\mathrm{div}, \mathbb{R}^3)$ such that $\mathrm{div}\,w = \mathrm{div}\,u$ on Ω. Lemma 3.1 finishes the proof. ∎

LEMMA 3.8 (regular decomposition of $\mathbf{H}_0(\mathrm{div}, \Omega)$). *There are continuous linear operators* $\mathsf{R} : \mathbf{H}_0(\mathrm{div}, \Omega) \mapsto (H^1_0(\Omega))^3$, $\mathsf{Z} : \mathbf{H}_0(\mathrm{div}, \Omega) \mapsto (H^1_0(\Omega))^3$ *such that* $\mathsf{R} + \mathbf{curl} \circ \mathsf{Z} = \mathsf{Id}$ *on* $\mathbf{H}_0(\mathrm{div}, \Omega)$ *and* $\mathsf{R}u = 0 \Leftrightarrow \mathrm{div}\,u = 0$.

Proof. Observe that $\mathrm{div}\,\mathbf{H}_0(\mathrm{div}, \Omega) \subset L^2_0(\Omega)$ and use Lemmas 3.6 and 3.1. ∎

COROLLARY 3.9. *Both operators* R *introduced in Lemmas* 3.7 *and* 3.8 *satisfy*

$$\exists C = C(\Omega) > 0 : \quad \|\mathsf{R}\mathbf{v}\|_{H^1(\Omega)} \leq C \|\mathrm{div}\,\mathbf{v}\|_{L^2(\Omega)} \quad \forall \mathbf{v} \in \mathcal{H}(\mathrm{div}).$$

Using the operator symbols from Table 3.1, we can summarize the above assertions in the following lemma.

LEMMA 3.10 (stable regular decomposition).

$$\begin{array}{l} \exists \mathsf{R} \in L(\mathcal{H}(\mathsf{D}, \Omega), \boldsymbol{\mathcal{H}}^1), \\ \exists \mathsf{Z} \in L(\mathcal{H}(\mathsf{D}, \Omega), \mathcal{H}(\mathsf{D}^-, \Omega)), \\ \exists C = C(\Omega) > 0 \end{array} : \quad \begin{cases} \mathsf{R} + \mathsf{D}^- \circ \mathsf{Z} = \mathsf{Id} \quad \text{on } \mathcal{H}(\mathsf{D}, \Omega), \\ \|\mathsf{R}\mathbf{v}\|_{H^1(\Omega)} \leq C \|\mathsf{D}\mathbf{v}\|_{L^2(\Omega)} \quad \forall \mathbf{v} \in \mathcal{H}(\mathsf{D}, \Omega), \\ \|\mathsf{Z}\mathbf{v}\|_{\mathcal{H}(\mathsf{D}^-, \Omega)} \leq C \|\mathbf{v}\|_{\mathcal{H}(\mathsf{D}, \Omega)} \quad \forall \mathbf{v} \in \mathcal{H}(\mathsf{D}, \Omega). \end{cases}$$

Proof. The assertion about R rephrases those of Corollaries 3.5 and 3.9. Then

$$\begin{aligned} \left\|\mathsf{D}^-\mathsf{Z}\mathbf{v}\right\|_{L^2(\Omega)} &\leq \|\mathsf{R}\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{v}\|_{L^2(\Omega)} \\ &\leq C \|\mathsf{R}\mathbf{v}\|_{H^1(\Omega)} + \|\mathbf{v}\|_{L^2(\Omega)} \leq C \|\mathsf{D}\mathbf{v}\|_{L^2(\Omega)} + \|\mathbf{v}\|_{L^2(\Omega)}, \end{aligned}$$

and the estimate for the norm $\|\cdot\|_{\mathcal{H}(\mathsf{D}^-, \Omega)}$ can be inferred from the fact that $\mathsf{D}^- : \mathcal{H}(\mathsf{D}^-, \Omega) \mapsto L^2(\Omega)$ has closed range. ∎

In light of the fictitious space lemma Theorem 2.2, the result of Lemma 3.10 can be read as follows (we refer the reader to Table 3.1 for the meaning of $\boldsymbol{\mathcal{H}}^1$ and write

$(Id - \boldsymbol{\Delta}) : \boldsymbol{\mathcal{H}}^1 \mapsto (\boldsymbol{\mathcal{H}}^1)'$ for the (second order elliptic) operator associated with the inner product of $\boldsymbol{\mathcal{H}}^1$): the regular decomposition confirms that

$$\mathsf{B} := \mathsf{I} \circ (\mathsf{Id} - \boldsymbol{\Delta})^{-1} \circ \mathsf{I}^* + \mathsf{D}^- \circ (\mathsf{Id} + (\mathsf{D}^-)^* \mathsf{D}^-)^{-1} \circ (\mathsf{D}^-)^* \tag{3.2}$$

will supply a "preconditioner" for the operator $\mathsf{A} : \mathcal{H}(\mathsf{D}, \Omega) \mapsto \mathcal{H}(\mathsf{D}, \Omega)'$ induced by the bilinear form of (1.3). Here, I designates the trivial injection $\mathsf{I} : \boldsymbol{\mathcal{H}}^1 \mapsto \mathcal{H}(\mathsf{D}, \Omega)$ arising from the continuous embedding $\boldsymbol{\mathcal{H}}^1 \subset \mathcal{H}(\mathsf{D}, \Omega)$. Applying Theorem 2.2 and (2.6) and recalling that $\|\boldsymbol{\Psi}\|_{\mathcal{H}(\mathsf{D},\Omega)} \leq \|\boldsymbol{\Psi}\|_{H^1(\Omega)}$, $\boldsymbol{\Psi} \in \mathcal{H}(\mathsf{D}^-, \Omega)$, and $\|\mathsf{D}^- \varphi\|_{\mathcal{H}(\mathsf{D},\Omega)} \leq \|\varphi\|_{\mathcal{H}(\mathsf{D}^-,\Omega)}$, $\varphi \in \mathcal{H}(\mathsf{D}^-, \Omega)$, we readily conclude for $\tau = 1$

$$\kappa(\mathsf{BA}) \leq \|\mathsf{R}\|^2 + \|\mathsf{Z}\|^2 . \tag{3.3}$$

Of course, "preconditioning" an operator equation set in infinite-dimensional function spaces is hardly relevant for practical computations. Hence, the main objective of this paper is to get a discrete version of the above result.

Remark 3. All the above regular decompositions are global in the sense that the estimates of Corollaries 3.9 and 3.5 do not hold on subsets of Ω.

Remark 4. The regular decompositions outlined above have been widely used in functional analysis and numerical analysis connected with $\mathcal{H}(\mathsf{D}, \Omega)$. In the study of function spaces and traces, some of them first appeared in [6] and were later used in [12, 15]. They have found their way into the theoretical analysis of multilevel methods [22,26], domain decomposition methods [32], and boundary element methods [23].

A major shortcoming of the regular decompositions summarized in Lemma 3.10 is their lack of $L^2(\Omega)$-stability. This is obviously guaranteed in the case of the classical $L^2(\Omega)$-orthogonal Helmholtz decomposition

$$\mathcal{H}(\mathsf{D}, \Omega) = \mathcal{H}(\mathsf{D}0, \Omega) \oplus \mathcal{H}(\mathsf{D}0, \Omega)^{\perp}. \tag{3.4}$$

However, the $L^2(\Omega)$-orthogonal complement generally fails to belong to $H^1(\Omega)$ or $(H^1(\Omega))^3$, respectively. According to [17, section 3.4] this can only be taken for granted if $-\Delta$ with suitable homogeneous Dirichlet or Neumann boundary conditions is 2-*regular* on Ω (see [1] for details). We will refer to this situation as *the* 2-*regular case*. Conversely, Table 3.1 gives the meaning of the symbols in the *generic case*.

Remark 5. Convexity of Ω will ensure the 2-regular case. Moreover, for a convex Ω, formulas (2.8) and (2.3) in [16] involve the estimate

$$|\mathsf{R}\mathbf{v}|_{H^1(\Omega)} \leq \|\mathsf{D}\mathbf{v}\|_{L^2(\Omega)} , \tag{3.5}$$

where $|\cdot|_{H^1(\Omega)}$ designates the componentwise $H^1(\Omega)$-seminorm of a vector field.

However, use of the Helmholtz decomposition (3.4) entails relaxing the boundary conditions in $\mathcal{H}(\mathsf{D}^-, \Omega)$, when boundary conditions are imposed on $\mathcal{H}(\mathsf{D}, \Omega)$. More precisely, in the 2-regular case the following slightly modified meanings of the notation from Table 3.1 will be assumed. They are given in Table 3.2, where

$$\mathbf{H}^1_t(\Omega) := (H^1(\Omega))^3 \cap \mathbf{H}_0(\mathbf{curl}, \Omega), \qquad \mathbf{H}^1_n(\Omega) := (H^1(\Omega))^3 \cap \mathbf{H}_0(\mathrm{div}, \Omega).$$

With (3.4) in mind, in the 2-regular case, the operators Z and R from Lemma (3.10) can be chosen to satisfy

$$\left\|\mathsf{D}^- \mathsf{Z}\mathbf{v}\right\|^2_{L^2(\Omega)} + \|\mathsf{R}\mathbf{v}\|^2_{L^2(\Omega)} = \|\mathbf{v}\|^2_{L^2(\Omega)} \quad \forall \mathbf{v} \in \mathcal{H}(\mathsf{D}, \Omega). \tag{3.6}$$

TABLE 3.2
Symbols with altered meaning in the 2-regular case.

D	$\mathcal{H}(\mathsf{D},\Omega)$	D^-	$\mathcal{H}(\mathsf{D}^-,\Omega)$	D^+	$\boldsymbol{\mathcal{H}}^1$
curl	$\mathbf{H}_0(\mathbf{curl},\Omega)$	**grad**	$H^1_0(\Omega)$	div	$\mathbf{H}^1_t(\Omega)$
div	$\mathbf{H}_0(\mathrm{div},\Omega)$	**curl**	$\mathbf{H}_0(\mathbf{curl},\Omega)$	0	$\mathbf{H}^1_n(\Omega)$

In particular, Z and $\mathsf{D}^- \circ \mathsf{Z}$ turn out to be the $L^2(\Omega)$-orthogonal projections parallel to $\mathcal{H}(\mathsf{D}0,\Omega)$ and onto $\mathcal{H}(\mathsf{D}0,\Omega)$, respectively. The estimates of Corollaries 3.9 and 3.5 remain valid [1, section 2]. Thus, in the 2-regular case, Lemma 3.10 still holds with operators Z and R satisfying (3.6).

Remark 6. If Ω is a *polyhedron* (i.e., has flat faces), then we learn from Theorem 2.3 in [16] that

$$\|\mathbf{curl}\,\mathbf{u}\|^2_{L^2(\Omega)} + \|\mathrm{div}\,\mathbf{u}\|^2_{L^2(\Omega)} = \|\mathbf{grad}\,\mathbf{u}\|^2_{L^2(\Omega)} \quad \forall \mathbf{u} \in \mathbf{H}^1_t(\Omega) \cup \mathbf{H}^1_n(\Omega). \tag{3.7}$$

Whenever the definition of R is based on the Helmholtz decomposition it will map into either $\mathbf{H}^1_t(\Omega)$ or $\mathbf{H}^1_n(\Omega)$ in the 2-regular case. Thus, we conclude that on a polyhedron in the 2-regular case

$$\|\mathsf{R}\mathbf{v}\|_{H^1(\Omega)} \le \|\mathbf{v}\|_{\mathcal{H}(\mathsf{D},\Omega)} \quad \forall \mathbf{v} \in \mathcal{H}(\mathsf{D},\Omega) \quad \Rightarrow \quad \|\mathsf{R}\| = 1. \tag{3.8}$$

4. Finite element spaces. Essentially, the analysis of this paper applies to all the finite element subspaces of $\mathbf{H}(\mathbf{curl},\Omega)$ and $\mathbf{H}(\mathrm{div},\Omega)$ that can be viewed as discrete differential forms. This includes the so-called first and second families of edge elements [29, 30] and Raviart–Thomas elements and the Brezzi–Douglas–Marini (BDM) elements [11, Chapter 4]. To keep the presentation focused, we discuss only the lowest order cases.

Examples for the lowest order $\mathcal{H}(\mathsf{D},\Omega)$-conforming finite element spaces on a tetrahedral mesh $\mathcal{T}_h$ of Ω are listed in Table 4.1. They can be defined by

$$V_h(\mathbf{grad}) := \left\{ v_h \in \begin{matrix} H^1(\Omega) \\ H^1_0(\Omega) \end{matrix} \;:\; v_{h|K}(\mathbf{x}) = a + \mathbf{b}\cdot\mathbf{x},\, a \in \mathbb{R}, \mathbf{b} \in \mathbb{R}^3,\ \forall K \in \mathcal{T}_h \right\},$$

$$V_h(\mathbf{curl}) := \left\{ \mathbf{v}_h \in \begin{matrix} \mathbf{H}(\mathbf{curl},\Omega) \\ \mathbf{H}_0(\mathbf{curl},\Omega) \end{matrix} \;:\; \mathbf{v}_{h|K}(\mathbf{x}) = \mathbf{a} + \mathbf{x}\times\mathbf{b},\, \mathbf{a},\mathbf{b} \in \mathbb{R}^3,\ \forall K \in \mathcal{T}_h \right\},$$

$$V_h(\mathrm{div}) := \left\{ \mathbf{v}_h \in \begin{matrix} \mathbf{H}(\mathrm{div},\Omega) \\ \mathbf{H}_0(\mathrm{div},\Omega) \end{matrix} \;:\; \mathbf{v}_{h|K}(\mathbf{x}) = \mathbf{a} + \beta\mathbf{x},\, \mathbf{a} \in \mathbb{R}^3, \beta \in \mathbb{R}, \forall K \in \mathcal{T}_h \right\}.$$

$$V_h(0) := \left\{ v_h \in \begin{matrix} L^2(\Omega) \\ L^2_0(\Omega) \end{matrix} \;:\; v_{h|K}(\mathbf{x}) = a,\, a \in \mathbb{R},\ \forall K \in \mathcal{T}_h \right\}.$$

For a thorough discussion the reader is referred to [21, Chapter 3] and [29]. Resorting to a unified notation, we use the symbol $V_h(\mathsf{D})$ for these spaces. Its concrete meaning in different contexts is specified in Table 4.1.

A fundamental property of these families of finite element spaces is that they permit a discrete counterpart of (3.1):

$$\text{Assumption 3.1} \quad \Rightarrow \quad V_h(\mathsf{D}0) := \{\mathbf{v}_h \in V_h(\mathsf{D}) :\ \mathsf{D}\mathbf{v}_h = 0\} = \mathsf{D}^- V_h(\mathsf{D}^-). \tag{4.1}$$

TABLE 4.1
Finite element spaces of Whitney forms.

D	$\mathcal{H}(\mathsf{D},\Omega)$	$V_h(\mathsf{D}) \subset \mathcal{H}(\mathsf{D},\Omega)$	FE space	Reference
grad	$H^1(\Omega)$ $H^1_0(\Omega)$	$V_h(\mathbf{grad})$	linear Lagrangian FE	[13]
curl	$\mathbf{H}(\mathbf{curl},\Omega)$ $\mathbf{H}_0(\mathbf{curl},\Omega)$	$V_h(\mathbf{curl})$	edge elements	[29]
div	$\mathbf{H}(\mathrm{div},\Omega)$ $\mathbf{H}_0(\mathrm{div},\Omega)$	$V_h(\mathrm{div})$	face elements	[29]
0	$L^2(\Omega)$ $L^2_0(\Omega)$	$V_h(0)$	p.w. constants	

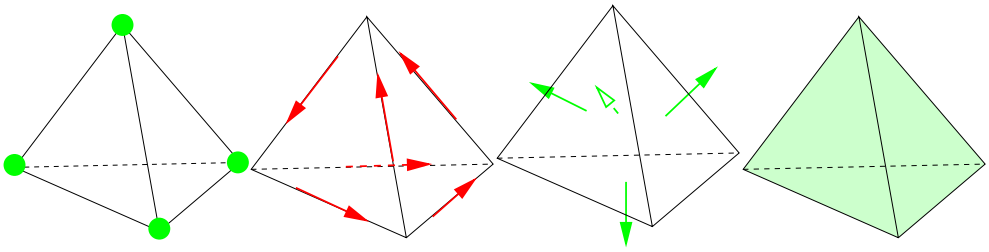


FIG. 4.1. *Symbolic notation for local degrees of freedom for* $V_h(\mathbf{grad}), V_h(\mathbf{curl}), V_h(\mathrm{div})$*, and* $V_h(0)$ *(left to right).*

These discrete potentials can even be chosen in a stable manner: with constants depending only on Ω, D, and the shape regularity of $\mathcal{T}_h$,

(4.2)
$$\forall \mathbf{v}_h \in V_h(\mathsf{D}0): \quad \exists p_h \in V_h(\mathsf{D}^-): \quad \mathbf{v}_h = \mathsf{D}^- p_h \quad \text{and} \quad \|p_h\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{L^2(\Omega)}.$$

For face elements this is a consequence of discrete Poincaré-type inequalities for $V_h(\mathbf{curl})$; see [21, Theorem 4.7]. For $\mathsf{D} = \mathbf{curl}$ and $\mathsf{D}^- = \mathbf{grad}$, (4.2) is just standard Poincaré–Friedrichs inequalities in $H^1(\Omega)/\mathbb{R}$ and $H^1_0(\Omega)$, respectively.

All the finite element spaces $V_h(\mathsf{D})$ are equipped with bases $\mathcal{B}(\mathsf{D})$ comprising locally supported functions; see [21, section 3.2]. These bases are L^2-stable in the sense that

(4.3)
$$\mathbf{v}_h = \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \mathbf{v}_\mathbf{b},\ \mathbf{v}_\mathbf{b} \in \mathrm{span}\{\mathbf{b}\}, \qquad \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v}_\mathbf{b}\|^2_{L^2(\Omega)} \approx \|\mathbf{v}_h\|^2_{L^2(\Omega)} \quad \forall \mathbf{v}_h \in V_h(\mathsf{D}),$$

with constants[1] depending only on the shape regularity of $\mathcal{T}_h$; see [24, section 2].

The spaces $V_h(\mathsf{D})$ also feature idempotent *nodal interpolation operators* Π^D_h whose range is $V_h(\mathsf{D})$. In the case $\mathsf{D} = \mathbf{grad}$ this is plain linear interpolation. For $\mathsf{D} = \mathbf{curl}$

[1] By the symbols $\approx$, $\lesssim$, and $\gtrsim$ we designate two- or one-sided inequalities, respectively, that hold up to multiplication of one side with a positive constant. In inequalities involving norms on function spaces this constant must not depend on the choice of functions. It may not depend on other problem and discretization parameters, and this will always be made clear.

the interpolation is based on path integrals along edges

$$\Pi_h^{\mathbf{curl}}\mathbf{v} = \sum_{e\in\mathcal{E}_h}\int_e \mathbf{v}\cdot \mathrm{d}\vec{s}\cdot \mathbf{b}_e, \tag{4.4}$$

where $\mathcal{E}_h$ is the set of (interior) edges of $\mathcal{T}_h$ and $\mathbf{b}_e$ is the edge element basis function associated with the edge e. For $\mathsf{D} = \mathrm{div}$, the interpolation relies on face fluxes:

$$\Pi_h^{\mathrm{div}}\mathbf{v} = \sum_{f\in\mathcal{F}_h}\int_f \mathbf{v}\cdot \mathrm{d}S\cdot \mathbf{b}_f, \tag{4.5}$$

with $\mathcal{F}_h$ designating the set of (interior) faces of $\mathcal{T}_h$. The relevant domains of integration for interpolation are depicted in Figure 4.1. Finally, the "interpolation" onto $V_h(0)$ agrees with $L^2(\Omega)$-projection. All these operators are well defined for continuous functions/vector fields, unbounded on $\mathcal{H}(\mathsf{D},\Omega)$ (except for $V_h(0)$), and possess the exceptional *commuting diagram property*

$$\mathsf{D}\circ\Pi_h^{\mathsf{D}} = \Pi_h^{+}\circ\mathsf{D} \quad \text{on domain of } \Pi_h^{\mathsf{D}}, \quad \Pi_h^{+} := \Pi_h^{\mathsf{D}^+}. \tag{4.6}$$

A concise way of writing (4.1) and (4.6) is through combined exact sequences and commuting diagrams:

$$\begin{array}{ccccccccccc}
0 & \longrightarrow & C^\infty(\Omega) & \xrightarrow{\mathbf{grad}} & (C^\infty(\Omega))^3 & \xrightarrow{\mathbf{curl}} & (C^\infty(\Omega))^3 & \xrightarrow{\mathrm{div}} & C^\infty(\Omega) & \longrightarrow & 0\\
 & & \downarrow{\scriptstyle \Pi_h^{\mathbf{grad}}} & & \downarrow{\scriptstyle \Pi_h^{\mathbf{curl}}} & & \downarrow{\scriptstyle \Pi_h^{\mathrm{div}}} & & \downarrow{\scriptstyle \Pi_h^{0}} & & \\
0 & \longrightarrow & V_h(\mathbf{grad}) & \xrightarrow{\mathbf{grad}} & V_h(\mathbf{curl}) & \xrightarrow{\mathbf{curl}} & V_h(\mathrm{div}) & \xrightarrow{\mathrm{div}} & V_h(0) & \longrightarrow & 0.
\end{array}$$

We write $h\in L^\infty(\Omega)$ for the piecewise constant meshwidth function, which assumes value $h_{|K} := \mathrm{diam}(K)$ in each cell K of $\mathcal{T}_h$. Using this function, we can state the following interpolation error estimate (see [21, section 3.6] and [21, Lemma 4.6]).

LEMMA 4.1. *The interpolation operators Π_h^{D} are bounded on $\{\mathbf{v}\in\boldsymbol{\mathcal{H}}^1,\ \mathsf{D}\mathbf{v}\in V_h(\mathsf{D}^+)\}\subset\boldsymbol{\mathcal{H}}^1$ and, with constants merely depending on D and the shape regularity of $\mathcal{T}_h$, they satisfy*

$$\left\|h^{-1}(Id-\Pi_h^{\mathsf{D}})\boldsymbol{\Psi}\right\|_{L^2(\Omega)} \lesssim \|\boldsymbol{\Psi}\|_{H^1(\Omega)} \quad \forall\boldsymbol{\Psi}\in\boldsymbol{\mathcal{H}}^1,\ \mathsf{D}\mathbf{v}\in V_h(\mathsf{D}^+). \tag{4.7}$$

Simple affine equivalence techniques also yield the inverse estimate

$$\|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \left\|h^{-1}\mathbf{v}_h\right\|_{L^2(\Omega)} \quad \forall\mathbf{v}\in V_h(\mathsf{D}), \tag{4.8}$$

with a constant depending only on D and the shape regularity of the mesh.

The role of the discrete auxiliary space will be played by the finite element space $S_h\subset\boldsymbol{\mathcal{H}}^1$ of continuous functions or vector fields, whose Cartesian components are piecewise linear. We point out that

$$\mathsf{D}S_h\subset\mathsf{D}V_h(\mathsf{D}), \quad \mathsf{D}\in\{\mathbf{grad},\mathbf{curl},\mathrm{div}\}. \tag{4.9}$$

Thanks to the commuting diagram property, we immediately conclude

$$\mathsf{D}\Pi_h^{\mathsf{D}}\boldsymbol{\Psi}_h = \Pi_h^{+}\mathsf{D}\boldsymbol{\Psi}_h = \mathsf{D}\boldsymbol{\Psi}_h \quad \forall\boldsymbol{\Psi}_h\in S_h. \tag{4.10}$$

Moreover, straightforward scaling arguments bear out that, with constants depending only on the shape regularity of $\mathcal{T}_h$,

$$\left\|\mathsf{D}\Pi_h^{\mathsf{D}}\boldsymbol{\Psi}_h\right\|_{L^2(\Omega)} \lesssim |\boldsymbol{\Psi}_h|_{H^1(\Omega)}, \quad \left\|\Pi_h^{\mathsf{D}}\boldsymbol{\Psi}_h\right\|_{L^2(\Omega)} \lesssim \|\boldsymbol{\Psi}_h\|_{L^2(\Omega)} \quad \forall \boldsymbol{\Psi}_h \in S_h. \tag{4.11}$$

Finally, we recall the surjective and idempotent quasi-interpolation operators for Lagrangian finite element spaces introduced in [34]. We may apply them to the components of vector fields separately. In the generic case (see Table 3.1), this gives rise to the projectors $\widetilde{\mathsf{Q}}_h : \boldsymbol{\mathcal{H}}^1 \mapsto S_h$, which inherit the continuity

$$\left\|\widetilde{\mathsf{Q}}_h\boldsymbol{\Psi}\right\|_{H^1(\Omega)} \lesssim \|\boldsymbol{\Psi}\|_{H^1(\Omega)} \quad \forall \boldsymbol{\Psi} \in \boldsymbol{\mathcal{H}}^1, \tag{4.12}$$

respect possible boundary values in the sense that $\widetilde{\mathsf{Q}}_h(H_0^1(\Omega))^3 \subset (H_0^1(\Omega))^3$, and satisfy the local projection error estimate

$$\left\|h^{-1}(\widetilde{\mathsf{Q}}_h - Id)\boldsymbol{\Psi}\right\|_{L^2(\Omega)} \lesssim \|\boldsymbol{\Psi}\|_{H^1(\Omega)} \quad \forall \boldsymbol{\Psi} \in \boldsymbol{\mathcal{H}}^1. \tag{4.13}$$

In the 2-regular case (see the end of section 3 for a discussion and Table 3.2 for the slightly changed meanings of symbols), we will replace $\widetilde{\mathsf{Q}}_h$ with the $L^2(\Omega)$-orthogonal projections $\mathsf{Q}_h : (L^2(\Omega))^3 \mapsto S_h$. From interpolation arguments we readily infer the estimate

$$\left\|h^{-1}(\mathsf{Q}_h - Id)\boldsymbol{\Psi}\right\|_{L^2(\Omega)} \lesssim \|\boldsymbol{\Psi}\|_{H^1(\Omega)} \quad \forall \boldsymbol{\Psi} \in \boldsymbol{\mathcal{H}}^1, \tag{4.14}$$

but this time, in contrast to (4.13), the constants will *also* depend on the quasi-uniformity of the mesh. So, h in (4.14) should be read as the global meshwidth of $\mathcal{T}_h$. The approximation property also involves the continuity

$$\|\mathsf{Q}_h\boldsymbol{\Psi}\|_{H^1(\Omega)} \lesssim \|\boldsymbol{\Psi}\|_{H^1(\Omega)} \quad \forall \boldsymbol{\Psi} \in \boldsymbol{\mathcal{H}}^1, \tag{4.15}$$

again, with quasi-uniformity of the mesh also entering the constants.

Summing up, whenever we can take quasi-uniformity of the meshes for granted, Q_h can replace $\widetilde{\mathsf{Q}}_h$ with the extra benefit of $L^2(\Omega)$-continuity.

5. Discrete regular decompositions. Now, following [25], let us derive a discrete version of the above regular decomposition results of section 3. First, we focus on the generic case; see Table 3.1. We fix a $\mathbf{v}_h \in V_h(\mathsf{D})$ and use the stable regular decomposition of Lemma 3.10 to split it according to

$$\mathbf{v}_h = \boldsymbol{\Psi} + \mathsf{D}^- p, \quad \boldsymbol{\Psi} := \mathsf{R}\mathbf{v}_h \in \boldsymbol{\mathcal{H}}^1, \quad p := \mathsf{Z}\mathbf{v}_h \in \mathcal{H}(\mathsf{D}^-, \Omega). \tag{5.1}$$

We already know that the functions $\boldsymbol{\Psi}$ and p satisfy

$$\|\boldsymbol{\Psi}\|_{H^1(\Omega)} \lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}, \qquad \left\|\mathsf{D}^- p\right\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{\mathcal{H}(\mathsf{D},\Omega)}, \tag{5.2}$$

with constants depending only on Ω.

So far, (5.1) is useless in the context of practical fictitious space preconditioning, because both $\boldsymbol{\Psi}$ and p fail to be finite element functions. The challenge is to convert (5.1) into a purely discrete decomposition without squandering the stability expressed by (5.2). This can be achieved only by incorporating another "high-frequency" contribution. Eventually, that forces us to incorporate a smoothing procedure into the algorithm.

LEMMA 5.1. *For any* $\mathbf{v}_h$ *there is* $\mathbf{\Psi}_h \in S_h$, $p_h \in V_h(\mathsf{D}^-)$, *and* $\widetilde{\mathbf{v}}_h \in V_h(\mathsf{D})$ *such that*

$$\mathbf{v}_h = \widetilde{\mathbf{v}}_h + \Pi_h^{\mathsf{D}} \mathbf{\Psi}_h + \mathsf{D}^- p_h, \tag{5.3}$$

and

$$\left\| h^{-1} \widetilde{\mathbf{v}}_h \right\|_{L^2(\Omega)}^2 + \|\mathbf{\Psi}_h\|_{H^1(\Omega)}^2 \lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}^2, \qquad \|p_h\|_{\mathcal{H}(\mathsf{D}^-,\Omega)} \lesssim \|\mathbf{v}_h\|_{\mathcal{H}(\mathsf{D},\Omega)}. \tag{5.4}$$

The constants are allowed to depend on Ω *and the shape regularity of the mesh.*

Proof. First, note that in (5.1) $\mathsf{D}\mathbf{\Psi} = \mathsf{D}\mathbf{v}_h \in V_h(\mathsf{D}^+)$, and, owing to Lemma 4.1, $\Pi_h^{\mathsf{D}} \mathbf{\Psi}$ is well defined. Further, the commuting diagram property implies

$$\mathsf{D}\Pi_h^{\mathsf{D}} \mathbf{\Psi} = \Pi_h^+ \mathsf{D}\mathbf{\Psi} = \mathsf{D}\mathbf{\Psi} \quad \Rightarrow \quad \mathsf{D}(Id - \Pi_h^{\mathsf{D}})\mathbf{\Psi} = 0. \tag{5.5}$$

This confirms that the third term in the splitting

$$\mathbf{\Psi} = \Pi_h^{\mathsf{D}}(\mathbf{\Psi} - \widetilde{\mathsf{Q}}_h \mathbf{\Psi}) + \Pi_h^{\mathsf{D}} \widetilde{\mathsf{Q}}_h \mathbf{\Psi} + (Id - \Pi_h^{\mathsf{D}})\mathbf{\Psi} \tag{5.6}$$

actually belongs to the kernel of D. By (4.1), we conclude

$$\exists q \in \mathcal{H}(\mathsf{D}^-, \Omega): \quad (Id - \Pi_h^{\mathsf{D}})\mathbf{\Psi} = \mathsf{D}^- q, \tag{5.7}$$

and (4.7) together with (5.2) yields

$$\left\| h^{-1} \mathsf{D}^- q \right\|_{L^2(\Omega)} = \left\| h^{-1} (Id - \Pi_h^{\mathsf{D}})\mathbf{\Psi} \right\|_{L^2(\Omega)} \lesssim \|\mathbf{\Psi}\|_{H^1(\Omega)} \lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}. \tag{5.8}$$

Thus, we can define the terms in the decomposition (5.3) as

$$\widetilde{\mathbf{v}}_h := \Pi_h^{\mathsf{D}}(\mathbf{\Psi} - \widetilde{\mathsf{Q}}_h \mathbf{\Psi}) \in V_h(\mathsf{D}), \tag{5.9}$$

$$\mathbf{\Psi}_h := \widetilde{\mathsf{Q}}_h \mathbf{\Psi} \in S_h, \tag{5.10}$$

$$\mathsf{D}^- p_h := \mathsf{D}^-(p + q), \quad p_h \in V_h(\mathsf{D}^-). \tag{5.11}$$

Indeed, $\mathsf{D}^-(p+q) \in V_h(\mathsf{D})$ such that we can add a contribution from $\mathcal{H}(\mathsf{D}^-, \Omega)$ to $p + q$ and obtain a discrete function. Thanks to (4.2) we can guarantee that $\|p_h\|_{L^2(\Omega)} \lesssim \|\mathsf{D}^- p_h\|_{L^2(\Omega)}$; this will not affect the decomposition. The stability of the decomposition (5.3) can be established as follows: first, make use of Lemma 4.1 and (4.13) to obtain

$$\begin{aligned}
\left\| h^{-1} \widetilde{\mathbf{v}}_h \right\|_{L^2(\Omega)} &\leq \left\| h^{-1} (Id - \Pi_h^{\mathsf{D}})(\mathbf{\Psi} - \widetilde{\mathsf{Q}}_h \mathbf{\Psi}) \right\|_{L^2(\Omega)} + \left\| h^{-1} (Id - \widetilde{\mathsf{Q}}_h)\mathbf{\Psi} \right\|_{L^2(\Omega)} \\
&\lesssim \left\| (Id - \widetilde{\mathsf{Q}}_h)\mathbf{\Psi} \right\|_{H^1(\Omega)} + \|\mathbf{\Psi}\|_{H^1(\Omega)} \\
&\lesssim \|\mathbf{\Psi}\|_{H^1(\Omega)} \lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}.
\end{aligned}$$

Due to the definition (5.10), the next estimate is a simple consequence of (4.12) and Lemma 3.10:

$$\|\mathbf{\Psi}_h\|_{H^1(\Omega)} \lesssim \|\mathbf{\Psi}\|_{H^1(\Omega)} \lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}. \tag{5.12}$$

Finally, the estimates established so far plus the triangle inequality yield

$$\left\| \mathsf{D}^- p_h \right\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{L^2(\Omega)} + \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}. \tag{5.13}$$

Owing to the discrete Poincaré–Friedrichs inequality (4.2), this implies (5.4). ∎

Remark 7. It is worth noting (see Table 3.1) that homogeneous boundary conditions imposed on $\mathcal{H}(\mathsf{D},\Omega)$ permit us to choose $S_h \subset (H_0^1(\Omega))^3$. This means that $\mathbf{\Psi}_h$ will completely vanish on $\partial\Omega$, though it is only the tangential or normal components of $\mathbf{v}_h$, respectively, that vanish on $\partial\Omega$.

LEMMA 5.2. *In the* 2*-regular case (see section* 3*, Table* 3.2*) the splitting* (5.3) *from Lemma* 5.1 *can be chosen such that, in addition to the estimates asserted in Lemma* 5.1*, we have*

$$\left\|\mathsf{D}^- p_h\right\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{L^2(\Omega)}\,, \qquad \|\mathbf{\Psi}_h\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{L^2(\Omega)}\,,$$

with constants additionally depending on the quasi-uniformity of the mesh $\mathcal{T}_h$*.*

Proof. We rely on the $L^2(\Omega)$-orthogonal Helmholtz decomposition (3.4) to define p and $\mathbf{\Psi}$ in (5.1). Consequently, we can expect

$$\|\mathbf{\Psi}\|_{L^2(\Omega)} \le \|\mathbf{v}_h\|_{L^2(\Omega)}\,, \qquad \left\|\mathsf{D}^- p\right\|_{L^2(\Omega)} \le \|\mathbf{v}_h\|_{L^2(\Omega)}\,. \tag{5.14}$$

Then follow the proof of Lemma 5.1 and replace the quasi-interpolation $\widetilde{\mathsf{Q}}_h$ with the $L^2(\Omega)$-orthogonal projection Q_h. Taking into account that h designates a global meshwidth when the constants are allowed to depend on quasi-uniformity, a glance at (4.14) and (4.15) confirms that all estimates of Lemma 5.1 remain true.

The replacement of $\widetilde{\mathsf{Q}}_h$ is necessary, because $\widetilde{\mathsf{Q}}_h$ fails to be continuous with respect to the $L^2(\Omega)$-norm. When using Q_h instead, we arrive at the trivial estimate

$$\mathbf{\Psi}_h := \mathsf{Q}_h\mathbf{\Psi} \quad \Rightarrow \quad \|\mathbf{\Psi}_h\|_{L^2(\Omega)} \le \|\mathbf{\Psi}\|_{L^2(\Omega)} \le \|\mathbf{v}_h\|_{L^2(\Omega)}\,. \tag{5.15}$$

In addition, use the interpolation estimate (5.8) and the inverse inequality (4.8):

$$\left\|\mathsf{D}^- q\right\|_{L^2(\Omega)} \lesssim h\,\|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)} \lesssim \|\mathbf{v}_h\|_{L^2(\Omega)}\,. \tag{5.16}$$

Again, h denotes the (global) meshwidth of $\mathcal{T}_h$. Owing to (5.11) and (5.14), this finishes the proof. ∎

6. Stable splittings. We first discuss the case $0 < \tau \le 1$; that is, the second order term in the bilinear form is dominant. The notation refers to the generic case of Table 3.1.

THEOREM 6.1. *Assume* $0 < \tau \le 1$*. For any* $\mathbf{v}_h \in V_h(\mathsf{D})$ *there is* $p_h \in V_h(\mathsf{D}^-)$*,* $\mathbf{\Psi}_h \in S_h$ *such that, when* $\mathbf{v}_\mathbf{b} \in \operatorname{span}\{\mathbf{b}\}$*,* $\mathbf{b} \in \mathcal{B}(\mathsf{D})$*, a locally supported basis function,*

$$\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \mathbf{v}_\mathbf{b} + \Pi_h^{\mathsf{D}}\mathbf{\Psi}_h + \mathsf{D}^- p_h = \mathbf{v}_h, \tag{6.1}$$

$$\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v}_\mathbf{b}\|_A^2 + \|\mathbf{\Psi}_h\|_{H^1(\Omega)}^2 + \left\|\mathsf{D}^- p_h\right\|_A^2 \lesssim \|\mathbf{v}_h\|_A^2\,, \tag{6.2}$$

with a constant depending only on Ω*,* D*, and the shape regularity of* $\mathcal{T}_h$*, but independent of* $\tau \in\,]0,1]$ *and quasi-uniformity.*

Proof. The contributions $\mathbf{\Psi}_h$ and p_h are chosen as in Lemma 5.1. Hence, reusing the notation of (5.3),

$$\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \mathbf{v}_\mathbf{b} = \widetilde{\mathbf{v}}_h. \tag{6.3}$$

By the inverse estimate (4.8), (4.3), and the bound for $\|h^{-1}\widetilde{\mathbf{v}}_h\|$ from Lemma 5.1

$$\begin{aligned}\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v_b}\|_A^2 &= \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathsf{D}\mathbf{v_b}\|_{L^2(\Omega)}^2 + \tau \|\mathbf{v_b}\|_{L^2(\Omega)}^2 \\ &\lesssim \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \left\|h^{-1}\mathbf{v_b}\right\|_{L^2(\Omega)}^2 + \tau \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v_b}\|_{L^2(\Omega)}^2 \\ &\lesssim \left\|h^{-1}\widetilde{\mathbf{v}}_h\right\|_{L^2(\Omega)}^2 + \tau \|\widetilde{\mathbf{v}}_h\|_{L^2(\Omega)}^2 \lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}^2 + \tau \|\mathbf{v}_h\|_{L^2(\Omega)}^2 .\end{aligned} \tag{6.4}$$

The remaining bounds are immediate from Lemma 5.1 because

$$\left\|\mathsf{D}^- p_h\right\|_A^2 = \left\|\mathsf{D}\mathsf{D}^- p_h\right\|_{L^2(\Omega)}^2 + \tau \left\|\mathsf{D}^- p_h\right\|_{L^2(\Omega)}^2 = \tau \left\|\mathsf{D}^- p_h\right\|_{L^2(\Omega)}^2 .$$

All the constants merely depend on Ω and the shape regularity of $\mathcal{T}_h$. ∎

The case $\mathsf{D} = \operatorname{div}$ deserves special attention, because the discrete potential belongs to $V_h(\mathbf{curl})$. This is not entirely desirable, because it entails solving an $\mathbf{H}(\mathbf{curl},\Omega)$-elliptic problem in $V_h(\mathbf{curl})$ when evaluating the preconditioner. Yet, as $\mathbf{curl}\circ\mathbf{grad} = 0$, we can apply the decomposition of Theorem 6.1 recursively and replace $p_h \in V_h(\mathbf{curl})$ by a $\mathbf{\Phi}_h \in S_h$ and some "high-frequency" edge element function.

Thus, for $\mathsf{D} = \operatorname{div}$ and $\mathbf{v}_h \in V_h(\operatorname{div})$, we examine the decomposition

$$\begin{aligned}\mathbf{v}_h &= \sum_{\mathbf{b}\in\mathcal{B}(\operatorname{div})} \mathbf{v_b} + \Pi_h^{\operatorname{div}} \mathbf{\Psi}_h + \mathbf{curl}\, p_h \\ &= \sum_{\mathbf{b}\in\mathcal{B}(\operatorname{div})} \mathbf{v_b} + \Pi_h^{\operatorname{div}} \mathbf{\Psi}_h + \sum_{\mathbf{q}\in\mathcal{B}(\mathbf{curl})} \mathbf{curl}\, p_{\mathbf{q}} + \mathbf{curl}\, \mathbf{\Phi}_h,\end{aligned} \tag{6.5}$$

where $\mathbf{\Psi}_h, \mathbf{\Phi}_h \in S_h$ and

$$\mathbf{v_b} \in \operatorname{span}\{\mathbf{b}\}, \quad \mathbf{b}\in\mathcal{B}(\operatorname{div}), \qquad p_{\mathbf{q}} \in \operatorname{span}\{\mathbf{q}\}, \quad \mathbf{q}\in\mathcal{B}(\mathbf{curl}).$$

From Theorem 6.1 we conclude that for $0 < \tau \le 1$

$$\sum_{\mathbf{b}\in\mathcal{B}(\operatorname{div})} \|\mathbf{v_b}\|_A^2 + \|\mathbf{\Psi}_h\|_{H^1(\Omega)}^2 + \tau \sum_{\mathbf{q}\in\mathcal{B}(\mathbf{curl})} \|\mathbf{curl}\, p_{\mathbf{q}}\|_{L^2(\Omega)}^2 + \tau \|\mathbf{\Phi}_h\|_{H^1(\Omega)}^2 \lesssim \|\mathbf{v}_h\|_A^2 . \tag{6.6}$$

From now on we permit dependence of the constants on the variation of h. In other words, the estimates below hinge on the assumption of quasi-uniformity of the mesh $\mathcal{T}_h$, which permits us to assume a global meshwidth $h > 0$. Then, we can establish stability *uniformly* for all $\tau > 0$.

THEOREM 6.2. *Assume the* 2-regular case. *Then, for all* $\mathbf{v}_h \in V_h(\mathsf{D})$, *we can find* $p_h \in V_h(\mathsf{D}^-)$, $\mathbf{\Psi}_h \in S_h$ *such that, when* $\mathbf{v_b} \in \operatorname{span}\{\mathbf{b}\}$, $\mathbf{b}\in\mathcal{B}(\mathsf{D})$,

$$\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \mathbf{v_b} + \Pi_h^{\mathsf{D}} \mathbf{\Psi}_h + \mathsf{D}^- p_h = \mathbf{v}_h, \tag{6.7}$$

$$\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v_b}\|_A^2 + \|\mathbf{\Psi}_h\|_{H^1(\Omega)}^2 + \tau \|\mathbf{\Psi}_h\|_{L^2(\Omega)}^2 + \left\|\mathsf{D}^- p_h\right\|_A^2 \lesssim \|\mathbf{v}_h\|_A^2 , \tag{6.8}$$

with a constant depending only on Ω, D, *and the shape regularity and quasi-uniformity of* $\mathcal{T}_h$, *but independent of* $\tau \ge 0$.

Proof. We rely on the decomposition established in Lemma 5.2,

$$\mathbf{v}_h = \widetilde{\mathbf{v}}_h + \Pi_h^{\mathsf{D}} \mathbf{\Psi}_h + \mathsf{D}^- p_h, \tag{6.9}$$

and find, using (4.8), (6.3), and (3.6),

$$\begin{aligned} \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v}_\mathbf{b}\|_A^2 &\lesssim \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} h^{-2} \|\mathbf{v}_\mathbf{b}\|_{L^2(\Omega)}^2 + \tau \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \|\mathbf{v}_\mathbf{b}\|_{L^2(\Omega)}^2 \\ &\lesssim (h^{-2}+\tau) \|\widetilde{\mathbf{v}}_h\|_{L^2(\Omega)}^2 \lesssim (h^{-2}+\tau) h^2 \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}^2 \\ &\lesssim \|\mathsf{D}\mathbf{v}_h\|_{L^2(\Omega)}^2 + \tau \|\mathbf{v}_h\|_{L^2(\Omega)}^2 . \end{aligned} \tag{6.10}$$

Bounds for the other terms are straightforward from the estimates of Lemmas 5.2 and 5.1. ∎

As regards $\mathsf{D} = \operatorname{div}$, in the 2-regular case we also get a τ-uniform estimate for the splitting (6.5):

$$\begin{aligned} \sum_{\mathbf{b}\in\mathcal{B}(\operatorname{div})} \|\mathbf{v}_\mathbf{b}\|_A^2 + \|\mathbf{\Psi}_h\|_{H^1(\Omega)}^2 + \tau \|\mathbf{\Psi}_h\|_{L^2(\Omega)}^2 \\ + \tau \sum_{\mathbf{b}\in\mathcal{B}(\mathbf{curl})} \|\mathbf{curl}\, p_\mathbf{b}\|_{L^2(\Omega)}^2 + \tau \|\mathbf{\Phi}_h\|_{H^1(\Omega)}^2 \lesssim \|\mathbf{v}_h\|_A^2 . \end{aligned} \tag{6.11}$$

Remark 8. The theory seems to indicate increased robustness of the preconditioner with respect to $\tau \to \infty$ in the 2-regular case. However, the numerical experiments of section 8 send the unequivocal message that the "generic case" version of the preconditioner does not deteriorate as τ becomes large. Here, theory obviously falls short of capturing the actual behavior of the method.

7. Auxiliary space preconditioners. We start from the stable decompositions of $V_h(\mathsf{D})$ introduced in Theorems 6.1 and 6.2 and (6.5) and apply the abstract theory of section 2 for $V = V_h(\mathsf{D})$ and the energy bilinear form $a(\cdot,\cdot)$ from (1.3).

Throughout, let $\mathbf{A}_\mathsf{D}$ denote the Galerkin matrix arising from (1.3) with respect to the standard basis $\mathcal{B}(\mathsf{D})$ of $V_h(\mathsf{D})$. We write $\mathbf{L}$ for the matrix related to the bilinear form

$$(\mathbf{\Psi}, \mathbf{\Phi}) \mapsto (\mathbf{grad}\, \mathbf{\Psi}, \mathbf{grad}\, \mathbf{\Phi})_0 , \quad \mathbf{\Phi}, \mathbf{\Psi} \in \boldsymbol{\mathcal{H}}^1,$$

on S_h, which is endowed with the usual nodal basis of hat functions (for the components of vector fields). The positive definite mass matrix on S_h, that is, the Galerkin matrix for the $L^2(\Omega)$-inner product, will be designated by $\mathbf{M}$. Further, we adopt the notation $\mathbf{P}_\mathsf{D}$ for the matrix describing the mapping $\Pi_h^{\mathsf{D}} : \boldsymbol{\mathcal{H}}^1 \mapsto V_h(\mathsf{D})$ with respect to the "hat function basis" of $\boldsymbol{\mathcal{H}}^1$ and the basis $\mathcal{B}(\mathsf{D})$ of $V_h(\mathsf{D})$.

We restrict ourselves to Jacobi smoothing; that is, the smoothing operator is characterized by the inner product

$$s(\mathbf{v}_h, \mathbf{v}_h) = \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} a(\mathbf{v}_\mathbf{b}, \mathbf{v}_\mathbf{b}), \quad \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})} \mathbf{v}_\mathbf{b} = \mathbf{v}_h, \quad \mathbf{v}_\mathbf{b} \in \operatorname{span}\{\mathbf{b}\}, \tag{7.1}$$

and its matrix representation coincides with the diagonal $\mathbf{D}_A$ of $\mathbf{A}_\mathsf{D}$. More generally, one could use any $s(\cdot,\cdot)$ that features the spectral equivalence $s(\mathbf{v}_h, \mathbf{v}_h) \approx \left\|h^{-1}\mathbf{v}_h\right\|_{L^2(\Omega)}^2 + \tau\|\mathbf{v}_h\|^2$.

Since the square of the energy norm can be computed by summing local contributions from the cells K of the mesh $\mathcal{T}_h$, we find

$$\begin{aligned}\|\mathbf{v}_h\|_A^2 &= \left\|\sum\nolimits_{\mathbf{b}\in\mathcal{B}(\mathsf{D})}\alpha_{\mathbf{b}}\mathbf{b}\right\|_A^2 = \sum\nolimits_{K\in\mathcal{T}_h}\left\|\sum\nolimits_{j=1}^{M}\alpha_{K_j}\mathbf{b}_{K,j}\right\|_A^2 \\ &\le M\sum_K\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})}|\alpha_{\mathbf{b}}|^2\,\|\mathbf{b}\|_A^2 = M\sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})}|\alpha_{\mathbf{b}}|^2\,\|\mathbf{b}\|_A^2 = Ms(\mathbf{v},\mathbf{v})\end{aligned} \tag{7.2}$$

if $\mathbf{v}_h = \sum_{\mathbf{b}\in\mathcal{B}(\mathsf{D})}\alpha_{\mathbf{b}}\mathbf{b} \in V_h(\mathsf{D})$. Here, M bounds the (small) number of basis functions whose support overlaps with a single element K. This implies that c_s in (2.12) can be chosen as a small universal constant.

For the sake of simplicity, we continue the discussion for the cases $\mathsf{D} = \mathbf{curl}$ (edge elements) and $\mathsf{D} = \mathrm{div}$ (face elements) separately.

7.1. A preconditioner for $\mathbf{H(curl,\Omega)}$-elliptic problems. We rely on the splitting (6.1) to define the preconditioner. This means that, in terms of the concepts developed in section 2, we have two auxiliary spaces:

1. The space $W_1 := S_h$ with inner product $\overline{a}_1(\mathbf{\Psi}_h,\mathbf{\Psi}_h) := \|\mathbf{\Psi}_h\|^2_{H^1(\Omega)} + \tau\,\|\mathbf{\Psi}_h\|^2_{L^2(\Omega)}$, which is suggested by (6.2) and (6.8). The corresponding transfer operator is $\Pi_1 := \Pi_h^{\mathbf{curl}}$, and, thanks to (4.11), (2.11) holds with constant c_1 depending only on the shape regularity of the mesh.
2. The discrete potential space $W_2 := V_h(\mathsf{D}^-)$ equipped with inner product $\overline{a}_2(p_h,p_h) := \tau\,|p_h|^2_{H^1(\Omega)}$ and transfer operator $\Pi_2 := \mathbf{grad} : V_h(\mathsf{D}^-) \mapsto V_h(\mathsf{D})$, whose norm is uniformly bounded by 1.

We write $\mathbf{G}$ for the matrix related to $\mathbf{grad} : V_h(\mathsf{D}^-) \mapsto V_h(\mathsf{D})$ and $\mathbf{\Delta}$ for the discrete Laplacian (matrix) on linear Lagrangian finite element space $V_h(\mathbf{grad})$. Then the matrix of the resulting auxiliary space preconditioner for the $\mathbf{H}(\mathbf{curl},\Omega)$-elliptic problem (1.1) reads

$$\mathbf{B}_{\mathbf{curl}} := \mathbf{D}_A^{-1} + \mathbf{P}_{\mathbf{curl}}(\mathbf{L}+\tau\mathbf{M})^{-1}\mathbf{P}_{\mathbf{curl}}^T + \tau^{-1}\mathbf{G}(-\mathbf{\Delta})^{-1}\mathbf{G}^T. \tag{7.3}$$

THEOREM 7.1. *For $0 < \tau \le 1$ the spectral condition number $\kappa(\mathbf{B}_{\mathbf{curl}}\mathbf{A}_{\mathbf{curl}})$ depends only on Ω and the shape regularity of the mesh.*

In the 2-regular case $\kappa(\mathbf{B}_{\mathbf{curl}}\mathbf{A}_{\mathbf{curl}})$ is bounded independently of τ, but the quasi-uniformity of the mesh may affect the bound.

Proof. Theorems 6.1 and 6.2 provide the bound for c_0 from (2.13). The constants c_s and c_1, c_2 have been discussed before. Thus, (2.14) leads to the assertion of the theorem. ∎

The impact of switching to spectrally equivalent bilinear forms on W_1, W_2 can be gauged as in Corollary 2.3.

We point out that the transfers can be realized by purely local operations; see [5, section 3] and [5, section 5]. In detail, assuming the standard bases, gradient-matrix $\mathbf{G}$ will agree with the edge-vertex incidence matrix of the mesh. The matrix $\mathbf{P}_{\mathbf{curl}}$ connected with the interpolation $\Pi_h^{\mathbf{curl}}$ describes a local distribution of vectorial degrees of freedom attached to the nodes of the mesh to adjacent edges: the edge connecting vertices with values $\mathbf{w}_1$ and $\mathbf{w}_2$ receives the value

$$\tfrac{1}{2}(\mathbf{w}_1+\mathbf{w}_2)\cdot\mathbf{e}, \tag{7.4}$$

where $\mathbf{e}$ is the direction vector of the edge.

7.2. A preconditioner for H(div, Ω)-elliptic problems. In this case the decomposition (6.5) provides the starting point. It suggests that we choose the following auxiliary spaces:

1. $W_1 := S_h$ with inner product $\overline{a}_1(\mathbf{\Psi}_h, \mathbf{\Psi}_h) := \|\mathbf{\Psi}_h\|^2_{H^1(\Omega)} + \tau \|\mathbf{\Psi}_h\|^2_{L^2(\Omega)}$, which is suggested by (6.6) and (6.11). The corresponding transfer operator is $\Pi_1 := \Pi_h^{\mathrm{div}}$ and, thanks to (4.11), (2.11) holds with constant c_1 depending only on the shape regularity of the mesh. The related interpolation matrix $\mathbf{P}_{\mathrm{div}}$ assigns to each face of the mesh with unit normal $\mathbf{n}$ and area $|F|$ the number
$$\tfrac{1}{3}|F|\,(\mathbf{w}_1 + \mathbf{w}_2 + \mathbf{w}_3)\cdot\mathbf{n}, \tag{7.5}$$
where $\mathbf{w}_i$ is the vectorial nodal value at vertex i of the face.
2. $W_2 := V_h(\mathbf{curl})$ endowed with the localized inner product
$$\overline{a}_2(\mathbf{w}_h, \mathbf{w}_h) := \tau \sum_{\mathbf{q}\in\mathcal{B}(\mathbf{curl})} \|\mathbf{curl}\,\mathbf{w}_{\mathbf{q}}\|^2_{L^2(\Omega)}\,, \qquad \sum_{\mathbf{q}\in\mathcal{B}(\mathbf{curl})} \mathbf{w}_{\mathbf{q}} = \mathbf{w}_h \tag{7.6}$$
for $\mathbf{w}_h \in V_h(\mathbf{curl})$. Evidently, the Galerkin discretization of $\overline{a}_2$ leads to a diagonal matrix denoted by $\mathbf{D}_{\mathbf{curl}}$. A closer inspection of $\mathcal{B}(\mathbf{curl})$ verifies that $\mathbf{D}_{\mathbf{curl}}$ can never be singular.
The transfer operator associated with W_2 is $\mathbf{curl} : V_h(\mathbf{curl}) \mapsto V_h(\mathrm{div})$ and $c_2 = 1$ is obvious. Its matrix representation $\mathbf{C}$ coincides with the incidence matrix of (interior) edges and faces of the mesh; see [21, section 3.1].
3. $W_3 := S_h$ with norm $\sqrt{\tau}\,\|\cdot\|_{H^1(\Omega)}$ (cf. (6.6) and (6.11)) and transfer operator $\mathbf{curl} : S_h \mapsto V_h(\mathsf{D})$. Again, we immediately get $c_3 = 1$ for the constant from (2.11). Owing to the commuting diagram property (4.6) and (4.9), the matrix associated with this transfer is given by $\mathbf{CP}_{\mathbf{curl}}$.

Summing up, the matrix representation of the auxiliary space preconditioner for the variational problem (1.2) discretized on $V_h(\mathrm{div})$ is given by

$$\begin{aligned}\mathbf{B}_{\mathrm{div}} := {}& \mathbf{D}_A^{-1} + \mathbf{P}_{\mathrm{div}}(\mathbf{L} + \tau\mathbf{M})^{-1}\mathbf{P}_{\mathrm{div}}^T + \mathbf{C}\mathbf{D}_{\mathbf{curl}}^{-1}\mathbf{C}^T \\ & + \tau^{-1}\mathbf{C}\mathbf{P}_{\mathbf{curl}}(\mathbf{L} + \tau\mathbf{M})^{-1}\mathbf{P}_{\mathbf{curl}}^T\mathbf{C}^T.\end{aligned} \tag{7.7}$$

All transfer operators are purely local.

THEOREM 7.2. *For $0 < \tau \le 1$ the spectral condition number $\kappa(\mathbf{B}_{\mathrm{div}}\mathbf{A}_{\mathrm{div}})$ depends only on Ω and the shape regularity of the mesh.*

In the 2-regular case $\kappa(\mathbf{B}_{\mathrm{div}}\mathbf{A}_{\mathrm{div}})$ is bounded independently of τ, but the quasi-uniformity of the mesh may affect the bound.

Proof. We merely need to appeal to (6.6), (6.11), and (2.14), because good bounds for c_1, c_2, c_3, and c_s follow from the above arguments. ∎

Remark 9. If boundary conditions are imposed on $\mathcal{H}(\mathsf{D}, \Omega)$, the auxiliary space S_h should be chosen differently in the 2-regular case: it should comprise piecewise linear continuous vector fields, for which merely the tangential or normal components, respectively, vanish on $\partial\Omega$. Of course, this choice can be made in any case, because enlarging the auxiliary space will not affect the estimates adversely unless the continuity of Π is destroyed. On the other hand, tangential boundary conditions are awkward in terms of implementation; cf. the discussion in [5, section 5]. Moreover, as stressed in Remark 8, there is absolutely no numerical evidence that total zero boundary conditions for S_h do any harm.

7.3. Applications to problems with variable coefficients. So far we have skirted the case of variable coefficients, for instance, when we encounter the bilinear form

$$a(\mathbf{u},\mathbf{v}) := (\alpha\,\mathbf{curl}\,\mathbf{u},\mathbf{curl}\,\mathbf{v})_0 + (\beta\mathbf{u},\mathbf{v})_0\,, \quad \mathbf{u},\mathbf{v}\in\mathbf{H}_0(\mathbf{curl},\Omega), \tag{7.8}$$

with coefficient functions $\alpha,\beta\in L^\infty(\Omega)$, because the current theory fails to give any useful information about how strong variations of α and β affect the quality of the nodal auxiliary space preconditioners.

We can give only a heuristic recipe for how the algorithms may be adapted to the general bilinear form from (7.8). The idea is that the coefficient α will be used to define the matrix $\mathbf{L}$ in (7.3). This means that $\mathbf{L}$ agrees with the Galerkin matrix of the bilinear form $(\mathbf{u},\mathbf{v})\mapsto(\alpha\,\mathbf{grad}\,\mathbf{u},\mathbf{grad}\,\mathbf{v})_0$ on $\boldsymbol{\mathcal{H}}^1$. The coefficient β enters the matrices $\mathbf{M}$ and $\boldsymbol{\Delta}$; that is, they represent $(\mathbf{u},\mathbf{v})\mapsto(\beta\mathbf{u},\mathbf{v})_0$ and $(\varphi,\psi)\mapsto(\beta\,\mathbf{grad}\,\varphi,\mathbf{grad}\,\psi)_0$ on $\boldsymbol{\mathcal{H}}^1$ and $V_h(\mathbf{grad})$, respectively. Note that τ is now incorporated into the coefficient β.

8. Numerical experiments. The theory makes a statement about the asymptotic behavior of the nodal auxiliary, but information about concrete condition numbers remains hidden in several elusive constants. In this section we wish to demonstrate that the preconditioner actually achieves reasonably small condition numbers for relevant model problems. Moreover, we monitor the impact of the relative scaling of both parts of the bilinear form $a(\cdot,\cdot)$ from (1.3). Reaching beyond the scope of the theory, we will also examine the impact of strongly varying coefficients in (7.8). Throughout, nodal auxiliary spaces with zero boundary values for all vector components will be used ("generic case").

The first series of experiments is conducted in two dimensions. Note that in two dimensions the operators div and curl acting on vector fields merely differ by a rotation of $\frac{\pi}{2}$. Therefore, both variational problems (1.1) and (1.2) are covered when we consider the bilinear form

$$a(\mathbf{u},\mathbf{v}) := (\operatorname{curl}\mathbf{u},\operatorname{curl}\mathbf{v})_0 + \tau\,(\mathbf{u},\mathbf{v})_0\,, \quad \mathbf{u},\mathbf{v}\in\mathbf{H}_0(\operatorname{curl},\Omega), \tag{8.1}$$

where curl is the scalar-valued two-dimensional rotation $\operatorname{curl}\mathbf{u}=\frac{\partial u_1}{\partial x_2}-\frac{\partial u_2}{\partial x_1}$.

In two dimensions we can study the asymptotics with manageable computational effort. We emphasize that all the considerations underlying the nodal auxiliary subspace approach in three dimensions carry over to (8.1): Galerkin discretization can be based on edge elements on triangular meshes, for which curl-free functions can represented as gradients of piecewise linear Lagrangian finite element functions.

In most experiments we used the preconditioner given by the two-dimensional counterpart of (7.3). A direct solver was employed to realize the multiplications with the inverse matrices. Extremal eigenvalues were computed by means of a Lanczos procedure up to an accuracy of at least two digits.

Experiment I. A sequence of meshes of two polygonal domains was created by the regular refinement of the coarse meshes depicted in Figure 8.1. One domain is convex, that is, it satisfies the assumptions of the 2-regular case, while the other fails to do so. Spectral condition numbers of $\mathbf{B}_{\mathrm{curl}}\mathbf{A}_{\mathrm{curl}}$, $\mathbf{A}_{\mathrm{curl}}$ the edge element Galerkin matrix related to (8.1), were computed for different choices of τ; see Tables 8.1 and 8.3. A variant of the preconditioner relying on two steps of Gauss–Seidel smoothing (see section 2) was tested in the same setting. The measured condition numbers are listed in Tables 8.5 and 8.6.

We also keep track of the number of PCG iterations required to solve the discrete variational problems with bilinear form $a(\cdot,\cdot)$ from (8.1) and constant vector field $\mathbf{f} = \binom{1}{1}$ as the right-hand side. A relative reduction of the Euclidean norm of the residual vector by a factor of 10^6 was used as termination criterion. The results are recorded in Tables 8.2 and 8.4.

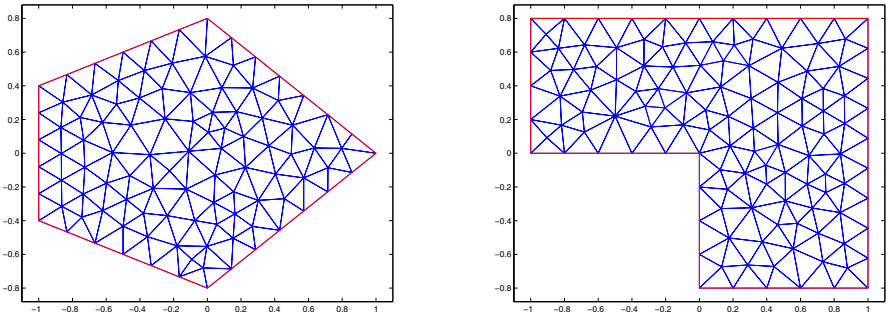


FIG. 8.1. *Coarsest meshes used in Experiment* I.

TABLE 8.1
Condition numbers for Experiment I: *Convex polygon.*

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
0	171	20.4	20.4	20.3	20.1	18.2	11.7	5.0	2.6	3.7
1	684	25.0	25.0	25.0	24.7	23.0	16.5	8.3	3.6	3.3
2	2736	28.4	28.4	28.4	28.2	26.6	20.8	12.5	6.2	2.9
3	10944	31.2	31.2	31.2	31.0	29.7	24.8	17.5	9.3	4.6
4	43776	33.5	33.5	33.5	33.3	32.2	28.3	22.1	14.2	7.1
5	175104	35.2	35.2	35.2	35.1	34.2	31.1	26.1	19.1	11.0

TABLE 8.2
Required PCG iterations for Experiment I: *Convex polygon.*

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
0	171	29	29	29	28	27	24	15	9	11
1	684	33	33	33	33	32	28	20	11	11
2	2736	36	36	36	36	35	31	25	16	10
3	10944	38	38	39	38	38	34	29	20	13
4	43776	41	41	41	41	40	37	32	25	17
5	175104	42	42	43	42	42	39	34	28	21

The condition numbers hardly deteriorate on successively finer meshes. The slight dependence on the refinement level is a commonly observed preasymptotic phenomenon; cf. Remark 2 in [8]. A similar statement applies to the number of CG iterations. Robustness in τ is evident though not covered by theory when using a nodal auxiliary space with zero boundary conditions ("generic setting," Table 3.1). Using a Gauss–Seidel smoother instead of Jacobi improves the performance at increased costs for a single application of the preconditioner.

Experiment II. Starting from a coarse mesh on the "L-shaped domain" (Figure 8.1, right) we generate a sequence of meshes by strictly local refinement; see

TABLE 8.3
Condition numbers for Experiment I: *L-shaped domain.*

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
0	205	63.0	62.9	62.6	59.9	42.5	14.9	4.9	2.6	4.2
1	820	69.5	69.4	69.1	66.2	47.7	18.4	8.7	3.2	3.4
2	3280	71.6	71.6	71.3	68.4	49.5	19.9	12.2	5.9	2.4
3	13120	72.3	72.3	72.0	69.0	50.1	21.2	15.6	9.2	4.0
4	52480	72.5	72.5	72.2	69.2	50.3	23.4	18.9	13.0	7.0
5	209920	72.6	72.6	72.3	69.3	50.4	25.2	21.7	16.7	10.3

TABLE 8.4
Required PCG iterations for Experiment I: *L-shaped domain.*

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
0	171	35	35	35	34	33	25	15	9	12
1	684	40	40	40	40	37	30	21	11	11
2	2736	43	43	43	43	40	33	25	16	9
3	10944	46	46	46	46	44	36	28	20	12
4	43776	49	49	49	48	47	38	31	25	17
5	175104	51	51	51	51	49	41	34	28	20

TABLE 8.5
Experiment I*, condition numbers on convex polygon with* symmetric GS smoothing.

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
0	171	6.5	6.5	6.5	6.4	5.6	3.1	1.9	2.0	2.0
1	684	9.4	9.4	9.4	9.3	8.4	5.3	2.4	2.0	2.0
2	2736	11.8	11.8	11.8	11.7	10.8	7.6	3.7	2.3	2.0
3	10944	13.7	13.7	13.7	13.6	12.7	9.7	5.7	2.9	2.2
4	43776	15.2	15.2	15.2	15.1	14.4	11.8	8.0	4.2	2.7
5	175104	16.4	16.4	16.4	16.3	15.7	13.6	10.4	6.4	3.2

TABLE 8.6
Experiment I*, condition numbers on L-shaped domain with* symmetric GS smoothing.

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
0	171	19.6	19.6	19.5	18.7	13.0	4.1	1.8	2.0	2.0
1	684	28.8	28.8	28.7	27.5	19.6	6.9	2.3	2.0	2.0
2	2736	33.5	33.5	33.4	32.0	23.1	8.8	3.9	2.2	2.0
3	10944	35.3	35.3	35.1	33.7	24.4	9.7	5.7	2.7	2.1
4	43776	35.9	35.9	35.7	34.3	24.9	10.3	7.3	4.1	2.5
5	175104	36.1	36.1	36.0	34.5	25.1	11.4	9.0	6.0	3.1

Figure 8.2. As in the previous experiment we recorded spectral condition numbers $\kappa(\mathbf{B}_{\mathrm{curl}}\mathbf{A}_{\mathrm{curl}})$; see Figure 8.3. This time, we monitor their dependence on the smallest size $h_{\min}$ of mesh cells.

The conclusions drawn in Experiment I carry over verbatim.

Experiment III. To study the response of the preconditioner to nonconstant coefficients, we apply it to the bilinear form

$$a(\mathbf{u},\mathbf{v}) := (\alpha \operatorname{curl}\mathbf{u}, \operatorname{curl}\mathbf{v})_0 + \tau\,(\beta\mathbf{u},\mathbf{v})_0\,, \quad \mathbf{u},\mathbf{v}\mathbf{H}_0(\operatorname{curl},\Omega), \tag{8.2}$$

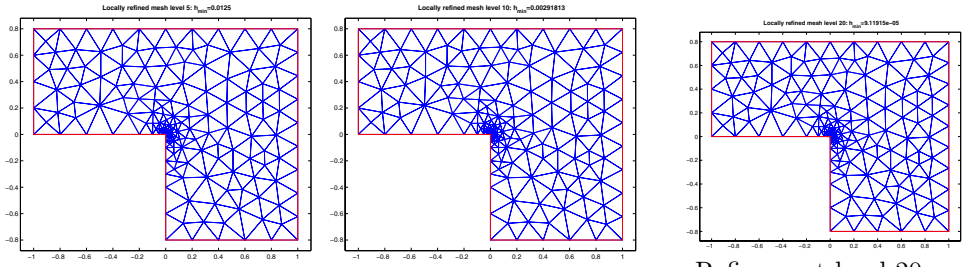


FIG. 8.2. *Sequence of locally refined meshes.*

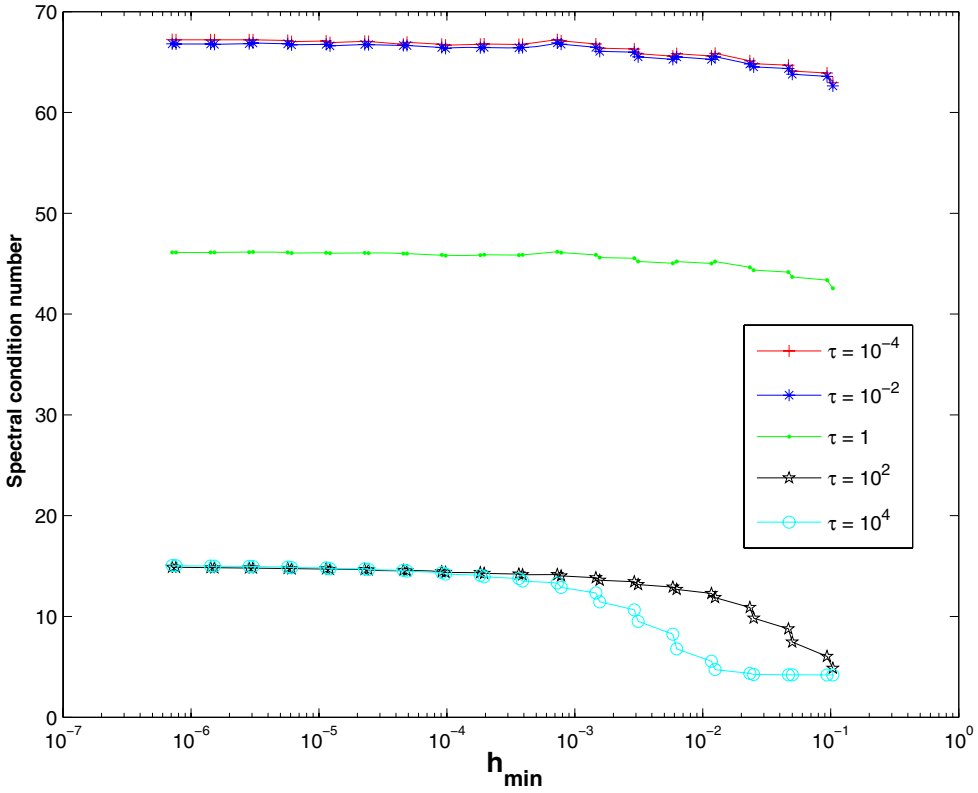


FIG. 8.3. *Behavior of auxiliary space preconditioner on locally refined meshes.*

with $\alpha, \beta \in L^\infty(\Omega)$. This is the two-dimensional analogue of (7.8). The implementation of the preconditioner $\mathbf{B}_{\mathrm{curl}}$ follows the policy outlined in section 7.3.

We consider $\Omega =]-1,1[^2$ with a triangular subdomain Ω_1 that is resolved by the mesh; see Figure 8.4. The coefficient functions behave like

$$\alpha(\mathbf{x}) := \begin{cases} \alpha_1 & \text{if } \mathbf{x} \in \Omega_1, \\ 1 & \text{elsewhere,} \end{cases} \qquad \beta(\mathbf{x}) := \begin{cases} \beta_1 & \text{if } \mathbf{x} \in \Omega_1, \\ 1 & \text{elsewhere.} \end{cases} \tag{8.3}$$

We recorded the condition numbers of the preconditioned stiffness matrices on sequences of meshes arising from successive regular refinement of the mesh depicted

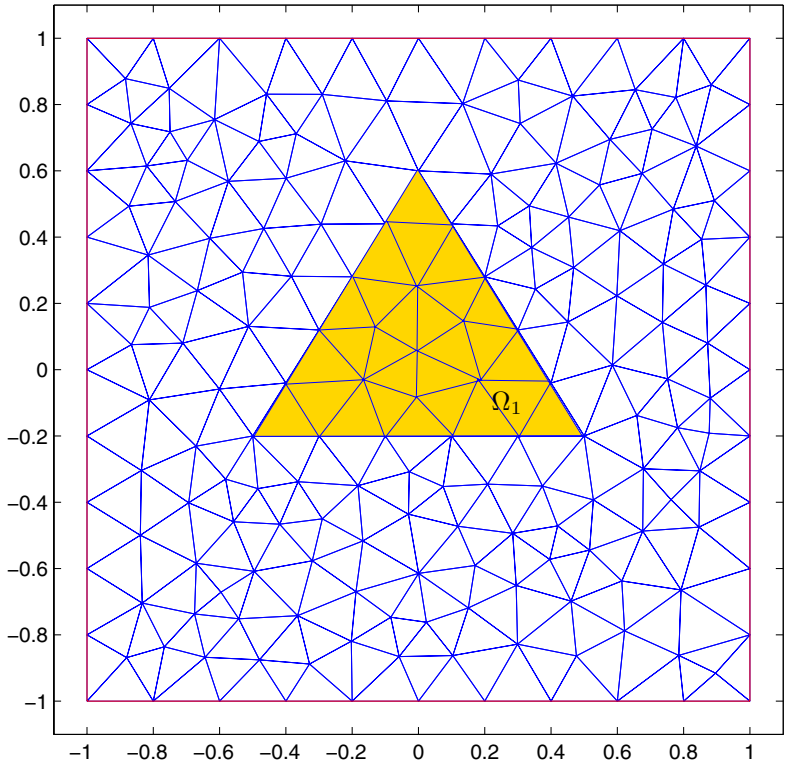


FIG. 8.4. *Subdomains with piecewise constant coefficients and associated coarsest triangular mesh.*

in Figure 8.4; see Tables 8.7 and 8.9. In addition, Tables 8.8 and 8.10 give the number of CG iterations required for a relative reduction of the Euclidean residual norm by a factor of 10^6. As before, zero was used as an initial guess and we chose the source field $\mathbf{f} = \binom{1}{1}$.

TABLE 8.7
Condition numbers *(two digits) recorded in Experiment* III: *Discontinuous coefficient* α, $\beta_1 = 1$.

		α_1										
Level	# cells	0.001	0.01	0.1	2	5	10	20	50	100	200	1000
0	332	19	18	16	18	21	24	28	31	33	34	35
1	1328	23	21	20	21	24	27	31	36	38	39	40
2	5312	28	27	22	23	25	29	33	37	40	41	42
3	21248	35	32	24	25	26	30	34	38	40	42	43
4	84992	41	36	26	26	27	30	34	38	41	42	43
5	339968	46	39	27	27	28	31	34	39	41	42	43

By and large, we observe that the condition number of the preconditioned system is not much affected by steep jumps in the coefficients α or β. A slight deterioration is caused by large values of β inside Ω_1. The same holds for the convergence of the preconditioned CG iterations. It seems that the behavior of the method surpasses the predictions of the theory.

Experiment IV. In this experiment we study the auxiliary space preconditioner for genuine three-dimensional boundary value problems of the form (1.1). Zero Dirich-

TABLE 8.8
Number of CG-iterations for Experiment III: *Discontinuous coefficient* α, $\beta_1 = 1$.

		α_1										
Level	# cells	0.001	0.01	0.1	2	5	10	20	50	100	200	1000
0	332	33	32	28	30	31	33	34	36	36	35	37
1	1328	37	35	34	33	36	37	38	39	41	42	42
2	5312	43	42	37	37	38	41	42	44	44	45	46
3	21248	49	48	39	39	41	43	45	47	48	48	50
4	84992	54	52	41	41	42	45	48	50	51	52	53
5	339968	58	56	44	43	45	48	50	52	53	54	55

TABLE 8.9
Condition numbers *(two digits) measured in Experiment* III: *Discontinuous coefficient* β, $\alpha_1 = 1$.

		β_1										
Level	# cells	0.001	0.01	0.1	2	5	10	20	50	100	200	1000
0	332	17	17	18	18	18	20	24	32	39	48	64
1	1328	21	21	21	21	21	22	26	35	44	53	73
2	5312	23	23	23	23	23	24	27	36	46	56	78
3	21248	24	25	25	25	25	25	28	37	46	57	80
4	84992	26	26	26	26	26	27	28	37	46	57	80
5	339968	27	27	27	27	27	28	29	37	46	57	81

TABLE 8.10
Number of CG-iterations for Experiment III: *Discontinuous coefficient* β, $\alpha_1 = 1$.

		β_1										
Level	# cells	0.001	0.01	0.1	2	5	10	20	50	100	200	1000
0	332	30	30	30	30	30	31	33	35	38	39	41
1	1328	35	35	34	34	34	34	38	40	42	43	46
2	5312	38	38	37	37	37	36	40	43	45	47	48
3	21248	39	39	38	38	39	38	40	45	47	50	53
4	84992	41	41	40	40	40	41	42	47	49	51	55
5	339968	43	43	42	42	42	43	44	49	52	54	56

let boundary conditions are used throughout; that is, $\mathcal{H}(\mathbf{curl}, \Omega) = \mathbf{H}_0(\mathbf{curl}, \Omega)$.

We consider two different domains; one is the unit cube $\Omega = \Omega_\square := (0,1)^3$, and the other is the unit ball $\Omega = \Omega_\circ := \{x = (x_1, x_2, x_3) \in R^3 : x_1^2 + x_2^2 + x_3^2 < 1\}$. Lowest order edge elements (cf. section 4) are applied to discretize (1.1) on quasi-uniform simplicial triangulations of both domains. We apply the preconditioner $\mathbf{B}_{\mathbf{curl}}$ given in (7.3) to the discretized systems with

1. $\mathbf{D}_A^{-1}$ replaced with the approximate inverse corresponding to three iterations of the symmetric point Gauss–Seidel method for $\mathbf{A}_{\mathbf{curl}}$,
2. $(\mathbf{L} + \tau\mathbf{M})^{-1}$ replaced with one V-cycle of an AMG method [35] for $\mathbf{L} + \tau\mathbf{M}$, and
3. $(-\boldsymbol{\Delta})^{-1}$ replaced with one V-cycle AMG method for matrix $-\boldsymbol{\Delta}$ of the discrete vector Laplacian.

The matrices $\mathbf{L}$ and $\mathbf{M}$ correspond to the generic case; see Table 3.1.

We study the condition number of $\mathbf{B}_{\mathbf{curl}}\mathbf{A}_{\mathbf{curl}}$ on sequences of uniformly and regularly refined triangulations of both domains. On the unit cube the coarsest mesh is obtained by splitting each cell of a uniform tensor product grid into six tetrahedra. For the sphere, a mesh generator is employed to get a sequence of increasingly finer meshes, whose tetrahedra all have about the same size and little distortion. Condition number

estimates are computed by means of the Lanczos method and listed in Tables 8.11 and 8.12 for different values of the scaling parameter τ.

TABLE 8.11
Unit cube: Spectral condition numbers of $\mathbf{B_{curl}A_{curl}}$.

		τ		
Level	# cells	10^{-4}	1	10^4
1	6×8^3	4.645	4.580	2.943
2	6×16^3	4.689	4.644	2.952
3	6×32^3	4.842	4.817	2.983
4	6×48^3	4.954	4.771	2.969

TABLE 8.12
Unit ball: Spectral condition numbers of $\mathbf{B_{curl}A_{curl}}$.

		τ		
Level	# cells	10^{-4}	1	10^4
1	2197	2.893	2.911	3.021
2	4462	3.334	3.372	3.317
3	8865	3.280	3.288	3.430
4	17260	3.499	3.494	3.329
5	66402	3.955	3.932	3.431
6	95593	4.132	4.102	5.022
7	148554	4.497	4.246	3.513
8	242588	4.340	4.552	3.391

As before, we also record the the number of iterations required for the PCG method with the above preconditioner to reduce the $\mathbf{B_{curl}}$-norm of the residual by a factor of 10^6. The iteration counts for different values of τ are given in Tables 8.13 and 8.14. In both cases a zero initial guess was used and the right-hand side was such that the corresponding exact solutions of the boundary value problems are

$$\mathbf{u}(x,y,z) = \begin{pmatrix} xyz(x-1)(y-1)(z-1) \\ \sin(\pi x)\sin(\pi y)\sin(\pi z) \\ (1-e^x)(1-e^{x-1})(1-e^y)(1-e^{y-1})(1-e^z)(1-e^{z-1}) \end{pmatrix} \text{ on } \Omega_\square,$$

$$\mathbf{u}(x,y,z) = (x^2+y^2+z^2-1)\mathbf{1} \quad \text{on } \Omega_\circ.$$

TABLE 8.13
Number of PCG iterations on unit cube.

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
1	6×8^3	14	14	14	14	14	13	10	10	10
2	6×16^3	14	14	14	14	14	13	11	10	9
3	6×32^3	14	14	14	14	14	13	12	10	9
4	6×48^3	14	14	14	14	14	13	12	10	9

The observations perfectly match those made in two dimensions: the condition numbers and iteration counts are essentially independent of the meshwidth and τ. The number of PCG iterations decreases slightly when τ gets larger.

To illustrate the importance of the extra smoothings $\mathbf{D}_A^{-1}$ in our preconditioner (7.3), we recorded the number of iterations of the corresponding PCG method with the term $\mathbf{D}_A^{-1}$ removed from the preconditioner (7.3). The results are given in Table 8.15. We can see that the number of iterations doubles as the meshwidth gets halved.

TABLE 8.14
Number of PCG-iterations on unstructured grids in the unit ball.

		τ								
Level	# cells	10^{-4}	10^{-3}	10^{-2}	10^{-1}	1	10	10^2	10^3	10^4
1	2197	9	10	10	10	11	11	11	11	12
2	4462	10	10	11	11	11	12	11	11	12
3	8865	10	10	11	11	11	11	11	11	11
4	17260	10	11	11	11	12	12	11	10	11
5	66402	11	12	13	13	13	12	11	10	11
6	95593	11	12	13	13	13	12	12	11	12
7	148554	12	12	13	13	13	13	12	12	10
8	242588	12	13	13	14	14	13	12	11	10

TABLE 8.15
Number of PCG-iterations on the cube without smoothing.

		τ		
Level	# cells	10^{-4}	1	10^4
1	6×8^3	28	28	138
2	6×16^3	52	53	384
3	6×32^3	106	107	770
4	6×48^3	155	156	

Concluding remarks. Nodal auxiliary space preconditioning for discrete $\mathbf{H}(\mathbf{curl}, \Omega)$- and $\mathbf{H}(\mathrm{div}, \Omega)$-elliptic variational problems has a solid theoretical foundation and proves satisfactory in numerical tests. It can pave the way for applying standard AMG methods to boundary value problems discretized by means of edge or face finite elements. Numerous improvements of the method that can make use of better smoothers and refined auxiliary spaces are conceivable.

Acknowledgments. The authors wish to thank Wang Mengyu, Hangzhou University, and Patrick Meury, ETH Zürich, for writing parts of the MATLAB code for the two-dimensional experiments and also Tan Lin and Shu Shi, Xiangtan University, for their help with the three-dimensional experiments.

REFERENCES

[1] C. AMROUCHE, C. BERNARDI, M. DAUGE, AND V. GIRAULT, *Vector potentials in three-dimensional nonsmooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.
[2] D. ARNOLD, R. FALK, AND R. WINTHER, *Multigrid Preconditioning in $H(\mathrm{div})$ on Non-convex Polygons*, Tech. rep., Penn State University, College Park, PA, 1997.
[3] D. ARNOLD, R. FALK, AND R. WINTHER, *Preconditioning in $H(\mathrm{div})$ and applications*, Math. Comp., 66 (1997), pp. 957–984.
[4] D. ARNOLD, R. FALK, AND R. WINTHER, *Multigrid in $H(\mathrm{div})$ and $H(\mathbf{curl})$*, Numer. Math., 85 (2000), pp. 175–195.
[5] R. BECK, *Algebraic Multigrid by Component Splitting for Edge Elements on Simplicial Triangulations*, Tech. rep. SC 99-40, ZIB, Berlin, Germany, 1999.
[6] M. BIRMAN AND M. SOLOMYAK, *L_2-theory of the Maxwell operator in arbitrary domains*, Russian Math. Surveys, 42 (1987), pp. 75–96.
[7] P. B. BOCHEV, C. J. GARASI, J. J. HU, A. C. ROBINSON, AND R. S. TUMINARO, *An improved algebraic multigrid method for solving Maxwell's equations*, SIAM J. Sci. Comput., 25 (2003), pp. 623–642.
[8] F. BORNEMANN, *A Sharpened Condition Number Estimate for the BPX-Preconditioner of Elliptic Finite Element Problems on Highly Non-uniform Triangulations*, Tech. rep. SC 91-9, ZIB, Berlin, Germany, 1991.

[9] A. Bossavit, *Computational Electromagnetism. Variational Formulation, Complementarity, Edge Elements*, Electromagnetism 2, Academic Press, San Diego, CA, 1998.
[10] W. Boyse, D. Lynch, K. Paulsen, and G. Minerbo, *Nodal-based finite-element modeling of Maxwell's equations*, IEEE Trans. Antennas and Propagation, 40 (1992), pp. 642–651.
[11] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New York, 1991.
[12] A. Buffa, M. Costabel, and D. Sheen, *On traces for* $\mathbf{H}(\mathbf{curl}, \Omega)$ *in Lipschitz domains*, J. Math. Anal. Appl., 276 (2002), pp. 845–867.
[13] P. Ciarlet, *The Finite Element Method for Elliptic Problems*, Stud. Math. Appl. 4, North–Holland, Amsterdam, 1978.
[14] M. Costabel, *A remark on the regularity of solutions of Maxwell's equations on Lipschitz domains*, Math. Methods Appl. Sci., 12 (1990), pp. 365–368.
[15] M. Costabel, *A coercive bilinear form for Maxwell's equations*, J. Math. Anal. Appl., 157 (1991), pp. 527–541.
[16] M. Costabel and M. Dauge, *Maxwell and Lamé eigenvalues on polyhedra*, Math. Methods Appl. Sci., 22 (1999), pp. 243–258.
[17] V. Girault and P. Raviart, *Finite Element Methods for Navier–Stokes Equations*, Springer-Verlag, Berlin, 1986.
[18] M. Griebel and P. Oswald, *On the abstract theory of additive and multiplicative Schwarz algorithms*, Numer. Math., 70 (1995), pp. 163–180.
[19] R. Hiptmair, *Multigrid method for Maxwell's equations*, SIAM J. Numer. Anal., 36 (1998), pp. 204–225.
[20] R. Hiptmair, *Discrete Hodge operators*, Numer. Math., 90 (2001), pp. 265–289.
[21] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.
[22] R. Hiptmair, *Analysis of multilevel methods for eddy current problems*, Math. Comp., 72 (2003), pp. 1281–1303.
[23] R. Hiptmair, *Coupling of finite elements and boundary elements in electromagnetic scattering*, SIAM J. Numer. Anal., 41 (2003), pp. 919–944.
[24] R. Hiptmair and A. Toselli, *Overlapping and multilevel Schwarz methods for vector valued elliptic problems in three dimensions*, in Parallel Solution of Partial Differential Equations, P. Bjorstad and M. Luskin, eds., IMA Vol. Math. Appl. 120, Springer-Verlag, Berlin, 1999, pp. 181–202.
[25] R. Hiptmair, G. Widmer, and J. Zou, *Auxiliary space preconditioning in* $\mathbf{H}_0(\mathbf{curl}, \Omega)$, Numer. Math., 103 (2006), pp. 435–459.
[26] R. Hiptmair and W.-Y. Zheng, *Local multigrid in* $\mathbf{H}(\mathbf{curl})$, Tech. rep. 2007-03, SAM, ETH Zürich, Zürich, Switzerland, 2007.
[27] T. Kolev and P. Vassilevski, *Parallel* H^1*-based Auxiliary Space AMG Solver for* $H(\mathbf{curl})$ *Problems*, Report UCRL-TR-2222763, LLNL, Livermore, CA, 2006.
[28] T. Kolev and P. Vassilevski, *Some Experience with a* H^1*-based Auxiliary Space AMG for* $H(\mathbf{curl})$ *Problems*, Report UCRL-TR-221841, LLNL, Livermore, CA, 2006.
[29] J. Nédélec, *Mixed finite elements in* $\mathbb{R}^3$, Numer. Math., 35 (1980), pp. 315–341.
[30] J. Nédélec, *A new family of mixed finite elements in* R^3, Numer. Math., 50 (1986), pp. 57–81.
[31] S. Nepomnyaschikh, *Decomposition and fictitious domain methods for elliptic boundary value problems*, in Proceedings of the Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations, D. Keyes, T. Chan, G. Meurant, J. Scroggs, and R. Voigt, eds., SIAM, Philadelphia, 1992, pp. 62–72.
[32] J. Pasciak and J. Zhao, *Overlapping Schwarz methods in H(curl) on polyhedral domains*, J. Numer. Math., 10 (2002), pp. 221–234.
[33] S. Reitzinger and J. Schöberl, *Algebraic multigrid for edge elements*, Numer. Linear Algebra Appl., 9 (2002), pp. 223–238.
[34] L. R. Scott and Z. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.
[35] K. Stüben, *An Introduction to Algebraic Multigrid*, Academic Press, London, 2001, pp. 413–528.
[36] U. Trottenberg, C. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, London, 2000.
[37] J. Xu, *Iterative methods by space decomposition and subspace correction*, SIAM Rev., 34 (1992), pp. 581–613.
[38] J. Xu, *The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids*, Computing, 56 (1996), pp. 215–235.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2510–2536

ERROR BOUNDS FOR APPROXIMATE EIGENVALUES OF PERIODIC-COEFFICIENT LINEAR DELAY DIFFERENTIAL EQUATIONS*

ED BUELER†

Abstract. We describe a new Chebyshev spectral collocation method for systems of variable-coefficient linear delay differential equations with a single fixed delay. Computable uniform a posteriori bounds are given for this method. When the coefficients are periodic, the system has a unique compact nonnormal monodromy operator whose spectrum determines the stability of the system. The spectral method approximates this operator by a dense matrix of modest size. In cases where the coefficients are smooth we observe spectral convergence of the eigenvalues of that matrix to those of the operator. Our main result is a computable a posteriori bound on the eigenvalue approximation error in the case that the coefficients are analytic.

Key words. Chebyshev, collocation, delay differential equations, eigenvalues, monodromy operator, a posteriori, spectral methods

AMS subject classifications. 34K06, 34K20, 65Q05, 65F15, 65L60

DOI. 10.1137/050633330

1. Introduction.

Consider the linear delay differential equation (DDE)

$$\ddot{x} + \dot{x} + (\delta + 0.1\cos(\pi t))x = b\,x(t-2), \tag{1.1}$$

with $\delta, b \in \mathbb{R}$. This is a delayed, damped Mathieu equation [21].

> *Question*: For which values of δ, b is (1.1) asymptotically stable in the sense that all solutions go to zero as $t \to \infty$?

A practical and visual answer to this question, for $(\delta, b) \in [0, 20] \times [-10, 10]$ in (1.1), is the numerically produced *stability chart* in Figure 1.1.

Questions like this one arise in the stability analysis of many DDE models, including nonlinear models. Indeed, linear periodic DDEs frequently occur as the "variational equation" for perturbations of a periodic solution of a nonlinear DDE [18], and thus questions of this type are important in nonlinear dynamics as well. Stability charts like Figure 1.1 are useful in applications, including biology [25] and engineering [29] stability problems. In the context of machine tool vibrations, for instance, stability charts allow the choice of parameters which minimize regenerative vibrations and maximize throughput and quality [20, 37].

The stability chart in Figure 1.1 is produced pixel-by-pixel by numerically approximating the spectral radius of the compact monodromy operator associated to DDE (1.1). This operator is defined in section 2. Its eigenvalues are called *multipliers*. For each (δ, b) parameter pair there is a monodromy operator, and the pixel is marked as stable if the computed spectral radius is less than one. By "computed spectral radius" we mean this: Numerical methods applied to equations like (1.1) can usually be formulated as giving a square matrix which "approximates" the monodromy

*Received by the editors June 9, 2005; accepted for publication (in revised form) May 14, 2007; published electronically November 28, 2007. This work was supported in part by NSF grant 0114500.
http://www.siam.org/journals/sinum/45-6/63333.html
†Department of Mathematics and Statistics, University of Alaska, Fairbanks, AK 99775 (ffelb@uaf.edu).

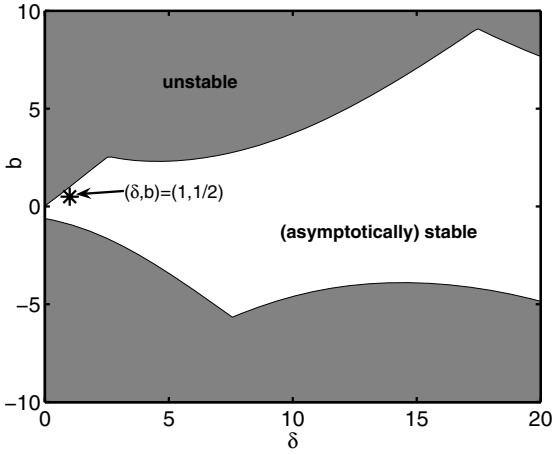


FIG. 1.1. *Stability chart for a delayed, damped Mathieu equation.*

operator, in the intended sense that the largest (magnitude) eigenvalues of the matrix approximate the largest multipliers of the operator [7, 10, 21, 24]. Standard numerical methods compute the eigenvalues of the matrix [17], and thus we get a computed spectral radius of the monodromy operator.

An example of the problem which motivates this paper is this: Based on Figure 1.1 we expect that the particular parameter pair $(\delta, b) = (1, 1/2)$ will be stable. In fact, for this parameter pair we get, from the spectral method introduced in this paper, the numerical value 0.612992319912 for the spectral radius of the monodromy operator. Based on many computations, by himself and others, of the eigenvalues of monodromy operators (the "multipliers") of periodic linear DDEs using spectral methods [8, 15, 21, 24], the author believes this value is trustworthy. In fact, the author believes that all twelve digits given are correct. But can one actually bound the difference between the actual multipliers and their approximations? Is the bound computable, and is it close to the actual error?

Figure 1.2 shows what we want, namely, to have the computed large eigenvalues of the matrix approximation within "error circles" in which the actual multipliers are known to reside. This figure results from applying our main result, which will be stated in Theorem I. It gives the desired kind of computable bounds for a class of DDEs which includes (1.1). We see, in particular, that the $(\delta, b) = (1, 1/2)$ case of (1.1) is stable because all error circles are within the unit disc.

This paper is concerned with linear periodic-coefficient DDEs with fixed delays. For such equations there is a Floquet theory [30] somewhat analogous to that for ordinary differential equations (ODEs). For constant-coefficient DDEs the multipliers can be determined, in theory or numerically, by finding the roots of a characteristic equation. For linear, periodic-coefficient DDEs with "integer" delays [19], if a Floquet transition matrix for certain linear ODEs associated to the DDE can be found exactly, then complex variable techniques can also in theory determine the multipliers [19, section 8.3]. (The Floquet transition matrix is the fundamental solution evaluated at one period of the coefficients of a periodic-coefficient linear ODE.) In general, however, ODE fundamental solutions must themselves be approximated. Furthermore, an approximation of the monodromy operator itself is of interest in many problems,

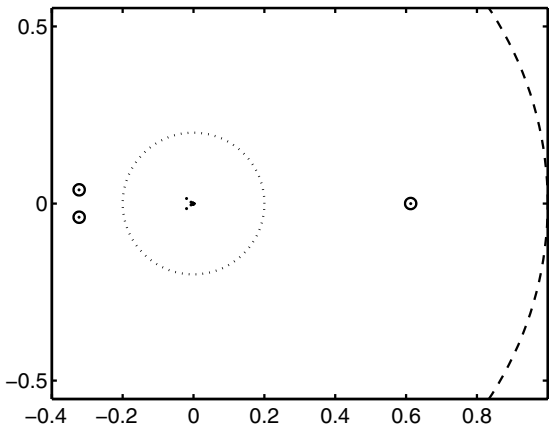


FIG. 1.2. *Computed multipliers (dots) of* (1.1) *with* $(\delta, b) = (1, 1/2)$. *Application of the technique of this paper (with* 76 *Chebyshev collocation points, in particular) gives an error radius of* 0.016 *around the computed multipliers (solid circles) for those actual multipliers* μ *such that* $|\mu| \geq 0.2$ *(dotted circle). The equation is seen to be stable because all eigenvalues are within the unit circle (dashed).*

including in cases where the eigenfunctions of the operator are important.

The DDEs addressed in this paper are d-dimensional systems of the form

$$\dot{y} = A(t)y + B(t)y(t-\tau), \tag{1.2}$$

where $\tau > 0$ is fixed and where $A(t), B(t)$ are $d \times d$ continuous matrix-valued coefficients. Higher-order scalar equations like (1.1) can, of course, be written as first-order systems of form (1.2). The Chebyshev collocation method for initial value problems (IVPs), described next and in section 3, makes no further restrictions on the form of the problem. To possess a unique monodromy operator, however, the coefficients A, B of the DDE must have common period T. We also assume $T = \tau$. The $T = \tau$ case is of importance in applications [20], and it is a technically easier first case in which to address the computation of error bounds on computed multipliers. To estimate the eigenvalue approximation error for our spectral method we will require the coefficients A, B to be analytic. Without some strong smoothness assumption we cannot, of course, expect spectral convergence of either solutions or eigenvalues.

This paper presents several new techniques for the numerical analysis of DDEs.

First we describe a new Chebyshev spectral collocation method for solving DDE IVPs of the form (1.2). The modest amount of notation necessary to describe the method in detail is given in section 3, but the essentials of the method can be communicated without it. Indeed, we choose N collocation points and then replace the DDE (1.2) by its collocation approximation

$$\hat{D}_N v = \hat{M}_A v + \hat{M}_B w, \tag{1.3}$$

where $\hat{D}_N$, $\hat{M}_A$, and $\hat{M}_B$ are square matrices of typically modest size—e.g., 30 to 300 rows—defined in section 3, and where v, w are vectors of collocation values of $y(t)$ and $y(t-\tau)$, respectively. The solution of the linear system (1.3) by Gauss elimination (or other dense matrix methods) completes the spectral method. We give computable a posteriori error bounds on the uniform error of this method (Theorem 3.4).

Second, we make an observation which is almost trivial, but which nonetheless gives a new spectral method for approximating the monodromy operator associated to DDE (1.2). We observe that the matrix which computes v from w, in (1.3), is a matrix approximation of the monodromy operator:

$$U_N = \left(\hat{D}_N - \hat{M}_A\right)^{-1} \hat{M}_B. \tag{1.4}$$

We have seen in many examples [9, 10] that the eigenvalues of U_N are excellent approximations of the multipliers.

In fact, spectral methods have been applied to DDEs many times in the past [5, 13, 23, 24]. Though the above specific spectral method is new to our knowledge, in that it uses collocation at the Chebyshev extreme points, the results just stated should not be surprising to many readers.

Our third new technique is very different from the existing literature, however. It says that one can compute bounds on the errors in the eigenvalues of U_N.

THEOREM I (Theorem 6.4 gives precise statement and proof). *If the coefficient functions $A(t)$, $B(t)$ in* (1.2) *are analytic on a closed interval of length T, and if the delay is equal to the period ($\tau = T$), then the distance from a large multiplier μ to the nearest of the computed eigenvalues λ_i of U_N is bounded*:

$$\min |\mu - \lambda_i| \leq \omega \ \mathrm{cond}(\tilde{V}).$$

Here "$\mathrm{cond}(\tilde{V})$*" is a computable condition number coming from a numerical diagonalization of U_N. The other factor ω comes from a combination of a priori bounds, derived from properties of the coefficients A, B in* (1.2), *and a posteriori estimates on IVPs for* (1.2) *with predetermined initial functions (namely, Chebyshev polynomials).*

We observe in many examples that ω decays exponentially, while $\mathrm{cond}(\tilde{V})$ grows slowly, as the number N of Chebyshev collocation points increases. Thus the estimate is usably small in these cases, and we give two examples here.

The a posteriori estimates on IVPs just mentioned come from an apparently new kind of result for spectral methods applied to ODEs (Theorem 3.4). A new "bootstrapping" method for achieving good a posteriori bounds on the growth of fundamental solutions to variable-coefficient ODEs is given in section 4. To prove Theorem I we also give an apparently new theorem on eigenvalue perturbation for diagonalizable operators on Hilbert spaces (Theorem 5.1 and Corollary 5.2). These are all technical results for our current purposes, but they may be of independent interest.

As noted, we address the case where the period of the coefficients T is equal to the (single, fixed) delay τ. A preprint by the author [8] sketches the generalization of the results to the situation in which $T \geq \tau$. Further generalization to multiple "integer" delays [19] presents only resolvable bookkeeping difficulties [9].

Regarding previous work in this area, spectral methods were first applied to DDE IVPs in [5] and [23]. More recently, collocation using the Gauss–Legendre points has been applied to the problem of finding periodic solutions to nonlinear DDEs in [13]. (Note that our problems have periodic coefficients but they do not, generally, have periodic solutions.) There is also a substantial engineering literature addressing stability—as opposed to solutions of IVPs—of ODEs and DDEs by spectral methods. For ODE problems, Sinha and Wu [28] use a Chebyshev Galerkin method, for example. Engineering applications led the author and colleagues to use spectral methods to address DDE stability questions [10]. Luzyanina and Engelborghs address the

application of spectral methods, among others, to the stability of linear DDEs and functional differential equations, and numerical evidence for spectral convergence of the multipliers is given [24, Table 2].

A recent paper by Breda, Maset, and Vermiglio [7] shows that for constant-coefficient DDEs with fixed discrete and distributed delays, a class of Runge–Kutta methods approximate the multipliers with a polynomial (in the mesh/step size) rate of convergence; spectral convergence does not, of course, occur. No computable error bounds are given.

The Chebyshev spectral method used here for DDEs first appeared in preprint form in [8]. Gilsinn and Potra [15] have recently sketched an a priori proof of the convergence of the numerical multipliers by this method. In [15] convergence is under mesh-refinement ("h-refinement"), based on the proof techniques of [5]. The regularity of the DDE to which the convergence argument applies is unspecified. No estimate of the rate of convergence is given.

Mesh-refinement has not been considered in the current paper, as one cannot achieve the spectral convergence necessary to have good eigenvalue estimation; in this paper we address only "p-refinement," where one increases the degree of polynomial approximation.

We believe that this paper represents the first quantitative technique for bounding the eigenvalue approximation error for a large class of continuous, infinite-dimensional systems whose dynamics are represented by *nonnormal* compact operators which are *not* already represented as integral operators with known kernels (compare [1]), or as Toeplitz operators (see [35] and the literature cited there), or which come from constant-coefficient equations (e.g., [7]). Note that the monodromy operators here are described by formula (2.4), so that they are operators of the form (finite rank operator) *plus* (integral operator). Finding the kernel of the integral operator part generally requires approximation of the fundamental solution of a variable-coefficient ODE, however. This is itself a nontrivial task in general [22].

2. The monodromy operator for a linear, periodic DDE. Consider the linear DDE system

$$\dot{y}(t) = A(t)\, y(t) + B(t)\, y(t-2) \tag{2.1}$$

for $y(t) \in \mathbb{C}^d$. Suppose A, B are continuous matrix-valued functions on $I = [-1, 1]$ which extend to not-necessarily-continuous periodic functions on the real line with period $T = 2$. The normalization of the period and the delay to 2, and the choice of interval $[-1, 1]$, are convenient for describing the Chebyshev collocation method (below). This normalization can be achieved by scaling and shifting the independent variable t in any linear periodic DDE with $T = \tau$, of course.

The *monodromy operator* U for (2.1) is defined as follows by considering the IVP [18]. Suppose $y(t)$ solves (2.1) with initial condition $y(t) = f(t)$ for $t \in I$; $y(t)$ is then defined for $t \in [-1, \infty)$. For $s \geq 1$, define $y_s(t) = y(t + s - 1)$ for $t \in I$; this is standard shift notation for DDEs [18]. Note that y_s is a $\mathbb{C}^d$-valued function defined on I and that $y_1 = f$. Define U to act on a soon-to-be-specified space of functions on I:

$$Uf = y_3. \tag{2.2}$$

Note U maps a function on I back to a function on I; such a form is essential if U is to have a spectrum.

Thus the monodromy operator acts by solving an ODE IVP

$$\dot{z}(t) = A(t)z(t) + B(t)f(t), \qquad z(-1) = f(1), \tag{2.3}$$

for $t \in I$, to give $Uf = z$. (Also, $z = y_3$, but we want to clearly show the ODE IVP.) This linear, nonhomogeneous ODE problem can be solved by integration *if* the solution to the corresponding homogeneous ODE problem is known. In fact,

$$(Uf)(t) = \Phi_A(t)f(1) + \int_{-1}^{t} \Phi_A(t)\Phi_A(s)^{-1}B(s)f(s)\,ds, \tag{2.4}$$

where $\Phi_A(t)$ is the fundamental solution of $\dot{z} = A(t)z$. (By definition $\Phi_A(t)$ solves $\dot{\Phi}_A = A\Phi_A$ on I and $\Phi_A(-1) = I_d$, where I_d is the identity on $\mathbb{C}^d$.) Note that the second summand in this formula for the monodromy operator is an integral operator of Volterra type. Knowledge of the kernel of this integral operator, namely, $k(t,s) = \Phi_A(t)\Phi_A(s)^{-1}B(s)$, generally requires the numerical solution of an ODE problem, however.

We emphasize that $y_{2n+1} = U^n f$ solves, by "the method of steps," the IVP consisting of (2.1) and $y(t) = f(t)$, $t \in I$. More precisely, if $2n - 1 \le t \le 2n + 1$, then $y(t) = (U^n f)(t - 2n)$, and so $y(t)$ is determined by steps for all of $[-1, \infty)$. The (asymptotic) stability of (2.1) is therefore given by the condition that the spectral radius of U be less than unity. (On the other hand, the degree of nonnormality of U will determine how much caution is required in extracting meaning from the multipliers of the linearization (variational equation) of a nonlinear DDE around a periodic solution; compare [35].)

Note that the solutions of linear DDEs with periodic coefficients are *not* themselves periodic. Also, the solutions are generally only regular in pieces. For instance, even if the initial function f is smooth on the interval I, and if A, B are smooth on all of $\mathbb{R}$, nonetheless some derivative of $y(t)$ is generally not continuous at $t = 1$.

An alternative form for the monodromy operator uses a fundamental solution to the DDE itself [18, 15]. In the periodic-coefficient case of the current paper, formula (2.4) is equivalent to, but easier to use than, this alternate form. Our reasons for the "easier to use" claim are necessarily complicated, but they relate to the piecewise regularity of the eigenfunctions of the DDE problem. See Lemma 4.3 below.

We still have not specified the space on which U acts, and it is vital to choose a usable space. Inspection of formula (2.4) shows that if f is continuous on I, then Uf is continuous on I. It is appropriate to fix notation.

DEFINITION 2.1. *Let $C(I)$ be the space of scalar continuous functions on I. Let $\mathcal{C} = C(I) \otimes \mathbb{C}^d$ be the space of $\mathbb{C}^d$-valued continuous functions.*

Certainly a monodromy operator $U : \mathcal{C} \to \mathcal{C}$ is well defined from formula (2.4). Its spectrum determines the stability of DDE (2.1). On the other hand, it will be desirable to work in a Hilbert space because we need tools for eigenvalue perturbation. In addition, the output of U is more regular than the input. We will suffer no loss of generality if we restrict our attention to a Hilbert space which is a subspace of $\mathcal{C}$ because, in particular, the eigenfunctions of $U : \mathcal{C} \to \mathcal{C}$ will in every case be in our chosen subspace.

The Hilbert space we choose is defined via a well-behaved orthogonal basis of nonperiodic, smooth functions which do a good job of approximating functions in $\mathcal{C}$, namely, Chebyshev polynomials with a well-chosen normalization. As is well known, these polynomials also do an excellent job of interpolating smooth functions on I; see section 3. In the next two definitions, and the two lemmas which follow, we confine

ourselves to the scalar case ($d = 1$) for notational convenience. There will be no actual loss of generality (as explained after Lemma 2.5).

DEFINITION 2.2. *Let L^2 be the Hilbert space of complex-valued (measurable) functions f on I such that $\int_{-1}^{1} |f(t)|^2(1-t^2)^{-1/2}\,dt < \infty$, and define the inner product $\langle f, g\rangle_{L^2} = \int_{-1}^{1} \overline{f(t)}\, g(t)\, \left(1-t^2\right)^{-1/2}\,dt$. (Note that "$L^2$" will always refer to a weighted L^2 space.) The (L^2-)*normalized Chebyshev polynomials *are defined as $\hat{T}_0(t) = (1/\sqrt{\pi})T_0(t)$, $\hat{T}_k(t) = (\sqrt{2/\pi})T_k(t)$, where $T_k(t) = \cos(k \arccos t)$ are the* standard Chebyshev polynomials. *The set $\{\hat{T}_k\}_{k=0}^{\infty}$ is an orthonormal (ON) basis of L^2. For $f \in L^2$ let $\hat{f}_k = \langle \hat{T}_k, f\rangle_{L^2}$ be the (L^2-)*Chebyshev expansion coefficients *of f. Thus $f(t) = \sum_{k=0}^{\infty} \hat{f}_k \hat{T}_k(t)$, with convergence in L^2, and $\|f\|_{L^2}^2 = \sum_{k=0}^{\infty} |\hat{f}_k|^2$.*

Unfortunately, the monodromy operator U is not bounded on L^2. In particular, (2.4) refers to the point values of the input function. Therefore, we turn to a Hilbert subspace of $L^2 \cap \mathcal{C}$ introduced by Tadmor [31].

DEFINITION 2.3. *$H^1 := \{f \in L^2 | \sum_{k=0}^{\infty}(1+k)^2|\hat{f}_k|^2 < \infty\}$. For $f, g \in H^1$ we define the inner product $\langle f, g\rangle_{H^1} = \sum_{k=0}^{\infty}(1+k)^2 \overline{\hat{f}_k}\, \hat{g}_k$. Let $\tilde{T}_k(t) = (1+k)^{-1}\hat{T}_k(t)$ be the H^1-*normalized Chebyshev polynomials. *The set $\{\tilde{T}_k\}_{k=0}^{\infty}$ is an ON basis of H^1.*

The space H^1 is a Sobolev space. It is not actually equivalent to $W_T^{1,2} = \left\{f \in L^2 \middle| \|f\|_{L^2} + \|\dot{f}\|_{L^2} < \infty\right\}$ [31], but we will have no need for such an equivalence. Next, we see that $H^1 \subset C(I)$ and pointwise evaluation is bounded.

LEMMA 2.4. *If $f \in H^1$, then f is continuous (it has a continuous representative) and $|f(t)| \le 0.9062\|f\|_{H^1}$. In particular,*

$$\delta_1 f \equiv \sum_{k=0}^{\infty} \hat{f}_k \hat{T}_k(1) = \sum_{k=0}^{\infty} \int_{-1}^{1} \hat{T}_k(1)\hat{T}_k(t)(1-t^2)^{-1/2} f(t)\,dt \tag{2.5}$$

is a bounded linear functional $f \mapsto f(1)$ on H^1.

Proof. The proof is a standard exercise for Sobolev spaces. See [8] for a complete proof. ∎

To use H^1 we also need to know that if f is sufficiently regular on I, then $f \in H^1$. We give a criterion via the Fourier series of $f(\cos\theta)$. In fact, if $f \in C^1(I)$, then the even function $\underline{f}(\theta) = f(\cos\theta)$ is in $C^1_{per}[-\pi, \pi]$; that is, $\underline{f}$ can be periodically extended with period 2π to be C^1 on $\mathbb{R}$. For $k \in \mathbb{Z}$ let $\hat{\underline{f}}(k)$ be the kth Fourier coefficient of $\underline{f}$; that is, $\hat{\underline{f}}(k) = (2\pi)^{-1/2}\int_{-\pi}^{\pi} e^{-ik\theta}\underline{f}(\theta)\,d\theta$. Then $\hat{f}_k = \hat{\underline{f}}(k) = \hat{\underline{f}}(-k)$ for $k > 0$ and $\hat{f}_0 = 2^{-1/2}\hat{\underline{f}}(0)$. For functions g on $[-\pi, \pi]$ define the norm $\|g\|_F^2 := \int_{-\pi}^{\pi} |g(\theta)|^2\,d\theta$.

LEMMA 2.5.

$$\|f\|_{L^2}^2 = \frac{1}{2}\sum_{k=-\infty}^{\infty} |\hat{\underline{f}}(k)|^2 = \frac{1}{2}\|\underline{f}\|_F^2, \text{ and} \tag{2.6}$$

$$\|f\|_{H^1}^2 = \frac{1}{2}\sum_{k=-\infty}^{\infty} (1+k)^2|\hat{\underline{f}}(k)|^2 \le \|\underline{f}\|_F^2 + \|\underline{f}'\|_F^2. \tag{2.7}$$

Thus $C^1(I) \subset H^1 \subset C(I)$ and

$$\|f\|_{H^1}^2 \le 2\pi\|f\|_{\infty}^2 + 2\pi\|\dot{f}\|_{\infty}^2 \tag{2.8}$$

if $f \in C^1(I)$.

Proof. Again the proof is standard and is found in [8]. ∎

Now we return to the general $\mathbb{C}^d$-valued (nonscalar) case. By mild abuse of notation, we let L^2 be the space of $\mathbb{C}^d$-valued measurable functions f for which $\int_{-1}^1 |f(t)|^2(1-t^2)^{-1/2}\,dt < \infty$; there is a corresponding inner product. Again by abuse of the notation we let H^1 be the subspace of L^2 for which $\sum_{k=0}^\infty (1+k)^2|\hat{f}_k|^2 < \infty$, where $\hat{f}_k = \langle \hat{T}_k, f\rangle_{L^2} \in \mathbb{C}^d$. We give H^1 the obvious inner product.

Since Uf solves IVP (2.3), U maps $\mathcal{C}$ to C^1 and indeed $U : H^1 \to H^1$. We can extract from (2.4) a useful estimate of $\|U\|_{H^1}$.

LEMMA 2.6. *Suppose* $\left|\Phi_A(t)\Phi_A(s)^{-1}\right| \le C_A$ *for all* $-1 \le s \le t \le 1$. *Let* $a^2 = 1 + \|A\|_\infty^2$ *and* $c = 0.9062$. *Then*

$$\|U\|_{H^1} \le \sqrt{2\pi d}\left(caC_A + \|B\|_\infty\left(c^2 + \pi a^2 C_A^2/2\right)^{1/2}\right). \tag{2.9}$$

Proof. Suppose $f \in H^1$ and let $g(t) = \int_{-1}^t \Phi_A(s)^{-1}B(s)f(s)\,ds$. Note that $\|Uf\|_{H^1} \le \|\Phi_A(t)f(-1)\|_{H^1} + \|\Phi_A(t)g(t)\|_{H^1}$. Letting $\varphi(t) = \Phi_A(t)f(-1)$, we have a bound from Lemma 2.5:

$$\|\varphi\|_{H^1}^2 \le 2\pi\sum_{k=1}^d \|\varphi_k\|_\infty^2 + \|(A\varphi)_k\|_\infty^2 \le 2\pi d\, C_A^2 a^2 |f(-1)|^2.$$

On the other hand, if $\omega(t) = \Phi_A(t)g(t)$, then $\dot\omega = A\omega + Bf$, and so by Lemma 2.5,

$$\|\omega\|_{H^1}^2 = 2\pi\sum_{k=1}^d \|\omega_k\|_\infty^2 + \left(\|(A\omega)_k\|_\infty + \|(Bf)_k\|_\infty\right)^2.$$

But

$$\begin{aligned}|\omega(t)_k| \le |\omega(t)| &\le \int_{-1}^1 \max_{-1\le s\le t\le 1}\left|\Phi_A(t)\Phi_A(s)^{-1}\right|\,|B(s)|\,|f(s)|\,ds\\ &\le C_A\|B\|_\infty\int_{-1}^1 |f(s)|\,ds \le \sqrt{\frac{\pi}{2}}C_A\|B\|_\infty\|f\|_{L^2} \le \sqrt{\frac{\pi}{2}}C_A\|B\|_\infty\|f\|_{H^1}\end{aligned}$$

by the Cauchy–Schwarz inequality with weight $(1-s^2)^{-1/2}\,ds$. Similarly, $|(A(t)\omega(t))_k| \le \sqrt{\frac{\pi}{2}}\|A\|_\infty C_A\|B\|_\infty\|f\|_{H^1}$. On the other hand, $|(B(t)f(t))_k| \le \|B\|_\infty|f(t)| \le c\|B\|_\infty\|f\|_{H^1}$. Thus

$$\|\omega(t)\|_{H^1}^2 \le 2\pi d\|B\|_\infty^2\left(\frac{\pi}{2}C_A^2 + \frac{\pi}{2}\|A\|_\infty^2 C_A^2 + c^2\right)\|f\|_{H^1}^2. \qquad \square$$

3. A Chebyshev collocation method for linear DDEs (and ODEs). First we consider IVPs for $(d \ge 1)$-dimensional ODE systems of the form

$$\dot y(t) = A(t)y(t) + u(t) \tag{3.1}$$

for $y(t), u(t) \in \mathbb{C}^d$, and $A(t)$ a $d \times d$ matrix.

Recall that $I = [-1, 1]$ and that $\mathcal{C}$ denotes $\mathbb{C}^d$-valued continuous functions on I. Denote $f \in \mathcal{C}$ by $f = (f_1, \ldots, f_d)^\top$. For $f \in \mathcal{C}$ let $\|f\|_\infty = \max_{t\in I}|f(t)|$, where "$|\cdot|$" is the Euclidean norm on $\mathbb{C}^d$. Note that $|\cdot|$ induces a norm on $d \times d$ matrices, also denoted "$|\cdot|$." For continuous matrix-valued functions $A(t) = (a_{ij}(t))$ define $\|A\|_\infty = \max_{t\in I}|A(t)|$.

To present our spectral method we need to recall the basics of, and give notation for, polynomial interpolation and collocation at the Chebyshev points.

DEFINITION 3.1. *Let $\mathcal{P}_N \subset \mathcal{C}$ be the space of $\mathbb{C}^d$-valued polynomials of degree at most N. Note that $\mathcal{P}_N$ has dimension $l = d(N+1)$. The* Chebyshev collocation (extreme) points *in I are $t_j = \{\cos(j\pi/N)\}$ for $j = 0, 1, \dots, N$* [33]. *Note that $t_0 = 1 > t_1 > \cdots > t_N = -1$.*

On the function (vector) spaces $\mathcal{C}$ and $\mathcal{P}_N$ we have collocation operators as follows. Evaluation at $N+1$ Chebyshev collocation points produces a vector in $\mathbb{C}^l$. We regard evaluation of continuous functions as a linear operator $G_N : \mathcal{C} \to \mathbb{C}^l$:

(3.2)
$$G_N f = (f_1(t_0), \dots, f_d(t_0), f_1(t_1), \dots, f_d(t_1), \dots, \dots, \dots, f_1(t_N), \dots, f_d(t_N))^\top.$$

We will always order the scalar components of an element of $\mathbb{C}^l$ consistently with the output of G_N (if the element refers to the collocation values of a $\mathbb{C}^d$-valued function). Restricting G_N to polynomials gives a *bijection* $E_N : \mathcal{P}_N \to \mathbb{C}^l$. The inverse of E_N, namely, $P_N : \mathbb{C}^l \to \mathcal{P}_N$, "creates" a $\mathbb{C}^d$-valued polynomial from collocation values; see (3.3) below for a method for computing P_N.

The composition of P_N and G_N is the interpolation operator $I_N = P_N \circ G_N : \mathcal{C} \to \mathcal{P}_N$. That is, if $p = I_N f$ for $f \in \mathcal{C}$, then p is a $\mathbb{C}^d$-valued polynomial of degree N such that $p(t_j) = f(t_j)$.

One may implement polynomial interpolation by "discrete Chebyshev series." Concretely, suppose $f \in C(I)$ is a scalar function. Recall that $T_k(t) = \cos(k \arccos t)$ are the standard Chebyshev polynomials. Then $p = I_N(f)$ is given by

$$p(t) = \sum_{k=0}^{N} \tilde{f}_k T_k(t), \qquad \tilde{f}_k = \sum_{j=0}^{N} C_{kj} f(t_j), \qquad C_{kj} = \frac{2}{N\gamma_j\gamma_k} \cos\left(\frac{\pi jk}{N}\right), \tag{3.3}$$

where $\gamma_j = 2$ if $j = 0$ or $j = N$ and $\gamma_j = 1$ otherwise [26]. These formulas can also be regarded as computing $p = P_N(\{f(t_j)\})$ if $\{f(t_j)\}$ is an arbitrary vector in $\mathbb{C}^l$. They may be implemented by a modification of the fast Fourier transform (FFT) [33].

A fundamental observation is that interpolation at the Chebyshev collocation points is spectrally accurate *for analytic functions*. Indeed, it converges exponentially with a known constant as follows. Note that an ellipse as described in the next lemma always exists because the region of analyticity is open and contains I by assumption.

LEMMA 3.2 (see [31]). *Suppose f is analytic on the closed set $I = [-1, 1]$. That is, suppose f is analytic in an open region $R \subset \mathbb{C}$ such that $I \subset R$. There exist constants $c > 0$ and $0 \le \rho < 1$ such that for all $N \ge 1$ the interpolant $p = I_N f$ satisfies $\|f - p\|_\infty \le c\rho^N$. Indeed, if E is an ellipse with foci ± 1, and if the interior of E is contained in R, and if the semimajor/-minor axes of E are of length S and s, respectively, then $\|f - p\|_\infty \le c'(S + s)^{-N}$ for some $c' > 0$.*

The "Chebyshev (spectral) differentiation matrix" is the map

$$D_N = E_N \circ \frac{d}{dt} \circ P_N : \mathbb{C}^l \to \mathbb{C}^l.$$

The entries of D_N are exactly known for all N [33]. In MATLAB, using `cheb.m` from [33], $D_N =$ `kron(cheb(N),eye(d))`. The action of D_N can also be computed efficiently by a modification of the FFT [33].

The matrix D_N is not invertible; it has kernel of dimension d. We do not quite seek its inverse, however, though solving the differential equation (3.1) by the Chebyshev spectral method we are about to describe is a closely related task. We need to

incorporate the initial condition from (3.1) before inverting, so we define $\hat{D}_N$ as the invertible $l \times l$ matrix which is equal to D_N in its first $dN = l - d$ rows but has

$$(\hat{D}_N)_{jk} = \begin{cases} 0, & 1 \le k \le dN, \\ \delta_{j-dN,k-dN}, & dN+1 \le k \le l, \end{cases}$$

for $dN+1 \le j \le l$. That is, the last d rows of $\hat{D}_N$ are zeroed, except that the identity I_d is inserted in the lower right $d \times d$ block.

Now, recalling ODE (3.1), define a block-diagonal matrix and a vector

$$\hat{M}_A = \begin{pmatrix} A(t_0) & & & \\ & \ddots & & \\ & & A(t_{N-1}) & \\ & & & 0_d \end{pmatrix}, \qquad \hat{u} = \begin{pmatrix} u(t_0) \\ \vdots \\ u(t_{N-1}) \\ y_0 \end{pmatrix}.$$

Here "0_d" denotes the $d \times d$ zero matrix. Both $\hat{D}_N, \hat{M}_A$ are $l \times l$, while $\hat{u}$ is $l \times 1$.

Our approximation of the IVP for (3.1) is described by the following lemma. The proof comes immediately from the definitions of $\hat{D}_N$, $\hat{M}_A$, and $\hat{u}$.

LEMMA 3.3 (our spectral method for ODE IVPs). *Fix $N \ge 1$. Suppose $A(t)$ is a continuous $d \times d$ matrix-valued function of $t \in I$, $u(t) \in \mathcal{C}$, and $y_0 \in \mathbb{C}^d$. The following are equivalent:*

- *$p \in \mathcal{P}_N$ satisfies*

$$\dot{p}(t_j) = A(t_j)p(t_j) + u(t_j) \quad \text{for } 0 \le j \le N-1 \qquad \text{and} \qquad p(-1) = y_0;$$

- *$v \in \mathbb{C}^l$ satisfies*

$$\hat{D}_N v = \hat{M}_A v + \hat{u}. \tag{3.4}$$

The equivalence is $p = P_N v$ and $v = E_N p$.

In practice one computes $v = (\hat{D}_N - \hat{M}_A)^{-1}\hat{u}$ by Gauss elimination to solve ODE (3.1). For this spectral method we have the a posteriori estimates given in the next theorem.

THEOREM 3.4. *Suppose $A(t)$ is a continuous $d \times d$ matrix-valued function of $t \in I$, $u(t) \in \mathcal{C}$, and $y_0 \in \mathbb{C}^d$. Suppose $y \in \mathcal{C}$ satisfies the IVP (3.1). Suppose the fundamental solution satisfies the bound*

$$\left|\Phi_A(t)\Phi_A(s)^{-1}\right| \le C_A \text{ for all } -1 \le s \le t \le 1. \tag{3.5}$$

Let $N \ge 1$, let $p \in \mathcal{P}_N$ be the $\mathbb{C}^d$-valued polynomial described by Lemma 3.3, and let $R_p = \dot{p}(-1) - A(-1)y_0 - u(-1)$. Then

$$\|y - p\|_\infty \le 2C_A\Big[\|Ap - I_N(Ap)\|_\infty + \|u - I_N(u)\|_\infty + |R_p|\Big] \tag{3.6}$$

and

$$\|\dot{y} - \dot{p}\|_\infty \le (2\|A\|_\infty C_A + 1)\,\Big[\|Ap - I_N(Ap)\|_\infty + \|u - I_N(u)\|_\infty + |R_p|\Big]. \tag{3.7}$$

The proof of Theorem 3.4 will be given momentarily, but some comments are in order. The estimates on the right sides of (3.6) and (3.7) have the structure

(stiffness) [(sum of uniform interpolation errors) + (initial residual)].

In fact, we interpret C_A as an estimate of the fastest possible exponential change in solving $\dot y = A(t)y$, and we interpret R_p as the amount by which (3.1) is not satisfied at the initial time -1. Section 8 gives practical methods for evaluating the uniform interpolation errors in (3.6) and (3.7).

To prove Theorem 3.4 we need a bound on the degree N monic polynomial with roots $t_0, \dots, t_{N-1}$. Such polynomials are uniformly exponentially small as $N \to \infty$. A proof of the following lemma, which seems not to appear in previous literature, is given in [8].

LEMMA 3.5. *For $N \geq 1$ let $Q_N(t) = (t-t_0)\dots(t-t_{N-1})$. For $t = \cos\theta \in I = [-1,1]$, $Q_N(t) = 2^{1-N}(\cos(\theta)-1)\sin(N\theta)/\sin(\theta)$. Thus $Q_N(-1) = (-1)^N N\, 2^{2-N}$ and $\|Q_N\|_\infty = N\, 2^{2-N}$ on I.*

Proof of Theorem 3.4. Let $q = I_N(Ap)$ and $w = I_N(u)$. Since $r = \dot p - q - w \in \mathcal{P}_N$ and $r(t_j) = 0$ for $j = 0, \dots, N-1$ by Lemma 3.3, it follows that there is $z \in \mathbb{C}^d$ such that $r = zQ_N$. Evaluating at $t = -1$, we find $R_p = z(-1)^N N 2^{2-N}$ so $|z| = |R_p|\, 2^{N-2}/N$. On the other hand,

$$\dot y - \dot p = A(y-p) + (Ap - q) + (u - w) - zQ_N, \tag{3.8}$$

so

$$y(t) - p(t) = \int_{-1}^{t} \Phi_A(t)\Phi_A(s)^{-1}\Big[(Ap(s) - q(s)) + (u(s) - w(s)) - zQ_N(s)\Big]\, ds.$$

Note that $y(-1) = p(-1)$. Taking norms, and using Lemma 3.5 to see that $\|zQ_N\|_\infty \leq |R_p|$,

$$|y(t) - p(t)| \leq 2C_A\Big[\|Ap - q\|_\infty + \|u - w\|_\infty + |R_p|\Big].$$

This inequality implies (3.6). Finally, using (3.8) and Lemma 3.5, we find that (3.6) implies (3.7). ∎

Example 1. We apply Theorem 3.4 to the ODE IVP

$$\dot y = (2t+1)y + (2t+1)\sin(3(t^2+t)), \quad y(-1) = 1. \tag{3.9}$$

The solution can be computed exactly by integration, and note that the solution $y(t)$ is entire because the coefficients are entire. The effectiveness of estimate (3.6) is seen in Figure 3.1 below. As N increases, spectral convergence starts to occur only when polynomial interpolation can "handle" the (complex) exponential rates present in the fundamental solution $\Phi_A(t)$ and the nonhomogeneity $u(t)$ on the interval I; this happens at $N \approx 10$ here. The error $\|y-p\|_\infty$ decreases to approximately 10^{-14}, and this is the level of rounding error because $\|y\|_\infty = 9.35$. The main point, however, is that the estimate from (3.6) nicely follows the error, though with a slowly growing overestimation factor of roughly 10^3 when $N = 40$. Both the combined interpolation error (the sum $\|Ap - I_N(Ap)\|_\infty + \|u - I_N(u)\|_\infty$) and the initial residual R_p contribute nontrivially to the estimate. Note that we use the optimal bound $C_A = e^{9/4}$ when applying Theorem 3.4 in this example.

Now we return to DDE (2.1), which is our actual interest. Define

$$\hat M_B = \begin{pmatrix} B(t_0) & & & \\ & \ddots & & \\ & & B(t_{N-1}) & \\ I_d & & & 0_d \end{pmatrix},$$

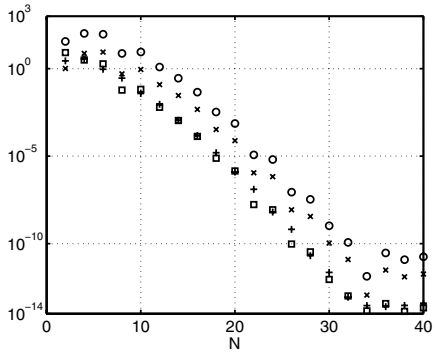


FIG. 3.1. *Theorem* 3.4 *applied to scalar ODE* (3.9) *reveals spectral convergence in uniform error* $\|y-p\|_\infty$ *on* I *(squares), where* y *is the solution and* p *is its polynomial approximation. The a posteriori estimates (circles) (the right side of* (3.6)*) closely track the actual error. The combined interpolation error ("+") and the initial residual ("×") are also shown; see the text.*

where I_d and 0_d are the $d \times d$ identity and zero matrices, respectively. The insertion of "I_d" in the lower left position represents the connection condition for the DDE method of steps. The Chebyshev collocation approximation of DDE (2.1), or, more particularly, of formula (2.3), is

$$\hat{D}_N v = \hat{M}_A v + \hat{M}_B w \tag{3.10}$$

if $v_j \approx z(t_j)$ and $w_j = f(t_j)$. Note the obvious fact that the IVP for DDE (2.1) is of the same form as ODE (3.1), but with $u = Bf$. Theorem 3.4 can therefore be applied to estimating errors in IVPs for (2.1).

The periodicity of the coefficients is not really relevant to solving a DDE of form (2.1) by the method of steps, when we use method (3.10). (Without periodicity there is no unique monodromy operator, however.) All that is required to use (3.10) to solve a nonperiodic DDE is that we must re-evaluate the collocated coefficient matrices $\hat{M}_A$ and $\hat{M}_B$ on each successive interval. The length of the interval would usually be equal to the delay if there is only one delay, in particular.

When the coefficients of (2.1) have period $T = \tau$, the matrix which approximates U is

$$U_N = \left(\hat{D}_N - \hat{M}_A\right)^{-1} \hat{M}_B \in \mathbb{C}^{l\times l}. \tag{3.11}$$

In practice we need to assume that the matrix inverse in (3.11) exists in order to find the entries of U_N, but we naturally observe the computed condition number for inversion. On the other hand, to compute the eigenvalues and eigenvectors of U_N we do not actually need the inverse. Vectors v such that $U_N v = \lambda v$ evidently solve the generalized eigenproblem

$$\hat{M}_B v = \lambda \left(\hat{D}_N - \hat{M}_A\right) v. \tag{3.12}$$

The eigenvalues can be approximated by the QZ algorithm [17], for instance.

4. Bounds on ODE fundamental solutions. Before we move on to the main goal of numerically approximating the monodromy operator and its eigenvalues, we

must find usable bounds on ODE fundamental solutions. In fact, we believe the results of this section are of independent interest for those needing computable bounds on the growth of solutions to variable-coefficient linear ODEs.

We recall existence and analytic continuation. The following result is proven by a straightforward Picard iteration argument [8].

LEMMA 4.1. *Suppose $A(z) \in \mathbb{C}^{d\times d}$ is (entrywise) analytic on a convex open set $E \supset I = [-1,1]$. Then there is a unique function $\Phi_A(z)$, analytic on E, satisfying*

$$\Phi_A(z) = I_d + \int_{-1}^{z} A(\zeta)\Phi_A(\zeta)\,d\zeta. \tag{4.1}$$

If $|A(z)| \le \alpha$ for $z \in E$, then $|\Phi_A(z)| \le e^{\alpha|z+1|}$ for $z \in E$. Furthermore, $\dot\Phi_A(t) = A(t)\Phi_A(t)$ for $t \in I$.

One can easily show, in addition, that the transition matrix $\Omega(t) = \Phi_A(t)\Phi_A(s)^{-1}$, for $-1 \le s \le t \le 1$, has an a priori estimate. (Note that Ω satisfies $\dot\Omega(t) = A(t)\Omega(t)$, $\Omega(s) = I_d$.) By Gronwall's inequality, in fact, $|\Omega(t)| \le e^{\tilde\alpha(t-s)}$ if $|A(\tau)| \le \tilde\alpha$ for $\tau \in [s,t]$. Unfortunately, this bound on $\Omega(t)$ is frequently not close to the actual maximum of $|\Omega(t)|$. We can, however, use the collocation algorithm to approximate the fundamental solution and then use a posteriori estimates from Theorem 3.4 to bound the fundamental solution. There is a "bootstrapping" aspect to such a technique: one must have some bound on the fundamental solution in order to apply Theorem 3.4, which leads to an improved bound. This is the content of the next lemma, for which one should recall that $\Psi_A(t) = (\Phi_A(t)^{-1})^\top$ satisfies the "adjoint equation" $\dot\Psi_A(t) = -A(t)^\top\Psi_A(t)$ with $\Psi_A(-1) = I_d$.

LEMMA 4.2. *Consider the following IVPs*:

$$\dot y_s(t) = A(t)y_s(t), \qquad y_s(-1) = e_s, \qquad s = 1,\dots,d, \tag{4.2}$$

where $\{e_s\}$ is the standard basis for $\mathbb{C}^d$.

Suppose that for each s we have a polynomial approximation p_s of y_s, and that we have estimates $\|y_s - p_s\|_\infty \le \mu_s$; see Theorem 3.4. Note that $\Phi_A(t) = \left[\, y_1(t) \;\middle|\; \dots \;\middle|\; y_d(t) \,\right]$ is the fundamental solution to $\dot y = A(t)y$. Let $\Phi_N(t) = \left[\, p_1(t) \;\middle|\; \dots \;\middle|\; p_d(t) \,\right]$, the approximation of $\Phi_A(t)$. If $\xi^2 = \sum_{s=1}^d \mu_s^2$, then $|\Phi_A(t) - \Phi_N(t)| \le \xi$ for all $t \in I$. Similarly, if $\Psi_A(t) = \left[\, z_1(t) \;\middle|\; \dots \;\middle|\; z_d(t) \,\right]$ is the fundamental solution to the adjoint equation $\dot z = -A(t)^\top z$, and if $\Psi_N(t) = \left[\, q_1(t) \;\middle|\; \dots \;\middle|\; q_d(t) \,\right]$, where $q_s(t)$ are polynomial approximations to $z_s(t)$ satisfying $\|z_s - q_s\|_\infty \le \nu_s$, so that $\Psi_N(t)$ is the collocation approximation of $\Psi_A(t)$, then $|\Psi_A(t) - \Psi_N(t)| \le \omega$ for $t \in I$, where $\omega^2 = \sum_{s=1}^d \nu_s^2$.

Also,

$$\left|\Phi_A(t)\Phi_A(s)^{-1}\right| \le \left(\xi + \|\Phi_N\|_\infty\right)\left(\omega + \|\Psi_N\|_\infty\right).$$

Proof. Note that

$$|\Phi_A(t)\Phi_A(s)^{-1}| \le \left(\|\Phi_A - \Phi_N\|_\infty + \|\Phi_N\|_\infty\right)\left(\|\Psi_A - \Psi_N\|_\infty + \|\Psi_N\|_\infty\right).$$

But

$$|\Phi_A(t) - \Phi_N(t)| = \max_{|u|=1}\left|\sum_{s=1}^d u_s\left(y_s(t) - p_s(t)\right)\right| \le \max_{|u|=1}|u|\left(\sum_{s=1}^d \mu_s^2\right)^{1/2} = \xi$$

using the Cauchy–Schwarz inequality, and similarly for $|\Psi_A(t) - \Psi_N(t)|$. ∎

Example 2. Consider the second-order ODE $\ddot{x} + \dot{x} + (10 + 9\cos(\pi t))x = 0$. It is a relatively stiff damped Mathieu equation corresponding to

$$A(t) = \begin{pmatrix} 0 & 1 \\ -10 - 9\cos(\pi t) & -1 \end{pmatrix}$$

in first-order form. A quick calculation gives $\|A\|_\infty \approx 19.026$. Thus $C_1 = e^{2\|A\|_\infty} \approx 3.5381 \times 10^{16}$ is the a priori bound on $|\Phi_A(t)\Phi_A(s)^{-1}|$ from Gronwall. We use the collocation algorithm with $N = 50$ to find $\Phi_N(t), \Psi_N(t)$ approximating $\Phi_A(t), \Psi_A(t)$. The a posteriori estimates from Theorem 3.4 are computed using $C_A = C_1$, and, as in Lemma 4.2, we find $|\Phi_A(t)\Phi_A(s)^{-1}| \le C_2 = 2.7117 \times 10^9$. This is a significant improvement, but also we can now iterate, using $C_A = C_2$ in the a posteriori estimates to generate C_3, and so on. The result is a sequence of bounds

$$|\Phi_A(t)\Phi_A(s)^{-1}| \le 3.5381 \times 10^{16},\ 2.7117 \times 10^9,\ 19.627,\ 19.587,\ 19.587, \ldots.$$

Each number on the right *is* a bound, with an a priori argument for the first and a posteriori arguments for the remainder. Evidently they converge superlinearly to about 19.6. By looking at the numerically approximated fundamental solution we see that this is a nearly optimal bound. In any case, such improvements by many orders of magnitude make further error estimation using Theorem 3.4 very practical.

In addition to the Lemma 2.6 estimate on the norm of U, we require, for the a posteriori estimation of eigenvalues, an a priori result on the polynomial approximation of eigenfunctions of U.

These eigenfunctions each solve a homogeneous ODE. In fact, if $Ux = \mu x$ for $\mu \in \mathbb{C}$, then $y = \mu x$ solves $\dot{y} = Ay + Bx$. Thus if $\mu \neq 0$, then $\dot{x} = \left(A + \mu^{-1}B\right)x$. Let $\Phi_\mu(t)$ be the fundamental solution to this ODE, so $\dot{\Phi}_\mu = (A + \mu^{-1}B)\Phi_\mu$ and $\Phi_\mu(-1) = I_d$. Note that $x(t) = \Phi_\mu(t)x(-1)$.

Suppose $A(z), B(z)$ are analytic on an ellipse E with foci ± 1. (If A, B are analytic on the compact set $I = [-1, 1]$, then they have such a common regularity ellipse.) Define $\|M\|_{\infty E} = \max_{z \in E} |M(z)|$ for continuous matrix-valued functions $M(z)$. From Lemma 4.1

$$|\Phi_\mu(z)| \le \exp\left(\|A + \mu^{-1}B\|_{\infty E}\,|z + 1|\right)$$

is an a priori bound on the analytic continuation of $\Phi_\mu(t)$ to $z \in E$. In fact, when we consider below those multipliers μ with magnitude greater than a chosen level $\sigma > 0$ we will use the bound

$$|\Phi_\mu(z)| \le C_\sigma := \exp\left(\left(\|A\|_{\infty E} + \sigma^{-1}\|B\|_{\infty E}\right)(S + 1)\right), \tag{4.3}$$

where S is the major semiaxis of a common regularity ellipse E for A and B. This a priori bound for $|\Phi_\mu(z)|$ turns out to be one of the two troublesome constants in the main Theorem 6.4 on multiplier estimation.

As announced, we now bound the interpolation error for eigenfunctions of U.

LEMMA 4.3. *Suppose A, B are analytic $d \times d$ matrix-valued functions with a common regularity ellipse $E \subset \mathbb{C}$ with foci ± 1 and sum of semiaxes $e^\eta = S + s > 1$. Suppose $Ux = \mu x$ for $\mu \neq 0$ and $\|x\|_{H^1} = 1$. Let $\Phi_\mu(z)$ be the unique analytic continuation of $\Phi_\mu(t)$ for $z \in E$ and suppose C_μ is a bound of $|\Phi_\mu(z)|$ for all $z \in E$. If $p = I_k x$, then*

$$\|x - p\|_{H^1} \le 8\sqrt{d}\,C_\mu(\sinh\eta)^{-1}k\,e^{-k\eta}. \tag{4.4}$$

Proof. First, $x(z) = \Phi_\mu(z)x(-1)$ is the analytic continuation of $x(t)$, $t \in I$, to $z \in E$. Furthermore, $|x(z)| \le C_\mu |x(-1)|$. It follows from (4.16) of [31] that

$$\|x - p\|_{H^1}^2 \le \sum_{j=1}^{d} \left(8C_\mu |x(-1)| (\sinh \eta)^{-1} k e^{-k\eta}\right)^2 = d \left(8C_\mu (\sinh \eta)^{-1} k e^{-k\eta}\right)^2 |x(-1)|^2.$$

Note that $|x(-1)| \le 0.9062 \|x\|_{H^1} = 0.9062$ by Lemma 2.4. ∎

Estimate (4.4), which bounds the polynomial approximation error for eigenfunctions of U, will be of great importance in controlling the *eigenvalue* approximation error from our spectral method (Theorem I).

DEFINITION 4.4. *Let $\sigma > 0$. The* a priori eigenfunction approximation error bound for large eigenvalues of U, *for $k+1$ point Chebyshev interpolation, is*

$$\epsilon_k = 8\sqrt{d} C_\sigma (\sinh \eta)^{-1} k\, e^{-k\eta}, \tag{4.5}$$

where $C_\sigma > 0$ is a bound on the analytic continuation of fundamental solutions:

$$|\Phi_\mu(z)| \le C_\sigma \quad \textit{for all } z \in E \textit{ and } |\mu| \le \sigma.$$

The crucial fact to note is that ϵ_k decays exponentially with increasing k and that the rate of decay is related to the size of the regularity ellipse E. Not surprisingly, in other words, if the coefficients of the DDE are more regular, then the spectral method is better at approximating the eigenfunctions. On the other hand, the constant C_σ generally increases rapidly as E is expanded.

5. Eigenvalue perturbation for operators on Hilbert spaces. The eigenvalue perturbation result in this section generalizes the well-known Bauer–Fike theorem for matrices [4]. It may not be new. It is fairly close to the definition of the eigenvalue condition number itself, introduced in [36]. It is also close to Proposition 1.15 in [11]. Nonetheless we need a precise statement, especially of Corollary 5.2, and we cannot find such in the literature.

Recall that a separable infinite-dimensional Hilbert space is isometrically isomorphic to the space of sequences $l^2 = \{a = (a_1, a_2, \dots) \,\big|\, a_j \in \mathbb{C}, \sum |a_j|^2 < \infty\}$. Denote the standard basis elements of l^2 by δ_j. By definition, a bounded operator $\Lambda \in \mathcal{L}(l^2)$ is *diagonal in the standard basis* if for each j, $\Lambda\delta_j = \lambda_j \delta_j$ for some $\lambda_j \in \mathbb{C}$.

THEOREM 5.1. *Let $\mathcal{X}$ be a separable infinite-dimensional complex Hilbert space. Suppose $A \in \mathcal{L}(\mathcal{X})$ is diagonalizable in the sense that there is a linear, bounded map $V : l^2 \to \mathcal{X}$ with bounded inverse and a diagonal (in the standard basis) operator $\Lambda \in \mathcal{L}(l^2)$ such that $A = V\Lambda V^{-1}$. Let $B \in \mathcal{L}(\mathcal{X})$. If $Bx = \mu x$ for $\mu \in \mathbb{C}$ and $\|x\| = 1$, then*

$$\min_{\lambda \in \sigma(A)} |\mu - \lambda| \le \|V\| \|V^{-1}\| \|(B - A)x\|. \tag{5.1}$$

Proof. If $\mu \in \sigma(A)$ there is nothing to prove. Otherwise, let $R_A = (A - \mu I)^{-1}$ and $R_\Lambda = (\Lambda - \mu I)^{-1}$. Note that $R_A = V R_\Lambda V^{-1}$ so that $\|R_\Lambda\|^{-1} \le \|V\|\, \|V^{-1}\| \,/\, \|R_A\|$. Also, R_Λ is diagonal and bounded with

$$\|R_\Lambda\| \le \sup_j |\lambda_j - \mu|^{-1} = \max_{\lambda \in \sigma(A)} |\lambda - \mu|^{-1} = \left(\min_{\lambda \in \sigma(A)} |\lambda - \mu|\right)^{-1},$$

where $\{\lambda_i\}$ are the diagonal entries of Λ. Thus

$$\min_{\lambda\in\sigma(A)} |\lambda-\mu| \le \|R_A\|^{-1}\|V\|\|V^{-1}\|. \tag{5.2}$$

On the other hand, if μ is an eigenvalue of B and $x \in H^1$ satisfies $Bx = \mu x$ and $\|x\| = 1$, then $(B-A)x = -(A-\mu I)x$, or $x = -(A-\mu I)^{-1}(B-A)x$. Taking norms, it follows that $1 \le \|R_A\|\|(B-A)x\|$. Combine this with inequality (5.2) to get estimate (5.1). ∎

An easy modification of Theorem 5.1 applies to $\mathcal{X} \cong \mathbb{C}^d$, and the result obviously implies the classical Bauer–Fike theorem. The hypotheses of Theorem 5.1 imply that $\sigma(A)$ is the closure of the set of diagonal entries of Λ. Because the spectrum $\sigma(A) \subset \mathbb{C}$ is compact for any bounded operator $A \in \mathcal{L}(\mathcal{X})$, the "min" in estimate (5.1) is appropriate.

Theorem 5.1 can be generalized to Banach spaces which are isomorphic to sequence spaces (i.e., l^p spaces). Specifically, we need to hypothesize an operator Λ, similar to A in the sense used in Theorem 5.1, for which $\|(\Lambda-\mu I)^{-1}\| \le \sup_{\lambda_i} |\lambda_i-\mu|^{-1}$, where $\{\lambda_i\}$ is a dense subset of $\sigma(A)$. Assuming A is similar to a diagonal operator on l^p suffices, for instance. (We will not use this generalization.)

Recall that if the coefficients A, B in DDE (2.1) are analytic, then there exist polynomials very close to the eigenfunctions of the monodromy operator U (see Lemma 4.3). This motivates the following corollary.

COROLLARY 5.2. *Suppose the hypotheses of Theorem* 5.1 *hold. Suppose also that* $p \in \mathcal{X}$ *satisfies* $\|x-p\| < \epsilon$. *Then*

$$\min_{\lambda\in\sigma(A)} |\mu-\lambda| \le \|V\|\|V^{-1}\|\Big[\epsilon(\|B\|+\|A\|) + \|(B-A)p\|\Big].$$

Proof. Note that $\|(B-A)x\| \le \|(B-A)(x-p)\| + \|(B-A)p\|$. ∎

Let us describe how this corollary will be used, and in so doing sketch the strategy for proving Theorem I. In section 3 we described a spectral method which produces an $l \times l$ matrix approximation U_N which approximates the monodromy operator U for DDE (2.1). We will numerically diagonalize U_N by standard methods for matrices of modest size, for instance, by the QR or QZ algorithms for eigenvalues and inverse iteration for eigenvectors [17]. Typically l is in the range 30 to 300 and U_N is dense. Thus it is a reasonable task to fully diagonalize U_N in practice. (Note that in stating our results we do not include errors made in these standard computations of finite-dimensional linear algebra.)

We want to know how close the unknown eigenvalues of U (the multipliers) are to the computed eigenvalues of U_N. The operators U and U_N act on different spaces, the former on a Hilbert–Sobolev space H^1 and the latter on $\mathbb{C}^l$. It turns out, however, that we can *extend* U_N to an operator $\tilde{U}_N$ which acts on H^1. Furthermore, it turns out (next section) that a diagonalization of U_N can be boosted to an operator diagonalization of $\tilde{U}_N$.

A numerical diagonalization of U_N is, of course, an invertible matrix $V \in \mathbb{C}^{l\times l}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{l\times l}$ such that $U_N = V\Lambda V^{-1}$. We do not assert that U_N is diagonalizable in every case. Rather, we assert that diagonalizability is the generic case [34]. We necessarily incorporate the conditioning of the eigenvalue problem for U_N, and specifically the condition number $\|\tilde{V}\|\|\tilde{V}^{-1}\|$ where $\tilde{V}$ is the operator formed from the eigenfunctions of $\tilde{U}_N$. The above informal definitions, made precise in the next section, allow us to outline our strategy for estimating the error in the approximate multipliers:

(i) given a DDE of form (2.1), compute U_N as described in section 3;
(ii) numerically diagonalize $U_N = V\Lambda V^{-1}$; denote the polynomial described by the kth column of V by "p_k";
(iii) thereby diagonalize $\tilde{U}_N = \tilde{V}\tilde{\Lambda}\tilde{V}^{-1}$, where $\tilde{U}_N$ is an operator on the same space as U; compute bounds for $\|\tilde{V}\|, \|\tilde{V}^{-1}\|$ (Lemma 6.2);
(iv) consider only the eigenvalues μ of U which satisfy $|\mu| \geq \sigma$ for some $\sigma > 0$, and, for the finitely many normalized eigenfunctions of U with multipliers μ satisfying $|\mu| \geq \sigma$, note the a priori bound on degree k ($\leq N$) polynomial interpolation at the Chebyshev points (Lemma 4.3 and Definition 4.4);
(v) write each p_k as a linear combination of the H^1-normalized Chebyshev polynomials $\tilde{T}_0, \ldots, \tilde{T}_N$;
(vi) by simply applying U_N, approximately solve DDE IVPs with each initial function $\tilde{T}_0, \ldots, \tilde{T}_N$; record a posteriori estimates from Theorem 3.4;
(vii) use Corollary 5.2 with $\mathcal{X} = H^1$, $A = \tilde{U}_N$, and $B = U$; estimate norm $\|U\|$ from Lemma 2.6; bound $\|(B-A)p\| = \|(U-\tilde{U}_N)p\|$ by estimates in steps (iv) and (vi);
(viii) conclude with an upper bound on the distance $\min|\mu - \lambda_i|$ as λ_i ranges over the approximate multipliers (Theorem 6.4).

6. Complete statement and proof of Theorem I. The expression (3.11) for U_N, or of the corresponding generalized eigenproblem (3.12), is all that is needed to rapidly approximate the monodromy operator and compute approximate multipliers. This section is devoted to the additional task of producing computable error estimates for the approximate multipliers.

We start with some definitions and a technical lemma in which we make the diagonalized $l \times l$ matrix U_N into a diagonalized operator $\tilde{U}_N$ which acts on H^1 (as does the monodromy operator U). This will allow us to apply Corollary 5.2 with $A = \tilde{U}_N$ and $B = U$.

Recall that the polynomials $\mathcal{P}_N$ are a subspace of our Hilbert space H^1. Let $\Pi_N \in \mathcal{L}(H^1)$ be orthogonal projection onto $\mathcal{P}_N$. Using the operator notation of section 3, define the (finite rank) approximate monodromy operator

$$\tilde{U}_N = P_N U_N E_N \Pi_N \in \mathcal{L}(H^1). \tag{6.1}$$

We now have two operators on H^1 (U and $\tilde{U}_N$) and an $l \times l$ matrix (U_N). Lemma 6.2 below shows that a numerical diagonalization of U_N yields a diagonalization of $\tilde{U}_N$. To prove this we will need to consider how to expand a $\mathbb{C}^d$-valued polynomial $p(t)$ of degree N into the basis $\{\tilde{T}_j \otimes e_s\}$ for $j = 0, \ldots, N$ and $s = 1, \ldots, d$, where $\{e_s\}$ is the standard basis for $\mathbb{C}^d$. In fact, if

$$p(t) = \sum_{\substack{j=0,\ldots,N \\ s=1,\ldots,d}} \gamma_{j,s}\tilde{T}_j(t) \otimes e_s = \sum_{j,\,s} \gamma_{j,s} \frac{z_j}{\sqrt{\pi}(1+j)} T_j(t) \otimes e_s, \tag{6.2}$$

where $z_j = \sqrt{2}$ if $j \geq 1$ and $z_0 = 1$, and if $\{e_s^*\}$ is the dual basis to $\{e_s\}$, then

$$\gamma_{j,s} = \sqrt{\pi}(1+j) z_j^{-1} \sum_{r=0}^{N} C_{jr} e_s^*\left(p(t_r)\right), \tag{6.3}$$

where $C = \{C_{jr}\}$ is the matrix in (3.3). Thus we can find the expansion coefficients of $p(t)$ given the values of $p(t)$ at the collocation points. Formulas (6.2) and (6.3) are

perhaps memorable if we call them "H^1-normalized discrete Chebyshev series for $\mathbb{C}^d$-valued polynomials." The next definition gives notation for the expansion coefficients associated to the diagonalization of U_N.

DEFINITION 6.1. *Suppose $U_N = V\Lambda V^{-1}$ with $\Lambda = (\lambda_j)_{j=1}^l$ diagonal. Let v_k be the kth column of V $(1 \le k \le l)$. Define $p_k(t) = P_N v_k$, a $\mathbb{C}^d$-valued polynomial of degree N. Expand $p_k(t)$ in discrete Chebyshev series*

$$p_k(t) = \sum \Gamma^V_{jd+s,k} \tilde{T}_j(t) \otimes e_s,$$

where the matrix of coefficients is defined by

$$\Gamma^V_{jd+s,k} = \sqrt{\pi}(1+j) z_j^{-1} \sum_{r=0}^{N} C_{jr} e_s^* \left(p_k(t_r)\right) = \sqrt{\pi}(1+j) z_j^{-1} \sum_{r=0}^{N} C_{jr} (v_k)_{rd+s},$$

where $C = \{C_{jr}\}$ is given by (3.3). *Equivalently,*

$$\Gamma^V = \left(\tilde{C} \otimes I_d\right) V, \text{ where } \tilde{C} = \sqrt{\pi/2}\, \operatorname{diag}\left(2^{-1/2}, 2, 3, \ldots, N+1\right) C.$$

Note that Γ^V is an invertible $(l \times l)$ matrix because V and C are invertible.

In the next lemma we diagonalize the operator $\tilde{U}_N$. Recall that our meaning for such a diagonalization is given in the statement of Theorem 5.1. Denote a typical element of l^2 by $a = (a_j^s)$, $j = 0, 1, \ldots$, $1 \le s \le d$, with $a_j^s \in \mathbb{C}$. Informally, the next lemma describes an operator $\tilde{V} : l^2 \to H^1$ which has matrix form

$$\tilde{V} = \left(\begin{array}{c|c} \Gamma^V & 0 \\ \hline 0 & I \end{array}\right)$$

in the basis $\{\tilde{T}_j \otimes e_s\}$ for H^1 and the standard basis for l^2. Here "I" is the isometry on rows $l+1, l+2, \ldots, \infty$.

LEMMA 6.2. *Let $\tilde{V} : l^2 \to H^1$ be defined by*

$$\tilde{V}a = \sum_{\substack{j=0,\ldots,N \\ s=1,\ldots,d}} a_j^s\, p_{jd+s}(t) + \sum_{\substack{j>N \\ s=1,\ldots,d}} a_j^s\, \tilde{T}_j(t) \otimes e_s.$$

Then $\tilde{V}$ is bounded and boundedly invertible, and, recalling that "$|\cdot|$" is the matrix 2-norm,

$$\|\tilde{V}\| \le |\Gamma^V| + 1, \qquad \|\tilde{V}^{-1}\| \le |(\Gamma^V)^{-1}| + 1. \tag{6.4}$$

Define $\tilde{\Lambda} \in \mathcal{L}(l^2)$ by $(\tilde{\Lambda}a)_{jd+s} = \lambda_{jd+s} a_j^s$ if $j < N$ and $s = 1, \ldots, d$, while if $k > l$ then $(\tilde{\Lambda}a)_k = 0$. Then $\tilde{\Lambda}$ is a diagonal operator on l^2 of rank at most l, and we have the diagonalization $\tilde{U}_N = \tilde{V}\tilde{\Lambda}\tilde{V}^{-1}$.

Proof. First we compute $\tilde{V}^{-1} : H^1 \to l^2$ by its action on the basis $\{\tilde{T}_j \otimes e_s\}$:

$$\tilde{V}^{-1}\left(\tilde{T}_j \otimes e_s\right) = \begin{cases} \left((\Gamma^V)^{-1}_{1,jd+s}, \ldots, (\Gamma^V)^{-1}_{l,jd+s}, 0, \ldots\right), & 0 \le j \le N, \\ \delta_{jd+s}, & j > N, \end{cases}$$

if $\{\delta_k\}_{k=1}^\infty$ is the standard basis for l^2. It is routine to check $\tilde{V}^{-1}\tilde{V}a = a$ for all $a \in l^2$. Estimates (6.4) follow from the definition of Γ^V and the fact that $\tilde{V}$ and $\tilde{V}^{-1}$ act isometrically between $\mathcal{P}_N^\top \subset H^1$ and the corresponding subspace of l^2.

Now, $U_N V = V\Lambda$ if and only if $U_N E_N p_k = \lambda_k E_N p_k$ for $k = 1, \dots, l$. Note that $\tilde{V}\delta_k = p_k$ if $k \le l$ and $\tilde{V}\delta_k = \tilde{T}_j \otimes e_s$, where $k = jd + s$, if $k > l$. Note that $k = jd + s > l$ if and only if $j > N$. Thus if $k \le l$, then

$$\tilde{U}_N \tilde{V}\delta_k = P_N U_N E_N \Pi_N p_k = P_N U_N E_N p_k = \lambda_k p_k = \left(\tilde{V}\Lambda\right)\delta_k,$$

while if $k > l$, then $\tilde{U}_N \tilde{V}\delta_k = P_N U_N E_N \Pi_N (\tilde{T}_j \otimes e_s) = 0$. Thus $\tilde{U}_N = \tilde{V}\tilde{\Lambda}\tilde{V}^{-1}$. ∎

We will use the obvious bound $\|\tilde{U}_N\| \le (\max_{1\le j\le l} |\lambda_j|)\, \|\tilde{V}\|\|\tilde{V}^{-1}\|$.

To get a posteriori estimates on eigenvalues we need a posteriori estimates on specific IVPs for DDE (2.1). The following lemma summarizes the application of Theorem 3.4 for this purpose.

LEMMA 6.3. *Suppose $f \in H^1$ and let $u(t) = B(t)f(t)$. Let $p \in \mathcal{P}_N$ be the approximation of y found from the spectral method applied to $\dot{y} = Ay + u$, $y(-1) = f(1)$ (Lemma 3.3). Then*

$$\|Uf - \tilde{U}_N f\|_{H^1} \le \sqrt{2\pi d}\left((2C_A)^2 + (2\|A\|_\infty C_A + 1)^2\right)^{1/2} Z, \tag{6.5}$$

where C_A satisfies (3.5) *and where*

$$Z = \|Ap - I_N(Ap)\|_\infty + \|Bf - I_N(Bf)\|_\infty + |\dot{p}(-1) - A(-1)f(1) - B(-1)f(-1)|.$$

Proof. Note that $Uf = y$ and $\tilde{U}_N f = p$. Now combine the result of Theorem 3.4 with inequality (2.8) in Lemma 2.5. ∎

We now give our main theorem on the multipliers of DDE (2.1).

THEOREM 6.4 (precise form of Theorem I). *Suppose that A, B in* (2.1) *are analytic $d \times d$ matrix-valued functions with common regularity ellipse $E \supset I$. Let $N \ge 1$ and $l = d(N+1)$. Let $\sigma > 0$ and recall the definition of the a priori eigenfunction approximation error bounds for large eigenvalues, denoted ϵ_k (Definition 4.4). Assume that U_N, defined by formula* (3.11), *is diagonalized: $U_N = V\Lambda V^{-1}$ with V invertible and $\Lambda = (\lambda_i)_{i=1}^l$ diagonal. Order the eigenvalues $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge 0$. Suppose we have a posteriori estimates for IVPs using the Chebyshev polynomials as initial functions:*

$$\|U(\tilde{T}_j \otimes e_s) - \tilde{U}_N(\tilde{T}_j \otimes e_s)\|_{H^1} \le \nu_j^s, \tag{6.6}$$

where $j = 0, \dots, N$ and $s = 1, \dots, d$ (Lemma 6.3; note that $\tilde{U}_N$ is defined by (6.1)*). Let $\xi_k^2 = \sum_{j=0}^k \sum_{s=1}^d (\nu_j^s)^2$.*

Consider a large eigenvalue $\mu \in \mathbb{C}$ of U: $Ux = \mu x$ for $x \in H^1$, $\|x\| = 1$, and $|\mu| \ge \sigma$. Let

$$\omega_k = \epsilon_k\left(\|U\| + |\lambda_1|\operatorname{cond}(\tilde{V})\right) + (1 + \epsilon_k)\xi_k$$

for $k = 1, \dots, N$. Then

$$\min_{i=1,\dots,l} |\mu - \lambda_i| \le \min\{\omega_1, \dots, \omega_N\}\operatorname{cond}(\tilde{V}). \tag{6.7}$$

Note that $\|U\|$ is estimated in Lemma 2.6 and $\operatorname{cond}(\tilde{V}) = \|\tilde{V}\|\|\tilde{V}^{-1}\|$ is estimated in Lemma 6.2.

Proof. By Lemma 4.3, for each $k = 1, \ldots, N$ we have $\|x - q_k\|_{H^1} < \epsilon_k$, where $q_k = I_k x$. Apply Corollary 5.2 with $\mathcal{X} = H^1$, $A = \tilde{U}_N$, $B = U$, $p = q_k$, and $\epsilon = \epsilon_k$:

$$\min |\mu - \lambda_i| \le \operatorname{cond}(\tilde{V}) \left(\epsilon_k (\|U\| + \|\tilde{U}_N\|) + \|(U - \tilde{U}_N) q_k\| \right). \tag{6.8}$$

On the other hand, $q_k = \sum_{j=0}^{k} \sum_{s=1}^{d} \alpha_k^{js} \tilde{T}_j \otimes e_s$ is a linear combination of Chebyshev polynomials, so from the a posteriori bounds ν_j^s,

$$\begin{aligned} \|(U - \tilde{U}_N) q_k\|_{H^1} &\le \sum_{j=0}^{k} \sum_{s=1}^{d} |\alpha_k^{js}| \nu_j^s \le \|q_k\|_{H^1} \xi_k \\ &\le \left(\|x\|_{H^1} + \|q_k - x_k\|_{H^1} \right) \xi_k \le (1 + \epsilon_k) \xi_k \end{aligned}$$

by Cauchy–Schwarz. From (6.8) and Lemma 6.2 we conclude with estimate (6.7). ∎

The idea in Theorem 6.4 is that all quantities on the right side of inequality (6.7) are known a priori, can be computed by a matrix eigenvalue package, or can be computed a posteriori (Theorem 3.4). Furthermore, the quantities ϵ_k and ξ_k are exponentially small, though for the latter this is in an a posteriori and fixed-precision-limited sense. Thus the computed eigenvalues of the matrix U_N can be shown to be exponentially close to those of the operator U as long as U_N is "well diagonalizable" in the sense that $\operatorname{cond}(\tilde{V})$ is small (which follows if $\operatorname{cond}(V)$ is small). Unfortunately, $\operatorname{cond}(\tilde{V})$, which depends on N, and C_σ, which does not, can be large. We observe below in two examples that $\operatorname{cond}(\tilde{V})$ grows slowly with N, however.

Bounds on fundamental solutions appear several times, at least implicitly, in the statement of Theorem 6.4. Recall that Lemma 2.6 estimates $\|U\|$ in terms of a bound C_A on $|\Phi_A(t)\Phi_A(s)^{-1}|$, $s, t \in I$, where $\Phi_A(t)$ is the fundamental solution to $\dot{x}(t) = A(t)x(t)$. Also, Theorem 3.4 uses C_A in the a posteriori bounds on IVPs. Fortunately, section 4 gives a technique for computing good bounds C_A. Thus the estimate for $\|U\|$ is reasonable and the a posteriori quantities ν_j^s are very small in practice (within a few orders of magnitude of machine precision). On the other hand, as noted, the bound C_σ on the analytic continuation $\Phi_\mu(z)$ of the fundamental solution $\Phi_\mu(t)$ to $\dot{x}(t) = (A(t) + B(t)/\mu)\, x(t)$, for the large multipliers with $|\mu| \ge \sigma$, which appears in Definition 4.4 of ϵ_k, is still a priori. Such a bound will inevitably be large. The result is that ϵ_k is large for small k but that ϵ_k decreases exponentially with k.

The a posteriori quantities ξ_k are, in practice, bounded below because precision is fixed. In Example 3 below using double precision, $\xi_k \approx 10^{-10}$ for $k << N$ and thus $\omega_k \ge 10^{-10}$. But $\operatorname{cond}(\tilde{V}) \approx 10^8$, so $\min\{\omega_k\} \operatorname{cond}(\tilde{V}) \approx 10^{-2}$, which gives only two digits of accuracy for the largest eigenvalue in that example. Application of Theorem 6.4 is a situation where 128-bit or 256-bit floating point representation would have substantial consequences.

7. An example of the application of Theorem I. Let us apply Theorem 6.4 to the damped delayed Mathieu equation (1.1).

Example 3. Consider the parameter values $(\delta, b) = (1, 1/2)$ in (1.1). Using the same method as in Example 2, we find an a priori bound $C_A^{(0)} = e^{2\sqrt{2+1.1^2}} \approx 35.99$ on the norm of $\Omega = \Phi_A(t)\Phi_A(s)^{-1}$. We improve this bound by a posteriori iterations as in section 4 to a new bound $C_A = 4.613$.

Recall that if x is an eigenfunction of U corresponding to eigenvalue μ, then $\dot{x} = (A + B/\mu)x$. Suppose we consider those eigenvalues $|\mu| \ge \sigma = 0.2$. Note that

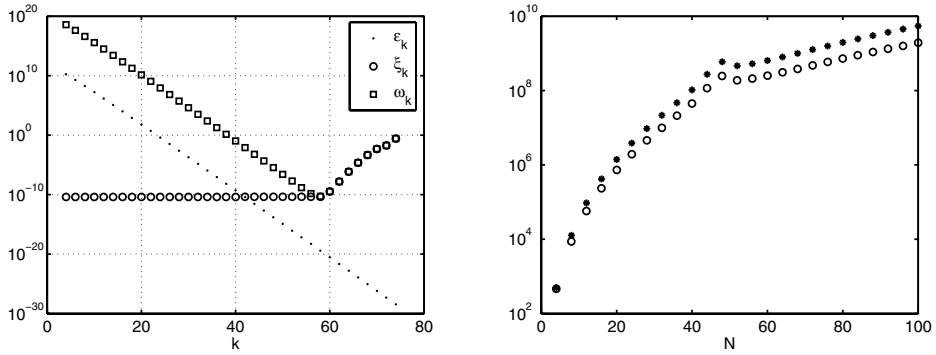


FIG. 7.1. *Behavior of constants (left figure) when Theorem* 6.4 *is applied to the delayed, damped Mathieu equation* (1.1) *with* $(\delta, b) = (1, 1/2)$ *and* $N = 75$. *The minimum of* ω_k *is about* 5×10^{-11}. *Right figure*: *The estimate on* $\operatorname{cond}(\tilde{V})$ *from Lemma* 6.2 *(circles) and the* 2*-norm condition number of* V *(dots), as functions of* N.

A, B are entire, so we are free to choose any common regularity ellipse $E \supset I$ with foci ± 1. The analytic continuation $\Phi_\mu(z)$ for $z \in E$ of the fundamental solution $\Phi_\mu(t)$ of $\dot{x} = (A + B/\mu)x$ is bounded by C_σ as in (4.3). Let $S = 1.996$ (a choice explained below) and $s = \sqrt{S^2 - 1}$ be semiaxes of E. Then from (4.3), $C_\sigma = 1.303 \times 10^{11}$; see below. Note that $e^\eta = S + s = 3.724$ determines the exponential rate of decrease of ϵ_k.

We now apply Theorem 6.4 with $N = 75$. The approximate multipliers are shown in Figure 1.2. The right side of (6.7) gives 0.01595, which is the radius of the small eigenvalue error bound circles in Figure 1.2.

The constants ϵ_k, ξ_k, and ω_k which appear in the statement of Theorem 6.4 are shown in Figure 7.1. For this value of N the condition number estimate for $\tilde{V}$ given in Lemma 6.2 is 3.5×10^8. The estimate of $\operatorname{cond}(\tilde{V})$ from Lemma 6.2 depends on N, as shown in Figure 7.1. We observe that it grows relatively slowly for large N. (The conditioning of the multipliers will be addressed in section 8.) In any case, the error bound (6.7) shows that the actual multipliers which exceed $\sigma = 0.2$ in magnitude lie within the error bound circles around the approximate multipliers in Figure 7.1. The DDE is, in particular, stable.

It is revealing to ask how to choose the regularity ellipse $E \supset I$ in the case when $A(z), B(z)$ are analytic in large neighborhoods of I or even when A, B are both entire. Such is the case if A, B are constant, for instance. The significance of the choice of E to Theorem 6.4 is that the sum of its semiaxes $e^\eta = S + s$ controls the rate of exponential decay of the a priori quantities ϵ_k, but also that the size of E affects the a priori bound C_σ which appears in Definition 4.4.

Fix $\sigma > 0$ and define C_σ as in Definition 4.4. Then, noting that the minor semiaxis $s = \sqrt{S^2 - 1}$ is a function of S, from (4.5) we find that ϵ_k is a function of S:

$$\epsilon_k(S) = 16\sqrt{d}\, k\, \frac{\exp\left([\|A\|_{\infty E} + \|B\|_{\infty E}/\sigma](S+1)\right)}{((S+s)^2 - 1)\ (S+s)^{k-1}}. \tag{7.1}$$

In general, $\|A\|_{\infty E}, \|B\|_{\infty E}$ depend on S.

Based on the evidence in examples like the one above, for sufficiently large N the a posteriori estimates ξ_k on IVPs first increase significantly at about $k = 3N/4$. We therefore choose S by seeking to minimize $\epsilon_{3N/4}(S)$.

Example 4 (continuation of Example 3). For (1.1) with $(\delta, b) = (1, 1/2)$ we have $|A(z)|_F^2 \le 2 + (1 + |\cos \pi z|)^2$. On E with foci ± 1 and semiaxes S, s, it follows that

$$\|A\|_{\infty E} \le \max_{z \in E} |A(z)|_F \le \left(2 + (1 + \sqrt{1 + e^{\pi s}})^2\right)^{1/2}.$$

The minor semiaxis s of E appears in this expression because the magnitude of $\cos(\pi z)$ grows exponentially in the imaginary direction. Note that $\|B\|_{\infty E} = 1/2$. To find the best S one must solve the problem

$$\min_{S>1} \quad \frac{\exp\left(\left[\left(2 + (1 + \sqrt{1 + e^{\pi s}})^2\right)^{1/2} + (2\sigma)^{-1}\right] (S+1)\right)}{((S+s)^2 - 1)\ (S+s)^{(3N/4)-1}}.$$

Upon inspection we see that this is a smooth and convex single-variable minimization problem. The minimum $S = 1.996$ is used above.

8. Discussion. We start with another example of the use of the spectral method to compute the multipliers of a DDE, and of Theorem 6.4 to estimate the errors in the multipliers.

Example 5. Consider the simple scalar DDE

$$\dot{x} = -x + x(t-2). \tag{8.1}$$

The multipliers are exactly known in the following sense [18, Theorem A.5]: The characteristic equation of (8.1) is $\lambda = -1 + e^{-2\lambda}$. The roots of this equation are the "exponents," and $\mu = e^{2\lambda}$ are the multipliers. Supposing $\lambda = \alpha + i\beta$, the characteristic equation can be written as a pair of real equations

$$\alpha = -1 + e^{-2\alpha} \cos 2\beta, \qquad \beta = -e^{-2\alpha} \sin 2\beta. \tag{8.2}$$

We find a single real exponent $\lambda_0 = 0$. All other exponents are complex and have imaginary parts β satisfying the transcendental equation $-\beta = \sin(2\beta) \exp(2 + 2\beta \cot 2\beta)$, which follows from (8.2). The solutions of this equation are easily seen to be simple roots β_k lying in the intervals $(k - 1/2)\pi < \beta_k < k\pi$ for $k = 1, 2, \ldots$. They can be found by the usual highly reliable numerical methods (e.g., bisection). The multipliers are then $\mu_0 = 1$ and $\mu_{\pm k} = \exp(\alpha_k \mp i\beta_k)$, where α_k is found from β_k by the first equation of (8.2). Note that these multipliers are already ordered by decreasing magnitude.

On the other hand, we can form U_N by the methods of this paper and find its $N+1$ eigenvalues $\mu_k^{(N)}$. The comparison of these numerical multipliers to the correct values for $N = 40$ ($N = 80$) is seen in Figure 8.1. We observe that the largest 10 (30, respectively) multipliers are accurate to full precision. The remaining smaller multipliers are quite inaccurate. This occurrence is no surprise as the eigenfunctions corresponding to the small multipliers are highly oscillatory and exponentially damped. (It is well known that Chebyshev interpolation requires somewhat more than 2 points per wavelength to achieve spectral accuracy [33, section 7].)

How does the impressive accuracy of the largest approximate eigenvalues compare to estimate (6.7) from Theorem 6.4? As also shown in Figure 8.1, if $\sigma = 0.1$, then estimate (6.7) decreases exponentially as a function of N to a minimum of 10^{-7} or so at $N \approx 55$. Thus the estimate remains roughly 8 orders of magnitude above the actual errors in the large multipliers. This difference has two sources. First, the a posteriori error estimates (6.6) coming from IVPs are too large by a few orders of

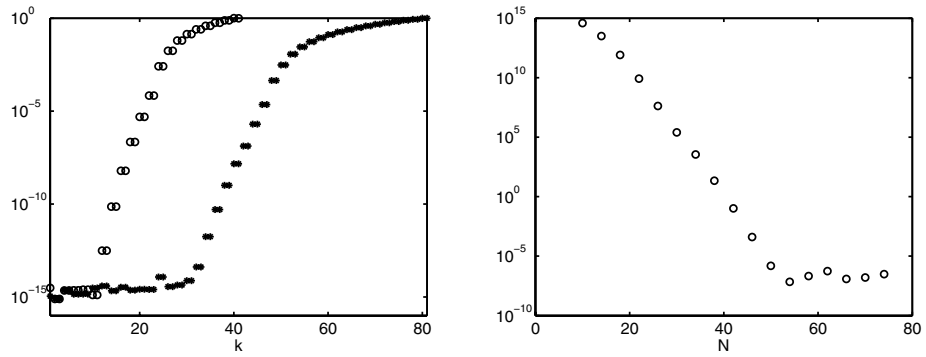


FIG. 8.1. *Left: Error in numerical multipliers for DDE* (8.1) *when* $N = 40$ *(circles) and* $N = 80$ *(dots). Right: Error bound* (6.7) *from Theorem* 6.4*, as a function of* N*, when* $\sigma = 0.1$.

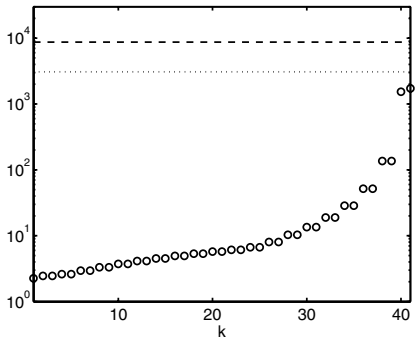


FIG. 8.2. *Condition numbers* $s_k^{(N)}$ *of the approximate multipliers* $\mu_k^{(N)}$ *for* $N = 40$; $\mathrm{cond}_2(V) = |V|\,|V^{-1}|$ *(dashed line) and estimate of* $\mathrm{cond}(\tilde{V})$ *(dotted line) from Lemma* 6.2 *are also shown.*

magnitude; compare Example 1 in section 3. Second, the estimate of $\mathrm{cond}(\tilde{V})$ is in the range 10^3–10^5 for $30 \le N \le 80$.

In fact, how badly conditioned are the computed multipliers in the above example? Recall that $s_k^{(N)} = |w_k^* v_k|^{-1}$ are the condition numbers of the eigenvalues $\mu_k^{(N)}$ of U_N if w_k and v_k are the associated normalized left and right eigenvectors, respectively [36, p. 68]. Figure 8.2 shows the computed $s_k^{(N)}$ for $N = 40$. We see that for $k \le N/2$ the condition numbers $s_k^{(N)}$ are small ($\approx 10^1$). For $k \approx N$, however, the condition numbers reach a maximum greater than 10^3. Recalling that these eigenvalues are ordered by decreasing size, we see that the difficult-to-approximate eigenvalues are the very small ones. These eigenvalues correspond to rapidly oscillating eigenfunctions for which we cannot expect our spectral method to do a good job anyway. Note that a priori estimates of these condition numbers, for the approximating matrix U_N or for the monodromy operator U, are not known.

A follow-up question is this: How does the conditioning of the individual eigenvalues relate to the matrix condition number used in Theorem 6.4? Figure 8.2 also shows $\mathrm{cond}_2(V) = |V|\,|V^{-1}|$ for the numerical diagonalization $U_N = V\Lambda V^{-1}$. We have an illustration of a quite general phenomenon: for any matrix, $\max s_k$ is always within a small factor of $\mathrm{cond}_2(V)$ [12]. Here the factor is about five.

The figure also shows our estimate of $\mathrm{cond}(\tilde{V}) = \|\tilde{V}\|\,\|\tilde{V}^{-1}\|$ from Lemma 6.2. We see that our estimate of $\mathrm{cond}(\tilde{V})$ is not problematic in this example. Indeed, $\max s_k \le$ (estimate of $\mathrm{cond}(\tilde{V})$ from Lemma 6.2) $\le \mathrm{cond}_2(V)$ in this and all other examples computed by the author. (A general proof of this fact is not known.)

We are led to a question which we can state somewhat imprecisely as follows: Let μ_k be the eigenvalues of a compact operator U, and let $\mu_k^{(N)}$ be the eigenvalues of a matrix U_N which approximates U. Is there a theorem which uses the computed condition numbers $s_k^{(N)}$ to bound the errors $|\mu_k - \mu_k^{(N)}|$? One might say that estimate (6.7) is uniform in k, because our analysis is based on a Bauer–Fike-type result (Corollary 5.2), while Figure 8.2 suggests $|\mu_k - \mu_k^{(N)}|$ should depend on k.

Let us address the practical computation of the bounds in this paper. First, a common task in computing the bound in Theorem 6.4 is the evaluation of a polynomial on the interval $I = [-1, 1]$ given its Chebyshev collocation values. This should be done by barycentric interpolation [27, 6]. A comparison of barycentric interpolation to more naive methods in computing the bounds here, which clearly shows the better performance of the barycentric method, is given in [8].

Also, in using Theorem 3.4 we need estimates of $\|\alpha\|_\infty$ where $\alpha(t)$ is a polynomial or an analytic function on I. For such norms a transform technique is useful. Consider the polynomial case first, as follows. Recall that $T_k(t) = \cos(k \arccos t)$ is the standard kth Chebyshev polynomial and that $|T_k(t)| \le 1$ for $t \in I$. If p is a polynomial and we want $\|p\|_\infty$, then we start by expanding $p(t) = \sum_{k=0}^{N} a_k T_k(t)$ by discrete Chebyshev series (3.3). It follows that

$$\|\alpha\|_\infty \le \sum_{k=0}^{N} |a_k|. \tag{8.3}$$

Estimate (8.3) uses the coefficients a_k in the Chebyshev expansion of p. These are easily calculated by the FFT [33]. Indeed, if `v` is a column vector of the collocation values of $p(t)$, then $\|p\|_\infty \le$ `sum(abs(coefft(v)))`, where "`coefft`" is the MATLAB function

```
function a = coefft(v)
   N = length(v)-1; if N==0, a=v; return, end
   U = fft([v; flipud(v(2:N))])/N;            % do t -> theta then FFT
   a = ([.5 ones(1,N-1) .5])'.*U(1:N+1);
```

If $\alpha(t)$ is analytic on I, then we estimate $\|\alpha\|_\infty$ by using (8.3) on a high degree polynomial interpolant of $\alpha(t)$. Concretely, we start with a modest value of M for Chebyshev interpolation and evaluate $\alpha(t)$ at collocation points $t_j^{(M)} = \cos(\pi j/M)$, $j = 0, \ldots, M$. Efficiency in the FFT suggests using M which is one less than a power of two, so perhaps $M = 15$. We use the FFT as above to calculate the coefficients a_k of the corresponding polynomial. We then determine whether the last few coefficients are small. For example, if $\max\{|a_{M-3}|, \ldots, |a_M|\} < 10\,\epsilon_m$, where ϵ_m is machine precision, then we accept the estimate $\|\alpha\|_\infty \approx \|p\|_\infty$ and use (8.3). If not, we double M—actually, $M_{\mathrm{new}} = 2(M_{\mathrm{old}} + 1) - 1$ for efficiency in the FFT—and try again. We might stop if $M = 2^{12} - 1$, for example. (A very similar issue to the one addressed in this paragraph appears in [3], and the technique here mimics the one there.)

Example 6. Consider estimating $\|u - I_5 u\|_\infty$ on I for $u(t) = \sin(2t)$. Our procedure with $\alpha = u - I_5 u$ stops at $N = 31$ with estimate $\|u - I_5 u\|_\infty \le 7.1 \times 10^{-4}$. By contrast, the use of 1000 equally spaced points in $[-1, 1]$ and barycentric interpolation of $p(t) = (I_5 u)(t)$ yields $\|u - I_5 u\|_\infty \approx 6.8 \times 10^{-4}$ at substantially greater cost.

A final technical concern is worth raising. It relates to the application of this paper to some of the periodic-coefficient linear DDEs which appear in practice. If DDE (2.1) came originally from an *approximate* periodic solution of a nonlinear DDE, as, for instance, when using the techniques of [13] or [14], then the period T is known only approximately (in general). The coefficients A, B are also known only approximately. Errors in the multipliers of the linear DDE in such cases may well be dominated by errors in computing the periodic solution to the original (nonlinear) DDE or by errors in computing the coefficients in (2.1).

9. Conclusion. We have introduced what we believe is a new spectral method for linear DDEs with a single fixed delay. The method uses collocation at the Chebyshev extreme points [33]. Extension of this method to multiple fixed delays is straightforward and has been implemented in MATLAB [9]. In the periodic-coefficient case, with period equal to the delay for technical convenience, we use the spectral method to compute a square matrix approximation to the monodromy operator associated to the DDE. The eigenvalues of this matrix approximation are seen in examples to be spectrally convergent to the eigenvalues of the monodromy operator (the multipliers).

Our main result uses new eigenvalue perturbation and a posteriori estimation techniques to give computable error bounds on the eigenvalue approximation error.

Now, one would obviously like an a priori proof of the spectral convergence of our collocation method when the coefficients of the DDE are analytic, but this is open. That is, one would want to show spectral convergence, as the degree of polynomial collocation increases, for the approximate solutions of ODEs and DDEs *and* for the approximate eigenvalues of the monodromy operator.

We can list some open questions related to the approximation of the multipliers by the technique of this paper:

- How can we get a better norm bound C_σ on $\Phi_\mu(z)$, the analytical continuation of the fundamental solution to $\dot{x} = (A + B/\mu)x$, for $|\mu| \geq \sigma > 0$ and z in a common regularity ellipse of A and B? (See Definition 4.4.)
- What is the best norm in which to compute $\operatorname{cond}(\tilde{V})$, the condition number of the infinite "matrix" of approximate eigenfunctions?

An additional question about the conditioning of the multipliers appears in section 8.

We have questions about the the monodromy operator U itself. For example, under what conditions does U actually diagonalize? Better yet, what a priori estimates can be computed for the conditioning of its eigenvalue problem? What can generally be said about the pseudospectra [32] of U? Do the eigenfunctions of U generically form a Riesz basis [2, 16] or some other strongly linearly independent basis for H^1? These latter questions do not affect our a posteriori methods, however.

Acknowledgments. This paper grows out of enjoyable work on engineering applications of DDEs initiated and sustained by Eric Butcher. Our ideas developed through work with Venkatesh Deshmukh and students Victoria Averina, Haitao Ma, Praveen Nindujarla, and Jacob Stroh. Correspondence with David Gilsinn on ideas and algorithms has been valuable and fun.

REFERENCES

[1] K. Atkinson, *Convergence rates for approximate eigenvalues of compact integral operators*, SIAM J. Numer. Anal., 12 (1975), pp. 213–222.

[2] S. A. Avdonin and O. P. Germanovich, *The basis property of a family of Floquet solutions of a linear periodic equation of neutral type in a Hilbert space*, Sibirsk. Mat. Zh., 36 (1995), pp. 992–997.
[3] Z. Battles and L. N. Trefethen, *An extension of MATLAB to continuous functions and operators*, SIAM J. Sci. Comput., 25 (2004), pp. 1743–1770.
[4] F. L. Bauer and C. T. Fike, *Norms and exclusion theorems*, Numer. Math., 2 (1960), pp. 137–141.
[5] A. Bellen, *One-step collocation for delayed differential equations*, J. Comput. Appl. Math., 10 (1984), pp. 275–283.
[6] J.-P. Berrut and L. N. Trefethen, *Barycentric Lagrange interpolation*, SIAM Rev., 46 (2004), pp. 501–517.
[7] D. Breda, S. Maset, and R. Vermiglio, *Computing the characteristic roots for delay differential equations*, IMA J. Numer. Anal., 24 (2004), pp. 1–19.
[8] E. Bueler, *Chebyshev Collocation for Linear, Periodic Ordinary and Delay Differential Equations: A Posteriori Estimates*, http://arxiv.org/ps_cache/math/pdf/0409/0409464v1.pdf, 2004.
[9] E. Bueler, *Guide to* `ddec`*: Stability of linear, periodic DDEs using the* `ddec` *suite of* MATLAB *codes*, http://www.dms.uaf.edu/∼bueler/DDEcharts.htm, 2005.
[10] E. A. Butcher, H. Ma, E. Bueler, V. Averina, and Z. Szabo, *Stability of linear time-periodic delay-differential equations via Chebyshev polynomials*, Internat. J. Numer. Methods Engrg., 59 (2004), pp. 895–922.
[11] F. Chatelin, *Spectral Approximation of Linear Operators*, Academic Press, New York, 1983.
[12] J. Demmel, *The condition number of equivalence transformations that block diagonalize matrix pencils*, SIAM J. Numer. Anal., 20 (1982), pp. 599–610.
[13] K. Engelborghs, T. Luzyanina, K. J. in 'T Hout, and D. Roose, *Collocation methods for the computation of periodic solutions of delay differential equations*, SIAM J. Sci. Comput., 22 (2000), pp. 1593–1609.
[14] D. E. Gilsinn, *Approximating limit cycles of a Van der Pol equation with delay*, in Dynamic Systems and Applications. Vol. 4, Dynamic, Atlanta, GA, 2004, pp. 270–276.
[15] D. E. Gilsinn and F. A. Potra, *Integral operators and delay differential equations*, J. Integral Equations Appl., 18 (2006), pp. 297–336.
[16] I. C. Gohberg and M. G. Kreĭn, *Introduction to the Theory of Linear Nonselfadjoint Operators*, Translated from the Russian by A. Feinstein. Translations of Mathematical Monographs 18, AMS, Providence, RI, 1969.
[17] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed., Johns Hopkins Studies in the Mathematical Sciences, Johns Hopkins University Press, Baltimore, MD, 1996.
[18] J. K. Hale, *Theory of Functional Differential Equations*, Appl. Math. Sci., 3, Springer-Verlag, New York, 1977.
[19] J. K. Hale and S. M. V. Lunel, *Introduction to Functional-Differential Equations*, Appl. Math. Sci. 99, Springer-Verlag, New York, 1993.
[20] T. Insperger and G. Stépán, *Stability of high-speed milling*, in Proceedings of the Symposium on Nonlinear Dynamics and Stochastic Mechanics (Orlando, 2000), Vol. AMD-241, ASME, New York, 2000, pp. 119–123.
[21] T. Insperger and G. Stépán, *Stability chart for the delayed Mathieu equation*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., 458 (2002), pp. 1989–1998.
[22] A. Iserles, *Expansions that grow on trees*, Notices Amer. Math. Soc., 49 (2002), pp. 430–440.
[23] K. Ito, H. T. Tran, and A. Manitius, *A fully discrete spectral method for delay-differential equations*, SIAM J. Numer. Anal., 28 (1991), pp. 1121–1140.
[24] T. Luzyanina and K. Engelborghs, *Computing Floquet multipliers for functional differential equations*, Internat. J. Bifur. Chaos Appl. Sci. Engrg., 12 (2002), pp. 2977–2989.
[25] N. MacDonald, *Biological Delay Systems: Linear Stability Theory*, Cambridge Studies in Mathematical Biology 9, Cambridge University Press, Cambridge, UK, 1989.
[26] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in Fortran: The Art of Scientific Computing*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.
[27] H. E. Salzer, *Lagrangian interpolation at the Chebyshev points* $x_{n,\nu} = \cos(\nu\pi/n)$, $\nu = 0(1)n$; *some unnoted advantages*, Computer J., 15 (1972), pp. 156–159.
[28] S. C. Sinha and D.-H. Wu, *An efficient computational scheme for the analysis of periodic systems*, J. Sound Vibration, 151 (1991), pp. 91–117.
[29] G. Stépán, *Retarded Dynamical Systems: Stability and Characteristic Functions*, Pitman Res. Notes in Math. 210, Longman, Harlow, UK, 1989.

[30] A. Stokes, *A Floquet theory for functional differential equations*, Proc. Natl. Acad. Sci. USA, 48 (1962), pp. 1330–1334.
[31] E. Tadmor, *The exponential accuracy of Fourier and Chebyshev differencing methods*, SIAM J. Numer. Anal., 23 (1986), pp. 1–10.
[32] L. N. Trefethen, *Pseudospectra of linear operators*, SIAM Rev., 39 (1997), pp. 383–406.
[33] L. N. Trefethen, *Spectral Methods in MATLAB*, Software Environ. Tools 10, SIAM, Philadelphia, 2000.
[34] L. N. Trefethen and D. Bau, *Numerical Linear Algebra*, SIAM, Philadelphia, 1997.
[35] L. N. Trefethen and M. Embree, *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*, Princeton University Press, Princeton, NJ, 2005.
[36] J. H. Wilkinson, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, UK, 1965.
[37] M.-X. Zhao and B. Balachandran, *Dynamics and stability of the milling process*, Internat. J. Solids Structures, 38 (2001), pp. 2233–2248.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2537–2563

A LEAST-SQUARES FINITE ELEMENT METHOD FOR THE MAGNETOSTATIC PROBLEM IN A MULTIPLY CONNECTED LIPSCHITZ DOMAIN*

HUO-YUAN DUAN†, PING LIN†, P. SAIKRISHNAN†, AND ROGER C. E. TAN†

Abstract. A new least-squares finite element method is developed for the curl-div magnetostatic problem in Lipschitz and multiply connected domains filled with anisotropic nonhomogeneous materials. In order to deal with possibly low regularity of the solution, local L^2 projectors are introduced to standard least-squares formulation and applied to both curl and div operators. Coercivity is established by adding suitable mesh-dependent bilinear terms. As a result, any continuous finite elements (lower-order elements are enriched with suitable bubbles) can be employed. A desirable error bound is obtained: $||\mathbf{u}-\mathbf{u}_h||_0 \le C\,||\mathbf{u}-\tilde{\mathbf{u}}||_0$, where $\mathbf{u}_h$ and $\tilde{\mathbf{u}}$ are the finite element approximation and the finite element interpolant of the exact solution $\mathbf{u}$, respectively. Numerical tests confirm the theoretical results.

Key words. least-squares continuous finite element method, L^2 projector, curl-div magnetostatic problem

AMS subject classifications. 65M60, 74S05, 78A30, 65N30, 65N15

DOI. 10.1137/050640102

1. Introduction.

Recently, there has been increasing interest in seeking finite element solutions of Maxwell's equations; see [5, 7, 16, 21, 22, 23, 26] and references therein. As a typical model, the curl-div magnetostatic problem plays a central role in the study of finite element methods for Maxwell's equations and other mathematical subjects such as existence-uniqueness and regularity-singularity [1, 7, 15, 28]. Let us first recall this model. For a domain Ω of $\mathbb{R}^3$ filled with anisotropic nonhomogeneous materials described by a tensor ε, given two functions $\mathbf{g} \in (L^2(\Omega))^3$ and $f \in L^2(\Omega)$, with $\mathbf{u}$ the unknown field, the curl-div magnetostatic problem reads as follows [6, 7]:

$$\mathbf{curl}\,\mathbf{u} = \mathbf{g}, \quad \operatorname{div}(\varepsilon\,\mathbf{u}) = f \quad \text{in } \Omega. \tag{1.1}$$

In this paper we shall consider a least-squares C^0 finite element method to solve (1.1) numerically. Although (1.1) is simple in appearance, the C^0 finite element discretization is not straightforward generally. To illustrate this we consider a simple case below. Let Ω be a Lipschitz polyhedron of $\mathbb{R}^3$ with boundary $\partial\Omega$, $\varepsilon = 1$, and assuming a boundary condition $\mathbf{u}\cdot\mathbf{n} = 0$ on $\partial\Omega$. Let $X_T := \{\mathbf{v} \in (L^2(\Omega))^3; \mathbf{curl}\,\mathbf{v} \in (L^2(\Omega))^3, \operatorname{div}\mathbf{v} \in L^2(\Omega), \mathbf{v}\cdot\mathbf{n}|_{\partial\Omega} = 0\}$, equipped with the norm $||\mathbf{v}||^2_{X_T} = ||\mathbf{curl}\,\mathbf{v}||^2_0 + ||\operatorname{div}\mathbf{v}||^2_0 + ||\mathbf{v}||^2_0$, where $||\cdot||_0$ is the L^2 norm. Equation (1.1) can be formulated as a standard least-squares variational problem (see [22, 23]): Find $\mathbf{u} \in X_T$ such that, for all $\mathbf{v} \in X_T$,

$$\mathcal{L}(\mathbf{u},\mathbf{v}) := (\mathbf{curl}\,\mathbf{u}, \mathbf{curl}\,\mathbf{v}) + (\operatorname{div}\mathbf{u}, \operatorname{div}\mathbf{v}) = (\mathbf{g}, \mathbf{curl}\,\mathbf{v}) + (f, \operatorname{div}\mathbf{v}), \tag{1.2}$$

*Received by the editors September 12, 2005; accepted for publication (in revised form) May 17, 2007; published electronically November 28, 2007. This work was supported by NUS academic research grant R-146-000-064-112.

http://www.siam.org/journals/sinum/45-6/64010.html

†Department of Mathematics, National University of Singapore, 2 Science Drive 2, Singapore 117543, Republic of Singapore (scidhy@nus.edu.sg, matlinp@nus.edu.sg, saikrishnan.p@gmail.com, scitance@nus.edu.sg).

where $(\cdot,\cdot)$ denotes the L^2 inner product. As is well known, $\mathcal{L}$ is coercive on X_T with respect to the norm $||\cdot||_{X_T}$ (cf. [2, 15, 20]). Then it seems to be natural to employ the C^0 finite element method for problem (1.2). This is indeed true for smooth domains or for convex polyhedra (see [12, 22, 23, 27]). However, when the domain is nonsmooth (Ω contains reentrant corners or edges), it turns out that the C^0 finite element method of problem (1.2) does not usually yield correct approximations (cf. [5, 15, 16]). Here we provide an intuitive interpretation. A more accurate interpretation may be found in [5, 15, 16, 24]. Let $\mathbf{u}_h$ be the C^0 finite element solution of (1.2), where $h > 0$ denotes the mesh-parameter of the simplex partition of Ω. The classical C^0 finite element and interpolation theory [8, 13] leads to an error estimate: $||\mathbf{u}-\mathbf{u}_h||_{X_T} \le C\,||\mathbf{u}-\tilde{\mathbf{u}}||_{X_T} \le C\,h^r\,||\mathbf{u}||_{1+r}$, for all $r \ge 0$, where $\tilde{\mathbf{u}}$ is an interpolant of $\mathbf{u}$ in a continuous piecewise $\mathcal{P}_r$ polynomial of order not greater than r. This error estimate indicates that $\mathbf{u}$ should be at least in H^1 in order to have a convergence. But, for nonsmooth domains, $\mathbf{u}$ may not be in H^1 (see [15]).

Nonetheless, this is not necessarily the problem of the use of C^0 finite elements. In fact, any function in L^2 (even in L^1) can be well approximated by C^0 elements, and we have

$$||\mathbf{u}-\tilde{\mathbf{u}}||_0 \le C\,h^s\,||\mathbf{u}||_s \tag{1.3}$$

when $\mathbf{u}$ is in H^s with $0 \le s < 1$ (see [3, 8, 13, 14, 29]). Since C^0 finite elements are good enough for H^s $(0 \le s < 1)$ functions, the use of the formulation (1.2) is the source of trouble when C^0 finite elements are used for the problem with singular (nonsmooth) solutions whose singularities generally result from nonsmoothness of domains or heterogeneities of materials filling the domain or both. This motivates the design of more suitable formulations to replace (1.2) so that C^0 finite elements may concurrently work for nonsmooth solutions (not in H^1) as well as smooth solutions (at least in H^1). Meanwhile, new formulations should still have the same merits as (1.2): (a) the resulting linear system is symmetric and positive definite; (b) a globally continuous solution can be produced. See a survey [4] on least-squares finite element methods.

There are a few modified formulations available. A weighted formulation, proposed in [16], may be employed. See also an earlier paper [17]. The weighted least-squares method is theoretically and numerically proven to be convergent correctly. A property of this method is that a positive weight function of nonpolynomials is applied to the div operator appearing in (1.2). In two dimensions the weight function may be taken in the form of r^γ, where r is the distance to the reentrant corner with opening angle greater than π, and γ is an index that characterizes the singularity of the exact solution. It becomes, however, rather complicated to determine the weight function in three dimensions, due to the more complex characterization of the singularity information for three-dimensional domains. Another property is that the C^0 finite element space is required to contain the gradient of some C^1 finite element space. This excludes the use of some simpler C^0 finite elements (e.g., the linear element). Several C^1 finite element spaces are available in two dimensions (cf. [13]), but, to our knowledge, few C^1 finite elements are known in three dimensions. So it is unclear how to choose a reasonable three-dimensional C^0 finite element space. There are other least-squares methods available—for example, the FOSLL* method [11, 25] and the *negative* norm method [9, 10]. Of a scaled version of (1.1) (setting $\mathbf{u}^* = \varepsilon^{1/2}\,\mathbf{u}$ and $\mathbf{g} = \mathbf{0}$) the FOSLL* method first seeks the solution of a dual problem associated with the dual operator of the differential operator of the scaled problem of (1.1), and

then the solution $\mathbf{u}^*$ is obtained by differentiating the dual solution. This FOSLL* method [25] is in essence a scalar potential method and the approximation of $\mathbf{u}^*$ is always discontinuous because of the differentiation. The negative norm method [9], to accommodate the case of nonsmooth solutions, formulates (1.1) in suitable dual norms (assuming that Ω is simply connected, $\partial\,\Omega$ has no disconnected components, $\varepsilon = 1$, and the boundary condition $\mathbf{u}\cdot\mathbf{n}|_{\partial\,\Omega} = 0$):

$$(\mathbf{curl}\,\mathbf{u}, \mathbf{curl}\,\mathbf{v})_{(H^{-1}(\Omega))^3} + (\operatorname{div}\mathbf{u}, \operatorname{div}\mathbf{v})_{(H^1(\Omega))^*}$$
$$= (\mathbf{g}, \mathbf{curl}\,\mathbf{v})_{(H^{-1}(\Omega))^3} + (f, \operatorname{div}\mathbf{v})_{(H^1(\Omega))^*},$$

where $(H^{-1}(\Omega))^3$ is the dual space of $(H_0^1(\Omega))^3$, with $H_0^1(\Omega) = \{v \in H^1(\Omega); v|_{\partial\,\Omega} = 0\}$, and $(H^1(\Omega))^*$ is the dual space of $H^1(\Omega)$, and $(\cdot,\cdot)_{(H^{-1}(\Omega))^3}$ and $(\cdot,\cdot)_{(H^1(\Omega))^*}$ denote the inner products of $(H^{-1}(\Omega))^3$ and $(H^1(\Omega))^*$, respectively. In the discrete level this method is something like adding the inverse of the discrete Laplace operator or its preconditioner in front of both curl and div operators in (1.2). The error estimate of the finite element approximation for nonsmooth solutions may be obtained from this method, but at the expense of multiple applications of the inverse of the discrete Laplacian or its preconditioner. The programming is rather tricky in practice as well.

In this paper, we develop new least-squares methods with the use of C^0 finite elements. The main idea is to apply local L^2 projectors to both curl and div operators appearing in (1.2), with a few extra mesh-dependent stabilization terms added. Specifically, let R_h and $\breve{R}_h$ be L^2 projectors defined relative to L^2 inner products $(\cdot,\cdot)_h$, and let $S_h(\cdot,\cdot)$ be a mesh-dependent bilinear form; we define a new least-squares bilinear form:

$$\mathcal{L}_h(\mathbf{u},\mathbf{v}) = (R_h(\mathbf{curl}\,\mathbf{u}), R_h(\mathbf{curl}\,\mathbf{v}))_h + (\breve{R}_h(\operatorname{div}(\varepsilon\,\mathbf{u})), \breve{R}_h(\operatorname{div}(\varepsilon\,\mathbf{v})))_h + S_h(\mathbf{u},\mathbf{v}). \tag{1.4}$$

These R_h and $\breve{R}_h$ are defined as local L^2 projectors or pseudolocal L^2 projectors. Local L^2 projectors are defined element-by-element onto the discontinuous piecewise constant finite element spaces and pseudolocal L^2 projectors are defined onto the continuous piecewise linear finite element spaces with respect to the trapezoidal quadrature scheme of the standard L^2 inner product (Note that the L^2 projectors defined in this way are essentially local. See Remark 3.1 of this paper.) We prove that both L^2 projected least-squares methods, labeled as the local L^2 projection method and the pseudolocal L^2 projection method, are coercive:

$$\mathcal{L}_h(\mathbf{v},\mathbf{v}) \geq C\,||\mathbf{v}||_0^2. \tag{1.5}$$

As a result, any C^0 finite elements can be employed. We also prove that the condition number of the resulting linear system is $\mathcal{O}(h^{-2})$. To show the idea, we focus on the error analysis of linear C^0 finite elements in three dimensions. We employ the linear element enriched face bubbles for the local L^2 projection method and the linear element enriched with element bubbles for the pseudolocal L^2 projection method. We can construct an interpolant $\tilde{\mathbf{u}}$ of the exact solution $\mathbf{u}$ that satisfies not only the usual interpolation error estimation (1.3) but also the exclusive interpolation property:

$$||R_h(\mathbf{curl}\,(\mathbf{u}-\tilde{\mathbf{u}}))||_h = ||\breve{R}_h(\operatorname{div}(\varepsilon\,(\mathbf{u}-\tilde{\mathbf{u}})))||_h = 0. \tag{1.6}$$

We thus obtain mainly from (1.5) and (1.6) the following desirable error estimates:

$$||\mathbf{u}-\mathbf{u}_h||_0 \leq C\,||\mathbf{u}-\tilde{\mathbf{u}}||_0, \tag{1.7}$$

where $\mathbf{u}_h$ is the finite element solution associated with (1.4). Equation (1.7) means that the finite element solution is almost the best approximation to the exact solution. In the case that $\mathbf{u} \in H^s$, an L^2 optimal error bound $\mathcal{O}(h^s)$ follows directly from (1.7) and (1.3). We point out that the role of the face bubbles and the element bubbles is to make (1.6) hold. For higher order C^0 finite elements (cubic elements and above), additional bubbles are not always needed, since they usually have face or/and element bubbles of their own.

Before closing this section, we would like to give several remarks. The implementation of the L^2 projected least-squares method of this paper is almost as easy as that of the standard least-squares method (1.2). But, the former allows less regular solution. In comparison with the weighted least-squares method, it does not need any a priori singularity information of the solution and allows the use of both lower-order (maybe enriched with suitable bubbles) and higher-order C^0 elements in both two and three dimensions. Also, it is not clear if there is an improved L^2 error bound for both the standard and the weighted methods in the case where the solution is smooth but the domain is nonsmooth, since to obtain an improved L^2 error bound one has to resort to the well-known Aubin–Nitsche duality argument [8, 13], but this argument usually requires the domain to be smooth enough in order that the associated auxiliary variational problem admits a solution with an appropriate regularity. Unlike the negative norm least-squares method which involves a preconditioner for second-order elliptic problems, the method here deals only with local L^2 projectors, so the practical implementation is simpler. Compared with the FOSLL* method, the method here avoids the differentiation of approximate solutions of potentials and obtains continuous approximate solutions.

The outline of this paper is as follows. In section 2, we review the curl-div magnetostatic problem and recall the L^2 orthogonal decomposition of vector fields. In section 3, two L^2 projected least-squares C^0 finite element methods are described. In section 4, coercivity is established and condition number is estimated. In section 5, the error estimate of the method is obtained. In the last section, some numerical tests are performed to demonstrate the theoretical results obtained.

2. The magnetostatic problem and L^2 decomposition. Let Ω of $\mathbb{R}^3$ be an open, bounded, and possible multiconnected Lipschitz polyhedron, with boundary $\Gamma = \partial\Omega$ and $\mathbf{n}$ the outward unit normal vector to Γ. Let $\varepsilon = (\varepsilon_{ij}) \in \mathbb{R}^{3\times 3}$ satisfy $\varepsilon_{ij} = \varepsilon_{ji}$, $1 \le i, j \le 3$, and

$$C \sum_{i=1}^{3} \xi_i^2 \le \sum_{i,j=1}^{3} \varepsilon_{ij}\xi_i\xi_j \le C^{-1} \sum_{i=1}^{3} \xi_i^2 \quad \text{a.e. in } \bar{\Omega} \quad \forall \xi = (\xi_i) \in \mathbb{R}^3.$$

For the sake of simplicity, we always assume that ε is Lipschitz continuous over $\bar{\Omega}$. With a few modifications, the method of this paper can deal with the case ε being piecewise Lipschitz continuous but not globally continuous (see Remark 5.4).

We now describe the curl-div system of magnetostatic problems.

Given $\mathbf{g} \in (L^2(\Omega))^3$ and $f \in L^2(\Omega)$, the curl-div magnetostatic problem is to find $\mathbf{u}$ such that

$$\mathbf{curl}\,\mathbf{u} = \mathbf{g}, \quad \operatorname{div}(\varepsilon\,\mathbf{u}) = f \quad \text{in } \Omega, \tag{2.1}$$

$$(\varepsilon\,\mathbf{u}) \cdot \mathbf{n} = 0 \quad \text{on } \Gamma, \tag{2.2}$$

where $\mathbf{curl}\,\mathbf{v} = \bigtriangledown \times \mathbf{v}$, $\operatorname{div}\mathbf{v} = \bigtriangledown \cdot \mathbf{v}$, and $\bigtriangledown$ is the gradient operator.

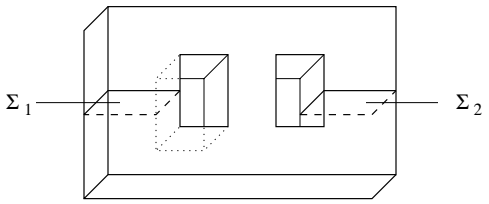


FIG. 1. *A multiply connected domain* Ω.

The solution of problem (2.1)–(2.2) may not be unique when Ω is multiconnected. In that case, to ensure the uniqueness of the solution we need to introduce additional constraints on the so-called *cuts*. (Roughly speaking, these cuts "cut" the multiconnected domain into a simply connected one.) Note that when Ω is simply connected, problem (2.1)–(2.2) has a unique solution. (Of course, $\mathbf{g}$ and f are required to satisfy necessary compatible conditions, e.g., $\operatorname{div}\mathbf{g} = 0$, $\int_\Omega f = 0$, etc.)

To that goal, as done in [2, 28], we assume that there is a set of N cuts Σ_j, $1 \le j \le N$, such that $\overset{\circ}{\Omega} = \Omega \setminus \Sigma$ (where $\Sigma = \bigcup_{j=1}^{N} \Sigma_j$) is pseudoLipschitz and simply connected, where $\bar{\Sigma}_j \subset \Omega$ is a compact and connected two-dimensional Lipschitz manifold with boundary $\partial\Sigma_j \subset \Gamma$, and $\bar{\Sigma}_i \cap \bar{\Sigma}_j = \emptyset$ if $i \ne j$, and Σ_j is globally two-sided, denoted by Σ_j^+ and Σ_j^-, and $\partial\overset{\circ}{\Omega} = \Gamma \cup \Sigma^+ \cup \Sigma^-$. As an illustrating example we consider a multiply connected domain shown in Figure 1. Cutting along Σ_1 and Σ_2 we get a simply connected domain $\overset{\circ}{\Omega}$ with boundary $\Gamma \cup \Sigma_1^+ \cup \Sigma_1^- \cup \Sigma_2^+ \cup \Sigma_2^-$, where $\Sigma_i^\pm$ are the upper and lower (relative to $\mathbf{n}_i$, the unit normals in Σ_i) sheets of Σ_i.

Additional constraints can be thus given by [2, 28]

$$\int_{\Sigma_j} (\varepsilon\,\mathbf{u}) \cdot \mathbf{n} = 0, \quad 1 \le j \le N. \tag{2.3}$$

Problem (2.1)–(2.3) then admits a unique solution, with compatible conditions satisfied by $\mathbf{g}$ and f. Readers may refer to [1, 6, 7, 28] for more details.

We next introduce some Hilbert spaces.

Let $D \subseteq \Omega$. Denote by $H^1(D)$, $H^1(D)/\mathbb{R}$, and $H_0^1(D)$ the usual Hilbert spaces. We also need $H(\operatorname{div};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^3, \operatorname{div}\mathbf{v} \in L^2(\Omega)\}$, $H(\operatorname{div}^0;\Omega) = \{\mathbf{v} \in H(\operatorname{div};\Omega); \operatorname{div}\mathbf{v} = 0\}$, $H_\Gamma(\operatorname{div};\Omega) = \{\mathbf{v} \in H(\operatorname{div};\Omega), \mathbf{v}\cdot\mathbf{n}|_\Gamma = 0\}$, $H_\Gamma(\operatorname{div}^0;\Omega) = H_\Gamma(\operatorname{div};\Omega) \cap H(\operatorname{div}^0;\Omega)$, $H(\mathbf{curl};\Omega) = \{\mathbf{v} \in (L^2(\Omega))^3, \mathbf{curl}\,\mathbf{v} \in (L^2(\Omega))^3\}$, $H_\Gamma(\mathbf{curl};\Omega) = \{\mathbf{v} \in H(\mathbf{curl};\Omega), \mathbf{v}\times\mathbf{n}|_\Gamma = \mathbf{0}\}$, and

$$U = \{\mathbf{v} \in (L^2(\Omega))^3; \operatorname{div}(\varepsilon\mathbf{v}) \in L^2(\Omega), \mathbf{curl}\mathbf{v} \in (L^2(\Omega))^3, (\varepsilon\mathbf{v})\cdot\mathbf{n}|_\Gamma \in L^2(\Gamma)\}, \tag{2.4}$$

$$H_{\text{flux},\Sigma}(\operatorname{div}^0;\Omega) = \left\{\mathbf{v} \in H(\operatorname{div}^0;\Omega); \int_{\Sigma_j} \mathbf{v}\cdot\mathbf{n} = 0, \quad 1 \le j \le N\right\}, \tag{2.5}$$

$$\mathbb{H} = \{\mathbf{v} \in U; \mathbf{curl}\,\mathbf{v} = 0, \operatorname{div}(\varepsilon\,\mathbf{v}) = 0, (\varepsilon\,\mathbf{v})\cdot\mathbf{n}|_\Gamma = 0\}. \tag{2.6}$$

The $\mathbb{H}$, referred to as "harmonic space," may not be trivial in the case of multiconnected domains and accounts for why the solution of problem (2.1)–(2.2) may not be unique. The $\mathbb{H}$ has a finite dimension and can be characterized as the space of gradients of a finite number of scalar functions (See Proposition 2.1 below).

PROPOSITION 2.1 (see [2, 28]). *For any $\Pi \in \mathbb{H}$, there is a $q \in H^1\big(\overset{\circ}{\Omega}\big)$ such that $\bigtriangledown q = \Pi$ in $\overset{\circ}{\Omega}$ and $||q||_{H^1\left(\overset{\circ}{\Omega}\right)} \leq C\,||\Pi||_0$, where q satisfies*

$$\operatorname{div}(\varepsilon \bigtriangledown q) = 0 \quad \text{in } \overset{\circ}{\Omega}, \quad (\varepsilon \bigtriangledown q)\cdot \mathbf{n}\Big|_{\Gamma\cap\partial\,\overset{\circ}{\Omega}} = 0,$$

$$[q]|_{\Sigma_j} = \text{constant}, \quad [(\varepsilon \bigtriangledown q)\cdot \mathbf{n}]\,\Big|_{\Sigma_j} = 0, \quad 1 \leq j \leq N,$$

where $[v]|_{\Sigma_j}$ denotes the jump in v across Σ_j.

Denote by $(\varepsilon\,\cdot,\cdot)$ the ε-weighted L^2 inner product, i.e.,

$$(\varepsilon\,\mathbf{u},\mathbf{v}) := \int_\Omega \varepsilon\,\mathbf{u}\,\mathbf{v}.$$

We now recall the L^2 orthogonal decomposition.

PROPOSITION 2.2 (see [2, 28]). *For any $\mathbf{v} \in (L^2(\Omega))^3$, it can be written as the following L^2 orthogonal decomposition, with respect to $(\varepsilon\,\cdot,\cdot)$:*

$$\mathbf{v} = \bigtriangledown p + \Pi + \varepsilon^{-1}\,\mathbf{curl}\,\psi,$$

where $p \in H^1(\Omega)/\mathbb{R}$, $\Pi \in \mathbb{H}$, $\psi \in H_\Gamma(\mathbf{curl};\Omega) \cap H_{\text{flux},\Sigma}(div^0;\Omega)$, and

$$||\psi||_0 \leq C\,||\mathbf{curl}\,\psi||_0, \quad ||\varepsilon^{\frac{1}{2}}\,\mathbf{v}||_0^2 = ||\varepsilon^{\frac{1}{2}}\bigtriangledown p||_0^2 + ||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2 + ||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0^2.$$

Above and below, the letter C (with or without subscripts) stands for a generic constant which is independent of the mesh-parameter h and may take different values at different occurrences. Denote by $(\cdot,\cdot)_{0,D}$ and $||\cdot||_{0,D}$ the inner product and the norm of $L^2(D)$ or $(L^2(D))^3$, and $(\cdot,\cdot) := (\cdot,\cdot)_{0,\Omega}$, $||\cdot||_0 := ||\cdot||_{0,\Omega}$.

We finally recall Green's formula of integrating by parts:

$$(\operatorname{div}\mathbf{v}, q) = -(\mathbf{v}, \bigtriangledown q) + \int_\Gamma \mathbf{v}\cdot\mathbf{n}\,q \quad \forall \mathbf{v} \in H(\operatorname{div};\Omega), \forall q \in H^1(\Omega);$$

$$(\mathbf{curl}\,\mathbf{u},\mathbf{v}) = (\mathbf{u},\mathbf{curl}\,\mathbf{v}) + \int_\Gamma \mathbf{u}\times\mathbf{n}\cdot\mathbf{v} \quad \forall \mathbf{u} \in H(\mathbf{curl};\Omega), \forall \mathbf{v} \in (H^1(\Omega))^3.$$

3. The L^2 projected least-squares finite element methods. In this section, we shall describe two L^2 projected least-squares finite element methods: (1) the local L^2 projection method; (2) the pseudolocal L^2 projection method.

3.1. Bubbles and some finite dimensional spaces. Let $\mathcal{C}_h$ denote a regular triangulation of $\bar{\Omega}$ into tetrahedra, with diameters h_K for all $K \in \mathcal{C}_h$ bounded by h [8, 13]. We assume that the closure of each cut in $\{\Sigma_j; 1 \leq j \leq N\}$ is the union of the closure of some faces of tetrahedra in $\mathcal{C}_h$.

Let $\mathcal{E}_h^0$ be the set of all the interior faces in $\mathcal{C}_h$, $\mathcal{E}_h^\partial$ the set of all the faces on Γ, and $\mathcal{E}_h = \mathcal{E}_h^0 \bigcup \mathcal{E}_h^\partial$ the set of all faces in $\mathcal{C}_h$. We define $\mathcal{M}_h$ as the collection of macroelements in the following way. Each macroelement in $\mathcal{M}_h$ corresponds to a face $F \in \mathcal{E}_h$ one-by-one: (1) if $F \in \mathcal{E}_h^0$, then the macroelement in $\mathcal{M}_h$ is the union of the two tetrahedra sharing F; (2) if $F \in \mathcal{E}_h^\partial$, then the macroelement is the tetrahedron including F as a face. Note that some macroelements in $\mathcal{M}_h$, corresponding to different faces, are allowed to be identical and that the number of all macroelements is

the same as that of all faces. To emphasize the dependence on F, sometimes we write M as M_F.

We now introduce some bubbles. If K is a tetrahedron with vertices a_i, $1 \le i \le 4$, we denote by λ_i the barycoordinate of a_i, and by F_i the face opposite a_i, and then we introduce the face bubbles

$$b_{F_1} = \lambda_2\lambda_3\lambda_4, \quad b_{F_2} = \lambda_3\lambda_4\lambda_1, \quad b_{F_3} = \lambda_4\lambda_1\lambda_2, \quad b_{F_4} = \lambda_1\lambda_2\lambda_3 \tag{3.1}$$

and the usual element bubble on K,

$$b_K = \lambda_1\lambda_2\lambda_3\lambda_4 \in H_0^1(K). \tag{3.2}$$

We would have

$$b_{F_i} \in H_0^1(F_i), \quad b_{F_i}|_{F_j} = 0 \quad \forall j \neq i. \tag{3.3}$$

For any $M_F \in \mathcal{M}_h$, corresponding to $F \in \mathcal{E}_h$, we introduce the macroelement bubble b_{M_F} as follows.

(1) If $F \in \mathcal{E}_h^0$, i.e., $M_F = K_1 \cup K_2$ with $K_1, K_2 \in \mathcal{C}_h$ sharing F, we denote by $b_F^{K_1}$ and $b_F^{K_2}$ the face bubble of F in K_1 and K_2, respectively. We set

$$b_{M_F}(x), = \begin{cases} b_F^{K_1}(x), & x \in K_1, \\ b_F^{K_2}(x), & x \in K_2, \\ 0 \quad \text{elsewhere.} \end{cases} \tag{3.4}$$

It can be seen that $b_{M_F} \in H_0^1(M_F)$ and

$$b_{M_F}|_F \in H_0^1(F), \quad b_{M_F}|_{F'} = 0 \quad \forall F' (\neq F) \in \mathcal{E}_h. \tag{3.5}$$

(2) If $F \in \mathcal{E}_h^\partial$, i.e., $M_F = K$, with $K \in \mathcal{C}_h$ sharing F with Γ, we set

$$b_{M_F}(x) = \begin{cases} b_F(x), & x \in K, \\ 0 \quad \text{elsewhere.} \end{cases} \tag{3.6}$$

Also, we have (3.5), but $b_{M_F} \notin H_0^1(M_F)$.

Let $\mathcal{P}_r$ be the space of polynomials of order not greater than $r \ge 0$. We define

$$\mathbf{P}(M_F) := \operatorname{span}\{\varepsilon\,(\mathcal{P}_0(M_F))^3, (\mathcal{P}_0(M_F))^3\} = \operatorname{span}\{\mathbf{p}_{F,l}; 1 \le l \le m_F\}, \tag{3.7}$$

$$\mathbf{P}(K) := \operatorname{span}\{\varepsilon\,(\mathcal{P}_0(K))^3, (\mathcal{P}_0(K))^3\} = \operatorname{span}\{\mathbf{p}_l; 1 \le l \le m_K\}, \tag{3.8}$$

where m_F and m_K are positive integers standing for the dimensions of $\mathbf{P}(M_F)$ and $\mathbf{P}(K)$, respectively, and we define the following bubble spaces:

$$\Phi_h := \left\{ \mathbf{v} \in (H^1(\Omega))^3; \mathbf{v} = \sum_{F \in \mathcal{E}_h} \sum_{l=1}^{m_F} c_{F,l}\, \mathbf{p}_{F,l}\, b_{M_F} \ \forall c_{F,l} \in \mathbb{R} \right\}, \tag{3.9}$$

$$\Psi_h := \{\mathbf{v} \in (H_0^1(\Omega))^3; \mathbf{v}|_K \in \mathbf{P}(K)\, b_K \ \forall K \in \mathcal{C}_h\}. \tag{3.10}$$

We also need some additional finite dimensional spaces which will be used in the next section. First, let

$$P_h := \{q \in H^1(\Omega); q|_K \in \mathcal{P}_1(K) \ \forall K \in \mathcal{C}_h\}. \tag{3.11}$$

Second, let q_i^0, $1 \le i \le N$, be the piecewise linear polynomial function, taking value "1" at the nodes on one side of Σ_i, say, Σ_i^+, and "0" at all other nodes (including those on Σ_i^-, the other side of Σ_i). Let $A_h := \mathrm{span}\{q_i^0, i = 1, \ldots, N\}$; we define

$$V_h := P_h + A_h, \tag{3.12}$$

$$W_h := (P_h \cap H_0^1(\Omega))^3, \tag{3.13}$$

where V_h and W_h will be used only for the pseudolocal L^2 projection method below.

3.2. Finite element method. Let U_h be the finite element space. The L^2 projected least-squares finite element method is to find $\mathbf{u}_h \in U_h$ such that

$$\mathcal{L}_h(\mathbf{u}_h, \mathbf{v}) = \mathcal{F}_h(\mathbf{v}) \quad \forall \mathbf{v} \in U_h, \tag{3.14}$$

where

$$\begin{aligned} \mathcal{L}_h(\mathbf{u}, \mathbf{v}) := {} & (\breve{R}_h(\mathrm{div}(\varepsilon\mathbf{u})), \breve{R}_h(\mathrm{div}(\varepsilon\mathbf{v})))_h \\ & + (R_h(\mathbf{curl}\,\mathbf{u}), R_h(\mathbf{curl}\,\mathbf{v}))_h + S_h(\mathbf{u}, \mathbf{v}), \end{aligned} \tag{3.15}$$

$$\mathcal{F}_h(\mathbf{v}) := (f, \breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{v}))) + (\mathbf{g}, R_h(\mathbf{curl}\,\mathbf{v})) + Z_h(f, \mathbf{g}; \mathbf{v}), \tag{3.16}$$

$S_h(\mathbf{u}, \mathbf{v})$ is a mesh-dependent semipositive definite bilinear form on $U_h \times U_h$, $Z_h(f, \mathbf{g}; \mathbf{v})$ is a mesh-dependent linear form on U_h, $(\cdot, \cdot)_h$ is an approximation of L^2 inner product $(\cdot, \cdot)$, and $\breve{R}_h, R_h$ are L^2 projectors. These are to be defined below.

We first describe the *local L^2 projection method*.

We define $(\cdot, \cdot)_h := (\cdot, \cdot)$ and

$$U_h := (P_h)^3 + \Phi_h, \tag{3.17}$$

$$\breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{u}))|_K := \frac{1}{|K|} \int_K \mathrm{div}\,(\varepsilon\,\mathbf{u}) \quad \forall K \in \mathcal{C}_h, \tag{3.18}$$

$$R_h(\mathbf{curl}\,\mathbf{u})|_K := \frac{1}{|K|} \int_K \mathbf{curl}\,\mathbf{u} \quad \forall K \in \mathcal{C}_h, \tag{3.19}$$

$$R_h^\Gamma((\varepsilon\,\mathbf{u}) \cdot \mathbf{n})|_F := \frac{1}{|F|} \int_F (\varepsilon\,\mathbf{u}) \cdot \mathbf{n} \quad \forall F \subset \Gamma, \tag{3.20}$$

where $\mathbf{u}$ is assumed to belong to U defined in (2.4), $|K|$ and $|F|$ respectively denote the volumes of K and F,

$$\begin{aligned} S_h(\mathbf{u}, \mathbf{v}) := {} & \int_\Gamma R_h^\Gamma((\varepsilon\,\mathbf{u}) \cdot \mathbf{n})\, R_h^\Gamma((\varepsilon\,\mathbf{v}) \cdot \mathbf{n}) + S_{h,\mathrm{div}}(\mathbf{u}, \mathbf{v}) \\ & + S_{h,\mathbf{curl}}(\mathbf{u}, \mathbf{v}) + S_{h,\Gamma}(\mathbf{u}, \mathbf{v}) + S_{\mathrm{flux},\Sigma}(\mathbf{u}, \mathbf{v}), \end{aligned} \tag{3.21}$$

$$Z_h(f, \mathbf{g}; \mathbf{v}) := Z_{h,\mathrm{div}}(f; \mathbf{v}) + Z_{h,\mathbf{curl}}(\mathbf{g}; \mathbf{v}) + Z_{h,\Gamma}(f; \mathbf{v}), \tag{3.22}$$

where the definitions of those mesh-dependent bilinear and linear forms of

$$S_{h,\mathrm{div}}(\mathbf{u}, \mathbf{v}), S_{h,\mathbf{curl}}(\mathbf{u}, \mathbf{v}), S_{h,\Gamma}(\mathbf{u}, \mathbf{v}), S_{\mathrm{flux},\Sigma}(\mathbf{u}, \mathbf{v}) \tag{3.23}$$

and

$$Z_{h,\mathrm{div}}(f;\mathbf{v}), Z_{h,\mathbf{curl}}(\mathbf{g};\mathbf{v}), Z_{h,\Gamma}(f;\mathbf{v}) \tag{3.24}$$

will be concretely given in section 3.3.

We next describe the *pseudolocal L^2 projection method.*

We define

$$(p,q)_h := \sum_{K\in\mathcal{C}_h} \frac{|K|}{4} \sum_{i=1}^{4} p(a_i)\, q(a_i) \quad \text{(trapezoidal quadrature)}, \tag{3.25}$$

$$U_h := (P_h)^3 + \Psi_h, \tag{3.26}$$

$$(\breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{u})), q)_h = -\sum_{K\in\mathcal{C}_h} (\varepsilon\,\mathbf{u}, \bigtriangledown\, q)_{0,K} \quad \forall q \in V_h, \tag{3.27}$$

$$(R_h(\mathbf{curl}\,\mathbf{u}), \mathbf{v})_h = (\mathbf{u}, \mathbf{curl}\,\mathbf{v}) \quad \forall \mathbf{v} \in W_h, \tag{3.28}$$

where $\mathbf{u}$ is assumed to belong to $(L^2(\Omega))^3$, V_h and W_h are given by (3.12) and (3.13), and

$$S_h(\mathbf{u},\mathbf{v}) := S_{h,\mathrm{div}}(\mathbf{u},\mathbf{v}) + S_{h,\mathbf{curl}}(\mathbf{u},\mathbf{v}) + S_{h,\Gamma}(\mathbf{u},\mathbf{v}), \tag{3.29}$$

$$Z_h(f,\mathbf{g};\mathbf{v}) := Z_{h,\mathrm{div}}(f;\mathbf{v}) + Z_{h,\mathbf{curl}}(\mathbf{g};\mathbf{v}) + Z_{h,\Gamma}(f;\mathbf{v}). \tag{3.30}$$

Remark 3.1. Noting that the resulting matrix of the trapezoidal quadrature scheme defined by (3.25) is diagonal, for this reason we call R_h and $\breve{R}_h$, defined by (3.27) and (3.28), *pseudolocal L^2* projectors.

3.3. Mesh-dependent bilinear and linear forms. In this subsection we shall define those mesh-dependent bilinear and linear forms as in (3.23) and (3.24).

We need to introduce some local spaces of some suitable functions.

Let $\mathcal{F}_K : \hat{K} \to K$ denote the invertible mapping from the reference element $\hat{K}$ onto $K \in \mathcal{C}_h$, i.e., $K = \mathcal{F}_K(\hat{K})$, which associates the function q defined on K with the function $\hat{q}$ defined on $\hat{K}$ by $q = \hat{q} \circ \mathcal{F}_K^{-1}$. On $\hat{K}$ we introduce three spaces of some suitable functions as follows:

$$\begin{cases} S_{\mathrm{div}}(\hat{K}) := \mathrm{span}\{\hat{v}_l; 1 \le l \le m_{\mathrm{div}}\} \subset L^2(\hat{K}), \\ S_{\Gamma}(\hat{K}) := \mathrm{span}\{\hat{z}_l; 1 \le l \le m_{\Gamma}\} \subset L^2(\hat{K}), \\ S_{\mathbf{curl}}(\hat{K}) := \mathrm{span}\{\hat{\mathbf{w}}_l; 1 \le l \le m_{\mathbf{curl}}\} \subset (L^2(\hat{K}))^3, \end{cases} \tag{3.31}$$

where three integers $m_{\mathrm{div}}, m_{\Gamma}, m_{\mathbf{curl}}$ denoting the dimensions of corresponding spaces and these functions $\hat{v}, \hat{z}, \hat{\mathbf{w}}$ are determined according to Hypothesis H1 in section 4.1. Using $\mathcal{F}_K$ we obtain on $K \in \mathcal{C}_h$ three local spaces as follows:

$$\begin{cases} S_{\mathrm{div}}(K) := S_{\mathrm{div}}(\hat{K}) \circ \mathcal{F}_K^{-1} = \mathrm{span}\{v_{K,l} := \hat{v}_l \circ \mathcal{F}_K^{-1}; 1 \le l \le m_{\mathrm{div}}\}, \\ S_{\Gamma}(K) := S_{\Gamma}(\hat{K}) \circ \mathcal{F}_K^{-1} = \mathrm{span}\{z_{K,l} := \hat{z}_l \circ \mathcal{F}_K^{-1}; 1 \le l \le m_{\Gamma}\}, \\ S_{\mathbf{curl}}(K) := S_{\mathbf{curl}}(\hat{K}) \circ \mathcal{F}_K^{-1} = \mathrm{span}\{\mathbf{w}_{K,l} := \hat{\mathbf{w}}_l \circ \mathcal{F}_K^{-1}; 1 \le l \le m_{\mathbf{curl}}\}. \end{cases} \tag{3.32}$$

Remark 3.2. Consider $\varepsilon = 1$ on K. For the local L^2 projection method we may choose

$$S_{\mathrm{div}}(K) = \mathcal{P}_2(K), \quad S_{\Gamma}(K) = \mathcal{P}_3(K), \quad S_{\mathbf{curl}}(K) = (\mathcal{P}_2(K))^3.$$

For the pseudolocal L^2 projection method we can choose

$$S_{\mathrm{div}}(K) = \mathcal{P}_0(K) + \mathrm{span}\left\{\frac{\partial b_K}{\partial x}, \frac{\partial b_K}{\partial y}, \frac{\partial b_K}{\partial z}\right\}, \quad S_\Gamma(K) = \mathcal{P}_1(K),$$

$$S_{\mathbf{curl}}(K) = (\mathcal{P}_0(K))^3 + \mathrm{span}\left\{\left(c_3\frac{\partial b_K}{\partial y} - c_2\frac{\partial b_K}{\partial z}, c_1\frac{\partial b_K}{\partial z} - c_3\frac{\partial b_K}{\partial x}, c_2\frac{\partial b_K}{\partial x} - c_1\frac{\partial b_K}{\partial y}\right)^t \right.$$
$$\left. \forall (c_1, c_2, c_3) \in \mathbb{R}^3\right\}.$$

Note that, in general, $S_{\mathrm{div}}(K), S_\Gamma(K)$, and $S_{\mathbf{curl}}(K)$ are not spaces of polynomials. However, if ε is piecewise smooth, we may replace ε by its suitable piecewise polynomial approximation, say, ε_h. With this ε_h, then $S_{\mathrm{div}}(K), S_\Gamma(K)$, and $S_{\mathbf{curl}}(K)$ are of course chosen as piecewise polynomials. See also Remarks 5.4 and 5.5.

We are now in a position to define the mesh-dependent bilinear and linear forms:

$$S_{h,\mathrm{div}}(\mathbf{u},\mathbf{v}) := \sum_{K\in\mathcal{C}_h} \frac{\sum_{l=1}^{m_{\mathrm{div}}} (\varepsilon\,\mathbf{u}, \bigtriangledown\,(v_{K,l}\,b_K))_{0,K}\,(\varepsilon\,\mathbf{v}, \bigtriangledown\,(v_{K,l}\,b_K))_{0,K}}{\sum_{l=1}^{m_{\mathrm{div}}} ||\bigtriangledown\,(v_{K,l}\,b_K)||_{0,K}^2}, \tag{3.33}$$

$$Z_{h,\mathrm{div}}(f;\mathbf{v}) := -\sum_{K\in\mathcal{C}_h} \frac{\sum_{l=1}^{m_{\mathrm{div}}} (f, v_{K,l}\,b_K)_{0,K}\,(\varepsilon\,\mathbf{v}, \bigtriangledown\,(v_{K,l}\,b_K))_{0,K}}{\sum_{l=1}^{m_{\mathrm{div}}} ||\bigtriangledown\,(v_{K,l}\,b_K)||_{0,K}^2}, \tag{3.34}$$

$$S_{h,\mathbf{curl}}(\mathbf{u};\mathbf{v}) := \sum_{K\in\mathcal{C}_h} \frac{\sum_{l=1}^{m_{\mathbf{curl}}} (\mathbf{u}, \mathbf{curl}\,(\mathbf{w}_{K,l}\,b_K))_{0,K}\,(\mathbf{v}, \mathbf{curl}\,(\mathbf{w}_{K,l}\,b_K))_{0,K}}{\sum_{l=1}^{m_{\mathbf{curl}}} \|\mathbf{curl}\,(\mathbf{w}_{K,l}\,b_K)\|_{0,K}^2}, \tag{3.35}$$

$$Z_{h,\mathbf{curl}}(\mathbf{g};\mathbf{v}) := \sum_{K\in\mathcal{C}_h} \frac{\sum_{l=1}^{m_{\mathbf{curl}}} (\mathbf{g}, \mathbf{w}_{K,l}\,b_K)_{0,K}\,(\mathbf{v}, \mathbf{curl}\,(\mathbf{w}_{K,l}\,b_K))_{0,K}}{\sum_{l=1}^{m_{\mathbf{curl}}} ||\mathbf{curl}\,(\mathbf{w}_{K,l}\,b_K)||_{0,K}^2}, \tag{3.36}$$

$$S_{h,\Gamma}(\mathbf{u},\mathbf{v}) := \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} (\varepsilon\,\mathbf{u}, \bigtriangledown\,(z_{K,l}\,b_F))_{0,K}\,(\varepsilon\,\mathbf{v}, \bigtriangledown\,(z_{K,l}\,b_F))_{0,K}}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2}, \tag{3.37}$$

$$Z_{h,\Gamma}(f;\mathbf{v}) := -\sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} (f, z_{K,l}\,b_F)_{0,K}\,(\varepsilon\,\mathbf{v}, \bigtriangledown\,(z_{K,l}\,b_F))_{0,K}}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2}, \tag{3.38}$$

$$S_{\mathrm{flux},\Sigma}(\mathbf{u},\mathbf{v}) := \sum_{j=1}^{N} \frac{1}{|\Sigma_j|}\int_{\Sigma_j} (\varepsilon\,\mathbf{u})\cdot\mathbf{n} \int_{\Sigma_j} (\varepsilon\,\mathbf{v})\cdot\mathbf{n}. \tag{3.39}$$

Remark 3.3. Note that $b_K \in H_0^1(K)$ is the element bubble defined as in (3.2) and that $b_F \in H_0^1(F)$ is the face bubble defined as in (3.1); we have $v_{K,l}\,b_K \in H_0^1(K)$, $\mathbf{w}_{K,l}\,b_K \in (H_0^1(K))^3$ and $z_{K,l}\,b_F \in H_0^1(F)$, $z_{K,l}\,b_F|_{F'} = 0$ for $F'(\neq F) \subset \partial\,K$, so all the denominators above are not zero on any tetrahedron.

4. Coercivity and condition number. In this section we shall investigate the coercivity property and the condition number associated with the finite element problem described in the previous section.

4.1. Mesh-dependent norm. In this subsection we give some properties of the mesh-dependent bilinear forms in Propositions 4.1 and 4.2 below.

PROPOSITION 4.1. *We have*

$$|S_{h,\mathrm{div}}(\mathbf{u},\mathbf{v})| \le ||\varepsilon\,\mathbf{u}||_0\,||\varepsilon\,\mathbf{v}||_0,$$

$$|S_{h,\mathbf{curl}}(\mathbf{u},\mathbf{v})| \le ||\mathbf{u}||_0\,||\mathbf{v}||_0,$$

$$|S_{h,\Gamma}(\mathbf{u},\mathbf{v})| \le ||\varepsilon\,\mathbf{u}||_0\,||\varepsilon\,\mathbf{v}||_0,$$

$$0 \le S_{h,\mathrm{div}}(\mathbf{v},\mathbf{v}) \le C \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2,$$

$$0 \le S_{h,\mathbf{curl}}(\mathbf{v},\mathbf{v}) \le C \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathbf{curl}\,\mathbf{v}||_{0,K}^2,$$

$$0 \le S_{h,\Gamma}(\mathbf{v},\mathbf{v}) \le C \left(\sum_{K\in\mathcal{C}_h} h_K^2\,||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2 + \sum_{F\subset\Gamma} h_F \int_F |(\varepsilon\,\mathbf{v})\cdot\mathbf{n}|^2 \right).$$

Here h_F stands for the diameter of F.

Proof. The first three inequalities and the left-hand sides of the last three inequalities easily follow from the definitions given as in (3.33), (3.35), and (3.37). Here we show only the right-hand side of the last inequality as an example. Using Green's formula of integrating by parts, we have

$$\begin{aligned}
S_{h,\Gamma}(\mathbf{v},\mathbf{v}) &= \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left((\varepsilon\,\mathbf{v}, \bigtriangledown\,(z_{K,l}\,b_F))_{0,K}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2} \\
&= \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left(-(\mathrm{div}\,(\varepsilon\,\mathbf{v}), z_{K,l}\,b_F)_{0,K} + ((\varepsilon\,\mathbf{v})\cdot\mathbf{n}, z_{K,l}\,b_F)_{0,F}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2} \\
&\le C \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left(((\varepsilon\,\mathbf{v})\cdot\mathbf{n}, z_{K,l}\,b_F)_{0,F}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2} \\
&\quad + C \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left((\mathrm{div}\,(\varepsilon\,\mathbf{v}), z_{K,l}\,b_F)_{0,K}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2},
\end{aligned}$$

where, by a standard scaling argument [13, 8],

$$\begin{aligned}
\sum_{l=1}^{m_\Gamma} \left((\mathrm{div}\,(\varepsilon\,\mathbf{v}), z_{K,l}\,b_F)_{0,K}\right)^2 &\le ||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2 \sum_{l=1}^{m_\Gamma} ||z_{K,l}\,b_F||_{0,K}^2 \\
&\le C\,h_K^3\,||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2,
\end{aligned}$$

$$\sum_{l=1}^{m_\Gamma} \left(((\varepsilon\,\mathbf{v})\cdot\mathbf{n}, z_{K,l}\,b_F)_{0,F}\right)^2 \leq ||(\varepsilon\,\mathbf{v})\cdot\mathbf{n}||_{0,F}^2 \sum_{l=1}^{m_\Gamma} ||z_{K,l}\,b_F||_{0,F}^2$$

$$\leq C\,h_F^2\,||(\varepsilon\,\mathbf{v})\cdot\mathbf{n}||_{0,F}^2,$$

$$\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2 \geq C\,h_K,$$

where the last three constants C depend on the L^2 or H^1 norms of given functions $\hat{z}_l\,\hat{b}_{\hat{F}}$ on the reference element $\hat{K}$, $1 \leq l \leq m_\Gamma$, with $\hat{b}_{\hat{F}} = b_F \circ (\mathcal{F}_K|_F)$, but they are independent of h and K. It follows that the right-hand side of the last inequality in Proposition 4.1 holds. ∎

We define

$$||\mathbf{v}||_{\mathcal{C}_h}^2 := \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2 + \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathbf{curl}\,\mathbf{v}||_{0,K}^2 + \sum_{F\subset\Gamma} h_F \int_F |(\varepsilon\,\mathbf{v})\cdot\mathbf{n}|^2. \tag{4.1}$$

Hypothesis H1. We assume that, for any $\mathbf{u} \in U_h$, the following *local* inclusions hold:

$$\mathrm{div}\,((\varepsilon\,\mathbf{u})|_K) \in S_{\mathrm{div}}(K) = \mathrm{span}\{v_{K,l}; 1 \leq l \leq m_{\mathrm{div}}\} \quad \forall K \in \mathcal{C}_h, \tag{4.2}$$

$$(\varepsilon\,\mathbf{u})\cdot\mathbf{n}|_F \in S_\Gamma(K)|_F = \mathrm{span}\{z_{F,l} = z_{K,l}|_F; 1 \leq l \leq m_\Gamma\} \quad \forall F \subset \Gamma, \tag{4.3}$$

$$\mathbf{curl}\,(\mathbf{u}|_K) \in S_{\mathbf{curl}}(K) = \mathrm{span}\{\mathbf{w}_{K,l}; 1 \leq l \leq m_{\mathbf{curl}}\} \quad \forall K \in \mathcal{C}_h. \tag{4.4}$$

We additionally assume that $v_{K,l}$, $1 \leq l \leq m_{\mathrm{div}}$, constitutes a group of linearly independent basis and assume the same for $z_{F,l}, 1 \leq l \leq m_\Gamma$, and $\mathbf{w}_{K,l}, 1 \leq l \leq m_{\mathbf{curl}}$.

Remark 4.1. Considering the example in Remark 3.2, we see that Hypothesis H1 holds.

Proposition 4.2. *Assume Hypothesis* H1 holds. *We have on* U_h

$$S_{h,\mathrm{div}}(\mathbf{v},\mathbf{v}) \geq C \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2, \tag{4.5}$$

$$S_{h,\mathbf{curl}}(\mathbf{v},\mathbf{v}) \geq C \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathbf{curl}\,\mathbf{v}||_{0,K}^2, \tag{4.6}$$

$$S_{h,\Gamma}(\mathbf{v},\mathbf{v}) \geq C_1 \sum_{F\subset\Gamma} h_F \int_F |(\varepsilon\,\mathbf{v})\cdot\mathbf{n}|^2 - C_2 \sum_{K\in\mathcal{C}_h} h_K^2\,||\mathrm{div}\,(\varepsilon\,\mathbf{v})||_{0,K}^2. \tag{4.7}$$

Consequently, we have

$$S_{h,\mathrm{div}}(\mathbf{v},\mathbf{v}) + S_{h,\mathbf{curl}}(\mathbf{v},\mathbf{v}) + S_{h,\Gamma}(\mathbf{v},\mathbf{v}) \geq C\,||\mathbf{v}||_{\mathcal{C}_h}^2 \quad \forall \mathbf{v} \in U_h. \tag{4.8}$$

Proof. We first show (4.5). From (4.2) we write $\operatorname{div}(\varepsilon\,\mathbf{v})|_K = \sum_{l=1}^{m_{\rm div}} c_l\, v_{K,l}$. We have

$$\sum_{l=1}^{m_{\rm div}} \left((\varepsilon\,\mathbf{v}, \bigtriangledown(v_{K,l}\, b_K))_{0,K}\right)^2 = \sum_{l=1}^{m_{\rm div}} \left((\operatorname{div}(\varepsilon\,\mathbf{v}), v_{K,l}\, b_K)_{0,K}\right)^2$$

$$= \sum_{l=1}^{m_{\rm div}} (\mathbf{c}^t\, \mathbf{d}_l)^2 = \mathbf{c}^t\, A_K^2\, \mathbf{c},$$

where $\mathbf{c} = (c_1, \ldots, c_{m_{\rm div}})^t \in \mathbb{R}^{m_{\rm div}}$, $\mathbf{d}_l = (d_{1l}, \ldots, d_{m_{\rm div}l})^t \in \mathbb{R}^{m_{\rm div}}$, with $d_{il} = (v_{K,i}, v_{K,l}\, b_K)_{0,K}$, $i = 1, \ldots, m_{\rm div}$, and A_K is the mass matrix and $A_K = [\mathbf{d}_1, \ldots, \mathbf{d}_{m_{\rm div}}] \in \mathbb{R}^{m_{\rm div}\times m_{\rm div}}$. Clearly, A_K is symmetric and positive definite. Let $T \in \mathbb{R}^{m_{\rm div}\times m_{\rm div}}$ be the orthogonal matrix such that $A_K = T^t \operatorname{diag}(\lambda_1, \ldots, \lambda_{m_{\rm div}})\, T$, where $0 < \lambda_1 < \cdots < \lambda_{m_{\rm div}}$ are the eigenvalues of A_K. It can be easily seen that $\lambda_1 \geq C\,|K|$. Let $\bar{\mathbf{c}} := T\,\mathbf{c} = (\bar{c}_1, \ldots, \bar{c}_{m_{\rm div}})^t \in \mathbb{R}^{m_{\rm div}}$; we have $\sum_{l=1}^{m_{\rm div}} \left((\varepsilon\,\mathbf{v}, \bigtriangledown(v_{K,l}\, b_K))_{0,K}\right)^2 = \sum_{l=1}^{m_{\rm div}} (\bar{c}_l\, \lambda_l)^2$. By a similar argument we have $(\operatorname{div}(\varepsilon\,\mathbf{v}), \operatorname{div}(\varepsilon\,\mathbf{v})\, b_K)_{0,K} = \sum_{l=1}^{m_{\rm div}} (\bar{c}_l)^2\, \lambda_l$. We then obtain

$$\sum_{l=1}^{m_{\rm div}} \left((\varepsilon\,\mathbf{v}, \bigtriangledown(v_{K,l}\, b_K))_{0,K}\right)^2 = \sum_{l=1}^{m_{\rm div}} (\bar{c}_l\, \lambda_l)^2 \geq \lambda_1 \sum_{l=1}^{m_{\rm div}} (\bar{c}_l)^2\, \lambda_l$$

$$= \lambda_1\, (\operatorname{div}(\varepsilon\,\mathbf{v}), \operatorname{div}(\varepsilon\,\mathbf{v})\, b_K)_{0,K} \geq C\,|K|\, (\operatorname{div}(\varepsilon\,\mathbf{v}), \operatorname{div}(\varepsilon\,\mathbf{v}))_{0,K}.$$

Here we used the equivalence $C^{-1} \int_K |g| \leq \int_K |g|\, b_K \leq C \int_K |g|$ for any function g in deriving the last inequality. Noting that $\sum_{l=1}^{m_{\rm div}} ||\bigtriangledown\,(v_{K,l}\, b_K)||_{0,K}^2 \leq C\, h_K$ (where C depends on the H^1 norms of $\hat{v}_l\, \hat{b}$, with $\hat{b} = b_K \circ \mathcal{F}_K$ and $1 \leq l \leq m_{\rm div}$, but it is independent of h and K), over $\mathcal{C}_h$ we take the sum of $\frac{\sum_{l=1}^{m_{\rm div}} \left((\varepsilon\,\mathbf{v}, \bigtriangledown(v_{K,l}\, b_K))_{0,K}\right)^2}{\sum_{l=1}^{m_{\rm div}} ||\bigtriangledown\,(v_{K,l}\, b_K)||_{0,K}^2} \geq C\, h_K^2\, ||\operatorname{div}(\varepsilon\,\mathbf{v})||_{0,K}^2$ to get (4.5), with $S_{h,{\rm div}}(\cdot,\cdot)$ defined by (3.33).

The inequality (4.6) can be similarly established from the local inclusion condition (4.4). We next show (4.7). From the definition of $S_{h,\Gamma}(\cdot,\cdot)$ as in (3.37), using Green's formula of integrating by parts, we have

$$S_{h,\Gamma}(\mathbf{v},\mathbf{v}) = \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial\, K}} \frac{\sum_{l=1}^{m_\Gamma} \left((\varepsilon\,\mathbf{v}, \bigtriangledown(z_{K,l}\, b_F))_{0,K}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\, b_F)||_{0,K}^2}$$

$$= \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left(-(\operatorname{div}(\varepsilon\,\mathbf{v}), z_{K,l}\, b_F)_{0,K} + ((\varepsilon\,\mathbf{v})\cdot\mathbf{n}, z_{K,l}\, b_F)_{0,F}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\, b_F)||_{0,K}^2}$$

$$\geq C_3 \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left(((\varepsilon\,\mathbf{v})\cdot\mathbf{n}, z_{K,l}\, b_F)_{0,F}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\, b_F)||_{0,K}^2}$$

$$- C_4 \sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left((\operatorname{div}(\varepsilon\,\mathbf{v}), z_{K,l}\, b_F)_{0,K}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\, b_F)||_{0,K}^2},$$

where, using a similar argument for proving (4.5), we have from the local inclusion condition (4.3)

$$\sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left(((\varepsilon\,\mathbf{v})\cdot\mathbf{n}, z_{F,l}\, b_F)_{0,F}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\, b_F)||_{0,K}^2} \geq C \sum_{F\subset\Gamma} h_F\, ||(\varepsilon\,\mathbf{v})\cdot\mathbf{n}||_{0,F}^2,$$

and from Proposition 4.1 we have

$$\sum_{\substack{F\subset\Gamma \\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} \left((\operatorname{div}(\varepsilon\,\mathbf{v}), z_{K,l}\, b_F)_{0,K}\right)^2}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\, b_F)||_{0,K}^2} \le C \sum_{K\in\mathcal{C}_h} h_K^2\, ||\operatorname{div}(\varepsilon\,\mathbf{v})||_{0,K}^2.$$

The estimate (4.7) thus follows. ∎

Remark 4.2. Propositions 4.1 and 4.2 imply that, instead of using $S_{h,\mathrm{div}}(\mathbf{u},\mathbf{v})$, $S_{h,\mathbf{curl}}(\mathbf{u},\mathbf{v})$, and $S_{h,\Gamma}(\mathbf{u},\mathbf{v})$, one may equivalently use $(\mathbf{u},\mathbf{v})_{\mathcal{C}_h}$ which corresponds to the mesh-dependent norm (4.1). However, when deriving the error estimates of C^0 finite elements, one can bound only $||\mathbf{curl}\,\tilde{\mathbf{u}}||_0$ by $||\mathbf{u}||_1$; here $\tilde{\mathbf{u}}$ is the C^0 finite element interpolant of $\mathbf{u}$, but $\mathbf{u}$ is not generally in H^1 in the case of nonsmooth domains. The use of $S_{h,\mathrm{div}}(\mathbf{u},\mathbf{v})$, $S_{h,\mathbf{curl}}(\mathbf{u},\mathbf{v})$, and $S_{h,\Gamma}(\mathbf{u},\mathbf{v})$ allows the solution to be in H^s with $0\le s<1$. This can easily be seen from the first three inequalities in Proposition 4.1.

LEMMA 4.1. *Under the same hypotheses as in Proposition* 4.2, *for the local* L^2 *projection method we have*

$$S_h(\mathbf{v},\mathbf{v}) \ge C\,\left\{||R_h^\Gamma((\varepsilon\,\mathbf{v})\cdot\mathbf{n})||_{0,\Gamma}^2 + S_{\mathrm{flux},\Sigma}(\mathbf{v},\mathbf{v}) + ||\mathbf{v}||_{\mathcal{C}_h}^2\right\} \quad \forall \mathbf{v}\in U_h.$$

We can have a similar estimate for the pseudolocal L^2 *projection method.*

4.2. Coercivity. This subsection is devoted to the coercivity property of $\mathcal{L}_h$.

Hypothesis H2. We assume that for any $\psi\in H_\Gamma(\mathbf{curl};\Omega)\cap H_{\mathrm{flux},\Sigma}(\mathrm{div}^0;\Omega)$, it can be written as a regular-singular decomposition,

$$\psi=\psi^0+\psi^1, \qquad \psi^0\in H_\Gamma(\mathbf{curl};\Omega)\cap(H^1(\Omega))^3, \quad \mathbf{curl}\,\psi^1=\mathbf{0},$$

where ψ^0 is the regular part and ψ^1 the singular part, with

$$||\psi^0||_1 \le C\,\{||\psi||_0 + ||\mathbf{curl}\,\psi||_0\}.$$

Remark 4.3. From [5] any $\psi\in H_\Gamma(\mathbf{curl};\Omega)\cap H(\mathrm{div};\Omega)$ can be written as the following "regular-singular" decomposition:

$$\psi=\psi^*+\bigtriangledown\, p, \quad p\in H_0^1(\Omega),$$

with $\psi^*\in H_\Gamma(\mathbf{curl};\Omega)\cap(H^1(\Omega))^3$ and

$$||\psi^*||_1 \le C\,\{||\psi||_0 + ||\mathbf{curl}\,\psi||_0 + ||\mathrm{div}\,\psi||_0\}.$$

We may take $\psi^0:=\psi^*$ and $\psi^1:=\bigtriangledown\, p$, and then verify Hypothesis H2.

THEOREM 4.1. *Let Hypotheses* H1 *and* H2 *hold. We have*

$$\mathcal{L}_h(\mathbf{u},\mathbf{u}) \ge C\,||\mathbf{u}||_0^2 \quad \forall \mathbf{u}\in U_h. \tag{4.9}$$

Therefore, the finite element problem has a unique solution.

Proof. We consider only the local L^2 projection method. The following argument can easily be applied to the pseudolocal L^2 projection method, with minor modifications.

We first show

$$\begin{cases} ||\breve{R}_h(\operatorname{div}(\varepsilon\,\mathbf{u}))||_0^2 + ||R_h(\mathbf{curl}\,\mathbf{u})||_0^2 \\ \ge C_5\,||\mathbf{u}||_0^2 - C_6\,(||R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}^2 + S_{\mathrm{flux},\Sigma}(\mathbf{u},\mathbf{u}) + ||\mathbf{u}||_{\mathcal{C}_h}^2), \end{cases} \tag{4.10}$$

and then use Lemma 4.1 to obtain (4.9).

To show (4.10), we need to use the L^2 orthogonal decomposition of $\mathbf{u}$. From Proposition 2.2 we write $\mathbf{u}$ as

$$\mathbf{u} = \bigtriangledown p + \Pi + \varepsilon^{-1}\,\mathbf{curl}\,\psi, \tag{4.11}$$

with $p \in H^1(\Omega)/\mathbb{R}$, $\Pi \in \mathbb{H}$, $\psi \in H_\Gamma(\mathbf{curl};\Omega) \cap H_{\text{flux},\Sigma}(\text{div}^0;\Omega)$, and

$$||\psi||_0 \le C\,||\mathbf{curl}\,\psi||_0, \quad ||\varepsilon^{\frac{1}{2}}\,\mathbf{u}||_0^2 = ||\varepsilon^{\frac{1}{2}}\bigtriangledown p||_0^2 + ||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2 + ||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0^2. \tag{4.12}$$

From Hypothesis H2 we further write ψ as

$$\psi = \psi^0 + \psi^1, \tag{4.13}$$

where

$$\psi^0 \in H_\Gamma(\mathbf{curl};\Omega) \cap (H^1(\Omega))^3, \quad \mathbf{curl}\,\psi^1 = \mathbf{0}, \tag{4.14}$$

$$||\psi^0||_1 \le C\,\{||\psi||_0 + ||\mathbf{curl}\,\psi||_0\} \le C\,||\mathbf{curl}\,\psi||_0. \tag{4.15}$$

In what follows we divide the proof of (4.10) into three steps according to the three components (p, Π, ψ) from the decomposition (4.11).

Step 1. We consider Π.

From Proposition 2.1, let $q \in H^1\big(\overset{\circ}{\Omega}\big)$ such that $\bigtriangledown q = \Pi$ in $\overset{\circ}{\Omega}$ and $||q||_{H^1(\overset{\circ}{\Omega})} \le C\,||\Pi||_0$. Let $\tilde{q}$ and $\bar{q}$ be piecewise constants such that [18, 19, 20]

$$\tilde{q}|_K = \frac{1}{|K|}\int_K q, \quad ||\tilde{q}||_{0,K} \le ||q||_{0,K} \quad \forall K \in \mathcal{C}_h, \tag{4.16}$$

$$\left(\sum_{K\in\mathcal{C}_h} h_K^{-2}\,||\tilde{q}-q||_{0,K}^2\right)^{1/2} \le C\,||q||_{H^1\left(\overset{\circ}{\Omega}\right)}, \tag{4.17}$$

and

$$\bar{q}|_F = \frac{1}{|F|}\int_F q, \quad ||\bar{q}||_{0,F} \le ||q||_{0,F} \quad \forall F \subset \Gamma, \tag{4.18}$$

$$\left(\sum_{F\subset\Gamma} h_F^{-1}\,||\bar{q}-q||_{0,F}^2\right)^{1/2} \le C\,||q||_{H^1\left(\overset{\circ}{\Omega}\right)}. \tag{4.19}$$

Let $\epsilon_1 > 0$ be a constant to be determined. We have

$$||\breve{R}_h(\mathrm{div}(\varepsilon\mathbf{u}))||_0^2 = ||\breve{R}_h(\mathrm{div}(\varepsilon\mathbf{u})) + \epsilon_1\tilde{q}||_0^2 - \epsilon_1^2||\tilde{q}||_0^2 - 2\epsilon_1\,(\breve{R}_h(\mathrm{div}(\varepsilon\,\mathbf{u})),\tilde{q}), \tag{4.20}$$

$$||\tilde{q}||_0 \le C\,||q||_{L^2\left(\overset{\circ}{\Omega}\right)} \le C\,||q||_{H^1\left(\overset{\circ}{\Omega}\right)} \le C\,||\Pi||_0 \le C\,||\varepsilon^{\frac{1}{2}}\,\Pi||_0, \tag{4.21}$$

$$\begin{aligned} -2\,\epsilon_1\,(\breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{u})),\tilde{q}) &= -2\,\epsilon_1\,\textstyle\sum_{K\in\mathcal{C}_h}(\mathrm{div}\,(\varepsilon\,\mathbf{u}),\tilde{q})_{0,K} \\ &= 2\,\epsilon_1\,\textstyle\sum_{K\in\mathcal{C}_h}(\mathrm{div}\,(\varepsilon\,\mathbf{u}),q-\tilde{q})_{0,K} \\ &\quad -2\,\epsilon_1\,\textstyle\sum_{K\in\mathcal{C}_h}(\mathrm{div}\,(\varepsilon\,\mathbf{u}),q)_{0,K}, \end{aligned} \tag{4.22}$$

$$\begin{aligned} -2\,\epsilon_1\,\textstyle\sum_{K\in\mathcal{C}_h}(\mathrm{div}\,(\varepsilon\,\mathbf{u}),q)_{0,K} &= 2\,\epsilon_1\,\textstyle\sum_{K\in\mathcal{C}_h}(\varepsilon\,\mathbf{u},\bigtriangledown q)_{0,K} \\ &\quad -2\,\epsilon_1\,\textstyle\sum_{F\subset\Gamma}\int_F(\varepsilon\,\mathbf{u})\cdot\mathbf{n}\,q \\ &\quad -2\,\epsilon_1\,\textstyle\sum_{j=1}^{N}\int_{\Sigma_j}(\varepsilon\,\mathbf{u})\cdot\mathbf{n}\,[q], \end{aligned} \tag{4.23}$$

$$2\,\epsilon_1\sum_{K\in\mathcal{C}_h}(\varepsilon\,\mathbf{u},\bigtriangledown q)_{0,K} = 2\,\epsilon_1\,(\varepsilon\,\mathbf{u},\Pi) = 2\,\epsilon_1\,(\varepsilon\,\Pi,\Pi) = 2\,\epsilon_1\,||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2, \tag{4.24}$$

$$-2\,\epsilon_1\sum_{F\subset\Gamma}\int_F(\varepsilon\,\mathbf{u})\cdot\mathbf{n}q = 2\epsilon_1\sum_{F\subset\Gamma}\int_F(\varepsilon\,\mathbf{u})\cdot\mathbf{n}(\bar{q}-q) - 2\epsilon_1\sum_{F\subset\Gamma}\int_F(\varepsilon\,\mathbf{u})\cdot\mathbf{n}\,\bar{q}, \tag{4.25}$$

and

$$\begin{aligned} -2\,\epsilon_1\sum_{F\subset\Gamma}\int_F(\varepsilon\,\mathbf{u})\cdot\mathbf{n}\,\bar{q} &= -2\,\epsilon_1\int_\Gamma R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})\,\bar{q} \\ &\ge -\epsilon_1\,C\,||R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}\,||\bar{q}||_{0,\Gamma} \\ &\ge -\epsilon_1\,C\,||R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}\,||q||_{0,\Gamma} \\ &\ge -\epsilon_1\,C\,||R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}\,||q||_{H^1\left(\overset{\circ}{\Omega}\right)} \\ &\ge -\epsilon_2\,||R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}^2 - \frac{C\,\epsilon_1^2}{\epsilon_2}\,||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2. \end{aligned} \tag{4.26}$$

Here we have used $||q||_{H^1(\overset{\circ}{\Omega})} \le C\,||\Pi||_0 \le C\,||\varepsilon^{\frac{1}{2}}\,\Pi||_0$ and the trace theorem,

$$||q||_{0,\Gamma} \le C\,||q||_{L^2\left(\partial\,\overset{\circ}{\Omega}\right)} \le C\,||q||_{H^1\left(\overset{\circ}{\Omega}\right)},$$

and Young's inequality,

$$|a|\,|b| \le \epsilon\,|a|^2 + \frac{1}{4\,\epsilon}\,|b|^2 \quad \forall a,b\in\mathbb{R},\forall\epsilon>0,$$

with $\epsilon = \epsilon_2$ a constant to be determined.

We also have

$$
\begin{aligned}
(4.27)\qquad & 2\,\epsilon_1 \sum_{K\in\mathcal{C}_h} (\operatorname{div}(\varepsilon\,\mathbf{u}), q-\tilde{q})_{0,K} + 2\,\epsilon_1 \sum_{F\subset\Gamma} \int_F (\varepsilon\,\mathbf{u})\cdot\mathbf{n}\,(\bar{q}-q) \\
& \geq -\epsilon_1\, C \left(\sum_{K\in\mathcal{C}_h} h_K^2\, ||\operatorname{div}(\varepsilon\,\mathbf{u})||_{0,K}^2 + \sum_{F\subset\Gamma} h_F\, ||(\varepsilon\,\mathbf{u})\cdot\mathbf{n}||_{0,F}^2 \right)^{\frac{1}{2}} \\
& \quad \times \left(\sum_{K\in\mathcal{C}_h} h_K^{-2}\, ||q-\tilde{q}||_{0,K}^2 + \sum_{F\subset\Gamma} h_F^{-1}\, ||\bar{q}-q||_{0,F}^2 \right)^{\frac{1}{2}} \\
& \geq -\epsilon_1\, C\, ||\mathbf{u}||_{\mathcal{C}_h}\, ||q||_{H^1\left(\overset{\circ}{\Omega}\right)} \geq -\epsilon_1\, C\, ||\mathbf{u}||_{\mathcal{C}_h}\, ||\varepsilon^{\frac{1}{2}}\,\Pi||_0 \\
& \geq -\epsilon_2\, ||\mathbf{u}||_{\mathcal{C}_h}^2 - \frac{C\,\epsilon_1^2}{\epsilon_2}\, ||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2
\end{aligned}
$$

and

$$
\begin{aligned}
(4.28)\quad -2\,\epsilon_1 \sum_{j=1}^{N} \int_{\Sigma_j} (\varepsilon\,\mathbf{u})\cdot\mathbf{n}\,[q] & \geq -2\,\epsilon_1 \sum_{j=1}^{N} |\Sigma_j|^{-1/2} \left| \int_{\Sigma_j} (\varepsilon\,\mathbf{u})\cdot\mathbf{n} \right| |\Sigma_j|^{1/2}\, |[q]| \\
& \geq -2\,\epsilon_1 \left(\sum_{j=1}^{N} \frac{1}{|\Sigma_j|} \left(\int_{\Sigma_j} (\varepsilon\,\mathbf{u})\cdot\mathbf{n} \right)^2 \right)^{1/2} \\
& \quad \times \left(\sum_{j=1}^{N} \int_{\Sigma_j} |[q]|^2 \right)^{1/2} \\
& = -2\,\epsilon_1\, \left(S_{\text{flux},\Sigma}(\mathbf{u},\mathbf{u})\right)^{1/2} \left(\sum_{j=1}^{N} \int_{\Sigma_j} |[q]|^2 \right)^{1/2} \\
& \geq -2\,\epsilon_1\, C\, \left(S_{\text{flux},\Sigma}(\mathbf{u},\mathbf{u})\right)^{1/2}\, ||q||_{L^2\left(\partial\,\overset{\circ}{\Omega}\right)} \\
& \geq -2\,\epsilon_1\, C\, \left(S_{\text{flux},\Sigma}(\mathbf{u},\mathbf{u})\right)^{1/2}\, ||q||_{H^1\left(\overset{\circ}{\Omega}\right)} \\
& \geq -\epsilon_2\, S_{\text{flux},\Sigma}(\mathbf{u},\mathbf{u}) - \frac{C\,\epsilon_1^2}{\epsilon_2}\, ||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2.
\end{aligned}
$$

Equations (4.20)–(4.28) yield

$$
\begin{aligned}
(4.29)\quad ||\breve{R}_h(\operatorname{div}(\varepsilon\,\mathbf{u}))||_0^2 & \geq \epsilon_1 \left(2 - C\,\epsilon_1 - \frac{C\,\epsilon_1}{\epsilon_2} \right) ||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2 \\
& \quad - \epsilon_2 \left(||R_h^{\Gamma}((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}^2 + S_{\text{flux},\Sigma}(\mathbf{u},\mathbf{u}) + ||\mathbf{u}||_{\mathcal{C}_h}^2 \right).
\end{aligned}
$$

Step 2. We consider p.

By a similar argument as that used in proving (4.29) with Π replaced by $\bigtriangledown p$, we have

$$\begin{cases} ||\breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{u}))||_0^2 & \geq \epsilon_1 \left(2 - C\,\epsilon_1 - \frac{C\,\epsilon_1}{\epsilon_2}\right) ||\varepsilon^{\frac{1}{2}} \bigtriangledown p||_0^2 \\ & \quad -\epsilon_2 \left(||R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}^2 + ||\mathbf{u}||_{\mathcal{C}_h}^2\right). \end{cases} \tag{4.30}$$

Step 3. We consider ψ.

Let $\tilde{\psi^0}$ be a piecewise constant vector such that

$$\tilde{\psi^0}|_K = \frac{1}{|K|}\int_K \psi^0, \quad ||\tilde{\psi^0}||_{0,K} \leq ||\psi^0||_{0,K} \quad \forall K \in \mathcal{C}_h, \tag{4.31}$$

$$\left(\sum_{K\in\mathcal{C}_h} h_K^{-2}\,||\psi^0 - \tilde{\psi^0}||_{0,K}^2\right)^{1/2} \leq C\,||\psi^0||_1. \tag{4.32}$$

We have

$$||R_h(\mathbf{curlu})||_0^2 = ||R_h(\mathbf{curlu}) - \epsilon_1\tilde{\psi^0}||_0^2 - \epsilon_1^2\,||\tilde{\psi^0}||_0^2 + 2\,\epsilon_1\,(R_h(\mathbf{curl}\,\mathbf{u}), \tilde{\psi^0}), \tag{4.33}$$

$$||\tilde{\psi^0}||_0 \leq ||\psi^0||_1 \leq C\,||\mathbf{curl}\,\psi||_0 \leq C\,||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0, \tag{4.34}$$

$$\begin{aligned} 2\,\epsilon_1\,(R_h(\mathbf{curl}\,\mathbf{u}), \tilde{\psi^0}) &= 2\,\epsilon_1 \textstyle\sum_{K\in\mathcal{C}_h} (\mathbf{curl}\,\mathbf{u}, \tilde{\psi^0})_{0,K} \\ &= 2\,\epsilon_1 \textstyle\sum_{K\in\mathcal{C}_h} (\mathbf{curl}\,\mathbf{u}, \tilde{\psi^0} - \psi^0)_{0,K} \\ &\quad + 2\,\epsilon_1 \textstyle\sum_{K\in\mathcal{C}_h} (\mathbf{curl}\,\mathbf{u}, \psi^0)_{0,K}, \end{aligned} \tag{4.35}$$

$$\begin{aligned} 2\,\epsilon_1 \textstyle\sum_{K\in\mathcal{C}_h} (\mathbf{curl}\,\mathbf{u}, \psi^0)_{0,K} &= 2\,\epsilon_1\,(\mathbf{u}, \mathbf{curl}\,\psi^0) = 2\,\epsilon_1\,(\mathbf{u}, \mathbf{curl}\,\psi) \\ &= 2\,\epsilon_1\,(\varepsilon^{-1}\,\mathbf{curl}\,\psi, \mathbf{curl}\,\psi) \\ &= 2\,\epsilon_1\,||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0^2, \end{aligned} \tag{4.36}$$

and using Young's inequality,

$$\begin{aligned} 2\,\epsilon_1 \sum_{K\in\mathcal{C}_h} (\tilde{\psi^0} - \psi^0, \mathbf{curl}\,\mathbf{u}) &\geq -\epsilon_1\,C \left(\sum_{K\in\mathcal{C}_h} h_K^{-2}\,||\tilde{\psi^0} - \psi^0||_{0,K}^2\right)^{\frac{1}{2}} \\ &\quad \times \left(\sum_{K\in\mathcal{C}_h} h_K^2\,||\mathbf{curl}\,\mathbf{u}||_{0,K}^2\right)^{\frac{1}{2}} \\ &\geq -\epsilon_1\,C\,||\psi^0||_1\,||\mathbf{u}||_{\mathcal{C}_h} \\ &\geq -\epsilon_1\,C\,||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0\,||\mathbf{u}||_{\mathcal{C}_h} \\ &\geq -\epsilon_2\,||\mathbf{u}||_{\mathcal{C}_h}^2 - \frac{C\,\epsilon_1^2}{\epsilon_2}\,||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0^2. \end{aligned} \tag{4.37}$$

Summarizing (4.33)–(4.37), we have

$$||R_h(\mathbf{curl}\,\mathbf{u})||_0^2 \geq \epsilon_1 \left(2 - C\,\epsilon_1 - \frac{C\,\epsilon_1}{\epsilon_2}\right) ||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0^2 - \epsilon_2\,||\mathbf{u}||_{\mathcal{C}_h}^2. \tag{4.38}$$

Combining (4.29), (4.30), and (4.38), we obtain

$$2\,||\breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{u}))||_0^2 + ||R_h(\mathbf{curl}\,\mathbf{u})||_0^2 \tag{4.39}$$

$$\geq \epsilon_1 \left(2 - C\,\epsilon_1 - \frac{C\,\epsilon_1}{\epsilon_2}\right) \times \left(||\varepsilon^{\frac{1}{2}} \bigtriangledown p||_0^2 + ||\varepsilon^{\frac{1}{2}}\,\Pi||_0^2 + ||\varepsilon^{-\frac{1}{2}}\,\mathbf{curl}\,\psi||_0^2\right)$$

$$-\epsilon_2\,C\,(||R_h^{\Gamma}((\varepsilon\,\mathbf{u})\cdot\mathbf{n})||_{0,\Gamma}^2 + S_{\mathrm{flux},\Sigma}(\mathbf{u},\mathbf{u}) + ||\mathbf{u}||_{\mathcal{C}_h}^2).$$

We therefore obtain (4.10), taking suitable values for ϵ_i, $i = 1, 2$. ∎

4.3. Condition number. We finally estimate the condition number of the resulting linear system. We here again consider only the local L^2 projection method.

THEOREM 4.2. *In addition to the same hypotheses as in Theorem* 4.1, *we assume uniform meshes. The condition number, associated with the resulting linear system, is* $\mathcal{O}(h^{-2})$.

Proof. Since, for all $\mathbf{v} \in U_h$,

$$(S_{h,\mathrm{div}}(\mathbf{v},\mathbf{v}))^{1/2} + (S_{h,\mathbf{curl}}(\mathbf{v},\mathbf{v}))^{1/2} + (S_{h,\Gamma}(\mathbf{v},\mathbf{v}))^{1/2} \leq C\,||\mathbf{v}||_0, \tag{4.40}$$

$$||R_h(\mathbf{curl}\,\mathbf{v})||_0 \leq ||\mathbf{curl}\,\mathbf{v}||_0 \leq C\,h^{-1}\,||\mathbf{v}||_0, \tag{4.41}$$

$$||\breve{R}_h(\mathrm{div}\,(\varepsilon\,\mathbf{v}))||_0 \leq C\,||\varepsilon||_\infty \left(\sum_{K\in\mathcal{C}_h} h_K^{-1}\,||\mathbf{v}||_{0,\partial\,K}^2\right)^{\frac{1}{2}} \leq C\,h^{-1}\,||\mathbf{v}||_0, \tag{4.42}$$

$$||R_h^{\Gamma}((\varepsilon\,\mathbf{v})\cdot\mathbf{n})||_{0,\Gamma} \leq \left(\sum_{F\subset\Gamma} ||(\varepsilon\,\mathbf{v})\cdot\mathbf{n}||_{0,F}^2\right)^{1/2} \leq C\,h^{-1}\,||\mathbf{v}||_0, \tag{4.43}$$

$$(S_{\mathrm{flux},\Sigma}(\mathbf{v},\mathbf{v}))^{1/2} \leq C\,h^{-1}\,||\mathbf{v}||_0, \tag{4.44}$$

we have

$$\mathcal{L}_h(\mathbf{v},\mathbf{v}) \leq C\,h^{-2}\,||\mathbf{v}||_0^2 \quad \forall \mathbf{v} \in U_h, \tag{4.45}$$

which, together with the coercivity property (4.9) and the symmetry property of $\mathcal{L}_h$, leads to the conclusion. ∎

5. Error bounds. In this section we shall establish the error bounds. We analyze only the local L^2 projection method. The analysis is similar for the pseudolocal L^2 projection method.

LEMMA 5.1. *Let* $\mathbf{u} \in U$ *and* $\mathbf{u}_h \in U_h$ *be the exact solution and the finite element solution to the local* L^2 *projection method. We have*

$$\mathcal{L}_h(\mathbf{u} - \mathbf{u}_h, \mathbf{v}_h) = 0 \quad \forall \mathbf{v}_h \in U_h. \tag{5.1}$$

Proof. Equation (5.1) holds, since

$$S_h(\mathbf{u},\mathbf{v}_h) = Z_h(f,\mathbf{g};\mathbf{v}_h), \tag{5.2}$$

$$(\breve{R}_h(\mathrm{div}(\varepsilon\mathbf{u})), \breve{R}_h(\mathrm{div}(\varepsilon\mathbf{v}_h))) = (\mathrm{div}(\varepsilon\,\mathbf{u}), \breve{R}_h(\mathrm{div}(\varepsilon\,\mathbf{v}_h))) = (f, \breve{R}_h(\mathrm{div}(\varepsilon\,\mathbf{v}_h))), \tag{5.3}$$

$$(R_h(\mathbf{curl}\,\mathbf{u}), R_h(\mathbf{curl}\,\mathbf{v}_h)) = (\mathbf{curl}\,\mathbf{u}, R_h(\mathbf{curl}\,\mathbf{v}_h)) = (\mathbf{g}, R_h(\mathbf{curl}\,\mathbf{v}_h)). \tag{5.4}$$

∎

LEMMA 5.2. *Assume that* $\mathbf{u} \in (H^s(\Omega))^3$ *with* $s > \frac{1}{2}$. *Then, there exists a* $\tilde{\mathbf{u}} \in U_h$ *such that*

$$||\breve{R}_h(\operatorname{div}(\varepsilon(\mathbf{u}-\tilde{\mathbf{u}})))||_0^2 = ||R_h(\mathbf{curl}(\mathbf{u}-\tilde{\mathbf{u}}))||_0^2 = 0, \tag{5.5}$$

$$||R_h^{\Gamma}((\varepsilon(\mathbf{u}-\tilde{\mathbf{u}}))\cdot\mathbf{n})||_{0,\Gamma}^2 = S_{\text{flux},\Sigma}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}}) = 0, \tag{5.6}$$

$$||\mathbf{u}-\tilde{\mathbf{u}}||_0 \le C\,h^s\,||\mathbf{u}||_s. \tag{5.7}$$

Proof. We first let $\mathbf{u}^0 \in (P_h)^3$ be such that [3, 14, 29]

$$||\mathbf{u}-\mathbf{u}^0||_0 + \left(\sum_{F\subset\mathcal{E}_h} h_F\,||\mathbf{u}-\mathbf{u}^0||_{0,F}^2\right)^{\frac{1}{2}} \le C\,h^s\,||\mathbf{u}||_s, \quad s>\frac{1}{2}. \tag{5.8}$$

We then define $\tilde{\mathbf{u}} \in U_h$ such that

$$\tilde{\mathbf{u}}(a) = \mathbf{u}^0(a) \qquad \text{for all vertices } a, \tag{5.9}$$

$$\int_F (\tilde{\mathbf{u}}-\mathbf{u})\cdot\mathbf{p} = \mathbf{0} \quad \forall \mathbf{p}\in\mathbf{P}(M_F), \forall F\in\mathcal{E}_h, \tag{5.10}$$

where $\mathbf{P}(M_F)$ is defined by (3.7). We write

$$\tilde{\mathbf{u}} = \mathbf{u}^0 + \sum_{F\in\mathcal{E}_h}\sum_{l=1}^{m_F} c_{F,l}\,\mathbf{p}_{F,l}\,b_{M_F}. \tag{5.11}$$

Noting that for any given $F\in\mathcal{E}_h$,

$$\sum_{F'\in\mathcal{E}_h, F'\neq F}\int_F \sum_{l=1}^{m_{F'}} c_{F',l}\,\mathbf{p}_{F',l}\,b_{M_{F'}} = 0, \tag{5.12}$$

from (5.10) we obtain

$$\sum_{l=1}^{m_F} c_{F,l}\int_F \mathbf{p}_{F,l}\cdot\mathbf{p}_{F,i}\,b_{M_F} = \int_F (\mathbf{u}-\mathbf{u}^0)\cdot\mathbf{p}_{F,i}, \quad 1\le i\le m_F\ \forall F\in\mathcal{E}_h, \tag{5.13}$$

which uniquely determines the coefficients $c_{F,l}$, $1\le l\le m_F$, $F\in\mathcal{E}_h$. Also, note that $\sum_{F\in\mathcal{E}_h}\sum_{l=1}^{m_F} c_{F,l}\,\mathbf{p}_{F,l}\,b_{M_F}$ is zero at all vertices, and (5.9) and (5.10) uniquely determine $\tilde{\mathbf{u}}$.

Let M be any given macroelement in $\mathcal{M}_h$ corresponding to a face $F\in\mathcal{E}_h$. We write M and F as M_{F_1} and F_1. We first consider the case $M_{F_1} = K_1\cup K_2$, with $K_1\cap K_2 = F_1\in\mathcal{E}_h^0$. For convenience we number all the faces in ∂M_{F_1} by F_i, one-to-one corresponding to M_{F_i}, $2\le i\le 7$. Denote $\mathbf{c}_i := (c_{F_i1},\dots,c_{F_im_{F_i}})^t \in \mathbb{R}^{m_{F_i}}$, and $\mathbf{L}_i := [\mathbf{p}_{F_i1},\dots,\mathbf{p}_{F_im_{F_i}}] \in \mathbb{R}^{m_{F_i}\times m_{F_i}}$; by a standard scaling argument we then have

from (5.12) and (5.13)

$$(5.14)\quad \int_{M_{F_1}} \left(\sum_{F\in\mathcal{E}_h} \sum_{l=1}^{m_F} c_{F,l}\, \mathbf{p}_{F,l}\, b_{M_F} \right)^2 = \int_{M_{F_1}} \left(\sum_{i=1}^{7} \sum_{l=1}^{m_{F_i}} c_{F_i,l}\, \mathbf{p}_{F_i,l}\, b_{M_{F_i}} \right)^2$$

$$= \int_{M_{F_1}} \left| \sum_{i=1}^{7} \mathbf{L}_i\, \mathbf{c}_i\, b_{M_{F_i}} \right|^2$$

$$\leq C \sum_{i=1}^{7} |\mathbf{c}_i|^2 \int_{M_{F_1}} \left(\sum_{i=1}^{7} \sum_{l=1}^{m_{F_i}} |\mathbf{p}_{F_i,l}\, b_{M_{F_i}}|^2 \right)$$

$$\leq C\, |M_{F_1}| \sum_{i=1}^{7} |\mathbf{c}_i|^2 \leq C \sum_{i=1}^{7} h_{F_i}\, ||\mathbf{u}-\mathbf{u}^0||^2_{0,F_i}.$$

We thus obtain

$$(5.15)\qquad ||\mathbf{u}-\tilde{\mathbf{u}}||_{0,M} \leq C\, ||\mathbf{u}-\mathbf{u}^0||_{0,M} + C \sum_{i=1}^{7} h_{F_i}^{\frac{1}{2}}\, ||\mathbf{u}-\mathbf{u}^0||_{0,F_i}.$$

Similarly, if M is the tetrahedron sharing an F with Γ, we have

$$(5.16)\qquad ||\mathbf{u}-\tilde{\mathbf{u}}||_{0,M} \leq C\, ||\mathbf{u}-\mathbf{u}^0||_{0,M} + C \sum_{F\subset\partial M} h_F^{\frac{1}{2}}\, ||\mathbf{u}-\mathbf{u}^0||_{0,F}.$$

Hence, from (5.15), (5.16), and (5.8), we obtain

$$(5.17)\qquad ||\mathbf{u}-\tilde{\mathbf{u}}||_0 \leq C\, ||\mathbf{u}-\mathbf{u}^0||_0 + C \left(\sum_{F\subset\mathcal{E}_h} h_F\, ||\mathbf{u}-\mathbf{u}^0||^2_{0,F} \right)^{\frac{1}{2}} \leq C\, h^s\, ||\mathbf{u}||_s.$$

Finally, noting that $\mathbf{n}$ is a constant vector and

$$(5.18)\qquad \mathcal{P}_0(M)\, \varepsilon\, \mathbf{n} \subset \mathbf{P}(M), \quad (\mathcal{P}_0(M))^3 \times \mathbf{n} \subset \mathbf{P}(M),$$

by virtue of (5.10) we can easily verify (5.5) and (5.6). □

THEOREM 5.1. *Assume that Hypotheses* H1 *and* H2 *hold and that* $\mathbf{u} \in (H^s(\Omega))^3$ *with* $s > \frac{1}{2}$. *Let* $\mathbf{u} \in U$ *and* $\mathbf{u}_h \in U_h$ *be the exact solution and the finite element solution to the local* L^2 *projection method. We have*

$$(5.19)\qquad ||\mathbf{u}-\mathbf{u}_h||_0 \leq C\, h^s\, ||\mathbf{u}||_s.$$

Proof. Let $\tilde{\mathbf{u}} \in U_h$ be constructed as in Lemma 5.2. We have from Lemma 5.1

$$(5.20)\qquad |||\mathbf{u}_h-\tilde{\mathbf{u}}|||^2 := \mathcal{L}_h(\mathbf{u}_h-\tilde{\mathbf{u}}, \mathbf{u}_h-\tilde{\mathbf{u}}) = \mathcal{L}_h(\mathbf{u}-\tilde{\mathbf{u}}, \mathbf{u}_h-\tilde{\mathbf{u}})$$

$$\leq |||\mathbf{u}-\tilde{\mathbf{u}}|||\, |||\mathbf{u}_h-\tilde{\mathbf{u}}|||;$$

that is,

$$(5.21)\qquad |||\mathbf{u}_h-\tilde{\mathbf{u}}||| \leq |||\mathbf{u}-\tilde{\mathbf{u}}|||.$$

On the other hand, from Lemma 5.2 and Proposition 4.1,

$$
\begin{aligned}
(5.22)\qquad |||\mathbf{u}-\tilde{\mathbf{u}}|||^2 &= \mathcal{L}_h(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}})\\
&= ||\breve{R}_h(\operatorname{div}(\varepsilon(\mathbf{u}-\tilde{\mathbf{u}})))||_0^2 + ||R_h(\mathbf{curl}(\mathbf{u}-\tilde{\mathbf{u}}))||_0^2\\
&\quad + ||R_h^{\Gamma}((\varepsilon(\mathbf{u}-\tilde{\mathbf{u}}))\cdot\mathbf{n})||_{0,\Gamma}^2 + S_{h,\operatorname{div}}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}})\\
&\quad + S_{h,\mathbf{curl}}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}})\\
&\quad + S_{h,\Gamma}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}}) + S_{\text{flux},\Sigma}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}})\\
&= S_{h,\operatorname{div}}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}}) + S_{h,\mathbf{curl}}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}})\\
&\quad + S_{h,\Gamma}(\mathbf{u}-\tilde{\mathbf{u}},\mathbf{u}-\tilde{\mathbf{u}})\\
&\quad \le C\,||\mathbf{u}-\tilde{\mathbf{u}}||_0^2.
\end{aligned}
$$

Using the triangle inequality, Theorem 4.1, (5.21), (5.22), and Lemma 5.2, we obtain

$$
\begin{aligned}
(5.23)\qquad ||\mathbf{u}-\mathbf{u}_h||_0 &\le ||\mathbf{u}-\tilde{\mathbf{u}}||_0 + ||\mathbf{u}_h-\tilde{\mathbf{u}}||_0\\
&\le ||\mathbf{u}-\tilde{\mathbf{u}}||_0 + C\,|||\mathbf{u}_h-\tilde{\mathbf{u}}|||\\
&\le ||\mathbf{u}-\tilde{\mathbf{u}}||_0 + C\,|||\mathbf{u}-\tilde{\mathbf{u}}|||\\
&\le C\,||\mathbf{u}-\tilde{\mathbf{u}}||_0 \le C\,h^s\,||\mathbf{u}||_s. \qquad \square
\end{aligned}
$$

Remark 5.1. From [2] we know that, in the case of $\varepsilon = 1$, the solution of problem (2.1)–(2.3) is in $(H^s(\Omega))^3$ with $s > \frac{1}{2}$.

Remark 5.2. Regarding the pseudolocal L^2 projected method, following a similar argument in Lemma 5.2, we can find an interpolant $\tilde{\mathbf{u}} \in U_h$ of $\mathbf{u} \in (H^s(\Omega))^3$ with $s \ge 0$ such that $||\mathbf{u}-\tilde{\mathbf{u}}||_0 \le C\,h^s\,||\mathbf{u}||_s$, and $\tilde{\mathbf{u}}$ satisfies interpolation properties similar to (5.5) and (5.6). Let $\mathbf{u}$ and $\mathbf{u}_h$ be the exact solution and the finite element solution to the pseudolocal L^2 projection method. Following the same argument in Theorem 5.1, we can obtain

$$
(5.24)\qquad ||\mathbf{u}-\mathbf{u}_h||_0 \le C\,h^s\,||\mathbf{u}||_s \quad \forall s \ge 0.
$$

Remark 5.3. The pseudolocal L^2 projection method admits $\mathbf{u} \in (H^s(\Omega))^3$ for all $s \ge 0$. The local L^2 projection method requires a little more regularity of $\mathbf{u}$, i.e., $\mathbf{u} \in (H^s(\Omega))^3$ for all $s > \frac{1}{2}$. This is because the latter error estimates on element faces are involved with the trace theorem that requires $s > \frac{1}{2}$; see [12, 14]. However, it is allowed that $\mathbf{u}$ has weaker regularity; i.e., $\mathbf{u} \in (W^{s,r}(\Omega))^3$ with $s > \frac{1}{r}$ and $2 \le r \le \infty$. Then (5.19) would become

$$
(5.25)\qquad ||\mathbf{u}-\mathbf{u}_h||_0 \le C\,h^s\,||\mathbf{u}||_{s,r}, \quad s > \frac{1}{r}, \quad \text{and} \quad 2 \le r \le \infty.
$$

Remark 5.4. In the case when ε is not globally continuous, we assume that ε is a piecewise Lipschitz continuous function. This determines a partition $\mathscr{P}$ of Ω into a finite set of subdomains $\Omega_1, \ldots, \Omega_L$ (which are assumed to be polyhedra). In each Ω_l the restriction of ε is Lipschitz continuous. We denote by Γ_{ij} the faces of $\Omega_i \cap \Omega_j$ and assume that all Γ_{ij} are contained in Ω. As usual, the triangulation $\mathcal{C}_h$ should be

conformed with these material interfaces so that each material interface is the union of the element faces in $\mathcal{C}_h$. Let $\mathcal{E}_{\text{inter},\Gamma} \subset \mathcal{E}_h$ denote the set of all element faces on Γ and on interfaces Γ_{ij}, and let $\mathcal{M}_{\text{inter},\Gamma} \subset \mathcal{M}_h$ be the set of macroelements deduced from $\mathcal{E}_{\text{inter},\Gamma}$ in a similar way as that for $\mathcal{M}_h$ in subsection 3.1. The few modifications to our L^2 projected methods for the case of discontinuous materials are as follows. We need only modify (3.20) by

$$R_h^{\text{inter},\Gamma}([(\varepsilon\,\mathbf{u})\cdot\mathbf{n}])|_F := \frac{1}{|F|}\int_F [(\varepsilon\,\mathbf{u})\cdot\mathbf{n}] \quad \forall F \in \mathcal{E}_{\text{inter},\Gamma}, \tag{5.26}$$

modify the mesh-dependent terms $S_{h,\Gamma}(\cdot,\cdot)$, $Z_{h,\Gamma}(\cdot;\cdot)$ in (3.37) and (3.38) by $S_{h,\text{inter},\Gamma}(\cdot,\cdot)$, $Z_{h,\text{inter},\Gamma}(\cdot;\cdot)$:

$$S_{h,\text{inter},\Gamma}(\mathbf{u},\mathbf{v}) := \sum_{\substack{F\in\mathcal{E}_{\text{inter},\Gamma}\\ M\in\mathcal{M}_{\text{inter},\Gamma}, M=K_1\cup K_2,\\ \text{and } \partial K_1\cap\partial K_2=F\\ \text{or } M=K,\partial K\cap\Gamma=F}} \frac{\sum_{l=1}^{m_{\text{inter},\Gamma}} (\varepsilon\,\mathbf{u}, \bigtriangledown(z_{M,l}\,b_M))_{0,M}\,(\varepsilon\,\mathbf{v}, \bigtriangledown(z_{M,l}\,b_M))_{0,M}}{\sum_{l=1}^{m_{\text{inter},\Gamma}} ||\bigtriangledown\,(z_{M,l}\,b_M)||_{0,M}^2}, \tag{5.27}$$

$$Z_{h,\text{inter},\Gamma}(f;\mathbf{v}) := -\sum_{\substack{F\in\mathcal{E}_{\text{inter},\Gamma}\\ M\in\mathcal{M}_{\text{inter},\Gamma}, M=K_1\cup K_2,\\ \text{and } \partial K_1\cap\partial K_2=F\\ \text{or } M=K,\partial K\cap\Gamma=F}} \frac{\sum_{l=1}^{m_{\text{inter},\Gamma}} (f, z_{M,l}\,b_M)_{0,M}\,(\varepsilon\,\mathbf{v}, \bigtriangledown(z_{M,l}\,b_M))_{0,M}}{\sum_{l=1}^{m_{\text{inter},\Gamma}} ||\bigtriangledown\,(z_{M,l}\,b_M)||_{0,M}^2}, \tag{5.28}$$

modify the term $\int_\Gamma R_h^\Gamma((\varepsilon\,\mathbf{u})\cdot\mathbf{n})\,R_h^\Gamma((\varepsilon\,\mathbf{v})\cdot\mathbf{n})$ in (3.21) by

$$\sum_{F\in\mathcal{E}_{\text{inter},\Gamma}} (R_h^{\text{inter},\Gamma}([(\varepsilon\,\boldsymbol{u})\cdot\boldsymbol{n}]), R_h^{\text{inter},\Gamma}([(\varepsilon\,\boldsymbol{v})\cdot\boldsymbol{n}]))_{0,F}, \tag{5.29}$$

and finally modify the notation S_Γ in (3.31), (3.32), and (4.3) by $S_{\text{inter},\Gamma}$. With these modifications, one may follow the same routine in the previous sections to obtain similar stability results and error estimates.

Remark 5.5. In Remark 3.2, we mentioned that, to consider that $S_{\text{div}}(K), S_\Gamma(K)$, and $S_{\mathbf{curl}}(K)$ are simpler polynomial spaces, one may replace ε by a suitable piecewise polynomial approximation, say, ε_h. This replacement does not affect the theory of stability analysis and error estimates. Simply, one need only work with ε_h, but note that such replacement introduces inconsistent error terms in Lemma 5.1 as follows:

$$(\text{div}\,((\varepsilon-\varepsilon_h)\,\mathbf{u}), \breve{R}_h(\text{div}(\varepsilon_h\,\mathbf{v}_h))) \quad \text{for the local } L^2 \text{ projection method,} \tag{5.30}$$

$$\sum_{K\in\mathcal{C}_h} ((\varepsilon-\varepsilon_h)\,\mathbf{u}, \bigtriangledown \breve{R}_h(\text{div}(\varepsilon_h\,\mathbf{v}_h)))_{0,K} \text{ for the pseudolocal } L^2 \text{ projection method,} \tag{5.31}$$

$$\sum_{K\in\mathcal{C}_h} \frac{\sum_{l=1}^{m_{\text{div}}} ((\varepsilon-\varepsilon_h)\,\mathbf{u}, \bigtriangledown(v_{K,l}\,b_K))_{0,K}\,(\varepsilon_h\,\mathbf{v}_h, \bigtriangledown(v_{K,l}\,b_K))_{0,K}}{\sum_{l=1}^{m_{\text{div}}} ||\bigtriangledown\,(v_{K,l}\,b_K)||_{0,K}^2}, \tag{5.32}$$

$$\sum_{\substack{F\subset\Gamma\\ \text{with } F\subset\partial K}} \frac{\sum_{l=1}^{m_\Gamma} ((\varepsilon-\varepsilon_h)\,\mathbf{u}, \bigtriangledown(z_{K,l}\,b_F))_{0,K}\,(\varepsilon_h\,\mathbf{v}_h, \bigtriangledown(z_{K,l}\,b_F))_{0,K}}{\sum_{l=1}^{m_\Gamma} ||\bigtriangledown\,(z_{K,l}\,b_F)||_{0,K}^2}, \tag{5.33}$$

$$\sum_{j=1}^{N} \frac{1}{|\Sigma_j|}\int_{\Sigma_j} ((\varepsilon-\varepsilon_h)\,\mathbf{u})\cdot\mathbf{n} \int_{\Sigma_j} (\varepsilon_h\,\mathbf{v}_h)\cdot\mathbf{n} \quad \text{for the local } L^2 \text{ projection method.} \tag{5.34}$$

Assume that ε is smooth enough, say, $\varepsilon_{ij} \in H^{\frac{5}{2}}$, $1 \le i,\ j \le 3$. Let ε_h be taken as a C^0 finite element interpolant to ε: for the pseudolocal L^2 projection method, $\varepsilon_{ij,h}|_K \in \mathcal{P}_2^+(K)$ for all $K \in \mathcal{C}_h$, where $\mathcal{P}_2^+(K)$ denotes the quadratic element $\mathcal{P}_2(K)$ plus *one* element bubble, while for the local L^2 projection method, $\varepsilon_{ij,h}|_K \in \mathcal{P}_2^{\square}(K)$ for all $K \in \mathcal{C}_h$, where $\mathcal{P}_2^{\square}(K)$ denotes the quadratic element $\mathcal{P}_2(K)$ plus *four* face bubbles. Recall that $K \in \mathcal{C}_h$ is a tetrahedron. These bubbles ensure that $\varepsilon_{ij,h}$, $1 \le i,\ j \le 3$, satisfy the interpolation property

$$\int_K (\varepsilon_{ij} - \varepsilon_{ij,h}) = 0 \quad \forall K \in \mathcal{C}_h \quad \text{for the pseudolocal } L^2 \text{ projection method} \tag{5.35}$$

and

$$\int_F (\varepsilon_{ij} - \varepsilon_{ij,h}) = 0 \quad \forall F \in \partial K, \forall K \in \mathcal{C}_h, \quad \text{for the local } L^2 \text{ projection method.} \tag{5.36}$$

We have from [13, 20] that

$$||\varepsilon - \varepsilon_h||_{0,K} + h_K\, |\varepsilon - \varepsilon_h|_{1,K} \le C\, h_K^{\frac{5}{2}}\, |\varepsilon|_{\frac{5}{2},K} \quad \forall K \in \mathcal{C}_h, \tag{5.37}$$

where $|\cdot|_1$ denotes the seminorm of H^1. Note that ε_h satisfies the same uniform ellipticity property as ε for a suitably small h.

It suffices to explain how to estimate (5.30) for the local L^2 projection method and (5.31) for the pseudolocal L^2 projection method. Error terms (5.32)–(5.34) can be estimated similarly. We first consider (5.30). Since we have assumed that $\mathbf{u} \in (H^s(\Omega))^3$ with $s > \frac{1}{2}$, letting

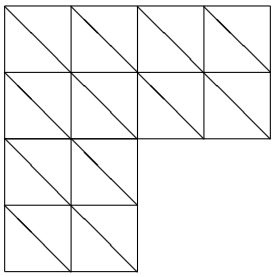
$$\bar{\mathbf{u}}|_K = \frac{1}{|K|} \int_K \mathbf{u} \quad \forall K \in \mathcal{C}_h, \tag{5.38}$$

we can obtain

$$\begin{aligned}
(\operatorname{div}((\varepsilon - \varepsilon_h)\,\mathbf{u}), \breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h))) &= \sum_{K \in \mathcal{C}_h} \breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h)) \sum_{F \subset \partial K} \int_F ((\varepsilon - \varepsilon_h)\,\mathbf{u}) \cdot \mathbf{n} \\
&= \sum_{K \in \mathcal{C}_h} \breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h)) \sum_{F \subset \partial K} \int_F ((\varepsilon - \varepsilon_h)\,(\mathbf{u} - \bar{\mathbf{u}})) \cdot \mathbf{n} \\
&\le C\, h^s\, ||\mathbf{u}||_s\, ||\breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h))||_0.
\end{aligned} \tag{5.39}$$

We next consider (5.31). Since $\mathbf{u} \in (H^s(\Omega))^3$ with $s \ge 0$, we have

$$\begin{aligned}
\sum_{K \in \mathcal{C}_h} ((\varepsilon - \varepsilon_h)\,\mathbf{u}, \bigtriangledown \breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h)))_{0,K} &= \sum_{K \in \mathcal{C}_h} \bigtriangledown \breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h)) \cdot \int_K (\varepsilon - \varepsilon_h)\,\mathbf{u} \\
&= \sum_{K \in \mathcal{C}_h} \bigtriangledown \breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h)) \cdot \int_K (\varepsilon - \varepsilon_h)\,(\mathbf{u} - \bar{\mathbf{u}}) \\
&\le C\, h^s\, ||\mathbf{u}||_s\, ||\breve{R}_h(\operatorname{div}(\varepsilon_h\,\mathbf{v}_h))||_h,
\end{aligned} \tag{5.40}$$



$h = 1/2$

FIG. 2. *L-domain and uniform partition* $\mathcal{C}_h$.

where we have used the norm equivalence between $||\cdot||_0$ and $||\cdot||_h$ (since $||\cdot||_h$ is a norm induced from the trapezoidal quadrature scheme in (3.25)); see [13, 8].

If assuming $\varepsilon_{ij} \in W^{1,\infty}$, $1 \le i, j \le 3$, then for the pseudolocal L^2 projection method, we may choose ε_h as a piecewise linear (plus one element bubble) continuous approximation with $\varepsilon_{ij,h}|_K \in \mathcal{P}_1^+(K)$, where $\mathcal{P}_1^+(K)$ denotes the linear polynomial space plus one element bubble (so $\varepsilon_{ij,h}|_K$ can satisfy (5.35)). From [3, 19, 13] we have $||\varepsilon - \varepsilon_h||_{0,\infty,K} \le C\, h_K\, |\varepsilon|_{1,\infty,K}$ for all $K \in \mathcal{C}_h$, and we can obtain an estimate similar to (5.40); while for the local L^2 projection method, we may choose ε_h as a piecewise linear (plus four face bubbles) continuous approximation of ε with $\varepsilon_{ij,h}|_K \in \mathcal{P}_1^{\square}(K)$, where $\mathcal{P}_1^{\square}(K)$ denotes the linear polynomial space plus four face bubbles (so $\varepsilon_{ij,h}|_K$ can satisfy (5.36)). From [3, 19, 13] we have $||\varepsilon - \varepsilon_h||_{0,K} + h_K\, |\varepsilon - \varepsilon_h|_{1,K} \le C\, h_K^{5/2}\, |\varepsilon|_{1,\infty,K}$ for all $K \in \mathcal{C}_h$, and we can obtain an estimate similar to (5.39).

Note that we may also choose ε_h as a suitable discontinuous piecewise polynomial approximation to ε, provided that (5.35) (or (5.36)) and the corresponding interpolation error estimates are satisfied. For example, for the pseudolocal L^2 projection method, we may choose a discontinuous piecewise constant ε_h, with $\varepsilon_{ij,h}|_K \in \mathcal{P}_0(K)$ for all $K \in \mathcal{C}_h$. In that case, one should work with ε_h following a similar modifying routine for discontinuous materials as highlighted in Remark 5.4.

6. Numerical experiments. In this section we perform some numerical tests for the local L^2 projection method. We consider a two-dimensional problem, with an L domain $\Omega = [-1,1] \times [-1,1] \setminus [0,1] \times [-1,0]$ (see Figure 2).

The continuous problem reads: Find $\mathbf{u}$ such that

$$\operatorname{curl} \mathbf{u} = g, \quad \operatorname{div} \mathbf{u} = f \quad \text{in } \Omega, \quad \mathbf{u} \cdot \mathbf{n} = \chi \quad \text{on } \Gamma = \partial\,\Omega.$$

We first consider a case of nonsmooth solution and take

$$\mathbf{u} = \bigtriangledown \left(r^{\frac{2}{3}} \cos\left(\frac{2\,\theta}{3}\right)\right),$$

where $x = r\,\cos(\theta), y = r\,\sin(\theta)$, and r is the distance to the reentrant corner $(0,0)$ (at the origin) of opening angle $3\,\pi/2$. We determine $g := \operatorname{curl} \mathbf{u}$, $f := \operatorname{div} \mathbf{u}$, and $\chi := \mathbf{u} \cdot \mathbf{n}|_\Gamma$. We also consider a case of smooth solution and take

$$\mathbf{u} = (\sin(\pi x)\cos(\pi y)/2\pi, \cos(\pi x)\sin(\pi y)/2\pi)^t.$$

The regularity for the first $\mathbf{u}$ is $(H^{\frac{2}{3}-\epsilon})^2$ for any $\epsilon \in (0,1)$, and from the theoretical result obtained the error reduction ratio should be approximately $2^{2/3} \approx 1.586$. The

TABLE 1
Relative errors in L^2 norm for nonsmooth solution.

	h=0.5	h=0.25	h=0.125	h=0.0625
$\frac{\|\mathbf{u}-\mathbf{u}_h\|_0}{\|\mathbf{u}\|_0}$	0.119837950	0.075507231	0.047692602	0.030055711

TABLE 2
Relative errors in L^2 norm for smooth solution.

	h=0.5	h=0.25	h=0.125	h=0.0625
$\frac{\|\mathbf{u}-\mathbf{u}_h\|_0}{\|\mathbf{u}\|_0}$	0.053623993	0.008008719	0.001067674	$1.361203187 \times 10^{-4}$

second $\mathbf{u}$ is infinitely smooth; the error reduction ratio should be around 8 since U_h corresponds to a quadratic element. (In two dimensions, the linear element enriched with edge bubbles is none other than the quadratic element.) The calculated results are listed in Tables 1 and 2 as follows. From Table 1 we see that the error reduction ratio is consistent with the predicted value 1.586, and from Table 2 we see that the error reduction ratio is approximately the predicted value 8 as h decreases. These computational results confirm our theoretical estimates.

Acknowledgment. The authors are very thankful to the anonymous referees for valuable comments which led to the improved presentation of this paper.

REFERENCES

[1] A. ALONSO AND A. VALLI, *Some remarks on the characterization of the space of tangential traces of $H(\mathrm{rot};\Omega)$ and the construction of an extension operator*, Manuscripta Math., 89 (1996), pp. 159–178.
[2] C. AMROUCHE, C. BERNARDI, M. DAUGE, AND V. GIRAULT, *Vector potentials in three-dimensional non-smooth domains*, Math. Methods Appl. Sci., 21 (1998), pp. 823–864.
[3] C. BERNARDI AND V. GIRAULT, *A local regularization operator for triangular and quadrilateral finite elements*, SIAM J. Numer. Anal., 35 (1998), pp. 1893–1916.
[4] P. B. BOCHEV AND M. D. GUNZBURGER, *Finite element methods of least-squares type*, SIAM Rev., 40 (1998), pp. 789–837.
[5] A.-S. BONNET-BEN DHIA, C. HAZARD, AND S. LOHRENGEL, *A singular field method for the solution of Maxwell's equations in polyhedral domains*, SIAM J. Appl. Math., 59 (1999), pp. 2028–2044.
[6] A. BOSSAVIT, *Magnetostatic problems in multiply connected regions: Some properties of the curl operator*, IEE Proc., 135 (1988), pp. 179–187.
[7] A. BOSSAVIT, *Computational Electromagnetism: Variational Formulations, Complementarity, Edge Elements*, Academic Press, New York, 1998.
[8] S. C. BRENNER AND L. R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, Berlin, 1996.
[9] J. BRAMBLE AND J. PASCIAK, *A new approximation technique for div-curl systems*, Math. Comp., 73 (2004), pp. 1739–1762.
[10] J. H. BRAMBLE, R. D. LAZAROV, AND J. E. PASCIAK, *A least-squares approach based on a discrete minus one inner product for first-order systems*, Math. Comp., 66 (1997), pp. 935–955.
[11] Z. CAI, T. A. MANTEUFFEL, S. F. MCCORMICK, AND J. RUGE, *First-order system $\mathcal{LL}^*$ (FOSLL*): Scalar elliptic partial differential equations*, SIAM J. Numer. Anal., 39 (2001), pp. 1418–1445.
[12] C. L. CHANG, *Finite element approximation for grad-div type of systems in the plane*, SIAM J. Numer. Anal., 29 (1992), pp. 452–461.
[13] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, Finite Element Methods (Part 1), P. G. Ciarlet and J.-L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.
[14] P. CLÉMENT, *Approximation by finite element functions using local regularization*, RAIRO Anal. Numér., 9 (1975), pp. 77–84.

[15] M. Costabel and M. Dauge, *Singularities of electromagnetic fields in polyhedral domains*, Arch. Ration. Mech. Anal., 151 (2000), pp. 221–276.
[16] M. Costabel and M. Dauge, *Weighted regularization of Maxwell equations in polyhedral domains*, Numer. Math., 93 (2002), pp. 239–277.
[17] C. L. Cox and G. J. Fix, *On the accuracy of least squares methods in the presence of corner singularities*, Comput. Math. Appl., 10 (1984), pp. 463–475.
[18] M. Crouzeix and P.-A. Raviart, *Conforming and nonconforming finite element methods for solving the stationary Stokes equations*, RAIRO Anal. Numér., 7 (1973), pp. 33–75.
[19] V. Girault, *A local projection operator for quadrilateral finite elements*, Math. Comp., 64 (1995), pp. 1421–1431.
[20] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, Berlin, 1986.
[21] R. Hiptmair, *Finite elements in computational electromagnetism*, Acta Numer., 11 (2002), pp. 237–339.
[22] B. N. Jiang, *The Least-Squares Finite Element Method, Theory and Applications in Computational Dynamics and Electromagnetics*, Springer-Verlag, Heidelberg, 1998.
[23] B. N. Jiang, J. Wu, and L. A. Povinelli, *The origin of spurious solutions in computational electromagnetics*, J. Comput. Phys., 125 (1995), pp. 104–123.
[24] U. Kangro and R. Nicolaides, *Divergence boundary conditions for vector Helmholtz equations with divergence constraints*, M2AN Math. Model. Numer. Anal., 33 (1999), pp. 479–492.
[25] T. A. Manteuffel, S. F. McCormick, J. Ruge, and J. G. Schmidt, *First-order system $\mathcal{LL}^*$ (FOSLL*) for general scalar elliptic problems in the plane*, SIAM J. Numer. Anal., 43 (2005), pp. 2098–2120.
[26] P. Monk, *Finite Element Methods for Maxwell Equations*, Clarendon Press, Oxford, UK, 2003.
[27] P. Neittaanmäki and R. Picard, *Error estimates for finite element approximation to a Maxwell-type boundary value problem*, Numer. Funct. Anal. Optim., 2 (1980), pp. 267–285.
[28] P. Pernandes and G. Gilardi, *Magnetostatic and electrostatic problems in inhomogeneous anisotropic media with irregular boundary and mixed boundary conditions*, Math. Models Methods Appl. Sci., 7 (1997), pp. 957–991.
[29] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2564–2576

REDUCED CANCELLATION IN THE EVALUATION OF ENTIRE FUNCTIONS AND APPLICATIONS TO THE ERROR FUNCTION*

W. GAWRONSKI†, J. MÜLLER†, AND M. REINHARD†

Abstract. A general concept for the reduction of cancellation problems in the evaluation of Taylor sections of certain entire functions is proposed. The resulting method is applied to and tested in the case of the (complementary) error function.

Key words. cancellation, error function, entire functions

AMS subject classifications. Primary, 65D20; Secondary, 33F05, 33B20

DOI. 10.1137/060669589

1. Taylor sections of entire functions: Cancellation and how to reduce it.

Let f be an entire function given by its Taylor series

$$f(z) = \sum_{\nu=0}^{\infty} a_\nu z^\nu = \sum_{\nu=0}^{\infty} \frac{f^{(\nu)}(0)}{\nu!} z^\nu \qquad (z \in \mathbb{C})$$

with respect to the origin. Of course, one idea to numerically evaluate $f(z)$ is to truncate the series, that is, to take, for n sufficiently large,

$$s_n(f, z) := \sum_{\nu=0}^{n} a_\nu z^\nu$$

as an approximation to $f(z)$. However, it turns out that, in many interesting cases, serious cancellation problems arise (for a discussion of cancellation effects in general, see, e.g., [SW]), if z is not restricted to a more or less small neighborhood of the origin. The main reason is that the maximal term

$$\mu(r) := \mu_f(r) := \max_{\nu \in \mathbb{N}_0} |a_\nu|\, r^\nu \qquad (r \geq 0)$$

may happen to be much larger than the modulus of the function value $f(z)$ (where $|z| = r$) itself.

How can this phenomenon be quantified?

We write

$$s_{n,p}(f, z) := \sum_{\nu=0}^{n} a_\nu z^\nu$$

for the nth partial sum of the above Taylor series of f if the computations are performed in floating point arithmetic with a precision of p decimal digits and with input data a_ν and z given with an accuracy of p digits. In this situation, the loss of exact

*Received by the editors September 13, 2006; accepted for publication (in revised form) August 3, 2007; published electronically December 5, 2007.
http://www.siam.org/journals/sinum/45-6/66958.html
†Department of Mathematics, University of Trier, D-54286 Trier, Germany (gawron@uni-trier.de, jmueller@uni-trier.de, rein4501@uni-trier.de).

digits in the evaluation of at least one of the values $\operatorname{Re} s_{N(z),p}(f,z)$ and $\operatorname{Im} s_{N(z),p}(f,z)$ may be approximately measured by

$$d_f(z) := \log_{10} \mu_f(|z|) - \log_{10} |f(z)|$$

for $N(z)$ sufficiently large (actually, as is seen, e.g., from [He, p. 27], a more precise lower bound would be $[d_f(z) - \log_{10}(2\sqrt{2})]$ with $[x]$ denoting the greatest integer not exceeding a real number x, but for our considerations this difference can be ignored).

We always assume that the number of terms $N(z) = N_f(z)$ is taken so large that errors do not result from truncation of the series (so, increasing the number of terms does not improve the exactness).

This implies that the number of exact digits (here and in what follows always understood in the sense of the minimal number in both the real and the imaginary part) is approximately given by $\max(p - d_f(z), 0)$ in the case of computations with p exact figures. So one can run into serious problems, if $d_f(z)$ turns out to be large.

Our aim is to reduce such problems by modifying f in an appropriate way. Before we go into the details, we first describe d_f, for entire functions of finite order and type and of regular growth, approximately in terms of the indicator function of f. For that purpose we recall some definitions and facts concerning the growth of entire functions.

The order $\rho = \rho_f$ of a (transcendental) entire function f is given by

$$\rho_f := \limsup_{r\to\infty} \frac{\log\log M_f(r)}{\log r},$$

where

$$M_f(r) := \max_{|z|=r} |f(z)|.$$

We suppose that $0 < \rho < \infty$. Then the type $\tau = \tau_f$ of f is defined as

$$\tau_f := \limsup_{r\to\infty} \frac{\log M_f(r)}{r^\rho}.$$

Since

$$\log \mu_f(r) / \log M_f(r) \to 1 \qquad (r \to \infty)$$

(see, e.g., [Ru, Theorem 10.1]), we can replace $M_f(r)$ in the above definitions by $\mu_f(r)$. Moreover, the growth of f along rays emanating from the origin is asymptotically described by its indicator function $h = h_f$, given by

$$h_f(\vartheta) := \limsup_{r\to\infty} \frac{\log |f(re^{i\vartheta})|}{r^\rho} \qquad (\vartheta \in [-\pi, \pi]).$$

It is well known that h_f is continuous and that $\max_{[-\pi,\pi]} h_f(\vartheta) = \tau$.

Finally, if f has completely regular growth (for a definition, see, e.g., [Le]), the $\limsup_{r\to\infty}$ in the above definitions can be replaced by $\lim^{\substack{r\to\infty \\ r\notin E}}$, where E is a so-called C^0-set, that is, E is the union of circles $\{z : |z - z_j| < r_j\}$ with

$$\lim_{R\to\infty} \frac{1}{R} \sum_{|z_j|<R} r_j = 0.$$

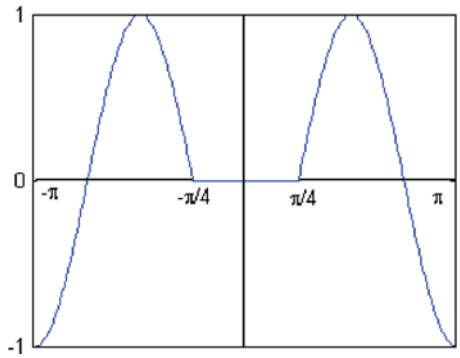


FIG. 1. *Indicator function of* $f(z) = \operatorname{erfc}(-z)$.

Hence

$$\lim_{\substack{r\to\infty \\ r\notin E}} \frac{\log(10)\cdot d_f(re^{i\vartheta})}{r^\rho} = \tau_f - h_f(\vartheta) =: \delta_f(\vartheta),$$

which means that the loss of exact decimal digits for $z = re^{i\vartheta}$ is (up to exceptional values, e.g., near the zeros of f) asymptotically described by $\delta_f(\vartheta)$ in the sense that

$$d_f(re^{i\vartheta}) \sim \delta_f(\vartheta) r^\rho / \log(10) \quad (r\to\infty,\ r\notin E).$$

Let us consider the following example.

Example. Suppose that $f(z) := 1 + \operatorname{erf}(z)$, where erf denotes the (complex) error function; that is,

$$\operatorname{erf}(z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2}\, dt\ .$$

Then, by the definition of the complementary error function erfc, we have

$$f(z) = \operatorname{erfc}(-z)\ .$$

Moreover, from the asymptotic behavior of erfc near ∞ (see, e.g., [AS, Eq. 7.1.23]), we obtain that $\rho_f = 2$ and

$$h_f(\vartheta) = \begin{cases} 0, & |\vartheta| \le \pi/4, \\ -\cos(2\vartheta), & |\vartheta| > \pi/4. \end{cases}$$

See Figure 1. In Algorithm 680 of Poppe and Wijers (see [PW1], [PW2]), which is based on Gautschi's algorithm [Ga] and which may be viewed as a benchmark for algorithms concerning the evaluation of the complex error function (see, e.g., [Wei]), truncation of the Taylor series

$$f(z) = 1 + \operatorname{erf}(z) = 1 + \frac{2}{\sqrt{\pi}} \sum_{\nu=0}^{\infty} \frac{(-1)^\nu z^{2\nu+1}}{(2\nu+1)\nu!} \tag{1}$$

is performed in the second quadrant

$$S = \{z = re^{i\vartheta} : r \ge 0,\ \pi/2 \le \vartheta \le \pi\}$$

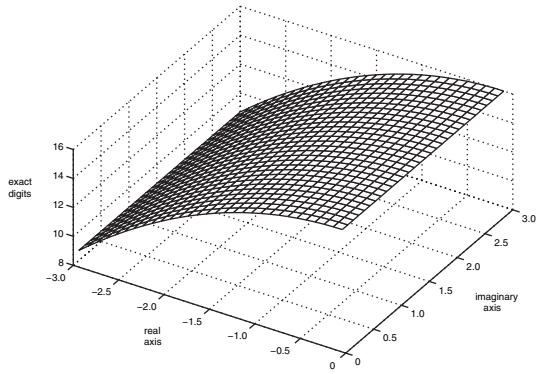


FIG. 2. $16 - \delta_f(\vartheta) r^2 / \log(10)$ *in the second quadrant* S.

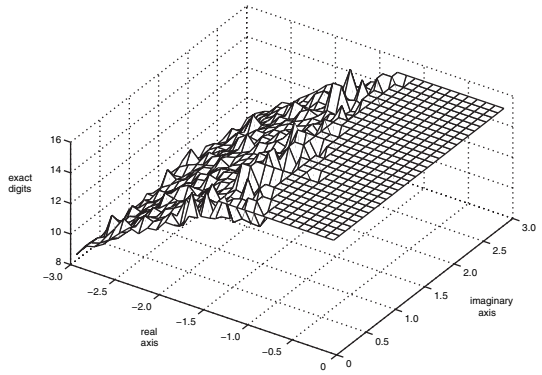


FIG. 3. $e_{f,s_{N(z),16}(f,\cdot)}(z)$ *in* S.

and for $|z| = r$ sufficiently small. Moreover, the calculation in the remaining quadrants is reduced to the second one by elementary operations.

According to the above considerations, we are faced with a loss of significant decimal digits of about

$$\delta_f(\vartheta) r^2 / \log(10) = \begin{cases} r^2 / \log(10), & |\vartheta| \le \pi/4, \\ (1 + \cos(2\vartheta)) r^2 / \log(10), & |\vartheta| > \pi/4, \end{cases}$$

for $z = re^{i\vartheta}$ and r large. So in S, the Taylor sections behave well on the imaginary axis, and the "worst case" appears in the neighborhood of the negative semiaxis (where $\operatorname{erfc}(-z)$ is very small).

Figure 2 shows the values of $16 - \delta_f(\vartheta) r^2 / \log(10)$ which, according to the above remarks, may be viewed as a theoretical measure for the exact digits if computations are performed in floating point arithmetic with a fixed precision of 16 decimal digits.

Moreover, Figure 3 shows $e_{f,s_{N(z),16}(f,\cdot)}(z)$, where

$$e_{f,g}(z) := -\log_{10}\left(\frac{|f(z) - g(z)|}{|f(z)|}\right), \tag{2}$$

that is, the decimal logarithm of the relative error when replacing $f(z)$ by $g(z)$. The value $e_{f,s_{N(z),16}(f,\cdot)}(z)$ measures approximately the smaller of the two numbers of

exact digits of the real part and the imaginary part of $f(z)$, if $f(z)$ is approximated by $s_{N(z),16}(f,z)$ and the computations are performed in double precision arithmetic, providing an accuracy of about 16 decimal digits.

The reference values for $f(z)$ were produced using exact arithmetic from IRRAM (see [Mue1]). Since the usual accuracy requirement of special function routines is 15 digits in double precision, we have cut off the error at a level of 15 digits. Therefore, all values on the 15-digit level represent approximations within the usual tolerance.

The numerical results shown in Figure 3 essentially fit to the theoretical values from Figure 2 and thus support the above considerations about the loss of exact decimal digits.

How can such cancellation problems be reduced? One idea is to modify the function f in such a way that, at least for certain parts of the complex plane, the order of magnitude of the modified $\tilde{f}$ and $\mu_{\tilde{f}}$ do not differ as much as for the function f. Then the Taylor sections $s_n(\tilde{f},z)$ can be computed with less cancellation, and an approximation of $f(z)$ may be obtained from $s_n(\tilde{f},z)$. Such a modification may consist in a multiplication of f by a certain elementary function. More precisely, the general idea is the following: Suppose that $S \subset \mathbb{C}$ is a given set (where f is to be numerically evaluated).

(i) Choose an elementary (entire) function $\varphi = \varphi_S$ so that

$$d_{f\varphi}(z) = \log_{10} \mu_{f\varphi}(|z|) - \log_{10} |f(z)\varphi(z)|$$

is "small" for $z \in S$ and $\varphi(z) \neq 0$ in S.

(ii) Take $\frac{1}{\varphi(z)} s_{N(z),p}(f\varphi,z)$ (for $N(z)$ sufficiently large) as an approximation of $f(z)$ for $z \in S$.

Of course, the question arises of how to choose φ appropriately. If f and φ are of the same order ρ (and of completely regular growth), then

$$h_{f\varphi} = h_f + h_\varphi,$$

and therefore

$$\delta_{f\varphi} = \tau_{f\varphi} - h_f - h_\varphi.$$

If $S = \{re^{i\vartheta} : r \geq 0, \vartheta \in \Theta\}$ is given, then $\varphi = \varphi_S$ should be chosen in such a way that

$$\max_{\vartheta\in\Theta} \delta_{f\varphi}(\vartheta) = \max_{\vartheta\in\Theta}\big(\max_{[-\pi,\pi]} (h_f + h_\varphi) - h_f(\vartheta) - h_\varphi(\vartheta)\big)$$

is small (compare also the considerations in [Mue2]).

If φ is taken from a parametrized family $\Phi = \{\varphi(a,\cdot) : a \in A\}$, we can try to solve the problem

$$\max_{\vartheta\in\Theta}\big(\max_{[-\pi,\pi]} (h_f + h_{\varphi(a,\cdot)}) - h_f(\vartheta) - h_{\varphi(a,\cdot)}(\vartheta)\big) \to \min_{a\in A}. \tag{3}$$

In the next section we study the case of f being the (complementary) error function in some detail.

2. The error function: Reducing errors. We again consider the complementary error function

$$f(z) = \operatorname{erfc}(-z) \qquad (z \in \mathbb{C}).$$

Since f is of order 2, an evident choice for Φ is

$$\Phi = \{\varphi(a,\cdot) : a \in A\}, \qquad \varphi(a,z) := e^{az^2} \qquad (z \in \mathbb{C}),$$

where $A \subset \mathbb{C}$. We (first) restrict our investigations to the case $A = \mathbb{R}$. Then

$$h_{\varphi(a,\cdot)}(\vartheta) = a\cos(2\vartheta) \qquad (\vartheta \in [-\pi,\pi]),$$

and thus, according to section 1,

$$(h_f + h_{\varphi(a,\cdot)})(\vartheta) = \begin{cases} a\cdot\cos(2\vartheta), & |\vartheta| \le \pi/4, \\ (a-1)\cos(2\vartheta), & |\vartheta| > \pi/4. \end{cases}$$

In particular, we obtain

$$\max_{[-\pi,\pi]}(h_f + h_{\varphi(a,\cdot)}) = \max(a, |a-1|) =: \tau(a),$$

and so the optimization problem (3) here reads as

$$\max_{\vartheta\in\Theta}\big(\tau(a) - (h_f + h_{\varphi(a,\cdot)})(\vartheta)\big) \to \min_{a\in\mathbb{R}}. \tag{4}$$

If Θ contains one of the points $\pm\pi/4$ or $\pm 3\pi/4$ (where $h_f - h_{\varphi(a,\cdot)}$ vanishes), then obviously the maximum in (4) is $\ge \tau(a)$, which is minimal exactly for $a = 1/2$. For $a = 1/2$ and $|\vartheta| \le 3\pi/4$ we find

$$\tau(a) - (h_f + h_{\varphi(a,\cdot)})(\vartheta) = \frac{1}{2}(1 - |\cos(2\vartheta)|) \le \frac{1}{2},$$

so $a = 1/2$ turns out to be the (unique) solution of (4) for all subsets Θ of the interval $[-3\pi/4, 3\pi/4]$, with $\{\pm\pi/4, \pm 3\pi/4\} \cap \Theta \ne \emptyset$.

As already mentioned above, in the algorithm of Poppe and Wijers, truncation of the Taylor series

$$f(z) = 1 + \frac{2}{\sqrt{\pi}}\sum_{\nu=0}^{\infty}\frac{(-1)^\nu z^{2\nu+1}}{(2\nu+1)\nu!}$$

is performed in the second quadrant $S = \{z = re^{i\vartheta} : r \ge 0,\ \pi/2 \le \vartheta \le \pi\}$ and for $|z| = r$ sufficiently small.

From the above considerations, the advice is to use Taylor sections of

$$(f\varphi)(z) := f(z)\varphi\left(\frac{1}{2}, z\right) = \operatorname{erfc}(-z)e^{z^2/2} \tag{5}$$

instead.

The following figures show the indicator function of $f\varphi$ and the corresponding approximation for the number of exact digits.

For the computation of the Taylor sections of (5) it is important to have a reasonable representation for the coefficients.

PROPOSITION 1. *For $z \in \mathbb{C}$ we have*

$$(f\varphi)(z) = f(z)\varphi\left(\frac{1}{2}, z\right) = \sum_{\nu=0}^{\infty} b_\nu z^\nu,$$

with

$$b_\nu = \begin{cases} \dfrac{1}{l!2^l} & \text{if} \quad \nu = 2l, \\ \dfrac{1}{\Gamma(l+3/2)\ 2^l} \displaystyle\sum_{k=0}^{l} \binom{k-1/2}{k} (-1)^k & \text{if} \quad \nu = 2l+1. \end{cases} \tag{6}$$

Proof. For $\alpha > 0$ the function

$$M_\alpha(z) := \sum_{k=0}^{\infty} \frac{z^k}{\Gamma(k/\alpha + 1)} \quad (z \in \mathbb{C})$$

is called the Mittag–Leffler function of order α. It is well known that for $\alpha = 2$ the Mittag–Leffler function is closely related to the complementary error function according to

$$\operatorname{erfc}(-z) = M_2(z) e^{-z^2} .$$

Hence, in terms of the Cauchy product we get

$$f(z)\varphi\left(\frac{1}{2}, z\right) = M_2(z) e^{-\frac{1}{2}z^2} = \sum_{\nu=0}^{\infty} b_\nu z^\nu,$$

with

$$b_\nu = \begin{cases} \dfrac{1}{l!2^l} & \text{if} \quad \nu = 2l, \\ \dfrac{1}{\Gamma(l+3/2)} \displaystyle\sum_{k=0}^{l} \binom{l+1/2}{k} \left(\frac{-1}{2}\right)^k & \text{if} \quad \nu = 2l+1. \end{cases} \tag{7}$$

So it remains to consider the case $\nu = 2l + 1$.

The binomial theorem and an index shift show that, for $x \in \mathbb{R}$, $z \in \mathbb{C}$, and $n \in \mathbb{N}_0$, we have

$$\sum_{k=0}^{n} \binom{x-n+k}{k} z^k (1+z)^{n-k} = \sum_{\mu=0}^{n} z^\mu \sum_{k=0}^{\mu} \binom{x-n+k}{k} \binom{n-k}{\mu-k}.$$

Furthermore, we obtain for $a,\ b \in \mathbb{C},\ \mu \in \mathbb{N}_0$

$$\sum_{k=0}^{\mu} \binom{a+k}{k} \binom{b+\mu-k}{\mu-k} = \binom{a+b+\mu+1}{\mu} .$$

Setting $a = x - n,\ b = n - \mu$ we have

$$\sum_{k=0}^{n} \binom{x+1}{k} z^k = \sum_{k=0}^{n} \binom{x-n+k}{k} z^k (1+z)^{n-k}. \tag{8}$$

Then applying (8) and setting $x = l - 1/2,\ n = l,\ z = -1/2$ in (7), we finally get the assertion. ∎

For

$$m_k := \binom{k-1/2}{k}$$

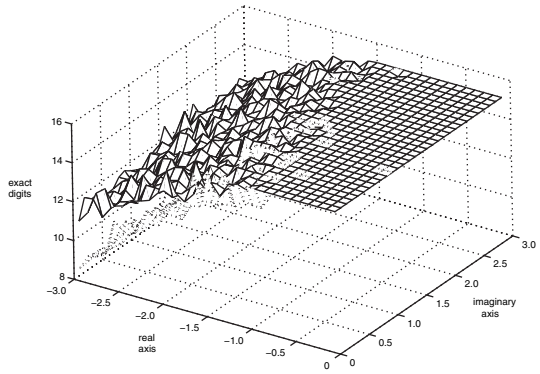


FIG. 4. $e_{f,s_{N(z),16}(f\varphi,\cdot)/\varphi}(z)$ *(solid line) and* $e_{f,s_{N(z),16}(f,\cdot)}(z)$ *(dotted line) in* S.

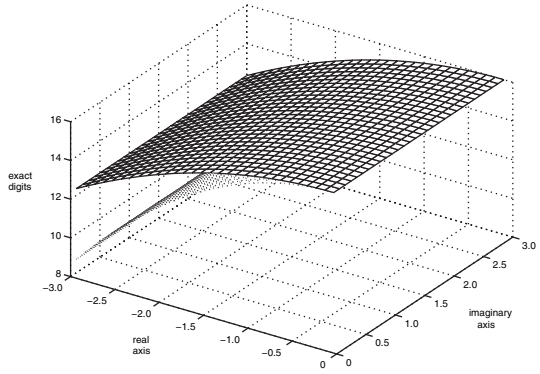


FIG. 5. $16-\delta_{f\varphi}(\vartheta)r^2/\log(10)$ *(solid line) and* $16-\delta_f(\vartheta)r^2/\log(10)$ *(dotted line) in the second quadrant* S.

we have $m_{k+1} = \frac{k+1/2}{k+1}\, m_k$, $k \in \mathbb{N}_0$. Thus, the sum $\sum_{k=0}^{l} \binom{k-1/2}{k}(-1)^k$ in (6) and therefore also the Taylor coefficients b_ν of (5) can be evaluated recursively (note that $\Gamma(l+\frac{3}{2}) = (l+\frac{1}{2})(l-\frac{1}{2})\ldots\frac{1}{2}\sqrt{\pi}$).

Even more suitable for numerical purposes is the two-term recursion

$$(\nu+1)(\nu+2)b_{\nu+2} = b_\nu + b_{\nu-2}, \quad b_{-2} := b_{-1} := 0,\ b_0 = 1,\ b_1 = \frac{2}{\sqrt{\pi}}$$

for the coefficients b_ν, which may be found by applying the above relations twice. More directly, this recursion follows from the fact that $F := f\varphi$ satisfies the differential equation $F'' = (1+z^2)F$ (which was pointed out by one of the referees).

We take $\frac{1}{\varphi(z)} s_{N(z),16}(f\varphi, z)$ (for $N(z)$ sufficiently large) as an approximation of $f(z)$ for $z \in S$ of small modulus and compare the results with the exact values of $f(z)$, which again were produced using exact arithmetic (IRRAM).

The numerical results shown in Figure 4 support our theoretical considerations which are illustrated in Figure 5 and demonstrate the advantages of this method compared to the computation of the Taylor sections of f (see Figures 2 and 3).

Although there is an improvement, the Taylor sections of $f\varphi$ also turn out to be numerically unstable with respect to cancellation near the negative axis, where $\delta_{f\varphi}$ is maximal (cf. Figure 6). So the question arises whether there is an appropriate entire function ψ such that $\delta_{f\psi}(\vartheta)$ is smaller than $\delta_{f\varphi}(\vartheta)$ for $\vartheta \approx \pi$.

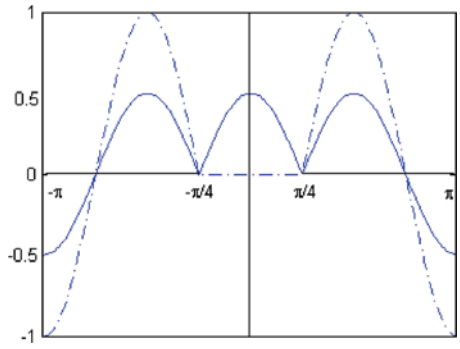


FIG. 6. *The indicator functions of $f\varphi$ (solid curve) and f (dotted curve).*

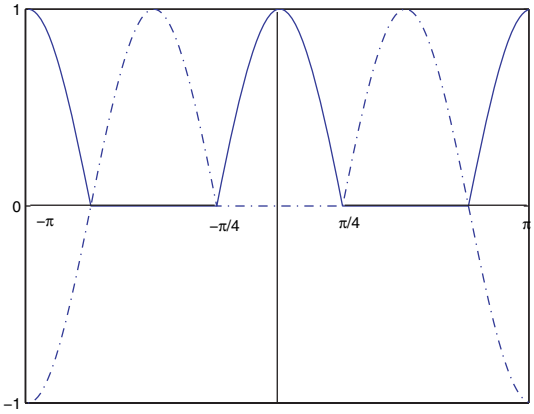


FIG. 7. *The indicator functions of $f\psi$ (solid curve) and f (dotted curve).*

If we take

$$\psi(z) = \operatorname{erfc}(z)\ e^{2z^2}\ ,$$

then we obtain for the indicator function

$$h_{f\psi}(\vartheta) = \begin{cases} 0, & \pi/4 \le |\vartheta| \le 3\pi/4, \\ \cos(2\vartheta), & \text{otherwise}, \end{cases}$$

and therefore $\delta_{f\psi}(\pi) = 0$. See Figure 7.

Thus it is reasonable to take $\frac{1}{\psi(z)} s_{N(z),16}(f\psi, z)$ as an approximation of $f(z)$ in particular close to the negative real axis (see Figure 8). The multiplication by $1/\psi(z)$ in the second quadrant S requires the evaluation of $\operatorname{erfc}(w)$ in the fourth (or the first) quadrant. In this part of the plane, the Taylor sections of $f\varphi$ from (5) turn out to be sufficiently well behaved. So we actually replace $\psi(z)$ by $\tilde{\psi}(z) := e^{2z^2} s_{N(-z),16}(f\varphi, -z)/\varphi(-z)$.

In order to find the Taylor coefficients of $f\psi$ with respect to the origin we just have to apply the Cauchy product again. This leads to

$$(f\psi)(z) = \operatorname{erfc}(-z)\psi(z) = M_2(z)\ M_2(-z) = \sum_{\nu=0}^{\infty} c_\nu z^\nu, \tag{9}$$

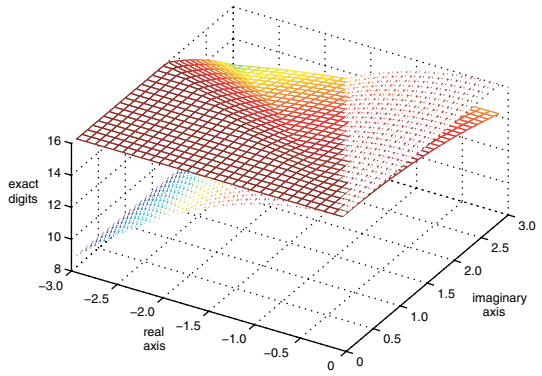


FIG. 8. $16 - \delta_{f\psi}(\vartheta)r^2/\log(10)$ *(solid line) and* $16 - \delta_f(\vartheta)r^2/\log(10)$ *(dotted line) in* S.

with

$$c_\nu = \begin{cases} 0 & \text{if } \nu \text{ is odd,} \\ \displaystyle\sum_{\mu=0}^{\nu} \frac{(-1)^\mu}{\Gamma(\frac{1}{2}\mu+1)\Gamma(\frac{1}{2}(\nu-\mu)+1)} & \text{if } \nu \text{ is even.} \end{cases}$$

As in the case of $f\varphi$ above, the coefficients c_ν, ν even, can be evaluated by recursion.

Setting $r := \nu/2$, we write

$$d_r := \sum_{\mu=0}^{2r} \frac{(-1)^\mu}{\Gamma(\frac{1}{2}\mu+1)\Gamma(r+1-\frac{1}{2}\mu)}, \quad r \in \mathbb{N}_0.$$

We get

$$d_r = \frac{2^r}{r!} - s_{r-1}, \tag{10}$$

with

$$s_n := \sum_{\nu=0}^{n} \frac{1}{\Gamma(\nu+\frac{3}{2})\Gamma(n-\nu+\frac{3}{2})}, \quad n \in \mathbb{N}_0, \text{ and } s_0 = \frac{4}{\pi}.$$

Then we can obtain that

$$s_{n+1} = \frac{2}{n+2}s_n + \frac{1}{(n+2)\Gamma(\frac{3}{2})\Gamma(n+\frac{5}{2})}, \; n \in \mathbb{N}_0, \; s_0 = \frac{4}{\pi}. \tag{11}$$

If we understand (11) as a difference equation, we get

$$s_n = \frac{2^n}{(n+1)!}\sum_{\nu=0}^{n} \frac{\nu!}{2^\nu\Gamma(\frac{3}{2})\Gamma(\nu+\frac{3}{2})}. \tag{12}$$

For the recursion to evaluate the coefficients d_r we obtain

$$d_{r+1} - \frac{2}{r+1}d_r = \frac{2^{r+1}}{(r+1)!} - s_r - \frac{2^{r+1}}{(r+1)!} + \frac{2}{r+1}s_{r-1} = -\frac{1}{(r+1)\Gamma(\frac{3}{2})\Gamma(r+\frac{3}{2})}$$

and therefore

$$d_{r+1} = \frac{2}{r+1} d_r - \frac{2}{(r+1)\pi} \prod_{\nu=0}^{r} \frac{2}{2\nu+1}, \ r \in \mathbb{N}_0, \ a_0 = 1.$$

It turns out, however, that this recursion tends to be unstable if performed upwards. Fortunately, this problem no longer occurs if we apply it in the *backward* direction, that is, we compute with an appropriate starting value $d_{R(z)}$ (or an approximation $\tilde{d}_{R(z)}$)

$$d_r = \frac{r+1}{2} d_{r+1} + \frac{1}{\pi} \prod_{\nu=0}^{r} \frac{2}{2\nu+1}$$

for $r = R(z) - 1, \ldots, 1, 0$. Of course, in this case the question arises of how to get the starting value. Since d_r tends to 0 very rapidly as r tends to ∞, it is possible to simply take $\tilde{d}_{R(z)} = 0$ for $R(z)$ sufficiently large.

In our case suitable values of $R(z)$ could be computed explicitly by using the representation

$$d_r = \frac{4}{\pi r!} \int_0^1 \frac{(1-\xi^2)^r}{1+\xi^2} d\xi, \quad r \in \mathbb{N}_0, \tag{13}$$

which follows from (10) and (12) after some routine calculations employing Euler's Beta integral. From (13) we also have the estimate

$$\sqrt{r}\ r!\ d_r \le \frac{2}{\sqrt{\pi}}$$

being asymptotically sharp as $r \to \infty$.

Similarly as in the case of $f\varphi$, a two-term recursion for the coefficients c_ν can be obtained from the fact that the function $G := f\psi$ satisfies the differential equation

$$G'' = 4(1-2z^2)G + 6zG' - \frac{8}{\pi}$$

(also pointed out by the referee), namely,

$$(\nu+1)(\nu+2)c_{\nu+2} = (4+6\nu)c_\nu - 8c_{\nu-2}.$$

This recursion is again stable (only) in the backward direction. Since now we have a second-order homogeneous equation, and since the exact value $c_0 = 1$ is known, the backward recurrence may be started with the (false) values $\hat{c}_K = 1$ and $\hat{c}_{K+2} = 0$ (for K large enough), and then the exact values are obtained by rescaling with factor $c_0/\hat{c}_0 = 1/\hat{c}_0$ (Miller's algorithm; cf. [Wi, section 4]).

The numerical results shown in Figure 9 demonstrate the efficiency of the proposed method for the evaluation $f(z)$ in particular near the negative real axis.

Obviously, the question arises of how the proposed method compares with the existing software for the computation of the error function. Algorithm 680 of Poppe and Wijers works with νth convergents of a certain continued fraction of $f(z)$ for z outside of a bounded region (and with ν depending on z). As already mentioned above, for z near the origin partial sums of the series (1) are used as an approximation.

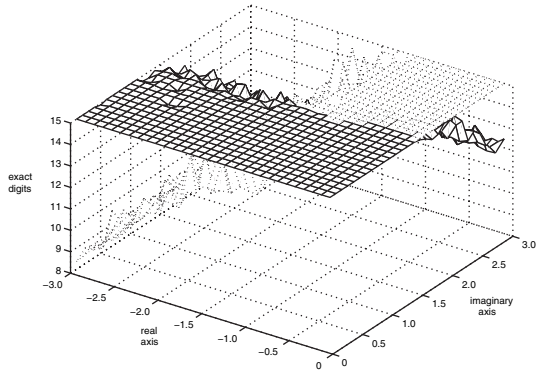


FIG. 9. $e_{f,s_{N(z),16}(f\psi,\cdot)/\tilde{\psi}}(z)$ *(solid line) and* $e_{f,s_{N(z),16}(f,\cdot)}(z)$ *(dotted line) in* S.

Moreover, there is an intermediate region in which Taylor sections not centered at the origin are applied. In this case, the computation of the coefficients is again based on continued fractions. Thus, the most challenging part is the intermediate region.

A combination of the continued fractions in the outer region as in Algorithm 680 and the approximations based on the Taylor expansions of $f\varphi$ and $f\psi$ as above yields an accuracy of at least 14 digits in the second quadrant near the axis but not in an intermediate part close to the line $\arg(z) = 3\pi/4$.

It turns out that a further improvement concerning the accuracy is possible. Multiplication of $f\psi$ with $\varphi(-1/2, z) = e^{-z^2/2}$ results in the indicator function

$$h_{f\psi\varphi(-1/2,\cdot)}(\vartheta) = \frac{1}{2}|\cos(2\vartheta)| \; .$$

The above theory shows that $f\psi\varphi(-1/2, \cdot)$ shares the advantages of $f\varphi$ and $f\psi$ concerning the reduction of cancellation. We were, however, not able to find a reasonable recursion relation for the coefficients, and, unfortunately, the computation from the coefficients of $f\psi$ and $\varphi(-1/2, \cdot)$ by convolution leads again to cancellation. So it seems necessary to evaluate the coefficients using exact arithmetic and then to implement them as data.

If we agree with the same disadvantage, then we can do even better in the area in which we are interested, namely, near the line $\arg(z) = 3\pi/4$.

With

$$a := e^{3\pi i/4} - 1/2$$

and $\varphi(a, z) = e^{az^2}$, we obtain for $f\psi\varphi(a, \cdot)$ the indicator function

$$h_{f\psi\varphi(a,\cdot)}(\vartheta) = h_{f\psi\varphi(-1/2,\cdot)}(\vartheta) + \cos(2\vartheta - 3\pi/2)$$

(see Figure 10), which is comparably "large" for $\vartheta \approx 3\pi/4$.

Numerical experiments confirm that a combination of the four types of approximants based on the continued fractions and the Taylor series of $f\varphi$, $f\psi$, and $f\psi\varphi(a, \cdot)$ provide a reasonable method in the sense that at least 14 exact figures are reached in the second quandrant (and thus in all of the complex plane; cf. [PW1]).

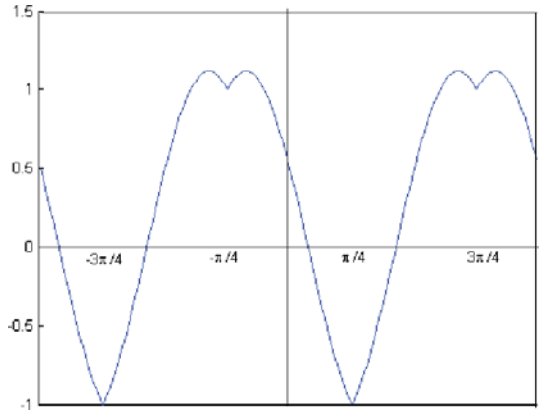


FIG. 10. *Indicator function of* $f\psi\varphi(a,\cdot)$.

Conclusion. The above numerical results have shown that an essential improvement with respect to cancellation is possible in the case of the error function if the growth of the function (measured in terms of the indicator function) is taken into account. Of course, similar ideas apply to other entire functions as, for example, confluent hypergeometric functions or Mittag–Leffler functions. In these cases, the corresponding indicator functions are also easily derived from the asymptotic behavior.

Acknowledgment. The authors express their gratitude to two referees whose expertise led to an improvement of the paper.

REFERENCES

[AS] M. ABRAMOVITZ AND I. A. STEGUN, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, National Bureau of Standards, Washington, D.C., 1964.

[Ga] W. GAUTSCHI, *Efficient computation of the complex error function*, SIAM J. Numer. Anal., 7 (1970), pp. 187–198.

[He] P. HENRICI, *Essentials of Numerical Analysis with Pocket Calculator Demonstrations*, Wiley, New York, 1982.

[Le] B. YA. LEVIN, *Distribution of Zeros of Entire Functions*, Transl. Math. Monogr. 5, American Mathematical Society, Providence, RI, 1980.

[Mue2] J. MÜLLER, *Accelerated polynomial approximation of finite order entire functions by growth reduction*, Math. Comp., 66 (1997), pp. 743–761.

[Mue1] N. TH. MÜLLER, *The iRRAM: Exact Arithmetic in C++*, Lecture Notes in Comput. Sci. 2064, Springer, Berlin, 2001, pp. 222–252.

[PW1] G. P. M. POPPE AND C. M. J. WIJERS, *More efficient evaluation of the complex error function*, ACM Trans. Math. Software, 16 (1990), pp. 38–46.

[PW2] G. P. M. POPPE AND C. M. J. WIJERS, *Algorithm* 680*: Evaluation of the complex error function*, ACM Trans. Math. Software, 16 (1990), p. 47.

[Ru] L. A. RUBEL, *Entire and Meromorphic Functions*, Springer, New York, 1996.

[SW] R. SCHABACK AND H. WENDLAND, *Numerische Mathematik*, Springer, Berlin, Heidelberg, New York, 2005.

[Wei] J. A. C. WEIDEMAN, *Computation of the complex error function*, SIAM J. Numer. Anal., 31 (1994), pp. 1497–1518.

[Wi] J. WIMP, *Computation with Recurrence Relations*, Pitman, Boston, London, Melbourne, 1984.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2577–2609

CONVERGENCE OF A VARIANT OF THE ZIPPER ALGORITHM FOR CONFORMAL MAPPING*

DONALD E. MARSHALL† AND STEFFEN ROHDE†

Abstract. In the early 1980s an elementary algorithm for computing conformal maps was discovered by R. Kühnau and the first author. The algorithm is fast and accurate, but convergence was not known. Given points $z_0, \ldots, z_n$ in the plane, the algorithm computes an explicit conformal map of the unit disk onto a region bounded by a Jordan curve γ with $z_0, \ldots, z_n \in \gamma$. We prove convergence for Jordan regions in the sense of uniformly close boundaries and give corresponding uniform estimates on the closed region and the closed disc for the mapping functions and their inverses. Improved estimates are obtained if the data points lie on a C^1 curve or a K-quasicircle. The algorithm was discovered as an approximate method for conformal welding; however, it can also be viewed as a discretization of the Loewner differential equation.

Key words. numerical conformal mapping, zipper algorithm, hyperbolic geodesics

AMS subject classifications. Primary, 30C30; Secondary, 65E05

DOI. 10.1137/060659119

Introduction. Conformal maps have applications to problems in physics, engineering, and mathematics, but how do you find a conformal map, say, of the upper-half plane $\mathbb{H}$ to a complicated region? Rather few maps can be given explicitly by hand, so that a computer must be used to find the map approximately. One reasonable way to describe a region numerically is to give a large number of points on the boundary (see Figure 1). One way to say that a computed map defined on $\mathbb{H}$ is "close" to a map to the region is to require that the boundary of the image be uniformly close to the polygonal curve through the data points. Indeed, the only information we may have about the boundary of a region is this collection of data points.

In the early 1980s an elementary algorithm was discovered independently by Kühnau [K] and the first author. The algorithm is fast and accurate, but convergence was not known. The purpose of this paper is to prove convergence in the sense of uniformly close boundaries and discuss related numerical issues. In many applications both the conformal map and its inverse are required. One important aspect of the algorithm that sets it apart from others is that this algorithm finds both maps simultaneously.

The algorithm can be viewed as an approximate solution to a conformal welding problem or as a discretization of the Loewner differential equation. The approximation to the conformal map is obtained as a composition of conformal maps onto slit half planes. Osculation methods also approximate a conformal map by repeated composition of simple maps. See Henrici [H] for a discussion of osculation methods and uniform convergence on compact sets. The algorithms of the present article follow the boundary of a given region much more closely than, for instance, the Koebe algorithm and give uniform convergence on all of $\mathbb{H}$ rather than just on compact subsets. Uniform convergence on the closure of the region is particularly important in applications

*Received by the editors May 5, 2006; accepted for publication (in revised form) April 13, 2007; published electronically December 7, 2007. The authors were supported in part by NSF grants DMS-0201435 and DMS-0244408.

http://www.siam.org/journals/sinum/45-6/65911.html

†Department of Mathematics, University of Washington, Seattle, WA 98195-4350 (marshall@math.washington.edu, rohde@math.washington.edu).

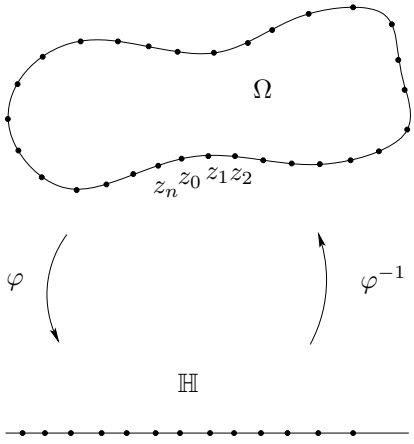


FIG. 1.

that involve boundary values of functions defined on the region. It is possible to use the techniques of this paper to prove that a version of the algorithm is an osculation method for smooth curves, and therefore by the results in [H] repeated applications converge, in the weaker sense, uniformly on compact subsets. However, prior to this article, even a proof that these methods satisfied the osculation family conditions was not known.

Depending on the type of slit (hyperbolic geodesic, straight line segment, or circular arc) we actually obtain different versions of this algorithm. These are described in section 1. We then focus our attention on the "geodesic algorithm" and study its behavior in different situations. The easiest case is discussed in section 2: If the data points $z_0, z_1, \ldots$ are the consecutive contact points of a chain of disjoint discs (see Figures 7 and 8 below), then a simple but very useful reinterpretation of the algorithm, together with the hyperbolic convexity of discs in simply connected domains (Jørgensen's theorem), implies that the curve produced by the algorithm is confined to the chain of discs (Theorem 2.2). One consequence is that for any bounded simply connected domain Ω, the geodesic algorithm can be used to compute a conformal map to a Jordan region Ω_c ("c" is for "computed") so that the Hausdorff distance between $\partial\Omega$ and $\partial\Omega_c$ is as small as desired (Theorem 2.4).

In section 3, we describe an extension of the ideas of section 2 that applies to a variety of domains such as smooth domains or quasiconformal discs with small constants, with better estimates. For instance, if $\partial\Omega$ is a C^1 curve, then the geodesic algorithm can be used to compute a conformal map to a Jordan region Ω_c with $\partial\Omega_c \in C^1$ so that the boundaries are uniformly close and so that the unit tangent vectors are uniformly close (Theorem 3.10). The heart of the convergence proof in these cases comprises the technical "self-improvement" in Lemmas 3.5 and 3.6. In fact, this approach constituted our first convergence proof.

In sections 4 and 5, we show how estimates on the distance between boundaries of Jordan regions give estimates for the uniform distance between the corresponding conformal maps to $\mathbb{D}$, and we apply these estimates to obtain bounds for the convergence of the conformal maps produced by the algorithm. We summarize some of our results as follows: If $\partial\Omega$ is contained in a chain of discs of radius $\leq \epsilon$ with the data points being the contact points of the discs, or if $\partial\Omega$ is a K-quasicircle with K close to one and the data points are consecutive points on $\partial\Omega$ of distance comparable to ϵ,

then the Hausdorff distance between $\partial\Omega$ and the boundary of the domain computed by the geodesic algorithm, $\partial\Omega_c$, is at most ε. Moreover, the conformal maps φ, φ_c onto $\mathbb{D}$ satisfy

$$\sup_{\Omega\cap\Omega_c} |\varphi - \varphi_c| \le C\epsilon^p,$$

where any $p < 1/2$ works in the disc-chain case, and p is close to one if K is close to one. In the case of quasicircles, we also have

$$\sup_{\mathbb{D}} |\varphi^{-1} - \varphi_c^{-1}| \le C\epsilon^p$$

with p close to one. Better estimates are obtained for regions bounded by smoother Jordan curves.

Section 6 contains a brief discussion of numerical results. The appendix has a simple self-contained proof of Jørgensen's theorem.

In a forthcoming paper we plan to address the convergence of the slit and zipper variants of the algorithm. The basic conformal maps and their inverses used in the geodesic algorithm are given in terms of linear fractional transformations, squares, and square roots. The slit and zipper algorithms use elementary maps whose inverses cannot be written in terms of elementary maps. In that paper we will discuss how to divide the plane into four regions so that Newton's method applied to variants of the inverses will converge quadratically in each region. Newton's method converges so rapidly that it virtually provides a formula for the inverses.

1. Conformal mapping algorithms. The geodesic algorithm. The most elementary version of the conformal mapping algorithm is based on the simple map $f_a : \mathbb{H} \setminus \gamma \longrightarrow \mathbb{H}$, where γ is an arc of a circle from 0 to $a \in \mathbb{H}$ which is orthogonal to $\mathbb{R}$ at 0.

This map can be realized by a composition of a linear fractional transformation, the square, and the square root map, as illustrated in Figure 2. The orthogonal circle

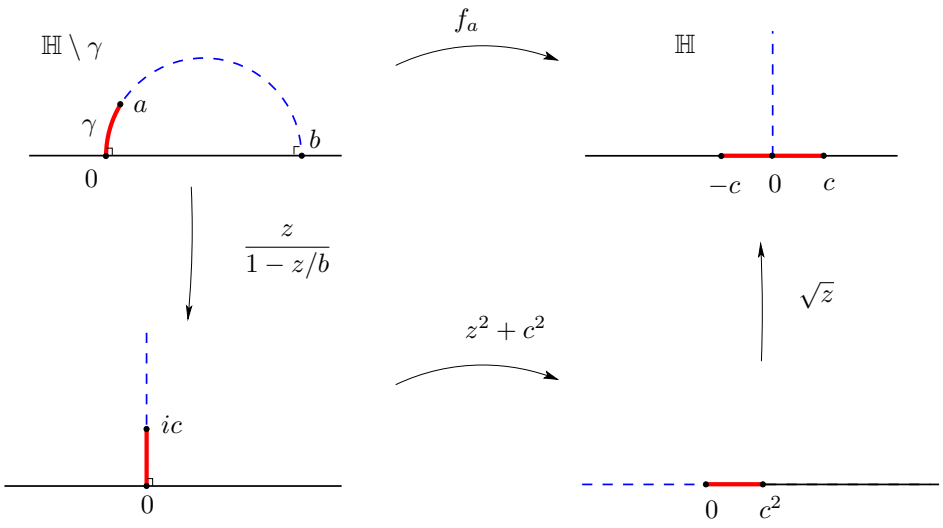


FIG. 2. *The basic map f_a.*

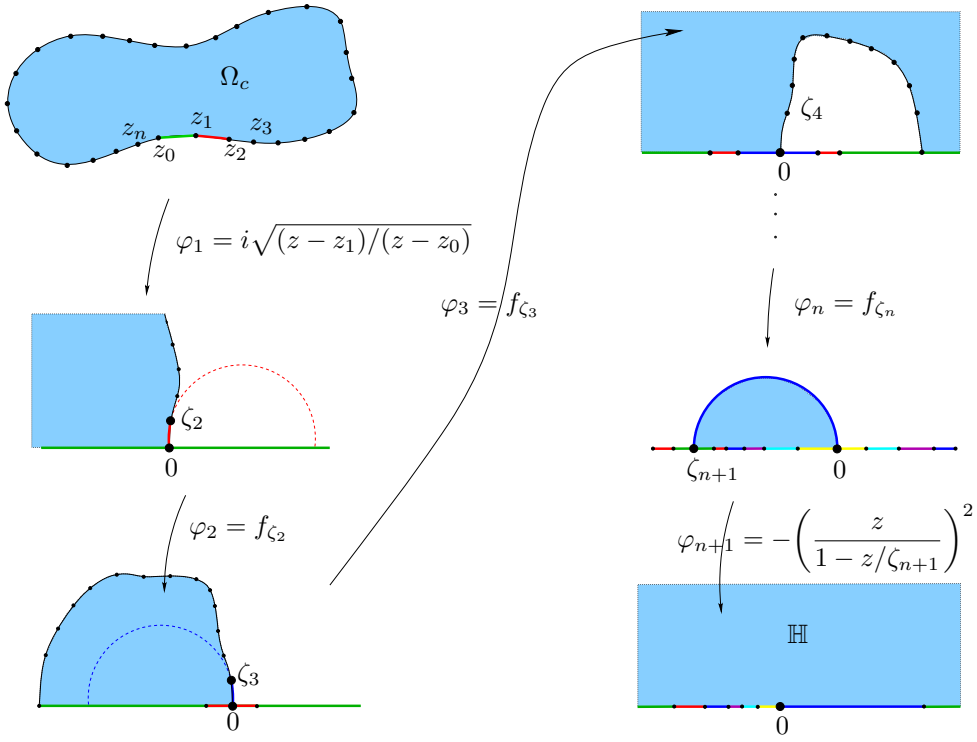


FIG. 3. *The geodesic algorithm.*

also meets $\mathbb{R}$ orthogonally at a point $b = |a|^2/\operatorname{Re} a$ and is illustrated by a dashed curve in Figure 2. In Figure 2, $c = |a|^2/\operatorname{Im} a$. Observe that the arc γ is opened to two adjacent intervals at 0 with a, the tip of γ, mapped to 0. The inverse f_a^{-1} can be easily found by composing the inverses of these elementary maps in the reverse order.

Now suppose that $z_0, z_1, \ldots, z_n$ are points in the plane. The basic maps f_a can be used to compute a conformal map of $\mathbb{H}$ onto a region Ω_c bounded by a Jordan curve which passes through the data points, as illustrated in Figure 3.

The complement in the extended plane of the line segment from z_0 to z_1 can be mapped onto $\mathbb{H}$ with the map

$$\varphi_1(z) = i\sqrt{\frac{z-z_1}{z-z_0}},$$

$\varphi_1(z_1) = 0$, and $\varphi_1(z_0) = \infty$. Set $\zeta_2 = \varphi_1(z_2)$ and $\varphi_2 = f_{\zeta_2}$. Repeating this process, define

$$\zeta_k = \varphi_{k-1} \circ \varphi_{k-2} \circ \cdots \circ \varphi_1(z_k)$$

and

$$\varphi_k = f_{\zeta_k}$$

for $k = 2, \ldots, n$. Finally, map a half disc to $\mathbb{H}$ by letting

$$\zeta_{n+1} = \varphi_n \circ \cdots \circ \varphi_1(z_0) \in \mathbb{R}$$

be the image of z_0, and set

$$\varphi_{n+1} = \pm\left(\frac{z}{1 - z/\zeta_{n+1}}\right)^2.$$

The + sign is chosen in the definition of φ_{n+1} if the data points have a negative winding number (clockwise) around an interior point of $\partial\Omega$, and otherwise the $-$ sign is chosen. Set

$$\varphi = \varphi_{n+1} \circ \varphi_n \circ \cdots \circ \varphi_2 \circ \varphi_1$$

and

$$\varphi^{-1} = \varphi_1^{-1} \circ \varphi_2^{-1} \circ \cdots \circ \varphi_{n+1}^{-1}.$$

Then φ^{-1} is a conformal map of $\mathbb{H}$ onto a region Ω_c such that $z_j \in \partial\Omega_c$, $j = 0, \ldots, n$. The portion γ_j of $\partial\Omega_c$ between z_j and z_{j+1} is the image of the arc of a circle in the upper-half plane by the analytic map $\varphi_1^{-1} \circ \cdots \circ \varphi_j^{-1}$. In more picturesque language, after applying φ_1, we grab the ends of the displayed horizontal line segment and pull, splitting apart or unzipping the curve at 0. The remaining data points move down until they hit 0, and then each splits into two points, one on each side of 0, moving further apart as we continue to pull.

As an aside, we make a few comments. As mentioned, $\partial\Omega_c$ is piecewise analytic. A curve is called C^1 if the arc length parameterization has a continuous first derivative. In other words, the direction of the unit tangent vector is continuous. It is easy to see that $\partial\Omega_c$ is also C^1 since the inverse of the basic map f_a in Figure 2 doubles angles at 0 and halves angles at $\pm c$. In fact, it is also $C^{\frac{3}{2}}$ (see Proposition 3.12). If the data points $\{z_j\}$ lie on the boundary of a given region $\partial\Omega$, the analyticity of $\partial\Omega_c$ also allows us in many situations (see Proposition 2.5 and Corollary 3.9) to extend φ_c analytically across $\partial\Omega_c$ so that the extended map is a conformal map of Ω onto a region with boundary very close to $\partial\mathbb{D}$. Note also that φ is a conformal map of the complement of Ω_c, $\mathbb{C}^* \setminus \overline{\Omega_c}$, onto the lower-half plane, $\mathbb{C} \setminus \overline{\mathbb{H}}$, where $\mathbb{C}^*$ denotes the extended plane. Simply follow the unshaded region in $\mathbb{H}$ in Figure 3. Finally, we remark that it is easier to use geodesic arcs in the right-half plane instead of in the upper-half plane when coding the algorithm because of the usual convention that $-\frac{\pi}{2} < \arg\sqrt{z} \leq \frac{\pi}{2}$.

The slit algorithm. Given a region Ω, we can select boundary points $z_0, \ldots, z_n$ on $\partial\Omega$ and apply the geodesic algorithm. We can view the circular arcs γ for the basic maps f_a as approximating the image of the boundary of Ω between 0 and a with a circular arc at each stage. We can improve the approximation by using straight lines instead of orthogonal arcs. So in the slit algorithm we replace the inverse of the maps f_a by conformal maps $g_a : \mathbb{H} \longrightarrow \mathbb{H} \setminus \mathrm{L}$, where L is a line segment from 0 to a. Explicitly

$$g_a(z) = C(z - p)^p (z + 1 - p)^{1-p},$$

where $p = \arg a/\pi$ and $C = |a|/p^p(1-p)^{1-p}$.

One way to see that g_a is a conformal map is to note that as x traces the real line from $-\infty$ to $+\infty$, $g_a(x)$ traces the boundary of $\mathbb{H} \setminus \mathrm{L}$ and $g_a(z) \sim Cz$ for large z, and then apply the argument principle. Another method would be to construct $\mathrm{Re}\log g_a$ using harmonic measure, as in the first two pages of [GM]. As in the basic maps of the

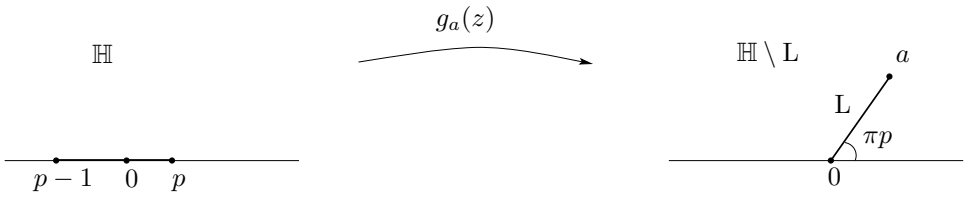


FIG. 4. *The slit maps.*

geodesic algorithm, the line segment from 0 to a is opened to two adjacent intervals on $\mathbb{R}$ by $f_a = g_a^{-1}$ with $f_a(a) = 0$ and $f_a(\infty) = \infty$. The map f_a cannot be written in terms of elementary functions, but an effective and rapid numerical inverse will be described in a subsequent paper.

We note that, as in the geodesic algorithm, the boundary of the region Ω_c computed with the slit algorithm will be piecewise analytic. However, it will not be C^1. Indeed, if g_a is the map illustrated by Figure 4, and if g_b is another such map, then $g_b \circ g_a$ forms a curve with angles $2\pi p$ and $2\pi(1-p)$ on either side of the curve at $b = g_b(0)$. Since analytic maps preserve angles, the boundary of the computed region consists of analytic arcs with endpoints at the data points, and angles determined by the basic maps. This will allow us to accurately compute conformal maps to regions with (a finite number of) "corners" or "bends."

The zipper algorithm. We can further improve the approximation by replacing the linear slits with arcs of (nonorthogonal) circles. In this version we assume there is an even number of boundary points, $z_0, z_1, \ldots, z_{2n+1}$. The first map is replaced by

$$\varphi_1(z) = \sqrt{\frac{(z-z_2)(z_1-z_0)}{(z-z_0)(z_1-z_2)}},$$

which maps the complement in the extended plane of the circular arc through z_0, z_1, z_2 onto $\mathbb{H}$. At each subsequent stage, instead of pulling down one point ζ_k, we can find a unique circular arc through 0 and the (images of) the next two data points ζ_{2k-1} and ζ_{2k}. By a linear fractional transformation ℓ_a which preserves $\mathbb{H}$, this arc is mapped to a line segment (assuming the arc is not tangent to $\mathbb{R}$ at 0). See Figure 5.

The complement of this segment in $\mathbb{H}$ can then be mapped to $\mathbb{H}$ as described in the slit algorithm, using g_d^{-1}, where $d = a/(1-a/b)$. The composition $h_{a,c} = g_d^{-1} \circ \ell_a$ then maps the complement of the circular arc in $\mathbb{H}$ onto $\mathbb{H}$. Thus at each stage we are giving a "quadratic approximation" instead of a linear approximation to the (image of) the boundary. The last map φ_{n+1} is a conformal map of the intersection of a disc with $\mathbb{H}$ where the boundary circular arc passes through 0, the image of z_{2n+1}, and the image of z_0 by the composition $\varphi_n \circ \cdots \circ \varphi_1$. See Figure 6.

If the zipper algorithm is used to approximate the boundary of a region with bends or angles at some boundary points, then better accuracy is obtained if the bends occur only at even numbered vertices $\{z_{2n}\}$, $n \neq 0$.

Conformal welding. The discovery of the slit algorithm by the first author came from considering conformal weldings. (The simpler geodesic algorithm was discovered later.) A decreasing continuous function $h : [0, +\infty) \to (-\infty, 0]$ with $h(0) = 0$ is called a *conformal welding* if there is a conformal map f of $\mathbb{H}$ onto $\mathbb{C} \setminus \gamma$, where γ is a Jordan arc from 0 to ∞ such that $f(x) = f(h(x))$ for $x \in [0, +\infty)$. In other words, the map f pastes the negative and positive real half lines together according to the

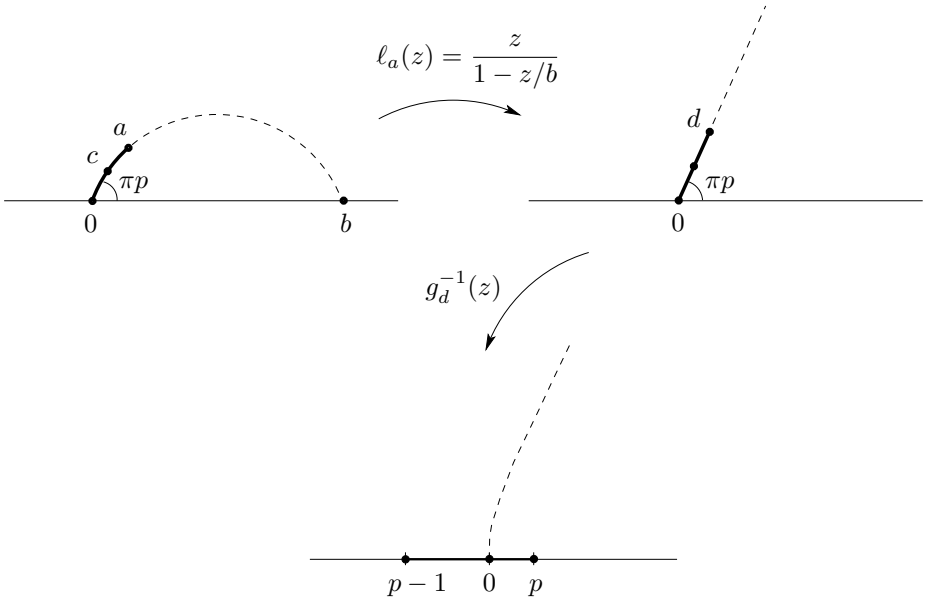


FIG. 5. *The circular slit maps.*

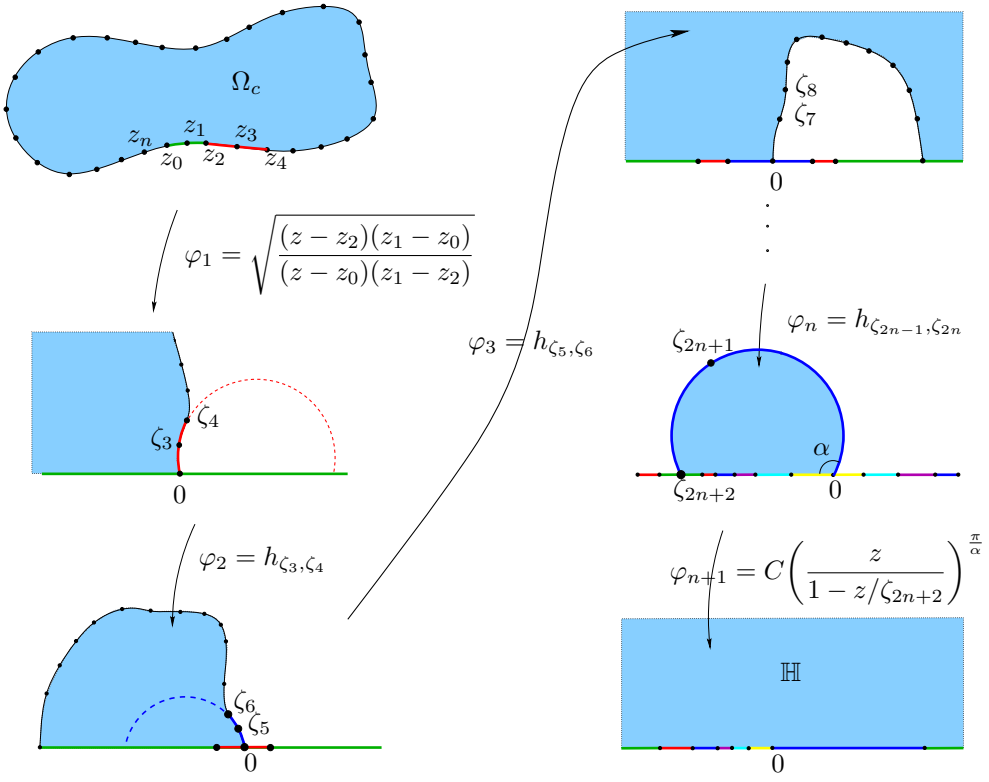


FIG. 6. *The zipper algorithm.*

prescription h to form a curve. One way to approximate a conformal welding is to prescribe the map h at finitely many points and then construct a conformal mapping of $\mathbb{H}$ which identifies the associated intervals.

A related problem, which the first author considered in joint work with L. Carleson, is as follows: Given angles $\alpha_1, \alpha_2, \ldots, \alpha_n$ and $0 < x_1 < x_2 < \cdots < x_n$, find points $y_n < \cdots < y_1 < 0$ so that there is a Schwarz–Christoffel map f of $\mathbb{H}$ onto a region bounded by a polygonal arc tending to ∞ with angles $\alpha_j, 2\pi - \alpha_j$ at the jth vertex $f(x_j) = f(y_j)$. This map welds the intervals $[x_j, x_{j+1}]$ and $[y_{j+1}, y_j]$, $j = 1, \ldots, n$. Unfortunately, at the time the best Schwarz–Christoffel method was only fast enough to do this problem with polygonal curves with up to 20 bends.

The basic maps g_a can be used to compute the conformal maps of weldings. Indeed, suppose $y_1 < 0 < x_1$, let $a = x_1/(x_1 - y_1)$, and apply the map $g_a(z/(x_1 - y_1))$. This map identifies the intervals $[y_1, 0]$ and $[0, x_1]$ by mapping them to the two "sides" of a line segment $L \subset \mathbb{H}$. Composing maps of this form will give a conformal map $\varphi : \mathbb{H} \to \mathbb{C} \setminus \gamma$ such that $\varphi([x_j, x_{j+1}]) = \varphi([y_{j+1}, y_j])$. The final intervals are welded together using the map z^2. The numerical computation of these maps is easily fast enough to compose 10^5 basic maps, thereby giving an approximation to almost any conformal welding. Conversely, given a Jordan arc γ connecting 0 to ∞, the associated welding can be found approximately by using the slit algorithm to approximate the conformal map from $\mathbb{H}$ to the complement of γ.

From this point of view, the slit or the geodesic algorithms find the conformal welding of a curve (approximately). From the point of view of increasing the boundary via a small curve γ_j from z_j to z_{j+1}, the algorithms are discrete solutions of Loewner's differential equation.

The idea of closing up such a region using a map of the form φ_{n+1} was suggested by L. Carleson, for which we thank him.

2. Disc-chains. The geodesic algorithm can be applied to any sequence of data points $z_0, z_1, \ldots, z_n$, unless the points are out of order in the sense that a data point z_j belongs to the geodesic from z_{k-1} to z_k for some $k < j$. In this section we will give a simple condition on the data points $z_0, z_1, \ldots, z_n$ which is sufficient to guarantee that the curve computed by the geodesic algorithm is close to the polygon with vertices $\{z_j\}$.

DEFINITION 2.1. *A* disc-chain $D_0, D_1, \ldots, D_n$ *is a sequence of pairwise disjoint open discs such that* ∂D_j *is tangent to* ∂D_{j+1} *for* $j = 0, \ldots, n-1$. *A* closed disc-chain *is a disc-chain such that* ∂D_n *is tangent to* ∂D_0.

Any closed Jordan polygon P, for example, can be covered by (the closure of) a closed disc-chain with arbitrarily small radii and centers on P (see Figure 7). There are several ways to accomplish this, but one straightforward method is the following: Given $\varepsilon > 0$, find pairwise disjoint discs $\{B_j\}$ centered at each vertex and of radius less than ε so that $B_j \cap P$ is connected for each j. Then

$$P \setminus \bigcup_j B_j = \bigcup L_k,$$

where $\{L_k\}$ are pairwise disjoint closed line segments. Cover each L_k with a disc-chain centered on L_k tangent to the corresponding B_j at the ends, and radius less than half the distance to any other L_i, and less than ε.

Another method for constructing a disc-chain is to draw a Jordan curve using only line segments of length 2^{-n} parallel to the coordinate axes. The circles with radius 2^{-n-1} centered at the endpoints of these segments form a disc-chain. The points of

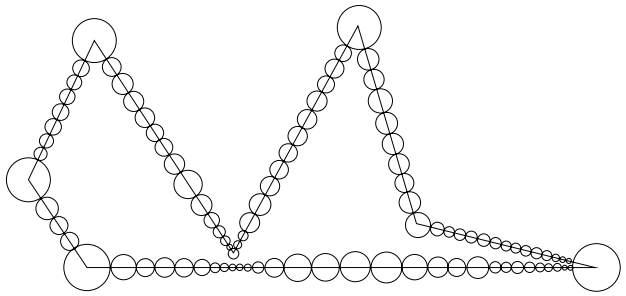


FIG. 7. *Disc-chain covering a polygon.*

tangency are the midpoints of the line segments. Such curves arise from the Whitney decomposition of a simply connected domain, which can be described as follows (see also [GM, p. 21]). If Q is a square, let $2Q$ denote the square with the same center and twice the diameter. Suppose Ω is a simply connected domain contained in the unit square. Divide the unit square into four equal squares.

(a) Discard any square which does not intersect Ω.
(b) If Q is one of the remaining squares for which $2Q \not\subset \Omega$, then divide Q into four equal squares.
(c) Repeat (a), (b), and (c) for the squares obtained in (b).

If this process is repeated ad infinitum, we obtain a decomposition of Ω into squares with the property that for each such square, the distance of the square to the boundary of Ω is comparable to the side length of the square: $2Q \subset \Omega$ and $5Q \cap \partial\Omega \neq \emptyset$. Fix n and $z_0 \in \Omega$ with $\operatorname{dist}(z_0, \partial\Omega) > 2^{-n}$. Let U_n be the union of all squares Q in the Whitney decomposition with side length at least 2^{-n} and let Ω_n be the component of the interior of U_n containing z_0. Then $\partial\Omega_n$ is a polygonal Jordan curve consisting of segments of length 2^{-n}. The discs of radius 2^{-n-1} centered at the endpoints of these segments form a disc-chain and the points of tangency are the midpoints of these segments.

Yet another method for constructing a disc-chain would be to start with a hexagonal grid of tangent discs, all of the same size, and then select a sequence of these discs which form a disc-chain. The boundary circles of a circle packing of a simply connected domain can also be used to form a disc-chain. See, for example, any of the pictures in Stephenson [SK].

If $D_0, D_1, \ldots, D_n$ is a closed disc-chain, set

$$z_j = \partial D_j \cap \partial D_{j+1}$$

for $j = 0, \ldots, n$, where $D_{n+1} \equiv D_0$.

THEOREM 2.2. *If $D_0, D_1, \ldots, D_n$ is a closed disc-chain, then the geodesic algorithm applied to the data $z_0, z_1, \ldots, z_n$ produces a conformal map φ_c^{-1} from the upper-half plane $\mathbb{H}$ to a region bounded by a C^1 and piecewise analytic Jordan curve γ with*

$$\gamma \subset \bigcup_0^n (D_j \cup z_j).$$

Proof. An arc of a circle which is orthogonal to $\mathbb{R}$ is a hyperbolic geodesic in the upper-half plane $\mathbb{H}$. Let γ_j denote the portion of the computed boundary, $\partial\Omega_c$,

between z_j and z_{j+1}. Since hyperbolic geodesics are preserved by conformal maps, γ_j is a hyperbolic geodesic in

$$\mathbb{C}^* \setminus \bigcup_{k=0}^{j-1} \gamma_k.$$

For this reason, we call the algorithm the "geodesic" algorithm.

Using the notation of Figure 2, each map f_a^{-1} is analytic across $\mathbb{R} \setminus \{\pm c\}$, where $f_a^{-1}(\pm c) = 0$, and f_a^{-1} is approximated by a square root near $\pm c$. If f_b^{-1} is another basic map, then f_b^{-1} is analytic and asymptotic to a multiple of z^2 near 0. Thus $f_b^{-1} \circ f_a^{-1}$ preserves angles at $\pm c$. The geodesic γ_j then is an analytic arc which meets γ_{j-1} at z_j with angle π. Thus the computed boundary $\partial\Omega$ is C^1 and piecewise analytic. The first arc γ_0 is a chord of D_0 and hence not tangent to ∂D_0. Since the angle at z_1 between γ_0 and γ_1 is π, γ_1 must enter D_1, and so by Jørgensen's theorem (see Theorem A.1 in the appendix)

$$\gamma_1 \subset D_1,$$

and γ_1 is not tangent to ∂D_1. By induction

$$\gamma_j \subset D_j,$$

$j = 0, 1, \ldots, n$. □

Disc-chains can be used to approximate the boundary of an arbitrary simply connected domain.

LEMMA 2.3. *Suppose that Ω is a bounded simply connected domain. If $\varepsilon > 0$, then there is a disc-chain $D_0, \ldots, D_n$ so that the radius of each D_j is at most ε and $\partial\Omega$ is contained in an ε-neighborhood of $\bigcup D_j$.*

Proof. We may suppose that Ω is contained in the unit square. Then for n sufficiently large, the disc-chain constructed using the Whitney squares with side length at least 2^{-n}, as described above, satisfies the conclusions of Lemma 2.3. □

The *Hausdorff distance* d_H in a metric ρ between two sets A and B is the smallest number d such that every point of A is within ρ-distance d of B, and every point of B is within ρ-distance d of A. The ρ-metrics we will consider in this article are the Euclidean and spherical metrics.

A consequence is the following theorem.

THEOREM 2.4. *If Ω is a bounded simply connected domain, then, for any $\varepsilon > 0$, the geodesic algorithm can be used to find a conformal map f_c of $\mathbb{D}$ onto a Jordan region Ω_c so that*

$$d_H(\partial\Omega, \partial\Omega_c) < \varepsilon, \tag{2.1}$$

where d_H is the Hausdorff distance in the Euclidean metric. If $\partial\Omega$ is a Jordan curve, then we can find f_c so that

$$\sup_{z \in \overline{\mathbb{D}}} |f(z) - f_c(z)| < \varepsilon,$$

where f is a conformal map of $\mathbb{D}$ onto Ω.

Proof. The first statement follows immediately from Theorem 2.2 and Lemma 2.3. To prove the second statement, note that the boundary of the regions constructed with

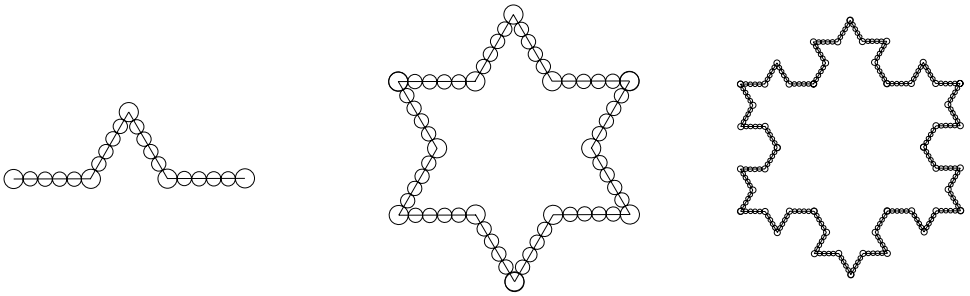


FIG. 8. *Approximating the von Koch snowflake.*

the Whitney decomposition converges to $\partial\Omega$ in the Fréchet sense. By a theorem of Courant [T, p. 383], the mapping functions can be chosen to be uniformly close. □

We note that if Ω is unbounded, Lemma 2.3 and Theorem 2.4 remain true if we use the spherical metric instead of the Euclidean metric to measure the radii of the discs and the distance to $\partial\Omega$.

There are other ways besides using the Whitney decomposition to approximate the boundary of a region by a disc-chain and hence to approximate the mapping function. However, Theorem 2.4 does not give an explicit estimate of the distance between mapping functions in terms of the geometry of the regions. This issue will be explored in greater detail in sections 4 and 5.

The von Koch snowflake is an example of a simply connected Jordan-domain whose boundary has Hausdorff dimension > 1. The standard construction of the von Koch snowflake provides a sequence of polygons which approximate it (see Figure 8). By Theorem 2.4 the mapping functions constructed from these disc-chains converge uniformly to the conformal map to the snowflake.

It is somewhat amusing and perhaps known that a constructive proof of the Riemann mapping theorem (without the use of normal families) then follows. Using linear fractional transformations and a square root map, we may suppose Ω is a bounded simply connected domain. Using the disc-chains associated with increasing levels of the Whitney decomposition, for instance, Ω can be exhausted by an increasing sequence of domains Ω_n for which the geodesic algorithm can be used to compute the conformal map φ_n of Ω_n onto $\mathbb{D}$ with $\varphi_n(z_0) = 0$ and $\varphi_n'(z_0) > 0$. Then by Schwarz's lemma

$$u_n(w) = \log \left| \frac{\varphi_m(w)}{\varphi_n(w)} \right|$$

for $n = m+1, m+2, \ldots$ is an increasing sequence of positive harmonic functions on Ω_m which is bounded above at z_0 by Schwarz's lemma applied to φ_n^{-1}, since Ω is bounded. By Harnack's estimate u_n is bounded on compact subsets of Ω, and by the Herglotz integral formula, $\log \frac{\varphi_m(w)}{\varphi_n(w)}$ converges uniformly on closed discs contained in Ω_m. Thus φ_n converges uniformly on compact subsets of Ω to an analytic function φ. By Hurwitz's theorem φ is one-to-one. Similarly, $\log|\varphi \circ \varphi_m^{-1}(z)/z|$ is an increasing sequence of negative harmonic functions on $\mathbb{D}$ which tend to 0 at $z = 0$. By Harnack again, $|\varphi \circ \varphi_m^{-1}(z)|$ converges to $|z|$ uniformly on compact subsets of $\mathbb{D}$. If $s < 1$, then $|\varphi \circ \varphi_m^{-1}(z)| > s$ for $|z|$ sufficiently close to 1, so by the argument principle, $\{w : |w| < s\} \subset \varphi(\Omega)$, and since s is arbitrary, $\varphi(\Omega) = \mathbb{D}$.

In the geodesic algorithm, we have viewed the maps φ_c and φ_c^{-1} as conformal maps

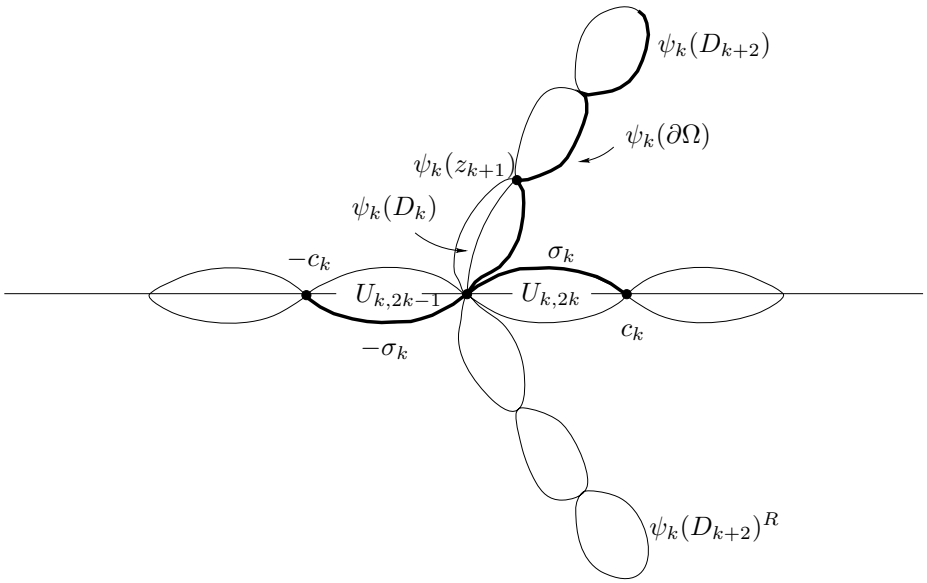


FIG. 9. *Proof of Proposition* 2.5.

between $\mathbb{H}$ and a region Ω_c whose boundary contains the data points. If we are given a region Ω and choose data points $\{z_k\} \in \partial\Omega$ properly, then the next proposition says that the computed maps φ_c and φ_c^{-1} are also conformal maps between the original region Ω and a region "close" to $\mathbb{H}$.

PROPOSITION 2.5. *If $D_0, \ldots, D_n$ is a closed disc-chain with points of tangency $\{z_k\}$, and if Ω is a simply connected domain such that*

$$\partial\Omega \subset \bigcup_{k=0}^{n} \overline{D_k},$$

then the computed map φ_c for the data points $\{z_k\}_0^n$ extends to be conformal on Ω.

We remark that changing the sign of the last map φ_{n+1} in the construction of φ_c gives a conformal map of the complement of the computed region onto $\mathbb{H}$. We choose the sign so that the computed boundary winds once around a given interior point of Ω.

Proof (see Figure 9). Without loss of generality $\Omega \supset \bigcup_{k=0}^{n} D_k$, and hence $\partial\Omega \subset \bigcup \partial D_k$. The basic map f_a in Figure 2 extends by reflection to be a conformal map of $\mathbb{C}^* \setminus (\gamma \cup \gamma^R)$ onto $\mathbb{C}^* \setminus [-c, c]$, where γ^R is the reflection of γ about $\mathbb{R}$. We will first describe the image of $\mathbb{C}^* \setminus \{D_0 \cup \cdots \cup D_n\}$ using these reflected maps. Set

$$\psi_k \equiv \varphi_k \circ \cdots \circ \varphi_1$$

and

$$W_k = \psi_k(\mathbb{C}^* \setminus \{D_0 \cup \cdots \cup D_n\}).$$

Then we claim $\mathbb{C}^* \setminus \{W_k \cup W_k^R\}$ consists of $2(n+1)$ pairwise disjoint simply connected regions:

$$\mathbb{C}^* \setminus \{W_k \cup W_k^R\} = \bigcup_{j=k}^{n} \psi_k(D_j) \cup \psi_k(D_j)^R \cup \bigcup_{j=1}^{2k} U_{k,j},$$

where each region $U_{k,j}$ is symmetric about $\mathbb{R}$ and $\mathbb{R} \subset \bigcup_{j=1}^{2k} \overline{U_{k,j}}$. The case $k = 1$ follows since $\psi_1(\mathbb{C}^* \setminus D_0)$ is bounded by two lines from 0 to ∞. The region $\psi_k(D_k)$ is a subset of $\mathbb{H}$ with 0 and $\psi_k(z_{k+1})$ on its boundary and containing the circular arc from 0 to $\psi_k(z_{k+1})$ which is orthogonal to $\mathbb{R}$. Then

$$\varphi_k\Big(\mathbb{C}^* \setminus (\psi_{k-1}(D_{k-1}) \cup \psi_{k-1}(D_{k-1})^R)\Big)$$

consists of two regions V and $-V = \{-z : z \in V\}$ with 0 and $c_k \in \partial V \cap \mathbb{R}$ and $-c_k \in \partial(-V) \cap \mathbb{R}$. Set $U_{k,2k} = V$, $U_{k,2k-1} = -V$, and $U_{k,p} = \varphi_k(U_{k-1,p})$ for $p \le 2k-2$. The claim now follows by induction.

Finally, we describe the extension of our maps to $\Omega \supset \bigcup_j D_j$. The map φ_k is the composition of a linear fractional transformation τ_k and the map $\sqrt{z^2 + c_k^2}$ (see Figure 2). Note that $\delta_k = \tau_k \circ \psi_{k-1}(\partial\Omega \cap \partial D_{k-1})$ is a curve in $\mathbb{H}$ connecting 0 to ic_k. The map $\sqrt{z^2 + c_k^2}$ is one-to-one and analytic on $\mathbb{C}^* \setminus (\delta_k \cup -\delta_k)$ with image $\mathbb{C}^* \setminus (\sigma_k \cup -\sigma_k)$, where σ_k is a curve connecting 0 to $c_k \in \mathbb{R}$. Thus φ_k extends to be one-to-one and analytic on Ω with image contained in $\mathbb{H} \cup \bigcup_{j=1}^{2k} U_{k,j}$. By induction, ψ_n is one-to-one and analytic on Ω. By direct inspection, the final map φ_{n+1} extends to be one-to-one and analytic, completing the proof of Proposition 2.5. □

As one might surmise from the proof of Proposition 2.5, care must be taken in any numerical implementation to ensure that the proper branch of $\sqrt{z^2 + c^2}$ is chosen at each stage in order to find the analytic extension of the computed map to all of Ω. For this reason, in the numerical implementation of the geodesic algorithm we define our maps using the right-half plane $\mathbb{H}^+ = \{z : \operatorname{Re} z > 0\}$ instead of $\mathbb{H}$.

3. Diamond-chains and pacmen. If we have more control than the disc-chain condition on the behavior of the boundary of a region, then we show in this section that the geodesic algorithm approximates the boundary with better estimates. The computed curve always has a continuously turning tangent direction. The goal in this section is to show that if enough data points are taken on a C^1 Jordan curve, then not only is the computed curve uniformly close, but also the tangent directions are uniformly close to the tangent directions of the given curve. If a subcollection $z_k, z_{k+1}, \ldots, z_p$ of the data points all lie along a line segment, then it is conceivable that the computed curve passing through the data points is oscillating alternately up and down between the data points, and then if z_{p+1} is off the line, it could conceivably cause subsequent oscillations to worsen. Over the long run, the oscillations might then become so large that the curve is no longer a C^1 approximation to the given curve. The key lemma, Lemma 3.5 below, shows that the tangent direction at the end of the geodesic arc actually improves if z_{p+1} is not too far from the line. It is this fixed fractional improvement which does not depend on the number of data points that allows us to iterate the argument.

We will first restrict our attention to domains of the form $\mathbb{C} \setminus \gamma$, where γ is a Jordan arc tending to ∞.

DEFINITION 3.1. *An* ε-diamond $D(a, b)$ *is an open rhombus with opposite vertices* a *and* b *and interior angle* 2ε *at* a *and at* b. *If* $a = \infty$, *then an* ε-diamond $D(\infty, b)$ *is a sector* $\{z : |\arg(z-b) - \theta| < \varepsilon\}$. *An* ε-diamond-chain *is a pairwise disjoint sequence of* ε*-diamonds* $D(z_0, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n)$. *A* closed ε-diamond-chain *is an* ε*-diamond-chain with* $z_n = z_0$.

See Figure 10. Let $B(z, R)$ denote the disc centered at z with radius R.

DEFINITION 3.2. *A* pacman *is a region of the form*

$$P = B(z_0, R) \setminus \{z : |\arg(\overline{\lambda}(z - z_0))| \le \varepsilon\}$$

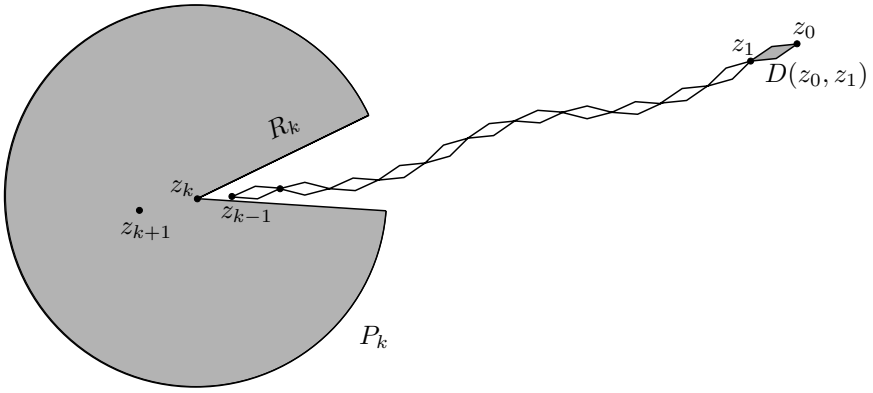


FIG. 10. *A diamond-chain and a pacman.*

for some radius $R < \infty$, center z_0, opening $2\varepsilon > 0$, *and* rotation λ, $|\lambda| = 1$.

Let C_1 be a constant to be chosen later (see Lemma 3.7), and let $z_0 = \infty$.

DEFINITION 3.3. *An ε-diamond-chain $D(\infty, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n)$ satisfies the* ε-pacman condition *if for each $1 \le k \le n-1$ the pacman*

$$P_k = B(z_k, R_k) \setminus \left\{ z : \left| \arg\left(\frac{z - z_k}{z_k - z_{k+1}} \right) \right| \le \varepsilon \right\}$$

with radius $R_k = C_1 |z_{k+1} - z_k| / \varepsilon^2$ satisfies

$$\left(\bigcup_{j=0}^{k-2} D(z_j, z_{j+1}) \right) \cap P_k = \emptyset.$$

The pacman P_k in Definition 3.3 is chosen to be symmetric about the segment between z_k and z_{k+1} with opening 2ε equal to the interior angle 2ε in the diamond-chain. Note that the ε-diamond $D(z_{k-1}, z_k)$ may intersect P_k.

The pacman condition is a quantitative method of estimating how "flat" the polygonal curve through the data points is and prevents the data point z_k from being too close to z_p for larger p (relative to $|z_k - z_{k+1}|$), as might happen if the polygon almost folded back onto itself as in Figure 7. The requirement is more stringent than the disc-chain condition, and it will allow us to control the smoothness of the unit tangent vector on the boundary of the computed region. If we start with a C^1 curve, then we can select data points that satisfy the pacman condition by making the spacing between successive data points smaller in places where the tangent vector is changing rapidly and where the curve almost folds back on itself.

When $z_0 = \infty$, the first map in the geodesic algorithm is replaced by $\varphi_1(z) = \lambda\sqrt{z - z_1}$. The argument of λ can be chosen so that $\varphi_1(z_2)$ is purely imaginary, in which case the boundary of the constructed region contains the half line from z_2 through z_1 and ∞. We will henceforth assume that

$$D(\infty, z_1) = \left\{ z : \left| \arg\left(\frac{z - z_1}{z_1 - z_2} \right) \right| < \varepsilon \right\}.$$

THEOREM 3.4. *There exist universal constants $\varepsilon_0 > 0$ and C_1 such that if an ε-diamond-chain*

$$D(\infty, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n)$$

satisfies the ε-pacman condition with $\varepsilon < \varepsilon_0$, and if

$$\left|\arg\left(\frac{z_{k+1}-z_k}{z_k-z_{k-1}}\right)\right| < \frac{\varepsilon}{10} \tag{3.1}$$

for $k = 2, \dots, n-1$, then the boundary curve γ_c computed by the geodesic algorithm with the data $z_0 = \infty, z_1, \dots, z_n$ satisfies

$$\gamma_c \subset \bigcup_{k=1}^{n} \Big(D(z_{k-1}, z_k) \cup \{z_k\} \Big).$$

Moreover, the argument θ of the tangent to γ_c between z_k and z_{k+1} satisfies

$$|\theta - \arg(z_{k+1} - z_k)| < 3\epsilon.$$

To prove Theorem 3.4, we first give several lemmas. To understand the motivation for the lemmas, perhaps it is helpful to point out that the computed boundary $\partial\Omega$ has a smoothly turning tangent, so that if $\gamma_j \subset D(z_j, z_{j+1})$ were tangent to $\partial D(z_j, z_{j+1})$ at z_{j+1}, then, if z_{j+2} were even slightly off the continuation of the straight line from z_j to z_{j+1} (on one side), γ_{j+1} would not be contained in $D(z_{j+1}, z_{j+2})$. This is why we need the improvement provided by the lemmas.

LEMMA 3.5. *There exists $\varepsilon_0 > 0$ such that if $\varepsilon < \varepsilon_0$, and if Ω is a simply connected region bounded by a Jordan arc $\partial\Omega$ from 0 to ∞ with*

$$\{z : |\arg z| < \pi - \varepsilon\} \subset \Omega,$$

then the conformal map f of $\mathbb{H}^+ = \{z : \operatorname{Re} z > 0\}$ onto Ω normalized so that $f(0) = 0$ and $f(\infty) = \infty$ satisfies

$$|\arg z_0^2 f'(z_0)| < \frac{5\varepsilon}{6}, \tag{3.2}$$

where $z_0 = f^{-1}(1)$.

The circle C_{z_0}, which is orthogonal to the imaginary axis at 0 and passes through z_0, has a tangent vector at z_0 with argument equal to $2 \arg z_0$. The quantity $\arg z_0^2 f'(z_0)$ in (3.2) is the argument of the tangent vector to $f(C_{z_0})$ at $f(z_0)$.

Proof. We may suppose that $|z_0| = 1$. Set

$$g(z) = \log \frac{f(z)}{z^2}.$$

Then $|\operatorname{Im} g(z)| \le \varepsilon$ on $\partial\mathbb{H}^+$ and hence also on $\mathbb{H}^+$, and $|\arg z_0| \le \frac{\varepsilon}{2}$, since $f(z_0) = 1$. Set $\alpha = \frac{\pi}{2\varepsilon}$ and

$$A = e^{\alpha g(z_0)} = z_0^{-2\alpha},$$

$$\varphi(z) = \frac{e^{\alpha z} - A}{e^{\alpha z} + \overline{A}},$$

and

$$\tau(z) = \frac{1+z}{1-z} \operatorname{Re} z_0 + i \operatorname{Im} z_0.$$

Then τ is a conformal map of $\mathbb{D}$ onto $\mathbb{H}^+$ such that $\tau(0) = z_0$ and φ is a conformal map of the strip $\{|\operatorname{Im} z| < \varepsilon\}$ onto $\mathbb{D}$ so that $\varphi(g(z_0)) = 0$. Thus $h = \varphi \circ g \circ \tau$ is analytic on $\mathbb{D}$ and bounded by 1 and $h(0) = 0$, so that by Schwarz's lemma

$$|\varphi'(g(z_0))|\,|g'(z_0)|\,|\tau'(0)| = |h'(0)| \le 1.$$

Consequently

$$\left|\frac{f'(z_0)}{f(z_0)} - \frac{2}{z_0}\right| = |g'(z_0)| \le \frac{2\varepsilon|\operatorname{Re} A|}{\pi \operatorname{Re} z_0} \le \frac{2\varepsilon}{\pi \cos\frac{\varepsilon}{2}},$$

and hence

$$\begin{aligned}|\arg z_0^2 f'(z_0)| &= \left|\arg z_0 + \arg \frac{z_0 f'(z_0)}{f(z_0)}\right| \\ &\le \frac{\varepsilon}{2} + \sin^{-1}\left(\frac{\varepsilon}{\pi \cos\frac{\varepsilon}{2}}\right) \\ &= \left(\frac{1}{2} + \frac{1}{\pi}\right)\varepsilon + \mathrm{O}(\varepsilon^2).\end{aligned}$$

This proves Lemma 3.5 if ε is sufficiently small. □

LEMMA 3.6. *Let Ω satisfy the hypotheses of Lemma 3.5. If $\varepsilon < \varepsilon_0/2$, then the hyperbolic geodesic γ from 0 to 1 for the region Ω lies in the kite*

$$P = \{z : |\arg z| < \varepsilon\} \cap \left\{z : |\arg(1-z)| < \frac{5\varepsilon}{6}\right\},$$

and the tangent vectors to γ have argument less than $\frac{8}{3}\varepsilon$.

Proof. By Jørgensen's theorem, γ is contained in the closed disc through 1 and 0 which has slope $\tan\varepsilon$ at 0. Likewise γ is contained in the reflection of this disc about $\mathbb{R}$, and hence $|\arg z| < \varepsilon$ on γ. This also shows that γ is contained in a kite like P but with angles ε at both 0 and 1. In the proof of Theorem 3.4, however, we need the improvement to $\frac{5\varepsilon}{6}$ of the angle at 1.

By Lemma 3.5, a portion of γ near 1 lies in P. Suppose $w_1 \in \gamma$ with $|\arg w_1| = \delta < \varepsilon$ and then apply Lemma 3.5 to the region $\frac{1}{w_1}\Omega$ with ε replaced by $\varepsilon + \delta$. Then the tangent vector to γ at w_1 has argument θ, where

$$|\theta - \arg w_1| < \frac{5}{6}(\varepsilon + |\arg w_1|). \tag{3.3}$$

Since $|\arg w_1| < \varepsilon$, we have $|\theta| \le \frac{8}{3}\varepsilon$. Moreover, (3.3) also implies $\theta < \frac{5}{6}\varepsilon$ when $\arg w_1 \le 0$ and $\theta > -\frac{5}{6}\varepsilon$ when $\arg w_1 \ge 0$. But if w_1 is the last point on $\gamma \cap \partial P$ before reaching 1, this is impossible. Thus $\gamma \subset P$, proving the lemma. □

The next lemma improves Lemma 3.5 by requiring only that the portion of $\partial\Omega$ in a large disc lies inside a small sector.

LEMMA 3.7. *There is a constant C_1 so that if $\varepsilon < \varepsilon_0/2$ and if $\partial\Omega$ is a Jordan arc such that $0 \in \partial\Omega$, $\partial\Omega \cap \{|z| > C_1/\varepsilon^2\} \ne \emptyset$, and*

$$P_\varepsilon = \left\{z : |\arg z| < \pi - \varepsilon \text{ and } |z| \le \frac{C_1}{\varepsilon^2}\right\} \subset \Omega,$$

then the conformal map $f : \mathbb{H}^+ \longrightarrow \Omega$ with $f(0) = 0$ and $|f(\infty)| > \frac{C_1}{\varepsilon^2}$ satisfies

$$|\arg z_0^2 f'(z_0)| < \frac{9\varepsilon}{10}, \tag{3.4}$$

where $z_0 = f^{-1}(1)$.

Proof. Set $R = \frac{C_1}{\varepsilon^2}$ and $B_R = B(0,R) = \{|z| < R\}$. Let U_R be the component of $\Omega \cap B_R$ containing the point 1. Then $f^{-1}(U_R) \subset \mathbb{H}^+$ is bounded by a set $F \subset i\mathbb{R}$ and curves $\sigma_j \subset \mathbb{H}^+$ on which $|f| = R$. Since $0 \in \partial f^{-1}(U_R)$ and $f(\infty) \notin B_R$, exactly one of the curves (call it σ_1) will connect the positive imaginary axis to the negative imaginary axis. The function $u(z) = \arg f(z) - \arg z^2$ is harmonic on the simply connected region $f^{-1}(U_R)$ with $|u| \le 2\pi + \varepsilon$. Then $\partial\Omega \cap B_R$ contains a subarc δ connecting 0 to ∂B_R and $|u| < \varepsilon$ on $f^{-1}(\delta)$. It is possible that B_R contains other subarcs of $\partial\Omega$, none of which intersect P_ε. We may suppose that $P_\varepsilon \cap \partial B_R \subset f(\sigma_1)$, for if $P_\varepsilon \cap \partial B_R \subset f(\sigma_j)$, $j \neq 1$, then σ_j separates a point $z_1 \in i\mathbb{R}$ from $f^{-1}(U_R)$. Then

$$g(\zeta) = f\left(\frac{\zeta}{1 + \zeta/z_1}\right)$$

satisfies the hypotheses of the lemma and $P_\varepsilon \cap \partial B_R$ is a subset of the image of the corresponding curve in $\mathbb{H}^+$ connecting the positive and negative imaginary axes. Moreover, a direct computation shows that

$$\zeta_0^2 g'(\zeta_0) = z_0^2 f'(z_0),$$

where $\zeta_0 = z_0/(1 - z_0/z_1)$.

We conclude $|u(z)| < \varepsilon$ at the endpoints of each σ_j because $P_\varepsilon \subset U_R$. Since u is continuous on the closure of $f^{-1}(U_R)$, and $|\arg f| > \pi - \varepsilon$ on $\partial f^{-1}(U_R) \cap i\mathbb{R}$, and $\arg z^2$ is the same at each endpoint of σ_j, $j > 1$, we conclude that $|u| < \varepsilon$ on $\partial f^{-1}(U_R) \cap i\mathbb{R}$. By the maximum principle

$$|u(z)| \le \varepsilon + (2\pi + \varepsilon)\omega(f(z), \partial B_r, U_R)$$

for $z \in f^{-1}(U_R)$, where $\omega(z, E, V)$ is the harmonic measure at z for $E \cap \overline{V}$ in $V \setminus E$. By Beurling's projection theorem [GM, p. 105] and a direct computation (see [GM, Corollary III.9.3])

$$\omega(1, \partial B_R, \Omega) \le \omega(1, \partial B_R, B_R \setminus [-R, 0]) = \frac{4}{\pi}\tan^{-1}\left(\frac{1}{R^{\frac{1}{2}}}\right). \tag{3.5}$$

Evaluating at $z_0 = f^{-1}(1)$, we obtain

$$|u(z_0)| = |-\arg z_0^2| \le \varepsilon + (2\pi + \varepsilon)\frac{4}{\pi}\tan^{-1}\left(\frac{\varepsilon}{C_1^{\frac{1}{2}}}\right) < \frac{11\varepsilon}{10}$$

for C_1 sufficiently large. Thus

$$|\arg z_0| \le \frac{11\varepsilon}{20}. \tag{3.6}$$

Next we show that there is a large half disc contained in $f^{-1}(\Omega \cap B_R)$. We may suppose that $|z_0| = 1$. Set

$$S = \inf\{|w - i\operatorname{Im} z_0| : w \in \mathbb{H}^+ \text{ and } f(w) \in \partial B_R\}.$$

Using the map

$$\frac{z - i\operatorname{Im} z_0 - S}{z - i\operatorname{Im} z_0 + S}$$

of $\mathbb{H}^+$ onto $\mathbb{D}$ and Beurling's projection theorem again,

$$\omega(z_0, f^{-1}(\partial B_R), \mathbb{H}^+) \geq \omega(z_0, [S, \infty) + i \operatorname{Im} z_0, \mathbb{H}^+).$$

Then by (3.5), (3.6), and an explicit computation

$$\frac{4}{\pi} \tan^{-1}\left(\frac{\varepsilon}{C_1^{\frac{1}{2}}}\right) \geq \frac{2}{\pi} \tan^{-1}\left(\frac{\operatorname{Re} z_0}{\sqrt{S^2 - \operatorname{Re} z_0^2}}\right).$$

For ε sufficiently small, this implies

$$B\left(0, \frac{C_1^{\frac{1}{2}}}{2\varepsilon}\right) \cap \mathbb{H}^+ \subset f^{-1}\left(\Omega \cap B\left(0, \frac{C_1}{\varepsilon^2}\right)\right).$$

Now follow the proof of Lemma 3.5 replacing τ with a conformal map of $\mathbb{D}$ onto $\mathbb{H}^+ \cap \{|z| < \frac{C_1^{1/2}}{2\varepsilon}\}$ such that $\tau(0) = z_0$. Then $\tau'(0) = 2 \operatorname{Re} z_0 + O(\frac{\varepsilon}{C_1^{1/2}})$, and for C_1 sufficiently large, (3.4) holds. □

Following the proof of Lemma 3.6 (replacing 5/6 by 9/10), the next corollary is obtained.

COROLLARY 3.8. *Suppose $\partial\Omega$ is a Jordan arc such that $0 \in \partial\Omega$, $\partial\Omega \cap \{|z| > C_1/\varepsilon^2\} \neq \emptyset$, and*

$$\left\{z : |\arg z| < \pi - \varepsilon \text{ and } |z| \leq \frac{C_1}{\varepsilon^2}\right\} \subset \Omega.$$

If $\varepsilon < \varepsilon_0/2$, then the hyperbolic geodesic γ from 0 to 1 for the region Ω lies in the kite

$$P = \{z : |\arg z| \leq \varepsilon\} \cap \left\{z : |\arg(1 - z)| \leq \frac{9\varepsilon}{10}\right\}. \tag{3.7}$$

Moreover, the tangent vectors to this geodesic have argument at most 3ε.

Proof of Theorem 3.4. Use the constant C_1 from Lemma 3.7 in Definition 3.3. As in Theorem 2.2, let γ_j denote the portion of the computed boundary $\partial\Omega_c$ between z_j and z_{j+1}. By construction $\gamma_0 \cup \gamma_1$ is a half line through $z_0 = \infty$, z_1, and z_2. Make the inductive hypotheses that

$$\bigcup_{j=0}^{k-1} \gamma_j \subset \bigcup_{j=0}^{k-1} D(z_j, z_{j+1}) \tag{3.8}$$

and

$$\gamma_{k-1} \cap P_k = \emptyset. \tag{3.9}$$

Since the ε-diamond-chain $D(\infty, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n)$ satisfies the ε-pacman condition, (3.8) and (3.9) show that the hypotheses of Corollary 3.8 hold for the curve $\gamma = \bigcup_0^{k-1} \gamma_j$ and hence $\gamma_k \subset D(z_k, z_{k+1})$. Also, by Corollary 3.8 and (3.1),

$$\gamma_k \cap P_{k+1} = \emptyset.$$

By induction, the theorem follows. □

If the hypotheses of Theorem 3.4 hold, then the proof of Proposition 2.5 gives the following corollary.

COROLLARY 3.9. *If Ω and the diamond-chain $D(z_k, z_{k+1})$ satisfy the hypotheses of Theorem 3.4, then the conformal map φ_c computed in the geodesic algorithm extends to be conformal on $\Omega \cup \bigcup_{k=0}^{n} D(z_k, z_{k+1})$.*

The next theorem says that for a region Ω bounded by a C^1 curve, the geodesic algorithm with data points $z_0, z_1, \ldots, z_n$ produces a region Ω_c whose boundary is a C^1 approximation to $\partial\Omega$.

THEOREM 3.10. *Suppose Ω is a Jordan region bounded by a C^1 curve $\partial\Omega$. Then there exists $\delta_0 > 0$ depending on $\partial\Omega$ so that for $\delta < \delta_0$,*

$$\partial\Omega \subset \bigcup_k \big(D(z_k, z_{k+1}) \cup \{z_k\}\big),$$

where $D = \bigcup D(z_k, z_{k+1})$ is a δ-diamond-chain, and so that $\partial\Omega_c$, the boundary of the region computed by the geodesic algorithm, is contained in $D \cup (\bigcup_k \{z_k\})$. Moreover, if $\zeta \in \partial\Omega_c$ and if $\alpha \in \partial\Omega$ with $|\zeta - \alpha| < \delta$, then

$$|\eta_\zeta - \eta_\alpha| < 6\delta, \tag{3.10}$$

where η_ζ and η_α are the unit tangent vectors to $\partial\Omega$ and $\partial\Omega_c$ at ζ and α, respectively.

Proof. There were two reasons for requiring that $z_0 = \infty$ in Theorem 3.4. The first reason was to ensure that

$$\left(\bigcup_0^{k-1} \gamma_j \right) \cap (\mathbb{C} \setminus B(z_k, R_k)) \neq \emptyset, \tag{3.11}$$

as needed for Lemma 3.7. The second reason is the difficulty in closing the curve, since Lemma 3.7 does not apply. The difficulty is that a pacman centered at z_n will contain z_0 if z_0 is too close to z_n. Since $\partial\Omega \in C^1$, we may suppose that the δ-diamond-chain $D(z_0, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n)$ satisfies the pacman condition. Note that this requires z_n to be much closer to z_{n-1} than to z_0. Since $\partial\Omega \in C^1$, if $|z_n - z_0|$ is sufficiently small, we can find two discs

$$\Delta_p \subset \mathbb{C} \setminus \bigcup_0^{n-1} D(z_k, z_{k+1})$$

for $p = 1, 2$ with

$$\{z_0, z_n\} = \partial\Delta_1 \cap \partial\Delta_2 \quad \text{and} \quad \Delta_1 \cap \Delta_2 \subset D(z_n, z_0),$$

where $D(z_n, z_0)$ is a δ-diamond. By Jørgensen's theorem, as in the proof of Theorem 2.2, the geodesic γ_n between z_n and z_0 is contained in $\Delta_1 \cap \Delta_2$. Then by the proof of Theorem 3.4, $\partial\Omega_c$ is contained in the δ-diamond-chain. The statement about tangent vectors now follows from Corollary 3.8. ∎

We say that $\{z_k\}$ are *locally evenly spaced* if

$$\frac{1}{D} \leq \left| \frac{z_k - z_{k-1}}{z_k - z_{k+1}} \right| \leq D \tag{3.12}$$

for some constant $D < \infty$. Note that the spacing between points can still grow or decay geometrically. We define the *mesh size* μ of the data points $\{z_j\}$ to be

$$\mu(\{z_j\}) = \sup_k |z_k - z_{k+1}|.$$

We say that a Jordan curve Γ in the extended plane $\mathbb{C}^*$ is a *K-quasicircle* if for some linear fractional transformation τ

$$\frac{|w_1 - w| + |w - w_2|}{|w_1 - w_2|} \le K \tag{3.13}$$

for all $w_1, w_2 \in \tau(\Gamma)$ and for all w on the subarc of $\tau(\Gamma)$ with smaller diameter. Thus circles and lines are 1-quasicircles. Quasicircles look very flat on all scales if K is close to 1, but for any $K > 1$ they can contain a dense set of spirals. See, for example, Figure 8.

If Γ satisfies (3.13) with $K = 1+\delta$ and small δ and if $\{z_k\} \subset \tau(\Gamma)$ is locally evenly spaced, then

$$\left|\arg\left(\frac{z_k - z_{k-1}}{z_{k+1} - z_k}\right)\right| \le C\delta^{\frac{1}{2}} \tag{3.14}$$

for some constant C, depending on D. The referee suggested that a proof of this fact might help the reader. Note that (3.12), (3.13), and (3.14) are invariant under translations and dilations, so that we may assume $z_{k-1} = -1$ and $z_k = 0$ and write $z_{k+1} = \zeta$. Then (3.13), with $w_1 = -1$, $w = 0$, and $w_2 = \zeta$, shows that

$$1 + |\zeta| \le (1 + \delta)|1 + \zeta|.$$

Writing $\zeta = re^{i\theta}$ and squaring yield

$$1 - \cos\theta \le (2\delta + \delta^2)\frac{(1 + r)^2}{2r}.$$

By (3.12) $D^{-1} \le |\zeta| = r \le D$ so that

$$\frac{\theta^2}{2} \le (2\delta + \delta^2)\frac{(1 + D)^2}{2D},$$

and

$$\left|\arg\left(\frac{z_k - z_{k-1}}{z_{k+1} - z_k}\right)\right| = |\theta| \le C\delta^{\frac{1}{2}}.$$

THEOREM 3.11. *There is a constant $K_0 > 1$ so that if Γ is a K-quasicircle with $K = 1 + \delta < K_0$ and if $\{z_k\}$ are locally evenly spaced on Γ, then the geodesic algorithm finds a conformal map of $\mathbb{H}$ onto a region Ω_c bounded by a $C(K)$-quasicircle containing the data points $\{z_k\}$, where $C(K)$ is a constant depending only on K.*

We can choose $C(K)$ so that $C(K) \to 1$ as $K \to 1$. Moreover, given $\eta > 0$, if the mesh size $\mu(\{z_k\})$ is sufficiently small, then

$$d_H(\Gamma, \partial\Omega_c) < \eta,$$

where d_H is the Hausdorff distance in the spherical metric.

Proof. We may suppose that Γ satisfies (3.13) with $K = 1 + \delta$ and δ small. Note that $\infty \in \Gamma$. If $\{z_k\}_1^n$ are locally evenly spaced points on $\partial\Omega$, with $\mu = \max|z_k - z_{k-1}|$ sufficiently small, then (3.14) holds and $D(\infty, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n), D(z_n, \infty)$ is a $C\delta^{\frac{1}{2}}$-diamond-chain, where the main axis of the cone $D(\infty, z_1)$ is in the direction $z_1 - z_2$ and the main axis of $D(z_n, \infty)$ is in the direction $z_n - z_{n-1}$. Moreover, $D(\infty, z_1), D(z_1, z_2), \ldots, D(z_{n-1}, z_n)$ satisfies the ε-pacman condition if

$$\varepsilon \ge C\delta^{\frac{1}{4}}$$

for some universal constant C. Now apply Theorem 3.4 to obtain $\gamma_j \subset D(z_{j-1}, z_j)$, $j = 1, \ldots, n-1$. By an argument similar to the proof of Theorem 3.10, we can also find a geodesic arc for $\mathbb{C} \setminus (\bigcup_0^{n-1} \gamma_j)$ from z_n to ∞ contained in $D(z_n, \infty)$. Then the computed curve will be a $C(K)$-quasicircle.

Note that if w_1, w, and w_2 are data points, then by assumption (3.13) holds with $K = 1 + \delta$. By Corollary 3.8, the tangent directions to the computed curve change by no more than $C\delta^{\frac{1}{4}}$ between the data points, and hence the computed curve is a $(1 + C'\delta^{\frac{1}{4}})$-quasicircle. □

As noted before, the boundary of the region computed with the geodesic algorithm, $\partial\Omega_c$, is a C^1 curve. We end this section by proving that $\partial\Omega_c$ is slightly better than C^1. If $0 < \alpha < 1$, we say that a curve Γ *belongs to* $C^{1+\alpha}$ if arc length parameterization $\gamma(s)$ of Γ satisfies

$$|\gamma'(s_1) - \gamma'(s_2)| \leq C|s_1 - s_2|^\alpha$$

for some constant $C < \infty$.

We say that a conformal map f defined on a region Ω *belongs to* $C^{1+\alpha}(\overline{\Omega})$, $0 < \alpha < 1$, provided f and f' extend to be continuous on $\overline{\Omega}$ and there is a constant C so that

$$|f'(z_1) - f'(z_2)| \leq C|z_1 - z_2|^\alpha$$

for all z_1, z_2 in $\overline{\Omega}$.

PROPOSITION 3.12. *If the bounded Jordan region Ω_c is the image of the unit disc by the geodesic algorithm, then*

$$\partial\Omega_c \in C^{3/2},$$

and $\partial\Omega_c \notin C^{1+\alpha}$ for $\alpha > 1/2$, unless Ω_c is a circle or a line. Moreover, $\varphi \in C^{3/2}(\overline{\Omega_c})$ and $\varphi^{-1} \in C^{3/2}(\overline{\mathbb{D}})$.

Proof. To prove the first statement, it is enough to show that if γ is an arc of a circle in $\mathbb{H}$ which meets $\mathbb{R}$ orthogonally at 0 (constructed by application of one of the maps f_a^{-1} as in Figure 2), then the curve σ which is the image of $[-1, 1] \cup \gamma$ by the map $S(z) = \sqrt{z^2 - d^2}$ is $C^{\frac{3}{2}}$ (and no better class) in a neighborhood of $S(0) = id$. Indeed, subsequent maps in the composition φ^{-1} are conformal in $\mathbb{H}$ and hence preserve smoothness. For $d > 0$, the function

$$\psi(z) = \sqrt{\left(\frac{\sqrt{z^2 - c^2}}{1 + \sqrt{z^2 - c^2}/b}\right)^2 - d^2} = id + \frac{i}{2d}(z^2 - c^2) - \frac{i}{bd}(z^2 - c^2)^{\frac{3}{2}} + \mathrm{O}((z^2 - c^2)^2)$$

for some choice of $b \in \mathbb{R}$ and $c > 0$ is a conformal map of the upper-half plane onto a region whose complement contains the curve σ. Clearly $\psi \in C^{\frac{3}{2}}$ near $z = \pm c$, and so by a theorem of Kellogg (see [GM, p. 62]), $\sigma \in C^{\frac{3}{2}}$. The same theorem implies σ is not in C^α for $\alpha > \frac{3}{2}$ unless $1/b = 0$. This argument also shows that $\varphi_c \in C^{3/2}(\overline{\Omega})$. To prove $\varphi_c^{-1} \in C^{\frac{3}{2}}(\overline{\mathbb{D}})$, apply the same ideas above to the inverse maps. Alternatively, this last fact can be proved by following the proof of Lemma II.4.4 in [GM]. □

4. Estimates for conformal maps onto nearby domains. We begin this section with a discussion of the following question. Consider two simply connected planar domains Ω_j with $0 \in \Omega_j$ and conformal maps $\varphi_j : \Omega_j \to \mathbb{D}$ fixing 0, suitably normalized (for instance, positive derivative at 0). If Ω_1 and Ω_2 are "close," what can

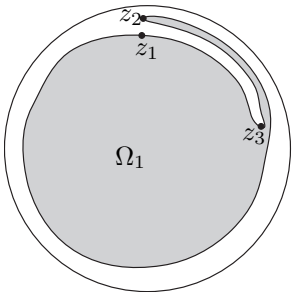


FIG. 11. *Small Hausdorff distance.*

be said about $|\varphi_1 - \varphi_2|$ on $\Omega_1 \cap \Omega_2$, or about $|\varphi_1^{-1} - \varphi_2^{-1}|$ on $\mathbb{D}$? The article [W] gives an overview and numerous results in this direction. How should "closeness" of the two domains be measured? Simple examples show that the Hausdorff distance in the Euclidean or spherical metric between the boundaries does not give uniform estimates for either $||\varphi_1 - \varphi_2||_\infty$ or $||\varphi_1^{-1} - \varphi_2^{-1}||_\infty$. For example, in Figure 11, Ω_1 contains a disc of radius $1 - \delta$, where δ is small, and hence for $\Omega_2 = \mathbb{D}$, $d_H(\Omega_1, \Omega_2) \leq \delta$, but $|\varphi_1(z_1) - \varphi_1(z_2)|$ is large and $|\varphi_1(z_2) - \varphi_1(z_3)|$ is small so that neither $||\varphi_1(z) - z||_\infty$ nor $||\varphi_1^{-1}(z) - z||_\infty$ is small.

Mainly for ease of notation, we will assume throughout this section that the Ω_j are Jordan-domains, and denote by $\gamma_j : \partial\mathbb{D} \to \partial\Omega_j$ an orientation preserving parameterization. Even the more refined distance

$$\inf_\alpha ||\gamma_1 - \gamma_2 \circ \alpha||_\infty,$$

where the infimum is over all homeomorphisms α of $\partial\mathbb{D}$, does not control $||\varphi_1^{-1} - \varphi_2^{-1}||_\infty$ or $||\varphi_1 - \varphi_2||_\infty$. For example, let Ω_2 be a small rotation of the region Ω_1 in Figure 11. What is needed is some control on the "roughness" of the boundary. Following [W], for a simply connected domain Ω we define

$$\eta(\delta) = \eta_\Omega(\delta) = \sup_C \operatorname{diam} T(C),$$

where the supremum is over all crosscuts of Ω with $\operatorname{diam} C \leq \delta$, and where $T(C)$ is the component of $\Omega \setminus C$ that does not contain 0. Notice that $\eta(\delta) \to 0$ as $\delta \to 0$ is equivalent to saying that $\partial\Omega$ is locally connected, and the condition $\eta(\delta) \leq K\delta$ for some constant K is equivalent to saying that Ω is a John-domain (see, e.g., [P, Chapter 5]). It is not difficult to control the modulus of continuity of $\varphi^{-1} : \mathbb{D} \to \Omega$ in terms of η; see [W, Theorem I]. This can be used to estimate $||\varphi_1^{-1} - \varphi_2^{-1}||_\infty$ in terms of the Hausdorff distance between the boundaries, for example.

THEOREM 4.1 (Warschawski [W, Theorem VI]). *If Ω_1 and Ω_2 are John-domains, $\eta_j(\delta) \leq \kappa\delta$ for $j = 1, 2$, and if $d_H(\partial\Omega_1, \partial\Omega_2) \leq \epsilon$, then*

$$||\varphi_1^{-1} - \varphi_2^{-1}||_\infty \leq C\epsilon^\alpha$$

with $\alpha = \alpha(\kappa)$ and $C = C(\kappa, \operatorname{dist}(0, \partial\Omega_1 \cup \partial\Omega_2))$.

In fact, Warschawski proves that every $\alpha < 2/(\pi^2\kappa^2)$ will work (with $C = C(\alpha)$). Using the Hölder continuity of quasiconformal maps, his proof can easily be modified to give the following better estimate if Ω_1 and Ω_2 are K-quasidiscs with K near 1. A K-quasidisc is a Jordan region bounded by a K-quasicircle.

COROLLARY 4.2. *If Ω_1 and Ω_2 are K-quasidiscs, and if $d_H(\partial\Omega_1, \partial\Omega_2) \le \epsilon$, then*

$$||\varphi_1^{-1} - \varphi_2^{-1}||_\infty \le C\epsilon^\alpha$$

with $\alpha = \alpha(K) \to 1$ as $K \to 1$.

As for estimates of $||\varphi_1 - \varphi_2||_\infty$, Warschawski shows [W, Theorem VII] that

$$\sup_{\Omega_1} |\varphi_1 - \varphi_2| \le C\epsilon^{1/2} \log \frac{2}{\epsilon}$$

if $\Omega_1 \subset \Omega_2$ and if Ω_1 is a John-domain, with C depending on κ and on dist$(0, \partial\Omega_1 \cup \partial\Omega_2)$. However, his result does not apply without the assumption of inclusion $\Omega_1 \subset \Omega_2$. To treat the general case the trick of controlling $|\varphi_1 - \varphi_2|$ by passing to the conformal map φ of the component Ω of $\Omega_1 \cap \Omega_2$ containing 0 (which now is included in Ω_j) does not seem to work, as the geometry of Ω cannot be controlled. Nevertheless, for the case of disc-chain domains, the above estimate can be proved, even without any further assumption on the geometry on the disk-chain.

THEOREM 4.3. *Let $D_1, D_2, \dots, D_n$ be a closed ε-disc-chain surrounding 0. Suppose $\partial\Omega_j \subset \bigcup_k \overline{D_k}$ for $j = 1, 2$, and let $\varphi_j : \Omega_j \to \mathbb{D}$ be conformal maps with $\varphi_1(0) = \varphi_2(0) = 0$ and $\varphi_1(p) = \varphi_2(p)$ for a point $p \in \partial\Omega_1 \cap \partial\Omega_2$. Then*

$$\sup_{w \in \Omega_1 \cap \Omega_2} |\varphi_1(w) - \varphi_2(w)| \le C\epsilon^{1/2} \log \frac{1}{\epsilon},$$

where C depends on dist$(0, \bigcup_k D_k)$ *only.*

In case we have control on the geometry of the domains, we have the following counterpart to Corollary 4.2.

THEOREM 4.4. *If Ω_1 and Ω_2 are K-quasidiscs, if $d_H(\partial\Omega_1, \partial\Omega_2) \le \epsilon$, and if $\varphi_1(p_1) = \varphi_2(p_2)$ for a pair of points $p_j \in \partial\Omega_j$ with $|p_1 - p_2| \le \epsilon$, then*

$$\sup_{w \in \Omega} |\varphi_1(w) - \varphi_2(w)| \le C\epsilon^\alpha$$

with $\alpha = \alpha(K) \to 1$ as $K \to 1$, where Ω is the component of $\Omega_1 \cap \Omega_2$ containing 0.

The proofs of both theorems rely on the following harmonic measure estimate, which is an immediate consequence of a theorem of Marchenko [M] (see [W, section 3] for the statement and a proof). To keep this paper self-contained, we include a simple proof, shown to us by John Garnett, for which we thank him.

LEMMA 4.5. *Let $0 < \theta < \pi$, $0 < \epsilon < 1/2$, and set $D = \mathbb{D} \setminus \{re^{it} : -\theta \le t \le \theta, 1 - \epsilon \le r < 1\}$, $A = \partial D \setminus \partial\mathbb{D}$. Then*

$$\omega(0, A, D) \le \frac{\theta}{\pi} + C\epsilon \log \frac{1}{\epsilon}$$

for some universal constant C.

Proof. Set $\omega(z) = \omega(z, A, D)$ for $z \in D$. By the mean value property, it is enough to show that

$$\omega(z) \le C\frac{\epsilon}{t - \theta}$$

for $z = (1 - \epsilon)e^{it}$ and $\theta + \epsilon \le t \le \pi$. To this end, set $I = \{e^{i\tau} : -\theta \le \tau \le \theta\}$ and consider the circular arc $\{\zeta : \omega(\zeta, I, \mathbb{D}) = \frac{1}{3}\}$. If $\epsilon < \epsilon_0$ for some universal ϵ_0 (for $\epsilon \ge \epsilon_0$ there is nothing to prove), then A is disjoint from this arc and it follows that

$\omega(\zeta, I, \mathbb{D}) \geq \frac{1}{3}$ on A. The maximum principle implies $\omega(\zeta) \leq 3\omega(\zeta, I, \mathbb{D})$ on D. Now the desired inequality follows from

$$\omega((1-\epsilon)e^{it}, I, \mathbb{D}) = \frac{1}{2\pi}\int_{-\theta}^{\theta} \frac{1-(1-\epsilon)^2}{|(1-\epsilon)e^{it}-e^{i\tau}|^2} d\tau \leq C\epsilon \int_{-\theta}^{\theta} \frac{1}{(t-\tau)^2} d\tau < C\frac{\epsilon}{t-\theta}. \qquad \square$$

Proof of Theorem 4.3. We may assume that $\varphi_j(p) = 1$. We will first assume that p is one of the points $D_k \cap D_{k+1}$. Denote by Ω the largest simply connected domain $\subset \mathbb{C}$ containing 0 whose boundary is contained in $\bigcup_k D_k$ (thus $\overline{\Omega}$ is the union of $\bigcup_k D_k$ and the bounded component of $\mathbb{C} \setminus \bigcup_k D_k$), and φ is the conformal map from Ω to $\mathbb{D}$ with $\varphi(0) = 0$ and $\varphi(p) = 1$. First, let $z \in \partial\Omega_1 \cap \partial\Omega$. Denote by B, respectively, B_1, the arc of $\partial\Omega$ ($\partial\Omega_1$) from p to z. By the Beurling projection theorem (or the distortion theorem), every $\varphi(D_j)$ has diameter $\leq C\sqrt{\epsilon}$. Therefore, $\varphi(B_1)$ is an arc in $\overline{\mathbb{D}}$, with the same endpoints as $\varphi(B)$, that is contained in $S = \{re^{it} : 1 - C\sqrt{\epsilon} \leq r < 1, -C\sqrt{\epsilon} < t < \arg\varphi(z) + C\sqrt{\epsilon}\}$. Denote $A = \partial S$. By Lemma 4.5,

$$\omega(0, B_1, \Omega_1) \leq \omega(0, B_1, \Omega \setminus B_1) \leq \omega(0, A, \mathbb{D} \setminus A) \leq \frac{1}{2\pi}\arg\varphi(z) + 2C\sqrt{\epsilon} + C\sqrt{\epsilon}\log\frac{1}{\sqrt{\epsilon}}$$

and we obtain

$$\arg\varphi_1(z) = 2\pi\omega(0, B_1, \Omega_1) \leq \arg\varphi(z) + C\epsilon^{1/2}\log\frac{1}{\epsilon}.$$

The same argument, applied to the other arc from p to z, gives the opposite inequality, and together it follows that

$$|\varphi(z) - \varphi_1(z)| \leq C\epsilon^{1/2}\log\frac{1}{\epsilon}.$$

Now let $z \in \partial\Omega_1$ be arbitrary. If z' is a point of $\partial\Omega_1 \cap \partial\Omega$ in the same disc D_j as z, then we have

$$|\varphi(z)-\varphi_1(z)| \leq |\varphi(z)-\varphi(z')|+|\varphi(z')-\varphi_1(z')|+|\varphi_1(z)-\varphi_1(z')| \leq 2C\sqrt{\epsilon}+C\epsilon^{1/2}\log\frac{1}{\epsilon}.$$

The maximum principle yields $|\varphi - \varphi_1| \leq C\epsilon^{1/2}\log\frac{1}{\epsilon}$ on Ω_1. The same argument applies to $|\varphi - \varphi_2|$, and the theorem follows from the triangle inequality.

If $p \in \partial\Omega_1 \cap \partial\Omega_2$ is arbitrary, let p' be one of the points $D_k \cap D_{k+1}$ in the same disc D_j as p. Then the above estimate, applied to a rotation of φ_1, φ_2, and p', gives $|\varphi_2(p')/\varphi_1(p')\varphi_1 - \varphi_2| \leq C\epsilon^{1/2}\log\frac{2}{\epsilon}$, and the theorem follows from $|\varphi_j(p) - \varphi_j(p')| \leq C\sqrt{\epsilon}$. $\square$

The following lemma is another easy consequence of the aforementioned theorem of Marchenko [M] (see [W, section 3]).

LEMMA 4.6. *Let $H \subset \mathbb{D}$ be a K-quasidisc with $0 \in H$ such that $\partial H \subset \{1 - \epsilon < |z| < 1\}$, and let h be a conformal map from $\mathbb{D}$ to H with $h(0) = 0$ and $|h(p) - p| < \epsilon$ for some $p \in \partial\mathbb{D}$. Then*

$$|h(z) - z| \leq C\epsilon\log\frac{1}{\epsilon},$$

where C depends on K only.

Proof. We may assume that $p = 1$. Let $z = e^{i\tau}$ and consider the arc $A = \{h(e^{it}) : 0 \leq t \leq \tau\} \subset \partial H$ of harmonic measure $\tau/2\pi$. For a suitable constant C, depending

on K, we have that $D = \mathbb{D} \setminus \{re^{it} : -C\epsilon \le t \le \arg h(z) + C\epsilon,\ 1-\epsilon \le r < 1\}$ contains A. By the maximum principle and Lemma 4.5,

$$\tau/2\pi = \omega(0, A, H) \le \omega(0, \partial D \cap \mathbb{D}, D) \le \arg h(z)/2\pi + C\epsilon \log \frac{1}{\epsilon}.$$

Applying the same reasoning to $\partial H \setminus A$, the lemma follows for all $z \in \partial\mathbb{D}$ and thus for all $z \in \mathbb{D}$. □

Note that the conclusion of Lemma 4.6 is true if, instead of assuming that H is a K-quasidisc, we assume only that $\arg z$ is increasing on ∂H.

Proof of Theorem 4.4. Because Ω_1 and Ω_2 are K-quasidiscs, φ_1 and φ_2 have K^2-quasiconformal extensions to $\mathbb{C}$ (see [L, Chapter I.6]). In particular, they are Hölder continuous with exponent $1/K^2$ (see [A]), and it follows that with $\alpha = 1/K^2$ and $r = 1 - C\epsilon^\alpha$, we have $\varphi_1^{-1}(\{|z| \le r\}) \subset \Omega_2$. In particular, $h(z) = \varphi_2(\varphi_1^{-1}(rz))$ is a conformal map from $\mathbb{D}$ onto a K^4-quasidisc $H \subset \mathbb{D}$, and by the Hölder continuity of φ_2 and φ_1^{-1} we have $\partial H \subset \{1 - C\epsilon^{\alpha^3} < |z| < 1\}$. Now Lemma 4.6 yields $|h(z) - z| \le C\epsilon^\beta$ for any $\beta < \alpha^3$ and $C = C(\beta)$. For $w \in \Omega \subset \Omega_1 \cap \Omega_2$, let $z = \varphi_1(w)$; then

$$\begin{aligned} |\varphi_1(w) - \varphi_2(w)| &= |z - \varphi_2(\varphi_1^{-1}(z))| \\ &\le |z - \varphi_2(\varphi_1^{-1}(rz))| + |\varphi_2(\varphi_1^{-1}(rz)) - \varphi_2(\varphi_1^{-1}(z))| \le C\epsilon^\beta, \end{aligned}$$

where again we have used the Hölder continuity of φ_2 and φ_1^{-1}. The theorem follows. □

5. Convergence of the mapping functions. We will now combine the results of sections 2 and 3 with the estimates of the previous section to obtain quantitative estimates on the convergence of the geodesic algorithm. Throughout this section, Ω will denote a given simply connected domain containing 0, bounded by a Jordan curve $\partial\Omega$, $z_0, \dots, z_n$ are consecutive points on $\partial\Omega$, Ω_c is the domain and $\varphi_c : \Omega_c \to \mathbb{D}$ is the map computed by the geodesic algorithm, and $\varphi : \Omega \to \mathbb{D}$ is a conformal map, normalized so that $\varphi_c(0) = \varphi(0) = 0$ and $\varphi_c(p_0) = \varphi(p_0)$ for some $p_0 \in \partial\Omega \cap \partial\Omega_c$.

Combining Theorems 2.2 and 4.3 and Propositions 2.5 and 3.12 we immediately obtain the following theorem.

THEOREM 5.1. *If $\partial\Omega$ is contained in a closed ϵ-disc-chain $\bigcup_{j=0}^n \overline{D_j}$ and if $z_j = \partial D_j \cap \partial D_{j+1}$, then $\partial\Omega_c$ is a smooth ($C^{\frac{3}{2}}$) piecewise analytic Jordan curve contained in $\bigcup_{j=0}^n D_j \cup z_j$, the map φ_c extends to be conformal on $\Omega \cup \Omega_c$, and*

$$\sup_{w\in\Omega} |\varphi(w) - \varphi_c(w)| \le C\epsilon^{1/2} \log \frac{1}{\epsilon}.$$

Now assume that $\partial\Omega$ is a K-quasicircle with $K < K_0$, and assume approximate equal spacing of the z_j, say, $\frac{1}{2}\epsilon < |z_{j+1} - z_j| < 2\epsilon$. Then

$$\frac{C}{\epsilon} \le n \le \frac{C}{\epsilon^d}, \tag{5.1}$$

where d (essentially the Minkowski dimension) is close to 1 when K is close to 1. Combining Theorem 3.11 with Corollary 4.2 and Theorem 4.4, we have the following theorem.

THEOREM 5.2. *Suppose $\partial\Omega$ is a K-quasicircle with $K < K_0$. The Hausdorff distance between $\partial\Omega$ and $\partial\Omega_c$ is bounded by $C'(K)\epsilon$, where $C'(K)$ is a constant depending upon K that tends to 0 as K tends to 1 and n to infinity. Furthermore,*

$$||\varphi^{-1} - \varphi_c^{-1}||_\infty \le C\epsilon^\alpha$$

and

$$\sup_{w\in\Omega_0} |\varphi(w) - \varphi_c(w)| \le C\epsilon^\alpha$$

with $\alpha = \alpha(K) \to 1$ *as* $K \to 1$, *where* Ω_0 *is the component of* $\Omega \cap \Omega_c$ *containing* 0.

The best possible exponent in (5.1) in terms of the standard definition of $K(\partial\Omega)$, which slightly differs from our geometric definition, is given by Smirnov's (unpublished) proof of Astala's conjecture,

$$d \le 1 + \left(\frac{K-1}{K+1}\right)^2.$$

This allows us to easily convert estimates given in terms of ε, as in Theorem 5.2, into estimates involving n.

Finally, assume that $\partial\Omega$ is a smooth closed Jordan curve. Then Ω is a K-quasicircle and a John-domain by the uniform continuity of the derivative of the arc length parameterization of $\partial\Omega$. The quasiconformal norm $K(\partial\Omega)$ and the John constant depend on the global geometry, as does the ε-pacman condition when there are not very many data points. As the example in Figure 11 shows, even an infinitely differentiable boundary can have a large quasiconformal constant and a large John constant. However, the ε-pacman condition becomes a local condition if the mesh size $\mu(\{z_k\}) = \max_k |z_{k+1} - z_k|$ of the data points is sufficiently small. The radii of the balls in the definition of the ε-pacman condition

$$R_k = C_1 \frac{|z_{k+1} - z_k|}{\varepsilon^2} \tag{5.2}$$

increase as ε decreases but can be chosen small for a fixed ε if the mesh size μ is small. To apply the geodesic algorithm we suppose that the data points have small mesh size and, as in the proof of Theorem 3.10, $|(z_0 - z_n)/(z_{n-1} - z_n)|$ is sufficiently large so that the ε-diamond-chain $D(z_0, z_1), \dots, D(z_{n-1}, z_n)$ satisfies the ε-pacman condition and

$$\partial\Omega \subset \bigcup_{k=0}^{n} D(z_k, z_{k+1}),$$

where $D(z_n, z_{n+1}) = D(z_n, z_0)$ is an ε-diamond. This can be accomplished for smooth curves by taking data points $z_0, \dots, z_n, z_0$ with small mesh size and discarding the last few $z_{n-n_1}, \dots, z_n$, where n_1 is an integer depending on ε and on $\partial\Omega$. The remaining subset still has small mesh size (albeit larger). This process of removing the last few data points is necessary to apply the proof of Theorem 3.10, but in practice it is omitted. We view it only as a defect in the method of proof.

If $\partial\Omega \in C^1$ and if φ is a conformal map of Ω onto $\mathbb{D}$, then $\arg(\varphi^{-1})'$ is continuous. Indeed, it gives the direction of the unit tangent vector. However, there are examples of C^1 boundaries where φ' and $(\varphi^{-1})'$ are not continuous. In fact, it is possible for both to be unbounded. If we make the slightly stronger assumption that $\partial\Omega \in C^{1+\alpha}$ for some $0 < \alpha < 1$, then $\varphi \in C^{1+\alpha}$ and $\varphi^{-1} \in C^{1+\alpha}$ by Kellogg's theorem (see [GM, p. 62]). In particular, the derivatives are bounded above and below on $\overline{\Omega}$ and $\overline{\mathbb{D}}$, respectively. Because of Proposition 3.12, we will consider the case $1 + \alpha = 3/2$. Similar results are true for $1 + \alpha < 3/2$.

THEOREM 5.3. *Suppose $\partial\Omega$ is a closed Jordan curve in $C^{3/2}$ and φ is a conformal map of Ω onto $\mathbb{D}$. Suppose $z_0, z_1, \ldots, z_n, z_0$ are data points on $\partial\Omega$ with mesh size $\mu = \max |z_j - z_{j+1}|$. Then there is a constant C_1 depending on the geometry of $\partial\Omega$, so that the Hausdorff distance between $\partial\Omega$ and $\partial\Omega_c$ satisfies*

$$d_H(\partial\Omega, \partial\Omega_c) \le C_1 \mu^{3/2} \tag{5.3}$$

and the conformal map φ_c satisfies

$$||\varphi^{-1} - \varphi_c^{-1}||_\infty \le C\mu^p \tag{5.4}$$

and

$$\sup_{z \in \Omega \cap \Omega_c} |\varphi(z) - \varphi_c(z)| \le C\mu^p \tag{5.5}$$

for every $p < 3/2$.

For example, if n data points are approximately evenly spaced on $\partial\Omega$, so that $\mu = C/n$, then the error estimates are of the form $C/n^{3/2}$ in (5.3) and C/n^p for $p < 3/2$ in (5.4) and (5.5). While Theorem 5.3 gives simple estimates in terms of the mesh size or the number of data points, smaller error estimates can be obtained with fewer data points if the data points are distributed so that there are fewer on subarcs where $\partial\Omega$ is flat and more where the boundary bends or where it folds back on itself. In other words, construct diamond-chains with angles ε_k satisfying the ε_k-pacman condition centered at z_k for each k. The errors will then be given by

$$\max_k \big(\varepsilon_k |z_k - z_{k+1}|\big)^p.$$

Proof. It is not hard to see from (5.2) that $\partial\Omega$ satisfies the ϵ-pacman condition with

$$\epsilon = C\mu^{1/2}$$

for C sufficiently large. By the proof of Theorem 3.10, $\partial\Omega_c$ is contained in the union of the diamonds. The diamonds $D(z_k, z_{k+1})$ have angle $C\mu^{1/2}$ and width bounded by $C\mu$, and therefore (5.3) holds.

Let ψ be a conformal map of $\mathbb{D}$ onto the complement of $\overline{\Omega}$, $\mathbb{C}^* \setminus \overline{\Omega}$. Then by Kellogg's theorem, as mentioned above, $\psi \in C^{3/2}$. In particular, $|\psi'|$ is bounded above and below on $1/2 < |z| < 1$. By the Koebe distortion theorem there are constants C_1, C_2 so that

$$C_1(1 - |z|) \le \operatorname{dist}(\psi(z), \partial\Omega) \le C_2(1 - |z|)$$

for all z with $1/2 < |z| < 1$. Thus we can choose $r = 1 - C_3\mu^{3/2}$ so that the image of the circle of radius r, $I_r = \psi(\{|z| = r\})$, does not intersect the diamond-chain and $d_H(I_r, \partial\Omega) \sim \mu^{3/2}$. Then the bounded component of the complement of I_r is a Jordan region U_r containing Ω and bounded by $I_r \in C^{3/2}$, with $C^{3/2}$ norm dependent only on $\partial\Omega$, and the bounds on $|\psi'|$.

Let σ be a conformal map of U_r onto $\mathbb{D}$. Inequality (5.4) now follows from [W, Theorem VIII] by comparing the conformal maps φ^{-1} and φ_c^{-1} to the conformal map σ^{-1}, where $\sigma : U_r \to \mathbb{D}$ and where all three (inverse) conformal maps are normalized to have positive derivative at 0 and map 0 to the same point in Ω.

To see (5.5), note that

$$\sigma(\partial\Omega \cup \partial\Omega_c) \subset \{z : 1 - |z| < c\mu^{3/2}\}.$$

Moreover, because $\partial\Omega \cup \partial\Omega_c$ is contained in the diamond-chain, and because both $\sigma \in C^{3/2}$ and $\sigma^{-1} \in C^{3/2}$, $\arg \sigma(\zeta)$ is increasing along $\partial\Omega$ for μ sufficiently small. By the remark after the proof of Lemma 4.6,

$$|\omega(0, \gamma, \sigma(\Omega)) - \omega(0, \gamma^*, \mathbb{D})| \le C\mu^{3/2} \log \mu$$

for every subarc γ of $\sigma(\partial\Omega)$, where γ^* denotes the radial projection of γ onto $\partial\mathbb{D}$. The same statements are true for $\partial\Omega_c$. Then (5.5) follows because the harmonic measure of the subarc γ_p of $\partial\Omega$ from p_0 to p is given by

$$\omega(0, \gamma_p, \Omega) = \frac{1}{2\pi} \arg\left(\frac{\varphi(p)}{\varphi(p_0)}\right),$$

and a similar statement is true for φ_c. □

The constant C in Theorem 5.3 depends on the quasiconformality constant $K = K(\partial\Omega)$, p, $\operatorname{diam}(\Omega)$, $\operatorname{dist}(0, \partial\Omega)$, and

$$M = \sup_{1/2<|z|<1} \big(|\psi'|, 1/|\psi'|\big),$$

where ψ is a conformal map of the complement of Ω to $\mathbb{D}$. If $I_r = \psi(\{|z| = r\})$ is replaced by a $C^{3/2}$ curve which is constructed geometrically instead of using the conformal map ψ, then the constant C can be taken to depend only on the geometry of the region Ω.

Similar results, albeit more complicated, for uniform convergence of the derivatives of the computed maps and the derivatives of their inverses could also be obtained from the results in [W2, Theorems III and V].

6. Some numerical results. An in-depth comparision of the algorithms in this article with other methods of conformal mapping and convergence rates will be written separately. To give the reader a sense of the speed and accuracy of computations, if 10,000 data points are given, it takes about 25 seconds with the geodesic algorithm to compute the conformal maps to the interior, the exterior, and their inverses on a 3.2 GHz Pentium IV computer. Since all of the basic maps are given explicitly in terms of elementary maps, the speed depends only on the number of points and not the shape of the region or the distribution of the data points. The accuracy can be measured if the true conformal map is known. For example,

$$f(z) = \frac{rz}{1 - (rz)^2},$$

where $r < 1$ maps the unit disc into an inverted ellipse. See Figure 12.

The region was chosen because it almost pinches off at 0, and because the stretching/compression given by $\max|f'| / \min|f'|$ is big for r near 1. This is sometimes called the "crowding phenomenon." We chose $r = .95$ and used as data points the image by f of 10,000 equally spaced points on the unit circle, and we compared the corresponding points on the unit circle computed by the geodesic algorithm with 10,000 equally spaced points. The errors were less than $1.8 \cdot 10^{-6}$. The same procedure using the zipper algorithm took 84 seconds and had errors less than $9.2 \cdot 10^{-8}$. When

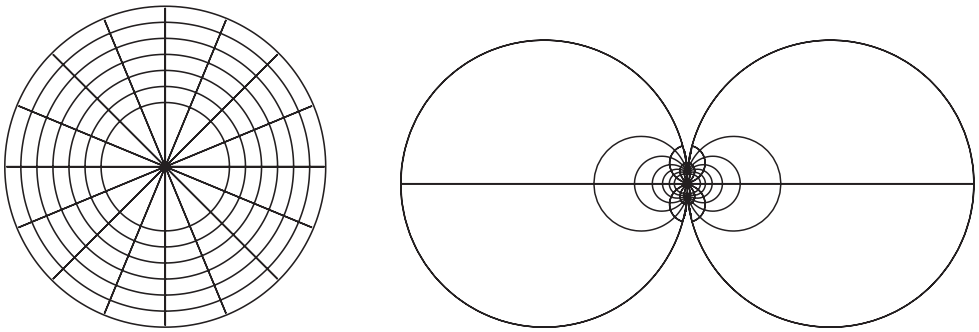


FIG. 12. *Inverted ellipse with* $r = .95$.

the number of data points was increased to 100,000, the time to run the geodesic algorithm increased to 25 minutes with errors less than $2 \cdot 10^{-8}$. In this example, the difference between successive (given) boundary data points on the inverted ellipse ranged from .025 to $3 \cdot 10^{-6}$ so that perhaps a better distribution of data points would have given even smaller errors.

In practice, the choice of data points corresponding to equally spaced points on the circle is not available. An alternative approach to this example is to select data points on the inverted ellipse which are approximately equally spaced in arc length. However, if we choose 10,000 points in this manner, then three consecutive points at the inward pointing "tips" of the region form a "turning angle" of more than 100° because the curvature is so large. This leads to relatively large errors in the map since the tip has large harmonic measure. Another method is to select data points so that the "turning angle"

$$\left|\arg\left(\frac{z_{k+1}-z_k}{z_k-z_{k-1}}\right)\right|$$

is not too big. This results in inaccuracies for this region because the curvature rapidly decreases to zero near the tips, and hence the data points are not very "evenly spaced."

A better method is to use a combination of these ideas. We generated a list of 10^6 points on the boundary of the inverted ellipse and then selected a subset using the following criteria: Having selected $z_1, \ldots, z_k$, choose z_{k+1} to be the first data point in the list after z_k satisfying

$$\left|\arg\left(\frac{z_{k+1}-z_k}{z_k-z_{k-1}}\right)\right| > \delta$$

or

$$\log\left|\frac{z_{k+1}-z_k}{z_k-z_{k-1}}\right| > \delta.$$

To compare with our previous results, we selected $\delta = .0025$ and thereby obtained 9,890 data points with the property that the "turning angle" is small and the ratio of lengths of successive arcs is close to 1. We compared the points on the unit circle obtained from the geodesic algorithm with the true inverse images. The maximal error was less than 5.3×10^{-6}. It is interesting to note that the maximal distance between successive points on the unit circle is 4.2×10^{-2} so that the errors are much

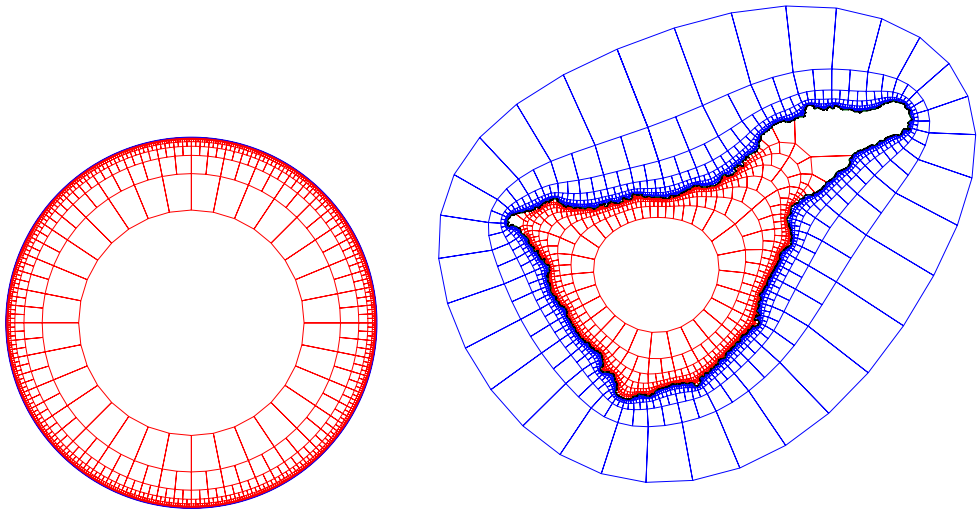


FIG. 13. *Tenerife.*

smaller than the harmonic measure of the corresponding arcs. This technique can be applied to any region where the boundary is known at a very large number of points.

Figure 13 shows the conformal map of a Carleson grid on the disc to both the interior and exterior of the island Tenerife (Canary Islands). We chose this region to illustrate the method on a nonsmooth region where no symmetry is involved. The center of the interior is the volcano Teide. It also shows both the original data for the coastline, connected with straight line segments, and the boundary curve connecting the data points using the zipper algorithm. At this resolution, it is not possible to see the difference between these curves. The zipper algorithm was applied to 6,168 data points and took 36 seconds. The image of 24,673 points on the unit circle took 48 seconds, and all of these points were within $9 \cdot 10^{-5}$ of the polygon formed by connecting the 6,168 data points. The points on the circle corresponding to the 6,168 vertices were mapped to points within 10^{-10} of the vertices. This error is due to the tolerance set for Newton's method, round-off error, and the compression/expansion of harmonic measure. The image of 8,160 vertices in the Carleson grid took 25 seconds to be mapped to the interior and 25 seconds to the exterior.

The first objection one might have in applying these algorithms with a large number of data points is that compositions of even very simple analytic maps can be quite chaotic. Indeed, this is the subject of the field of complex dynamics. We could redefine the basic maps f_a by composing with a linear fractional transformation of the upper-half plane so that the composed map is asymptotic to z as $z \to \infty$. This will not affect the computed curve in these algorithms since the next basic map begins with a linear fractional transformation (albeit altered). However, if we formulate the basic maps in this way, then because the maps are nearly linear near ∞, the numerical errors will accumulate only linearly.

Banjai and Trefethen [BT] adapted fast multipole techniques to the Schwarz–Christoffel algorithm and successfully computed the conformal map to a region which is bounded by a polygon with about 10^5 edges. They used a 12-fold symmetry in the region to immediately reduce the parameter problem to size 10^4. Any other

conformal mapping technique can also use symmetry and obtain a 12-fold reduction in the number of data points required; however, their work does show at least that Schwarz–Christoffel is possible with 10^4 vertices, though convergence of the algorithm to solve the parameter problem is not always ensured. The time it takes to run the zipper algorithm and the resulting accuracy for these snowflake regions is very close to the timing and accuracy for the fast multipole improvements in the Schwarz–Christoffel method. The geodesic algorithm is almost as good and has the advantage that it is very easy to code and convergence can be proved. For a region bounded by a polygon with a small number of vertices, where high accuracy is desired (for instance, errors on the order of 10^{-14}), the Schwarz–Christoffel method is preferable. It would be interesting to try to prove convergence of the technique used in [BT] to find the prevertices in the Schwarz–Christoffel representation for polygons which are K-quasicircles in terms of K. It would be interesting as well to apply multipole techniques to the zipper algorithm. A first step in this direction can be found in Kennedy [KT].

One additional observation worth repeating in this context is that the geodesic and zipper algorithms *always* compute a conformal map of $\mathbb{H}$ to a region bounded by a Jordan curve passing through the data points, even if the disc-chain or pacman conditions are not met. The image region can be found by evaluating the function at a large number of points on the real line. By Proposition 2.5 and Corollary 3.9, if the data points $\{z_j\}$ satisfy the hypotheses of Theorem 2.2 or 3.4, then φ can be analytically extended to be a conformal map of the original region Ω to a region very close to $\mathbb{D}$. To do so requires careful consideration of the appropriate branch of $\sqrt{z}$ at each stage of the composition.

Theorems 2.2 and 3.4 and their proofs suggest how to select points on the boundary of a region to give good accuracy for the mapping functions. Roughly speaking, points need to be chosen closer together where the region comes close to folding back on itself. See Figure 12, for example. Greater accuracy can be obtained by placing more points on the boundary near the center and fewer on the big lobes. See also the remarks after Theorem 5.3 in this regard. In practice, the zipper map works well if points are distributed so that

$$B(z_k, 5|z_{k+1} - z_k|) \cap \partial\Omega \tag{6.1}$$

is connected.

When the boundary of the given region is not smooth, then one of the processes described in section 2 should be used to generate the boundary data, if the geodesic algorithm is to be used. For example, if nothing is known about the boundary except for a list of data points, then we preprocess the data by taking data points along the line segments between the original data points, so that these new points correspond to points of tangency of disjoint circles centered on the line segments, including circles centered at the original data points. Note that the original boundary points are not among these new data points. The geodesic algorithm then finds a conformal map to a region with the new data points on the boundary. The boundary of the new region will be close to the polygonal curve through the original data points but will not pass through the original data points. This boundary is "rounded" near the original data points. Indeed, it is a smooth curve.

When the boundary of the desired region is less smooth, for example, with "corners," then the zipper or slit algorithms should be used. In this case additional points are placed along the line segments between the data points, with at least five points

per edge and satisfying (6.1). In practice, at least 500 points are chosen on the boundary so that the image of the circle will be close to the polygonal line through the data points. Since two data points are pulled down to the real line with each basic map in the zipper algorithm, the original data points should occur at even numbered indices in the resulting data set (the first data point is called z_0). Then the computed boundary Ω_c will have corners at each of the original data points, with angles very close to the angles of the polygon through the original data points.

Fortran programs for a version of the zipper algorithm can be obtained from [MD]. Also included is a graphics program, written in C with X-11 graphics by Mike Stark, for the display of the conformal maps. There are also several demo programs applying the algorithm to problems in elementary fluid flow, extremal length, and hyperbolic geometry. Extensive testing of the geodesic algorithm [MM] and an early version of the zipper algorithm was done in the 1980s with Morrow. In particular, that experimentation suggested the initial function φ_0 in the zipper algorithm which maps the complement of a circular arc through z_0, z_1, and z_2 onto $\mathbb{H}$.

Appendix. Jørgensen's theorem. Since Jørgensen's theorem is a key component of the proof of the convergence of the geodesic algorithm, we include a short self-contained proof. It says that discs are strictly convex in the hyperbolic geometry of a simply connected domain Ω (unless $\partial\Omega$ is contained in the boundary of the disk).

THEOREM A.1 (Jørgensen [J]). *Suppose Ω is a simply connected domain. If Δ is an open disc contained in Ω and if γ is a hyperbolic geodesic in Ω, then $\gamma \cap \Delta$ is connected, and if it is nonempty, it is not tangent to $\partial\Delta$ in Ω.*

Proof (see [P, pp. 91–93]). Applying a linear fractional transformation to Ω, we replace the disc Δ by the upper-half plane $\mathbb{H}$. Suppose $x \in \mathbb{R}$ and suppose that f is a conformal map of $\mathbb{D}$ onto Ω such that $f(0) = x$ and $f'(0) > 0$. We will use the auxiliary function $z + 1/z$, which is real-valued on $\partial\mathbb{D} \cup (-1, 1)$. Then

$$\operatorname{Im}\left(\frac{f'(0)}{f(z) - x} - \left(\frac{1}{z} + z \right) \right)$$

is a bounded harmonic function on $\mathbb{D}$ which is greater than or equal to 0 by the maximum principle. Thus $\operatorname{Im} \frac{f'(0)}{f(z)-x} \geq 0$ on $(-1, 1)$, and hence $\operatorname{Im} f(z) \leq 0$ on the diameter $(-1, 1)$. The condition $f'(0) > 0$ means that the geodesic $f\big((-1, 1)\big)$ is tangent to $\mathbb{R}$ at x. Two circles which are orthogonal to $\partial\mathbb{D}$ can meet in $\mathbb{D}$ in at most one point, and hence hyperbolic geodesics in simply connected domains (images of orthogonal circles) meet in at most one point and are not tangent. Thus if γ is a geodesic in Ω which intersects $\mathbb{H}$ and contains the point x, then it cannot be tangent to $\mathbb{R}$ at x and cannot re-enter $\mathbb{H}$ after leaving it at x because it is separated from $\mathbb{H}$ by the geodesic $f\big((-1, 1)\big)$. The theorem follows. □

In section 2, we commented that a constructive proof of the Riemann mapping theorem followed from the proof of Theorem 2.2. The application of Jørgensen's theorem in the proof of Theorem 2.2 is only to domains for which the Riemann map has been explicitly constructed.

Acknowledgments. The first author would like to express his deep gratitude to L. Carleson for our exciting investigations at The Mittag-Leffler Institute (1982–1983) which led to the discovery of the zipper algorithms. We would like to thank the referees for careful reading and useful comments.

REFERENCES

[A] L. Ahlfors, *Lectures on Quasiconformal Mappings*, Van Nostrand, Princeton, NJ, 1966.

[BT] L. Banjai and L. N. Trefethen, *A multipole method for Schwarz–Christoffel mapping of polygons with thousands of sides*, SIAM J. Sci. Comput., 25 (2003), pp. 1042–1065.

[GM] J. Garnett and D. E. Marshall, *Harmonic Measure*, Cambridge University Press, New York, 2005.

[H] P. Henrici, *Applied and Computational Complex Analysis, Vol.* 3, John Wiley & Sons, New York, 1986.

[J] V. Jørgensen, *On an inequality for the hyperbolic measure and its applications in the theory of functions*, Math. Scand., 4 (1956), pp. 113–124.

[KT] T. Kennedy, *Computing the Loewner Driving Process of Random Curves in the Half Plane*, preprint, http://arxiv.org/PS_cache/math/pdf/0702/0702071v1.pdf.

[K] R. Kühnau, *Numerische Realisierung konformer Abbildungen durch "Interpolation,"* Z. Angew. Math. Mech., 63 (1983), pp. 631–637.

[L] O. Lehto, *Univalent Functions and Teichmüller Spaces*, Springer-Verlag, New York, 1987.

[M] A. R. Marchenko, *Sur la representation conforme*, C. R. Acad. Sci. USSR, 1 (1935), pp. 289–290.

[MD] D. E. Marshall, *Zipper*, Fortran Programs for Numerical Computation of Conformal Maps, and C Programs for X-11 Graphics Display of the Maps. Sample pictures, Fortran, and C code available online at http://www.math.washington.edu/~marshall/personal.html.

[MM] D. E. Marshall and J. A. Morrow, *Compositions of Slit Mappings*, manuscript, 1987.

[P] C. Pommerenke, *Boundary Behaviour of Conformal Maps*, Springer-Verlag, Berlin, 1992.

[SK] K. Stephenson, *Circle packing: A mathematical tale*, Notices Amer. Math. Soc., 50 (2003), pp. 1376–1388.

[T] M. Tsuji, *Potential Theory in Modern Function Theory*, Chelsea, New York, 1975.

[W] S. Warschawski, *On the degree of variation in conformal mapping of variable regions*, Trans. Amer. Math. Soc., 69 (1950), pp. 335–356.

[W2] S. Warschawski, *On the distortion in conformal mapping of variable domains*, Trans. Amer. Math. Soc., 82 (1956), pp. 300–322.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2610–2636

CONVERGENCE RATES OF GENERAL REGULARIZATION METHODS FOR STATISTICAL INVERSE PROBLEMS AND APPLICATIONS*

N. BISSANTZ†, T. HOHAGE‡, A. MUNK‡, AND F. RUYMGAART§

Abstract. Previously, the convergence analysis for linear statistical inverse problems has mainly focused on spectral cut-off and Tikhonov-type estimators. Spectral cut-off estimators achieve minimax rates for a broad range of smoothness classes and operators, but their practical usefulness is limited by the fact that they require a complete spectral decomposition of the operator. Tikhonov estimators are simpler to compute but still involve the inversion of an operator and achieve minimax rates only in restricted smoothness classes. In this paper we introduce a unifying technique to study the mean square error of a large class of regularization methods (spectral methods) including the aforementioned estimators as well as many iterative methods, such as ν-methods and the Landweber iteration. The latter estimators converge at the same rate as spectral cut-off but require only matrix-vector products. Our results are applied to various problems; in particular we obtain precise convergence rates for satellite gradiometry, L^2-boosting, and errors in variable problems.

Key words. statistical inverse problems, iterative regularization methods, Tikhonov regularization, nonparametric regression, minimax convergence rates, satellite gradiometry, Hilbert scales, boosting, errors in variable

AMS subject classifications. 62G05, 62J05, 62P35, 65J10, 35R30

DOI. 10.1137/060651884

1. Introduction.

This paper is concerned with estimating an element f of a Hilbert space $\mathbb{H}_1$ from indirect noisy measurements

$$Y = Kf + \text{“noise”} \tag{1.1}$$

related to f by a (known) operator $K : \mathbb{H}_1 \to \mathbb{H}_2$ mapping $\mathbb{H}_1$ to another Hilbert space $\mathbb{H}_2$. The operator K is assumed to be linear, bounded, and injective, but *not* necessarily compact. We are interested in the case that the operator equation (1.1) is ill-posed in the sense that the Moore–Penrose inverse of K is unbounded. The analysis of regularization methods for the stable solution of (1.1) depends on the mathematical model for the noise term on the right-hand side of (1.1): If the noise is considered as a deterministic quantity, it is natural to study the worst-case error. In the literature a number of efficient methods for the solution of (1.1) have been developed, and it has been shown under certain conditions that the worst-case error converges at optimal order as the noise level tends to 0 (see Engl, Hanke, and Neubauer [14]). If the noise is modeled as a random quantity, the convergence of estimators $\hat{f}$ of f should be studied in statistical terms. Here we consider the expected square error $\mathbf{E}\,\|\hat{f} - f\|^2$, also

*Received by the editors February 10, 2006; accepted for publication (in revised form) April 23, 2007; published electronically December 7, 2007.
http://www.siam.org/journals/sinum/45-6/65188.html

†Lehrstuhl für Stochastik, Ruhr Universität Bochum, D-44780 Bochum, Germany (nicolai.bissantz@ruhr-uni-bochum.de).

‡Institute of Numerical and Applied Mathematics, University of Göttingen, D-37083 Göttingen, Germany (hohage@math.uni-goettingen.de, munk@makt.uni-goettingen.de). The work of these authors was supported by the Graduiertenkolleg 1023 “Identification in Mathematical Models,” DFG grant MU 1230/8-1, and by the DFG “Sonderforschungsbereich” 475.

§Department of Mathematics and Statistics, Texas Tech University, Lubbock, TX 79409 (h.ruymagaart@ttu.edu).

called the *mean integrated square error* (MISE). This problem has also been studied extensively in the statistical literature, but the numerical efficiency has not been a major issue so far. It is the purpose of this paper to provide an analysis of a class of computationally efficient regularization methods including Landweber iteration, ν-methods, and iterated Tikhonov regularization, which is applicable to linear inverse problems with random noise as they occur, for example, in parameter identification problems in partial differential equations (PDEs), deconvolution, or errors in variable models.

There exists a considerable amount of literature on regularization methods for linear inverse problems with random noise. For surveys we refer to O'Sullivan [37], Nychka and Cox [36], Evans and Stark [16], and Kaipio and Somersalo [26]. A large part of the literature focuses on methods which require the explicit knowledge of a spectral decomposition of the operator K^*K. The simplest of these methods is spectral cut-off (or truncated singular value decomposition (SVD) for compact operators) where an estimator is constructed by a truncated expansion of f with respect to (w.r.t.) the eigenfunctions of K^*K (see, e.g., Diggle and Hall [10] and Healy, Hendriks, and Kim [21]). It has been shown in a number of papers that spectral cut-off estimators are order optimal in a minimax sense under certain conditions (see, e.g., Mair and Ruymgaart [30], Efromovich [13], and Kim and Koo [27]). Based on an SVD of K it is also possible to construct exact minimax estimators for given smoothness classes (see Johnstone and Silverman [25]).

Another major approach is wavelet-vaguelette (and vaguelette-wavelet)-based methods which lead to estimators of a functional form similar to that of SVD methods. In general these estimators are based on expansions of f and Kf w.r.t. different bases of the respective function spaces than those provided by the SVD of K (see, e.g., Donoho [12], Abramovich and Silverman [1], and Johnstone et al. [24]).

A well-known method both in the statistical and the deterministic inverse problems literature is Tikhonov regularization. This has been studied for certain classes of linear statistical inverse problems by Cox [9], Nychka and Cox [36], and Mathé and Pereverzev [31, 33].

The main restriction of the usefulness of spectral cut-off and related estimators is the need of the spectral data of the operator (i.e., an SVD if K is compact) to implement these estimators. This is known explicitly only in a limited number of special cases, and numerical computation of the spectral data is prohibitively expensive for many situations. Although Tikhonov regularization does not require the spectral data of the operator, there is still the requirement of setting up and inverting a matrix representing the operator. For iterative regularization methods such as Landweber iteration or ν-methods (see Nemirovskii and Polyak [35], Brakhage [6], and Engl, Hanke, and Neubauer [14]) only matrix-vector multiplications are required. Furthermore, it is known that Tikhonov regularization achieves minimax rates of convergence only in a restricted number of smoothness classes, which is highlighted by the fact that its qualification number is 1, whereas Landweber iteration has infinitely large qualification, and ν-methods with qualification ν are available for every $\nu > 0$ (see [14]).

Iterative regularization methods are particularly attractive for inverse problems in PDEs. Here the operator K maps an unknown parameter f in a PDE to (part of) the solution to this PDE. Hence, applying K to a vector f simply means solving the PDE with the parameter f, whereas inverting or even setting up the matrix for K is often not feasible. We will discuss two linear inverse problem for PDEs (the backwards

heat equation and satellite gradiometry) in section 5. However, most inverse problems for PDEs are nonlinear even if the PDE is linear. Such problems are often solved by regularized Newton methods. In this case the methods and the analysis of this paper can be applied to the linearized operator equations in each Newton step as discussed in the forthcoming paper [2].

In this paper we will show that general spectral regularization methods as defined in section 2 achieve the same rates of convergence of the MISE as spectral cut-off, which is known to be optimal in most cases (see above). Whereas the bias or approximation error is exactly the same in a deterministic and a statistical framework, the analysis significantly differs in the estimation of the noise term. In spectral cut-off for compact operators, the noise (or variance) part of the estimators $\hat{f}_\alpha$ belongs to a finite-dimensional space of "low-frequencies." The main difficulty in the analysis of general spectral regularization methods is the estimation of the "high frequency" components of the noise. Unlike in a deterministic framework, the bound on the noise term depends not only on the regularization parameter, but also on the distribution of the singular values of K (if K is compact). Therefore, a statistical analysis has to impose additional conditions on the operator. We will verify these conditions for several important problems, including inverse problems in PDEs and errors in variable models. As an example of particular interest in the machine learning context we obtain optimal rates of convergence of L^2-boosting by interpreting L^2-boosting as a Landweber iteration (see also Bühlmann and Yu [7] and Yao, Rosasco, and Caponnetto [42]).

The plan of this paper is as follows: Section 2 gives a brief overview of regularization methods and source conditions and introduces an abstract noise model. Section 3 contains the main results of this paper on the rates of convergence of general spectral regularization methods. In section 4 we demonstrate how a number of commonly used statistical noise models fit into our general framework. Finally, in section 5 we discuss the application of our results to the backwards heat equation, satellite gradiometry, errors in variable models with dependent random variables, L^2-boosting, and operators in Hilbert scales. Proofs of section 3 are collected in section 6.

2. Framework.

We first review some basic notions of regularization theory.

2.1. Spectral theorem.

Halmos's version of the spectral theorem (see, for instance, Halmos [20] and Taylor [40]) turns out to be particularly convenient for the construction and statistical analysis of regularized inverses of a self-adjoint operator. This has been demonstrated by Mair and Ruymgaart [30] for the spectral cut-off estimator. The theorem claims that for a (not necessarily bounded) self-adjoint operator $A : D(A) \to \mathbb{H}$ defined on a dense subset $D(A)$ of a separable Hilbert space $\mathbb{H}$ there exists a σ-compact space $\mathbb{S}$, a Borel measure Σ on $\mathbb{S}$, a unitary operator $U : \mathbb{H} \to L^2(\Sigma)$, and a measurable function $\rho : \mathbb{S} \to \mathbb{R}$ such that

$$UAf = \rho \cdot Uf, \qquad \Sigma\text{-a.e.}, \tag{2.1}$$

for all $f \in D(A)$. Introducing the multiplication operator $M_\rho : D(M_\rho) \to L^2(\Sigma)$, $M_\rho\varphi := \rho \cdot \varphi$ defined on $D(M_\rho) := \{\varphi \in L^2(\Sigma) : \rho\varphi \in L^2(\Sigma)\}$, we can rewrite (2.1) as $A = U^* M_\rho U$, i.e., A is unitarily equivalent to a multiplication operator. The essential range of ρ is the spectrum $\sigma(A)$ of A. If A is bounded and positive definite as below, then $0 < \rho \le \|A\|$, Σ-a.e.

Remark 1. In the special case that A is compact, a well-known version of the spectral theorem states that A has a complete orthonormal system of eigenvectors

u_i with corresponding eigenvalues ρ_i, and $Af = \sum_{j=0}^{\infty} \rho_j \langle u_j, f\rangle u_j$. This can be rewritten in the multiplicative form (2.1) by choosing Σ as the counting measure on $\mathbb{S} = \mathbb{N}$, i.e., $L^2(\Sigma) = l^2(\mathbb{N})$, the multiplicator function as $\rho(i) = \rho_i$, $i \in \mathbb{N}$, and defining the unitary operator $U : \mathbb{H} \to l^2(\mathbb{N})$ by $(Uf)(i) := \langle f, u_i\rangle$, $i \in \mathbb{N}$.

2.2. Regularized estimators. Recall Halmos's spectral theorem from section 2.1. For a self-adjoint operator $A : D(A) \to \mathbb{H}$ and a bounded, measurable function $\Phi : \sigma(A) \to \mathbb{R}$ one defines an operator $\Phi(A) \in L(\mathbb{H})$ by

$$\Phi(A) = U^* M_{\Phi(\rho)} U \tag{2.2}$$

(see, e.g., Taylor [40]). The mapping $\Phi \mapsto \Phi(A)$, called the *functional calculus* at A, is an algebra homomorphism from the algebra of bounded measurable functions on $\sigma(A)$ to the algebra $L(\mathbb{H})$ of bounded linear operators on $\mathbb{H}$, and

$$\|\Phi(A)\| \le \sup_{\lambda \in \sigma(A)} |\Phi(\lambda)|, \tag{2.3}$$

with equality if Φ is continuous. We will construct estimators of the input function by regularization methods of the form

$$\hat{f}_{\alpha,\sigma} = \Phi_\alpha(K^*K)K^*Y. \tag{2.4}$$

Here $\Phi_\alpha : \sigma(K^*K) \to \mathbb{R}$ is a collection of bounded filter functions approximating the unbounded function $t \mapsto \frac{1}{t}$ on $\sigma(K^*K)$, which are parametrized by a *regularization parameter* $\alpha > 0$.

A particular example of a regularization method of the form (2.4) is the *spectral cut-off* estimator (also known as *truncated SVD*) described by the functions

$$\Phi_\alpha^{\mathrm{SC}}(t) := \begin{cases} t^{-1}, & t \ge \alpha, \\ 0, & t < \alpha. \end{cases}$$

As explained in the introduction, we will focus on regularization methods which can be implemented without explicit knowledge of the spectral decomposition of the operator K^*K. This includes both *implicit methods* such as Tikhonov regularization ($\Phi_\alpha(t) = (\alpha + t)^{-1}$), iterated Tikhonov regularization, and Lardy's method, which involve the inversion of an operator, and *explicit methods* such as Landweber iteration ($\Phi_{1/(k+1)}(t) = \sum_{j=0}^{k-1}(1 - \beta t)^j$, where $\beta \in (0, \|K^*K\|^{-2})$ is a step-length parameter, and ν-methods, which require only matrix-vector products in a discrete setting. For a derivation and discussion of these methods we refer to the monograph [14].

2.3. Smoothness classes. We will measure the smoothness of the input function f relative to the smoothing properties of K in terms of *source conditions*: Let $\Lambda : [0, \infty) \to [0, \infty)$ be a continuous, strictly increasing function with $\Lambda(0) = 0$, and assume that there exists a "source" $w \in \mathbb{H}_1$ such that

$$f = \Lambda(K^*K)w \tag{2.5}$$

(see [14, 15, 32]). The set of all f satisfying this condition with $\|w\|_{\mathbb{H}_1} \le \overline{w}$, $\overline{w} > 0$ will be denoted by $F_{\Lambda,\overline{w},K^*K} := \{\Lambda(K^*K)w : w \in \mathbb{H}_1, \|w\| \le \overline{w}\}$. We will shorten this to $F_{\Lambda,\overline{w}} := F_{\Lambda,\overline{w},K^*K}$ if there is no ambiguity. The most common choice, which is usually appropriate for finitely smoothing operators K, is

$$\Lambda(t) = t^\nu, \qquad \nu > 0. \tag{2.6}$$

In particular, (2.5) with $\Lambda(t) = \sqrt{t}$ is equivalent to $f = K^*v$, $\|v\|_{\mathbb{H}_2} \leq 1$ (see Engl, Hanke, and Neubauer [14, Prop. 2.18]). For exponentially ill-posed problems such as the backwards heat equation, (2.6) is usually too restrictive, and *logarithmic source conditions* corresponding to the choice

$$\Lambda(t) = (-\ln t)^{-p}, \qquad p > 0, \tag{2.7}$$

are more appropriate (see Hohage [23] and Mair [29]). Since Λ is singular at $t = 1$, we assume that the norms in $\mathbb{H}_1$ and $\mathbb{H}_2$ are scaled such that $\|K^*K\| < 1$ in this case. For a further discussion of source conditions and interpretations as smoothness conditions in Sobolev spaces we refer to the applications in section 5.

If f belongs to the smoothness class $F_{\Lambda,\overline{w}}$ and we are given exact data $Y = g$, then the error is bounded by

$$\begin{aligned} \|\Phi_\alpha(K^*K)K^*g - f\| &= \|(\Phi_\alpha(K^*K)K^*K - I)\Lambda(K^*K)w\| \\ &\leq \sup_{t\in\sigma(K^*K)} |(\Phi_\alpha(t)t - 1)\Lambda(t)|\overline{w}, \end{aligned} \tag{2.8}$$

where we have used (2.3).

2.4. Assumptions on smoothness and the regularization method. In the following we discuss a number of standard assumptions on the filter functions Φ_α satisfied for all commonly used regularization methods, in particular those discussed in section 2.2 (see [14]). First, we assume that there exists a constant $C_2 > 0$ such that

$$\sup_{t\in\sigma(K^*K)} |t\Phi_\alpha(t)| \leq C_2, \qquad \text{uniformly in } \alpha > 0. \tag{2.9a}$$

To bound the so-called propagated deterministic noise error $\tau\|\Phi_\alpha(K^*K)K^*\xi\|$, we impose the following condition:

$$\text{there exists } C_3 > 0 : \sup_{\alpha>0} \sup_{t\in\sigma(K^*K)} |\alpha\Phi_\alpha(t)| \leq C_3. \tag{2.9b}$$

In view of the bound (2.8) on the approximation error, we also assume that there exists a number $\nu_0 > 0$ called *qualification* of the method and constants $\gamma_\nu > 0$ such that

$$\sup_{t\in\sigma(K^*K)} |t^\nu(1 - t\Phi_\alpha(t))| \leq \gamma_\nu\alpha^\nu \qquad \text{for all } \alpha \text{ and all } 0 \leq \nu \leq \nu_0. \tag{2.9c}$$

The qualification of a method is a measure of the maximal degree of smoothness in terms of the Hölder-type conditions (2.5), (2.6) under which the approximation error (2.8) converges at optimal order. The following are qualifications of some commonly used methods: Tikhonov regularization: 1; K-times iterated Tikhonov regularization: K; Landweber iteration: ∞ (in the sense that it is greater than any real number); ν-methods: ν (where $\nu > 0$ is a parameter in the method); see references in the introduction.

Note that the condition (2.9c) with $\nu_0 > 0$ implies that $\lim_{\alpha\searrow 0} \Phi_\alpha(t) = \frac{1}{t}$ for all $t \in \sigma(K^*K)$. For $\nu = 0$ the condition (2.9c) implies (2.9a) with $C_2 = 1+\gamma_0$. However, this value of C_2 is usually not optimal since for most regularization methods (2.9a) holds true with $C_2 = 1$.

For general source conditions we assume that there exists a constant γ_Λ such that

$$\sup_{t\in\sigma(K^*K)} |\Lambda(t)(1-t\Phi_\alpha(t))| \le \gamma_\Lambda \Lambda(\alpha), \qquad \alpha \searrow 0. \tag{2.10}$$

Under Hölder-type source conditions (2.6) this holds true for $\nu \le \nu_0$ by assumption (2.9c). For the choice $\Lambda(t) = (-\ln t)^{-p}$, it has been shown in Hohage [23] that (2.9c) with $\nu_0 > 0$ implies (2.10). For more general functions Λ we refer to Mathé and Pereverzev [32] for similar implications.

2.5. Noise model. In this subsection we introduce an abstract noise model which will be used in the proof of our main result. In section 4 we will demonstrate that several noise models commonly encountered in statistical modeling fit into this general framework.

Following Mathé and Pereverzev [31], we assume that our given data can be written as

$$Y = g + \sigma\varepsilon + \tau\xi, \qquad g := Kf, \tag{2.11}$$

where $\xi \in \mathbb{H}_2$, $\|\xi\| \le 1$ is a deterministic error, ε is a stochastic error, and $\tau, \sigma > 0$ are the corresponding noise levels. Note that model (2.11) allows for stochastic and deterministic noise, simultaneously.

Often, the stochastic error is modeled as a Hilbert space–valued random variable Ξ, i.e., a measurable function $\Xi : \Omega \to \mathbb{H}_2$, where $(\Omega, \mathcal{P}, P)$ is the underlying probability space. However, we will assume more generally that it is a Hilbert-space process, i.e., a continuous linear operator

$$\varepsilon : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P).$$

Every Hilbert space–valued random variable Ξ with finite second moments, $\mathbf{E}\,\|\Xi\|^2 < \infty$, can be identified with a Hilbert-space process $\varphi \mapsto \langle \Xi, \varphi\rangle$, $\varphi \in \mathbb{H}_2$, but not vice versa. We will use the notation $\langle \varepsilon, \varphi\rangle := \varepsilon\varphi$, $\varphi \in \mathbb{H}_2$. The covariance $\mathbf{Cov}_\varepsilon : \mathbb{H}_2 \to \mathbb{H}_2$ of a Hilbert-space process $\varepsilon : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P)$ is the bounded linear operator defined implicitly by $\langle \mathbf{Cov}_\varepsilon \varphi_1, \varphi_2\rangle = \mathbf{Cov}\left(\langle \varepsilon, \varphi_1\rangle, \langle \varepsilon, \varphi_2\rangle\right)$, $\varphi_1, \varphi_2 \in \mathbb{H}_2$. We call ε a *white noise process* if $\mathbf{Cov}_\varepsilon = I$ and $\mathbf{E}\,\langle \varepsilon, \varphi\rangle = 0$ for all $\varphi \in \mathbb{H}_2$. Note that a Gaussian white noise process in an infinite-dimensional Hilbert space cannot be identified with a Hilbert space–valued random variable.

If $\varepsilon : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P)$ is a Hilbert-space process and $A : \mathbb{H}_2 \to \mathbb{H}_1$ is a bounded linear operator, we define the Hilbert-space process $A\varepsilon : \mathbb{H}_1 \to L^2(\Omega, \mathcal{P}, P)$ by $\langle A\varepsilon, \varphi\rangle := \langle \varepsilon, A^*\varphi\rangle$, $\varphi \in \mathbb{H}_1$. Its covariance operator is given by $\mathbf{Cov}_{A\varepsilon} = A\mathbf{Cov}_\varepsilon A^*$.

Assumption 1. In the noise model (2.11) $\xi \in \mathbb{H}_2$ is a deterministic vector with $\|\xi\| = 1$, and ε is a Hilbert-space process such that

$$\mathbf{E}\,\langle \varepsilon, \varphi\rangle = 0, \qquad \|\mathbf{Cov}_\varepsilon\| \le 1 \tag{2.12}$$

for all $\varphi \in \mathbb{H}_2$. Moreover, $K^*\varepsilon$ is a Hilbert space–valued random variable satisfying

$$\mathbf{E}\,\|K^*\varepsilon\|^2 < \infty, \tag{2.13}$$

and there exists a spectral decomposition (2.1) of K^*K such that for almost all $s \in \mathbb{S}$

$$\mathbf{Var}\,(UK^*\varepsilon(s)) \le \rho(s). \tag{2.14}$$

The first condition in (2.12) is not a restriction since an expected value different from zero can be included in $\tau\xi$, and the second condition is a scaling condition analogous to $\|\xi\| \le 1$. Assumption (2.13) ensures that the estimators defined in (2.4) are Hilbert space–valued random variables with finite second moments. Inequality (2.13) is usually a mild assumption, but it excludes, e.g., very mildly ill-posed problems in combination with white noise. The following lemma implies that (2.14) is a condition on the choice of U in the Halmos representation (2.1) rather than a condition on the noise model. Moreover, we can arrange that $\rho \in L^1(\Sigma)$, as required in section 3 below. Noise models with a finite number of observations satisfying Assumption 1 are discussed in section 4 below.

LEMMA 2. *If ε is a Hilbert-space process satisfying* (2.12), *$K^*\varepsilon$ is a Hilbert space–valued random variable satisfying* (2.13), *and K is injective, then there exists a spectral decomposition* (2.1) *of K^*K such that* (2.14) *holds true, and $\rho \in L^1(\Sigma)$.*

Proof. According to Halmos's spectral theorem there exists a Borel measure $\tilde\Sigma$ on a σ-compact space $\mathbb{S}$, and a unitary operator $\tilde U : L^2(\mathbb{R}^d) \to L^2(\tilde\Sigma)$ such that $K^*K = \tilde U^* M_\rho \tilde U$. For any $\tilde\Sigma$-measurable function $\chi > 0$ on $\mathbb{S}$ we can construct another Halmos representation of K^*K by introducing the Borel measure $\Sigma := \chi\tilde\Sigma$ on $\mathbb{S}$ and the mapping $U : L^2(\mathbb{R}^d) \to L^2(\Sigma)$, $Uf := \chi^{-1/2} \cdot \tilde U f$ since U is unitary and $UK^*Kf = \rho \cdot Uf$, Σ-a.e., for all $f \in \mathbb{H}_1$. In particular, we may define

(2.15)
$$\chi(s) := \frac{\mathbf{Var}\,(\tilde U K^*\varepsilon)(s)}{\rho(s)} \qquad \text{for } s \in M, \quad M := \{s \in \mathbb{S} : \mathbf{Var}\,(\tilde U K^*\varepsilon)(s) > 0\}.$$

Here we use that $\rho > 0$, $\tilde\Sigma$-a.e., since K and hence K^*K is injective by assumption. We first consider the case $\tilde\Sigma(M^c) = 0$, where $M^c := \mathbb{S} \setminus M$. Then (2.14) holds true for $s \in M$ as $\mathbf{Var}\,(UK^*\varepsilon)(s) = \chi(s)^{-1}\mathbf{Var}\,(\tilde U K^*\varepsilon)(s) = \rho(s)$. Moreover,

$$\int \rho \,\mathrm{d}\Sigma = \int \mathbf{Var}\,(UK^*\varepsilon)\,\mathrm{d}\Sigma = \mathbf{E}\int |UK^*\varepsilon|^2\,\mathrm{d}\Sigma = \mathbf{E}\,\|K^*\varepsilon\|^2 < \infty, \tag{2.16}$$

which is the assertion. Now assume that $\tilde\Sigma(M^c) > 0$. Let ψ be an arbitrary strictly positive function in $L^1(\tilde\Sigma)$, e.g., $\psi(s) := \big(j(s)^2\tilde\Sigma(A_{j(s)})\big)^{-1}$, where $j(s) := \min\{j : s \in A_j\}$ for a sequence $A_1 \subset A_2 \subset \cdots \subset \Omega$ with $\tilde\Sigma(A_j) < \infty$ and $\tilde\Sigma(\mathbb{S} \setminus \bigcup_j A_j) = 0$. Such a sequence exists because $\tilde\Sigma$ is σ-finite. We define $\chi(s)$ by (2.15) for $s \in M$ and $\chi(s) := \frac{\psi(s)}{\rho(s)}$ for $s \in M^c$. Then (2.14) is trivially satisfied for $s \in M^c$, and $\rho \in L^1(\Sigma)$ since $\int_M \rho\,\mathrm{d}\Sigma < \infty$ as in (2.16) and $\int_{M^c} \rho\,\mathrm{d}\Sigma \le \int \psi\,\mathrm{d}\tilde\Sigma < \infty$. This finishes the proof. ∎

3. MISE estimates. In this section the main results of this paper are presented. Recall the definition of the estimator $\hat f_{\alpha,\sigma}$ of the input function f in (2.4). Since $\mathbf{E}\,\Phi_\alpha(K^*K)K^*\varepsilon = 0$, the MISE satisfies the bias-variance decomposition

$$\mathbf{E}\,\|\hat f_{\alpha,\sigma} - f\|^2 = \mathbf{B}\,(\hat f_{\alpha,\sigma})^2 + \mathbf{E}\,\|\hat f_{\alpha,\sigma} - \mathbf{E}\,\hat f_{\alpha,\sigma}\|^2, \tag{3.1}$$

with the bias term $\mathbf{B}\,(\hat f_{\alpha,\sigma}) := \|\mathbf{E}\,\hat f_{\alpha,\sigma} - f\|$. As discussed in the introduction, the bias term can be bounded by standard estimates, whereas the variance term requires a special treatment involving a splitting in the frequency domain.

3.1. Estimation of the bias. The bias in our model coincides with the error in a deterministic setting and can be estimated by standard techniques (see [14]).

Using the triangle inequality, the noise model (2.11), (2.12), and the definition (2.4) of $\hat{f}_{\alpha,\sigma}$, we get

$$\mathbf{B}\,(\hat{f}_{\alpha,\sigma}) \le \|\Phi_\alpha(K^*K)K^*Kf - f\| + \tau\|\Phi_\alpha(K^*K)K^*\xi\|.$$

The first term (called the approximation error) is bounded by $\gamma_\Lambda\Lambda(\alpha)\overline{w}$ due to (2.8) and (2.10). For the second term (called the propagated deterministic noise error) we obtain the bound

$$\|\Phi_\alpha(K^*K)K^*\xi\|^2 = \langle\Phi_\alpha(KK^*)\xi, KK^*\Phi_\alpha(KK^*)\xi\rangle \le \frac{C_2C_3}{\alpha} \tag{3.2}$$

using the identity $\Phi_\alpha(K^*K)K^* = K^*\Phi_\alpha(KK^*)$ (see [14, eq. (2.43)]) and (2.9). Hence,

$$\mathbf{B}\,(\hat{f}_{\alpha,\sigma}) \le \gamma_\Lambda\Lambda(\alpha)\overline{w} + \sqrt{\frac{C_2C_3}{\alpha}}\tau. \tag{3.3}$$

Since we aim to show optimality of general regularization methods by comparison to spectral cut-off (see the introduction and section 3.3), we now compare the approximation errors of general regularization methods and spectral cut-off. To this end, we introduce the following notation.

Notation. For two real-valued functions f, g defined on an interval $(0, \bar{\alpha}]$ we write

$$f(\alpha) \sim g(\alpha) \qquad (\text{or } f(\alpha) \lesssim g(\alpha)), \qquad \text{as } \alpha \searrow 0,$$

if $g(\alpha) \neq 0$ for α in some neighborhood of 0 and $\lim_{\alpha\searrow 0}\frac{f(\alpha)}{g(\alpha)} = 1$ or $\limsup_{\alpha\searrow 0}\frac{f(\alpha)}{g(\alpha)} \le 1$. Furthermore, we write

$$f(\alpha) \asymp g(\alpha), \qquad \text{as } \alpha \searrow 0,$$

if there exist constants $\bar{\alpha} > 0$ and $C_{\bar{\alpha}} \ge 1$ such that $(1/C_{\bar{\alpha}})f(\alpha) \le g(\alpha) \le C_{\bar{\alpha}}f(\alpha)$ for $0 < \alpha \le \bar{\alpha}$.

Recall that $\Lambda : [0,\infty) \to [0,\infty)$ is assumed to be a strictly increasing, continuous function with $\Lambda(0) = 0$ and that $1 - t\Phi_\alpha^{\mathrm{SC}}(t) = \chi_{[0,\alpha]}(t)$, i.e., $(I - K^*K\Phi_\alpha^{\mathrm{SC}}(K^*K))$ is an orthogonal projection operator. Therefore,

$$\sup_{f\in F_{\Lambda,\overline{w}}}\|(I - K^*K\Phi_\alpha^{\mathrm{SC}}(K^*K))f\| = \sup_{t\in\sigma(K^*K)}(1 - t\Phi_\alpha^{\mathrm{SC}}(t))\Lambda(t)\overline{w} \sim \Lambda(\alpha)\overline{w}, \quad \alpha \searrow 0.$$

The last relation holds since 0 is not an isolated point of the spectrum $\sigma(K^*K)$ for ill-posed operator equations. Using (2.8) and (2.10) we obtain the estimate

$$\begin{aligned}\sup_{f\in F_{\Lambda,\overline{w}}}&\|(I - K^*K\Phi_\alpha(K^*K))f\| \\ &\le \gamma_\Lambda\Lambda(\alpha)\overline{w} \sim \gamma_\Lambda \sup_{f\in F_{\Lambda,\overline{w}}}\|(I - K^*K\Phi_\alpha^{\mathrm{SC}}(K^*K))f\|\end{aligned} \tag{3.4}$$

as $\alpha \searrow 0$. For many regularization methods and smoothness classes we have $\gamma_\Lambda \le 1$.

3.2. Estimation of the integrated variance and rate of convergence of the MISE. The more difficult part is the estimation of the integrated variance of the error $\hat{f}_{\alpha,\sigma} - f$. Under Assumption 1 we have

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2 = \sigma^2\mathbf{E}\,\|\Phi_\alpha(\rho)UK^*\varepsilon\|^2 \le \sigma^2\int_{\mathbb{S}}\Phi_\alpha^2(\rho)\rho\,\mathrm{d}\Sigma. \tag{3.5}$$

A crucial point in the following analysis is the estimation of the tails of the spectral function ρ. To this end, we bound the variance in terms of the function

$$R(\alpha) := \Sigma(\{\rho \geq \alpha\}) \qquad \text{for } \alpha > 0. \tag{3.6}$$

In order to control the MISE of $\hat{f}_{\alpha,\sigma}$ as $\alpha \searrow 0$ it is tempting to assume that R is smooth in a neighborhood around 0. However, this is not true in general. Therefore, we will pose instead that R can be approximated suitably by a smooth function S with properties similar to those of R as $\alpha \searrow 0$. Obviously, R is monotonically decreasing (see (3.8a) below). If $\rho \geq 0$ belongs to $L^1(\Sigma)$, then $-\int_0^\infty \alpha \, \mathrm{d}R(\alpha) = \int_{\mathbb{S}} \rho \, \mathrm{d}\Sigma < \infty$ (see (3.8b)), and it follows from Lebesgue's dominated convergence theorem that $\lim_{\alpha\searrow 0} \alpha R(\alpha) = \lim_{\alpha\searrow 0} \int_{\mathbb{S}} \alpha \, 1_{\{\rho\geq\alpha\}} \, \mathrm{d}\Sigma = 0$ (see (3.8c)).

Assumption 2. There exists a constant $\bar{\alpha} \in (0, \|\rho\|_\infty]$ and a function $S \in C^2((0,\bar{\alpha}])$ such that

$$R(\alpha) \sim S(\alpha), \qquad \text{as } \alpha \searrow 0, \tag{3.7}$$

with R defined by (3.6) in terms of the spectral decomposition (2.1), and S satisfies

$$S' < 0, \tag{3.8a}$$

$$-\alpha S'(\alpha) \text{ is integrable on } (0,\bar{\alpha}], \tag{3.8b}$$

$$\lim_{\alpha\searrow 0} \alpha S(\alpha) = 0, \tag{3.8c}$$

$$\exists \gamma_{\mathrm{S}} \in (0,2) \text{ for all } \alpha \in (0,\bar{\alpha}] : \frac{S''(\alpha)}{-S'(\alpha)} \leq \frac{\gamma_{\mathrm{S}}}{\alpha}. \tag{3.8d}$$

We will show in section 5 for a number of examples that this assumption is satisfied. Now we are in a position to give an estimate of the MISE. The estimate of the MISE in the image space $\mathbb{H}_2$ in (3.10) is needed in the analysis of L^2-boosting (section 5.4) and for nonlinear inverse problems.

THEOREM 3. *Consider the model* (2.11), *and let Assumptions* 1 *and* 2 *hold true. We define a general spectral estimator* $\hat{f}_{\alpha,\sigma}$ *by* (2.4) *and assume that* Φ_α *satisfies* (2.9).

1. *If condition* (2.10) *is satisfied for the function* Λ *defining the smoothness class* $F_{\Lambda,\overline{w},K^*K}$, *then for all* $f \in F_{\Lambda,\overline{w},K^*K}$ *the MISE can be asymptotically bounded by*

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|_{\mathbb{H}_1}^2 \lesssim \left(\gamma_\Lambda \Lambda(\alpha)\overline{w} + \sqrt{\frac{C_2 C_3}{\alpha}}\tau\right)^2 + \frac{(C_2^2 + C_3^2)\sigma^2}{\alpha^2} \int_0^\alpha S(\beta)\,\mathrm{d}\beta, \qquad \text{as } \alpha \searrow 0. \tag{3.9}$$

2. *Assume that* $g \in F_{\tilde{\Lambda},\overline{w},KK^*} \subset \mathbb{H}_2$ *and that* $\tilde{\Lambda}$ *satisfies* (2.10). *(If* $g = Kf$ *with* $f \in F_{\Lambda,\overline{w},K^*K}$, *then* $\tilde{\Lambda}(t) := \sqrt{t}\Lambda(t)$, *but we do* not *assume* $g \in R(K)$ *here!) Then*

$$\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - g\|_{\mathbb{H}_2}^2 \lesssim \left(\gamma_{\tilde{\Lambda}} \tilde{\Lambda}(\alpha)\overline{w} + C_2\tau\right)^2 + \frac{(C_2^2 + C_3^2)\sigma^2}{\alpha^2} \int_0^\alpha \beta S(\beta)\,\mathrm{d}\beta, \qquad \text{as } \alpha \searrow 0. \tag{3.10}$$

Note that for statistical inverse problems, as opposed to deterministic inverse problems, the estimates of the noise term and hence the rates of convergence of the MISE do not depend only on the relative smoothness of the solution (i.e., on Λ), but also on the operator (i.e., on S).

Remark 4. We comment on the choice of the regularization parameter $\alpha > 0$. If the noise levels σ and τ, the spectral properties of K^*K (i.e., S), and the smoothness of f (i.e., Λ) are known, one can choose α by minimizing the right-hand side of (3.9). Since typically the smoothness of the solution is not known a priori, so-called *adaptive* methods must be employed for the selection of α. We do not intend to review the considerable amount of literature on this topic here but want to mention that the explicit bounds on the variance given in Theorem 3 allow the application of the Lepskij balancing principle as proposed for inverse problems by Mathé and Pereverzev [32, 33] and Bauer and Pereverzev [3]. We will discuss this in more detail elsewhere. With this method one typically loses a log factor in the asymptotic rates of convergence. In most cases this can be avoided by using Akaike's method as studied for spectral cut-off and related methods by Cavalier et al. [8]. Unfortunately, Assumption 2 in this paper is not satisfied for the methods discussed here.

3.3. Comparison with spectral cut-off. To show that with an optimal choice of α our estimators can achieve the best possible order of convergence among all estimators as $\sigma \searrow 0$, we compare them to the spectral cut-off estimator for which minimax results are known in many situations (see references in the introduction). Since we are mainly interested in the case that the statistical noise is asymptotically dominant, we will assume that $\tau = 0$ for simplicity. Moreover, we assume in addition to (2.14) that the lower bound

$$\mathbf{Var}\left(UK^*\varepsilon(s)\right) \geq \gamma_{\mathbf{var}}\rho(s) \tag{3.11}$$

holds true for some constant $\gamma_{\mathbf{var}} > 0$. For the white noise model this is satisfied with $\gamma_{\mathbf{var}} = 1$ and for the inverse regression model with $\gamma_{\mathbf{var}} = C_{v,l}/C_1$ (see (4.13)). Moreover, we need the following assumption to prove optimal rates in many mildly ill-posed problems.

Assumption 3. There exists a constant $C_4 > 0$ such that for all $\alpha \in (0, \bar{\alpha}]$

$$\frac{C_4}{\alpha} \leq \frac{-S'(\alpha)}{S(\alpha)}. \tag{3.12}$$

THEOREM 5. *Let Assumptions* 1 *and* 2 *and the lower bound* (3.11) *hold true and assume that the family of functions* $\{\Phi_\alpha\}$ *satisfies* (2.9). *Moreover, assume that either* $S = R$ *or Assumption* 3 *holds true. Then the integrated variance of the estimator* $\hat{f}_{\alpha,\sigma}$ *is bounded by the integrated variance of the spectral cut-off estimator* $\hat{f}^{\mathrm{SC}}_{\alpha,\sigma}$,

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2 \lesssim \frac{C_2^2 + \kappa C_3^2}{\gamma_{\mathbf{var}}}\,\mathbf{E}\,\|\hat{f}^{\mathrm{SC}}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}^{\mathrm{SC}}_{\alpha,\sigma}\|^2, \qquad \text{as } \alpha \searrow 0, \tag{3.13}$$

with C_2 *and* C_3 *as in Theorem* 3 *and* $\kappa := \gamma_{\mathrm{S}}/(2-\gamma_{\mathrm{S}})$, γ_{S} *defined in* (3.8d). *Moreover, if condition* (2.10) *is satisfied for the function* Λ *defining the smoothness class* $F_{\Lambda,\overline{w}}$ *and if* $\tau = 0$, *then there exists a constant* $C > 0$ *such that*

$$\sup_{f\in F_{\Lambda,\overline{w}}} \mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|^2 \leq C \sup_{f\in F_{\Lambda,\overline{w}}} \mathbf{E}\,\|\hat{f}^{\mathrm{SC}}_{\alpha,\sigma} - f\|^2 \tag{3.14}$$

for all $\sigma > 0$ *and all* $\alpha > 0$ *sufficiently small.*

Whereas condition (3.12) is usually satisfied for mildly ill-posed problems, it is not satisfied for exponentially ill-posed problems where $S(\alpha) \sim c(-\ln\alpha)^q$ for constants $c, q > 0$. Nevertheless, the error bounds in Theorem 3 yield optimal rates of convergence in the limit $\sigma \searrow 0$ for logarithmic source conditions after taking the infimum over all α. This is made precise in the following result, which relies on a comparison of the rates for general regularization methods and bounds on the spectral cut-off rates, which are known to be optimal in many situations (see, e.g., Mair and Ruymgaart [30]).

THEOREM 6. *Under the assumptions of Theorem* 3, *Part* 1 *with* $\tau = 0$, *define the increasing functions* $\gamma_1(\alpha) := -\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}R(\beta)$ *and* $\gamma_2(\alpha) := \frac{1}{\alpha^2}\int_0^\alpha S(\beta)\,\mathrm{d}\beta$ *and assume that*

$$\Lambda\left(\overline{\gamma_2}(\gamma_1(\alpha))\right) \lesssim C\Lambda(\alpha), \qquad \text{as } \alpha \searrow 0, \tag{3.15}$$

with the inverse function $\overline{\gamma_2}$ *of* γ_2 *and a constant* $C > 0$. *Then*

$$\inf_{\alpha>0} \mathbf{E}\,\|\hat f_{\alpha,\sigma} - f\|^2 \lesssim \inf_{\alpha>0}\left((C\gamma_\Lambda\Lambda(\alpha)\overline{w})^2 + (C_3^2 + C_2^2)\sigma^2\gamma_1(\alpha)\right), \qquad \text{as } \sigma \searrow 0;$$

i.e., if we choose the optimal value of α *for every noise level* σ, *all spectral regularization methods achieve the same rate of convergence of the MISE as spectral cut-off.*

Assumption (3.15) is satisfied if $\Lambda(t) = (-\ln t)^{-p}$ and $\gamma_1(\alpha) \le \gamma_2(\alpha^2)$ since

$$\Lambda(\overline{\gamma_2}(\gamma_1(\alpha))) \le \Lambda(\alpha^2) = (-2\ln\alpha)^{-p} = 2^{-p}\Lambda(\alpha).$$

4. Noise models satisfying Assumption 1. In this section we show that several commonly used noise models fit into the general framework described in Assumption 1. We start with an (infinite-dimensional) white noise model, and then continue with several models based on finitely many observations.

4.1. White noise. A frequently used model is to assume that ε in (2.11) is a white noise process in $\mathbb{H}_2$ (see, e.g., Donoho [11, 12] and Mathé and Pereverzev [31]). Moreover, we assume that K^*K is a trace-class operator; i.e., it is compact and the eigenvalues ρ_j of K^*K satisfy $\operatorname{tr}(K^*K) := \sum_{j=0}^\infty \rho_j < \infty$. Then $\mathbf{Cov}_{K^*\varepsilon} = K^*K$, so

$$\mathbf{E}\,\|K^*\varepsilon\|^2 = \operatorname{tr}(\mathbf{Cov}_{K^*\varepsilon}) = \operatorname{tr}(K^*K) < \infty.$$

Therefore, $K^*\varepsilon$ can be identified with a Hilbert space–valued random variable. Using the notation introduced in Remark 1 and defining $e_j : \mathbb{N} \to \mathbb{R}$ by $e_j(k) := \delta_{jk}$, $u_j = U^*e_j \in \mathbb{H}_1$ is a unit-length eigenvector of K^*K to the eigenvalue ρ_j, and

$$\mathbf{Var}\,(UK^*\varepsilon(j)) = \mathbf{Var}\,\langle UK^*\varepsilon, e_j\rangle = \mathbf{Var}\,\langle\varepsilon, Ku_j\rangle = \|Ku_j\|^2 = \rho_j$$

for $j = 0, 1, 2, \ldots$. Therefore, (2.14) is satisfied with equality.

4.2. Quasi deconvolution, errors in variable, noncompact operators. Suppose we want to estimate the density f of a random variable Z with values in $\mathbb{R}^d$, but we can observe only a random variable $X = Z + W$ perturbed by a random variable W. Hence, our data are

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} X = Z + W. \tag{4.1}$$

The density g of X is given by

$$g = \int_{\mathbb{R}^d} h(\cdot - z|z) f(z)\,\mathrm{d}z =: Kf, \tag{4.2}$$

where $h(\cdot|z)$ is the conditional density of W given $Z = z$. If Z and W are stochastically independent, K is a convolution operator. The recovering of f is known as the *deconvolution problem* and has been studied extensively (see, e.g., Stefanski and Carroll [38], Fan [17], and Diggle and Hall [10]). Dependent Z and W in (4.1) occur in many scientific applications, e.g., brightness determination of extragalactic star clusters in astrophysics, where the variance σ^2 of the noise W increases monotonically with decreasing brightness of the object Z. Here, a reasonable model is described by $h(y|z) = (2\pi\sigma^2(z))^{-1/2}\exp(-y^2/\sigma^2(z))$ (see Bissantz [4]).

We assume that $f \in L^2(\mathbb{R}^d)$ and that K is a bounded, injective operator in $L^2(\mathbb{R}^d)$. As opposed to the previous section, in general, K is not compact here. Obviously, an unbiased estimator of $q := K^*g$ is given by

$$\hat{q}_n(y) := \frac{1}{n}\sum_{j=1}^{n} h(X_j - y|y). \tag{4.3}$$

To fit this into our general framework, we show that $\hat{q}_n = q + K^*\tilde{\varepsilon}$ for a Hilbert-space process $\tilde{\varepsilon}: L^2(\mathbb{R}^d) \to L^2(\Omega, \mathcal{P}, P)$ defined by

$$\langle \tilde{\varepsilon}, \varphi\rangle := \frac{1}{n}\sum_{j=1}^{n}\varphi(X_j) - \langle g, \varphi\rangle . \tag{4.4}$$

In fact, for $\psi \in L^2(\mathbb{R}^d)$,

$$\langle K^*\tilde{\varepsilon}, \psi\rangle = \langle \tilde{\varepsilon}, K\psi\rangle = \frac{1}{n}\sum_{j=1}^{n}\int_{\mathbb{R}^d} h(X_j - z|z)\psi(z)\,\mathrm{d}z - \langle K^*g, \psi\rangle = \langle \hat{q}_n - q, \psi\rangle .$$

The next result states that Assumption 1 is satisfied.

PROPOSITION 7. *Assume that the operator K defined by* (4.2) *is injective and satisfies $\|K\|_{2,2} < \infty$ and $\|K\|_{2,\infty} < \infty$, where $\|K\|_{r,s}$ is defined as the operator norm of $K: L^r(\mathbb{R}^d) \to L^s(\mathbb{R}^d)$. Moreover, let $\hat{q}_n$ and $\tilde{\varepsilon}$ be defined by* (4.3) *and* (4.4), *and let*

$$\sigma := \frac{1}{\sqrt{n}}\left(\|g\|_{L^\infty} + \|g\|_{L^2}^2\right)^{1/2} \quad \text{and} \quad \varepsilon := \tilde{\varepsilon}/\sigma. \tag{4.5}$$

Then ε satisfies Assumption 1, *and $\hat{q}_n = q + \sigma K^*\varepsilon$.*

Proof. We have to show that (2.12)–(2.14) hold true. Since the X_j are assumed to be independent, it suffices to consider the case $n = 1$. The first part of (2.12), i.e., $\langle \varepsilon, \varphi\rangle = 0$ for $\varphi \in L^2(\mathbb{R}^d)$, follows from $E\varphi(X) = \int \varphi g\,\mathrm{d}x$. Since

$$\mathbf{Cov}(\langle\tilde{\varepsilon}, \varphi_1\rangle, \langle\tilde{\varepsilon}, \varphi_2\rangle) = \int_{\mathbb{R}^d}\varphi_1\varphi_2 g\,\mathrm{d}x - \langle g, \varphi_1\rangle\langle g, \varphi_2\rangle \quad \text{for all } \varphi_1, \varphi_2 \in \mathbb{H}_2,$$

the covariance operator of $\tilde{\varepsilon}$ is given by $\mathbf{Cov}_{\tilde{\varepsilon}} = M_g - g \otimes g$, where M_g means multiplication by g, and $g \otimes g: L^2(\mathbb{R}^d) \to L^2(\mathbb{R}^d)$ is the rank-1 operator defined by $(g \otimes g)\varphi := g\langle\varphi, g\rangle$. Now $\|\mathbf{Cov}_\varepsilon\| \le 1$ follows from the estimate $\|\mathbf{Cov}_{\tilde{\varepsilon}}\| \le \|g\|_{L^\infty} + \|g\|_{L^2}^2$, which completes the proof of (2.12).

To show (2.13), i.e., $\mathbf{E}\,\|\hat{q}_n - q\|^2 < \infty$, note that

$$\mathbf{Cov}_{\hat{q}_1} = K^*\mathbf{Cov}_{\tilde{\varepsilon}}K = K^*M_gK - (K^*g)\otimes(K^*g).$$

We have to show that this is a trace-class operator. Obviously $(K^*g)\otimes(K^*g)$ is trace class as a rank-1 operator. It is not obvious, however, that K^*M_gK is trace class since neither K nor M_g are even compact in general. To show this, we rewrite the kernel of K as $k(x,z) := h(x-z|z)$ and note that $\operatorname{ess\,sup}\|k(x,\cdot)\|_{L^2} = \|K\|_{2,\infty} < \infty$. Since $g \geq 0$, the operator K^*M_gK is self-adjoint and positive semidefinite. Let $\{\varphi_j : j \in \mathbb{N}\}$ be a complete orthonormal system in the separable Hilbert space $L^2(\mathbb{R}^d)$. The B. Levi theorem yields

$$\sum_{j\in\mathbb{N}} \langle \varphi_j, K^*M_gK\varphi_j\rangle = \sum_{j\in\mathbb{N}} \int g(x)|(K\varphi_j)(x)|^2 dx$$

$$= \sum_{j\in\mathbb{N}} \int g(x)|\,\langle k(x,\cdot),\varphi_j\rangle\,|^2 dx \leq \|g\|_{L^1} \operatorname*{ess\,sup}_{x\in\mathbb{X}_2} \|k(x,\cdot)\|_{L^2}^2 < \infty,$$

which implies that K^*M_gK is trace class with $\operatorname{tr}(K^*M_gK) \leq \|K\|_{2,\infty}^2$. Finally, (2.14) follows from Lemma 2. □

If K is a convolution operator with convolution kernel $w(x-z)$, then the canonical choice of the unitary operator U in the Halmos decomposition is the Fourier transform

$$(U\varphi)(\xi) = (\mathcal{F}\varphi)(\xi) = \int_{\mathbb{R}^d} \varphi(x) e^{-2\pi i \xi\cdot x} dx, \tag{4.6}$$

and the multiplier function is then $\rho = |\mathcal{F}w|^2$. In this case the condition (2.14) in Assumption 1 can be verified explicitly; see Mair and Ruymgaart [30].

4.3. Inverse regression. We now review another commonly used noise model (see Wabha [41], O'Sullivan [37], Nychka and Cox [36], and Bissantz, Hohage, and Munk [5]) and show how it is related to the model (2.11). Suppose that $\mathbb{H}_i = L^2(\mu_i)$ are L^2-spaces w.r.t. measure spaces $(\mathbb{X}_i, \mathcal{X}_i, \mu_i)$, $i = 1, 2$, $\mathbb{H}_1$ is separable, and $K : L^2(\mu_1) \to L^2(\mu_2)$ is an integral operator

$$(Kf)(x) := \int_{\mathbb{X}_1} k(x,y) f(y)\,\mathrm{d}\mu_1(y), \qquad x \in \mathbb{X}_2, \tag{4.7}$$

with kernel k. Recall that K^*K is trace class if and only if K is Hilbert–Schmidt and that K is a Hilbert–Schmidt operator if and only if $k \in L^2(\mu_2 \times \mu_1)$ (see Taylor [39]). The latter condition is easy to verify in most applications.

We will assume in the following that the measure space $\mathbb{H}_2$ is finite. Then we can arrange that $\mu_2(\mathbb{X}_2) = 1$. We consider the regression model

$$Y_i = (Kf)(X_i) + \varepsilon_i, \qquad f \in \mathbb{H}_1, \quad i = 1, \dots, n, \tag{4.8}$$

where we assume for simplicity that the random variables $X_i \in \mathbb{X}_2$ have uniform distribution on $\mathbb{X}_2$ (see also Remark 9). Moreover, we assume that $(Y_i, X_i) \sim (Y, X)$, $i = 1, \dots, n$, are independent and identically distributed random variables with values in $\mathbb{R} \times \mathbb{X}_2$ such that

$$\mathbf{E}\,[Y|X] = (Kf)(X), \tag{4.9}$$

and hence $\mathbf{E}\,[\varepsilon|X] = 0$ for $\varepsilon := Y - (Kf)(X)$. Finally we assume that that $v(X) := \sqrt{\mathbf{E}\,[\varepsilon^2|X]}$ satisfies

$$0 < C_{v,l} \leq v(X) \leq C_{v,u} < \infty \qquad \text{a.s.} \tag{4.10}$$

for some constants $C_{v,l}, C_{v,u} > 0$. A straightforward computation shows that

$$\hat{q}_n = \frac{1}{n}\sum_{i=1}^{n} Y_i k(X_i, \cdot) \tag{4.11}$$

is an unbiased estimator of the vector $q := K^*Kf$. To fit the inverse regression model with random design in our general framework, we introduce the Hilbert-space (noise) process $\tilde{\varepsilon} : \mathbb{H}_2 \to L^2(\Omega, \mathcal{P}, P)$ by

$$\langle \tilde{\varepsilon}, \varphi \rangle := \frac{1}{n}\sum_{j=1}^{n} Y_j \varphi(X_j) - \langle g, \varphi \rangle, \qquad \varphi \in \mathbb{H}_2, \tag{4.12}$$

and show that

$$\langle K^*\tilde{\varepsilon}, \psi \rangle = \langle \tilde{\varepsilon}, K\psi \rangle = \frac{1}{n}\sum_{j=1}^{n} Y_j \int_{\mathbb{X}_j} k(X_j, y)\psi(y)\,\mathrm{d}\mu_1(y) - \langle K^*g, \psi \rangle = \langle \hat{q}_n - q, \psi \rangle$$

for all $\psi \in \mathbb{H}_1$, i.e., $\hat{q}_n = q + K^*\tilde{\varepsilon}$.

PROPOSITION 8. *Assume the inverse regression model* (4.7)–(4.10), *and let* $\hat{q}_n$ *and* $\tilde{\varepsilon}$ *be defined by* (4.11) *and* (4.12). *Moreover, let* $K : L^2(\mu_1) \to L^2(\mu_2)$ *be Hilbert–Schmidt, and* $\mu_2 - \operatorname{ess\,sup} \|k(x, \cdot)\|_{L^2(\mu_1)} < \infty$. *Define*

$$\sigma := \sqrt{\frac{C_1}{n}} \qquad \textit{and} \qquad \varepsilon := \tilde{\varepsilon}/\sigma,$$

with $C_1 := C_{v,u} + \|g\|^2_{L^\infty(\mu_2)} + \|g\|^2_{L^2(\mu_2)}$. *Then* ε *satisfies Assumption* 1 *for the unitary transform* U *defined in Remark* 1, *and* $\hat{q}_n = q + \sigma K^*\varepsilon$. *Moreover,*

$$\frac{C_{v,l}}{n}\rho(j) \le \mathbf{Var}\left((U\hat{q}_n)(j)\right), \qquad j = 0, 1, 2, \ldots. \tag{4.13}$$

Proof. It suffices to prove this for $n = 1$. Since X is uniformly distributed and (4.9) holds true, we have

$$\mathbf{E}\left(Y\varphi(X)\right) = \mathbf{E}\left(\mathbf{E}\left[\varepsilon|X\right]\varphi(X)\right) + \mathbf{E}\left(g(X)\varphi(X)\right) = \int g\varphi\,\mathrm{d}\mu_2 = \langle g, \varphi \rangle$$

for all $\varphi \in \mathbb{H}_2$, and hence the first part of (2.12) holds true. Using once more the same properties of X and Y we find that

$$\begin{aligned}
\mathbf{Cov}(\langle \tilde{\varepsilon}, \varphi_1 \rangle, \langle \tilde{\varepsilon}, \varphi_2 \rangle) &= \mathbf{E}\left\{Y^2\varphi_1(X)\varphi_2(X)\right\} - \langle g, \varphi_1 \rangle \langle g, \varphi_2 \rangle \\
&= \mathbf{E}\left\{(\varepsilon^2 + 2\varepsilon g(X) + g(X)^2)\varphi_1(X)\varphi_2(X)\right\} - \langle g, \varphi_1 \rangle \langle g, \varphi_2 \rangle \\
&= \int \varphi_1 \left(v^2 + g^2\right)\varphi_2\,\mathrm{d}\mu_2 - \langle g, \varphi_1 \rangle \langle g, \varphi_2 \rangle
\end{aligned}$$

for all $\varphi_1, \varphi_2 \in \mathbb{H}_2$. Hence, $\mathbf{Cov}_{\tilde{\varepsilon}} = M_{v^2+g^2} - g \otimes g$, where $M_{v^2+g^2}\varphi := (v^2 + g^2) \cdot \varphi$ and $(g \otimes g)\varphi := \langle g, \varphi \rangle\, g$. This implies $\|\mathbf{Cov}_{\tilde{\varepsilon}}\| \le C_1$ and finishes the proof of (2.12). Using the notation of Remark 1, condition (2.13) can be seen as follows:

$$\begin{aligned}
\mathbf{E}\,\|\hat{q}_1 - q\|^2 = \operatorname{tr}\left(\mathbf{Cov}_{\hat{q}_1 - q}\right) &= \sum_{j=0}^{\infty} \langle Ku_j, \mathbf{Cov}_{\tilde{\varepsilon}} Ku_j \rangle \\
&\le C_1 \sum_{j=0}^{\infty} \|Ku_j\|^2 = C_1 \operatorname{tr}(K^*K) < \infty.
\end{aligned}$$

Since

$$\mathbf{Var}\,(U\hat{q}_1)(j) = \langle u_j, \mathbf{Cov}_{\hat{q}_1} u_j\rangle = \langle Ku_j, \mathbf{Cov}_{\tilde{\varepsilon}} Ku_j\rangle \le C_1\|Ku_j\|^2 = C_1\rho_j,$$

we obtain the bound (2.14). The lower bound in (4.13) holds true since the operator $M_g^2 - g\otimes g$ is positive definite as covariance operator of $\tilde{\varepsilon}$ for the case $\varepsilon \equiv 0$. □

In contrast to [30] we do not need the assumption that the singular vectors $u_j \in \mathbb{H}_1$ and $v_j \in \mathbb{H}_2$ in the SVD $Kf = \sum_{j=0}^{\infty} \sqrt{\rho_j}\,\langle f, u_j\rangle_{\mathbb{H}_1} v_j$ are uniformly bounded sequences in $L^\infty(\mu_1)$ and $L^\infty(\mu_2)$, respectively. We require only that $\mu_2 - \operatorname{ess\,sup} \|k(x,\cdot)\|_{L^2(\mu_1)} < \infty$. This condition is often less restrictive and easier to verify.

Remark 9. Generalizations:

1. We can replace $L^2(\mu_1)$ by an arbitrary Hilbert space $\mathbb{H}_1$ (e.g., a Sobolev space) by replacing $k(x,\cdot)$ by $\tilde{k}(x) := \sum_{j=0}^{\infty} \sqrt{\rho}_j v(x) u_j$, $x \in \mathbb{X}_2$. Then (4.7) and (4.11) read $(Kf)(x) = \langle \tilde{k}(x), f\rangle_{\mathbb{H}_1}$ and $\hat{q}_n = \frac{1}{n}\sum_{i=1}^n Y_i \tilde{k}(X_i)$, respectively. Proposition 8 remains valid with literally the same proof if $L^2(\mathbb{X}_1)$ is replaced by $\mathbb{H}_1$.
2. (Deterministic and nonuniform design.) The noise model (2.11) also allows us to treat models of the form (4.8) where the measurement points are either nonuniformly distributed on $\mathbb{X}_2$ or $x_i = x_i^{(n)}$ are deterministic quantities (see, for instance, Nychka and Cox [36] and O'Sullivan [37]). For conditions on the design density see Munk [34].

5. Applications. In this section we discuss how Assumption 2 of our main result (Theorem 5) can be verified for some specific operators A of practical interest.

A remarkable number of interesting inverse problem can be expressed in the form

$$K^*K = \Theta(-\Delta) \tag{5.1}$$

in terms of the Laplace operator Δ on some compact, smooth d-dimensional Riemannian manifold M with a (possibly empty) boundary ∂M. Our first three examples are of this form. Here $\Theta : [0,\infty) \to (0,\infty)$ is a function satisfying $\lim_{\lambda\to\infty}\Theta(\lambda) = 0$. Under the given assumptions the Laplace operator $-\Delta$ defined on $D(-\Delta) := H_0^1(M) \cap H^2(M) \subset L^2(M)$ (i.e., with Dirichlet condition on ∂M) is a positive, self-adjoint operator, which has a complete orthonormal system of eigenvectors u_i in $L^2(M)$ with corresponding eigenvalues λ_i (see, e.g., Taylor [40, Chap. 8.2]). Hence the operator on the right-hand side of (5.1) defined in (2.2) can be written as $\Theta(-\Delta)f = \sum_i \Theta(\lambda_i)\,\langle f, u_i\rangle\, u_i$ for $f \in L^2(M)$. Due to a famous result of Weyl (see Taylor [40, Chap. 8, Thm. 3.1., and Cor. 3.5]), the distribution of the eigenvalues

$$N(\lambda) := \#\{\lambda_i : \lambda_i \le \lambda\}, \qquad \lambda \ge 0,$$

has the asymptotic behavior

$$N(\lambda) \sim c_M \lambda^{d/2}, \qquad c_M := \frac{\operatorname{vol} M}{\Gamma\left(\frac{d}{2}+1\right)(4\pi)^{d/2}} \tag{5.2}$$

as $\lambda \to \infty$, where $\operatorname{vol} M = \int_M 1\,dx$ denotes the volume of M. Under the given assumptions the operator A is compact as operator norm limit of the finite rank operators $\sum_{i=1}^k \Theta(\lambda_i)\,\langle u_i, \cdot\rangle\, u_i$ as $k \to \infty$. Assume that $\Theta(\lambda)$ is monotonically decreasing for $\lambda \ge \lambda_0$ and that $\Theta(\lambda) > \alpha_0 := \Theta(\lambda_0)$ for $\lambda < \lambda_0$. As $\lim_{\lambda\to\infty}\Theta(\lambda) = 0$, the inverse function $\overline{\Theta} : (0,\alpha_0] \to [\lambda_0,\infty)$ satisfies $\lim_{\alpha\searrow 0}\overline{\Theta}(\alpha) = \infty$. If the spectral decomposition of A is chosen as in Remark 1, then the function R defined in (3.6) satisfies

$$R(\alpha) = \#\{\lambda_i : \Theta(\lambda_i) \ge \alpha\} = N\left(\overline{\Theta}(\alpha)\right) \sim c_M\left(\overline{\Theta}(\alpha)\right)^{d/2}, \qquad \text{as } \alpha \searrow 0. \tag{5.3}$$

5.1. Backwards heat equation. We consider the inverse problem of reconstructing the temperature at time $t = 0$ on M from measurements of the temperature at time $t = T$. The forward problem is described by the parabolic equation

$$\begin{aligned} &\partial_t u(x,t) = \Delta u(x,t), && x \in M,\ t \in (0,T), \\ &u(x,t) = 0, && x \in \partial M,\ t \in (0,T], \\ &u(x,0) = f(x), && x \in M, \end{aligned} \tag{5.4}$$

with an initial temperature $f \in L^2(M)$ and the final temperature in $g(x) := u(x,T)$, $x \in M$. We have $g = \exp(-T\Delta)f$, i.e., $K = \exp(-T\Delta) \in L(L^2(M))$ and $K^*K = \exp(-2T\Delta)$. Hence,

$$\Theta(\lambda) = \exp(-2T\lambda)$$

in (5.1). By virtue of (5.3) the condition $R \sim S$ is satisfied for

$$S(\alpha) := c_M \left(-\frac{1}{2T}\ln\alpha\right)^{d/2}.$$

It is easy to check that this function satisfies the conditions (3.8). In particular

$$\frac{S''(\alpha)}{-S'(\alpha)} = \frac{1}{\alpha}\left(1 - \frac{d-2}{2}\frac{1}{\ln\alpha}\right), \tag{5.5}$$

so (3.8d) holds with any $\gamma_S \in (1,2)$ for sufficiently small $\bar\alpha$ if $d \geq 3$ and with $\gamma_S = 1$ for all $\bar\alpha < 1$ for $d \leq 2$.

If M is a compact Riemannian manifold without boundary, then the smoothness class $F_{\Lambda,1}$ for a logarithmic source condition (2.7) is the unit ball in the Sobolev space $H^{2p}(M)$ w.r.t. some equivalent norm (see Hohage [23]). Similar results hold true if M has a boundary with a Dirichlet or Neumann condition. In this case we additionally need to impose boundary conditions. Hence, if the initial temperature is bounded in some Sobolev norm, $\|f\|_{H^s} \leq C$, $s = 2p > 0$, if the regularization parameter is chosen such that $\alpha \asymp \sigma$, and if $\tau = O(\sigma^\mu)$ with $\mu > \frac{1}{2}$ as $\sigma \searrow 0$, then it follows from Theorem 3 after some elementary computations that the MISE decays like

$$\mathbf{E}\,\|\hat f_{\alpha,\sigma} - f\|_{L^2}^2 = \mathrm{O}\left((-\ln\sigma)^{-s}\right), \qquad \text{as } \sigma \searrow 0,$$

for all regularization methods satisfying (2.9).

5.2. Satellite gradiometry. In satellite gradiometry measurements of the gravitational force of the earth at a distance a from the center are used to reconstruct the gravitational potential u at the surface of the earth (see Hohage [23], Bauer and Pereverzev [3], and the references therein). Let the earth be described by $E := \{x : |x| < 1\}$, and let $M := \partial E$ denote the surface of the earth. Then u satisfies the Laplace equation

$$\Delta u(x) = 0, \qquad x \in \mathbb{R}^3 \setminus \overline{E},$$

and decays like $|u(x)| = O\left(|x|^{-1}\right)$ as $|x| \to \infty$. The available data consist of noisy measurements of the rate of change of the gravitational force $-\nabla u$ in radial direction $r = |x|$, i.e.,

$$g(x) := \frac{\partial^2 u}{\partial r^2}(x) \qquad \text{for } |x| = a.$$

A discussion of the measurement errors shows that they are mainly random in nature (see [3]). The problem is to determine the potential $f = u\big|_M$ at the surface M of the earth. Introducing the operator $K : L^2(M) \to L^2(aM)$ mapping the solution f to the data g, we can write K^*K in the form (5.1) with $\Theta(\lambda) = \Phi(\Lambda(\lambda))$ and

$$\Phi(t) := c\left(\frac{1}{2}+t\right)^2\left(\frac{3}{2}+t\right)^2 a^{-2t}, \qquad \Lambda(\lambda) := \sqrt{\lambda+\frac{1}{2}}$$

(see Hohage [23]). It is easy to show that $\Phi(t)$ is decreasing for sufficiently large t and that $\Lambda(\lambda)$ is monotonic increasing for all $\lambda > 0$. Obviously, $\overline{\Theta}(\alpha) = \overline{\Lambda}\left(\overline{\Phi}(\alpha)\right) = \overline{\Phi}(\alpha)^2 - \frac{1}{2}$. The function $\overline{\Phi}(\alpha)$ cannot be computed explicitly, but we can estimate its asymptotic behavior as $\alpha \searrow 0$. Writing $t = \overline{\Phi}(\alpha)$ for α sufficiently small and $p(t) := c\left(\frac{1}{2}+t\right)^2\left(\frac{3}{2}+t\right)^2$, we obtain

$$\begin{aligned}\overline{\Phi}(\alpha) &= -\log_a \alpha \frac{t}{-\log_a \alpha} = -\log_a \alpha \frac{t}{-\log_a\left(p(t)a^{-2t}\right)} \\ &= -\log_a \alpha \left(\frac{t}{-\log_a(p(t)) + 2t}\right) \sim -\frac{\ln \alpha}{2\ln a},\end{aligned}$$

as $\alpha \searrow 0$. Therefore, using (5.3), we get

$$R(\alpha) = c_M \overline{\Theta}(\alpha) \sim \left(-\frac{\ln \alpha}{2\ln a}\right)^2, \qquad \text{as } \alpha \searrow 0.$$

The function $S(\alpha) := \left(-\frac{\ln \alpha}{2\ln a}\right)^2$ satisfies the conditions (3.8) (see (5.5)). Moreover, the smoothness classes $F_{\Lambda,1}$ for logarithmic source conditions (2.7) are unit balls in the Sobolev spaces $H^p(M)$ w.r.t. equivalent norms (see Hohage [23]). Since the gravitational potential satisfies the Poisson equation $\Delta u = -\phi$ in $\mathbb{R}^3$ and since the mass density ϕ of the earth E belongs to $L^2(E)$, it follows from elliptic regularity results that $u \in H^2(E)$, so $f = u|_M \in H^{3/2}(M)$ in the sense of the trace operator (see, e.g., Taylor [39]). Therefore,

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - f\|_{L^2}^2 = \mathrm{O}\left((-\ln \sigma)^{-3}\right), \qquad \text{as } \sigma \searrow 0,$$

if $\tau = O(\sigma)$ and if we choose $\alpha \asymp \sigma$.

5.3. Operators in Hilbert scales. In the following we show that our assumptions are satisfied for operators acting in Hilbert scales (see Mair and Ruymgaart [30] and Mathé and Pereverzev [31]). Hence, spectral regularization methods yield optimal rates of convergence for this class of operators.

Let $L : D(L) \subset H \to H$ be an unbounded, positive, self-adjoint operator defined on a dense domain $D(L) \subset H$, and assume the inverse $L^{-1} : H \to H$ is bounded. Then L generates a scale of Hilbert spaces H_μ, $\mu \in \mathbb{R}$, defined as completion of $\bigcap_{n\in\mathbb{N}} D(L^n)$ under the norm generated by the inner product $\langle f, g\rangle_\mu := \langle L^\mu f, L^\mu g\rangle$. We have $H_\mu \subset H_\lambda$ for $\mu, \lambda \in \mathbb{R}$ with $\mu > \lambda$. A prototype is $L = \sqrt{I - \Delta}$ with the Laplace operator Δ on a closed manifold M, which leads to the usual Sobolev spaces on M.

We assume that K is a-times smoothing ($a > 0$) in (part of) the Hilbert scale (H_μ), i.e., $K : H_{\mu-a} \to H_\mu$ is a bounded operator for all $\mu \in [\underline{\mu}, \overline{\mu}]$ which has a bounded inverse $K^{-1} : H_\mu \to H_{\mu-a}$. This is equivalent to $\frac{1}{C_\mu}\|f\|_{\mu-a} \le \|Kf\|_\mu \le$

$C_\mu \|f\|_{\mu-a}$ for all $f \in H_{\mu-a}$ and some constants $C_\mu \geq 1$. Such conditions are satisfied for many boundary integral operators, multiplication operators, convolution operators, and compositions of such operators (see also the discussion after (5.10)). We do not assume here that K is self-adjoint or that K^*K and L commute, i.e., that they can be diagonalized by the same unitary operator U.

Usually the nature of the noise dictates the choice $\mathbb{H}_2 = H_0$, and one is interested in error bounds for the estimator in the positive norm, i.e., $\mathbb{H}_1 = H_{\mu-a}$ for $\mu \geq a$. Then the operator equation $Kf = g$ is ill-posed with $K = K_{0\leftarrow\mu-a}$ considered as an operator from $H_{\mu-a}$ to H_0.

To verify Assumptions 2 and 3 with $R \sim S$ in (3.7) replaced by $R \asymp S$ (see Remark 14), we assume that L has a complete orthonormal system of eigenvectors with eigenvalues $0 < \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \cdots$ tending to infinity. Then the embedding operator $J : H_\mu \hookrightarrow H_0$ is compact, and its singular values are given by $\sigma_j(J) = \lambda_j^{-\mu}$. It follows from the decomposition $K_{0\leftarrow\mu-a} = JK_{\mu\leftarrow\mu-a}$ and the Courant minimax characterization of the singular values $\sigma_j = \sigma_j(K_{0\leftarrow\mu-a})$ (see, e.g., Kress [28]) that

$$\frac{1}{\|K^{-1}\|_{\mu-a\leftarrow\mu}} \lambda_j^{-\mu} \leq \sigma_j(K_{0\leftarrow\mu-a}) \leq \|K\|_{\mu\leftarrow\mu-a} \lambda_j^{-\mu}, \qquad j = 0, 1, \ldots.$$

Hence if $N(\lambda) := \#\{\lambda_j : \lambda_j \leq \lambda\}$ and $C := \max(\|K\|_{\mu\leftarrow\mu-a}, \|K^{-1}\|_{\mu-a\leftarrow\mu})$, then $R(\alpha) := \{\sigma_j : \sigma_j^2 \geq \alpha\}$ satisfies

$$N\left((\alpha/C^2)^{-1/2\mu}\right) \leq R(\alpha) \leq N\left((C^2\alpha)^{-1/2\mu}\right).$$

If the counting function has the asymptotic behavior $N(\lambda) \asymp \lambda^d$ for some $d > 0$, then $R(\alpha) \asymp \alpha^{-d/2\mu}$. For the case $L = \sqrt{I - \Delta}$, d is the space dimension (see (5.2)). A straightforward computation shows that $S(\alpha) := \alpha^{-d/2\mu}$ satisfies (3.8) and (3.12) in Assumptions 2 and 3 if and only if $d/(2\mu) \in (0, 1)$. Under this condition, it follows from Remark 14 that Theorems 3 and 5 hold true with different constants.

It remains to discuss the Hölder-type source conditions (2.6) in this setting. To do this we assume for simplicity that $\mathbb{H}_1 = \mathbb{H}_2 = H_0$. Let K^* denote the adjoint of K w.r.t. the inner product in H_0. It is easy to show that $K^* : H_{-\mu} \to H_{-\mu+a}$ is bounded and boundedly invertible for all $\mu \in [\underline{\mu}, \overline{\mu}]$. Let $l \in \mathbb{N}$ such that $[-2al + 1, 2al - 1] \subset [\underline{\mu}, \overline{\mu}]$. Then there exists a constant $\gamma \geq 1$ such that

$$\gamma^{-1} \|L^{2al} f\|_{H_0} \leq \|(K^*K)^{-l} f\|_{H_0} \leq \gamma \|L^{2al} f\|_{H_0}$$

for all $f \in H_{2al}$. It follows from the Heinz inequality (see Engl, Hanke, and Neubauer [14] and Heinz [22]) that

$$\gamma^{-\sigma} \|L^{2a\sigma l} f\|_{H_0} \leq \|(K^*K)^{-\sigma l} f\|_{H_0} \leq \gamma^\sigma \|L^{2a\sigma l} f\|_{H_0}$$

for all $\sigma \in [0, 1]$ and $f \in H_{2a\sigma l}$. Therefore, the source condition $f = (K^*K)^\nu w$, $w \in H_0$, is equivalent to $f \in H_{2a\nu}$. Let $u := 2a\nu$ and $f \in H_u$. Then

$$\mathbf{E}\, \|\hat{f}_{\alpha,\sigma} - f\|_{H_0}^2 = \mathrm{O}\left(\sigma^{\frac{2u}{u+a+d/2}}\right), \qquad \text{as } \sigma \searrow 0,$$

for the choice $\alpha \asymp \sigma^{\frac{2a}{u+a+d/2}}$ if $\tau = \mathrm{O}\left(\sigma^{\frac{u+a}{u+a+d/2}}\right)$ and $\mu_0 \geq u/2a$.

5.4. L^2-boosting. Boosting algorithms include a large class of iterative procedures which improve stagewise the performance of estimators. They have attracted significant interest in the context of machine learning and more recently in statistics (see Freund and Schapire [18] or Friedman [19], among many others). One of the main challenges is to provide a proper convergence analysis and proper stopping rules for the iteration depth (see Zhang and Yu [43]). L^2-boosting has been introduced in the context of regression analysis by Bühlmann and Yu [7] for classification and more general learning problems. We consider the inverse regression problem described in section 4.3 if K is an embedding operator and $\mathbb{X}_2$ is a d-dimensional smooth, compact Riemannian manifold (e.g., a smooth compact subset of $\mathbb{R}^d$). Consider a *weak learner* of the form

$$\hat{f}_{0,n} = \frac{1}{n}\sum_{j=1}^{n} Y_i k(y, X_j), \tag{5.6}$$

with a continuous, symmetric kernel $k : \mathbb{X}_2 \times \mathbb{X}_2 \to \mathbb{R}$ such that the integral operator $\tilde{K} : L^2(\mathbb{X}_2) \to L^2(\mathbb{X}_2)$ with kernel k is compact and strictly positive definite with eigenvalues $\kappa_0 \geq \kappa_1 \geq \cdots$ and satisfies

$$\operatorname{ess\,sup}_{x\in\mathbb{X}_2} k(x,x) < \infty \qquad \text{and} \qquad \#\{\kappa_j \geq \alpha\} \asymp \alpha^{-d/(2\mu_0)}, \quad \text{as } \alpha \to 0, \tag{5.7}$$

for some $\mu_0 > 0$. Further, let H_μ, $\mu \in \mathbb{R}$ be the Hilbert scale generated by the operator $L := \tilde{K}^{-1/(2\mu_0)}$ as described in section 5.3. If we set $\mathbb{H}_1 := H_{\mu_0}$ and $\mathbb{H}_2 = H_0 = L^2(\mathbb{X}_2)$, then $\mathbb{H}_1 \subset \mathbb{H}_2$, and the adjoint of the embedding operator $K : \mathbb{H}_1 \hookrightarrow \mathbb{H}_2$ is given by $K^*\varphi = \tilde{K}\varphi$ since $\langle \varphi, \tilde{K}\psi\rangle_{\mathbb{H}_1} = \langle L^{\mu_0}\varphi, L^{\mu_0}\tilde{K}\psi\rangle_{L^2} = \langle \varphi, \psi\rangle_{L^2}$ for all $\psi \in L^2(\mathbb{X}_2)$ and all $\varphi \in \mathbb{H}_1$. By a similar reasoning one can show that $\mathbb{H}_1$ is a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k(\cdot, x)$. A typical example of a weak learner is a spline smoother which leads to Sobolev spaces H_μ (see [7]).

Note that the weak learner (5.6) can be abbreviated as $\hat{f}_{0,n} = K^*Y$. Boosting this learner results in a recursive iteration,

$$\hat{f}_{j+1,n} = \hat{f}_{j,n} - \beta K^*(Y - K\hat{f}_{j,n}), \qquad j = 0, 1, 2, \ldots, \tag{5.8}$$

which is in fact a Landweber iteration (see section 2.2). Hence, Theorem 3 gives the following bound.

COROLLARY 10. *Assume that k satisfies* (5.7) *with $\mu_0 > \frac{d}{2}$, let $g \in H_\mu$ with $\mu > 0$, and let $\beta \in (0, \|KK^*\|^2]$. Then the MISE is bounded by*

$$\mathbf{E}\,\|\hat{f}_{j,n} - g\|^2_{L^2(\mathbb{X}_2)} \leq C\left((j+1)^{-\mu/\mu_0} + n^{-1}(j+1)^{d/(2\mu_0)}\right). \tag{5.9}$$

For the optimal stopping index $j_*(n) \asymp n^{2\mu_0/(2\mu+d)}$ we obtain the rate $\mathbf{E}\,\|\hat{f}_{j_*(n),n} - f\|^2_{L^2(\mathbb{X}_2)} \leq Cn^{-2\mu/(2\mu+d)}$, which is the well-known minimax rate in the case of Sobolev spaces.

Proof. It follows easily from the definitions that $g \in H_\mu$ is equivalent to $g \in F_{\tilde{\Lambda},\overline{w}}$ with $\tilde{\Lambda}(t) = t^{\mu/2\mu_0}$ for some $\overline{w} > 0$. Since Landweber iteration has infinite qualification (see [14]), $\tilde{\Lambda}$ satisfies (2.10). Moreover, as the singular values of K are $\sigma_j(K) = \sqrt{\kappa_j}$, (5.7) implies that $R(\alpha) = \#\{\sigma_j(K)^2 \geq \alpha\} \asymp S(\alpha)$ with $S(\alpha) := \alpha^{-d/2\mu_0}$, and S satisfies (3.8) in Assumption 2 for $\mu_0 > \frac{d}{2}$. To verify the assumptions of Proposition 8, we note that $\operatorname{tr}(K^*K) = -\int_0^\infty \alpha\,\mathrm{d}R(\alpha) < \infty$ for $\mu_0 > \frac{d}{2}$ (i.e., K is Hilbert–Schmidt) and that $\operatorname{ess\,sup}_{x\in\mathbb{X}_2}\|k(x,\cdot)\|_{\mathbb{H}_1} = \operatorname{ess\,sup}_{x\in\mathbb{X}_2}\sqrt{\langle k(x,\cdot), k(x,\cdot)\rangle_{\mathbb{H}_1}} =$

$\operatorname{ess\,sup}_{x\in\mathbb{X}_2}\sqrt{k(x,x)} < \infty$. Therefore, Proposition 8 and Remark 9, Part 1 imply that Assumption 1 is satisfied with $\sigma \asymp n^{-1/2}$. Hence (5.9) follows from (3.10) in Theorem 3 with $\alpha = (j+1)^{-1}$ and Remark 14. □

Corollary 10 immediately applies to all other regularization methods covered by Theorem 3. In particular, ν-methods require only the square root of the number of Landweber iterations to achieve the optimal rate, but they seem to be unknown in statistics and machine learning.

Often a discretized sample variant of the iteration (5.8) is considered. Convergence of this algorithm has been analyzed by Yao, Rosasco, and Caponnetto [42], but without optimal rates. It is still an open problem whether this discretized version achieves the minimax rates of Corollary 10 in the general context of RKHS, as was shown in [7] for the particular case of spline learning.

5.5. Errors in variable problems. We now further discuss the errors in variable problems introduced in section 4.2. Our aim is to establish rates of convergence of estimators of the density f of $Z \in \mathbb{R}^d$ as the sample size n tends to infinity. Therefore, with a slight abuse of notation, we will write $\hat{f}_{\alpha,n} = \hat{f}_{\alpha,\sigma(n,g)}$ in this context. It follows from the definition (4.5) of σ and the boundedness of $\|\Lambda(K^*K)\|_{2,2}$ that

$$\begin{aligned}\sup_{f\in F_{\Lambda,\overline{w}}} \sigma(n,Kf) &= \sup_{\|w\|=\overline{w}} \sigma(n,K\Lambda(K^*K)w)\\ &\le \frac{\overline{w}}{\sqrt{n}}\left(\|K\Lambda(K^*K)\|_{2,\infty}^2 + \|K\Lambda(K^*K)\|_{2,2}^2\right),\end{aligned}$$

where the expression in parenthesis is finite under the assumptions of Proposition 7.

We first consider two important special cases

$$h_1(y|z) = w_1(y) := \exp(-\pi\|y\|_2^2), \qquad h_2(y|z) = w_2(y) := c_d\exp(-\|y\|_2), \quad y,z\in\mathbb{R}^d,$$

with normalization constant $c_d := \pi^{-d/2}\Gamma(d/2+1)/\Gamma(d+1)$ corresponding to an error variable W independent of Z. Here K is a convolution operator, the canonical unitary transformation U in the spectral decomposition is the Fourier transform $\mathcal{F}$ defined in (4.6), and the multiplier function is $\rho_j = |\mathcal{F}w_j|^2$, i.e., $\rho_1(\xi) = \exp(-2\pi\|\xi\|_2^2)$, and $\rho_2(\xi) = (1+4\pi^2\|\xi\|^2)^{-d-1}$. Hence, the corresponding functions R are given by

$$R_1(\alpha) = V_d\left(-\frac{1}{2\pi}\ln\alpha\right)^{d/2}, \quad R_2(\alpha) = V_d(2\pi)^{-d}\left(\alpha^{-1/(d+1)}-1\right)^{d/2}, \quad 0<\alpha<1,$$

where V_d denotes the volume of the unit ball in $\mathbb{R}^d$. Hence, Assumption 2 is satisfied for R_1 with $S = R_1$ (see (5.5)) and for R_2 with $S(\alpha) = V_d(2\pi)^{-d}\alpha^{-d/(2d+2)}$. Since the norm of the Sobolev space $H^s(\mathbb{R}^d)$ is defined by $\|f\|_{H^s(\mathbb{R}^d)} = (\int(1+|\xi|^2)^s|\mathcal{F}f(\xi)|^2\,\mathrm{d}\xi)^{1/2}$, a simple computation shows that in the first case a logarithmic source condition (2.7) is equivalent to $f\in H^{2p}(\mathbb{R}^d)$, and in the second case a Hölder-type source condition (2.6) is equivalent to $f\in H^{2(d+1)\nu}(\mathbb{R}^d)$. Suppose that $f\in H^s(\mathbb{R}^d)$. Then we find in the first case for the choice $\alpha\asymp n^{-1/2}$ the asymptotic rates

$$\mathbf{E}\,\|\hat{f}_{\alpha,n}-f\|_{L^2}^2 = \mathrm{O}\left((\ln n)^{-s}\right), \qquad \text{as } n\to\infty,$$

and in the second case the rate

$$\mathbf{E}\,\|\hat{f}_{\alpha,n}-f\|_{L^2}^2 = \mathrm{O}\left(n^{-\frac{s}{s+3d/2+1}}\right), \qquad \text{as } n\to\infty,$$

for the choice $\alpha \asymp n^{-\frac{d+1}{s+3d/2+1}}$. This generalizes results in Mair and Ruymgaart [30] for spectral cut-off to arbitrary regularization methods and to the multivariate setting.

We now consider the case that the random variables Z and W are not stochastically independent. We assume that the conditional density h is of the form

$$h(x-z|z) = w(x-z) + p(x,z), \tag{5.10}$$

where $\underline{c}(1+\|\xi\|_2^2)^{-a} \le |\mathcal{F}w(\xi)|^2 \le \overline{c}(1+\|\xi\|_2^2)^{-a}$ for some constants $a, \underline{c}, \overline{c} > 0$, and p is C^∞-smooth and decays exponentially as $\|x\|, \|z\| \to \infty$. Then the convolution operator $\tilde{K}$ with kernel w is bounded and boundedly invertible from $H^{\mu-a}(\mathbb{R}^d)$ to $H^\mu(\mathbb{R}^d)$ for all $\mu \in \mathbb{R}$, and the integral operator P with kernel p is compact from $H^{\mu-a}(\mathbb{R}^d)$ to $H^\mu(\mathbb{R}^d)$ for all $\mu \in \mathbb{R}$. Under our general assumption that $K = \tilde{K} + P$ is injective, it follows from Riesz theory that $K : H^{\mu-a}(\mathbb{R}^d) \to H^\mu(\mathbb{R}^d)$ has a bounded inverse. Hence, it follows from the arguments of the previous paragraph that Hölder source condition (2.6) for K is equivalent to $f \in H^{2a\nu}(\mathbb{R}^d)$. If we additionally assume periodicity of w and p with arbitrary size of the periodicity interval, then it follows from our results on operators in Hilbert scales that also Assumptions 1 and 2 are satisfied.

6. Appendix: Proofs and auxiliary results. This section contains the proofs of our main results on the MISE. First, we require some technical lemmas.

LEMMA 11. *If Assumption* 1 *holds true and the family of functions* $\{\Phi_\alpha\}$ *satisfies* (2.9), *then*

$$\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2 \le -\frac{(\sigma C_3)^2}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta) - (\sigma C_2)^2\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta), \tag{6.1a}$$

$$\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - \mathbf{E}\,K\hat{f}_{\alpha,\sigma}\|^2 \le (\sigma C_2)^2 R(\alpha) - \frac{(\sigma C_3)^2}{\alpha^2}\int_0^\alpha \beta^2\,\mathrm{d}R(\beta). \tag{6.1b}$$

(Recall that R is decreasing, i.e., the right-hand sides of the inequalities above are nonnegative.)

Proof. Recall the bound (3.5) on $\mathbf{E}\,\|\hat{f}_{\alpha,\sigma} - \mathbf{E}\,\hat{f}_{\alpha,\sigma}\|^2$. We split the integral on the right-hand side of (3.5) of the variance over the "frequency domain" $\mathbb{S}$ into low frequency components $\{\rho \ge \alpha\}$ and high frequency components $\{\rho < \alpha\}$. The low frequency components are bounded by

$$\int_{\{\rho\ge\alpha\}} \Phi_\alpha(\rho)^2\rho\,\mathrm{d}\Sigma \le C_2^2\int_{\{\rho\ge\alpha\}}\frac{1}{\rho}\,\mathrm{d}\Sigma = -C_2^2\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta),$$

where the latter equality holds by a transformation of the integral on the left-hand side to an integral w.r.t. the image measure Σ^ρ, and subsequent reformulation as the Lebesgue–Stieltjes integral given on the right-hand side of the equation. Similarly, the high frequency components of the variance can be estimated by

$$\int_{\{\rho<\alpha\}} \Phi_\alpha(\rho)^2\rho\,\mathrm{d}\Sigma \le \frac{C_3^2}{\alpha^2}\int_{\{\rho<\alpha\}}\rho\,\mathrm{d}\Sigma = -\frac{C_3^2}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta)$$

using (2.9b). This completes the proof of (6.1a).

In analogy to (3.5) we have $\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - \mathbf{E}\,K\hat{f}_{\alpha,\sigma}\|^2 \le \sigma^2\int_{\mathbb{S}}\Phi_\alpha(\rho)^2\rho^2\,\mathrm{d}\Sigma$, and the right-hand side of this inequality can be estimated as above to establish the bound (6.1b). ∎

The next lemma shows that for $R = S$ the high frequency components of the variance are asymptotically bounded by low frequency components and that the relative magnitude of these components is determined by the constant γ_S in (3.8d).

LEMMA 12. *Assume that* $S \in C^2((0,\bar\alpha])$ *satisfies* (3.8), *and define* $\kappa := \frac{\gamma_S}{2-\gamma_S}$, *i.e.*, $\frac{2\kappa}{\kappa+1} = \gamma_S$. *Then*

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta \,\mathrm{d}S(\beta) \le -\kappa \int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}S(\beta) - \frac{\kappa+1}{2}S'(\bar\alpha), \qquad \alpha \in (0,\bar\alpha]. \tag{6.2}$$

Proof. We rewrite (3.8d) as $(\kappa+1)S''(\alpha) \le 2\kappa\frac{-S'(\alpha)}{\alpha}$. Integrating this inequality from α to $\bar\alpha$ yields $(\kappa+1)(S'(\bar\alpha) - S'(\alpha)) \le 2\kappa\int_\alpha^{\bar\alpha} \frac{-S'(\beta)}{\beta}\,\mathrm{d}\beta$, or equivalently

$$0 \le \alpha S'(\alpha) + \kappa\alpha S'(\alpha) + 2\kappa\alpha \int_\alpha^{\bar\alpha} \frac{-S'(\beta)}{\beta}\,\mathrm{d}\beta - \alpha(\kappa+1)S'(\bar\alpha). \tag{6.3}$$

It follows that

$$0 \le \int_0^\alpha \beta\,\mathrm{d}S(\beta) - \kappa\alpha^2 \int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}S(\beta) - \frac{\alpha^2}{2}(\kappa+1)S'(\bar\alpha), \qquad \alpha \in (0,\bar\alpha]. \tag{6.4}$$

To verify this we check that the derivative of the right-hand side of (6.4) is the right-hand side of (6.3) and that the limit of the right-hand side of (6.4) as $\alpha \searrow 0$ is nonnegative by assumptions (3.8a) and (3.8b). Inequality (6.4) is equivalent to (6.2). ∎

Next we show under an additional assumption that the asymptotic balance between high and low frequency components of the variance also holds true if R is not smooth.

LEMMA 13. *If Assumption* 2 *holds true, then for* $j \in \{1,2\}$

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta^j\,\mathrm{d}S(\beta) \le \frac{1}{\alpha^2}\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta, \tag{6.5a}$$

$$\left|\frac{1}{\alpha^2}\int_0^\alpha \beta^j\,\mathrm{d}(R-S)(\beta)\right| = \mathrm{o}\left(\frac{1}{\alpha^2}\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta\right), \tag{6.5b}$$

$$-\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}S(\beta) \le \frac{1}{\alpha}S(\alpha), \tag{6.5c}$$

$$\left|\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}(R-S)(\beta)\right| = \mathrm{o}\left(\frac{1}{\alpha}S(\alpha)\right), \tag{6.5d}$$

as $\alpha \searrow 0$. *If additionally Assumption* 3 *is satisfied, then*

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta^j\,\mathrm{d}R(\beta) \sim -\frac{1}{\alpha^2}\int_0^\alpha \beta^j\,\mathrm{d}S(\beta), \tag{6.6a}$$

$$-\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}R(\beta) \sim -\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}S(\beta). \tag{6.6b}$$

Proof. Using (3.8c), a partial integration yields

$$-\int_0^\alpha \beta^j\,\mathrm{d}T(\beta) = -\alpha^j T(\alpha) + \int_0^\alpha j\beta^{j-1}T(\beta)\,\mathrm{d}\beta \qquad \text{for } T = S \text{ and } T = R-S. \tag{6.7}$$

Due to assumption (3.7) and (3.8b), the left-hand side of (6.7), and hence $\int_0^\alpha j\beta^{j-1} T(\beta)\,\mathrm{d}\beta$, is finite. Inequality (6.5a) follows from (6.7) with $T = S$ since $R(\alpha)$, and hence $S(\alpha)$, are positive for small α. By assumption (3.7), there exists for all $\epsilon > 0$ a $\delta = \delta(\epsilon) > 0$ such that

$$|R(\alpha) - S(\alpha)| \le \epsilon S(\alpha) \qquad \text{for } \alpha < \delta. \tag{6.8}$$

Therefore, using (6.7) with $T = S - R$,

$$\left| \int_0^\alpha \beta^j \,\mathrm{d}(S(\beta) - R(\beta)) \right| \le \epsilon \alpha^j S(\alpha) + \epsilon \int_0^\alpha j\beta^{j-1} S(\beta)\,\mathrm{d}\beta$$

for $\alpha < \delta$. As $\alpha^j S(\alpha) = \int_0^\alpha j\beta^{j-1} S(\alpha)\,\mathrm{d}\beta \le \int_0^\alpha j\beta^{j-1} S(\beta)\,\mathrm{d}\beta$ due to (3.8a), we obtain (6.5b).

To prove (6.5c) and (6.5d), again partial integration yields for $T = S$ or $T = R - S$

$$-\int_\alpha^{\bar\alpha} \frac{1}{\beta}\,\mathrm{d}T(\beta) = \frac{1}{\alpha} T(\alpha) - \frac{1}{\bar\alpha} T(\bar\alpha) - \int_\alpha^{\bar\alpha} \frac{1}{\beta^2} T(\beta)\,\mathrm{d}\beta. \tag{6.9}$$

For $T = S$ this yields (6.5c). Let $\epsilon > 0$ and choose $\delta_1 := \delta(\epsilon)$ according to (6.8) and $\delta_2 := \delta_1 \epsilon$. Then

$$\left| \int_\alpha^{\bar\alpha} \frac{R(\beta) - S(\beta)}{\beta^2}\,\mathrm{d}\beta \right| \le \epsilon \int_\alpha^{\delta_1} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta + \int_{\delta_1}^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta + \int_{\delta_1}^{\bar\alpha} \frac{R(\beta)}{\beta^2}\,\mathrm{d}\beta$$

for $\alpha \le \delta_2$. Due to the monotonicity of S we have

$$\int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \ge \int_{\delta_2}^{\delta_1} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \ge S(\delta_1) \int_{\delta_2}^{\delta_1} \frac{\mathrm{d}\beta}{\beta^2} = S(\delta_1) \left(\frac{1}{\delta_2} - \frac{1}{\delta_1} \right) = \frac{1-\epsilon}{\epsilon} \frac{S(\delta_1)}{\delta_1},$$

so

$$\int_{\delta_1}^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \le S(\delta_1) \int_{\delta_1}^{\infty} \frac{\mathrm{d}\beta}{\beta^2} = \frac{S(\delta_1)}{\delta_1} \le \frac{\epsilon}{1-\epsilon} \int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta,$$

$$\int_{\delta_1}^{\bar\alpha} \frac{R(\beta)}{\beta^2}\,\mathrm{d}\beta \le \frac{1}{\delta_1} R(\delta_1) \le (1+\epsilon) \frac{S(\delta_1)}{\delta_1} \le \epsilon \frac{1+\epsilon}{1-\epsilon} \int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta.$$

Since $S(\alpha) > 0$ for all $\alpha \in (0, \bar\alpha]$ due to (3.12) we can extend the integrals over $[\delta_2, \delta_1]$ and $[\alpha, \delta_1]$ to $[\alpha, \bar\alpha]$ and obtain

$$\left| \int_\alpha^{\bar\alpha} \frac{R(\beta) - S(\beta)}{\beta^2}\,\mathrm{d}\beta \right| \le \epsilon \left(1 + \frac{1}{1-\epsilon} + \frac{1+\epsilon}{1-\epsilon} \right) \int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta$$

for $\epsilon < 1$ and $\alpha \le \delta_2$. Since $\int_\alpha^{\bar\alpha} \frac{S(\beta)}{\beta^2}\,\mathrm{d}\beta \le S(\alpha)/\alpha - S(\bar\alpha)/\bar\alpha \lesssim S(\alpha)/\alpha$ due to (6.9) and (3.8a), we obtain (6.5d).

Assume now that Assumption 3 holds true. Written as $-\alpha^j S'(\alpha) \ge C_4 \alpha^{j-1} S(\alpha)$ and integrated from 0 to α, (3.12) yields

$$-\int_0^\alpha \beta^j \,\mathrm{d}S(\beta) \ge C_4 \int_0^\alpha \beta^{j-1} S(\beta)\,\mathrm{d}\beta \qquad \text{for } \alpha \in (0, \bar\alpha].$$

Together with (6.5b) this implies (6.6a). Writing (3.12) as $-S'(\alpha)/\alpha \geq C_4(S(\alpha)/\alpha^2)$ and adding $C_4(-S'(\alpha)/\alpha)$ on both sides, we obtain

$$(C_4+1)\frac{-S'(\alpha)}{\alpha} \geq C_4\left(\frac{-S'(\alpha)}{\alpha}+\frac{S(\alpha)}{\alpha^2}\right) = C_4\frac{d}{d\alpha}\left\{-\frac{1}{\alpha}S(\alpha)\right\}.$$

Integrating this inequality from α to $\bar{\alpha}$ and multiplying by $(C_4+1)^{-1}$ yields

$$-\int_\alpha^{\bar{\alpha}} \frac{1}{\beta}\,\mathrm{d}S(\beta) \geq \frac{C_4}{C_4+1}\left(\frac{1}{\alpha}S(\alpha)-\frac{1}{\bar{\alpha}}S(\bar{\alpha})\right) \gtrsim \frac{C_4}{C_4+1}\frac{S(\alpha)}{\alpha}, \qquad \text{as } \alpha \searrow 0.$$

This together with (6.5d) implies (6.6b). □

Remark 14. Assume

$$R(\alpha) \asymp S(\alpha), \qquad \text{as } \alpha \searrow 0; \tag{6.10}$$

i.e., there exist constants $C \geq 1$ and $\bar{\alpha} > 0$ such that $(1/C)R(\alpha) \leq S(\alpha) \leq CR(\alpha)$ for $0 < \alpha \leq \bar{\alpha}$. In this case (6.8) holds true with $\delta = \bar{\alpha}$ and $\epsilon = \max(C-1, 1-1/C)$. Proceeding as in the proof of Lemma 13 and choosing $\delta_1 = \delta_2 = \bar{\alpha}$, we find that (6.5) holds true with $\mathrm{o}(\ldots)$ replaced by $\mathrm{O}(\ldots)$ if S satisfies (3.8). If additionally (3.12) holds true, then

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta) \asymp \frac{1}{\alpha^2}\int_0^\alpha S(\beta)\,\mathrm{d}\beta \asymp -\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}S(\beta),$$

$$-\int_\alpha^{\bar{\alpha}} \frac{1}{\beta}\,\mathrm{d}R(\beta) \asymp \frac{1}{\alpha}S(\alpha) \asymp -\int_\alpha^{\bar{\alpha}} \frac{1}{\beta}\,\mathrm{d}S(\beta).$$

Therefore similar convergence rate results with different constants can be shown if condition (3.7) in Assumption 2 is replaced by (6.10).

Proof of Theorem 3. To prove (3.9), we use the bias-variance decomposition (3.1) and the bound (3.3) of the bias. To bound the variance we start from (6.1a) in Lemma 11. From (6.5a) and (6.5b) we obtain $-\alpha^{-2}\int_0^\alpha \beta\,\mathrm{d}R(\beta) \lesssim \alpha^{-2}\int_0^\alpha S(\beta)\,\mathrm{d}\beta$. For the second term in (6.1a) the partial integration (6.9) with $T = R$ and $\bar{\alpha} > \|K^*K\|$ and (3.7) yield

$$-\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta) \leq \frac{1}{\alpha}R(\alpha) \sim \frac{1}{\alpha}S(\alpha), \quad \text{as } \alpha \searrow 0.$$

Using the partial integration (6.7) with $T = S$ and (3.8a) we obtain

$$\frac{1}{\alpha}S(\alpha) = \frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}S(\beta) + \frac{1}{\alpha^2}\int_0^\alpha S(\beta)\,\mathrm{d}\beta \leq \frac{1}{\alpha^2}\int_0^\alpha S(\beta)\,\mathrm{d}\beta.$$

This completes the proof of (3.9). The proof of (3.10) also relies on the bias-variance decomposition $\mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - g\|^2 = B^2 + V$, where the bias term satisfies

$$\begin{aligned} B &= \|\mathbf{E}\,K\hat{f}_{\alpha,\sigma} - g\| \leq \|K\Phi_\alpha(K^*K)K^*g - g\| + \tau\|K\Phi_\alpha(K^*K)K\xi\| \\ &\leq \|(\Phi_\alpha(KK^*)KK^* - I)\tilde{\Lambda}(KK^*)w\| + \tau\|\Phi_\alpha(KK^*)KK^*\| \leq \gamma_{\tilde{\Lambda}}\tilde{\Lambda}(\alpha)\overline{w} + \tau C_2. \end{aligned}$$

The bound on the variance term $V = \mathbf{E}\,\|K\hat{f}_{\alpha,\sigma} - \mathbf{E}\,K\hat{f}_{\alpha,\sigma}\|^2$ we start from (6.1b) in Lemma 11. By (6.5a) and (6.5b), the first term on the right-hand side satisfies

$\int_0^\alpha \beta^2 \,\mathrm{d}R(\beta) \lesssim \frac{1}{\alpha^2}\int_0^\alpha j\beta^{j-1}S(\beta)\,\mathrm{d}\beta$, and for the second term we obtain $R(\alpha) \sim S(\alpha) = \alpha^{-2}\int_0^\alpha 2\beta S(\alpha)\,\mathrm{d}\beta \le \alpha^{-2}\int_0^\alpha 2\beta S(\beta)\,\mathrm{d}\beta$ due to (3.8a). This shows that $V \le (\sigma/\alpha)^2(C_2^2+C_3^2)\int_0^\alpha 2\beta S(\beta)\,\mathrm{d}\beta$ and finishes the proof of (3.10). ∎

Proof of Theorem 5. Using (3.11) we can bound the variance of the spectral cut-off estimator as follows:

$$\sigma^{-2}\mathbf{E}\,\|\hat f_{\alpha,\sigma}^{\mathrm{SC}} - \mathbf{E}\,\hat f_{\alpha,\sigma}^{\mathrm{SC}}\|^2 = \int_{\mathbb{S}} \Phi_\alpha^{\mathrm{SC}}(\rho)^2 \mathbf{Var}\,(UK^*\varepsilon)\,\mathrm{d}\Sigma$$

$$\ge \gamma_{\mathbf{var}} \int_{\{\rho\ge\alpha\}} \frac{1}{\rho}\,\mathrm{d}\Sigma = -\gamma_{\mathbf{var}}\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta).$$

On the other hand, using Lemmas 12 and 13 we can bound the first term on the right-hand side of (6.1a) as follows:

$$-\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}R(\beta) \sim -\frac{1}{\alpha^2}\int_0^\alpha \beta\,\mathrm{d}S(\beta) \lesssim -\kappa\int_\alpha^{\bar\alpha}\frac{1}{\beta}\,\mathrm{d}S(\beta) \lesssim -\kappa\int_\alpha^\infty \frac{1}{\beta}\,\mathrm{d}R(\beta).$$

This yields (3.13). Inequality (3.14) follows from (3.1), (3.4), and (3.13). ∎

Proof of Theorem 6. Using the substitution $\alpha = \overline{\gamma_2}(\gamma_1(\beta))$ and Theorem 3, this follows from

$$\inf_{\alpha>0}\mathbf{E}\,\|\hat f_{\alpha,\sigma} - f\|^2 \lesssim \inf_{\alpha>0}\left(\gamma_\Lambda^2\Lambda(\alpha)^2\overline{w}^2 + (C_3^2+C_2^2)\sigma^2\gamma_2(\alpha)\right)$$

$$= \inf_{\beta>0}\left(\gamma_\Lambda^2\Lambda(\overline{\gamma_2}(\gamma_1(\beta)))^2\overline{w}^2 + (C_3^2+C_2^2)\sigma^2\gamma_1(\beta)\right)$$

$$\le \inf_{\beta>0}\left(C\gamma_\Lambda^2\Lambda(\beta)^2\overline{w}^2 + (C_3^2+C_2^2)\sigma^2\gamma_1(\beta)\right). \qquad ∎$$

Acknowledgments. We would like to thank P. Bühlmann and L. Rosasco for helpful discussions on L^2-boosting. Moreover, we are grateful to the anonymous referees for their valuable comments, which helped to improve the presentation of the material.

REFERENCES

[1] F. Abramovich and B. W. Silverman, *Wavelet decomposition approaches to statistical inverse problems*, Biometrika, 85 (1998), pp. 115–129.

[2] F. Bauer, T. Hohage, and A. Munk, *Regularized Newton methods for nonlinear inverse problems with random noise*, in preparation.

[3] F. Bauer and S. Pereverzev, *Regularization without preliminary knowledge of smoothness and error behavior*, European J. Appl. Math., 16 (2005), pp. 303–317.

[4] N. Bissantz, *Iterative inversion methods for statistical inverse problems*, in Phystat05: Proceedings of the Conference on Statistical Problems in Particle Physics, Astrophysics and Cosmology, L. Lyons and M. K. Ünel, eds., Imperial College Press, London, 2006, pp. 263–266.

[5] N. Bissantz, T. Hohage, and A. Munk, *Consistency and rates of convergence of nonlinear Tikhonov regularization with random noise*, Inverse Problems, 20 (2004), pp. 1773–1791.

[6] H. Brakhage, *On ill-posed problems and the method of conjugate gradients*, in Inverse and Ill-Posed Problems, H. W. Engl and C. W. Groetsch, eds., Academic Press, Orlando, FL, 1987, pp. 191–205.

[7] P. Bühlmann and B. Yu, *Boosting with L_2 loss: Regression and classification*, J. Amer. Statist. Assoc., 98 (2003), pp. 324–339.

[8] L. Cavalier, G. K. Golubev, D. Picard, and A. B. Tsybakov, *Oracle inequalities for inverse problems*, Ann. Statist., 30 (2002), pp. 843–874.

[9] D. D. Cox, *Approximation of method of regularization estimators*, Ann. Statist., 16 (1988), pp. 694–712.
[10] P. J. Diggle and P. Hall, *A Fourier approach to nonparametric deconvolution of a density estimate*, J. Roy. Statist. Soc. Ser. B, 55 (1993), pp. 523–531.
[11] D. L. Donoho, *Statistical estimation and optimal recovery*, Ann. Statist., 22 (1994), pp. 238–270.
[12] D. Donoho, *Nonlinear solution of linear inverse problems by wavelet-vaguelette decomposition*, Appl. Comput. Harmon. Anal., 2 (1995), pp. 101–126.
[13] S. Efromovich, *Robust and efficient recovery of a signal passed through a filter and then contaminated by non-Gaussian noise*, IEEE Trans. Inform. Theory, 43 (1997), pp. 1184–1191.
[14] H. W. Engl, M. Hanke, and A. Neubauer, *Regularization of Inverse Problems*, Kluwer Academic, Dordrecht, Boston, London, 1996.
[15] H. W. Engl and J. Zou, *A new approach to convergence rate analysis of Tikhonov regularization for parameter identification problems in heat conduction*, Inverse Problems, 16 (2000), pp. 1907–1923.
[16] S. N. Evans and P. B. Stark, *Inverse problems as statistics*, Inverse Problems, 18 (2002), pp. R55–R97.
[17] J. Fan, *On the optimal rates of convergence for nonparametric deconvolution problems*, Ann. Statist., 19 (1991), pp. 1257–1272.
[18] Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, in Machine Learning: Proceedings of the Thirteenth International Conference, Morgan Kaufmann, San Francisco, CA, 1996, pp. 148–156.
[19] J. H. Friedman, *Greedy function approximation: A gradient boosting machine*, Ann. Statist., 29 (2001), pp. 1189–1232.
[20] P. R. Halmos, *What does the spectral theorem say?*, Amer. Math. Monthly, 70 (1963), pp. 241–247.
[21] D. M. Healy, H. Hendriks, and P. T. Kim, *Spherical deconvolution*, J. Multivariate Anal., 67 (1998), pp. 1–22.
[22] E. Heinz, *Beiträge zur Störungstheorie der Spektralzerlegung*, Math. Ann., 123 (1951), pp. 425–438.
[23] T. Hohage, *Regularization of exponentially ill-posed problems*, Numer. Funct. Anal. Optim., 21 (2000), pp. 439–464.
[24] I. M. Johnstone, G. Kerkyacharian, D. Picard, and M. Raimondo, *Wavelet deconvolution in a periodic setting*, J. Roy. Statist. Soc. Ser. B, 66 (2004), pp. 547–573.
[25] I. M. Johnstone and B. W. Silverman, *Speed of estimation in positron emission tomography and related inverse problems*, Ann. Statist., 18 (1990), pp. 251–280.
[26] J. P. Kaipio and E. Somersalo, *Statistical and Computational Inverse Problems*, Springer, New York, 2004.
[27] P. T. Kim and J.-Y. Koo, *Optimal spherical deconvolution*, J. Multivariate Anal., 80 (2002), pp. 21–42.
[28] R. Kress, *Linear Integral Equations*, 2nd ed., Springer, Berlin, Heidelberg, New York, 1999.
[29] B. A. Mair, *Tikhonov regularization for finitely and infinitely smoothing operators*, SIAM J. Math. Anal., 25 (1994), pp. 135–147.
[30] B. A. Mair and F. H. Ruymgaart, *Statistical inverse estimation in Hilbert scales*, SIAM J. Appl. Math., 56 (1996), pp. 1424–1444.
[31] P. Mathé and S. V. Pereverzev, *Optimal discretization of inverse problems in Hilbert scales. Regularization and self-regularization of projection methods*, SIAM J. Numer. Anal., 38 (2001), pp. 1999–2021.
[32] P. Mathé and S. Pereverzev, *Geometry of ill-posed problems in variable Hilbert scales*, Inverse Problems, 19 (2003), pp. 789–803.
[33] P. Mathé and S. Pereverzev, *Regularization of some linear ill-posed problems with discretized random noisy data*, Math. Comp., 75 (2006), pp. 1913–1929.
[34] A. Munk, *Testing the goodness of fit of parametric regression models with random Toeplitz forms*, Scand. J. Statist., 29 (2002), pp. 501–535.
[35] A. S. Nemirovskii and B. T. Polyak, *Iterative methods for solving linear ill-posed problems under precise information* I, Engrg. Cybernetics, 22 (1984), pp. 1–11.
[36] D. W. Nychka and D. Cox, *Convergence rates for regularized solutions of integral equations from discrete noisy data*, Ann. Statist., 17 (1989), pp. 556–572.
[37] F. O'Sullivan, *A statistical perspective on ill-posed inverse problems*, Statist. Sci., 4 (1986), pp. 502–527.
[38] L. Stefanski and R. J. Carroll, *Deconvoluting kernel density estimators*, Statistics, 21 (1990), pp. 169–184.

[39] M. Taylor, *Partial Differential Equations: Basic Theory*, Vol. 1, Springer, New York, 1996.
[40] M. Taylor, *Partial Differential Equations: Qualitative Studies of Linear Equations*, Vol. 2, Springer, New York, 1996.
[41] G. Wahba, *Practical approximate solutions to linear operator equations when data are noisy*, SIAM J. Numer. Anal., 14 (1977), pp. 651–667.
[42] Y. Yao, L. Rosasco, and A. Caponnetto, *On Early Stopping in Gradient Descent Learning*, available online from http://math.berkeley.edu/~yao/publications/earlystop.pdf.
[43] T. Zhang and B. Yu, *Boosting with early stopping: Convergence and consistency*, Ann. Statist., 33 (2005), pp. 1539–1579.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2637–2652

NUMERICAL ANALYSIS OF A UNILATERAL PROBLEM IN PLANAR THERMOELASTICITY*

M. I. M. COPETTI†

Abstract. We consider in this paper the numerical approximation of a quasi-static contact problem in linear thermoelasticity that models the evolution of the temperature and displacement of an elastic, homogeneous, and isotropic body that may come in contact with an elastic obstacle. We propose a finite element method to numerically approximate the continuous solution. Convergence without any regularity assumptions is proved and error estimates are obtained if the continuous solution is sufficiently regular.

Key words. planar thermoelasticity, contact problem, finite element approximation

AMS subject classifications. 65N30, 65N15

DOI. 10.1137/06066624X

1. Introduction.

Let $\Omega \subset \mathbf{R}^2$ be a convex polygonal bounded domain which represents the reference configuration of a thermoelastic, homogeneous, and isotropic body, and assume that the boundary $\partial\Omega$ is divided into three mutually disjoint parts Γ_0, Γ_t, Γ_c such that $\partial\Omega = \overline{\Gamma}_0 \cup \overline{\Gamma}_t \cup \overline{\Gamma}_c$, $\Gamma_0 \neq \emptyset$, and $\Gamma_c \neq \emptyset$.

The equations that describe the evolution of the system under consideration are (see [15], [17])

$$\theta_t - \triangle\, \theta = -m\, \mathrm{div}\, \boldsymbol{u}_t \qquad \text{in } \Omega,\ t > 0, \tag{1.1}$$

$$\mathrm{div}\, \boldsymbol{\sigma}(\boldsymbol{u}) = \mathbf{0} \qquad \text{in } \Omega,\ t > 0, \tag{1.2}$$

where $\theta = \theta(\mathbf{x}, t)$ is the body temperature, $\boldsymbol{u} = \boldsymbol{u}(\mathbf{x}, t) = (u_1(\mathbf{x}, t),\, u_2(\mathbf{x}, t))^t$ is the displacement from the reference configuration, and $\boldsymbol{\sigma}(\boldsymbol{u}) = \lambda\, \mathrm{div}\, \boldsymbol{u}\mathbf{I} + 2\,\mu\, \boldsymbol{\epsilon}(\boldsymbol{u}) - m\,\theta\,\mathbf{I}$ is the stress. Here $\lambda,\ \mu > 0$ are the Lamé constants, $m > 0$ is the coefficient of thermal expansion, $\mathbf{I}$ is the 2×2 identity matrix, $\boldsymbol{\epsilon} = \boldsymbol{\epsilon}(\boldsymbol{u}) = 0.5(\nabla \boldsymbol{u} + (\nabla \boldsymbol{u})^t)$ is the linearized 2×2 strain function, and $\mathrm{div}\, \boldsymbol{\sigma}$ is the 2-vector with ith component $\{\mathrm{div}\, \boldsymbol{\sigma}\}_i = \sum_{j=1}^{2} \boldsymbol{\sigma}_{ij,j}$. The index j that follows the comma indicates a partial derivative with respect to $\mathbf{x}_j$.

The body is held fixed on Γ_0 and tractions are zero on Γ_t so that

$$\boldsymbol{u} = \mathbf{0} \quad \text{on } \Gamma_0, \qquad \boldsymbol{\sigma\nu} = \mathbf{0} \quad \text{on } \Gamma_t,\ t > 0, \tag{1.3}$$

with $\boldsymbol{\nu} = (\nu_1, \nu_2)^t$ the normal unit vector pointing out of Ω. At Γ_c the body may come in contact with an elastic obstacle with rigidity $\frac{1}{\epsilon} > 0$ located initially at distance $g(\mathbf{x}) \geq 0$ from Γ_c. At the region of contact the normal stress is compressive, the contact is frictionless, and the body can penetrate the obstacle:

$$\left.\begin{array}{l} \boldsymbol{\sigma}_T = \boldsymbol{\sigma\nu} - \sigma_\nu \boldsymbol{\nu} = \mathbf{0}, \\ \sigma_\nu = -\frac{1}{\epsilon}[u_\nu - g]_+ \end{array}\right\} \quad \text{on } \Gamma_c,\ \ t > 0, \tag{1.4}$$

*Received by the editors July 27, 2006; accepted for publication (in revised form) May 9, 2007; published electronically December 7, 2007. This research was partially supported by the Brazilian institution CNPq.

http://www.siam.org/journals/sinum/45-6/66624.html

†Laboratório de Análise Numérica e Astrofísica, Departamento de Matemática, Universidade Federal de Santa Maria, 97105-900 Santa Maria, RS, Brazil (mimc@lana.ccne.ufsm.br).

where $\sigma_\nu = (\boldsymbol{\sigma\nu})\cdot\boldsymbol{\nu}$, $u_\nu = \boldsymbol{u}\cdot\boldsymbol{\nu}$, and $[\chi]_+ \equiv \max\{\chi, 0\}$. For simplicity, we assume that the temperature is equal to zero at the boundary of Ω and it is initially prescribed:

$$\theta = 0 \quad \text{on } \partial\Omega, \qquad \theta(\mathbf{x}, 0) = \theta_0(\mathbf{x}), \qquad \mathbf{x} \in \Omega. \tag{1.5}$$

This problem, for a nonisotropic, nonhomogeneous elastic body was used by Rivera and Racke [15] as a penalization to the contact problem with a rigid obstacle where (1.4) is replaced by the so-called Signorini contact condition

$$\left.\begin{array}{ll} u_\nu \le g, & \sigma_\nu = (\boldsymbol{\sigma\nu})\cdot\boldsymbol{\nu} \le 0, \\ \sigma_\nu(u_\nu - g) = 0, & \boldsymbol{\sigma}_T = \boldsymbol{\sigma\nu} - \sigma_\nu\boldsymbol{\nu} = \mathbf{0} \end{array}\right\} \quad \text{on } \Gamma_c, \quad t > 0. \tag{1.6}$$

One dimensional quasi-static contact problems with various boundary conditions for the temperature have been extensively studied by mathematicians and engineers both theoretically and numerically. In particular, existence and uniqueness results were obtained by Gilbert, Shi, and Shillor [12] and Shi and Shillor [16]. Finite element approximations were analyzed by Copetti and Elliott [10] and Copetti [8], [9]. Error estimates were given and numerical experiments performed. In [14], the stability of steady-state solutions was considered.

Multidimensional contact problems are more complex due to the presence of the elasticity equations. An existence result for the Signorini contact problem, with either a Dirichlet or a heat exchange condition for the temperature, was proved by Shi and Shillor [17] using truncation and compactness arguments, while uniqueness was established by Ames and Payne [1]. Quasi-static and dynamic problems with Barber's heat exchange condition were studied by Xu [19] and Bien [3], respectively. The papers [2], [5], and [6] provide error analysis for some contact problems in elasticity. A static problem in thermoelasticity with many bodies in contact is numerically approximated by the finite element method in [13]. To our knowledge, the present paper is the first work dealing with the numerical approximation of problem (1.1)–(1.5).

Let us introduce the spaces $\mathbf{L}^2(\Omega) = \{L^2(\Omega)\}^2$, $\mathbf{H}^s(\Omega) = \{H^s(\Omega)\}^2$, and $\mathbf{H}^1_E(\Omega) = \{\boldsymbol{v} \in \mathbf{H}^1(\Omega) \mid \boldsymbol{v} = \mathbf{0} \text{ on } \Gamma_0\}$. We indicate the inner product in $\{L^2(\Omega)\}^n$, $n = 1, 2$, by $(\cdot,\cdot)$ and the norms of $\{L^2(\Omega)\}^n$ and $\{H^s(\Omega)\}^n$ by $\|\cdot\|$ and $\|\cdot\|_s$, respectively.

Throughout the paper, the letter C is used to denote a positive constant which depends on the data and is independent of mesh sizes.

If $\boldsymbol{\sigma}$ and $\boldsymbol{\tau}$ are 2×2 matrix-valued functions, we define

$$\boldsymbol{\sigma} : \boldsymbol{\tau} = \sum_{i,j=1}^{2} \sigma_{ij}\tau_{ij},$$

and we consider on $\mathbf{H}^1_E(\Omega)$ the inner product

$$b(\boldsymbol{u}, \boldsymbol{v}) = \int_\Omega \boldsymbol{\epsilon}(\boldsymbol{u}) : \boldsymbol{\epsilon}(\boldsymbol{v}) d\mathbf{x}$$

and let $\|v\|_b = (b(\boldsymbol{v}, \boldsymbol{v}))^{\frac{1}{2}}$ be the associated norm. Since $\Gamma_0 \neq \emptyset$, the Korn inequality [4]

$$\|\boldsymbol{v}\|_b \ge C\|\boldsymbol{v}\|_1 \quad \forall\, \boldsymbol{v} \in \mathbf{H}^1_E(\Omega) \tag{1.7}$$

implies that the bounded bilinear form $a(\boldsymbol{v}, \mathbf{w}) \equiv \lambda(\operatorname{div}\boldsymbol{v}, \operatorname{div}\mathbf{w}) + 2\mu\, b(\boldsymbol{v}, \mathbf{w})$ is coercive.

The following result holds [1], [15].

THEOREM 1.1. *If $\theta_0 \in H_0^1(\Omega)$, $g \in H^{\frac{1}{2}}(\Gamma_c)$, and m is sufficiently small, there exists a unique solution to* (1.1)–(1.5) *such that*

$$\begin{aligned} \theta \in L^2(0,T;H^2(\Omega)\cap H_0^1(\Omega)), \quad \theta_t \in L^2(0,T;L^2(\Omega)),\\ \boldsymbol{u} \in L^2(0,T;\mathbf{H}_E^1(\Omega)), \quad \boldsymbol{u}_t \in L^2(0,T;\mathbf{H}^1(\Omega)),\\ \{\operatorname{div}\boldsymbol{\sigma}(\boldsymbol{u})\}_i \in L^2(0,T;L^2(\Omega)), \quad i=1,2. \end{aligned}$$

Given any $w \in H_0^1(\Omega)$ and $\boldsymbol{v} \in \mathbf{H}_E^1(\Omega)$, the weak form

$$(\theta_t, w) + (\nabla\theta, \nabla w) = -m(\operatorname{div}\boldsymbol{u}_t, w), \tag{1.8}$$

$$\lambda(\operatorname{div}\boldsymbol{u}, \operatorname{div}\boldsymbol{v}) + 2\mu\, b(\boldsymbol{u},\boldsymbol{v}) - m(\theta, \operatorname{div}\boldsymbol{v}) + \frac{1}{\epsilon}([u_\nu - g]_+, v_\nu)_{\Gamma_c} = 0 \tag{1.9}$$

will be the base of our numerical method. We define $\boldsymbol{u}(\mathbf{x},0) \equiv \boldsymbol{u}_0(\mathbf{x})$ as the unique solution in $\mathbf{H}_E^1(\Omega)$ of (1.2)–(1.4) when $\theta = \theta_0$.

2. Numerical approximation. Let T^h be a regular family of triangulations of Ω, $\overline{\Omega} = \bigcup_{\tau\in T^h}\tau$ with mesh size h such that any point where the boundary condition changes is a vertex of the triangulations. Let $S^h \subset H^1(\Omega)$ and $\mathbf{S}^h \subset \mathbf{H}^1(\Omega)$ be finite element spaces of continuous functions on $\overline{\Omega}$ which are linear on each $\tau \in T^h$. We introduce the spaces

$$S_0^h = \{v \in S^h \mid v = 0 \text{ on } \partial\Omega\} \text{ and } \mathbf{S}_E^h = \{\boldsymbol{v} \in \mathbf{S}^h \mid \boldsymbol{v} = \mathbf{0} \text{ on } \Gamma_0\}$$

and the projection $P_E^h : \mathbf{H}_E^1 \to \mathbf{S}_E^h$ defined by $a(\boldsymbol{v} - P_E^h\boldsymbol{v}, \mathbf{w}) = 0\ \forall\, \mathbf{w} \in \mathbf{S}_E^h$. As a consequence of Korn's inequality (1.7) we have

$$\|\boldsymbol{v} - P_E^h\boldsymbol{v}\|_1 \le C \inf_{\mathbf{w}\in\mathbf{S}_E^h} \|\boldsymbol{v} - \mathbf{w}\|_1. \tag{2.1}$$

We will also use the elliptic projection $P_0^h : H_0^1 \to S_0^h$, $(\nabla(\eta - P_0^h\eta), \nabla\chi) = 0$ $\forall\,\chi \in S_0^h$, and the L^2-projection $P^h : \mathbf{L}^2 \to \mathbf{S}_E^h$, $(\boldsymbol{\zeta} - P^h\boldsymbol{\zeta}, \boldsymbol{\phi}) = 0\ \forall\boldsymbol{\phi} \in \mathbf{S}_E^h$, which satisfy [18]

$$\|\eta - P_0^h\eta\| + h\|\nabla(\eta - P_0^h\eta)\| \le Ch^s\|\eta\|_s, \quad 1 \le s \le 2, \tag{2.2}$$

$$\|\boldsymbol{\zeta} - P^h\boldsymbol{\zeta}\| \le Ch^s\|\boldsymbol{\zeta}\|_s, \quad 1 \le s \le 2. \tag{2.3}$$

For continuous functions $\boldsymbol{v}$, we denote by $I^h\boldsymbol{v}$ the interpolant of $\boldsymbol{v}$ in $\mathbf{S}_E^h$ with respect to the triangulation T^h. The following approximation property holds [18]:

$$\|\boldsymbol{v} - I^h\boldsymbol{v}\| + h\|\nabla(\boldsymbol{v} - I^h\boldsymbol{v})\| \le Ch^s\|\boldsymbol{v}\|_s, \quad 1 \le s \le 2. \tag{2.4}$$

The finite element approximation to (1.1)–(1.5) is to find $\Theta^n \in S_0^h$, $\boldsymbol{U}^n \in \mathbf{S}_E^h$, $n = 1, 2, \ldots, N$, such that $\forall\, W \in S_0^h$ and $\forall\,\boldsymbol{V} \in \mathbf{S}_E^h$

$$\frac{1}{\Delta t}(\Theta^n - \Theta^{n-1}, W) + (\nabla\Theta^n, \nabla W) = -\frac{m}{\Delta t}(\operatorname{div}(\boldsymbol{U}^n - \boldsymbol{U}^{n-1}), W), \tag{2.5}$$

$$\lambda(\operatorname{div}\boldsymbol{U}^n, \operatorname{div}\boldsymbol{V}) + 2\mu\, b(\boldsymbol{U}^n, \boldsymbol{V}) - m(\Theta^n, \operatorname{div}\boldsymbol{V}) + \frac{1}{\epsilon}([U_\nu^n - g]_+, V_\nu)_{\Gamma_c} = 0. \tag{2.6}$$

Here $\Delta t = T/N$ and Θ^0, $\boldsymbol{U}^0$ are approximations to θ_0 and $\boldsymbol{u}_0$, respectively.

Suppose that $\{\Theta^n, \boldsymbol{U}^n\}$ and $\{\tilde{\Theta}^n, \tilde{\boldsymbol{U}}^n\}$ are two solutions of (2.5)–(2.6), and define $\psi = \Theta^n - \tilde{\Theta}^n, \boldsymbol{\eta} = \boldsymbol{U}^n - \tilde{\boldsymbol{U}}^n$. Thus, $\forall\, W \in S_0^h$ and $\forall\, \boldsymbol{V} \in \mathbf{S}_E^h$,

$$(\psi, W) + \Delta t(\nabla\psi, \nabla W) = -m(\operatorname{div}\boldsymbol{\eta}, W),$$
$$\lambda(\operatorname{div}\boldsymbol{\eta}, \operatorname{div}\boldsymbol{V}) + 2\mu\, b(\boldsymbol{\eta}, \boldsymbol{V}) - m(\psi, \operatorname{div}\boldsymbol{V}) + \frac{1}{\epsilon}([U_\nu^n - g]_+ - [\tilde{U}_\nu^n - g]_+, V_\nu)_{\Gamma_c} = 0,$$

and

$$\|\psi\|^2 + \Delta t\|\nabla\psi\|^2 + \lambda\|\operatorname{div}\boldsymbol{\eta}\|^2 + 2\mu\,\|\boldsymbol{\eta}\|_b^2 = -\frac{1}{\epsilon}([U_\nu^n - g]_+ - [\tilde{U}_\nu^n - g]_+, U_\nu^n - \tilde{U}_\nu^n)_{\Gamma_c} \le 0$$

since the functional $[\cdot]_+$ is monotone. It follows that (2.5)–(2.6) has a unique solution.

2.1. Implementation. To compute the numerical solution we used the iterative procedure

$$\frac{1}{\Delta t}(\Theta^{n,l} - \Theta^{n-1}, W) + (\nabla\Theta^{n,l}, \nabla W) = -\frac{m}{\Delta t}(\operatorname{div}(\boldsymbol{U}^{n,l-1} - \boldsymbol{U}^{n-1}), W),$$
$$\lambda(\operatorname{div}\boldsymbol{U}^{n,l}, \operatorname{div}\boldsymbol{V}) + 2\mu\, b(\boldsymbol{U}^{n,l}, \boldsymbol{V}) - m(\Theta^{n,l}, \operatorname{div}\boldsymbol{V}) + \frac{1}{\epsilon}([U_\nu^{n,l-1} - g]_+, V_\nu)_{\Gamma_c} = 0,$$

where $\boldsymbol{U}^{n,0} = \boldsymbol{U}^{n-1}$. At each iteration l, two systems of linear equations need to be solved. Since the coefficient matrices are symmetric positive definite, these systems have unique solutions and the Gauss–Seidel method can be used to solve them.

THEOREM 2.1. *If μ is sufficiently large and m is sufficiently small, the sequences $\{\Theta^{n,l}\}$ and $\{\boldsymbol{U}^{n,l}\}$ converge to the unique solution of* (2.5)–(2.6).

Proof. Let $p^l = \Theta^{n,l} - \Theta^{n,l-1}$ and $\boldsymbol{q}^l = \boldsymbol{U}^{n,l} - \boldsymbol{U}^{n,l-1}$. We have

$$\|p^l\|^2 + \Delta t\|\nabla p^l\|^2 + \lambda\|\operatorname{div}\boldsymbol{q}^l\|^2 + 2\mu\,\|\boldsymbol{q}^l\|_b^2 = m(\operatorname{div}(\boldsymbol{q}^l - \boldsymbol{q}^{l-1}), p^l)$$
$$- \frac{1}{\epsilon}([U_\nu^{n,l-1} - g]_+ - [U_\nu^{n,l-2} - g]_+, q_\nu^l)_{\Gamma_c}.$$

Using Sobolev's trace theorem and (1.7) we find that, for $\alpha > 0$,

$$\frac{1}{\epsilon}([U_\nu^{n,l-1} - g]_+ - [U_\nu^{n,l-2} - g]_+, q_\nu^l)_{\Gamma_c} \le \frac{1}{\epsilon}\|q_\nu^{l-1}\|_{L^2(\Gamma_c)}\,\|q_\nu^l\|_{L^2(\Gamma_c)}$$
$$\le \frac{C}{\epsilon}\|\boldsymbol{q}^{l-1}\|_b\,\|\boldsymbol{q}^l\|_b \le \frac{\alpha C}{2\epsilon}\|\boldsymbol{q}^{l-1}\|_b^2 + \frac{C}{2\alpha\epsilon}\,\|\boldsymbol{q}^l\|_b^2.$$

Therefore, for $\delta > 0$,

$$\|\ p^l\|^2 + \Delta t\|\nabla p^l\|^2 + \lambda\|\operatorname{div}\boldsymbol{q}^l\|^2 + 2\mu\,\|\boldsymbol{q}^l\|_b^2 \le \frac{\delta m^2}{2}\|\operatorname{div}(\boldsymbol{q}^l - \boldsymbol{q}^{l-1})\|^2$$
$$+\ \frac{1}{2\delta}\|p^l\|^2 + \frac{\alpha C}{2\epsilon}\|\boldsymbol{q}^{l-1}\|_b^2 + \frac{C}{2\alpha\epsilon}\|\boldsymbol{q}^l\|_b^2$$
$$\le \delta m^2 C\left(\|\boldsymbol{q}^l\|_b^2 + \|\boldsymbol{q}^{l-1}\|_b^2\right) + \frac{1}{2\delta}\|p^l\|^2 + \frac{\alpha C}{2\epsilon}\|\boldsymbol{q}^{l-1}\|_b^2 + \frac{C}{2\alpha\epsilon}\,\|\boldsymbol{q}^l\|_b^2.$$

It follows that

$$\left(1 - \frac{1}{2\delta}\right)\|p^l\|^2 + \Delta t\|\nabla p^l\|^2 + \lambda\|\operatorname{div}\boldsymbol{q}^l\|^2 + \left(2\mu - \delta m^2 C - \frac{C}{2\alpha\epsilon}\right)\|\boldsymbol{q}^l\|_b^2$$
$$\le \left(\delta m^2 C + \frac{\alpha C}{2\epsilon}\right)\|\boldsymbol{q}^{l-1}\|_b^2.$$

Taking $\delta = \alpha = 1$ results in

$$\left(2\mu - m^2C - \frac{C}{2\epsilon}\right)\|\boldsymbol{q}^l\|_b^2 \le \left(m^2C + \frac{C}{2\epsilon}\right)\|\boldsymbol{q}^{l-1}\|_b^2.$$

Thus, for sufficiently large μ and sufficiently small m, $2\mu - m^2C - C/(2\epsilon) > 0$ and there exists M, $0 < M < 1$, such that

$$\|\boldsymbol{q}^l\|_b^2 \le M\|\boldsymbol{q}^{l-1}\|_b^2,$$

which proves the result. □

We choose the initial displacement $\boldsymbol{U}^0$ as the unique solution of (2.6) when $\Theta^n = \Theta^0$. To find $\boldsymbol{U}^0$ we iterate in the nonlinear term. Convergence follows as in the previous theorem.

3. Convergence. In this section, a convergence analysis and an error estimate are presented. We follow the work of Copetti and French [11].

Let us define $t_n = n\Delta t$, $\theta^n = \theta(\cdot, t_n)$, $\boldsymbol{u}^n = \boldsymbol{u}(\cdot, t_n)$, and

$$\hat{\theta}^n = \int_0^{t_n} \theta(\cdot,t)dt, \quad \bar{\boldsymbol{u}}^n = \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} \boldsymbol{u}(\cdot,t)dt, \quad \bar{\theta}^n = \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} \theta(\cdot,t)dt,$$

$$\beta_\epsilon(\chi) = \frac{1}{\epsilon}[\chi]_+, \quad \bar{\beta}_\epsilon^n(\chi) = \frac{1}{\Delta t}\int_{t_{n-1}}^{t_n} \beta_\epsilon(\chi)dt,$$

$$\varepsilon^j = \Delta t\sum_{i=1}^{j}\Theta^i - P_0^h\hat{\theta}^j, \ j = 1,\dots,n, \quad \varepsilon^0 = 0,$$

and note that

$$\varepsilon^n - \varepsilon^{n-1} = \Delta t(\Theta^n - P_0^h\bar{\theta}^n). \tag{3.1}$$

THEOREM 3.1. *Let $E(\theta, \boldsymbol{u})$ be defined by*

$$E(\theta,\boldsymbol{u}) = \left\|\int_0^T \nabla\theta(\cdot,t)dt - \Delta t\sum_{i=1}^{N}\nabla\Theta^i\right\|^2 + \Delta t\sum_{i=1}^{N}\|\Theta^i - \theta(\cdot,t_i)\|^2$$
$$+ \lambda\Delta t\sum_{i=1}^{N}\|\mathrm{div}\,(\boldsymbol{U}^i - \boldsymbol{u}(\cdot,t_i))\|^2 + \mu\Delta t\sum_{i=1}^{N}\|\boldsymbol{U}^i - \boldsymbol{u}(\cdot,t_i)\|_b^2.$$

Then,

$$E(\theta,\boldsymbol{u}) \le C(\lambda,\mu,\epsilon)\bigg(\|\Theta^0 - \theta_0\|^2 + \frac{1}{\Delta t}\|\boldsymbol{U}^0 - \boldsymbol{u}_0\|^2 + (\Delta t)^2 + h^4$$
$$+ \left(\frac{1}{\Delta t} + 1\right)\Delta t\sum_{i=1}^{N}\|\boldsymbol{u}^i - P_E^h\boldsymbol{u}^i\|_1^2\bigg).$$

Proof. Integrating (1.8) from 0 to t_n and adding (2.5) from 1 to n ($1 \le n \le N$), we find, $\forall\, W \in S_0^h$,

$$(\Theta^n - \theta^n, W) + \left(\Delta t\sum_{i=1}^{n}\nabla\Theta^i - \nabla\hat{\theta}^n, \nabla W\right) = -m(\mathrm{div}\,(\boldsymbol{U}^n - \boldsymbol{u}^n), W)$$
$$+ \ (\Theta^0 - \theta_0, W) + m(\mathrm{div}\,(\boldsymbol{U}^0 - \boldsymbol{u}_0), W),$$

which can be written as

$$\frac{1}{\Delta t}(\varepsilon^n-\varepsilon^{n-1},W)+(\nabla\varepsilon^n,\nabla W)=-m(\operatorname{div}(\boldsymbol{U}^n-\boldsymbol{u}^n),W)$$
$$+(\Theta^0-\theta_0,W)+m(\operatorname{div}(\boldsymbol{U}^0-\boldsymbol{u}_0),W)+(\theta^n-P_0^h\bar\theta^n,W).$$

Taking $W=\frac{\varepsilon^n-\varepsilon^{n-1}}{\Delta t}$ results in

$$\frac{1}{(\Delta t)^2}\|\varepsilon^n-\varepsilon^{n-1}\|^2+\frac{1}{2\Delta t}\left(\|\nabla\varepsilon^n-\nabla\varepsilon^{n-1}\|^2+\|\nabla\varepsilon^n\|^2-\|\nabla\varepsilon^{n-1}\|^2\right)$$
$$=-\frac{m}{\Delta t}(\operatorname{div}(\boldsymbol{U}^n-\boldsymbol{u}^n),\varepsilon^n-\varepsilon^{n-1})+\frac{1}{\Delta t}(\Theta^0-\theta_0,\varepsilon^n-\varepsilon^{n-1})$$
$$+\frac{m}{\Delta t}(\operatorname{div}(\boldsymbol{U}^0-\boldsymbol{u}_0),\varepsilon^n-\varepsilon^{n-1})+\frac{1}{\Delta t}(\theta^n-P_0^h\bar\theta^n,\varepsilon^n-\varepsilon^{n-1}). \tag{3.2}$$

Integrating (1.9) from t_{n-1} to t_n and subtracting the result from (2.6), we get

$$\lambda(\operatorname{div}(\boldsymbol{U}^n-\bar{\boldsymbol{u}}^n),\operatorname{div}\boldsymbol{V})+2\mu\,b(\boldsymbol{U}^n-\bar{\boldsymbol{u}}^n,\boldsymbol{V})-m(\Theta^n-\bar\theta^n,\operatorname{div}\boldsymbol{V})$$
$$+\left(\beta_\epsilon(U_\nu^n-g)-\bar\beta_\epsilon^n(u_\nu-g),V_\nu\right)_{\Gamma_c}=0,$$

and the definition of P_E^h yields

$$\lambda(\operatorname{div}\boldsymbol{e}^n,\operatorname{div}\boldsymbol{V})+2\mu\,b(\boldsymbol{e}^n,\boldsymbol{V})-m(\Theta^n-\bar\theta^n,\operatorname{div}\boldsymbol{V})$$
$$+\left(\beta_\epsilon(U_\nu^n-g)-\bar\beta_\epsilon^n(u_\nu-g),V_\nu\right)_{\Gamma_c}$$
$$=\lambda(\operatorname{div}(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n),\operatorname{div}\boldsymbol{V})+2\mu\,b(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n,\boldsymbol{V}), \tag{3.3}$$

where $\boldsymbol{e}^n=\boldsymbol{U}^n-P_E^h\boldsymbol{u}^n$. Let us observe that

$$\beta_\epsilon(U_\nu^n-g)-\bar\beta_\epsilon^n(u_\nu-g)=\beta_\epsilon(U_\nu^n-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g)+\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g)-\bar\beta_\epsilon^n(u_\nu-g)$$

and that

$$(\beta_\epsilon(\chi)-\beta_\epsilon(\eta),\chi-\eta)\geq 0.$$

Taking $\boldsymbol{V}=\boldsymbol{e}^n$ in (3.3), we can write

$$\lambda\|\operatorname{div}\boldsymbol{e}^n\|^2+2\mu\|\boldsymbol{e}^n\|_b^2+\left(\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g)-\bar\beta_\epsilon^n(u_\nu-g),e_\nu^n\right)_{\Gamma_c}$$
$$\leq\lambda(\operatorname{div}(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n),\operatorname{div}\boldsymbol{e}^n)+2\mu\,b(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n,\boldsymbol{e}^n)+m(\Theta^n-\bar\theta^n,\operatorname{div}\boldsymbol{e}^n)$$
$$\leq\lambda(\operatorname{div}(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n),\operatorname{div}\boldsymbol{e}^n)+2\mu\,b(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n,\boldsymbol{e}^n)$$
$$+\ m(\Theta^n-P_0^h\bar\theta^n,\operatorname{div}(\boldsymbol{U}^n-\boldsymbol{u}^n))+m(\bar\theta^n-P_0^h\bar\theta^n,\operatorname{div}(\boldsymbol{u}^n-\boldsymbol{U}^n))$$
$$+\ m(\Theta^n-P_0^h\bar\theta^n+P_0^h\bar\theta^n-\bar\theta^n,\operatorname{div}(\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n)). \tag{3.4}$$

Adding (3.2) and (3.4) and recalling (3.1) result in

$$\frac{1}{(\Delta t)^2}\|\varepsilon^n-\varepsilon^{n-1}\|^2+\frac{1}{2\Delta t}\left(\|\nabla\varepsilon^n-\nabla\varepsilon^{n-1}\|^2+\|\nabla\varepsilon^n\|^2-\|\nabla\varepsilon^{n-1}\|^2\right)+\lambda\|\operatorname{div}\boldsymbol{e}^n\|^2$$
$$+\ 2\mu\|\boldsymbol{e}^n\|_b^2\leq\frac{1}{\Delta t}(\Theta^0-\theta_0,\varepsilon^n-\varepsilon^{n-1})+\frac{m}{\Delta t}(\operatorname{div}(\boldsymbol{U}^0-\boldsymbol{u}_0),\varepsilon^n-\varepsilon^{n-1})$$
$$+\ \frac{1}{\Delta t}(\theta^n-P_0^h\bar\theta^n,\varepsilon^n-\varepsilon^{n-1})+\lambda(\operatorname{div}(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n),\operatorname{div}\boldsymbol{e}^n)$$
$$+\ 2\mu\,b(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n,\boldsymbol{e}^n)+m(\bar\theta^n-P_0^h\bar\theta^n,\operatorname{div}(P_E^h\boldsymbol{u}^n-\boldsymbol{U}^n))$$
$$+\ \frac{m}{\Delta t}(\varepsilon^n-\varepsilon^{n-1},\operatorname{div}(\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n))+\left(\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g),e_\nu^n\right)_{\Gamma_c}.$$

Thus

$$\frac{1}{2(\Delta t)^2}\|\varepsilon^n-\varepsilon^{n-1}\|^2+\frac{1}{4\Delta t}\left(\|\nabla\varepsilon^n-\nabla\varepsilon^{n-1}\|^2\right)+\frac{1}{2\Delta t}\left(\|\nabla\varepsilon^n\|^2-\|\nabla\varepsilon^{n-1}\|^2\right)$$
$$+\ \frac{\lambda}{2}\|\operatorname{div}\boldsymbol{e}^n\|^2+\mu\|\boldsymbol{e}^n\|_b^2\le\|\Theta^0-\theta_0\|^2+\frac{2m^2}{\Delta t}\|\boldsymbol{U}^0-\boldsymbol{u}_0\|^2+\|\theta^n-P_0^h\bar\theta^n\|^2$$
$$+\ \lambda\|\operatorname{div}(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n)\|^2+\mu\|\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n\|_b^2+\frac{2m^2}{\Delta t}\|\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n\|^2+\frac{m^2}{\lambda}\|\bar\theta^n-P_0^h\bar\theta^n\|^2$$
$$+\ \frac{1}{2}\|\operatorname{div}(\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n)\|^2+\left(\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g),e_\nu^n\right)_{\Gamma_c}, \tag{3.5}$$

where we used that $(\operatorname{div}\boldsymbol{v},w)=-(\boldsymbol{v},\nabla w),\ w\in H_0^1(\Omega)$.

Let us estimate the last seven terms in the latter inequality. Thus

$$I_1\equiv\|\theta^n-P_0^h\bar\theta^n\|^2=\|\theta^n-\bar\theta^n+\bar\theta^n-P_0^h\bar\theta^n\|^2$$
$$\le 2\left\|\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n}\int_t^{t_n}\theta_t(\cdot,s)dsdt\right\|^2+2\left\|\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n}(\theta-P_0^h\theta)(\cdot,t)dt\right\|^2$$
$$\le C\left(\Delta t\int_{t_{n-1}}^{t_n}\|\theta_t\|^2dt+\frac{h^4}{\Delta t}\int_{t_{n-1}}^{t_n}\|\theta\|_2^2dt\right),$$
$$I_2\equiv\lambda\|\operatorname{div}(\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n)\|^2\le C\Delta t\int_{t_{n-1}}^{t_n}\|\operatorname{div}\boldsymbol{u}_t\|^2dt,$$
$$I_3\equiv\mu\|\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n\|_b^2\le C\mu\|\bar{\boldsymbol{u}}^n-\boldsymbol{u}^n\|_1^2\le C\mu\Delta t\int_{t_{n-1}}^{t_n}\|\boldsymbol{u}_t\|_1^2dt,$$
$$I_4\equiv\frac{2m^2}{\Delta t}\|\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n\|^2\le\frac{C}{\Delta t}\|\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n\|_1^2,$$
$$I_5\equiv\frac{m^2}{\lambda}\|\bar\theta^n-P_0^h\bar\theta^n\|^2\le\frac{Ch^4}{\lambda\Delta t}\int_{t_{n-1}}^{t_n}\|\theta\|_2^2dt,$$
$$I_6\equiv\frac{1}{2}\|\operatorname{div}(\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n)\|^2\le C\|\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n\|_1^2,$$
$$I_7\equiv\left(\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g),e_\nu^n\right)_{\Gamma_c}$$
$$=\left(\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)+\beta_\epsilon(u_\nu^n-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g),e_\nu^n\right)_{\Gamma_c}.$$

We have, using Sobolev's trace theorem and (1.7),

$$\left(\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g),e_\nu^n\right)_{\Gamma_c}\le\|\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)\|_{L^2(\Gamma_c)}\|e_\nu^n\|_{L^2(\Gamma_c)}$$
$$\le\|\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)\|_{L^2(\Gamma_c)}\|\boldsymbol{e}^n\|_{\mathbf{L}^2(\Gamma_c)}\le C\|\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)\|_{L^2(\Gamma_c)}\|\boldsymbol{e}^n\|_1$$
$$\le\frac{C}{\mu}\|\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)\|_{L^2(\Gamma_c)}^2+\frac{\mu}{4}\|\boldsymbol{e}^n\|_b^2$$

and

$$\|\bar\beta_\epsilon^n(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)\|_{L^2(\Gamma_c)}^2=\left\|\frac{1}{\Delta t}\int_{t_{n-1}}^{t_n}\left(\beta_\epsilon(u_\nu-g)-\beta_\epsilon(u_\nu^n-g)\right)dt\right\|_{L^2(\Gamma_c)}^2$$
$$\le\left\|\frac{1}{\epsilon\Delta t}\int_{t_{n-1}}^{t_n}|u_\nu-u_\nu^n|dt\right\|_{L^2(\Gamma_c)}^2\le\left\|\frac{1}{\epsilon\Delta t}\int_{t_{n-1}}^{t_n}\left|\int_{t_n}^t u_{\nu t}(\cdot,s)ds\right|dt\right\|_{L^2(\Gamma_c)}^2$$

$$\leq \frac{C\Delta t}{\epsilon^2}\int_{t_{n-1}}^{t_n}\|\boldsymbol{u}_t\|_1^2 dt.$$

On the other hand,

$$\begin{aligned}&\left(\beta_\epsilon(u_\nu^n-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g),e_\nu^n\right)_{\Gamma_c}\\ &\leq\frac{1}{\mu}\left\|\beta_\epsilon(u_\nu^n-g)-\beta_\epsilon((P_E^h\boldsymbol{u}^n)_\nu-g)\right\|_{L^2(\Gamma_c)}^2+\frac{\mu}{4}\|e_\nu^n\|_{L^2(\Gamma_c)}^2\\ \leq\frac{1}{\epsilon^2\mu}\left\|u_\nu^n-(P_E^h\boldsymbol{u}^n)_\nu\right\|_{L^2(\Gamma_c)}^2+\frac{\mu}{4}\|e_\nu^n\|_{L^2(\Gamma_c)}^2&\leq\frac{C}{\epsilon^2\mu}\left\|\boldsymbol{u}^n-P_E^h\boldsymbol{u}^n\right\|_1^2+\frac{\mu}{4}\|\boldsymbol{e}^n\|_b^2.\end{aligned}$$

Collecting all the bounds and summing (3.5) over n result in

$$\begin{aligned}&\Delta t\sum_{i=1}^n\left\|\frac{\varepsilon^i-\varepsilon^{i-1}}{\Delta t}\right\|^2+\sum_{i=1}^n\|\nabla(\varepsilon^i-\varepsilon^{i-1})\|^2+\|\nabla\varepsilon^n\|^2+\lambda\Delta t\sum_{i=1}^n\|\operatorname{div}\boldsymbol{e}^i\|^2\\ &+\mu\Delta t\sum_{i=1}^n\|\boldsymbol{e}^i\|_b^2\leq C(\lambda,\mu,\epsilon)\bigg(\|\nabla\varepsilon^0\|^2+\|\Theta^0-\theta_0\|^2+\frac{1}{\Delta t}\|\boldsymbol{U}^0-\boldsymbol{u}_0\|^2\\ &+(\Delta t)^2\int_0^{t_n}\|\theta_t\|^2dt+h^4\int_0^{t_n}\|\theta\|_2^2dt+(\Delta t)^2\int_0^{t_n}\|\operatorname{div}\boldsymbol{u}_t\|^2dt\\ &+(\Delta t)^2\int_0^{t_n}\|\boldsymbol{u}_t\|_1^2dt+\left(\frac{1}{\Delta t}+1\right)\Delta t\sum_{i=1}^n\|\boldsymbol{u}^i-P_E^h\boldsymbol{u}^i\|_1^2\bigg).\end{aligned}$$

The definition of ε^0 and the regularity of θ and $\boldsymbol{u}$ yield

$$\begin{aligned}&\Delta t\sum_{i=1}^n\left\|\frac{\varepsilon^i-\varepsilon^{i-1}}{\Delta t}\right\|^2+\|\nabla\varepsilon^n\|^2+\lambda\Delta t\sum_{i=1}^n\|\operatorname{div}\boldsymbol{e}^i\|^2+\mu\Delta t\sum_{i=1}^n\|\boldsymbol{e}^i\|_b^2\\ &\leq C(\lambda,\mu,\epsilon)\bigg(\|\Theta^0-\theta_0\|^2+\frac{1}{\Delta t}\|\boldsymbol{U}^0-\boldsymbol{u}_0\|^2+(\Delta t)^2+h^4\\ &+\left(\frac{1}{\Delta t}+1\right)\Delta t\sum_{i=1}^n\|\boldsymbol{u}^i-P_E^h\boldsymbol{u}^i\|_1^2\bigg),\end{aligned}$$

and the result follows observing that

$$\begin{aligned}\left\|\int_0^{t_n}\nabla\theta(\cdot,t)dt-\Delta t\sum_{i=1}^n\nabla\Theta^i\right\|^2&=\left\|\nabla(\hat\theta^n-P_0^h\hat\theta^n)-\nabla\varepsilon^n\right\|^2,\\ \Delta t\sum_{i=1}^n\|\Theta^i-\theta(\cdot,t_i)\|^2=\Delta t\sum_{i=1}^n\left\|\frac{\varepsilon^i-\varepsilon^{i-1}}{\Delta t}+P_0^h\bar\theta^i-\theta^i\right\|^2,&\\ \lambda\Delta t\sum_{i=1}^n\|\operatorname{div}(\boldsymbol{U}^i-\boldsymbol{u}(\cdot,t_i))\|^2\leq\lambda\Delta t\sum_{i=1}^n\|\operatorname{div}(\boldsymbol{e}^i+P_E^h\boldsymbol{u}^i-\boldsymbol{u}^i)\|^2&\end{aligned}$$

and

$$\mu\Delta t\sum_{i=1}^n\|\boldsymbol{U}^i-\boldsymbol{u}(\cdot,t_i)\|_b^2\leq C\mu\Delta t\sum_{i=1}^n\|\boldsymbol{e}^i+P_E^h\boldsymbol{u}^i-\boldsymbol{u}^i\|_b^2.\qquad\square$$

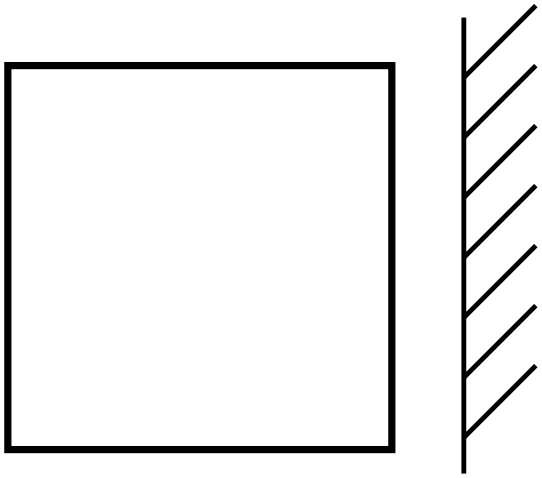


FIG. 1. *The contact problem setting.*

COROLLARY 3.2. *Let* $\Theta^0 = P_0^h\theta_0$, $\boldsymbol{U}^0 = P^h\boldsymbol{u}_0$, *and* $\Delta t = O(h)$. *Under the assumptions of Theorem* 3.1,

$$E(\theta, \boldsymbol{u}) \to 0 \quad as \quad h \to 0.$$

Proof. We follow Ciarlet [7]. Given $\varepsilon^{'} > 0$, since $\mathbf{H}^2(\Omega) \cap \mathbf{H}_E^1(\Omega)$ is dense in $\mathbf{H}_E^1(\Omega)$, there exists $\tilde{\boldsymbol{u}} \in \mathbf{H}^2(\Omega) \cap \mathbf{H}_E^1(\Omega)$ such that

$$\|\boldsymbol{u}^n - \tilde{\boldsymbol{u}}\|_1^2 \leq \varepsilon^{'}.$$

Thus, using (2.1) and (2.4),

$$\|\boldsymbol{u}^n - P_E^h\boldsymbol{u}^n\|_1^2 \leq C\|\boldsymbol{u}^n - \tilde{\boldsymbol{u}} + \tilde{\boldsymbol{u}} - I^h\tilde{\boldsymbol{u}}\|_1^2 \leq C(\varepsilon^{'} + h^2\|\tilde{\boldsymbol{u}}\|_2^2).$$

Recalling (2.2) and (2.3), the convergence result is obtained taking $\varepsilon^{'} = O(h^2)$. ∎

COROLLARY 3.3. *Suppose that* $\boldsymbol{u} \in L^\infty(0, T; \mathbf{H}^2(\Omega))$, $\theta_0 \in H^2(\Omega)$, *and* $\boldsymbol{u}_0 \in \mathbf{H}^2(\Omega)$. *Let* $\Theta^0 = P_0^h\theta_0$ *and* $\boldsymbol{U}^0 = P^h\boldsymbol{u}_0$. *Then the error estimate*

$$E(\theta, \boldsymbol{u}) \leq C(\lambda, \mu, \epsilon)\left(\frac{h^4}{\Delta t} + (\Delta t)^2 + h^4 + \frac{h^2}{\Delta t} + h^2\right)$$

holds and the error bound is $O(h)$ *if* $\Delta t = h$.

Proof. The estimate is a consequence of the additional assumed regularity. ∎

4. Numerical experiments. In our experiments, the reference configuration is the square $\Omega = (0,1) \times (0,1)$ with $\Gamma_C = \{(1, x_2) \,|\, 0 < x_2 < 1\}$ and Γ_0 the remaining boundary of Ω. The elastic obstacle is located at distance $g(\mathbf{x}) = 0.02$ from Γ_C (see Figure 1) and the initial temperature Θ^0 is the linear interpolant of $\theta_0(\mathbf{x}) = 20000\, x_1 x_2 (x_1 - 1)(x_2 - 1)$. We let $\lambda = 1$, $\mu = 30$, and $m = 0.008$, and tolerance of 1×10^{-7} was used to stop the iterative procedures. A uniform triangulation of Ω was obtained by dividing the square into $M \times M$ squares with side $h = 1/M$ and connecting the north-west vertex to the south-east vertex of each square.

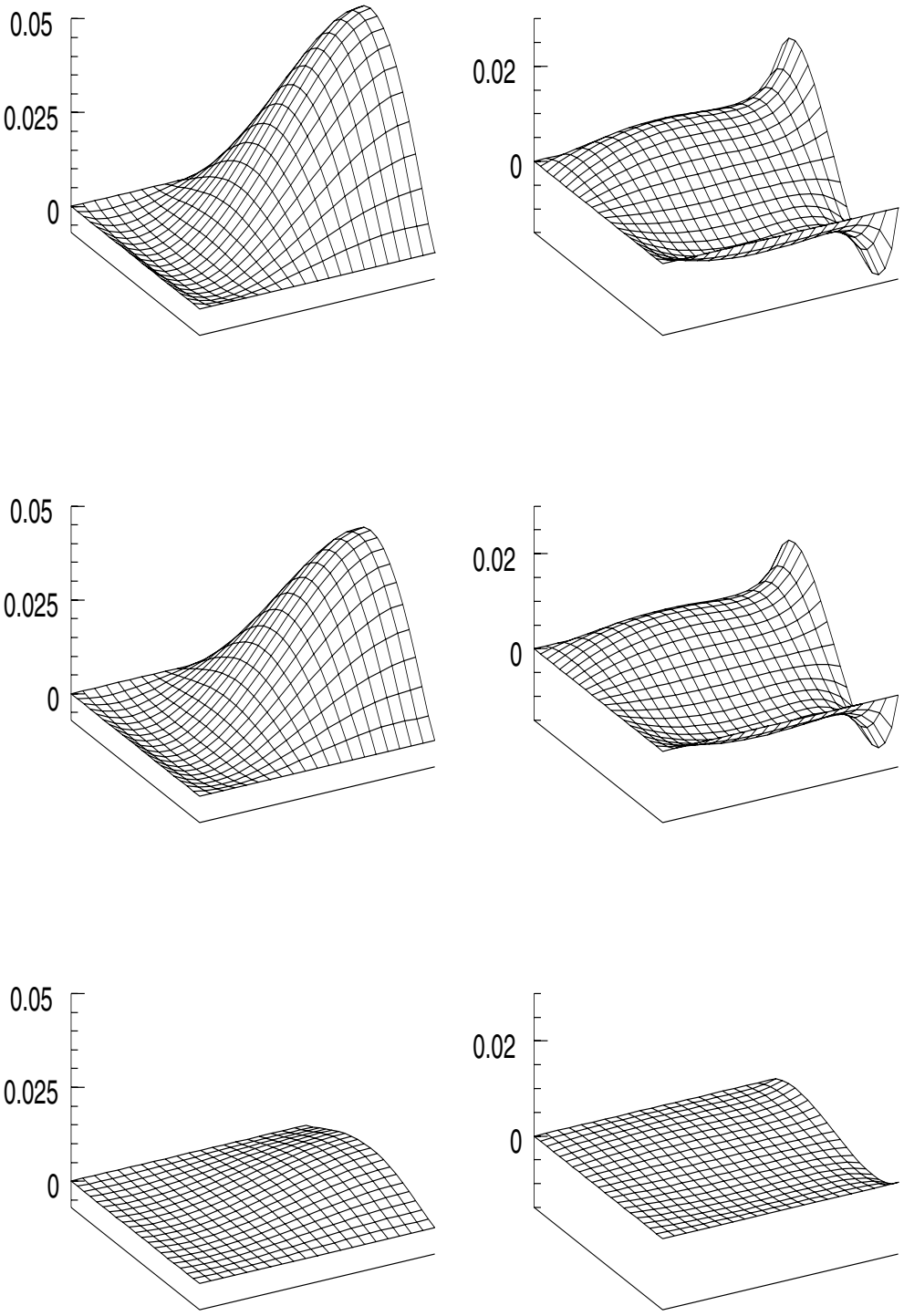


FIG. 2. *The evolution of the displacement components when* $\epsilon = 0.1$. *The left and right columns show* u_1 *and* u_2, *respectively, at* $t = 0,\ 0.01$, *and* 0.1.

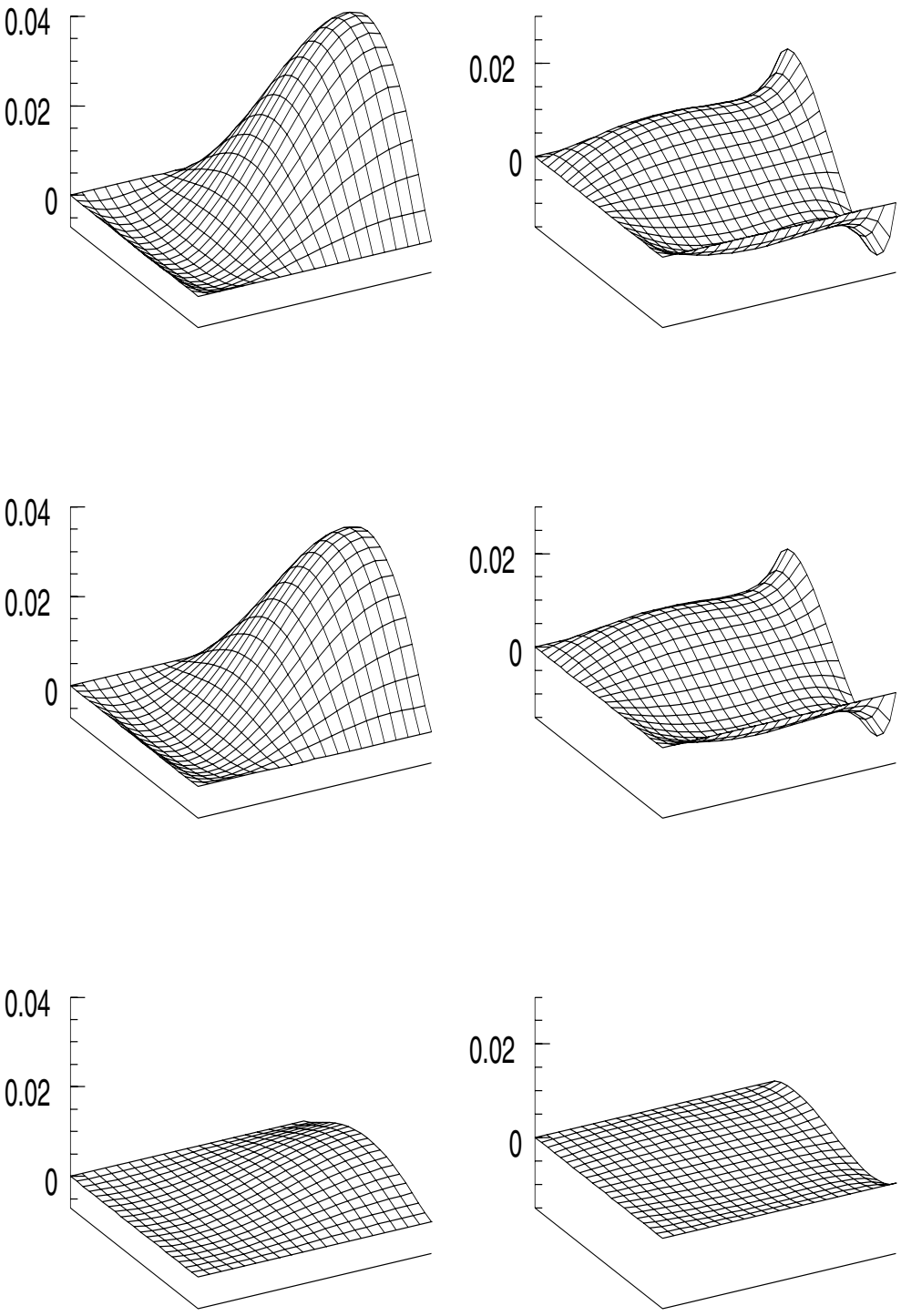


FIG. 3. *The evolution of the displacement components when* $\epsilon = 0.01$. *The left and right columns show* u_1 *and* u_2, *respectively, at* $t = 0,\ 0.01,$ *and* 0.1.

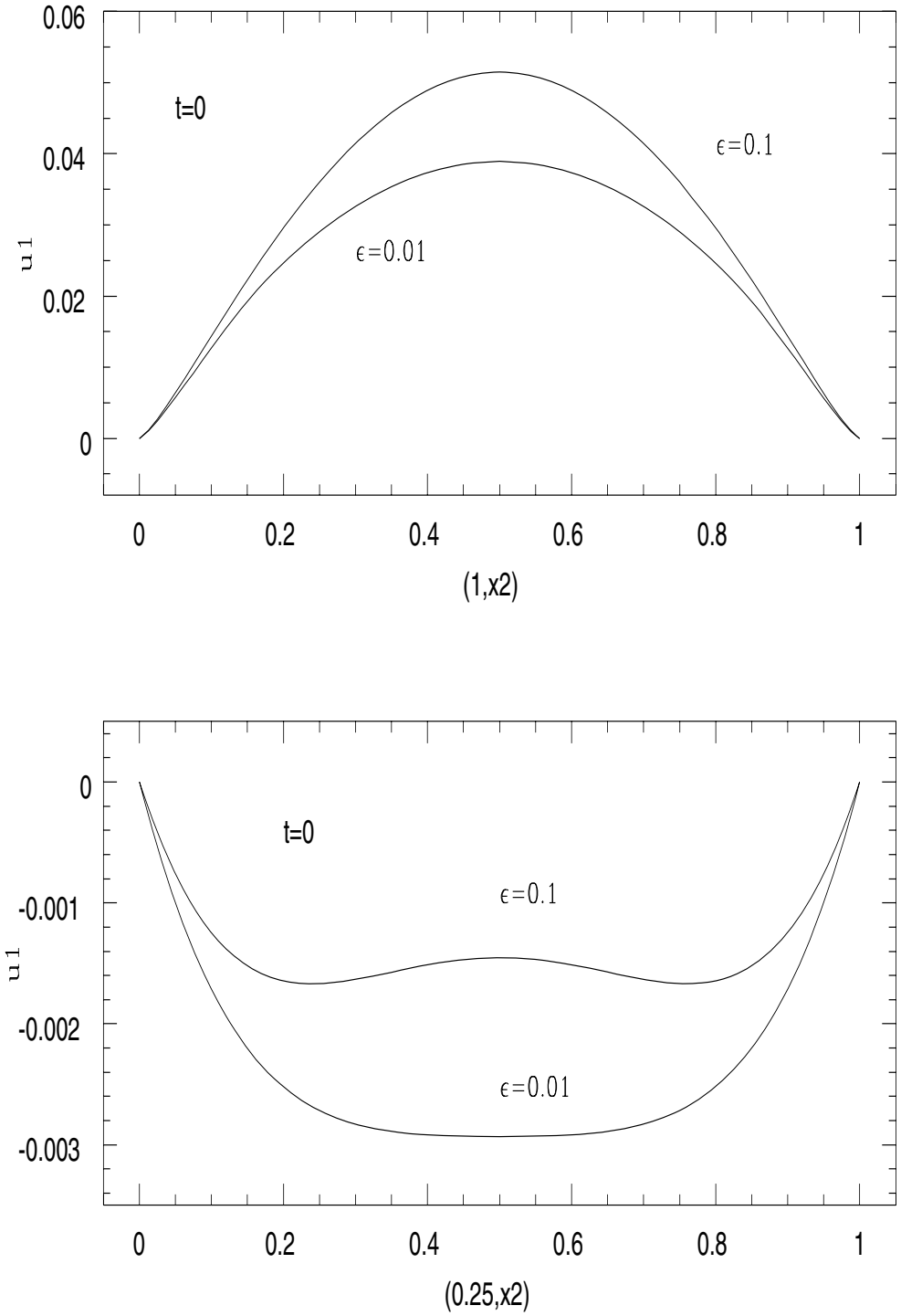


FIG. 4. *The displacement component* u_1 *at* $t = 0$ *for* $\mathbf{x} = (1, x_2)$ *and* $\mathbf{x} = (0.25, x_2)$.

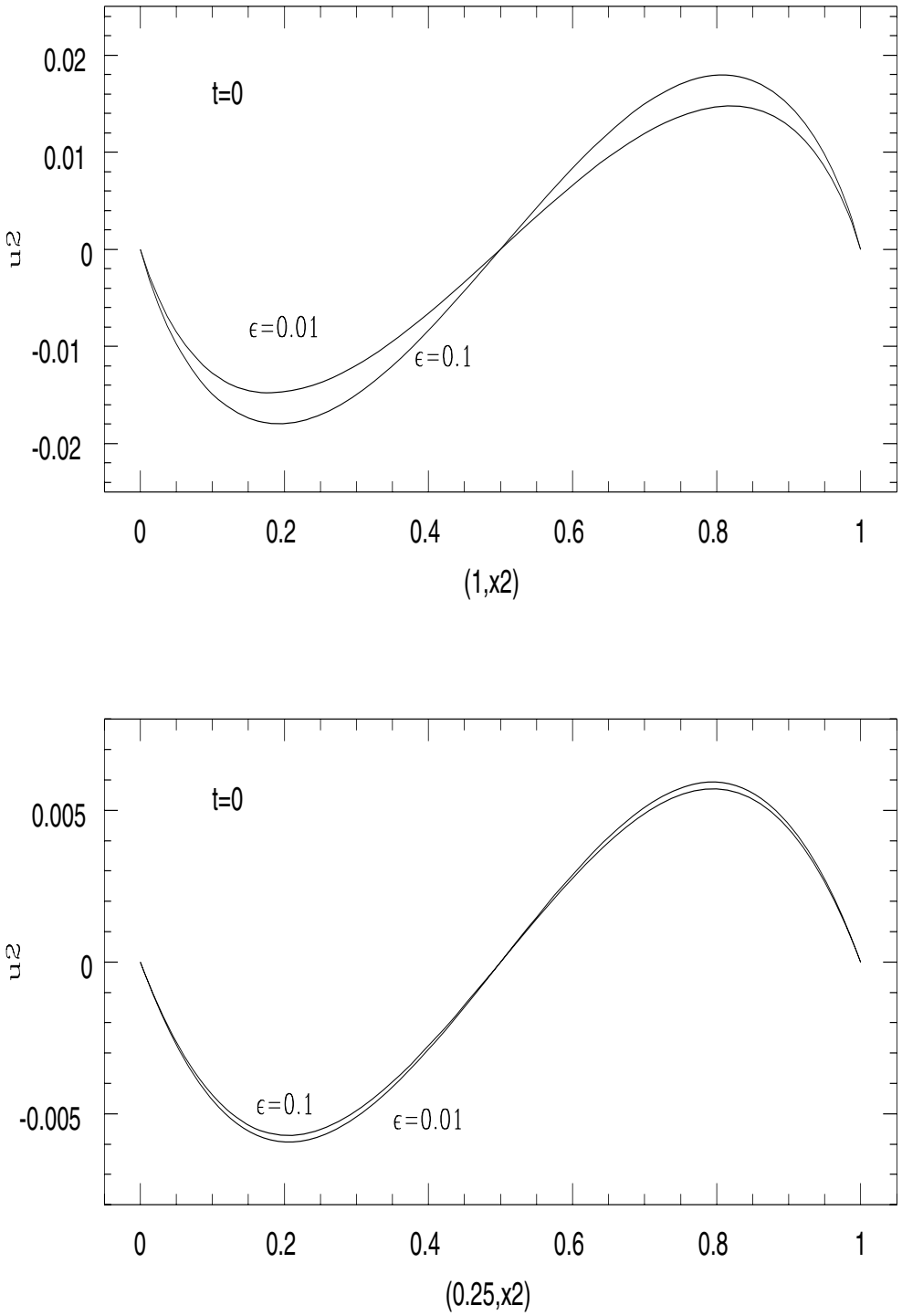


FIG. 5. *The displacement component* u_2 *at* $t = 0$ *for* $\mathbf{x} = (1, x_2)$ *and* $\mathbf{x} = (0.25, x_2)$.

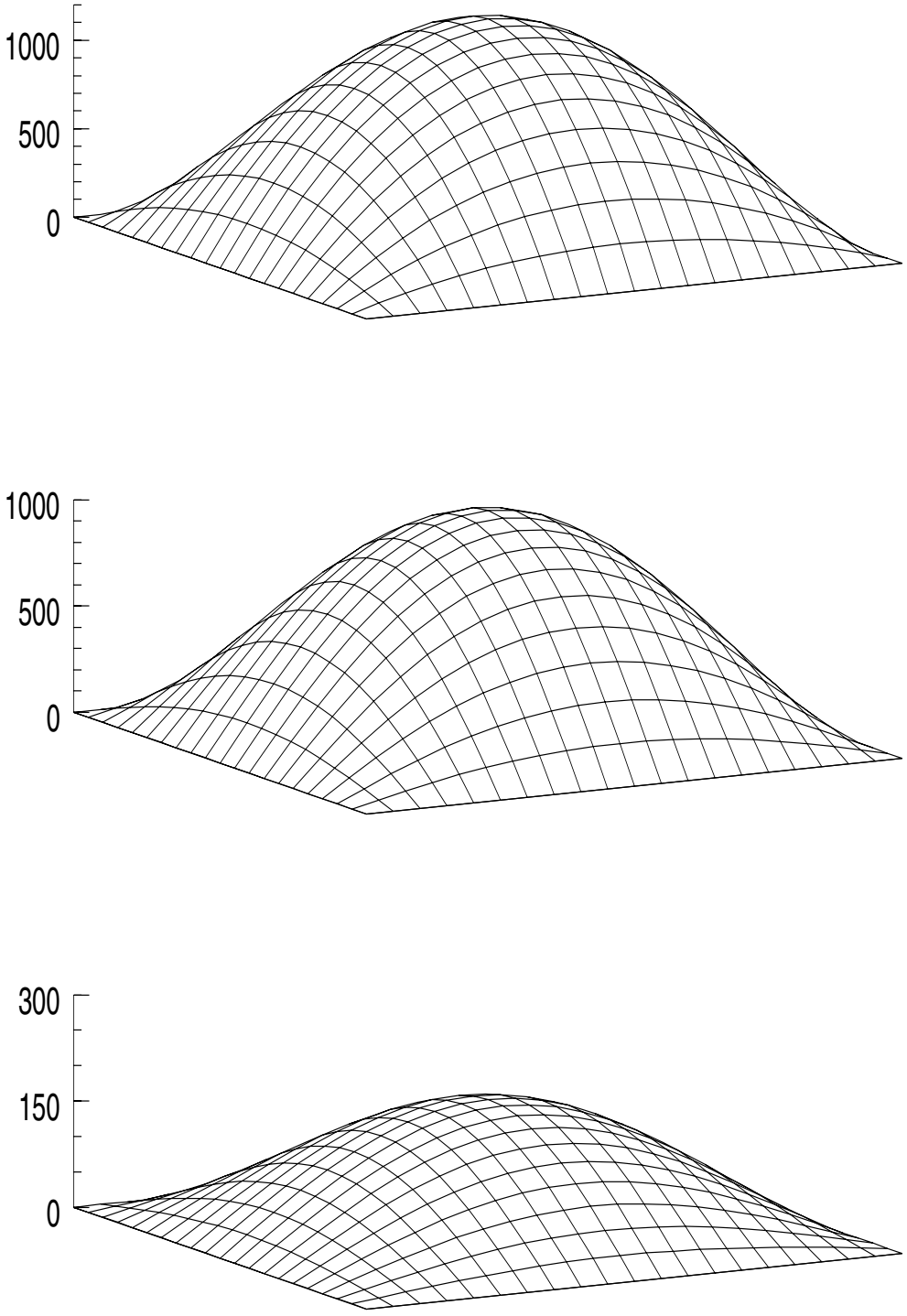


FIG. 6. *The temperature at* $t = 0,\ 0.01,$ *and* 0.1.

TABLE 1
The computed errors for $T = 0.2$ and $\epsilon = 0.1$.

$h = \Delta t$	$\Delta t \sum_{i=1}^{N} \|\Theta^i - \theta(\cdot, t_i)\|^2$	Rate	$\Delta t \sum_{i=1}^{N} \|U^i - u(\cdot, t_i)\|_b^2$	Rate
1/10	2.05×10^3		1.68×10^{-5}	
1/20	8.29×10^2	1.30	6.21×10^{-6}	1.44
1/40	2.61×10^2	1.67	1.64×10^{-6}	1.92
1/80	7.39×10	1.82	5.25×10^{-7}	1.64
1/160	1.98×10	1.90	2.51×10^{-7}	1.07

In the first experiment, $M = 80$, $\Delta t = 1 \times 10^{-5}$, and we took two values for the parameter ϵ, $\epsilon = 0.1$ and $\epsilon = 0.01$. The results are shown in Figures 2–6. At $t = 0$ the body is in contact with the obstacle and the evolution is toward the state $\theta = 0$, $\boldsymbol{u} = \mathbf{0}$. As expected, when ϵ decreases the obstacle becomes more rigid and penetration is more difficult. Note that there is a region where the body has contracted. The temperature profiles were virtually the same in both cases.

In order to test the error estimates, we compared the numerical solutions on coarse meshes with the solution obtained with the finer mesh ($M = 320$ and $\Delta t = 1 \times 10^{-5}$). The computed errors for $T = 0.2$ and $\epsilon = 0.1$ are reported in Table 1, where we observe convergence rates larger than those given by Corollary 3.3.

5. Remark on the Signorini problem. In [11], Copetti and French considered the numerical approximation of the one dimensional Signorini problem with viscosity effects and proved error estimates using the decomposition

$$\theta - \Theta^n = \theta - \theta^\epsilon + \theta^\epsilon - \Theta^n,$$
$$u - U^n = u - u^\epsilon + u^\epsilon - U^n,$$

where $\{\theta, u\}$ represents the solution to the Signorini problem, $\{\theta^\epsilon, u^\epsilon\}$ the solution to the penalized problem, and $\{\Theta^n, U^n\}$ the numerical solution based on the penalization.

In the two dimensional case, similar results can be obtained if the solutions to the continuous problems are sufficiently smooth. We can show that

$$\int_0^T \|\theta - \theta^\epsilon\|^2 dt + \frac{1}{2} \left\| \int_0^T \nabla(\theta - \theta^\epsilon) dt \right\|^2 + \lambda \int_0^T \|\mathrm{div}\,(\boldsymbol{u} - \boldsymbol{u}^\epsilon)\|^2 dt$$
$$+ 2\mu \int_0^T \|\boldsymbol{u} - \boldsymbol{u}^\epsilon\|_b^2 dt + \frac{1}{\epsilon} \int_0^T \|u_\nu - u_\nu^\epsilon\|_{L^2(\Gamma_c)}^2 dt \le \frac{\epsilon}{2} \int_0^T \|\sigma_\nu\|_{L^2(\Gamma_c)}^2 dt,$$

which together with Corollary 3.3 yields error bounds for the numerical approximation of (1.1)–(1.3) and (1.5)–(1.6) by the discrete solution $\{\Theta^n, \boldsymbol{U}^n\}$.

Acknowledgment. The author wishes to thank the referees for their valuable comments which improved the manuscript.

REFERENCES

[1] K. A. AMES AND L. E. PAYNE, *Uniqueness and continuous dependence of solutions to a multidimensional thermoelastic contact problem*, J. Elasticity, 34 (1994), pp. 139–148.
[2] M. BARBOTEU, W. HAN, AND M. SOFONEA, *Numerical analysis of a bilateral frictional contact problem for linearly elastic materials*, IMA J. Numer. Anal., 22 (2002), pp. 407–436.
[3] M. BIEN, *Existence of global weak solutions for coupled thermoelasticity with Barber's heat exchange condition*, J. Appl. Anal., 9 (2003), pp. 163–186.

[4] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, New York, 1994.
[5] M. Campo, J. R. Fernández, and J. M. Viaño, *Numerical analysis and simulations of a quasi-static frictional contact problem with damage in viscoelasticity*, J. Comput. Appl. Math., 192 (2006), pp. 30–39.
[6] O. Chau and J. R. Fernández, *A convergence result in elastic-viscoplastic contact problems with damage*, Math. Comput. Modelling, 37 (2003), pp. 301–321.
[7] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II, P. G. Ciarlet and J. L. Lions, eds., Handb. Numer. Anal. II, North–Holland, Amsterdam, 1991, pp. 17–351.
[8] M. I. M. Copetti, *Finite element approximation to a contact problem in linear thermoelasticity*, Math. Comp., 68 (1999), pp. 1013–1024.
[9] M. I. M. Copetti, *A one-dimensional thermoelastic problem with unilateral constraint*, Math. Comput. Simulation, 59 (2002), pp. 361–376.
[10] M. I. M. Copetti and C. M. Elliott, *A one-dimensional quasi-static contact problem in linear thermoelasticity*, European J. Appl. Math., 4 (1993), pp. 151–174.
[11] M. I. M. Copetti and D. A. French, *Numerical solution of a thermoviscoelastic contact problem by a penalty method*, SIAM J. Numer. Anal., 41 (2003), pp. 1487–1504.
[12] R. P. Gilbert, P. Shi, and M. Shillor, *A quasi-static contact problem in linear thermoelasticity*, Rend. Mat. Appl. (7), 10 (1990), pp. 785–808.
[13] I. Hlaváček and J. Nedoma, *On a solution of a generalized semi-coercive contact problem in thermo-elasticity*, Math. Comput. Simulation, 60 (2002), pp. 1–17.
[14] J. A. Pelesko, *Nonlinear stability, thermoelastic contact, and the Barber condition*, Trans. ASME J. Appl. Mech., 68 (2001), pp. 28–33.
[15] J. E. M. Rivera and R. Racke, *Multidimensional contact problems in thermoelasticity*, SIAM J. Appl. Math., 58 (1998), pp. 1307–1337.
[16] P. Shi and M. Shillor, *Uniqueness and stability of the solution to a thermoelastic contact problem*, European J. Appl. Math., 1 (1990), pp. 371–387.
[17] P. Shi and M. Shillor, *Existence of a solution to the n dimensional problem of thermoelastic contact*, Comm. Partial Differential Equations, 17 (1992), pp. 1597–1618.
[18] V. Thomée, *Galerkin Finite Element Method for Parabolic Problems*, Lecture Notes in Math. 1054, Springer-Verlag, Berlin, 1984.
[19] X. Xu, *The N-dimensional quasistatic problem of thermoelastic contact with Barber's heat exchange condition*, Adv. Math. Sci. Appl., 6 (1996), pp. 559–587.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2653–2670

SPECTRAL DISCRETIZATION OF A NAGHDI SHELL MODEL*

CHRISTINE BERNARDI† AND ADEL BLOUZA‡

Abstract. We consider the Naghdi equations which model a thin three-dimensional shell. We propose a spectral discretization of this problem in the case where the midsurface of the shell is weakly regular. We perform the numerical analysis of the discrete problem and prove optimal error estimates.

Key words. spectral methods, Naghdi shell model

AMS subject classifications. 65N35, 74K25, 74S25

DOI. 10.1137/050642058

1. Introduction.

We consider a formulation of Naghdi's shell model in Cartesian coordinates which is appropriate for linearly elastic shells that present curvature discontinuities. The aim of this paper is to propose a spectral discretization of the mixed formulation of this problem and to perform its numerical analysis.

The formulation of Naghdi's model which is used here was introduced in [7], [11]. This formulation relies on the idea of using a local basis-free formulation in which the unknowns are described in Cartesian coordinates instead of covariant or contravariant components as is usually done in shell theory; see, for example, [2]. Such a formulation is able to handle shells with a $W^{2,\infty}$-midsurface, thus allowing for curvature discontinuities, as opposed to $\mathscr{C}^3$ in the classical formalism (see [15, Chap. 7] and the references therein), and leads to much simpler expressions. Even though it is proved in [11] to be well-posed and to be the natural limit of the classical formulation when a sequence of regular midsurfaces converges to a $W^{2,\infty}$-midsurface, the new formulation has not been used in a numerical spectral setting to the best of our knowledge. For simplicity, we only consider the case of a shell with a $W^{2,\infty}$-midsurface.

The literature on finite element approximation of two-dimensional shell models is large. Let us mention a few approaches. Concerning conforming methods, the Ganev and Argyris triangles provide interpolation by polynomials of degree 4 and 5, with high order convergence in $c\,h^4$ when the solution is smooth enough. These elements are used, for example, to study the linear Koiter model for $\mathscr{C}^3$-shells in the classical covariant formulation; see [1, Part II, Chap. 1]. This method is applied to approximate geometrically exact shell models in [13]. The Argyris elements are also used in [18] for numerical analysis of Koiter's model with little regularity in the Cartesian formulation proposed in [10]. We also mention the three-dimensional shell element approach; see [14]. Still in the context of shells with little regularity, i.e., when the midsurface is of $W^{2,\infty}$-regularity, a nonconforming discrete Kirchhoff triangle (DKT) element is used in [21] to approximate a Koiter equation similar to the model introduced in [10]. Other works [19], [20] concern the finite element discretization of shell problems with domain decomposition. The main difficulty of all these discretizations is that, in

*Received by the editors October 6, 2005; accepted for publication (in revised form) June 22, 2007; published electronically December 7, 2007.
http://www.siam.org/journals/sinum/45-6/64205.html

†Laboratoire Jacques-Louis Lions, C.N.R.S. & Université Pierre et Marie Curie, B.C. 187, 4 place Jussieu, 75252 Paris Cedex 05, France (bernardi@ann.jussieu.fr).

‡Laboratoire de Mathématiques Raphaël Salem (U.M.R. 6085 C.N.R.S.), Université de Rouen, avenue de l'Université, B.P. 12, 76801 Saint-Étienne-du-Rouvray, France (Adel.Blouza@univ-rouen.fr).

most situations, a locking phenomenon appears when the choice of the discretization parameter is not sufficient to handle the thickness of the shell.

In this paper, we propose a spectral discretization of Naghdi's model. It relies on a mixed variational formulation of the corresponding equation proposed in [9]: A Lagrange multiplier is introduced to enforce the tangency requirement on one of the unknowns. A further penalization term is also added in order to stabilize the system. We first describe the discrete problem which is constructed from the variational formulation of the model by the Galerkin method with numerical integration (see [4, sect. 15] or [6, Chap. V] for a detailed presentation of this procedure). Under some further but likely regularity assumptions on the midsurface of the shell, we prove that it is well-posed. Finally, relying on standard polynomial approximation and interpolation results, we prove error estimates which are fully optimal from a numerical point of view for a fixed thickness of the shell.

The extension of this study to the case of a piecewise regular shell discretized by the spectral element method is under consideration. Numerical experiments should confirm the interest of this discretization.

An outline of the paper is as follows.

- In section 2, we recall the geometry of the midsurface and Naghdi's shell formulation. We introduce a mixed version of Naghdi's model intended to approximate the above mentioned tangency. We prove the existence and uniqueness of the solution.
- Section 3 is devoted to the description of the spectral discrete problem. We also prove its well-posedness.
- Error estimates are derived in section 4.

2. Presentation of the model. Greek indices and exponents take their values in the set $\{1,2\}$, and Latin indices and exponents take their values in the set $\{1,2,3\}$. Unless otherwise specified, the summation convention for repeated indices and exponents according to this set of values is assumed. Let $\{\boldsymbol{e}_1,\boldsymbol{e}_2,\boldsymbol{e}_3\}$ be the canonical orthonormal basis of the Euclidean space $\mathbb{R}^3$. We denote by $\boldsymbol{u}\cdot\boldsymbol{v}$ the inner product of $\mathbb{R}^3$, $|\boldsymbol{u}|=\sqrt{\boldsymbol{u}\cdot\boldsymbol{u}}$ the associated Euclidean norm, and $\boldsymbol{u}\wedge\boldsymbol{v}$ the vector product of $\boldsymbol{u}$ and $\boldsymbol{v}$.

Let ω be a bounded connected domain of $\mathbb{R}^2$ with a Lipschitz-continuous boundary $\partial\omega$. We consider a shell whose midsurface is given by $S=\varphi(\bar\omega)$, where φ is a one-to-one mapping in $W^{2,\infty}(\omega)^3$ such that the two vectors $\boldsymbol{a}_\alpha(\boldsymbol{x})=(\partial_\alpha\varphi)(\boldsymbol{x})$ are linearly independent at each point $\boldsymbol{x}$ of $\bar\omega$. Thus,

$$\boldsymbol{a}_3(\boldsymbol{x})=\frac{\boldsymbol{a}_1(\boldsymbol{x})\wedge\boldsymbol{a}_2(\boldsymbol{x})}{|\boldsymbol{a}_1(\boldsymbol{x})\wedge\boldsymbol{a}_2(\boldsymbol{x})|}$$

is the unit normal vector on the midsurface at point $\varphi(\boldsymbol{x})$. The vectors $\boldsymbol{a}_i(\boldsymbol{x})$ define the local covariant basis at point $\varphi(\boldsymbol{x})$. The contravariant basis $\boldsymbol{a}^i(\boldsymbol{x})$ is defined by the relations $\boldsymbol{a}_i\cdot\boldsymbol{a}^j=\delta_i^j$, where δ_i^j is the Kronecker symbol. In particular $\boldsymbol{a}^3(\boldsymbol{x})$ coincides with $\boldsymbol{a}_3(\boldsymbol{x})$. Note that all of these vectors belong to $W^{1,\infty}(\omega)^3$. We set $a(\boldsymbol{x})=|\boldsymbol{a}_1(\boldsymbol{x})\wedge\boldsymbol{a}_2(\boldsymbol{x})|^2$ so that $\sqrt{a(\boldsymbol{x})}$ is the area element of the midsurface in the chart φ. Finally, the first fundamental form of the surface is given in covariant components by $a_{\alpha\beta}=\boldsymbol{a}_\alpha\cdot\boldsymbol{a}_\beta$.

Let $\boldsymbol{u}$ be a midsurface displacement in $H^1(\omega)^3$ and $\boldsymbol{r}$ be a rotation vector in $H^1(\omega)^3$ such that $\boldsymbol{r}$ is tangential to the midsurface. These functions are given in covariant and Cartesian components by

$$\boldsymbol{u}(\boldsymbol{x})=u_i(\boldsymbol{x})\boldsymbol{a}^i(\boldsymbol{x})=u_i^c(\boldsymbol{x})\boldsymbol{e}_i,\qquad\text{with}\quad u_i=\boldsymbol{u}\cdot\boldsymbol{a}_i\quad\text{and}\quad u_i^c=\boldsymbol{u}\cdot\boldsymbol{e}_i,$$

and

$$\boldsymbol{r}(\boldsymbol{x}) = r_\alpha(\boldsymbol{x})\boldsymbol{a}^\alpha(\boldsymbol{x}) = r_i^c(\boldsymbol{x})\boldsymbol{e}_i, \qquad \text{with} \quad r_\alpha = \boldsymbol{r}\cdot\boldsymbol{a}_\alpha \quad \text{and} \quad r_i^c = \boldsymbol{r}\cdot\boldsymbol{e}_i.$$

Note that the tangency requirement is easily expressed in covariant coordinates, as it simply reads $r_3 = 0$, whereas it becomes $r_i^c(\boldsymbol{x})a_{3,i}^c(\boldsymbol{x}) = 0$ in ω in Cartesian coordinates.

Let $a^{\alpha\beta\rho\sigma}$ denote the coefficients of the elasticity tensor. In the case of homogeneous, isotropic material with Young modulus $E > 0$ and Poisson coefficient ν, $0 \le \nu < \frac{1}{2}$, these coefficients are given by

$$a^{\alpha\beta\rho\sigma} = \frac{E}{2(1+\nu)}(a^{\alpha\rho}a^{\beta\sigma} + a^{\alpha\sigma}a^{\beta\rho}) + \frac{E\nu}{1-\nu^2}a^{\alpha\beta}a^{\rho\sigma}, \tag{2.1}$$

where $a^{\alpha\beta} = \boldsymbol{a}^\alpha\cdot\boldsymbol{a}^\beta$ are the contravariant components of the first fundamental form. We note that each coefficient of this tensor belongs to $L^\infty(\omega)$. Moreover, it satisfies the usual symmetry properties

$$a^{\alpha\beta\rho\sigma}(\boldsymbol{x}) = a^{\rho\sigma\alpha\beta}(\boldsymbol{x}) = a^{\beta\alpha\rho\sigma}(\boldsymbol{x}) \qquad \text{for a.e. } \boldsymbol{x}\in\omega \tag{2.2}$$

and is uniformly strictly positive: There exists a positive constant c_0 such that, for all symmetric tensors $\boldsymbol{\tau} = (\tau_{\alpha\beta})$ in $\mathbb{R}^{2\times 2}$,

$$a^{\alpha\beta\rho\sigma}(\boldsymbol{x})\tau_{\alpha\beta}\tau_{\rho\sigma} \ge c_0\,|\boldsymbol{\tau}|^2 \qquad \text{for a.e. } \boldsymbol{x}\in\omega. \tag{2.3}$$

In this context, the covariant components of the change of metric tensor read

$$\gamma_{\alpha\beta}(\boldsymbol{u}) = \frac{1}{2}(\partial_\alpha\boldsymbol{u}\cdot\boldsymbol{a}_\beta + \partial_\beta\boldsymbol{u}\cdot\boldsymbol{a}_\alpha), \tag{2.4}$$

the covariant components of the change of transverse shear tensor read

$$\delta_{\alpha 3}(\boldsymbol{u},\boldsymbol{r}) = \frac{1}{2}(\partial_\alpha\boldsymbol{u}\cdot\boldsymbol{a}_3 + \boldsymbol{r}\cdot\boldsymbol{a}_\alpha), \tag{2.5}$$

and the covariant components of the change of curvature tensor read

$$\chi_{\alpha\beta}(\boldsymbol{u},\boldsymbol{r}) = \frac{1}{2}(\partial_\alpha\boldsymbol{u}\cdot\partial_\beta\boldsymbol{a}_3 + \partial_\beta\boldsymbol{u}\cdot\partial_\alpha\boldsymbol{a}_3 + \partial_\alpha\boldsymbol{r}\cdot\boldsymbol{a}_\beta + \partial_\beta\boldsymbol{r}\cdot\boldsymbol{a}_\alpha); \tag{2.6}$$

see [7], [11]. Note that all these quantities make sense for shells with little regularity and are easily expressed with the Cartesian coordinates of the unknowns and geometrical data. For instance, we have

$$\partial_\alpha\boldsymbol{u}\cdot\boldsymbol{a}_\beta = (\partial_\alpha u_i^c)\,a_{\beta,i}^c,$$

and so on.

We assume that the boundary $\partial\omega$ of the chart domain is divided into two parts: γ_0, which has a finite number of connected components and a strictly positive one-dimensional measure and on which the shell is clamped, and the complementary part $\gamma_1 = \partial\omega\setminus\gamma_0$, on which the shell is subjected to applied tractions and moments.

To take into account the boundary conditions, we define the space

$$H^1_{\gamma_0}(\omega) = \big\{\mu\in H^1(\omega);\ \mu = 0 \text{ on } \gamma_0\big\}. \tag{2.7}$$

We also denote by $H^{\frac{1}{2}}_{00}(\gamma_1)$ the space of functions in $H^{\frac{1}{2}}(\gamma_1)$ such that their extension by zero to $\partial\omega$ belongs to $H^{\frac{1}{2}}(\partial\omega)$; see [22, Chap. 1, sect. 11]. Let us now consider the function space, introduced in [7], [11], which is appropriate in the context of

shells with little regularity

$$\mathbb{V}(\omega) = \big\{(\boldsymbol{v},\boldsymbol{s}) \in H^1_{\gamma_0}(\omega)^3 \times H^1_{\gamma_0}(\omega)^3;\ \boldsymbol{s}\cdot\boldsymbol{a}_3 = 0 \text{ in } \omega\big\}. \tag{2.8}$$

This space is endowed with the natural Hilbert norm

$$\|(\boldsymbol{v},\boldsymbol{s})\|_{\mathbb{V}(\omega)} = \big(\|\boldsymbol{v}\|^2_{H^1(\omega)^3} + \|\boldsymbol{s}\|^2_{H^1(\omega)^3}\big)^{1/2}. \tag{2.9}$$

We now recall the variational formulation of the problem corresponding to the linear Naghdi model for shells with little regularity. It reads:

Find $(\boldsymbol{u},\boldsymbol{r})$ *in* $\mathbb{V}(\omega)$ *such that*

$$\forall (\boldsymbol{v},\boldsymbol{s}) \in \mathbb{V}(\omega), \quad a\big((\boldsymbol{u},\boldsymbol{r});(\boldsymbol{v},\boldsymbol{s})\big) = \mathcal{L}((\boldsymbol{v},\boldsymbol{s})), \tag{2.10}$$

where the bilinear form $a(\cdot;\cdot)$ is defined by

$$\begin{aligned} a\big((\boldsymbol{u},\boldsymbol{r});(\boldsymbol{v},\boldsymbol{s})\big) = \int_\omega \Big\{ e a^{\alpha\beta\rho\sigma}\Big[\gamma_{\rho\sigma}(\boldsymbol{v}) + \frac{e^2}{12}\chi_{\alpha\beta}(\boldsymbol{u},\boldsymbol{r})\chi_{\rho\sigma}(\boldsymbol{v},\boldsymbol{s})\Big] \\ + 2e\frac{E}{1+\nu} a^{\alpha\beta}\delta_{\alpha 3}(\boldsymbol{u},\boldsymbol{r})\delta_{\beta 3}(\boldsymbol{v},\boldsymbol{s})\Big\}\sqrt{a}\, d\boldsymbol{x}, \end{aligned} \tag{2.11}$$

and the linear form $\mathcal{L}(\cdot)$ is given by

$$\mathcal{L}((\boldsymbol{v},\boldsymbol{s})) = \int_\omega \boldsymbol{f}\cdot\boldsymbol{v}\,\sqrt{a}\, d\boldsymbol{x} + \int_{\gamma_1} (\boldsymbol{M}\cdot\boldsymbol{v} + \boldsymbol{N}\cdot\boldsymbol{s})\, d\tau. \tag{2.12}$$

The three terms in $a(\cdot,\cdot)$ represent the membrane, bending, and shear deformations, respectively. The data $\boldsymbol{f}$, $\boldsymbol{M}$, and $\boldsymbol{N}$ represent a given resultant force density, an applied moment density, and an applied traction density (so that usually $\boldsymbol{N}\cdot\boldsymbol{a}_3 = 0$), respectively. Finally the thickness of the shell which is assumed to be constant is denoted by e, $0 < e < 1$.

We refer to [7], [11] for the proof of the well-posedness of this problem which is stated in the next theorem. Since $\boldsymbol{a}_3$ belongs to $W^{1,\infty}(\omega)^3$, the form $a(\cdot;\cdot)$ is obviously continuous on $\mathbb{V}(\omega)\times\mathbb{V}(\omega)$, with the norm smaller than $c\,e$. Similarly, the form $\mathcal{L}$ is continuous on $\mathbb{V}(\omega)$, and its norm satisfies, with obvious notation,

$$\|\mathcal{L}\| \le c\,\big(\|\boldsymbol{f}\|_{H^1_{\gamma_0}(\omega)^{3\prime}} + \|\boldsymbol{M}\|_{H^{\frac12}_{00}(s\gamma_1)^{3\prime}} + \|\boldsymbol{N}\|_{H^{\frac12}_{00}(\gamma_1)^{3\prime}}\big). \tag{2.13}$$

So the well-posedness mainly relies on the following ellipticity property which is proved in [11, Lem. 3.6]: There exists a constant $c_* > 0$ such that

$$\forall (\boldsymbol{v},\boldsymbol{s}) \in \mathbb{V}(\omega), \quad a\big((\boldsymbol{v},\boldsymbol{s});(\boldsymbol{v},\boldsymbol{s})\big) \ge c_*\, e^3\, \|(\boldsymbol{v},\boldsymbol{s})\|^2_{\mathbb{V}(\omega)}. \tag{2.14}$$

THEOREM 2.1. *For any data* $(\boldsymbol{f},\boldsymbol{M},\boldsymbol{N})$ *in* $H^1_{\gamma_0}(\omega)^{3\prime} \times H^{\frac12}_{00}(\gamma_1)^{3\prime} \times H^{\frac12}_{00}(\gamma_1)^{3\prime}$, *problem* (2.10) *admits a unique solution* $(\boldsymbol{u},\boldsymbol{r})$ *in* $\mathbb{V}(\omega)$. *Moreover this solution satisfies*

$$\|(\boldsymbol{u},\boldsymbol{r})\|_{\mathbb{V}(\omega)} \le c\, e^{-3}\, \|\mathcal{L}\|. \tag{2.15}$$

However, since the purpose of the present work is to approximate the solution of problem (2.10) with a spectral method and to proceed in the simplest possible way, we immediately encounter a problem: The tangency constraint $\boldsymbol{s}\cdot\boldsymbol{a}_3 = 0$ which appears in the definition of $\mathbb{V}(\omega)$ clearly cannot be implemented in a standard way for a general shell. So the idea, already proposed in [9], consists in handling this constraint via the introduction of a Lagrange multiplier. We thus consider a mixed Naghdi problem in which the unknowns are the displacement $\boldsymbol{u}$ and the rotation $\boldsymbol{r}$, which belong to $H^1_{\gamma_0}(\omega)^3$ without any orthogonality constraint on $\boldsymbol{r}$, and the Lagrange

multiplier λ, which belongs to the space $H^1_{\gamma_0}(\omega)$ and is aimed to enforce the tangency constraint $\boldsymbol{r}\cdot\boldsymbol{a}_3=0$. In view of the discretization, we also possibly add a stabilizing term. Its usefulness appears in the next section.

Let us introduce the relaxed function space

$$\mathbb{X}(\omega)=H^1_{\gamma_0}(\omega)^3\times H^1_{\gamma_0}(\omega)^3, \tag{2.16}$$

still equipped with the norm defined in (2.9) which is now denoted by $\|\cdot\|_{\mathbb{X}(\omega)}$. We also set $\mathbb{M}(\omega)=H^1_{\gamma_0}(\omega)$. For simplicity, we use an extension of the forms $a(\cdot;\cdot)$ and $\mathcal{L}(\cdot)$ defined in (2.11) and (2.12), respectively, to $\mathbb{X}(\omega)\times\mathbb{X}(\omega)$ and $\mathbb{X}(\omega)$.

For a nonnegative parameter η, we consider the variational problem:

Find (U^η,ψ^η) *in* $\mathbb{X}(\omega)\times\mathbb{M}(\omega)$ *such that*

$$\begin{array}{ll}\forall V\in\mathbb{X}(\omega), & a(U^\eta;V)+\eta\,\tilde a(U^\eta;V)+b(V;\psi^\eta)=\mathcal{L}(V),\\ \forall\chi\in\mathbb{M}(\omega), & b(U^\eta;\chi)=0,\end{array} \tag{2.17}$$

where the bilinear forms $\tilde a(\cdot;\cdot)$ and $b(\cdot;\cdot)$ are defined by, with the notation $U=(\boldsymbol{u},\boldsymbol{r})$ and $V=(\boldsymbol{v},\boldsymbol{s})$,

$$\tilde a(U;V)=\int_\omega \partial_\alpha(\boldsymbol{r}\cdot\boldsymbol{a}_3)\partial_\alpha(\boldsymbol{s}\cdot\boldsymbol{a}_3)\,d\boldsymbol{x},\quad b(V;\chi)=\int_\omega\partial_\alpha(\boldsymbol{s}\cdot\boldsymbol{a}_3)\partial_\alpha\chi\,d\boldsymbol{x}. \tag{2.18}$$

Remark. Since we are aiming for simplicity of implementation, we have made no attempt to make the duality term intrinsic. In fact, it does depend on the chart, whereas the other terms do not. This could arguably be considered to be a poor choice, especially if a chart was used that gave much more weight to one part of the shell compared to the rest. An intrinsic choice that obviously works is

$$\tilde b(V;\chi)=\int_\omega a^{\alpha\beta}\partial_\alpha(\boldsymbol{s}\cdot\boldsymbol{a}_3)\partial_\beta\chi\,\sqrt{a}\,d\boldsymbol{x}.$$

Remark. In the case $\eta=0$, it can be checked that $\psi=\psi^0$ is the solution of the Laplace equation

$$-\sum_{\alpha=1}^{2}\partial^2_\alpha\psi=\frac{e^3}{12}a^{\alpha\beta\rho\sigma}\,\chi_{\alpha\beta}(\boldsymbol{u},\boldsymbol{r})\,b_{\rho\sigma}\,\sqrt{a}\quad\text{in }\omega,$$

with mixed Dirichlet boundary conditions on γ_0 and Neumann conditions on γ_1 and the $b_{\rho\sigma}$ are the components of the second fundamental form of the surface. Thus the function ψ seems to have no physical meaning and is only useful to handle the tangency condition.

It must be observed that, since $\boldsymbol{a}_3$ belongs to $W^{1,\infty}(\omega)^3$, the forms $\tilde a(\cdot;\cdot)$ and $b(\cdot;\cdot)$ are continuous on $\mathbb{X}(\omega)\times\mathbb{X}(\omega)$ and $\mathbb{X}(\omega)\times\mathbb{M}(\omega)$, respectively. Moreover, the following identity is readily checked:

$$\mathbb{V}(\omega)=\big\{V=(\boldsymbol{v},\boldsymbol{s})\in\mathbb{X}(\omega);\ \forall\chi\in\mathbb{M}(\omega),\ b(V;\chi)=0\big\}. \tag{2.19}$$

The next ellipticity property

$$\forall V\in\mathbb{V}(\omega),\quad a(V;V)+\eta\,\tilde a(V;V)\ge c_*\,e^3\,\|V\|^2_{\mathbb{X}(\omega)}, \tag{2.20}$$

is an obvious consequence of (2.14), and it can be checked thanks to exactly the same argument as in [9] that an analogous property still holds for all V in $\mathbb{X}(\omega)$ whenever η is positive. We now investigate the inf-sup condition on the form $b(\cdot;\cdot)$.

PROPOSITION 2.2. *There exists a positive constant $c_\sharp$ such that the following inf-sup condition holds:*

$$\forall \chi \in \mathbb{M}(\omega), \quad \sup_{V \in \mathbb{X}(\omega)} \frac{b(V;\chi)}{\|V\|_{\mathbb{X}(\omega)}} \geq c_\sharp \, \|\chi\|_{H^1(\omega)}. \tag{2.21}$$

Proof. Let χ be an arbitrary element of $\mathbb{M}(\omega)$. Since χ vanishes on γ_0 and $\boldsymbol{a}_3$ belongs to $W^{1,\infty}(\omega)^3$, it is readily checked that $V = (\mathbf{0}, \chi \mathbf{a}_3)$ belongs to $\mathbb{X}(\omega)$. Using the fact that $\chi \, \mathbf{a}_3 \cdot \mathbf{a}_3$ is equal to χ, we have with this choice of V

$$b(V;\chi) \geq |\chi|^2_{H^1(\omega)},$$

so that, thanks to a generalized Poincaré–Friedrichs inequality,

$$b(V;\chi) \geq c \, \|\chi\|^2_{H^1(\omega)}.$$

On the other hand, we observe that, owing to the regularity of $\mathbf{a}_3$,

$$\|V\|_{\mathbb{X}(\omega)} \leq \|\chi \mathbf{a}_3\|_{H^1(\omega)^3} \leq c \, \|\chi\|_{H^1(\omega)}.$$

Combining the last two inequalities gives the desired inf-sup condition. ∎

We are now in a position to prove the main result of this section.

THEOREM 2.3. *For any data $(\mathbf{f}, \mathbf{M}, \mathbf{N})$ in $H^1_{\gamma_0}(\omega)^{3\prime} \times H^{\frac{1}{2}}_{00}(\gamma_1)^{3\prime} \times H^{\frac{1}{2}}_{00}(\gamma_1)^{3\prime}$, problem (2.17) admits a unique solution (U^η, ψ^η) in $\mathbb{X}(\omega) \times \mathbb{M}(\omega)$. Moreover this solution satisfies*

$$\|U^\eta\|_{\mathbb{X}(\omega)} + \|\psi^\eta\|_{H^1(\omega)} \leq c \, e^{-3} \, \|\mathcal{L}\|. \tag{2.22}$$

Moreover, the part U^η of this solution is equal to the solution U of problem (2.10).

Proof. The existence and uniqueness of the solution (U^η, ψ^η), together with estimate (2.22), are a direct consequence of properties (2.20) and (2.21); see [12, Chap. II, Thm. 1.1] and [16, Chap. I, Cor. 4.1], for instance. Moreover, U^η belongs to $\mathbb{V}(\omega)$, and it is readily checked that, for any V in $\mathbb{V}(\omega)$, $\tilde{a}(U^\eta; V)$ is zero, so that U^η satisfies (2.10). ∎

3. The spectral discrete problem and its well-posedness. To describe the discrete problem, we now assume that ω is the square $]-1,1[^2$ (this can induce a further diffeomorphism; however, for simplicity we keep the notation φ for the chart). We assume that γ_0 is the union of one, two, three, or four whole edges of ω.

For each nonnegative integer n, we denote by $\mathbb{P}_n(\omega)$ the space of restrictions to ω of polynomials with two variables and degree $\leq n$ with respect to each variable. In order to take into account the boundary conditions of the problem, we introduce the space $\mathbb{P}^{\gamma_0}_n(\omega) = \mathbb{P}_n(\omega) \cap H^1_{\gamma_0}(\omega)$. Next, for a fixed integer $N \geq 2$ and another integer L, $2 \leq L \leq N$, we define the discrete spaces

$$\mathbb{X}_N = \mathbb{P}^{\gamma_0}_N(\omega)^3 \times \mathbb{P}^{\gamma_0}_N(\omega)^3, \qquad \mathbb{M}_N = \mathbb{P}^{\gamma_0}_L(\omega). \tag{3.1}$$

The reason for using two different parameters L and N is explained later.

We also make use of the Gauss–Lobatto formula on the interval $]-1,1[$. Let $\mathbb{P}_n(-1,1)$ denote the space of restrictions to $]-1,1[$ of polynomials with degree $\leq n$. For a third integer $M \geq N$, we set: $\xi_0 = -1$ and $\xi_M = 1$. We recall that there exist

$M-1$ nodes ξ_j, $1 \le j \le M-1$, in $]-1,1[$, with $\xi_0 < \xi_1 < \cdots < \xi_M$, and $M+1$ positive weights ρ_j, $0 \le j \le M$, such that

$$\forall \Phi \in \mathbb{P}_{2M-1}(-1,1), \quad \int_{-1}^{1} \Phi(\zeta)\, d\zeta = \sum_{j=0}^{M} \Phi(\xi_j)\, \rho_j. \tag{3.2}$$

Moreover the following property holds [4, Form. (13.20)]:

$$\forall \varphi \in \mathbb{P}_M(-1,1), \quad \|\varphi\|^2_{L^2(-1,1)} \le \sum_{j=0}^{M} \varphi^2(\xi_j)\, \rho_j \le 3\, \|\varphi\|^2_{L^2(-1,1)}. \tag{3.3}$$

The interest of using "overintegration," i.e., taking $M > N$, in the case of nonconstant coefficients has been fully brought to light in [23].

This leads to define a discrete product on ω: For any continuous functions u and v on $\bar{\omega}$,

$$(u,v)_M = \sum_{i=0}^{M} \sum_{j=0}^{M} u(\xi_i,\xi_j) v(\xi_i,\xi_j)\, \rho_i \rho_j. \tag{3.4}$$

It follows from (3.3) that this product is a scalar product on $\mathbb{P}_M(\omega)$. We also introduce the Lagrange interpolation operator $\mathcal{I}_M$ at the nodes (ξ_i,ξ_j), $0 \le i,j \le M$, with values in $\mathbb{P}_M(\omega)$. Finally, the discrete product $(\cdot,\cdot)^{\gamma_1}_M$ is defined according to the geometry of γ_1: For any continuous functions u and v on $\bar{\gamma}_1$, if γ_1 is the edge $\{-1\}\times]-1,1[$,

$$(u,v)^{\gamma_1}_M = \sum_{j=0}^{M} u(-1,\xi_j) v(-1,\xi_j)\, \rho_j, \tag{3.5}$$

while, if γ_1 is the union of the two edges $\{-1\}\times]-1,1]$ and $[-1,1[\times\{1\}$,

$$(u,v)^{\gamma_1}_M = \sum_{j=0}^{M} u(-1,\xi_j) v(-1,\xi_j)\, \rho_j + \sum_{j=0}^{M} u(\xi_j,1) v(\xi_j,1)\, \rho_j, \tag{3.6}$$

and so on. The Lagrange interpolation operator $i^{\gamma_1}_M$ is simply defined as the trace of $\mathcal{I}_M$ on γ_1.

From now on, we make the further nonrestrictive hypothesis.

Assumption 3.1. The $\mathbf{a}_\alpha$, $\alpha = 1$ and 2, belong to $H^{s_0}(\omega)^3$ and $\mathbf{a}_3$ belongs to $H^{s_0+1}(\omega)^3$ for a real number $s_0 > 1$.

In order to take into account this rather weak regularity, we introduce the $H^1(\omega)$-projection $\mathbf{a}_{kN}$ of each $\mathbf{a}_k$ onto $\mathbb{P}_N(\omega)^3$, which satisfies

$$\begin{aligned} &\forall \mathbf{v}_N \in \mathbb{P}_N(\omega)^3, \qquad (\partial_\alpha \mathbf{a}_{kN}, \partial_\alpha \mathbf{v}_N)_M = \int_\omega (\partial_\alpha \mathbf{a}_k)(\partial_\alpha \mathbf{v}_N)\, d\mathbf{x}, \\ &(\mathbf{a}_{kN}, 1)_M = \int_\omega \mathbf{a}_k(\mathbf{x})\, d\mathbf{x}. \end{aligned} \tag{3.7}$$

In a similar way, we define the $\mathbf{c}_{\beta N}$ as the solution of the same problem with $\mathbf{a}_k$ replaced by $\partial_\beta \mathbf{a}_3$. It follows from (3.3) that the $\mathbf{a}_{kN}$ and $\mathbf{c}_{\beta N}$ are uniquely defined from these equations. This leads to the following discrete forms of the tensors:

$$\gamma^N_{\alpha\beta}(\mathbf{u}) = \frac{1}{2}(\partial_\alpha \mathbf{u} \cdot \mathbf{a}_{\beta N} + \partial_\beta \mathbf{u} \cdot \mathbf{a}_{\alpha N}), \tag{3.8}$$

$$\delta^N_{\alpha 3}(U) = \frac{1}{2}(\partial_\alpha \mathbf{u}\cdot\mathbf{a}_{3N} + \mathbf{r}\cdot\mathbf{a}_{\alpha N}), \tag{3.9}$$

$$\chi^N_{\alpha\beta}(U) = \frac{1}{2}(\partial_\alpha \mathbf{u}\cdot\mathbf{c}_{\beta N} + \partial_\beta \mathbf{u}\cdot\mathbf{c}_{\alpha N} + \partial_\alpha \mathbf{r}\cdot\mathbf{a}_{\beta N} + \partial_\beta \mathbf{r}\cdot\mathbf{a}_{\alpha N}). \tag{3.10}$$

Up to the replacement of the $\mathbf{a}_k$ by $\mathbf{a}_{kN}$ and also of the $\partial_\beta \mathbf{a}_3$ by $\mathbf{c}_{\beta N}$, the discrete problem is now constructed from (2.17) by the Galerkin method with numerical integration. It reads:

Find (U_N, ψ_N) *in* $\mathbb{X}_N \times \mathbb{M}_N$ *such that*

$$\begin{array}{ll} \forall V_N \in \mathbb{X}_N, & a_M(U_N;V_N) + \eta\, \tilde{a}_M(U_N;V_N) + b_M(V_N;\psi_N) = \mathcal{L}_M(V_N), \\ \forall \chi_N \in \mathbb{M}_N, & b_M(U_N;\chi_N) = 0, \end{array} \tag{3.11}$$

where the bilinear forms $a_M(\cdot;\cdot)$, $\tilde{a}_M(\cdot;\cdot)$, and $b_M(\cdot;\cdot)$ are defined, with the notation $U_N = (\mathbf{u}_N, \mathbf{r}_N)$ and $V_N = (\mathbf{v}_N, \mathbf{s}_N)$, by

$$\begin{aligned} a_M(U_N;V_N) &= e\left(a^{\alpha\beta\rho\sigma}\,\gamma^N_{\alpha\beta}(\mathbf{u}_N), \gamma^N_{\rho\sigma}(\mathbf{v}_N)\sqrt{a}\right)_M \\ &\quad + \frac{e^3}{12}\left(a^{\alpha\beta\rho\sigma}\,\chi^N_{\alpha\beta}(U_N), \chi^N_{\rho\sigma}(V_N)\sqrt{a}\right)_M \\ &\quad + 2e\frac{E}{1+\nu}\left(a^{\alpha\beta}\delta^N_{\alpha 3}(U_N), \delta^N_{\beta 3}(V_N)\sqrt{a}\right)_M, \\ \tilde{a}_M(U_N;V_N) &= \left(\partial_\alpha \mathcal{I}_M(\mathbf{r}_N\cdot\mathbf{a}_{3N}), \partial_\alpha \mathcal{I}_M(\mathbf{s}_N\cdot\mathbf{a}_{3N})\right)_M, \\ b_M(V_N;\chi_N) &= \left(\partial_\alpha \mathcal{I}_M(\mathbf{s}_N\cdot\mathbf{a}_{3N}), \partial_\alpha \chi_N\right)_M. \end{aligned} \tag{3.12}$$

The linear form $\mathcal{L}_M(\cdot)$ is given by

$$\mathcal{L}_M(V_N) = (\mathbf{f}, \mathbf{v}_N\sqrt{a})_M + (\mathbf{M}, \mathbf{v}_N)^{\gamma_1}_M + (\mathbf{N}, \mathbf{s}_N)^{\gamma_1}_M. \tag{3.13}$$

Remark. The discrete problem (3.11) differs from the standard spectral discretization of elliptic problems in two ways.

- The idea for the replacement of the $\mathbf{a}_k$ by $\mathbf{a}_{kN}$ and of the $\partial_\beta \mathbf{a}_3$ by $\mathbf{c}_{\beta N}$ in the definition of the forms $a_M(\cdot,\cdot)$ and $b_M(\cdot,\cdot)$ comes from the lack of regularity of the $\mathbf{a}_k$. Indeed, if one of the $\mathbf{a}_\alpha$ is not replaced by $\mathbf{a}_{\alpha N}$, the continuity of these forms would require the boundedness of $\mathcal{I}_M\mathbf{a}_\alpha$ at least in $H^1(\omega)^3$, which would require that $\mathbf{a}_\alpha$ belongs to $H^s(\omega)^3$ for some $s > \frac{3}{2}$. We prefer to avoid this restriction.
- It is usual in spectral methods to take M equal to N, in order that the mass matrix is diagonal. The choice of an M possibly larger than N here is due to the fact that the coefficients involved in the previous forms depend on the space variable and are not very smooth (see [23] for more details). However, if ξ^*_j denote the nodes ξ_j for M equal to N and φ^*_j are the associated Lagrange polynomials, the unknown U_N admits the expansion

$$U_N(x,y) = \sum_{i=0}^{N}\sum_{j=0}^{N} U_N(\xi^*_i, \xi^*_j)\,\varphi^*_i(x)\varphi^*_j(y). \tag{3.14}$$

  So computing the values of U_N and $\partial_\alpha U_N$ at the nodes (ξ_j, ξ_j) only requires the knowledge of the two matrices made of the $\varphi^*_i(\xi_k)$ and of the $\varphi^{*\prime}_i(\xi_k)$, respectively.

The analysis of problem (3.11) relies on a number of properties of the previous forms. We begin with their continuity. In a preliminary step, we establish some results concerning the new coefficients $\mathbf{a}_{kN}$ and $\mathbf{c}_{\beta N}$.

LEMMA 3.2. *There exists a constant c independent of N such that, for any real number p, $2 \le p \le \frac{2}{2-s_0}$,*

$$\|\mathbf{a}_\alpha - \mathbf{a}_{\alpha N}\|_{L^\infty(\omega)^3} \le c\, N^{1-s_0}\, (\log\, N)^{\frac{1}{2}}\, \|\mathbf{a}_\alpha\|_{H^{s_0}(\omega)^3}, \tag{3.15}$$

$$\|\mathbf{a}_3 - \mathbf{a}_{3N}\|_{L^\infty(\omega)^3} \le c\, N^{-s_0}\, (\log\, N)^{\frac{1}{2}}\, \|\mathbf{a}_3\|_{H^{s_0+1}(\omega)^3}, \tag{3.16}$$

$$\|\mathbf{a}_3 - \mathbf{a}_{3N}\|_{W^{1,p}(\omega)^3} \le c\, N^{4(\frac{1}{2}-\frac{1}{p})-s_0}\, \|\mathbf{a}_3\|_{H^{s_0+1}(\omega)^3}, \tag{3.17}$$

$$\|\partial_\beta \mathbf{a}_3 - \mathbf{c}_{\beta N}\|_{L^\infty(\omega)^3} \le c\, N^{1-s_0}\, (\log\, N)^{\frac{1}{2}}\, \|\mathbf{a}_3\|_{H^{s_0+1}(\omega)^3}. \tag{3.18}$$

Proof. There exists [4, Thm. 7.4] a polynomial $\tilde{\mathbf{a}}_{kN}$ in $\mathbb{P}_N(\omega)^3$ such that, for all real numbers s, $0 \le s \le s_0$,

$$\|\mathbf{a}_\alpha - \tilde{\mathbf{a}}_{\alpha N}\|_{H^s(\omega)^3} \le c\, N^{s-s_0}\, \|\mathbf{a}_\alpha\|_{H^{s_0}(\omega)^3}. \tag{3.19}$$

To prove the first estimate, we use the triangle inequality

$$\|\mathbf{a}_\alpha - \mathbf{a}_{\alpha N}\|_{L^\infty(\omega)^3} \le \|\mathbf{a}_\alpha - \tilde{\mathbf{a}}_{\alpha N}\|_{L^\infty(\omega)^3} + \|\mathbf{a}_{\alpha N} - \tilde{\mathbf{a}}_{\alpha N}\|_{L^\infty(\omega)^3}.$$

To bound the first quantity, we recall from [17] that, for all ε, $0 < \varepsilon < s_0 - 1$, the norm of the Sobolev embedding of $H^{1+\varepsilon}(\omega)$ into $L^\infty(\omega)$ is bounded by $c\,\varepsilon^{-\frac{1}{2}}$ for a constant c independent of ε. Combining this result with (3.20) with $s = 1 + \varepsilon$ gives

$$\|\mathbf{a}_\alpha - \tilde{\mathbf{a}}_{\alpha N}\|_{L^\infty(\omega)^3} \le c\,\varepsilon^{-\frac{1}{2}}\, \|\mathbf{a}_\alpha - \tilde{\mathbf{a}}_{\alpha N}\|_{H^{1+\varepsilon}(\omega)^3} \le c'\,\varepsilon^{-\frac{1}{2}}\, N^{\varepsilon+1-s_0}\, \|\mathbf{a}_\alpha\|_{H^{s_0}(\omega)^3}.$$

Evaluating the second one relies on the inverse inequality (see [3, Chap. III, Prop. 3.1]), valid for any $p < +\infty$,

$$\|\mathbf{a}_{\alpha N} - \tilde{\mathbf{a}}_{\alpha N}\|_{L^\infty(\omega)^3} \le c\, N^{\frac{4}{p}}\, \|\mathbf{a}_{\alpha N} - \tilde{\mathbf{a}}_{\alpha N}\|_{L^p(\omega)^3}.$$

We recall from [25] that the norm of the embedding of $H^1(\omega)$ into $L^p(\omega)$ behaves like $c\,p^{\frac{1}{2}}$ for a constant c independent of p. This yields

$$\begin{aligned}\|\mathbf{a}_{\alpha N} - \tilde{\mathbf{a}}_{\alpha N}\|_{L^\infty(\omega)^3} &\le c\,p^{\frac{1}{2}}\, N^{\frac{4}{p}}\, \|\mathbf{a}_{\alpha N} - \tilde{\mathbf{a}}_{\alpha N}\|_{H^1(\omega)^3} \\ &\le c\,p^{\frac{1}{2}}\, N^{\frac{4}{p}}\, \big(\|\mathbf{a}_\alpha - \mathbf{a}_{\alpha N}\|_{H^1(\omega)^3} + \|\mathbf{a}_\alpha - \tilde{\mathbf{a}}_{\alpha N}\|_{H^1(\omega)^3}\big).\end{aligned}$$

It follows from the definition of $\mathbf{a}_{\alpha N}$ that

$$\|\mathbf{a}_\alpha - \mathbf{a}_{\alpha N}\|_{H^1(\omega)^3} \le c\, N^{1-s_0}\, \|\mathbf{a}_\alpha\|_{H^{s_0}(\omega)^3}. \tag{3.20}$$

Combining this with (3.19) for $s = 1$ thus leads to

$$\|\mathbf{a}_{\alpha N} - \tilde{\mathbf{a}}_{\alpha N}\|_{L^\infty(\omega)^3} \le c\,p^{\frac{1}{2}}\, N^{\frac{4}{p}+1-s_0}\ \|\mathbf{a}_\alpha\|_{H^{s_0}(\omega)^3}.$$

Thus estimate (3.15) follows from the previous lines by taking $\varepsilon = \frac{4}{p} = \frac{1}{\log\, N}$. Estimates (3.16) and (3.18) rely on exactly the same arguments with s_0 replaced or not by $s_0 + 1$. To prove (3.17), we first take s such that $\frac{1}{p} = \frac{2-s}{2}$, so that $H^s(\omega)$ is embedded in $W^{1,p}(\omega)$ and use an analogous inverse inequality as previously (see again [3, Chap. III, Prop. 3.1]), which gives

$$\|\mathbf{a}_3 - \mathbf{a}_{3N}\|_{W^{1,p}(\omega)^3} \le c\,\|\mathbf{a}_3 - \tilde{\mathbf{a}}_{3N}\|_{H^s(\omega)^3} + c'\, N^{4(\frac{1}{2}-\frac{1}{p})}\, \|\mathbf{a}_{3N} - \tilde{\mathbf{a}}_{3N}\|_{H^1(\omega)^3}.$$

Thus, we deduce from the analogues of (3.19) and (3.20), with s_0 replaced by s_0+1, that

$$\|\mathbf{a}_3-\mathbf{a}_{3N}\|_{W^{1,p}(\omega)^3}\le c\,(N^{s-s_0-1}+N^{4(\frac12-\frac1p)-s_0})\|\mathbf{a}_3\|_{H^{s_0+1}(\omega)^3}.$$

Noting that $s-s_0-1$ is equal to $2(\frac12-\frac1p)-s_0$, hence smaller than $4(\frac12-\frac1p)-s_0$, gives the desired result. ∎

As a consequence of the previous lemma, the norms of the coefficients $\mathbf{a}_{kN}$ and $\mathbf{c}_{\beta N}$ in $L^\infty(\omega)^3$ and also of $\mathbf{a}_{3N}$ in $W^{1,p}(\omega)^3$ are bounded independently of N. This leads to the following continuity results.

LEMMA 3.3. *There exists a constant c independent of N and $M\ge N$ such that the following continuity property holds:*

$$\forall U_N\in\mathbb{X}_N,\forall V_N\in\mathbb{X}_N,\quad |a_M(U_N;V_N)|\le c\,e\,\|U_N\|_{\mathbb{X}(\omega)}\|V_N\|_{\mathbb{X}(\omega)}. \tag{3.21}$$

Proof. Since the coefficients $a^{\alpha\beta\rho\sigma}$ and also $\sqrt{a}$ are bounded, we derive by a Cauchy–Schwarz inequality

$$e\,\big|\big(a^{\alpha\beta\rho\sigma}\,\gamma^N_{\alpha\beta}(\mathbf{u}_N),\gamma^N_{\rho\sigma}(\mathbf{v}_N)\,\sqrt{a}\big)_M\big|\le c\,e\,(\gamma^N_{\alpha\beta}(\mathbf{u}_N),\gamma^N_{\alpha\beta}(\mathbf{u}_N))^{\frac12}_M(\gamma^N_{\rho\sigma}(\mathbf{v}_N),\gamma^N_{\rho\sigma}(\mathbf{v}_N))^{\frac12}_M.$$

Thanks to Lemma 3.2, we observe that, at each node (ξ_i,ξ_j), $0\le i,j\le M$,

$$\gamma^N_{\alpha\beta}(\mathbf{u}_N)(\xi_i,\xi_j)\le c\,\big(|(\partial_\alpha\mathbf{u}_N)(\xi_i,\xi_j)|+|(\partial_\beta\mathbf{u}_N)(\xi_i,\xi_j)|\big),$$

so that

$$(\gamma^N_{\alpha\beta}(\mathbf{u}_N),\gamma^N_{\alpha\beta}(\mathbf{u}_N))_M\le c'\,\big((\partial_\alpha\mathbf{u}_N,\partial_\alpha\mathbf{u}_N)_M+(\partial_\beta\mathbf{u}_N,\partial_\beta\mathbf{u}_N)_M\big).$$

Using (3.3) and a similar estimate for $(\gamma^N_{\rho\sigma}(\mathbf{v}_N),\gamma^N_{\rho\sigma}(\mathbf{v}_N))_M$ leads to

$$e\,\big|\big(a^{\alpha\beta\rho\sigma}\,\gamma^N_{\alpha\beta}(\mathbf{u}_N),\gamma^N_{\rho\sigma}(\mathbf{v}_N)\,\sqrt{a}\big)_M\big|\le c\,e\,\|\mathbf{u}_N\|_{H^1(\omega)^3}\|\mathbf{v}_N\|_{H^1(\omega)^3}.$$

Similar arguments also yield the desired estimates for the second and third terms in $a_M(U_N,V_N)$. ∎

LEMMA 3.4. *There exists a constant c independent of N and $M\ge N$ such that the following continuity property holds:*

$$\forall U_N\in\mathbb{X}_N,\forall V_N\in\mathbb{X}_N,\quad |\tilde a_M(U_N;V_N)|\le c\,\|U_N\|_{\mathbb{X}(\omega)}\|V_N\|_{\mathbb{X}(\omega)}. \tag{3.22}$$

Proof. We derive from (3.3) that

$$|\tilde a_M(U_N;V_N)|\le 3\,|\mathcal{I}_M(\mathbf{r}_N\cdot\mathbf{a}_{3N})|_{H^1(\omega)}|\mathcal{I}_M(\mathbf{s}_N\cdot\mathbf{a}_{3N})|_{H^1(\omega)}.$$

We recall from [4, Forms. (13.27) and (13.28)] the following property, valid for all integers K:

$$\forall w_K\in\mathbb{P}_K(\omega),\quad |\mathcal{I}_M w_K|_{H^1(\omega)}\le c\,\left(1+\frac{K}{M}\right)|w_K|_{H^1(\omega)}. \tag{3.23}$$

Since both $\mathbf{r}_N\cdot\mathbf{a}_{3N}$ and $\mathbf{s}_N\cdot\mathbf{a}_{3N}$ are polynomials with degree smaller than $2N$, this implies

$$|\tilde a_M(U_N;V_N)|\le c\,|\mathbf{r}_N\cdot\mathbf{a}_{3N}|_{H^1(\omega)}|\mathbf{s}_N\cdot\mathbf{a}_{3N}|_{H^1(\omega)}.$$

Then we observe that

$$\partial_\alpha(\mathbf{r}_N \cdot \mathbf{a}_{3N}) = \partial_\alpha \mathbf{r}_N \cdot \mathbf{a}_{3N} + \mathbf{r}_N \cdot \partial_\alpha \mathbf{a}_{3N}.$$

This yields, for the p such that $4(\frac{1}{2} - \frac{1}{p}) < s_0$ and q given by $\frac{1}{p} + \frac{1}{q} = \frac{1}{2}$,

$$|\mathbf{r}_N \cdot \mathbf{a}_{3N}|_{H^1(\omega)} \le |\mathbf{r}_N|_{H^1(\omega)^3} \|\mathbf{a}_{3N}\|_{L^\infty(\omega)^3} + \|\mathbf{r}_N\|_{L^q(\omega)^3} \|\mathbf{a}_{3N}\|_{W^{1,p}(\omega)^3}.$$

By combining Lemma 3.2 with the embedding of $H^1(\omega)$ into $L^q(\omega)$, we obtain

$$|\mathbf{r}_N \cdot \mathbf{a}_{3N}|_{H^1(\omega)} \le c \, \|\mathbf{r}_N\|_{H^1(\omega)^3} \|\mathbf{a}_3\|_{H^{s_0+1}(\omega)^3}.$$

A similar estimate holds for $|\mathbf{s}_N \cdot \mathbf{a}_{3N}|_{H^1(\omega)}$, which yields the desired continuity property. ∎

The continuity of $b_M(\cdot;\cdot)$ relies on the same arguments, so we omit the proof of the next lemma.

LEMMA 3.5. *There exists a constant c independent of N and $M \ge N$ such that the following continuity property holds:*

$$\forall V_N \in \mathbb{X}_N, \forall \chi_N \in \mathbb{M}_N, \quad |b_M(V_N;\chi_N)| \le c \, \|V_N\|_{\mathbb{X}(\omega)} \|\chi_N\|_{H^1(\omega)}. \tag{3.24}$$

Finally we derive from (3.3) that, if $\|\mathcal{L}_M\|_N$ denotes the norm of $\mathcal{L}_M$ in the space of linear forms on $\mathbb{X}_N$,

$$\|\mathcal{L}_M\|_N \le c \left(\|\mathcal{I}_M \mathbf{f}\|_{L^2(\omega)^3} + \|i_M^{\gamma_1} \mathbf{M}\|_{L^2(\gamma_1)^3} + \|i_M^{\gamma_1} \mathbf{N}\|_{L^2(\gamma_1)^3} \right). \tag{3.25}$$

To go further, we introduce the kernel

$$\mathbb{V}_N = \left\{ V_N = (\mathbf{v}_N, \mathbf{s}_N) \in \mathbb{X}_N; \ \forall \chi_N \in \mathbb{M}_N, \ b_M(V_N;\chi_N) = 0 \right\}. \tag{3.26}$$

It is readily checked that $\mathbb{V}_N$ is not contained in $\mathbb{V}(\omega)$ in the general case. So the proof of the next ellipticity property relies on the following result, due to [9, Lem. 3.3]: For a constant $c_\flat > 0$,

$$\forall V \in \mathbb{X}(\omega), \quad [V]^2 + \tilde{a}(V;V) \ge c_\flat \, \|V\|^2_{\mathbb{X}(\omega)}, \tag{3.27}$$

where the seminorm $[\cdot]$ is defined by

$$[V] = \left(\sum_{\alpha=1}^{2} \sum_{\beta=1}^{2} \|\gamma_{\alpha\beta}(\mathbf{v})\|^2_{L^2(\omega)} + \sum_{\alpha=1}^{2} \sum_{\beta=1}^{2} \|\chi_{\alpha\beta}(\mathbf{v},\mathbf{s})\|^2_{L^2(\omega)} + \sum_{\alpha=1}^{2} \|\delta_{\alpha 3}(\mathbf{v},\mathbf{s})\|^2_{L^2(\omega)} \right)^{\frac{1}{2}}.$$

Let us also consider its discrete analogue on $\mathbb{X}_N$:

$$\begin{aligned} [V_N]_M = \Big(&\left(\gamma^N_{\alpha\beta}(\mathbf{v}_N), \gamma^N_{\alpha\beta}(\mathbf{v}_N)\right)_M + \left(\chi^N_{\alpha\beta}(\mathbf{v}_N,\mathbf{s}_N), \chi^N_{\alpha\beta}(\mathbf{v}_N,\mathbf{s}_N)\right)_M \\ &+ \left(\delta^N_{\alpha 3}(\mathbf{v}_N,\mathbf{s}_N), \delta^N_{\alpha 3}(\mathbf{v}_N,\mathbf{s}_N)\right)_M \Big)^{\frac{1}{2}}. \end{aligned} \tag{3.28}$$

From now on, denoting by $\lfloor \cdot \rfloor$ the integer part, we choose L and M such that, for fixed real numbers λ and μ, $0 < \lambda < 1$ and $0 < \mu \le 1$,

$$L = \lfloor (1-\lambda) N \rfloor \qquad \text{and} \qquad M = \lfloor (1+\mu) N \rfloor. \tag{3.29}$$

PROPOSITION 3.6. *There exist a positive integer N_* and a positive constant $\tilde{c}_*$ such that, for all $N \geq N_*$, the following ellipticity property holds:*

$$\forall V_N \in \mathbb{X}_N, \quad a_M(V_N;V_N) + \eta\, \tilde{a}_M(V_N;V_N) \geq \tilde{c}_*\, \min\{e^3,\eta\}\, \|V_N\|^2_{\mathbb{X}(\omega)}. \tag{3.30}$$

Proof. It is readily checked from (2.3) that, for all V_N in $\mathbb{X}_N$,

$$a_M(V_N;V_N) + \eta\, \tilde{a}_M(V_N;V_N) \geq \tilde{c}_*\, \min\{e^3,\eta\} \Big([V_N]^2_M + \tilde{a}_M(V_N;V_N) \Big). \tag{3.31}$$

On the other hand, since $\mathbb{X}_N$ is contained in $\mathbb{X}(\omega)$, it follows from (3.27) that

$$[V_N]^2 + \tilde{a}(V_N;V_N) \geq c_\flat\, \|V_N\|^2_{\mathbb{X}(\omega)}. \tag{3.32}$$

So it remains to compare $[V_N]$ and $[V_N]_M$ and also $\tilde{a}(V_N;V_N)$ and $\tilde{a}_M(V_N;V_N)$. Let K denote the integer part of $\mu\, N - 1$ (we assume N large enough for K to be positive).

(1) Let $\mathbf{a}_{\alpha K}$ denote an approximation of $\mathbf{a}_\alpha$ in $\mathbb{P}_K(\omega)^3$ which still satisfies (3.15), and let $\gamma^K_{\alpha\beta}(\mathbf{v}_N)$ be defined as in (3.8) with $\mathbf{a}_{\alpha N}$ replaced by $\mathbf{a}_{\alpha K}$. It follows from the exactness property (3.2) that

$$\big(\gamma^K_{\alpha\beta}(\mathbf{v}_N), \gamma^K_{\alpha\beta}(\mathbf{v}_N)\big)_M = \|\gamma^K_{\alpha\beta}(\mathbf{v}_N)\|^2_{L^2(\omega)^{2\times 2}}.$$

Then we use the inequalities

$$\big(\gamma^N_{\alpha\beta}(\mathbf{v}_N), \gamma^N_{\alpha\beta}(\mathbf{v}_N)\big)_M \geq \big(\gamma^K_{\alpha\beta}(\mathbf{v}_N), \gamma^K_{\alpha\beta}(\mathbf{v}_N)\big)_M + 2\big((\gamma^N_{\alpha\beta} - \gamma^K_{\alpha\beta})(\mathbf{v}_N), \gamma^K_{\alpha\beta}(\mathbf{v}_N)\big)_M,$$

and

$$\|\gamma^K_{\alpha\beta}(\mathbf{v}_N)\|^2_{L^2(\omega)^{2\times 2}} \geq \|\gamma_{\alpha\beta}(\mathbf{v}_N)\|^2_{L^2(\omega)^{2\times 2}} - 2\int_\Omega \big(\gamma_{\alpha\beta}(\mathbf{v}_N) - \gamma^K_{\alpha\beta}(\mathbf{v}_N)\big)\gamma_{\alpha\beta}(\mathbf{v}_N)\, d\mathbf{x}.$$

On the other hand, the same arguments as for Lemma 3.3 yield that

$$\big|\big((\gamma^N_{\alpha\beta} - \gamma^K_{\alpha\beta})(\mathbf{v}_N), \gamma^K_{\alpha\beta}(\mathbf{v}_N)\big)_M\big| \leq c \sum_{\alpha=1}^2 \|\mathbf{a}_{\alpha N} - \mathbf{a}_{\alpha K}\|_{L^\infty(\omega)^3}\, \|\mathbf{v}_N\|^2_{H^1(\omega)^3},$$

while it is readily checked that

$$\left|\int_\Omega \big(\gamma_{\alpha\beta}(\mathbf{v}_N) - \gamma^K_{\alpha\beta}(\mathbf{v}_N)\big)\gamma_{\alpha\beta}(\mathbf{v}_N)\, d\mathbf{x}\right| \leq c \sum_{\alpha=1}^2 \|\mathbf{a}_\alpha - \mathbf{a}_{\alpha K}\|_{L^\infty(\omega)^3}\, \|\mathbf{v}_N\|^2_{H^1(\omega)^3}.$$

All of this yields

$$\begin{aligned}\big(\gamma^N_{\alpha\beta}(\mathbf{v}_N), \gamma^N_{\alpha\beta}(\mathbf{v}_N)\big)_M \geq\ & \|\gamma_{\alpha\beta}(\mathbf{v}_N)\|^2_{L^2(\omega)^{2\times 2}} \\ & - c\sum_{\alpha=1}^2 \big(\|\mathbf{a}_\alpha - \mathbf{a}_{\alpha N}\|_{L^\infty(\omega)^3} + \|\mathbf{a}_\alpha - \mathbf{a}_{\alpha K}\|_{L^\infty(\omega)^3}\big)\|\mathbf{v}_N\|^2_{H^1(\omega)^3}.\end{aligned}$$

Using the same arguments for estimating the two other terms together with Lemma 3.2 leads to

$$[V_N]^2_M \geq [V_N]^2 - c\, N^{1-s_0}\, (\log\, N)^{\frac{1}{2}}\, \|V_N\|^2_{\mathbb{X}(\omega)}. \tag{3.33}$$

(2) Similarly, let $\mathbf{a}_{3K}$ denote an approximation of $\mathbf{a}_3$ in $\mathbb{P}_K(\omega)^3$ which still satisfies (3.17). Since $\mathbf{s}_N \cdot \mathbf{a}_{3K}$ now belongs to $\mathbb{P}_M(\omega)$, it is equal to $\mathcal{I}_M(\mathbf{s}_N \cdot \mathbf{a}_{3K})$ and, moreover,

$$\big(\partial_\alpha \mathcal{I}_M(\mathbf{s}_N \cdot \mathbf{a}_{3K}), \partial_\alpha \mathcal{I}_M(\mathbf{s}_N \cdot \mathbf{a}_{3K})\big)_M = \int_\omega \partial_\alpha(\mathbf{s}_N \cdot \mathbf{a}_{3K}) \partial_\alpha(\mathbf{s}_N \cdot \mathbf{a}_{3K})\, d\mathbf{x}.$$

Thus, the same arguments as for Lemma 3.4 (see (3.23)) yield that

$$\tilde{a}_M(V_N;V_N) \geq \tilde{a}(V_N,V_N) - c\sum_{\alpha=1}^{2}\big(\|\partial_\alpha(\mathbf{s}_N\cdot(\mathbf{a}_3-\mathbf{a}_{3N}))\|_{L^2(\omega)^3}+\|\partial_\alpha(\mathbf{s}_N\cdot(\mathbf{a}_3-\mathbf{a}_{3K}))\|_{L^2(\omega)^3}\big)\,\|\mathbf{s}_N\|_{H^1(\omega)^3}.$$

Using once more Lemma 3.2 (with p such that $4(\frac{1}{2}-\frac{1}{p})\leq 1$) leads to

$$\tilde{a}_M(V_N;V_N) \geq \tilde{a}(V_N,V_N) - c\,N^{1-s_0}\,(\log\,N)^{\frac{1}{2}}\,\|\mathbf{s}_N\|^2_{H^1(\omega)^3}. \tag{3.34}$$

Combining (3.32) to (3.34) gives

$$[V_N]^2_M + \tilde{a}_M(V_N;V_N) \geq \big(c_\flat - c\,N^{1-s_0}\,(\log\,N)^{\frac{1}{2}}\big)\,\|V_N\|^2_{\mathbb{X}(\omega)}.$$

We choose N_* such that $c\,N_*^{1-s_0}\,(\log\,N_*)^{\frac{1}{2}} \leq \frac{c_\flat}{2}$. Thus, the desired property follows from (3.31). ∎

Proving the inf-sup condition on $b_M(\cdot,\cdot)$ is performed in two steps. Note that it requires the choice of L made in (3.30) (we refer to [5, sect. 3] for proving an inf-sup condition with such a choice in a completely different context).

LEMMA 3.7. *There exist a positive integer $N_\sharp$ and a positive constant $\tilde{c}_\sharp$ such that, for all $N \geq N_\sharp$, the following inf-sup condition holds:*

$$\forall \chi_N \in \mathbb{M}_N, \quad \sup_{V\in\mathbb{X}_N}\frac{b(V_N;\chi_N)}{\|V_N\|_{\mathbb{X}(\omega)}} \geq \tilde{c}_\sharp\,\|\chi_N\|_{H^1(\omega)}. \tag{3.35}$$

Proof. Let χ_N be any element of $\mathbb{M}_N$. Let now K' stand for the integer part of $\lambda N - 1$. We introduce an approximation $\mathbf{a}_{3K'}$ of $\mathbf{a}_3$ in $\mathbb{P}_{K'}(\omega)^3$ which satisfies (3.16) and (3.17). It follows from the definition (3.1) of $\mathbb{M}_N$ and the choice (3.29) of L that, for any χ_N in $\mathbb{M}_N$, the function $\mathbf{s}_N = \chi_N\,\mathbf{a}_{3K'}$ belongs to $\mathbb{P}_N(\omega)^3$. Since χ_N vanishes on γ_0, the function $V_N = (\mathbf{0},\mathbf{s}_N)$ belongs to $\mathbb{X}_N$ and satisfies

$$b(V_N;\chi_N) = \int_\omega \partial_\alpha(\chi_N\,\mathbf{a}_{3K'}\cdot\mathbf{a}_3)\partial_\alpha\chi_N\,d\mathbf{x}.$$

Since $\mathbf{a}_3\cdot\mathbf{a}_3$ is equal to 1, this gives

$$b(V_N;\chi_N) = |\chi_N|^2_{H^1(\omega)} - \int_\omega \partial_\alpha(\chi_N\,(\mathbf{a}_3-\mathbf{a}_{3K'})\cdot\mathbf{a}_3)\partial_\alpha\chi_N\,d\mathbf{x}.$$

Combining the Poincaré–Friedrichs inequality with the continuity of $b(\cdot;\cdot)$ yields

$$b(V_N;\chi_N) \geq c\,\|\chi_N\|^2_{H^1(\omega)} - |\chi_N\,(\mathbf{a}_3-\mathbf{a}_{3K'})\cdot\mathbf{a}_3)|_{H^1(\omega)}|\chi_N|_{H^1(\omega)}.$$

We have, for an appropriate value of p and with $\frac{1}{p}+\frac{1}{q}=\frac{1}{2}$,

$$\begin{aligned}\|\partial_\alpha(\chi_N\,(\mathbf{a}_3-\mathbf{a}_{3K'})\cdot\mathbf{a}_3)\|_{L^2(\omega)} \leq\ & |\chi_N|_{H^1(\omega)}\|\mathbf{a}_3\|_{L^\infty(\omega)^3}\|\mathbf{a}_3-\mathbf{a}_{3K'}\|_{L^\infty(\omega)^3}\\ &+\|\chi_N\|_{L^2(\omega)}\|\mathbf{a}_3\|_{W^{1,\infty}(\omega)^3}\|\mathbf{a}_3-\mathbf{a}_{3K'}\|_{L^\infty(\omega)^3}\\ &+\|\chi_N\|_{L^q(\Omega)}\|\mathbf{a}_3\|_{L^\infty(\omega)^3}\|\mathbf{a}_3-\mathbf{a}_{3K'}\|_{W^{1,p}(\omega)^3}.\end{aligned}$$

Using the fact [25] that the norm of the embedding of $H^1(\omega)$ into $L^q(\omega)$ behaves like $c\,q^{\frac{1}{2}}$, combined with (3.16) and (3.17), and taking q equal to $\log N$ lead to

$$b(V_N;\chi_N) \geq \big(c - c'\,N^{-s_0}\,(\log\,N)^{\frac{1}{2}}\big)\,\|\chi_N\|^2_{H^1(\omega)},$$

whence, for N large enough,

$$b(V_N;\chi_N) \geq \frac{c}{2}\,\|\chi_N\|^2_{H^1(\omega)}.$$

On the other hand, the usual arguments combined with (3.16) and (3.17) give $\|\mathbf{s}_N\|_{H^1(\omega)^3} \leq c\,\|\chi_N\|_{H^1(\omega)}$, which yields the desired inf-sup condition. ∎

Thanks to Lemma 3.7, the proof of the next proposition relies on the same arguments as for Proposition 3.6.

PROPOSITION 3.8. *There exist a positive integer $N_{\sharp\sharp}$ and a positive constant $\tilde{c}_{\sharp\sharp}$ such that, for all $N \geq N_{\sharp\sharp}$, the following inf-sup condition holds:*

$$\forall \chi_N \in \mathbb{M}_N, \quad \sup_{V\in\mathbb{X}_N} \frac{b_M(V_N;\chi_N)}{\|V_N\|_{\mathbb{X}(\omega)}} \geq \tilde{c}_{\sharp\sharp}\,\|\chi_N\|_{H^1(\omega)}. \tag{3.36}$$

The well-posedness result is now a direct consequence of Propositions 3.6 and 3.8. The stability estimate also requires (3.25).

THEOREM 3.9. *There exists a positive integer N_0 such that, for any data $(\mathbf{f},\mathbf{M},\mathbf{N})$ in $\mathscr{C}^0(\overline{\omega})^3 \times \mathscr{C}^0(\overline{\gamma}_1)^3 \times \mathscr{C}^0(\overline{\gamma}_1)^3$ and for $N \geq N_0$, problem (3.11) admits a unique solution (U_N,ψ_N) in $\mathbb{X}_N \times \mathbb{M}_N$. Moreover this solution satisfies*

$$\|U_N\|_{\mathbb{X}(\omega)} + \|\psi_N\|_{H^1(\omega)} \leq c\,\max\{e^{-3},\eta^{-1}\}\,\|\mathcal{L}_M\|_N. \tag{3.37}$$

4. Error estimates. The error estimate that we now prove is derived from Proposition 3.6 and requires the integer N_* introduced there. Indeed we are not interested in the evaluation of the error concerning the Lagrange multiplier ψ. We first prove the following version of the second Strang's lemma.

PROPOSITION 4.1. *For any integer $N \geq N_*$, the following error estimate holds between the solution (U^η,ψ^η) of problem (2.17) and the solution (U_N,ψ_N) of problem (3.11):*

$$\begin{aligned}\|U^\eta - U_N\|_{\mathbb{X}(\omega)} \leq c\,\max\{e^{-3},\eta^{-1}\}\\ \Big(\inf_{W_N\in\mathbb{V}_N}\Big(\max\{e,\eta\}\,\|U^\eta - W_N\|_{\mathbb{X}(\omega)} + \sup_{V_N\in\mathbb{X}_N}\frac{E^a_M(W_N;V_N)+\eta\,\tilde{E}^a_M(W_N;V_N)}{\|V_N\|_{\mathbb{X}(\omega)}}\Big)\\ + \inf_{\chi_N\in\mathbb{M}_N}\Big(\|\psi^\eta-\chi_N\|_{H^1(\omega)} + \sup_{V_N\in\mathbb{X}_N}\frac{E^b_M(V_N;\chi_N)}{\|V_N\|_{\mathbb{X}(\omega)}}\Big)\\ + \sup_{V_N\in\mathbb{X}_N}\frac{E^{\mathcal{L}}_M(V_N)}{\|V_N\|_{\mathbb{X}(\omega)}}\Big),\end{aligned} \tag{4.1}$$

where the four quantities E^a_M, $\tilde{E}^a_M$, E^b_M, and $E^{\mathcal{L}}_M$ are defined by

$$\begin{aligned} E^a_M(W_N;V_N) &= (a-a_M)(W_N;V_N), \quad & \tilde{E}^a_M(W_N;V_N) &= (\tilde{a}-\tilde{a}_M)(W_N;V_N),\\ E^b_M(V_N;\chi_N) &= (b-b_M)(V_N;\chi_N), & E^{\mathcal{L}}_M(V_N) &= (\mathcal{L}-\mathcal{L}_M)(V_N).\end{aligned} \tag{4.2}$$

Proof. Let V_N and W_N be any functions in $\mathbb{V}_N$. We derive from problem (3.11) that

$$a_M(U_N-W_N;V_N)+\eta\,\tilde{a}_M(U_N-W_N;V_N) = \mathcal{L}_M(V_N)-a_M(W_N;V_N)-\eta\,\tilde{a}_M(W_N;V_N).$$

Then, using problem (2.17) yields for any χ_N in $\mathbb{M}_N$ (note that $b_M(V_N;\chi_N)$ is equal to zero)

$$\begin{aligned} a_M(U_N-W_N;V_N)+\eta\,\tilde{a}_M(U_N-W_N;V_N)\\ = -E^{\mathcal{L}}_M(V_N) + a(U^\eta-W_N;V_N) + E^a_M(W_N;V_N)\\ + \eta\,\tilde{a}(U^\eta-W_N;V_N) + \eta\,\tilde{E}^a_M(W_N;V_N) + b(V_N;\psi^\eta-\chi_N) + E^b_M(V_N;\chi_N).\end{aligned}$$

Next, we take V_N equal to $U_N - W_N$ (which belongs to $\mathbb{V}_N$) and use the ellipticity property (3.30). Since the norm of $a(\cdot;\cdot)$ is smaller than $c\,e$, this gives the desired estimate for the term $\|U_N - W_N\|_{\mathbb{X}(\omega)}$. We conclude thanks to a triangle inequality by noting that $e\,\max\{e^{-3}, \eta^{-1}\}$ is larger than 1. ∎

The two terms $\inf_{W_N\in\mathbb{V}_N} \|U^\eta - W_N\|_{\mathbb{X}(\omega)}$ and $\inf_{\chi_N\in\mathbb{M}_N} \|\psi^\eta - \chi_N\|_{H^1(\omega)}$ represent the approximation errors, while the four other terms are issued from numerical integration and the replacement of the coefficients of the initial problem by discrete ones. We begin with the first approximation error.

LEMMA 4.2. *For any integer $N \geq N_{\sharp\sharp}$, there exists a constant c independent of N such that, for all U in $\mathbb{V}(\omega)$,*

$$\inf_{W_N\in\mathbb{V}_N} \|U - W_N\|_{\mathbb{X}(\omega)} \leq c \inf_{Z_N\in\mathbb{X}_N} \left(\|U - Z_N\|_{\mathbb{X}(\omega)} + \sup_{\omega_N\in\mathbb{M}_N} \frac{E_M^b(Z_N;\omega_N)}{\|\omega_N\|_{H^1(\omega)}} \right). \tag{4.3}$$

Proof. Let Z_N be any element of $\mathbb{X}_N$. It follows from the inf-sup condition (3.36) (see [16, Chap. I, Lem. 4.1]) that there exists a $\tilde{Z}_N$ in $\mathbb{X}_N$ such that

$$\begin{aligned} &\forall \omega_N \in \mathbb{M}_N, \quad b_M(\tilde{Z}_N;\omega_N) = b_M(Z_N;\omega_N) \\ &\text{and} \quad \|\tilde{Z}_N\|_{\mathbb{X}(\omega)} \leq \tilde{c}_{\sharp\sharp}^{-1} \sup_{\omega_N\in\mathbb{M}_N} \frac{b_M(Z_N;\omega_N)}{\|\omega_N\|_{H^1(\omega)}}. \end{aligned} \tag{4.4}$$

Then the function $W_N = Z_N - \tilde{Z}_N$ belongs to $\mathbb{V}_N$. Moreover, we obtain the desired estimate by using the triangle inequality

$$\|U - W_N\|_{\mathbb{X}(\omega)} \leq \|U - Z_N\|_{\mathbb{X}(\omega)} + \|\tilde{Z}_N\|_{\mathbb{X}(\omega)},$$

combined with (4.4), the identity

$$b_M(Z_N;\omega_N) = -b(U - Z_N;\omega_N) - E_M^b(Z_N;\omega_N),$$

and the continuity of $b(\cdot;\cdot)$.

Since γ_0 is the union of whole edges of ω, the following estimates are standard; see [4, sect. 7], for instance: If the solution (U^η, ψ^η) belongs to $H^S(\omega)^{3\times3} \times H^S(\omega)$ for a real number $S \geq 1$,

$$\begin{aligned} \inf_{Z_N\in\mathbb{X}_N} \|U^\eta - Z_N\|_{\mathbb{X}(\omega)} &\leq c\,N^{1-S}\,\|U^\eta\|_{H^S(\omega)^{3\times3}}, \\ \inf_{\chi_N\in\mathbb{M}_N} \|\psi^\eta - \chi_N\|_{H^1(\omega)} &\leq c\,N^{1-S}\,\|\psi^\eta\|_{H^S(\omega)}. \end{aligned} \tag{4.5}$$

So it remains to investigate the four terms defined in (4.2). This involves the real number s_0 introduced in Assumption 3.1. ∎

LEMMA 4.3. *There exists a constant c only depending on the norms of the $\mathbf{a}_\alpha$ in $H^{s_0}(\omega)^3$ and of $\mathbf{a}_3$ in $H^{s_0+1}(\omega)^3$ such that*

$$\forall W_N \in \mathbb{X}_N, \quad \sup_{V_N\in\mathbb{X}_N} \frac{E_M^a(W_N;V_N)}{\|V_N\|_{\mathbb{X}(\omega)}} \leq c\,e\,N^{1-s_0}\,(\log\,N)^{\frac{1}{2}}\,\|W_N\|_{\mathbb{X}(\omega)}. \tag{4.6}$$

Proof. Let K'' now denote the integral part of $\frac{\mu\,N-1}{2}$. It follows from the definition of the $a^{\alpha\beta\rho\sigma}$ and also of a that all of these coefficients and also $\sqrt{a}$ belong to $H^{s_0}(\omega)$. We denote by $\mathbf{a}_{\alpha K''}$, $a_{K''}^{\alpha\beta\rho\sigma}$ and $(\sqrt{a})_{K''}$ some approximations of the $\mathbf{a}_\alpha$, $a^{\alpha\beta\rho\sigma}$, and $\sqrt{a}$ in $\mathbb{P}_{K''}(\omega)^3$ or in $\mathbb{P}_{K''}(\omega)$ which still satisfy (3.15). We then derive from the exactness

property (3.2) and with obvious notation

$$\big(a_{K''}^{\alpha\beta\rho\sigma}\,\gamma_{\alpha\beta}^{K''}(\mathbf{w}_N),\gamma_{\rho\sigma}^{K''}(\mathbf{v}_N)\,(\sqrt{a})_{K''}\big)_M=\int_\omega a_{K''}^{\alpha\beta\rho\sigma}\gamma_{a\beta}^{K''}(\mathbf{w}_N)\gamma_{\rho\sigma}^{K''}(\mathbf{v}_N)\,(\sqrt{a})_{K''}\,d\mathbf{x}.$$

Inserting this equality in the definition of $a^M(\cdot;\cdot)$ and similar ones for the other terms of $a(\cdot;\cdot)$ and $a_M(\cdot;\cdot)$ leads to the following bound:

$$E_M^a(W_N;V_N)\le c\,\kappa_N\,\|W_N\|_{\mathbb{X}(\omega)}\|V_N\|_{\mathbb{X}(\omega)},$$

where the quantity κ_N is equal to

$$\begin{aligned}\kappa_N=\max\Big\{&\|\mathbf{a}_k-\mathbf{a}_{kK''}\|_{L^\infty(\omega)^3},\|\mathbf{a}_k-\mathbf{a}_{kN}\|_{L^\infty(\omega)^3},\\&\|\partial_\alpha\mathbf{a}_3-\mathbf{c}_{\alpha K''}\|_{L^\infty(\omega)^3},\|\partial_\alpha\mathbf{a}_3-\mathbf{c}_{\alpha N}\|_{L^\infty(\omega)^3},\\&\|a^{\alpha\beta\rho\sigma}-a_{K''}^{\alpha\beta\rho\sigma}\|_{L^\infty(\omega)},\|\sqrt{a}-(\sqrt{a})_{K''}\|_{L^\infty(\omega)}\Big\}.\end{aligned}$$

So the desired estimate is obviously derived from (3.15), (3.16), and (3.18). ∎

LEMMA 4.4. *There exists a constant c only depending on the norms of $\mathbf{a}_3$ in $H^{s_0+1}(\omega)^3$ such that*

$$\forall W_N\in\mathbb{X}_N,\quad\sup_{V_N\in\mathbb{X}_N}\frac{\tilde E_M^a(W_N;V_N)}{\|V_N\|_{\mathbb{X}(\omega)}}\le c\,N^{-s_0}\,(\log\,N)^{\frac12}\,\|W_N\|_{\mathbb{X}(\omega)}.\tag{4.7}$$

Proof. We set $W_N=(\mathbf{w}_N,\mathbf{t}_N)$ and $V_N=(\mathbf{v}_N,\mathbf{s}_N)$. As in the proof of Proposition 3.6, we take K equal to the integer part of $\mu\,N-1$ and consider an approximation $\mathbf{a}_{3K}$ of $\mathbf{a}_3$ in $\mathbb{P}_K(\omega)^3$ which still satisfies (3.16) and (3.17). Thus, $\mathcal{I}_M(\mathbf{t}_N\cdot\mathbf{a}_{3K})$ and $\mathcal{I}_M(\mathbf{s}_N\cdot\mathbf{a}_{3K})$ are equal to $\mathbf{t}_N\cdot\mathbf{a}_{3K}$ and $\mathbf{s}_N\cdot\mathbf{a}_{3K}$, respectively, whence

$$\big(\partial_\alpha\mathcal{I}_M(\mathbf{t}_N\cdot\mathbf{a}_{3K}),\partial_\alpha\mathcal{I}_M(\mathbf{s}_N\cdot\mathbf{a}_{3K})\big)_M=\int_\omega\partial_\alpha(\mathbf{t}_N\cdot\mathbf{a}_{3K})\partial_\alpha(\mathbf{s}_N\cdot\mathbf{a}_{3K})\,d\mathbf{x}.$$

Adding and subtracting this equality and using the continuity of $\tilde a_M(\cdot;\cdot)$ proved in Lemma 3.4, we derive

$$\begin{aligned}\frac{\tilde E_M^a(W_N;V_N)}{\|V_N\|_{\mathbb{X}(\omega)}}\le c\,\Big(&\|W_N\|_{\mathbb{X}(\omega)}\big(\|\mathbf{s}_N\cdot(\mathbf{a}_{3N}-\mathbf{a}_{3K})\|_{H^1(\omega)}+\|\mathbf{s}_N\cdot(\mathbf{a}_3-\mathbf{a}_{3K})\|_{H^1(\omega)}\big)\\&+\|V_N\|_{\mathbb{X}(\omega)}\big(\|\mathbf{t}_N\cdot(\mathbf{a}_{3N}-\mathbf{a}_{3K})\|_{H^1(\omega)}+\|\mathbf{t}_N\cdot(\mathbf{a}_3-\mathbf{a}_{3K})\|_{H^1(\omega)}\big)\Big).\end{aligned}$$

Then we use the inequality, with $\frac1p+\frac1q=\frac12$,

$$\|\mathbf{s}_N\cdot(\mathbf{a}_3-\mathbf{a}_{3K})\|_{H^1(\omega)}\le\|\mathbf{s}_N\|_{H^1(\omega)^3}\|\mathbf{a}_3-\mathbf{a}_{3K}\|_{L^\infty(\omega)^3}+\|\mathbf{s}_N\|_{L^q(\omega)^3}\|\mathbf{a}_3-\mathbf{a}_{3K}\|_{W^{1,p}(\omega)^3}$$

and similar ones for the other terms. Combining this with (3.16) and (3.17), using once more the fact that the norm of the embedding of $H^1(\omega)$ into $L^q(\omega)$ is smaller than $c\,q^{\frac12}$ and taking q equal to $\log N$, gives the desired result. ∎

The proof of the next lemma relies on exactly the same arguments as for Proposition 3.6. So we omit it.

LEMMA 4.5. *There exists a constant c only depending on the norms of $\mathbf{a}_3$ in $H^{s_0+1}(\omega)^3$ such that*

$$\forall \chi_N \in \mathbb{M}_N, \quad \sup_{V_N \in \mathbb{X}_N} \frac{E_M^b(V_N;\chi_N)}{\|V_N\|_{\mathbb{X}(\omega)}} \leq c\, N^{-s_0}\, (\log\, N)^{\frac{1}{2}}\, \|\chi_N\|_{H^1(\omega)}, \tag{4.8}$$

and

$$\forall Z_N \in \mathbb{X}_N, \quad \sup_{\omega_N \in \mathbb{M}_N} \frac{E_M^b(Z_N;\omega_N)}{\|\omega_N\|_{H^1(\omega)}} \leq c\, N^{-s_0}\, (\log\, N)^{\frac{1}{2}}\, \|Z_N\|_{\mathbb{X}(\omega)}. \tag{4.9}$$

LEMMA 4.6. *Assume that the data $(\mathbf{f},\mathbf{M},\mathbf{N})$ belong to $H^{s_1}(\omega)^3 \times H^{s_1}(\gamma_1)^3 \times H^{s_1}(\gamma_1)^3$ for a real number $s_1 > 1$. There exists a constant $\tilde{c}$ only depending on the norms of the $\mathbf{a}_\alpha$ in $H^{s_0}(\omega)^3$ such that*

$$\sup_{V_N \in \mathbb{X}_N} \frac{E_M^L(V_N)}{\|V_N\|_{\mathbb{X}(\omega)}} \leq c\left(\tilde{c}\, N^{1-s_0}\, (\log\, N)^{\frac{1}{2}} + c(\mathbf{f},\mathbf{M},\mathbf{N})\, N^{-s_1}\right), \tag{4.10}$$

with the quantity $c(\mathbf{f},\mathbf{M},\mathbf{N})$ equal to $\|\mathbf{f}\|_{H^{s_1}(\omega)^3} + \|\mathbf{M}\|_{H^{s_1}(\gamma_1)^3} + \|\mathbf{N}\|_{H^{s_1}(\gamma_1)^3}$.

Proof. If K denotes the integer part of $\mu\, N - 1$ and $(\sqrt{a})_K$ an approximation of $\sqrt{a}$ in $\mathbb{P}_K(\omega)$ which satisfies (3.15) (we recall that $\sqrt{a}$ belongs to $H^{s_0}(\omega)$), we derive from (3.2) the identity, for all $V_N = (\mathbf{v}_N, \mathbf{s}_N)$ in $\mathbb{X}_N$,

$$\left(\mathbf{f}, \mathbf{v}_N\, (\sqrt{a})_K\, \right)_M = \int_\omega \mathcal{I}_M \mathbf{f} \cdot \mathbf{v}_N (\sqrt{a}\,)_K\, d\mathbf{x},$$

whence

$$\left| \int_\omega \mathbf{f} \cdot \mathbf{v}_N \sqrt{a}\, d\mathbf{x} - \left(\mathbf{f}, \mathbf{v}_N\, \sqrt{a}\right)_M \right|$$
$$\leq \left(\|\sqrt{a} - (\sqrt{a})_K\|_{L^\infty(\omega)} \|\mathbf{f}\|_{L^2(\omega)^3} + \|(\sqrt{a})_K\|_{L^\infty(\omega)} \|\mathbf{f} - \mathcal{I}_M \mathbf{f}\|_{L^2(\omega)^3} \right) \|\mathbf{v}_N\|_{L^2(\omega)^3}.$$

Similar but simpler arguments also yield analogous estimates for the terms involving $\mathbf{M}$ and $\mathbf{N}$. So estimate (4.10) is a direct consequence of (3.15) and the approximation properties of the operators $\mathcal{I}_M$ and $i_M^{\gamma_1}$; see [4, Thms. 13.4 and 14.2].

To conclude, we insert (4.3) into (4.1). Next, we use (4.5) and Lemmas 4.3–4.6 to bound all of the terms on the right-hand side. ∎

THEOREM 4.7. *Assume that:*

(i) *the solution (U^η, ψ^η) of problem (2.17) belongs to $H^S(\omega)^{3\times 3} \times H^S(\omega)$ for a real number $S \geq 1$, and*

(ii) *the data $(\mathbf{f},\mathbf{M},\mathbf{N})$ belongs to $H^{s_1}(\omega)^3 \times H^{s_1}(\gamma_1)^3 \times H^{s_1}(\gamma_1)^3$ for a real number $s_1 > 1$.*

Then, for any integer $N \geq N_0$, the following error estimate holds between this solution (U^η, ψ^η) and the solution (U_N, ψ_N) of problem (3.11):

$$\begin{aligned} \|U^\eta - U_N\|_{\mathbb{X}(\omega)} &\leq c\, \max\{e^{-3}, \eta^{-1}\} \\ &\left(c(U^\eta, \psi^\eta)\, \max\{e, \eta\}\, N^{1-S} + \tilde{c}\, N^{1-s_0}\, (\log\, N)^{\frac{1}{2}} + c(\mathbf{f},\mathbf{M},\mathbf{N})\, N^{-s_1}\right), \end{aligned} \tag{4.11}$$

where s_0 is introduced in Assumption 3.1, the quantity $c(U^\eta, \psi^\eta)$ is equal to $\|U^\eta\|_{H^S(\omega)^{3\times 3}} + \|\psi^\eta\|_{H^S(\omega)}$, the constant $\tilde{c}$ only depends on the coefficients involved in problem (2.10), and the quantity $c(\mathbf{f},\mathbf{M},\mathbf{N})$ is introduced in Lemma 4.6.

Estimate (4.11) is optimal and proves the convergence of the method without any restriction, for a fixed value of e. So taking N large enough gives an error smaller than any fixed tolerance.

REFERENCES

[1] M. Bernadou, *Méthodes d'éléments finis pour les problèmes de coques minces*, Collection "Recherches en Mathématiques Appliquée," 33, Masson, Paris, 1994.

[2] M. Bernadou and P. G. Ciarlet, *Sur l'ellipticité du modèle linéaire de coques de W. T. Koiter*, in Computing Methods in Applied Sciences and Engineering, R. Glowinski and J.-L. Lions, eds., Lecture Notes in Econom. and Math. Systems 134, Springer, Berlin, 1976, pp. 89–136.

[3] C. Bernardi, M. Dauge, and Y. Maday, *Polynomials in the Sobolev World*, Internal report, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie, 2003, Web Publications: www.ann.jussieu.fr/publications/R03038.html.

[4] C. Bernardi and Y. Maday, *Spectral Methods*, in The Handbook of Numerical Analysis V, P.G. Ciarlet and J.-L. Lions, eds., North-Holland, Amsterdam, 1997, pp. 209–485.

[5] C. Bernardi and Y. Maday, *Uniform inf-sup conditions for the spectral discretization of the Stokes problem*, Math. Models Methods Appl. Sci., 9 (1999), pp. 395–414.

[6] C. Bernardi, Y. Maday, and F. Rapetti, *Discrétisations variationnelles de problèmes aux limites elliptiques*, Collection "Mathématiques et Applications" 45, Springer, Berlin, 2004.

[7] A. Blouza, *Existence et unicité pour le modèle de Nagdhi pour une coque peu régulière*, C. R. Acad. Sci. Paris, Série I, 324 (1997), pp. 839–844.

[8] A. Blouza, F. Brezzi, and C. Lovadina, *Sur la classification des coques linéairement élastiques*, C. R. Acad. Sci. Paris, Série I, 328 (1999), pp. 831–836.

[9] A. Blouza, F. Hecht, and H. Le Dret, *Two finite element approximations of Naghdi's shell model in Cartesian coordinates*, SIAM J. Numer. Anal., 44 (2006), pp. 636–654.

[10] A. Blouza and H. Le Dret, *Existence and uniqueness for the linear Koiter model for shells with little regularity*, Quart. Appl. Math., LVII (1999), pp. 317–337.

[11] A. Blouza and H. Le Dret, *Nagdhi's shell model: Existence, uniqueness and continuous dependence on the midsurface*, J. Elasticity, 64 (2001), pp. 199–216.

[12] F. Brezzi and M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer Ser. Comput. Math. 15, Springer, Berlin, 1991.

[13] M. Carrive, P. Le Tallec, and J. Mousso, *Approximation par éléments finis d'un modèle de coques géométriquement exact*, Revue Européenne des Éléments Finis, 4 (1995), pp. 633–662.

[14] D. Chapelle, A. Ferent, and K. J. Bathe, *3D-shell elements and their underlying mathematical model*, Math. Models Methods Appl. Sci., 14 (2004), pp. 105–142.

[15] P. G. Ciarlet, *Mathematical Elasticity, Volume* III: *Theory of Shells*, North-Holland, Amsterdam, 2000.

[16] V. Girault and P.-A. Raviart, *Finite Element Methods for Navier–Stokes Equations, Theory and Algorithms*, Springer, Berlin, 1986.

[17] A. Haraux, *A Sharp Norm Estimate for an Almost Critical Sobolev Imbedding*, manuscript.

[18] N. Kerdid and P. M. Eiroa, *Conforming finite element approximation for shells with little regularity*, Comput. Methods Appl. Mech. Engrg., 188 (2000), pp. 95–107.

[19] C. Lacour and Y. Maday, *La méthode des éléments avec joint appliquée aux méthodes d'approximations "discrete Kirchhoff triangles*," C. R. Acad. Sci. Paris, Série I, 326 (1998), pp. 1237–1242.

[20] P. Le Tallec, J. Mandel, and M. Vidrascu, *A Neumann–Neumann domain decomposition algorithm for solving plate and shell problems*, SIAM J. Numer. Anal., 35 (1998), pp. 836–867.

[21] P. Le Tallec and S. Mani, *Analyse numérique d'un modèle de coque de Koiter discrétisé en base cartésienne par éléments finis DKT*, Modél. Math. Anal. Numér., 32 (1998), pp. 433–450.

[22] J.-L. Lions and E. Magenes, *Problèmes aux Limites Non Homogènes et Applications* I, Dunod, Paris, 1968.

[23] Y. Maday and E. M. Rønquist, *Optimal error analysis of spectral methods with emphasis on non-constant coefficients and deformed geometries*, Comput. Methods Appl. Mech. Engrg., 80 (1990), pp. 91–115.

[24] É. Sanchez-Palencia, *Statique et dynamique des coques minces.* I. *Cas de flexion pure non inhibée*, C. R. Acad. Sci. Paris, Série I, 309 (1989), pp. 411–417.

[25] G. Talenti, *Best constant in Sobolev inequality*, Ann. Math. Pura ed Appl., 110 (1976), pp. 353–372.

SIAM J. NUMER. ANAL.
Vol. 45, No. 6, pp. 2671–2696

ADJOINT CONSISTENCY ANALYSIS OF DISCONTINUOUS GALERKIN DISCRETIZATIONS*

RALF HARTMANN†

Abstract. This paper is concerned with the adjoint consistency of discontinuous Galerkin (DG) discretizations. Adjoint consistency—in addition to consistency—is the key requirement for DG discretizations to be of optimal order in L^2 as well as measured in terms of target functionals. We provide a general framework for analyzing the adjoint consistency of DG discretizations which is also useful for the derivation of adjoint consistent methods. This analysis will be performed for the DG discretizations of the linear advection equation, the interior penalty DG method for elliptic problems, and the DG discretization of the compressible Euler equations. This framework is then used to derive an adjoint consistent DG discretization of the compressible Navier–Stokes equations. Numerical experiments demonstrate the link of adjoint consistency to the accuracy of numerical flow solutions and the smoothness of discrete adjoint solutions.

Key words. discontinuous Galerkin discretization, adjoint consistency, compressible Navier–Stokes equations, continuous adjoint problem, discrete adjoint problem

AMS subject classifications. 65N12, 65N15, 65N30

DOI. 10.1137/060665117

1. Introduction.

The past few years have seen considerable progress in the development and analysis of discontinuous Galerkin (DG) methods. In addition to consistency in numerical analysis the so-called adjoint consistency property of discretizations has experienced increasing interest [1, 11, 12, 24]. Adjoint consistency is the key property of discretizations that ensures optimal order of convergence of the error measured in L^2 as well as in terms of specific target functionals $J(\cdot)$. A typical situation is given by the interior penalty discontinuous Galerkin methods. While both the symmetric version (SIPG) and the nonsymmetric version (NIPG) are of optimal order $\mathcal{O}(h^p)$ measured in H^1, only the symmetric version is adjoint consistent which allows employment of a duality argument resulting in an optimal $\mathcal{O}(h^{p+1})$ order of convergence in L^2. In contrast to that, the nonsymmetric interior penalty method is suboptimal, $\mathcal{O}(h^p)$ in L^2. Similarly, adjoint consistency in conjunction with a duality argument leads to the so-called order doubling in the error measured in target functionals $J(\cdot)$. Whereas the adjoint consistency of the SIPG method results in $\mathcal{O}(h^{2p})$ of the error in $J(\cdot)$, the nonsymmetric version (NIPG) lacks adjoint consistency and target functionals behave like $\mathcal{O}(h^p)$. Adjoint consistency is closely linked to the smoothness of the discrete adjoint solutions. For adjoint consistent discretizations the discrete adjoint problem represents a consistent discretization of the continuous adjoint problem. Consequently, discrete adjoint solutions inherit the smoothness properties of the continuous adjoint solutions. Conversely, it has been seen that adjoint inconsistent discretizations exhibit some nonsmoothness. In particular, in [12] it has been shown that the discrete adjoint solutions arising from the SIPG method are essentially continuous. In contrast to that, the adjoint solutions arising from the

*Received by the editors July 17, 2006; accepted for publication (in revised form) July 20, 2007; published electronically December 7, 2007. This work was supported by the President's Initiative and Networking Fund of the Helmholtz Association of German Research Centres.
http://www.siam.org/journals/sinum/45-6/66511.html

†Institute of Aerodynamics and Flow Technology, DLR (German Aerospace Center), Lilienthalplatz 7, 38108 Braunschweig, Germany, and Institute of Scientific Computing, TU Braunschweig, 38092 Braunschweig, Germany (Ralf.Hartmann@dlr.de).

NIPG method are discontinuous between element interfaces where the jumps in the adjoint solutions persist even as the mesh is refined. The lack of regularity of the adjoint solution leads to the suboptimal rate of convergence of the NIPG method.

Adjoint consistency has been considered (cf. [1]) for linear elliptic problems with homogeneous boundary conditions which results in a characterization of element and interior face terms while ignoring the discretization of boundary terms. However, adjoint consistency is equally important for the discretization of boundary terms and for the discretization of target functionals $J(\cdot)$. In [11] it has been shown that the interior penalty method for Poisson's equation with nonhomogeneous Dirichlet in combination with a specific target functional $J(\cdot)$ results in an adjoint inconsistent discretization of boundary conditions and a nonsmooth adjoint solution even for the SIPG discretization, which is known to be adjoint consistent in the interior of the domain. Only after an appropriate modification of the target functional have adjoint consistency, smoothness of the adjoint solution, and optimal convergence rates in $J(\cdot)$ been recovered; see also [23] for elliptic problems discretized by the BR2 scheme [5]. First results for the compressible Euler equations in [23, 24] indicate that adjoint consistency is of similar importance for nonlinear problems. Whereas adjoint inconsistent discretizations of boundary terms [4, 15] lead to irregular adjoint solutions near a reflective boundary, it has been shown in [23, 24, 14] that a specific discretization of boundary conditions and target functionals is required for recovering the adjoint consistency property and a smooth discrete adjoint solution.

We note that adjoint consistency is of importance also in continuous finite element methods. In [18] it was shown that, for the streamline diffusion (SD) discretization of the linear advection equation, the discrete adjoint problem is not a consistent discretization of the continuous adjoint problem; i.e., the SD discretization is not adjoint consistent. It is, however, asymptotically adjoint consistent.

As outlined so far, adjoint consistency is a desirable property which, however, involves several issues. In addition to the adjoint consistency of element and interior face terms, it involves the discretization of boundary conditions and target functionals. The purpose of this paper is to give a general framework for analyzing the adjoint consistency property of DG discretizations for linear as well as nonlinear problems. This framework includes the derivation of the continuous adjoint problems and boundary conditions provided the primal problems and the target functionals satisfy a compatibility condition. Furthermore, it includes the derivation of the discrete adjoint problems and of primal and adjoint residuals and a discussion of under which conditions the residuals vanish for the exact primal and adjoint solutions, respectively. Additionally, a so-called consistent modification of target functionals is introduced. The analysis is performed for various model problems, recovering properties and conclusions drawn in [1, 11, 24]. In addition, this framework is used to derive an adjoint consistent DG discretization of the compressible Navier–Stokes equations. Altogether, this publication provides a general framework of an adjoint consistency analysis which can be applied to a wide range of (more complex) linear and nonlinear problems.

The paper is structured as follows: We begin by outlining the main ingredients of the framework in section 2, including the definition of adjoint consistency for linear and nonlinear problems. Then in sections 3, 4, and 5 the DG discretization of the linear advection equation, the interior penalty DG method for Poisson's equation, and the DG discretization of the compressible Euler equations are analyzed. Then in section 6 it is shown that the interior penalty DG discretization of the compressible Navier–Stokes equations in [16] is not adjoint consistent. Within the framework appropriate modifications are derived for recovering adjoint consistency. These mod-

ifications include a specific treatment of convective and diffusive fluxes at boundaries and a consistent modification of total force coefficients. Then in section 7 we show some numerical results demonstrating the effect of adjoint consistency on the accuracy of the flow solution and on the smoothness of the discrete adjoint solution.

2. General framework. We begin by defining the adjoint consistency for linear and nonlinear problems. Let Ω be a bounded open domain in $\mathbb{R}^d$ with boundary Γ.

2.1. Linear problems. For $f \in L^2(\Omega)$ and $g \in L^2(\Gamma)$ consider the following linear problem:

$$Lu = f \quad \text{in } \Omega, \qquad\qquad Bu = g \quad \text{on } \Gamma, \tag{2.1}$$

where L denotes a linear differential operator on Ω and B denotes a linear differential (boundary) operator on Γ. Let J be a linear target functional given by

$$J(u) = \int_\Omega j_\Omega \, u \, \mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma \, Cu \, \mathrm{d}s, \tag{2.2}$$

where $j_\Omega \in L^2(\Omega)$, $j_\Gamma \in L^2(\Gamma)$, and C is a differential (boundary) operator on Γ. We say that the target functional (2.2) is *compatible* with (2.1), provided the following compatibility condition based on Green's formula holds (see, e.g., [2, 22] for elliptic problems):

$$(Lu, z)_\Omega + (Bu, C^* z)_\Gamma = (u, L^* z)_\Omega + (Cu, B^* z)_\Gamma, \tag{2.3}$$

where L^*, B^*, and C^* denote the adjoint operators to L, B, and C, respectively, and $(\cdot,\cdot)_\Omega$ and $(\cdot,\cdot)_\Gamma$ denote the $L^2(\Omega)$ and $L^2(\Gamma)$ scalar products, respectively. We note that, for given operators L and B associated with the primal problem (2.1), only some target functionals (2.2) with operators C are compatible, whereas others are not. However, assuming that (2.3) holds the adjoint problem associated to (2.1), (2.2) is given by

$$L^* z = j_\Omega \quad \text{in } \Omega, \qquad\qquad B^* z = j_\Gamma \quad \text{on } \Gamma. \tag{2.4}$$

In an adjoint-based optimization framework (see, e.g., [10]) this ensures that

$$\begin{aligned} J(u) &= (u, j_\Omega)_\Omega + (Cu, j_\Gamma)_\Gamma = (u, L^* z)_\Omega + (Cu, B^* z)_\Gamma \\ &= (Lu, z)_\Omega + (Bu, C^* z)_\Gamma = (f, z)_\Omega + (g, C^* z)_\Gamma. \end{aligned} \tag{2.5}$$

Let Ω be subdivided into shape-regular meshes $\mathcal{T}_h = \{\kappa\}$ consisting of elements κ, and let V_h be a discrete function space on $\mathcal{T}_h$. Furthermore, let V be a broken Sobolev space on $\mathcal{T}_h$ appropriately chosen such that $V_h \subset V$ and $u, z \in V$, where u and z are the solutions to (2.1) and (2.4), respectively. A typical situation is $V = H^2(\mathcal{T}_h)$, with $V_h \subset V$ and $u, z \in H^2(\Omega) \subset V$; see, e.g., [1, 6]. Finally, let $\mathcal{B} : V \times V \to \mathbb{R}$ be a bilinear form such that problem (2.1) is discretized as follows: Find $u_h \in V_h$ such that

$$\mathcal{B}(u_h, v) = \mathcal{F}(v) \quad \forall v \in V_h, \tag{2.6}$$

where $\mathcal{F} : V \to \mathbb{R}$ is a linear form including the prescribed force and boundary data functions f and g. Then the discretization (2.6) is said to be *consistent* if the exact solution $u \in V$ to the primal problem (2.1) satisfies:

$$\mathcal{B}(u, v) = \mathcal{F}(v) \quad \forall v \in V, \tag{2.7}$$

which can then be viewed as a (broken) weak formulation of (2.1); see [6]. We note that defining $\mathcal{B}$ via the discretization scheme (2.6) instead of the weak formulation (2.7) allows us to represent also inconsistent DG discretizations; see, e.g., [1].

Similarly, the discretization (2.6) is said to be *adjoint consistent* if the exact solution $z \in V$ to the adjoint problem (2.4) satisfies:

$$\mathcal{B}(w, z) = J(w) \quad \forall w \in V. \tag{2.8}$$

In the following we generalize this definition to nonlinear problems.

2.2. Nonlinear problems. We now consider the nonlinear problem

$$Nu = 0 \quad \text{in } \Omega, \qquad\qquad Bu = 0 \quad \text{on } \Gamma, \tag{2.9}$$

where N is a nonlinear differential (and Fréchet-differentiable) operator and B is a (possibly nonlinear) boundary operator. Let $J(\cdot)$ be a nonlinear target functional

$$J(u) = \int_\Omega j_\Omega(u)\,\mathrm{d}\mathbf{x} + \int_\Gamma j_\Gamma(Cu)\,\mathrm{d}s, \tag{2.10}$$

with Fréchet derivative

$$J'[u](w) = \int_\Omega j'_\Omega[u]\,w\,\mathrm{d}\mathbf{x} + \int_\Gamma j'_\Gamma[Cu]\,C'[u]w\,\mathrm{d}s, \tag{2.11}$$

where $j_\Omega(\cdot)$ and $j_\Gamma(\cdot)$ may be nonlinear with derivatives j'_Ω and j'_Γ, respectively, and C is a differential boundary operator on Γ and may be nonlinear with derivative C'. Here $'$ denotes the (total) Fréchet derivative, and the square bracket $[\cdot]$ denotes the state about which linearization is performed. Again, we say that the target functional (2.10) is *compatible* with (2.9) provided the following compatibility condition holds:

$$(N'[u]w, z)_\Omega + (B'[u]w, (C'[u])^* z)_\Gamma = (w, (N'[u])^* z)_\Omega + (C'[u]w, (B'[u])^* z)_\Gamma, \tag{2.12}$$

where $(N'[u])^*$, $(B'[u])^*$, and $(C'[u])^*$ denote the adjoint operators to $N'[u]$, $B'[u]$, and $C'[u]$, respectively. This condition is analogous to (2.3), with L, B, and C replaced by $N'[u]$, $B'[u]$, and $C'[u]$, respectively. Assuming that (2.12) holds, the continuous adjoint problem associated to (2.9) and (2.11) is given by

$$(N'[u])^* z = j'_\Omega[u] \quad \text{in } \Omega, \qquad\qquad (B'[u])^* z = j'_\Gamma[Cu] \quad \text{on } \Gamma. \tag{2.13}$$

We note that, in an optimization framework [10], this ensures, analogous to (2.5), that

$$\begin{aligned} J'[u](w) &= (w, j'_\Omega[u])_\Omega + (C'[u]w, j'_\Gamma[Cu])_\Gamma = (w, (N'[u])^* z)_\Omega + (C'[u]w, (B'[u])^* z)_\Gamma \\ &= (N'[u]w, z)_\Omega + (B'[u]w, (C'[u])^* z)_\Gamma. \end{aligned} \tag{2.14}$$

Let $\mathcal{N} : V \times V \to \mathbb{R}$ be a semilinear form, nonlinear in its first and linear in its second argument, such that the nonlinear problem (2.9) is discretized as follows: Find $u_h \in V_h$ such that

$$\mathcal{N}(u_h, v) = 0 \quad \forall v \in V_h. \tag{2.15}$$

Then the discretization (2.15) is said to be consistent if the exact solution $u \in V$ to the primal problem (2.9) satisfies the following equation:

$$\mathcal{N}(u, v) = 0 \quad \forall v \in V. \tag{2.16}$$

Furthermore, the discretization (2.15) is said to be adjoint consistent if the exact solutions $u, z \in V$ to the primal and adjoint problems (2.9) and (2.13), respectively, satisfy the following equation:

$$\mathcal{N}'[u](w,z) = J'[u](w) \quad \forall w \in V, \tag{2.17}$$

where $\mathcal{N}'[u]$ denotes the Fréchet derivatives of $\mathcal{N}(u,v)$ with respect to u.

In other words, a discretization is adjoint consistent if the discrete adjoint problem is a consistent discretization of the continuous adjoint problem. Finally, we note that, in the case of a linear problem and target functional, the definition of adjoint consistency in (2.17) reduces to the definition of linear adjoint consistency given in section 2.1.

2.3. The adjoint consistency analysis. Based on the definition of adjoint consistency in the previous two sections, we outline a framework of analyzing the adjoint consistency of discontinuous Galerkin discretizations. This framework can also be used to derive adjoint consistent DG discretizations.

2.3.1. Derivation of the continuous adjoint problem. Let the primal problem be given by (2.1) or by (2.9) in the nonlinear case. Furthermore, assume that $J(\cdot)$ is a linear (2.2) or linearized (2.11) compatible target functional. Then we derive the continuous adjoint problem (2.4) or (2.13) including adjoint boundary conditions.

We note that the derivation of the adjoint operator $(N'[u])^*$ for nonlinear systems is a considerably more complicated task than deriving L^* for scalar linear problems. Still more involved is the derivation of the adjoint boundary operators $(B'[u])^*$. In the framework of optimal design, [10] gives a general approach of deriving $(B'[u])^*$ and $(C'[u])^*$ assumed to be connected to B, C, N, and $(N'[u])^*$ through (2.12). This approach is based on a matrix representation of boundary operators which for systems of equations leads to lengthy and error prone derivations. In contrast to optimization where both $(B'[u])^*$ and C^* are required, in the following analysis we require only the adjoint operator $(B'[u])^*$. Due to this we can circumvent the approach described in [10] and use a simpler way of deriving the adjoint boundary operators $(B'[u])^*$.

2.3.2. Consistency analysis of the discrete primal problem. We rewrite the discontinuous Galerkin discretization (2.15) of problem (2.9) in the following element-based primal residual form: Find $u_h \in V_h$ such that

$$\sum_{\kappa\in\mathcal{T}_h} \int_\kappa R(u_h) v \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} r(u_h) v \,\mathrm{d}s + \int_\Gamma r_\Gamma(u_h) v \,\mathrm{d}s = 0 \quad \forall v \in V_h, \tag{2.18}$$

where $R(u_h)$, $r(u_h)$, and $r_\Gamma(u_h)$ denote the element, interior face, and boundary residuals, respectively. According to (2.16), the discretization (2.15) is consistent if the exact solution u to (2.9) satisfies

$$\sum_{\kappa\in\mathcal{T}_h} \int_\kappa R(u) v \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} r(u) v \,\mathrm{d}s + \int_\Gamma r_\Gamma(u) v \,\mathrm{d}s = 0 \quad \forall v \in V, \tag{2.19}$$

which holds, provided u satisfies

$$R(u) = 0 \ \text{ in } \kappa, \kappa \in \mathcal{T}_h, \quad r(u) = 0 \ \text{ on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \quad r_\Gamma(u) = 0 \ \text{ on } \Gamma. \tag{2.20}$$

2.3.3. Derivation of the discrete adjoint problem. Given the discretization (2.15), the target functional (2.10), and its linearization (2.11), we derive the discrete

adjoint problem: Find $z_h \in V_h$ such that

$$\mathcal{N}'[u_h](w, z_h) = J'[u_h](w) \quad \forall w \in V_h. \tag{2.21}$$

$\mathcal{N}'[u_h]$ is called the Jacobian of the numerical scheme and is required also for implicit and adjoint methods, e.g., Newton iteration, a posteriori error estimation, adjoint-based adaptation (see [13]), and for optimization.

2.3.4. Adjoint consistency of element, interior face, and boundary terms. We rewrite the discrete adjoint problem (2.21) in element-based adjoint residual form: Find $z_h \in V_h$ such that

$$\sum_{\kappa \in \mathcal{T}_h} \int_\kappa w\, R^*[u_h](z_h)\, \mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} w\, r^*[u_h](z_h)\, \mathrm{d}s + \int_\Gamma w\, r^*_\Gamma[u_h](z_h)\, \mathrm{d}s = 0 \tag{2.22}$$

for all $w \in V_h$, where $R^*[u_h](z_h)$, $r^*[u_h](z_h)$, and $r^*_\Gamma[u_h](z_h)$ denote the element, interior face, and boundary adjoint residuals, respectively. According to (2.17), the discretization (2.15) is adjoint consistent if the exact solutions u and z satisfy

$$\sum_{\kappa \in \mathcal{T}_h} \int_\kappa w\, R^*[u](z)\, \mathrm{d}\mathbf{x} + \sum_{\kappa \in \mathcal{T}_h} \int_{\partial\kappa \setminus \Gamma} w\, r^*[u](z)\, \mathrm{d}s + \int_\Gamma w\, r^*_\Gamma[u](z)\, \mathrm{d}s = 0 \quad \forall w \in V, \tag{2.23}$$

which holds, provided u and z satisfy

$$R^*[u](z) = 0 \ \text{ in } \kappa, \quad r^*[u](z) = 0 \ \text{ on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \quad r^*_\Gamma[u](z) = 0 \ \text{ on } \Gamma. \tag{2.24}$$

We note that the adjoint problem and consequently the adjoint consistency of a discretization depend on the specific target functional $J(\cdot)$ under consideration. Given a target functional of the form (2.10), we see that $R^*[u](z)$ depends on $j_\Omega(\cdot)$ and $r^*_\Gamma[u](z)$ depends on $j_\Gamma(\cdot)$. For obtaining an adjoint consistent discretization it is, in some cases (see following sections) necessary to modify the target functional as follows:

$$\tilde{J}(u_h) = J(i(u_h)) + \int_\Gamma r_J(u_h)\, \mathrm{d}s, \tag{2.25}$$

where $i(\cdot)$ and $r_J(\cdot)$ are functions to be specified. A modification of a target functional is called consistent if $\tilde{J}(u) = J(u)$ holds for the exact solution u. Thereby, the modification in (2.25) is consistent if the exact solution u satisfies $i(u) = u$ and $r_J(u) = 0$. Although the true value of the target functional is unchanged, $\tilde{J}(u) = J(u)$, the computed value $J(u_h)$ of the target functional is modified, and, more importantly, $\tilde{J}'[u_h]$ differs from $J'[u_h]$. This modification can be used to recover an adjoint consistent discretization. We note that (2.25) is not a unique choice of a consistent modification of $J(\cdot)$; other examples are $\tilde{J}(u_h) = J(u_h) + \int_\Omega R_J(u_h)\, \mathrm{d}\mathbf{x}$, with $R_J(u) = 0$, or $\tilde{J}(u_h) = m(J(u_h), J(i(u_h)))$, with $i(u) = u$ and $m(j, j) = j$. However, the consistent modification as given in (2.25) will be sufficient for the purposes of this work.

3. The linear advection equation. We consider the linear advection equation

$$\nabla \cdot (\mathbf{b}u) + cu = f \quad \text{in } \Omega, \qquad u = g \quad \text{on } \Gamma_-, \tag{3.1}$$

where $f \in L^2(\Omega)$ and $c \in L^\infty(\Omega)$ are real-valued and $\mathbf{b} = \{\beta_i\}_{i=1}^d$ is a vector function whose entries β_i are Lipschitz continuous real-valued functions on Ω. By $\Gamma_- = \{x \in \Gamma, \mathbf{b}\cdot\mathbf{n} < 0\}$ we denote the inflow part of the boundary $\Gamma = \partial\Omega$. Furthermore, we adopt the following hypothesis: There exists a $c_0 \in L^\infty(\Omega)$ and a number $\gamma_0 > 0$ such that $c(\mathbf{x}) + \frac{1}{2}\nabla\cdot\mathbf{b}(\mathbf{x}) = c_0^2(\mathbf{x}) \geq \gamma_0 > 0$. To demonstrate the similarities with the compressible Euler equations in section 5, we consider the linear advection equation in conservative form which is equivalent to the nonconservative form $\mathbf{b}\cdot\nabla u + \tilde{c}u = f$ (see, e.g., [19]), with $\tilde{c} - \frac{1}{2}\nabla\cdot\mathbf{b} = c_0^2$ and $\tilde{c} = c + \nabla\cdot\mathbf{b}$.

In order to derive the continuous adjoint problem, we multiply the left-hand side of (3.1) by z and integrate by parts over the domain Ω. Thereby, we obtain

$$(\nabla\cdot(\mathbf{b}u) + cu, z)_\Omega + (u, -\mathbf{b}\cdot\mathbf{n}z)_{\Gamma_-} = (u, -\mathbf{b}\cdot\nabla z + cz)_\Omega + (u, \mathbf{b}\cdot\mathbf{n}z)_{\Gamma_+}. \tag{3.2}$$

From (2.3) we see that, for $Lu = \nabla\cdot(\mathbf{b}u) + cu$ in Ω, and $Bu = u$ and $Cu = 0$ on Γ_-, and $Bu = 0$ and $Cu = u$ on Γ_+, we have $L^*z = -\mathbf{b}\cdot\nabla z + cz$ in Ω, $B^*z = 0$ and $C^*z = -\mathbf{b}\cdot\mathbf{n}z$ on Γ_-, and $B^*z = \mathbf{b}\cdot\mathbf{n}z$ and $C^*z = 0$ on Γ_+. In particular, the target functional $J(u) = \int_\Omega j_\Omega\, u\,\mathrm{d}\mathbf{x} + \int_{\Gamma_+} j_\Gamma\, u\,\mathrm{d}s$ is compatible, and the adjoint problem is given by

$$-\mathbf{b}\cdot\nabla z + cz = j_\Omega \quad \text{in } \Omega, \qquad\qquad \mathbf{b}\cdot\mathbf{n}\, z = j_\Gamma \quad \text{on } \Gamma_+. \tag{3.3}$$

Let Ω be subdivided into shape-regular meshes $\mathcal{T}_h = \{\kappa\}$ consisting of elements κ, and let V_h^p be the discrete function space consisting of discontinuous piecewise polynomial functions of degree $p \geq 0$. Suppose that $v|_\kappa \in H^1(\kappa)$ for each $\kappa \in \mathcal{T}_h$. Let κ^+ and κ^- be two adjacent elements of $\mathcal{T}_h$ and $\mathbf{x}$ be an arbitrary point on the interior edge $e = \partial\kappa^+ \cap \partial\kappa^- \subset \Gamma_\mathcal{I}$, where $\Gamma_\mathcal{I}$ denotes the union of all interior edges of $\mathcal{T}_h$. Moreover, let v and $\boldsymbol{\phi}$ be a scalar and a d-vector-valued function, respectively, that are smooth inside each element $\kappa^\pm$. By $u^\pm := u|_{\partial\kappa^\pm}$ and $\boldsymbol{\phi}^\pm := \boldsymbol{\phi}|_{\partial\kappa^\pm}$ we denote the traces of u and $\boldsymbol{\phi}$, respectively, on e taken from within the interior of $\kappa^\pm$. Then we define the averages at $\mathbf{x} \in e$ by $\{v\} = (v^+ + v^-)/2$ and $\{\boldsymbol{\phi}\} = (\boldsymbol{\phi}^+ + \boldsymbol{\phi}^-)/2$. Similarly, the jump at $\mathbf{x} \in e$ is given by $[\![v]\!] = v^+\mathbf{n}^+ + v^-\mathbf{n}^-$ and by $[\![\boldsymbol{\phi}]\!] = \boldsymbol{\phi}^+\cdot\mathbf{n}^+ + \boldsymbol{\phi}^-\cdot\mathbf{n}^-$. On a boundary edge $e \subset \Gamma$ the average and jump operators are defined by $\{v\} = v^+$, $\{\boldsymbol{\phi}\} = \boldsymbol{\phi}^+$, $[\![v]\!] = v^+\mathbf{n}^+$, and $[\![\boldsymbol{\phi}]\!] = \boldsymbol{\phi}^+\cdot\mathbf{n}^+$.

The DG discretization of (3.1) (e.g., [12, 7]) is given by: Find $u_h \in V_h^p$ such that

$$\begin{aligned}\mathcal{B}(u_h, v) \equiv &-\int_\Omega u_h\mathbf{b}\cdot\nabla_h v\,\mathrm{d}\mathbf{x} + \int_\Omega cu_h v\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_-\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\, u_h^- v^+\,\mathrm{d}s \\ &+ \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_+} \mathbf{b}\cdot\mathbf{n}\, u_h^+ v^+\,\mathrm{d}s = \int_\Omega fv\,\mathrm{d}\mathbf{x} - \int_{\Gamma_-}\mathbf{b}\cdot\mathbf{n}\, gv^+\,\mathrm{d}s\end{aligned} \tag{3.4}$$

for all $v \in V_h^p$. Then integration by parts on each $\kappa \in \mathcal{T}_h$ yields: Find $u_h \in V_h^p$:

$$\int_\Omega (\nabla_h\cdot(\mathbf{b}u_h) + cu_h)\, v\,\mathrm{d}\mathbf{x} - \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa_-}\mathbf{b}\cdot[\![u_h]\!]v^+\,\mathrm{d}s = \int_\Omega fv\,\mathrm{d}\mathbf{x} - \int_{\Gamma_-}\mathbf{b}\cdot\mathbf{n}\, gv^+\,\mathrm{d}s$$

for all $v \in V_h^p$. Hence we have the primal residual form (2.18) with

$$\begin{aligned} R(u_h) &= f - \nabla_h\cdot(\mathbf{b}u_h) - cu_h && \text{in } \kappa, \kappa \in \mathcal{T}_h, \\ r(u_h) &= \mathbf{b}\cdot[\![u_h]\!] && \text{on } \partial\kappa\setminus\Gamma, \kappa \in \mathcal{T}_h, \\ r_\Gamma(u_h) &= \mathbf{b}\cdot\mathbf{n}\,(u_h - g) && \text{on } \Gamma_-, \end{aligned}$$

and $r_\Gamma(u_h) \equiv 0$ on Γ_+. As (2.19) holds for the exact solution u to (3.1), we conclude that (3.4) is a consistent discretization of (3.1). Substituting

$$\sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa_-\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\, u_h^- v^+ \,\mathrm{d}s = -\sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa_+\setminus\Gamma} \mathbf{b}\cdot\mathbf{n}\, u_h^+ v^- \,\mathrm{d}s$$

in (3.4), we find that the discrete adjoint problem to the discretization (3.4) is given by: Find $z_h \in V_h$ such that

$$\mathcal{B}(w, z_h) \equiv \int_\Omega w\,(-\mathbf{b}\cdot\nabla_h z_h + c z_h)\,\mathrm{d}\mathbf{x} + \sum_\kappa \int_{\partial\kappa_+\setminus\Gamma} w^+\,\mathbf{b}\cdot[\![z_h]\!]\,\mathrm{d}s = J(w)$$

for all $w \in V_h$. Hence we have the adjoint residual form (2.22) with

$$\begin{aligned}
R^*(z_h) &= j_\Omega + \mathbf{b}\cdot\nabla_h z_h - c z_h && \text{in } \kappa, \kappa\in\mathcal{T}_h,\\
r^*(z_h) &= -\mathbf{b}\cdot[\![z_h]\!] && \text{on } \partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h, \qquad (3.5)\\
r_\Gamma^*(z_h) &= j_\Gamma - \mathbf{b}\cdot\mathbf{n}\, z_h && \text{on } \Gamma_+, \qquad (3.6)
\end{aligned}$$

and $r_\Gamma^*(z_h) \equiv 0$ on Γ_-. As the adjoint residuals vanish for the exact solution z to (3.3), we conclude that (3.4) is an adjoint consistent discretization of (3.1).

4. Poisson's equation. We now consider the elliptic model problem

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g_D \quad \text{on } \Gamma_D, \qquad \mathbf{n}\cdot\nabla u = g_N \quad \text{on } \Gamma_N, \tag{4.1}$$

where $f \in L^2(\Omega)$, $g_D \in L^2(\Gamma_D)$, and $g_N \in L^2(\Gamma_N)$ are given functions. We assume that Γ_D and Γ_N are disjoint subsets with union Γ. We also assume that Γ_D is nonempty.

In order to derive the continuous adjoint problem we multiply the left-hand side of (4.1) by z and integrate twice by parts over the domain Ω. Thereby, we obtain

$$(-\Delta u, z)_\Omega + (u, -\mathbf{n}\cdot\nabla z)_{\Gamma_D} + (\mathbf{n}\cdot\nabla u, z)_{\Gamma_N} = (u, -\Delta z)_\Omega + (\mathbf{n}\cdot\nabla u, -z)_{\Gamma_D} + (u, \mathbf{n}\cdot\nabla z)_{\Gamma_N}.$$

From (2.3) we see that, for $Lu = -\Delta u$ in Ω, $Bu = u$ and $Cu = \mathbf{n}\cdot\nabla u$ on Γ_D, and $Bu = \mathbf{n}\cdot\nabla u$ and $Cu = u$ on Γ_N, we have $L^*z = -\Delta z$ in Ω, $B^*z = -z$ and $C^*z = -\mathbf{n}\cdot\nabla z$ on Γ_D, and $B^*z = \mathbf{n}\cdot\nabla z$ and $C^*z = z$ on Γ_N. Then (2.2) reduces to

$$J(u) = \int_\Omega j_\Omega\, u\,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} j_D\,\mathbf{n}\cdot\nabla u\,\mathrm{d}s + \int_{\Gamma_N} j_N\, u\,\mathrm{d}s. \tag{4.2}$$

This target functional is compatible, and the continuous adjoint problem is given by

$$-\Delta z = j_\Omega \quad \text{in } \Omega, \qquad z = j_D \quad \text{on } \Gamma_D, \qquad \mathbf{n}\cdot\nabla z = j_N \quad \text{on } \Gamma_N. \tag{4.3}$$

The method by Baumann–Oden and the symmetric and nonsymmetric interior penalty methods can be expressed as follows (see, e.g., [25]): Find $u_h \in V_h^p$ such that

$$\begin{aligned}
\mathcal{B}(u_h, v) \equiv{}& \int_\Omega \nabla_h u_h\cdot\nabla_h v\,\mathrm{d}\mathbf{x}\\
&+ \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \left(\theta[\![u_h]\!]\cdot\{\nabla_h v\} - \{\nabla_h u_h\}\cdot[\![v]\!]\right)\mathrm{d}s + \int_{\Gamma_\mathcal{I}\cup\Gamma_D} \delta[\![u_h]\!]\cdot[\![v]\!]\,\mathrm{d}s\\
={}& \int_\Omega f v\,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} \theta g_D \mathbf{n}\cdot\nabla v\,\mathrm{d}s + \int_{\Gamma_D} \delta g_D v\,\mathrm{d}s + \int_{\Gamma_N} g_N v\,\mathrm{d}s \qquad (4.4)
\end{aligned}$$

for all $v \in V_h^p$, where the constants θ and δ are given by $\theta = 1, \delta = 0$ for Baumann–Oden, by $\theta = 1, \delta > 0$ for NIPG, and by $\theta = -1, \delta > \delta_0 > 0$ for SIPG.

The scheme in (4.4) is given in face-based form including integrals over all interior faces $\Gamma_{\mathcal{I}}$. Rewriting the interior face terms in element-based form, we obtain

$$\begin{aligned}\mathcal{B}(u_h, v) = &\int_\Omega \nabla_h u_h \cdot \nabla_h v \,\mathrm{d}\mathbf{x} + \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \frac{1}{2}\theta[\![u_h]\!] \cdot \nabla_h v \,\mathrm{d}s \\ &- \sum_\kappa \int_{\partial\kappa\setminus\Gamma_N} \{\nabla_h u_h\} \cdot \mathbf{n}\, v \,\mathrm{d}s + \sum_\kappa \int_{\partial\kappa} \delta[\![u_h]\!] \cdot \mathbf{n} v \,\mathrm{d}s + \int_{\Gamma_D} \theta u_h \mathbf{n} \cdot \nabla_h v \,\mathrm{d}s. \end{aligned} \tag{4.5}$$

Using integration by parts and the relation

$$\nabla_h u^+ \cdot \mathbf{n}^+ v^+ = \{\nabla_h u\} \cdot \mathbf{n}^+ v^+ + \frac{1}{2}[\![\nabla_h u]\!] v^+, \tag{4.6}$$

we can rewrite (4.4) as follows: Find $u_h \in V_h^p$ such that

$$\begin{aligned}\mathcal{B}(u_h, v) \equiv &-\int_\Omega \Delta_h u_h v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_N} \nabla_h u_h \cdot \mathbf{n}\, v \,\mathrm{d}s + \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \frac{1}{2}\theta[\![u_h]\!] \cdot \nabla_h v \,\mathrm{d}s \\ &+ \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \left(\frac{1}{2}[\![\nabla_h u_h]\!] + \delta[\![u_h]\!] \cdot \mathbf{n}\right) v \,\mathrm{d}s + \int_{\Gamma_D} \theta u_h \mathbf{n} \cdot \nabla_h v \,\mathrm{d}s + \int_{\Gamma_D} \delta u_h v \,\mathrm{d}s \\ = &\int_\Omega f v \,\mathrm{d}\mathbf{x} + \int_{\Gamma_D} \theta g_D \mathbf{n} \cdot \nabla v \,\mathrm{d}s + \int_{\Gamma_D} \delta g_D v \,\mathrm{d}s + \int_{\Gamma_N} g_N v \,\mathrm{d}s \end{aligned} \tag{4.7}$$

for all $v \in V_h^p$. Hence we have the element-based primal residual form

$$\begin{aligned}&\int_\Omega R(u_h) v \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} r(u_h) v + \boldsymbol{\rho}(u_h) \cdot \nabla v \,\mathrm{d}s \\ &\qquad + \int_\Gamma r_\Gamma(u_h) v + \boldsymbol{\rho}_\Gamma(u_h) \cdot \nabla v \,\mathrm{d}s = 0 \quad \forall v \in V_h, \end{aligned} \tag{4.8}$$

where the residuals are given by $R(u_h) = f + \Delta_h u_h$ in Ω, and

$$\begin{aligned} &r(u_h) = -\tfrac{1}{2}[\![\nabla_h u_h]\!] - \delta[\![u_h]\!] \cdot \mathbf{n}, && \boldsymbol{\rho}(u) = -\tfrac{1}{2}\theta[\![u_h]\!] && \text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\ &r_\Gamma(u_h) = \delta(g_D - u_h), && \boldsymbol{\rho}_\Gamma(u_h) = \theta(g_D - u_h)\mathbf{n} && \text{on } \Gamma_D, \\ &r_\Gamma(u_h) = g_N - \mathbf{n} \cdot \nabla_h u_h, && \boldsymbol{\rho}_\Gamma(u_h) = 0 && \text{on } \Gamma_N. \end{aligned} \tag{4.9}$$

We note that (4.8) is a generalization to (2.18) to include ∇v terms. Furthermore, we note that the discretization is consistent as the residuals in (4.9) vanish for the exact solution u to (4.1). Given the target functional defined in (4.2), the discrete adjoint problem (2.8) to the discretization (4.4) is given by: Find $z_h \in V_h$ such that

$$\int_\Omega \nabla_h w \cdot \nabla_h z_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\mathcal{I}}\cup\Gamma_D} \left(\theta[\![w]\!] \cdot \{\nabla_h z_h\} - \{\nabla_h w\} \cdot [\![z_h]\!] + \delta[\![w]\!] \cdot [\![z_h]\!]\right) \mathrm{d}s = J(w)$$

for all $w \in V_h$. Then in element-based form we have: Find $z_h \in V_h$ such that

$$\begin{aligned}&\int_\Omega \nabla_h w \cdot \nabla_h z_h \,\mathrm{d}\mathbf{x} + \sum_\kappa \int_{\partial\kappa\setminus\Gamma_N} w \left(\theta \mathbf{n} \cdot \{\nabla_h z_h\} + \delta[\![z_h]\!] \cdot \mathbf{n}\right) \mathrm{d}s \\ &\qquad - \frac{1}{2}\sum_\kappa \int_{\partial\kappa\setminus\Gamma} \nabla_h w \cdot [\![z_h]\!] \,\mathrm{d}s - \int_{\Gamma_D} \nabla_h w \cdot \mathbf{n} z_h \,\mathrm{d}s = J(w) \quad \forall w \in V_h. \end{aligned}$$

Using integration by parts and (4.6) with u_h and v replaced by z_h and w yields

$$
\begin{aligned}
-\int_\Omega w\Delta_h z_h \,\mathrm{d}x + \sum_\kappa \int_{\partial\kappa\setminus\Gamma} w\left(\frac{1}{2}[\![\nabla_h z_h]\!] + (1+\theta)\mathbf{n}\cdot\{\nabla_h z_h\} + \delta[\![z_h]\!]\cdot\mathbf{n}\right)\mathrm{d}s \\
- \sum_\kappa \int_{\partial\kappa\setminus\Gamma} \frac{1}{2}\nabla_h w\cdot[\![z_h]\!]\,\mathrm{d}s + \int_{\Gamma_N} w\mathbf{n}\cdot\nabla_h z_h\,\mathrm{d}s \\
+ \int_{\Gamma_D} w\left((1+\theta)\mathbf{n}\cdot\nabla_h z_h + \delta z_h\right)\mathrm{d}s - \int_{\Gamma_D}\nabla_h w\cdot\mathbf{n} z_h\,\mathrm{d}s \\
= \int_\Omega w\, j_\Omega\,\mathrm{d}\mathbf{x} + \int_{\Gamma_D}\nabla w\cdot\mathbf{n}\, j_D\,\mathrm{d}s + \int_{\Gamma_N} w\, j_N\,\mathrm{d}s,
\end{aligned}
$$

and we obtain the element-based adjoint residual form: Find $z_h \in V_h$ such that

$$
\begin{aligned}
&\int_\Omega w\, R^*(z_h)\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} w\, r^*(z_h) + \nabla w\cdot\boldsymbol{\rho}^*(z_h)\,\mathrm{d}s \\
&\qquad + \int_\Gamma w\, r_\Gamma^*(z_h) + \nabla w\cdot\boldsymbol{\rho}_\Gamma^*(z_h)\,\mathrm{d}s = 0 \quad \forall w\in V_h,
\end{aligned} \tag{4.10}
$$

where the adjoint residuals are given by $R^*(z_h) = j_\Omega + \Delta_h z_h$ in Ω, by

$$
r^*(z_h) = -\tfrac{1}{2}[\![\nabla_h z_h]\!] - (1+\theta)\mathbf{n}\cdot\{\nabla_h z\} - \delta[\![z_h]\!]\cdot\mathbf{n}, \qquad \boldsymbol{\rho}^*(z_h) = \tfrac{1}{2}[\![z_h]\!], \tag{4.11}
$$

on interior faces $\partial\kappa\setminus\Gamma, \kappa\in\mathcal{T}_h$, and by

$$
\begin{aligned}
&r_\Gamma^*(z_h) = -(1+\theta)\mathbf{n}\cdot\nabla_h z_h - \delta z_h, && \boldsymbol{\rho}_\Gamma^*(z_h) = (j_D + z_h)\mathbf{n} && \text{on } \Gamma_D, \\
&r_\Gamma^*(z_h) = j_N - \mathbf{n}\cdot\nabla_h z_h, && \boldsymbol{\rho}_\Gamma^*(z_h) = 0 && \text{on } \Gamma_N.
\end{aligned} \tag{4.12}
$$

From (4.11) we see that the exact solution z to the adjoint problem (4.3) satisfies $r^*(z) = 0$, provided $\theta = -1$. Furthermore, we have $R^*(z) = 0$. This shows the well-known result (see, e.g., [1]) that the method by Baumann–Oden and NIPG are not adjoint consistent whereas the interior face terms of SIPG are adjoint consistent. In fact, in [12] it has been demonstrated that the lack of adjoint consistency of the NIPG method leads to nonsmooth adjoint solutions and a suboptimal convergence of the method for the primal problem. In contrast to that, the adjoint consistent SIPG method shows an optimal order of convergence.

As $r_\Gamma^*(z) = 0$ and $\boldsymbol{\rho}_\Gamma^*(z) = 0$ on Γ_N the SIPG method is also adjoint consistent on Γ_N. However, on Γ_D the requirements $r_\Gamma^*(z) = 0$ and $\boldsymbol{\rho}_\Gamma^*(z) = 0$ reduce to the conditions $z = 0$ (note that $\theta = -1$) and $z = -j_D$, respectively, which are compatible for $j_D = 0$ but conflict for $j_D \neq 0$. This incompatibility can be resolved by modifying the target functional according to (2.25), with $i(u_h) = u_h$ and

$$
r_J(u_h) = -\delta(u_h - g_D)j_D, \tag{4.13}
$$

which in the following will be denoted by the *IP modification* of the target functional. This modification is consistent, as $i(u) = u$ and $r_J(u) = 0$ hold for the exact solution u to (4.1). As the modified functional is not linear in u_h (it is affine), the discrete adjoint problem includes its linearization as follows: Find $z_h \in V_h$ such that

$$
\mathcal{B}(w, z_h) = \tilde{J}'[u](w) \quad \forall w\in V_h, \tag{4.14}
$$

where

$$\tilde{J}'[u](w) = J'[u](w) + \int_{\Gamma_D} r'_J[u](w)\,\mathrm{d}s = J(w) - \int_{\Gamma_D} w\,\delta j_D\,\mathrm{d}s. \tag{4.15}$$

Then the adjoint residuals on Γ_D are given by

$$r^*_\Gamma(z_h) = -\delta j_D - (1+\theta)\mathbf{n}\cdot\nabla_h z_h - \delta z_h, \quad \boldsymbol{\rho}^*_\Gamma(z_h) = (j_D + z_h)\mathbf{n} \quad \text{on } \Gamma_D, \tag{4.16}$$

which vanish for $z = -j_D$. Hence the SIPG method is adjoint consistent also on Γ_D.

In contrast to the presentation in [11, 12], where the $\frac{1}{2}[\![\nabla_h z]\!]$ term in the interelement conditions analogous to (4.11) has been omitted, we see that for $\theta = -1$ there is a clear correspondence of the adjoint residuals to the primal residuals (4.9). In fact, the discrete adjoint equations correspond to the discrete primal equations with u, f, g_D, and g_N replaced by z, j_Ω, $-j_D$, and j_N, respectively; i.e., the discrete adjoint equation to the SIPG discretization represents an SIPG discretization of the continuous adjoint equation.

Furthermore, we note that [11] considers the target functional $J(u) = \int_{\Gamma_0} \mathbf{n}\cdot\nabla u j_D\,\mathrm{d}s$, $\Gamma_0 \subset \Gamma_D$, which is a special case of (4.2) with $j_\Omega \equiv 0$ in Ω, $j_N \equiv 0$ on Γ_N, and $j_D \equiv 0$ on $\Gamma_D \setminus \Gamma_0$. Numerical experiments in [11] have shown that the discrete adjoint solution associated with this target functional is nonsmooth near Γ_0. Furthermore, it has been demonstrated that, either by setting $\delta = 0$ on Γ_0 or by modifying the target functional appropriately, this effect vanishes, and the adjoint solution becomes smooth. We note that the modification of the target functional proposed in [11] is connected to (4.15). However, here we derive (4.15) in the more general framework of consistent modifications of target functionals; see (2.25).

5. The compressible Euler equations. In this section we consider the two-dimensional stationary compressible Euler equations

$$\nabla\cdot\mathcal{F}^c(\mathbf{u}) = 0 \quad \text{in } \Omega, \tag{5.1}$$

subject to various boundary conditions, e.g., slip-wall boundary conditions at solid wall boundaries $\Gamma_W \subset \Gamma$, with vanishing normal velocity $\mathbf{n}\cdot\mathbf{v} = n_1 v_1 + n_2 v_2 = 0$; i.e.,

$$B\mathbf{u} = n_1 u_2 + n_2 u_3 = 0 \quad \text{on } \Gamma_W \tag{5.2}$$

is imposed, where the vector of conservative variables $\mathbf{u} = (u_1, u_2, u_3, u_4)^\top$ and the convective fluxes $\mathcal{F}^c(\mathbf{u}) = (\mathbf{f}^c_1(\mathbf{u}), \mathbf{f}^c_2(\mathbf{u}))$ are defined by

$$\mathbf{u} = \begin{bmatrix} \rho \\ \rho v_1 \\ \rho v_2 \\ \rho E \end{bmatrix}, \quad \mathbf{f}^c_1(\mathbf{u}) = \begin{bmatrix} \rho v_1 \\ \rho v_1^2 + p \\ \rho v_1 v_2 \\ \rho H v_1 \end{bmatrix}, \quad \text{and} \quad \mathbf{f}^c_2(\mathbf{u}) = \begin{bmatrix} \rho v_2 \\ \rho v_1 v_2 \\ \rho v_2^2 + p \\ \rho H v_2 \end{bmatrix}, \tag{5.3}$$

and ρ, $\mathbf{v} = (v_1, v_2)^\top$, p, and E denote the density, velocity vector, pressure, and specific total energy, respectively. Additionally, H is the total enthalpy given by $H = E + \frac{p}{\rho} = e + \frac{1}{2}\mathbf{v}^2 + \frac{p}{\rho}$, where e is the specific static internal energy, and the pressure is determined by the equation of state of an ideal gas $p = (\gamma - 1)\rho e$, where $\gamma = c_p/c_v$ is the ratio of specific heat capacities at constant pressure c_p and constant volume c_v; for dry air $\gamma = 1.4$. Let us consider the target functional

$$J(\mathbf{u}) = \int_\Gamma j(C\mathbf{u})\,\mathrm{d}s = \int_{\Gamma_W} p(\mathbf{u})\,\mathbf{n}\cdot\boldsymbol{\psi}_{\Gamma_W}\,\mathrm{d}s, \tag{5.4}$$

with $C\mathbf{u} = p(\mathbf{u})$, $j(p) = p\,\mathbf{n}\cdot\boldsymbol{\psi}_{\Gamma_W}$ on Γ_W and $j(p) \equiv 0$ elsewhere, and $\boldsymbol{\psi}_{\Gamma_W} \in [L^2(\Gamma_W)]^2$. As we will see later, this target functional is compatible with (5.1) and (5.2). The most important target quantities of type (5.4) in inviscid compressible flows are the pressure-induced drag and lift coefficients c_{dp} and c_{lp}, where $\boldsymbol{\psi}_{\Gamma_W} = \frac{1}{C_\infty}\boldsymbol{\psi}$. Here $\boldsymbol{\psi}$ is given by $\boldsymbol{\psi}_d = (\cos(\alpha), \sin(\alpha))^\top$ and $\boldsymbol{\psi}_l = (-\sin(\alpha), \cos(\alpha))^\top$ for the drag and lift coefficients, respectively. Furthermore, $C_\infty = \frac{1}{2}\gamma p_\infty M_\infty^2 \bar{l} = \frac{1}{2}\gamma\frac{|\mathbf{v}_\infty|^2}{c_\infty^2} p_\infty \bar{l} = \frac{1}{2}\rho_\infty |\mathbf{v}_\infty|^2 \bar{l}$, where M_∞ denotes the Mach number at free-stream conditions, c_∞ is the speed of sound defined by $c_\infty^2 = \gamma p_\infty/\rho_\infty$, and $\bar{l}$ denotes a reference length.

In order to derive the continuous adjoint problem, we multiply the left-hand side of (5.1) by $\mathbf{z}$, integrate by parts, and linearize about $\mathbf{u}$ to obtain

$$(\nabla\cdot(\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}](\mathbf{w}))\,,\mathbf{z})_\Omega = -\,(\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}](\mathbf{w}), \nabla\mathbf{z})_\Omega + (\mathbf{n}\cdot\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}](\mathbf{w}), \mathbf{z})_\Gamma\,, \tag{5.5}$$

where $\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}] := (\mathcal{F}^c)'[\mathbf{u}]$ denotes the Fréchet derivative of $\mathcal{F}^c$ with respect to $\mathbf{u}$. Here we already use the subscript $_{\mathbf{u}}$ notation which we require in section 6 to distinguish from subscript $_{\nabla\mathbf{u}}$ denoting the derivative with respect to $\nabla\mathbf{u}$. Thereby, the variational formulation of the continuous adjoint problem is given by: Find $\mathbf{z}$ such that

$$-\left(\mathbf{w}, (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}])^\top\nabla\mathbf{z}\right)_\Omega + \left(\mathbf{w}, (\mathbf{n}\cdot\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}])^\top\mathbf{z}\right)_\Gamma = J'[\mathbf{u}](\mathbf{w}) \quad \forall\mathbf{w}\in V, \tag{5.6}$$

and the continuous adjoint problem is given by

$$-(\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}])^\top\nabla\mathbf{z} = 0 \quad \text{in } \Omega, \qquad\qquad (\mathbf{n}\cdot\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}])^\top\mathbf{z} = j'[\mathbf{u}] \quad \text{on } \Gamma. \tag{5.7}$$

Using $\mathcal{F}^c(\mathbf{u})\cdot\mathbf{n} = p(0, n_1, n_2, 0)^\top$ on Γ_W and the definition of j in (5.4), we obtain

$$p'[\mathbf{u}](0, n_1, n_2, 0)\cdot\mathbf{z} = \frac{1}{C_\infty}p'[\mathbf{u}]\mathbf{n}\cdot\boldsymbol{\psi} \quad \text{on } \Gamma_W,$$

which reduces to the boundary condition of the adjoint compressible Euler equations

$$(B'[\mathbf{u}])^*\mathbf{z} = n_1 z_2 + n_2 z_3 = \frac{1}{C_\infty}\mathbf{n}\cdot\boldsymbol{\psi} \quad \text{on } \Gamma_W. \tag{5.8}$$

Comparing (5.5) with (2.12) we see that the target functional (5.4) is compatible. Furthermore, we note that the adjoint boundary condition (5.8) has first been derived in [20]. In the framework of matrix representations of adjoint boundary operators related to (2.3), they have been derived in [10]; see also [9] for a more detailed derivation.

Let $\mathbf{V}_h^p$ be a discrete function space consisting of discontinuous piecewise vector-valued polynomial functions of degree $p \geq 0$. Then the discontinuous Galerkin discretization in element-based form of (5.1) is given by: Find $\mathbf{u}_h \in \mathbf{V}_h^p$ such that

$$\begin{aligned}\mathcal{N}(\mathbf{u}_h, \mathbf{v}) \equiv -\int_\Omega \mathcal{F}^c(\mathbf{u}_h) : \nabla_h\mathbf{v}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\,\mathbf{v}^+\,\mathrm{d}s \\ + \int_\Gamma \tilde{\mathcal{H}}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+)\,\mathbf{v}^+\,\mathrm{d}s = 0 \quad \forall\mathbf{v}\in\mathbf{V}_h^p,\end{aligned} \tag{5.9}$$

where $\mathcal{H}$ and $\tilde{\mathcal{H}}$ may be any Lipschitz continuous, consistent, and conservative numerical flux functions approximating the normal flux $\mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h)$. On interior faces, $\mathcal{H}$ takes into account the possible discontinuities of $\mathbf{u}_h$ at element interfaces. On the boundary Γ, $\tilde{\mathcal{H}}$ may depend on the interior trace $\mathbf{u}_h^+$ and a consistent boundary function $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$. We note that $\tilde{\mathcal{H}}$ may be different from $\mathcal{H}$. In fact, we will see below

that, depending on the specific choice of $\tilde{\mathcal{H}}$, the discontinuous Galerkin discretization (5.9) may be adjoint consistent or not.

Using integration by parts we obtain the residual form: Find $\mathbf{u}_h \in \mathbf{V}_h^p$ such that

$$\int_\Omega \mathbf{R}(\mathbf{u}_h)\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathbf{r}(\mathbf{u}_h)\cdot\mathbf{v}^+\,\mathrm{d}s + \int_\Gamma \mathbf{r}_\Gamma(\mathbf{u}_h)\cdot\mathbf{v}^+\,\mathrm{d}s = 0 \quad \forall\mathbf{v}\in\mathbf{V}_h^p, \tag{5.10}$$

where the primal residuals are given by

$$\begin{aligned} \mathbf{R}(\mathbf{u}_h) &= -\nabla\cdot\mathcal{F}^c(\mathbf{u}_h) && \text{in } \kappa,\kappa\in\mathcal{T}_h,\\ \mathbf{r}(\mathbf{u}_h) &= \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{H}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+) && \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h,\\ \mathbf{r}_\Gamma(\mathbf{u}_h) &= \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+) - \tilde{\mathcal{H}}(\mathbf{u}_h^+,\mathbf{u}_\Gamma(\mathbf{u}_h^+),\mathbf{n}^+) && \text{on } \Gamma. \end{aligned} \tag{5.11}$$

Given the consistency of the numerical flux $\mathcal{H}(\mathbf{w},\mathbf{w},\mathbf{n}) = \mathbf{n}\cdot\mathcal{F}^c(\mathbf{w})$ and the consistency of the boundary function, i.e., $\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ for the exact solution $\mathbf{u}$ to (5.1), we find that $\mathbf{u}$ satisfies the following equations:

$$\mathbf{R}(\mathbf{u}) = 0 \ \text{ in } \kappa,\kappa\in\mathcal{T}_h, \quad \mathbf{r}(\mathbf{u}) = 0 \ \text{ on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h, \quad \mathbf{r}_\Gamma(\mathbf{u}) = 0 \ \text{ on } \Gamma. \tag{5.12}$$

We conclude that (5.9) is a consistent discretization of (5.1).

For the target functional $J(\cdot)$ defined in (5.4) with Fréchet derivative $J'[\mathbf{u}](\cdot)$, the discrete adjoint problem is given by: Find $\mathbf{z}_h \in \mathbf{V}_h^p$ such that

$$\mathcal{N}'[\mathbf{u}_h](\mathbf{w},\mathbf{z}_h) = J'[\mathbf{u}_h](\mathbf{w}) \quad \forall\mathbf{w}\in\mathbf{V}_h^p, \tag{5.13}$$

where

$$\begin{aligned} \mathcal{N}'[\mathbf{u}_h](\mathbf{w},\mathbf{z}_h) \equiv& -\int_\Omega (\mathcal{F}^c_{\mathbf{u}}[\mathbf{u}_h]\mathbf{w}) : \nabla_h\mathbf{z}_h\,\mathrm{d}\mathbf{x}\\ &+ \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+)\mathbf{w}^+ + \mathcal{H}'_{\mathbf{u}^-}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+)\mathbf{w}^-\right)\mathbf{z}_h^+\,\mathrm{d}s\\ &+ \int_\Gamma\left(\tilde{\mathcal{H}}'_{\mathbf{u}^+}\left(\mathbf{u}_h^+,\mathbf{u}_\Gamma(\mathbf{u}_h^+),\mathbf{n}^+\right) + \tilde{\mathcal{H}}'_{\mathbf{u}^-}\left(\mathbf{u}_h^+,\mathbf{u}_\Gamma(\mathbf{u}_h^+),\mathbf{n}^+\right)\mathbf{u}'_\Gamma[\mathbf{u}_h^+]\right)\mathbf{w}^+\mathbf{z}_h^+\,\mathrm{d}s. \end{aligned} \tag{5.14}$$

Here $\mathbf{v}\to\mathcal{H}'_{\mathbf{u}^+}(\mathbf{v}^+,\mathbf{v}^-,\mathbf{n})$ and $\mathbf{v}\to\mathcal{H}'_{\mathbf{u}^-}(\mathbf{v}^+,\mathbf{v}^-,\mathbf{n})$ denote the derivatives of the flux function $\mathcal{H}(\cdot,\cdot,\cdot)$ with respect to its first and second arguments, respectively. As the numerical flux is conservative $\mathcal{H}(\mathbf{v},\mathbf{w},\mathbf{n}) = -\mathcal{H}(\mathbf{w},\mathbf{v},-\mathbf{n})$, we obtain $\mathcal{H}'_{\mathbf{u}^-}(\mathbf{v},\mathbf{w},\mathbf{n}) = \partial_{\mathbf{w}}\mathcal{H}(\mathbf{v},\mathbf{w},\mathbf{n}) = -\partial_{\mathbf{w}}\mathcal{H}(\mathbf{w},\mathbf{v},-\mathbf{n}) = -\mathcal{H}'_{\mathbf{u}^+}(\mathbf{w},\mathbf{v},-\mathbf{n})$, and

$$\begin{aligned} \int_{\Gamma_{\mathcal{I}}}\mathcal{H}'_{\mathbf{u}^-}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+)\mathbf{w}^-\mathbf{z}^+\,\mathrm{d}s &= -\int_{\Gamma_{\mathcal{I}}}\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^-,\mathbf{u}_h^+,\mathbf{n}^-)\mathbf{w}^-\mathbf{z}^+\,\mathrm{d}s\\ &= -\int_{\Gamma_{\mathcal{I}}}\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+)\mathbf{w}^+\mathbf{z}^-\,\mathrm{d}s, \end{aligned} \tag{5.15}$$

where we exchanged notations $^+$ and $^-$ on $\Gamma_{\mathcal{I}}$. Then the discrete adjoint problem (5.13) with (5.14) is given in adjoint residual form as follows: Find $\mathbf{z}_h \in \mathbf{V}_h^p$ such that

$$\int_\Omega \mathbf{w}\cdot\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathbf{w}^+\cdot\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}s + \int_\Gamma \mathbf{w}^+\cdot\mathbf{r}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}s = 0 \tag{5.16}$$

for all $\mathbf{w} \in \mathbf{V}_h^p$, where the adjoint residuals are given by

$$\begin{aligned}
\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) =& (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_h])^\top \nabla \mathbf{z}_h && \text{in } \kappa, \kappa \in \mathcal{T}_h, \\
\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) =& -\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\right)^\top [\![\mathbf{z}_h]\!] \cdot \mathbf{n} && \text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\
\mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h) =& j'[\mathbf{u}_h] - \left(\tilde{\mathcal{H}}'_{\mathbf{u}^+} + \tilde{\mathcal{H}}'_{\mathbf{u}^-} \mathbf{u}'_\Gamma[\mathbf{u}_h]\right)^\top \mathbf{z}_h^+ && \text{on } \Gamma,
\end{aligned} \tag{5.17}$$

where $\tilde{\mathcal{H}}'_{\mathbf{u}^+} := \tilde{\mathcal{H}}'_{\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+)$ and $\tilde{\mathcal{H}}'_{\mathbf{u}^-} := \tilde{\mathcal{H}}'_{\mathbf{u}^-}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}^+)$.

Comparing the discrete adjoint boundary condition

$$\left(\tilde{\mathcal{H}}'_{\mathbf{u}^+} + \tilde{\mathcal{H}}'_{\mathbf{u}^-} \mathbf{u}'_\Gamma[\mathbf{u}_h]\right)^\top \mathbf{z}_h^+ = j'[\mathbf{u}_h] \quad \text{on } \Gamma \tag{5.18}$$

and the continuous adjoint boundary condition in (5.7), we notice that not all choices of $\tilde{\mathcal{H}}$ give rise to an adjoint consistent discretization. In fact, we require $\tilde{\mathcal{H}}$ to have the following properties: In order to incorporate boundary conditions in the primal discretization (5.9), $\tilde{\mathcal{H}}$ must depend on $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$; hence, $\tilde{\mathcal{H}}'_{\mathbf{u}^-} \neq 0$. Furthermore, we require $\tilde{\mathcal{H}}'_{\mathbf{u}^+} = 0$ as otherwise the left-hand side in (5.17) involves two summands which is in contrast to the continuous adjoint boundary condition in (5.7). Finally, we recall that $\tilde{\mathcal{H}}$ is consistent $\tilde{\mathcal{H}}(\mathbf{v}, \mathbf{v}, \mathbf{n}) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{v})$ and conclude that $\tilde{\mathcal{H}}$ is given by $\tilde{\mathcal{H}}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+))$. Employing a modified target functional $\tilde{J}(\mathbf{u}_h) = J(\mathbf{i}(\mathbf{u}_h))$, i.e., (2.25) with $r_j(\mathbf{u}_h) \equiv 0$, (5.18) yields

$$\left(\mathbf{n} \cdot \left(\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_\Gamma(\mathbf{u}_h^+)]\right) \mathbf{u}'_\Gamma[\mathbf{u}_h^+]\right)^\top \mathbf{z} = j'[\mathbf{i}(\mathbf{u}_h^+)]\mathbf{i}'[\mathbf{u}_h^+]. \tag{5.19}$$

We find the modification $\mathbf{i}(\mathbf{u}_h) = \mathbf{u}_\Gamma(\mathbf{u}_h)$, which is consistent as $\mathbf{i}(\mathbf{u}) = \mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ holds for the exact solution $\mathbf{u}$. Thereby (5.19) reduces to

$$\left(\mathbf{n} \cdot \mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_\Gamma(\mathbf{u}_h^+)]\right)^\top \mathbf{z} = j'[\mathbf{u}_\Gamma(\mathbf{u}_h^+)], \tag{5.20}$$

which represents a discretization of the continuous adjoint boundary condition in (5.7). In order to obtain a discretization of the adjoint boundary condition at solid wall boundaries (5.8), we require $B\mathbf{u}_\Gamma(\mathbf{u}_h^+) = 0$ on Γ_W. This condition is satisfied by

$$\mathbf{u}_\Gamma(\mathbf{u}) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-n_1^2 & -n_1 n_2 & 0 \\ 0 & -n_1 n_2 & 1-n_2^2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \mathbf{u} \quad \text{on} \quad \Gamma_W, \tag{5.21}$$

which originates from $\mathbf{u}$ by subtracting the normal velocity component of $\mathbf{u}$; i.e., $\mathbf{v} = (v_1, v_2)$ is replaced by $\mathbf{v}_\Gamma = \mathbf{v} - (\mathbf{v} \cdot \mathbf{n})\mathbf{n}$ which ensures that the normal velocity component vanishes: $\mathbf{v}_\Gamma \cdot \mathbf{n} = 0$.

In summary, let $\mathbf{u}_\Gamma$ be given by (5.21) and $\tilde{\mathcal{H}}$ and $\tilde{J}$ be defined by

$$\tilde{\mathcal{H}}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}) = \mathbf{n} \cdot \mathcal{F}_\Gamma^c(\mathbf{u}_h^+), \qquad \tilde{J}(\mathbf{u}_h) = J_\Gamma(\mathbf{u}_h), \tag{5.22}$$

respectively, where $\mathcal{F}_\Gamma^c(\mathbf{u}_h^+) := \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+))$, $J_\Gamma(\mathbf{u}_h) := J(\mathbf{u}_\Gamma(\mathbf{u}_h))$, and $j_\Gamma(\mathbf{u}_h) := j(\mathbf{u}_\Gamma(\mathbf{u}_h))$; then the adjoint residuals (5.17) are given by

$$\begin{aligned}
\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) =& (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}_h])^\top \nabla \mathbf{z}_h && \text{in } \kappa, \kappa \in \mathcal{T}_h, \\
\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) =& -\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\right)^\top [\![\mathbf{z}_h]\!] \cdot \mathbf{n} && \text{on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \\
\mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h) =& j'_\Gamma[\mathbf{u}_h^+] - \left(\mathbf{n} \cdot \mathcal{F}_{\Gamma,\mathbf{u}}^c[\mathbf{u}_h^+]\right)^\top \mathbf{z}_h^+ && \text{on } \Gamma.
\end{aligned} \tag{5.23}$$

In particular, the discretization (5.9) together with (5.22) is adjoint consistent as the exact solutions u and $\mathbf{z}$ to (5.1) and (5.7), respectively, satisfy

$$\mathbf{R}^*[\mathbf{u}](\mathbf{z}) = 0 \ \text{ in } \kappa, \kappa \in \mathcal{T}_h, \quad \mathbf{r}^*[\mathbf{u}](\mathbf{z}) = 0 \ \text{ on } \partial\kappa \setminus \Gamma, \kappa \in \mathcal{T}_h, \quad \mathbf{r}^*_\Gamma[\mathbf{u}](\mathbf{z}) = 0 \ \text{ on } \Gamma.$$

Note that the adjoint residuals in (5.23) reduce to the adjoint residuals (3.5) of the linear advection equation with $b = 0$, when setting $\mathcal{F}^c(u) = \mathbf{b}u$ and $\mathcal{H}'_{u^+} = \mathbf{b} \cdot \mathbf{n}$.

Also note that the standard discontinuous Galerkin discretizations for the compressible Euler equations (see, e.g., [4, 15, 16] among several others) take the same numerical flux function on the boundary Γ as in the interior of the domain, and simply replace $\mathbf{u}_h^-$ in $\mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n})$ by the boundary function $\mathbf{u}_\Gamma(\mathbf{u}_h^+)$, resulting in $\tilde{\mathcal{H}}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n})$. Furthermore, the definition of $\mathbf{u}_\Gamma$ in [4, 15] based on $\mathbf{v}_\Gamma = \mathbf{v} - 2(\mathbf{v} \cdot \mathbf{n})\mathbf{n}$ ensures a vanishing average normal velocity $\bar{\mathbf{v}} \cdot \mathbf{n} = \frac{1}{2}(\mathbf{v} + \mathbf{v}_\Gamma) \cdot \mathbf{n} = 0$. However, $\mathbf{v}_\Gamma \cdot \mathbf{n} = 0$ and $B\mathbf{u}_\Gamma(\mathbf{u}_h^+) = 0$, as required in (5.20), are not satisfied. Thereby, the discontinuous Galerkin discretization based on the standard choice of $\tilde{\mathcal{H}}$ and $\mathbf{u}_\Gamma$ is not adjoint consistent. In fact, already the numerical experiments in [15] indicated large gradients, i.e., an irregular adjoint solution near solid wall boundaries. The lack of adjoint consistency of this standard approach was first analyzed by Lu [23, 24], who also proposed the adjoint consistent approach (5.22) and demonstrated that this approach gives rise to smooth adjoint solutions for an inviscid compressible flow over a Gaussian bump. The smoothness of the discrete adjoint has been confirmed in [14] for an inviscid compressible flow around a NACA0012 airfoil. Furthermore, [14] studies the effect of adjoint consistency on the accuracy of the flow solution and on error cancellation in an a posteriori error estimation approach.

6. The compressible Navier–Stokes equations. In this section we consider the two-dimensional stationary compressible Navier–Stokes equations

$$\nabla \cdot (\mathcal{F}^c(\mathbf{u}) - \mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u})) = 0 \quad \text{in } \Omega, \tag{6.1}$$

subject to various boundary conditions, e.g., no-slip wall boundary conditions with vanishing velocity $\mathbf{v} = (v_1, v_2)^\top = 0$ at isothermal walls Γ_{iso} where $T = T_{wall}$ or at adiabatic walls Γ_{adia} where $\mathbf{n} \cdot \nabla T = 0$. The vector of conservative variables $\mathbf{u}$ and the convective fluxes $\mathcal{F}^c(\mathbf{u})$ are as defined in (5.3). Furthermore, the viscous fluxes $\mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = (\mathbf{f}_1^v(\mathbf{u}, \nabla\mathbf{u}), \mathbf{f}_2^v(\mathbf{u}, \nabla\mathbf{u}))$ are defined by

$$\mathbf{f}_1^v(\mathbf{u}, \nabla\mathbf{u}) = \begin{bmatrix} 0 \\ \tau_{11} \\ \tau_{21} \\ \tau_{1j}v_j + \mathcal{K}T_{x_1} \end{bmatrix} \quad \text{and} \quad \mathbf{f}_2^v(\mathbf{u}, \nabla\mathbf{u}) = \begin{bmatrix} 0 \\ \tau_{12} \\ \tau_{22} \\ \tau_{2j}v_j + \mathcal{K}T_{x_2} \end{bmatrix}. \tag{6.2}$$

Here T denotes the temperature given by $e = c_v T$, $\mathcal{K}$ is the thermal conductivity coefficient, and $\underline{\tau}$ is the viscous stress tensor defined by $\underline{\tau} = \mu(\nabla\mathbf{v} + (\nabla\mathbf{v})^\top - \frac{2}{3}(\nabla \cdot \mathbf{v})I)$, where μ is the dynamic viscosity coefficient. Using the homogeneity tensor G (e.g., [16]) with $G_{ij}(\mathbf{u}) = \partial\mathbf{f}_i^v(\mathbf{u}, \nabla\mathbf{u})/\partial u_{x_j}$, for $i, j = 1, 2$, the viscous fluxes are $\mathbf{f}_i^v(\mathbf{u}, \nabla\mathbf{u}) = G_{ij}(\mathbf{u})\partial\mathbf{u}/\partial x_j$, $i = 1, 2$, and $\mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = G(\mathbf{u})\nabla\mathbf{u}$. Consider the target functional

$$J(\mathbf{u}) = \int_\Gamma j(C\mathbf{u})\,\mathrm{d}s = \int_{\Gamma_W} (p\,\mathbf{n} - \underline{\tau}\,\mathbf{n}) \cdot \boldsymbol{\psi}_{\Gamma_W}\,\mathrm{d}s, \tag{6.3}$$

where $(C\mathbf{u})_{ij} = p(\mathbf{u})\delta_{ij} - \tau_{ij}(\mathbf{u}, \nabla\mathbf{u})$, $j(C\mathbf{u}) = (C\mathbf{u})_{ij} n_j (\psi_{\Gamma_W})_i$ on Γ_W, and $j(C\mathbf{u}) \equiv 0$ elsewhere, with $\boldsymbol{\psi}_{\Gamma_W} \in [L^2(\Gamma_W)]^2$. Important target quantities of type (6.3) in viscous compressible flows are the (total) drag and lift coefficients c_d and c_l, which include both pressure-induced and viscous forces. Then $\boldsymbol{\psi}_{\Gamma_W} = \frac{1}{C_\infty}\boldsymbol{\psi}$, and C_∞ and $\boldsymbol{\psi}$ are as in (5.4).

In order to derive the adjoint problem, we multiply the left-hand side of (6.1) by $\mathbf{z}$, integrate by parts, and linearize about $\mathbf{u}$ to obtain

$$\begin{aligned}&\left(\nabla\cdot\left(\mathcal{F}^c_{\mathbf{u}}\mathbf{w} - \mathcal{F}^v_{\mathbf{u}}\mathbf{w} - \mathcal{F}^v_{\nabla\mathbf{u}}\nabla\mathbf{w}\right), \mathbf{z}\right)_\Omega \\ &\qquad = -\left(\left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)\mathbf{w} - \mathcal{F}^v_{\nabla\mathbf{u}}\nabla\mathbf{w}, \nabla\mathbf{z}\right)_\Omega + \left(\mathbf{n}\cdot\left(\mathcal{F}^c_{\mathbf{u}}\mathbf{w} - \mathcal{F}^v_{\mathbf{u}}\mathbf{w} - \mathcal{F}^v_{\nabla\mathbf{u}}\nabla\mathbf{w}\right), \mathbf{z}\right)_\Gamma,\end{aligned}$$

where $\mathcal{F}^v_{\mathbf{u}} := \partial_{\mathbf{u}}\mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = G'[\mathbf{u}]\nabla\mathbf{u}$ and $\mathcal{F}^v_{\nabla\mathbf{u}} := \partial_{\nabla\mathbf{u}}\mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = G(\mathbf{u})$ denote the derivatives of $\mathcal{F}^v$ with respect to $\mathbf{u}$ and $\nabla\mathbf{u}$, respectively. Using integration by parts once more we obtain the following variational formulation of the continuous adjoint problem: Find $\mathbf{z}$ such that

$$\begin{aligned}&-\left(\mathbf{w}, \left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)^\top \nabla\mathbf{z}\right)_\Omega - \left(\mathbf{w}, \nabla\cdot\left(\left(\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top \nabla\mathbf{z}\right)\right)_\Omega + \left(\mathbf{w}, \mathbf{n}\cdot\left(\left(\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top \nabla\mathbf{z}\right)\right)_\Gamma \\ &\qquad + \left(\mathbf{w}, \left(\mathbf{n}\cdot\left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)\right)^\top \mathbf{z}\right)_\Gamma - \left(\nabla\mathbf{w}, \left(\mathbf{n}\cdot\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top \mathbf{z}\right)_\Gamma = J'[\mathbf{u}](\mathbf{w}) \quad \forall\mathbf{w}\in\mathbf{V}.\end{aligned}$$

Given that

$$\begin{aligned}J'[\mathbf{u}](\mathbf{w}) &= \frac{1}{C_\infty}\int_{\Gamma_W} \left(p_{\mathbf{u}}[\mathbf{u}]\,\mathbf{n} - \underline{\tau}_{\mathbf{u}}[\mathbf{u}]\,\mathbf{n}\right)\cdot\boldsymbol{\psi}\,\mathbf{w} - \left(\underline{\tau}_{\nabla\mathbf{u}}[\mathbf{u}]\,\mathbf{n}\right)\cdot\boldsymbol{\psi}\,\nabla\mathbf{w}\,\mathrm{d}s \\ &= \left(\mathbf{w}, \frac{1}{C_\infty}\left(p_{\mathbf{u}}\,\mathbf{n} - \underline{\tau}_{\mathbf{u}}\,\mathbf{n}\right)\cdot\boldsymbol{\psi}\right)_{\Gamma_W} - \left(\nabla\mathbf{w}, \frac{1}{C_\infty}\left(\underline{\tau}_{\nabla\mathbf{u}}\,\mathbf{n}\right)\cdot\boldsymbol{\psi}\right)_{\Gamma_W},\end{aligned} \tag{6.4}$$

we see that the adjoint solution $\mathbf{z}$ satisfies the following equation:

$$-\left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)^\top \nabla\mathbf{z} - \nabla\cdot\left(\left(\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top \nabla\mathbf{z}\right) = 0, \tag{6.5}$$

subject to the boundary conditions on $\Gamma_W = \Gamma_{iso}\cup\Gamma_{adia}$,

$$\left(\mathbf{n}\cdot\left(\mathcal{F}^c_{\mathbf{u}} - \mathcal{F}^v_{\mathbf{u}}\right)\right)^\top \mathbf{z} + \mathbf{n}\cdot\left(\left(\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top \nabla\mathbf{z}\right) = \frac{1}{C_\infty}\left(p_{\mathbf{u}}\,\mathbf{n} - \underline{\tau}_{\mathbf{u}}\,\mathbf{n}\right)\cdot\boldsymbol{\psi}, \tag{6.6}$$

$$\left(\mathbf{n}\cdot\mathcal{F}^v_{\nabla\mathbf{u}}\right)^\top \mathbf{z} = \frac{1}{C_\infty}\left(\underline{\tau}_{\nabla\mathbf{u}}\,\mathbf{n}\right)\cdot\boldsymbol{\psi}. \tag{6.7}$$

At wall boundaries Γ_W where $\mathbf{v} = (v_1, v_2)^\top = 0$, the normal viscous flux reduces to $\mathbf{n}\cdot\mathcal{F}^v(\mathbf{u}, \nabla\mathbf{u}) = (0, (\tau\mathbf{n})_1, (\tau\mathbf{n})_2, \mathbf{n}\cdot\nabla T)^\top$. Hence (6.7) is fulfilled, provided $\mathbf{z}$ satisfies

$$\begin{pmatrix} 0 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_1\, z_2 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_2\, z_3 \\ (\mathbf{n}\cdot\nabla T_{\nabla\mathbf{u}})\, z_4 \end{pmatrix} = \frac{1}{C_\infty}\begin{pmatrix} 0 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_1\,\psi_1 \\ (\tau_{\nabla\mathbf{u}}\mathbf{n})_2\,\psi_2 \\ 0 \end{pmatrix}, \tag{6.8}$$

which reduces to the conditions $z_2 = \frac{1}{C_\infty}\psi_1$ on Γ_W, $z_3 = \frac{1}{C_\infty}\psi_2$ on Γ_W, and $z_4 = 0$ on Γ_{iso}. At adiabatic boundaries we have $\mathbf{n}\cdot\nabla T = 0$, and the last condition in (6.8) vanishes. Substituted into (6.6) we obtain $\mathbf{n}\cdot((\mathcal{F}^v_{\nabla\mathbf{u}})^\top\nabla\mathbf{z}) = 0$ on Γ_W, which at adiabatic boundaries reduces to $\mathbf{n}\cdot\nabla z_4 = 0$. On isothermal boundaries no additional

boundary condition is obtained. In summary, the boundary conditions of the adjoint problem (6.5) to the compressible Navier–Stokes equations are given by

$$z_2 = \frac{1}{C_\infty}\psi_1,\ z_3 = \frac{1}{C_\infty}\psi_2 \ \text{ on } \Gamma_W, \quad z_4 = 0 \ \text{ on } \Gamma_{iso}, \quad \mathbf{n}\cdot\nabla z_4 = 0 \ \text{ on } \Gamma_{adia}. \tag{6.9}$$

These boundary conditions have been derived by computing the adjoint equations of each of the four primal equations separately in [21] and [8].

In addition to the notation introduced in section 3, we use the standard notation $\underline{\sigma} : \underline{\tau} = \sum_{k=1}^{m}\sum_{l=1}^{n}\sigma_{kl}\tau_{kl}$ for matrices $\underline{\sigma}, \underline{\tau} \in \mathbb{R}^{m\times n}$, $m, n \geq 1$; additionally, for vectors $\mathbf{v} \in \mathbb{R}^m, \mathbf{w} \in \mathbb{R}^n$, the matrix $\mathbf{v}\otimes\mathbf{w} \in \mathbb{R}^{m\times n}$ is defined by $(\mathbf{v}\otimes\mathbf{w})_{kl} = v_k\, w_l$.

According to [16, 17] the interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations (6.1) is given by: Find $\mathbf{u}_h \in \mathbf{V}_h^p$ such that

$$\begin{aligned}\mathcal{N}(\mathbf{u}_h, \mathbf{v}) \equiv &- \int_\Omega \mathcal{F}^c(\mathbf{u}_h) : \nabla_h \mathbf{v}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_h^-, \mathbf{n}^+)\cdot\mathbf{v}^+\,\mathrm{d}s \\ &+ \int_\Omega \mathcal{F}^v(\mathbf{u}_h, \nabla_h\mathbf{u}_h) : \nabla_h\mathbf{v}\,\mathrm{d}\mathbf{x} - \int_{\Gamma_\mathcal{I}} \{\mathcal{F}^v(\mathbf{u}_h, \nabla_h\mathbf{u}_h)\} : \underline{[\![\mathbf{v}]\!]}\,\mathrm{d}s \\ &+ \int_{\Gamma_\mathcal{I}} \theta\{G^\top(\mathbf{u}_h)\nabla_h\mathbf{v}\} : \underline{[\![\mathbf{u}_h]\!]}\,\mathrm{d}s + \int_{\Gamma_\mathcal{I}} \delta\underline{[\![\mathbf{u}_h]\!]} : \underline{[\![\mathbf{v}]\!]}\,\mathrm{d}s + \mathcal{N}_\Gamma(\mathbf{u}_h, \mathbf{v}) = 0\end{aligned} \tag{6.10}$$

for all $\mathbf{v}$ in $\mathbf{V}_h^p$. Here $\mathcal{N}_\Gamma(\mathbf{u}_h, \mathbf{v})$ includes all boundary terms which will be specified in the following. Recalling the discussion at the end of section 5, we know that the discretization of boundary terms in [16] is not adjoint consistent. In fact, [16] uses the standard discretization of convective boundary fluxes as opposed to the adjoint consistent discretization given in (5.22). Thereby, in the following we consider the boundary terms like in [16] but with an adjoint consistent treatment of convective fluxes like in (5.22). Furthermore, we treat the viscous fluxes analogous to the convective fluxes; i.e., we replace the viscous boundary flux $\mathcal{F}^v$ by $\mathcal{F}_\Gamma^v$, where

$$\mathcal{F}_\Gamma^v(\mathbf{u}_h, \nabla\mathbf{u}_h) = \mathcal{F}^v(\mathbf{u}_\Gamma(\mathbf{u}_h), \nabla\mathbf{u}_h) = G_\Gamma(\mathbf{u}_h)\nabla\mathbf{u}_h = G(\mathbf{u}_\Gamma(\mathbf{u}_h))\nabla\mathbf{u}_h. \tag{6.11}$$

Thereby, the discretization of boundary terms is given by

$$\begin{aligned}\mathcal{N}_\Gamma(\mathbf{u}_h, \mathbf{v}) = &\int_\Gamma \mathbf{n}\cdot\mathcal{F}_\Gamma^c(\mathbf{u}_h^+)\mathbf{v}^+\,\mathrm{d}s + \int_\Gamma \delta\left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\cdot\mathbf{v}^+\,\mathrm{d}s \\ &- \int_\Gamma \mathbf{n}\cdot\mathcal{F}_\Gamma^v(\mathbf{u}_h^+, \nabla\mathbf{u}_h^+)\,\mathbf{v}^+\,\mathrm{d}s \\ &+ \theta\int_\Gamma \left(G_\Gamma^\top(\mathbf{u}_h^+)\nabla\mathbf{v}_h^+\right) : \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\otimes\mathbf{n}\,\mathrm{d}s,\end{aligned} \tag{6.12}$$

where on adiabatic boundaries $\Gamma_{adia} \subset \Gamma_W$ the viscous flux $\mathcal{F}_\Gamma^v$ and the corresponding homogeneity tensor G_Γ are modified such that $\mathbf{n}\cdot\nabla T = 0$. Using integration by parts in (6.10), we obtain the primal residual form as follows: Find $\mathbf{u}_h \in \mathbf{V}_h^p$ such that

$$\begin{aligned}\int_\Omega \mathbf{R}(\mathbf{u}_h)\cdot\mathbf{v}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathbf{r}(\mathbf{u}_h)\cdot\mathbf{v}^+ + \boldsymbol{\rho}(\mathbf{u}_h) : \nabla\mathbf{v}^+\,\mathrm{d}s \\ + \int_\Gamma \mathbf{r}_\Gamma(\mathbf{u}_h)\cdot\mathbf{v}^+ + \boldsymbol{\rho}_\Gamma(\mathbf{u}_h) : \nabla\mathbf{v}^+\,\mathrm{d}s = 0 \quad \forall\mathbf{v}\in\mathbf{V}_h^p,\end{aligned} \tag{6.13}$$

where the primal residuals are given by

$$
\begin{aligned}
\mathbf{R}(\mathbf{u}_h) &= -\nabla\cdot\mathcal{F}^c(\mathbf{u}_h) + \nabla\cdot\mathcal{F}^v(\mathbf{u}_h,\nabla_h\mathbf{u}_h) && \text{in } \kappa,\kappa\in\mathcal{T}_h,\\
\mathbf{r}(\mathbf{u}_h) &= \mathbf{n}\cdot\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{H}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+) - \frac{1}{2}[\![\mathcal{F}^v(\mathbf{u}_h,\nabla_h\mathbf{u}_h)]\!] - \delta[\![\mathbf{u}_h]\!]\cdot\mathbf{n},\\
\boldsymbol{\rho}(\mathbf{u}_h) &= -\frac{\theta}{2}\left(G(\mathbf{u}_h)\underline{[\![\mathbf{u}_h]\!]}\right)^\top && \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h,\\
\mathbf{r}_\Gamma(\mathbf{u}_h) &= \mathbf{n}\cdot(\mathcal{F}^c(\mathbf{u}_h^+) - \mathcal{F}_\Gamma^c(\mathbf{u}_h^+) - \mathcal{F}^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+) + \mathcal{F}_\Gamma^v(\mathbf{u}_h^+,\nabla\mathbf{u}_h^+)) - \delta(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)),\\
\boldsymbol{\rho}_\Gamma(\mathbf{u}_h) &= -\theta\left(G_\Gamma^\top(\mathbf{u}_h^+) : \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\otimes\mathbf{n}\right)^\top && \text{on } \Gamma.
\end{aligned}
$$

We see that the exact solution $\mathbf{u}$ to (6.1) satisfies

$$
\mathbf{R}(\mathbf{u}) = 0, \qquad \mathbf{r}(\mathbf{u}) = 0, \qquad \boldsymbol{\rho}(\mathbf{u}) = 0, \qquad \mathbf{r}_\Gamma(\mathbf{u}) = 0, \qquad \boldsymbol{\rho}_\Gamma(\mathbf{u}) = 0,
$$

where we used consistency of the numerical flux $\mathcal{H}(\mathbf{w},\mathbf{w},\mathbf{n}) = \mathbf{n}\cdot\mathcal{F}^c(\mathbf{w})$, continuity of $\mathbf{u}$, and the consistency of the boundary function, i.e., $\mathbf{u}$ satisfies $\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ on Γ. We conclude that the discretization given in (6.10) and (6.12) is consistent.

Given the target quantity $J(\cdot)$ defined in (6.3) with Fréchet derivative (6.4), we consider the following modification of $J(\cdot)$:

$$
\tilde{J}(\mathbf{u}_h) = J(\mathbf{i}(\mathbf{u}_h)) + \int_\Gamma r_J(\mathbf{u}_h)\,\mathrm{d}s = J_\Gamma(\mathbf{u}_h) + \int_\Gamma r_J(\mathbf{u}_h)\,\mathrm{d}s. \tag{6.14}
$$

As in section 5, here we set $\mathbf{i}(\mathbf{u}_h) = \mathbf{u}_\Gamma(\mathbf{u}_h)$ and $J_\Gamma(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h))$; $r_J(\mathbf{u}_h)$ will be specified later. Noting that $\mathbf{u}_\Gamma(\mathbf{u}) = \mathbf{u}$ holds for the exact solution $\mathbf{u}$, $\tilde{J}(\cdot)$ in (6.14) is a consistent modification of $J(\cdot)$, provided that $\mathbf{u}$ satisfies $r_J(\mathbf{u}) = 0$; see also (2.25).

Rewriting $\mathcal{N}(\mathbf{u}_h,\mathbf{v})$ in (6.10) in terms of the homogeneity tensor G and recalling (5.15), we see that the discrete adjoint problem is given by: Find $\mathbf{z}_h\in\mathbf{V}_h$ such that

$$
\mathcal{N}'[\mathbf{u}_h](\mathbf{w},\mathbf{z}_h) = \tilde{J}'[\mathbf{u}_h](\mathbf{w}) \quad \forall\mathbf{w}\in\mathbf{V}, \tag{6.15}
$$

where $\mathcal{N}'[\mathbf{u}](\mathbf{w},\mathbf{z})$ is given by

$$
\begin{aligned}
\mathcal{N}'[\mathbf{u}](\mathbf{w},\mathbf{z}) = &-\int_\Omega (\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}]\mathbf{w}) : \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma}\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n}^+)\mathbf{w}^+[\![\mathbf{z}]\!]\cdot\mathbf{n}\,\mathrm{d}\mathbf{s}\\
&+\int_\Omega (G'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{u}) : \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x} + \int_\Omega (G(\mathbf{u})\nabla_h\mathbf{w}) : \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x}\\
&-\int_{\Gamma_{\mathcal{I}}}\{G'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{u}\} : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s - \int_{\Gamma_{\mathcal{I}}}\{G(\mathbf{u})\nabla_h\mathbf{w}\} : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s\\
&+\int_{\Gamma_{\mathcal{I}}}\theta\{\left(G^\top\right)'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{z}\} : \underline{[\![\mathbf{u}]\!]}\,\mathrm{d}s + \int_{\Gamma_{\mathcal{I}}}\theta\{G^\top(\mathbf{u})\nabla_h\mathbf{z}\} : \underline{[\![\mathbf{w}]\!]}\,\mathrm{d}s\\
&+\int_{\Gamma_{\mathcal{I}}}\delta\underline{[\![\mathbf{w}]\!]} : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s + \mathcal{N}'_\Gamma[\mathbf{u}](\mathbf{w},\mathbf{z}).
\end{aligned} \tag{6.16}
$$

Using integration by parts, this can be rewritten as follows:

$$
\begin{aligned}
&- \int_\Omega \mathbf{w} \left(\mathcal{F}_{\mathbf{u}}^c[\mathbf{u}]\right)^\top \nabla_h \mathbf{z} \,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h} \int_{\partial\kappa\setminus\Gamma} \mathbf{w}^+ \left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}^+,\mathbf{u}^-,\mathbf{n}^+)\right)^\top [\![\mathbf{z}]\!]\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} \\
&+ \int_\Omega \mathbf{w} \left(G'[\mathbf{u}]\nabla_h\mathbf{u}\right)^\top \nabla_h\mathbf{z}\,\mathrm{d}\mathbf{x} - \int_\Omega \mathbf{w}\nabla_h\cdot\left(G^\top(\mathbf{u})\nabla_h\mathbf{z}\right)\,\mathrm{d}\mathbf{x} \\
&- \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} (G'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{u}) : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s - \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} (G(\mathbf{u})\nabla_h\mathbf{w}) : \underline{[\![\mathbf{z}]\!]}\,\mathrm{d}s \\
&+ \frac{1}{2}\theta\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \left(\left(G^\top\right)'[\mathbf{u}]\mathbf{w}\nabla_h\mathbf{z}\right) : \underline{[\![\mathbf{u}]\!]}\,\mathrm{d}s + \frac{1}{2}\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathbf{w}[\![G^\top(\mathbf{u})\nabla_h\mathbf{z}]\!]\,\mathrm{d}s \\
&+ (1+\theta)\sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} (\mathbf{w}\otimes\mathbf{n}) : \{G^\top(\mathbf{u})\nabla_h\mathbf{z}\}\,\mathrm{d}s + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \delta\mathbf{w}[\![\mathbf{z}]\!]\cdot\mathbf{n}\,\mathrm{d}\mathbf{s} \\
&+ \int_\Gamma (\mathbf{w}\otimes\mathbf{n}) : \left(G^\top(\mathbf{u})\nabla_h\mathbf{z}\right)\,\mathrm{d}s + \mathcal{N}'_\Gamma[\mathbf{u}](\mathbf{w},\mathbf{z}).
\end{aligned}
\tag{6.17}
$$

Hence the discrete adjoint problem (6.15) in adjoint residual form is given as follows: Find $\mathbf{z}_h \in \mathbf{V}_h$ such that

$$
\begin{aligned}
&\int_\Omega \mathbf{w}\cdot\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}\mathbf{x} + \sum_{\kappa\in\mathcal{T}_h}\int_{\partial\kappa\setminus\Gamma} \mathbf{w}\cdot\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) + \nabla\mathbf{w} : \boldsymbol{\rho}^*[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}s \\
&\qquad + \int_\Gamma \mathbf{w}\cdot\mathbf{r}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h) + \nabla\mathbf{w} : \boldsymbol{\rho}^*_\Gamma[\mathbf{u}_h](\mathbf{z}_h)\,\mathrm{d}s = 0 \quad \forall \mathbf{w}\in\mathbf{V}_h,
\end{aligned}
\tag{6.18}
$$

where the adjoint residuals are given by

$$
\begin{aligned}
\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h) &= \left(\mathcal{F}_{\mathbf{u}}^c(\mathbf{u}_h) - G'[\mathbf{u}_h]\nabla\mathbf{u}_h\right)^\top \nabla_h\mathbf{z}_h + \nabla_h\cdot\left(G^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right) \quad \text{in } \kappa,\kappa\in\mathcal{T}_h, \\
\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h) &= -\left(\mathcal{H}'_{\mathbf{u}^+}(\mathbf{u}_h^+,\mathbf{u}_h^-,\mathbf{n}^+)\right)^\top [\![\mathbf{z}_h]\!]\cdot\mathbf{n} \\
&\quad - \frac{1}{2}[\![G^\top(\mathbf{u}_h)\nabla\mathbf{z}_h]\!] - (1+\theta)\mathbf{n}\cdot\{G^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\} - \delta[\![\mathbf{z}_h]\!]\cdot\mathbf{n} \\
&\quad + \frac{1}{2}\left(G'[\mathbf{u}_h]\nabla\mathbf{u}_h\right)^\top \underline{[\![\mathbf{z}_h]\!]} - \frac{1}{2}\theta\left(G'[\mathbf{u}_h]\underline{[\![\mathbf{u}_h]\!]}\right)^\top \nabla_h\mathbf{z}_h \quad \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h, \\
\boldsymbol{\rho}^*[\mathbf{u}_h](\mathbf{z}_h) &= \frac{1}{2}G^\top[\mathbf{u}_h]\underline{[\![\mathbf{z}_h]\!]} \qquad \text{on } \partial\kappa\setminus\Gamma,\kappa\in\mathcal{T}_h.
\end{aligned}
\tag{6.19}
$$

The adjoint boundary residuals $\mathbf{r}^*_\Gamma$ and $\boldsymbol{\rho}^*_\Gamma$ will be specified below. Recalling that $\mathcal{F}^v_{\mathbf{u}} = G'[\mathbf{u}]\nabla\mathbf{u}$ and $\mathcal{F}^v_{\nabla\mathbf{u}} = G(\mathbf{u})$, we see that the exact solution $\mathbf{z}$ to the continuous adjoint problem (6.5) satisfies $\mathbf{R}^*[\mathbf{u}](\mathbf{z}) = 0$. In the three lines in (6.19) representing the face residual term $\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h)$, we recognize the jump $-(\mathcal{H}'_{\mathbf{u}^+})^\top[\![\mathbf{z}_h]\!]\cdot\mathbf{n}$ due to the convective part of the equations (cf. (5.23)); furthermore, the terms in the second line correspond to the adjoint face residuals of Poisson's equation (cf. (4.11)); and finally the two terms in the third line are due to the nonlinearity of the compressible Navier–Stokes equations. Whereas the last term in the third line vanishes for a smooth exact primal solution $\mathbf{u}$, all other terms vanish for the exact solution $\mathbf{z}$ to the adjoint problem (6.5), provided $\theta = -1$. Thereby, the adjoint solution $\mathbf{z}$ satisfies $\mathbf{r}^*[\mathbf{u}](\mathbf{z}) = 0$, provided that $\theta = -1$. Furthermore, $\mathbf{z}$ satisfies $\boldsymbol{\rho}^*[\mathbf{u}](\mathbf{z}) = 0$. In summary, we see that, like for Poisson's equation, the element and interior face terms of the IP discretization of the compressible Navier–Stokes equation are adjoint consistent for the symmetric ($\theta = -1$) but not for the nonsymmetric ($\theta = 1$) version.

The boundary terms of the discrete adjoint problem are given by

$$\begin{aligned}\mathcal{N}_\Gamma'[\mathbf{u}_h](\mathbf{w},\mathbf{z}_h)+&\int_\Gamma (\mathbf{w}\otimes\mathbf{n}) : \left(G_\Gamma^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right)\,\mathrm{d}s \equiv\\ &+\int_\Gamma \mathbf{n}\cdot\left(\mathcal{F}^c_{\Gamma,\mathbf{u}}[\mathbf{u}_h](\mathbf{w})\right)\mathbf{z}_h\,\mathrm{d}s+\int_\Gamma \delta\left(\mathbf{w}-\mathbf{u}_\Gamma'[\mathbf{u}_h]\mathbf{w}\right)\cdot\mathbf{z}\,\mathrm{d}s,\\ &-\int_\Gamma \mathbf{n}\cdot\left(\mathcal{F}^v_{\Gamma,\mathbf{u}}[\mathbf{u}_h,\nabla_h\mathbf{u}_h](\mathbf{w})+\mathcal{F}^v_{\Gamma,\nabla\mathbf{u}}[\mathbf{u}_h,\nabla_h\mathbf{u}_h](\nabla_h\mathbf{w})\right)\mathbf{z}_h\,\mathrm{d}s\\ &+\theta\int_\Gamma\left(\left(\left(G_\Gamma^\top\right)'[\mathbf{u}_h]\mathbf{w}\right)\nabla_h\mathbf{z}_h\right) : (\mathbf{u}_h-\mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n}\,\mathrm{d}s\\ &+\theta\int_\Gamma\left(G_\Gamma^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right) : (\mathbf{w}-\mathbf{u}_\Gamma'[\mathbf{u}_h]\mathbf{w})\otimes\mathbf{n}\,\mathrm{d}s\\ &+\int_\Gamma(\mathbf{w}\otimes\mathbf{n}) : \left(G_\Gamma^\top(\mathbf{u}_h)\nabla_h\mathbf{z}_h\right)\,\mathrm{d}s=\tilde{J}'[\mathbf{u}_h](\mathbf{w}).\end{aligned}$$

Thus the adjoint boundary residuals in (6.18) on Γ_W are given by

$$\begin{aligned}\mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h)=&\frac{1}{C_\infty}\left(p_\mathbf{u}\,\mathbf{n}-\underline{\tau}_\mathbf{u}\,\mathbf{n}\right)\cdot\psi-\left(\mathbf{n}\cdot\left(\mathcal{F}^c_{\Gamma,\mathbf{u}}-\mathcal{F}^v_{\Gamma,\mathbf{u}}\right)\right)^\top\mathbf{z}_h-\mathbf{n}\cdot\left(G_\Gamma^\top\nabla\mathbf{z}_h\right)\\ &+r_J'[\mathbf{u}_h]-\delta\left(I-\mathbf{u}_\Gamma'[\mathbf{u}_h]\right)^\top\mathbf{z}_h-\theta\left(G_\Gamma'[\mathbf{u}_h] : (\mathbf{u}_h-\mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_h\\ &-\theta\left(G_\Gamma(\mathbf{u}_h) : (I-\mathbf{u}_\Gamma'[\mathbf{u}_h])\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_h,\end{aligned}\tag{6.20}$$

$$\boldsymbol{\rho}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h)=-\frac{1}{C_\infty}\left(\underline{\tau}_{\nabla\mathbf{u}}\mathbf{n}\right)\cdot\psi+\left(\mathbf{n}\cdot\mathcal{F}^v_{\Gamma,\nabla\mathbf{u}}\right)^\top\mathbf{z}_h.\tag{6.21}$$

We recall (6.7), $\mathcal{F}^v_{\Gamma,\nabla\mathbf{u}}=G_\Gamma(\mathbf{u})$, and see that the exact solutions $\mathbf{u}$ and $\mathbf{z}$ to the primal problem (6.1) and the continuous adjoint problem (6.5)–(6.9) satisfy $\boldsymbol{\rho}_\Gamma^*[\mathbf{u}](\mathbf{z})=0$.

We now choose the modification $r_J(\mathbf{u}_h)$ of the target functional in (6.14) as follows:

$$r_J(\mathbf{u}_h)=\delta\left(\mathbf{u}_h^+-\mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\cdot\mathbf{z}_\Gamma+\theta\left(G_\Gamma^\top(\mathbf{u}_h^+)\nabla_h\mathbf{z}_\Gamma\right) : \left(\mathbf{u}_h^+-\mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)\otimes\mathbf{n},\tag{6.22}$$

with Fréchet derivative

$$\begin{aligned}r_J'[\mathbf{u}_h](\mathbf{w})=&\,\delta\left(I-\mathbf{u}_\Gamma'[\mathbf{u}_h]\right)\mathbf{w}\cdot\mathbf{z}_\Gamma+\theta\left(G_\Gamma'[\mathbf{u}_h] : (\mathbf{u}-\mathbf{u}_\Gamma(\mathbf{u}_h))\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_\Gamma\\ &+\theta\left(G_\Gamma(\mathbf{u}_h) : (I-\mathbf{u}_\Gamma'[\mathbf{u}_h])\otimes\mathbf{n}\right)^\top\nabla_h\mathbf{z}_\Gamma.\end{aligned}$$

As the exact solution $\mathbf{u}$ to the primal problem satisfies $\mathbf{u}_\Gamma(\mathbf{u})=\mathbf{u}$, we have $r_J(\mathbf{u})=0$. Hence (6.22) is a consistent modification of the target functional. Recalling (6.6) we see that the exact solutions $\mathbf{u}$ and $\mathbf{z}$ satisfy

$$\begin{aligned}\mathbf{r}_\Gamma^*[\mathbf{u}](\mathbf{z})=&\,\delta\left(I-\mathbf{u}_\Gamma'[\mathbf{u}]\right)^\top(\mathbf{z}_\Gamma-\mathbf{z})+\theta\left(G'[\mathbf{u}] : (\mathbf{u}-\mathbf{u}_\Gamma(\mathbf{u}))\otimes\mathbf{n}\right)^\top(\nabla\mathbf{z}_\Gamma-\nabla\mathbf{z})\\ &+\theta\left(G(\mathbf{u}) : (I-\mathbf{u}_\Gamma'[\mathbf{u}])\otimes\mathbf{n}\right)^\top(\nabla\mathbf{z}_\Gamma-\nabla\mathbf{z}).\end{aligned}\tag{6.23}$$

Furthermore, setting $\mathbf{z}_\Gamma=\mathbf{z}$ on Γ_W we obtain $\mathbf{r}_\Gamma^*[\mathbf{u}](\mathbf{z})=0$ and conclude that the discretization of boundary terms is adjoint consistent.

Due to $\mathbf{n}\cdot(G_\Gamma^\top(\mathbf{u}_h^+)\nabla\mathbf{z})=\mathbf{n}\cdot((\mathcal{F}^v_{\nabla\mathbf{u}})^\top\nabla\mathbf{z})=0$ on Γ_W the second term in (6.22) vanishes. Furthermore, on adiabatic boundaries Γ_{adia} we have $(\mathbf{u}_h^+-\mathbf{u}_\Gamma(\mathbf{u}_h^+))_i=0$,

$i = 1, 4$, and on isothermal boundaries Γ_{iso} we have $(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+))_1 = 0$. Together with (6.9) the consistent modification (6.22) reduces to

$$\begin{aligned} r_J(\mathbf{u}_h) &= \delta \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right) \cdot \mathbf{z}_\Gamma \\ &= \delta \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)_2 \frac{1}{C_\infty}\psi_1 + \delta \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right)_3 \frac{1}{C_\infty}\psi_2, \end{aligned} \tag{6.24}$$

which completes the adjoint consistency analysis of the interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations. Finally, we note that the consistent modification $r_J(\mathbf{u}_h)$ given in (6.24) corresponds to the IP modification of target functionals for Poisson's equation where $r_J(u_h) = \delta(u_h - g_D)z_\Gamma$ with $z_\Gamma = -j_D$; see (4.13).

In summary, we have shown that the adjoint element and interior residuals $\mathbf{R}^*[\mathbf{u}_h](\mathbf{z}_h)$, $\mathbf{r}^*[\mathbf{u}_h](\mathbf{z}_h)$, and $\boldsymbol{\rho}^*[\mathbf{u}_h](\mathbf{z}_h)$ (see (6.19)) vanish for the exact solutions $\mathbf{u}$ and $\mathbf{z}$ to (6.1) and (6.5), respectively, provided $\theta = -1$. Additionally, using an adjoint consistent treatment of convective and diffusive boundary fluxes,

$$\mathbf{n} \cdot \mathcal{F}_\Gamma^c(\mathbf{u}_h^+) = \mathbf{n} \cdot \mathcal{F}^c(\mathbf{u}_\Gamma(\mathbf{u}_h^+)), \qquad \mathbf{n} \cdot \mathcal{F}_\Gamma^v(\mathbf{u}_h^+) = \mathbf{n} \cdot \mathcal{F}^v(\mathbf{u}_\Gamma(\mathbf{u}_h^+), \nabla_h \mathbf{u}_h^+), \tag{6.25}$$

and using the following consistent modification of the target functional:

$$\tilde{J}(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h)) + \int_{\Gamma_W} \delta \left(\mathbf{u}_h^+ - \mathbf{u}_\Gamma(\mathbf{u}_h^+)\right) \cdot \mathbf{z}_\Gamma \, \mathrm{d}s, \tag{6.26}$$

with $\mathbf{z}_\Gamma = \frac{1}{C_\infty}(0, \psi_1, \psi_2, 0)^\top$, for $J(\cdot)$ representing a total force coefficient defined in (6.3), the adjoint boundary residuals $\mathbf{r}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h)$ and $\boldsymbol{\rho}_\Gamma^*[\mathbf{u}_h](\mathbf{z}_h)$ (see (6.20) and (6.21)) vanish for the exact solutions $\mathbf{u}$ and $\mathbf{z}$. Thereby, using the modifications given in (6.25) and (6.26) we recover an adjoint consistent symmetric interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations in conjunction with total force coefficients. Finally, we note that arguments given in [23] en route to obtaining an adjoint consistent discretization based on the BR2 scheme [5] can also be covered within the presented framework and lead to analogous modifications.

7. Numerical experiments for the compressible Navier–Stokes equations. In this section we will demonstrate the effect on the smoothness of the discrete adjoint solution when employing the adjoint consistent SIPG discretization based on (6.25) and (6.26) in comparison to the original SIPG discretization of the compressible Navier–Stokes equations [16], which instead uses

$$\mathcal{H}(\mathbf{u}_h^+, \mathbf{u}_\Gamma(\mathbf{u}_h^+), \mathbf{n}), \qquad \mathbf{n} \cdot \mathcal{F}^v(\mathbf{u}_h^+, \nabla_h \mathbf{u}_h^+) \quad \text{on } \Gamma, \qquad \text{and} \qquad J(\mathbf{u}_h). \tag{7.1}$$

Furthermore, we compare the accuracy of the original formulation and the adjoint consistent discretization on a sequence of globally refined meshes.

To this end, we revisit the standard test case [3, 16, 17] of a $M = 0.5$ viscous flow at $\mathrm{Re} = 5000$ and at zero angle of attack around the NACA0012 airfoil with adiabatic no-slip boundary conditions imposed on the profile. Figure 7.1 shows the primal flow solution based on the adjoint consistent discretization on a (locally) refined mesh created by repeated refinement of the coarse C-type mesh depicted in Figure 7.2. Then, in Figure 7.3 we show the components z_1–z_4 of the corresponding discrete adjoint solution $\mathbf{z}_h$. In particular, we find that the second and third components $z_2 \approx 1/C_\infty = 40/7$ and $z_3 \approx 0$ are constant on the profile. Furthermore, we see that $\mathbf{n} \cdot \nabla z_4 \approx 0$ as required by the continuous adjoint boundary conditions (6.9).

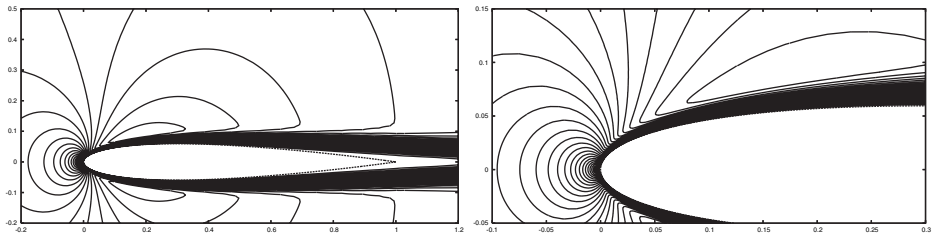


FIG. 7.1. *Adjoint consistent DG discretization of the* $M = 0.5, \alpha = 0°, \mathrm{Re} = 5000$ *flow around the NACA*0012 *airfoil: Mach isolines* $(0.02i, i \in \mathbb{N})$ *of the (primal) flow solution.*

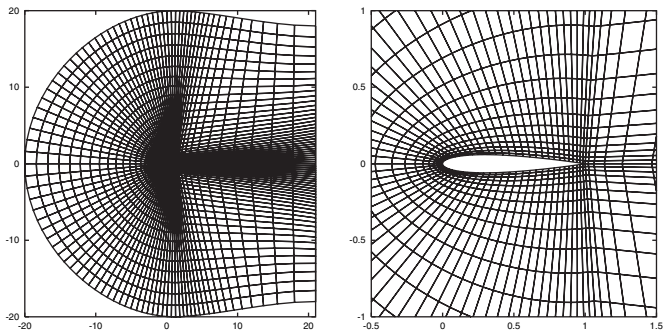


FIG. 7.2. *Computational mesh with* 3072 *element: (left) full and (right) detailed view.*

For comparison, Figure 7.4 shows the z_4 component of the discrete adjoint to the original DG discretization [16]. We clearly see an irregular adjoint solution at the profile boundary. In contrast, the discrete adjoint solution (see Figure 7.3) based on the adjoint consistent discretization is entirely smooth. The components z_i, $i = 1, 2, 3$, show a similar behavior.

For a quantitative comparison of the original and the adjoint consistent discretization, we collect the errors of the computed solutions on a sequence of globally refined meshes in Table 7.1. Here we show the number of elements, the number of degrees of freedom, and the error $J(\mathbf{u}) - J(\mathbf{u}_h)$ of the flow solution for three different discretizations. The error of $\mathbf{u}_h \in \mathbf{V}_h^1$ is measured in terms of the total drag coefficient c_d where the reference value $J(\mathbf{u}) \approx 0.05482$ is based on very fine grid computations. In columns (a) and (b) we collect the errors and rates of convergence of the original SIPG formulation (7.1) (cf. [16]) and for the adjoint consistent discretization based on (6.25) and (6.26), respectively. We see that the adjoint consistent discretization is by a factor of about 2–400 more accurate than the original discretization on the same mesh and with the same numerical complexity. Furthermore the adjoint consistent discretization shows an $\mathcal{O}(h^5)$ order of convergence which is significantly higher than that of the original discretization. In order to demonstrate the relevance of the IP modification (6.24) of the target functional, in column (c) of Table 7.1 we collect the respective errors based on the same discretization as in column (b) while omitting the IP modification; i.e., column (c) is based on (6.25) and $\tilde{J}(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h))$. The accuracy is significantly reduced, partly even below the accuracy of the original discretization. From this, we see that the IP modification is essential for the adjoint consistency of the discretization and the accuracy of the numerical flow solution. A similar behavior is seen for $\mathbf{u}_h \in \mathbf{V}_h^2$ in Table 7.2. Here the original discretization

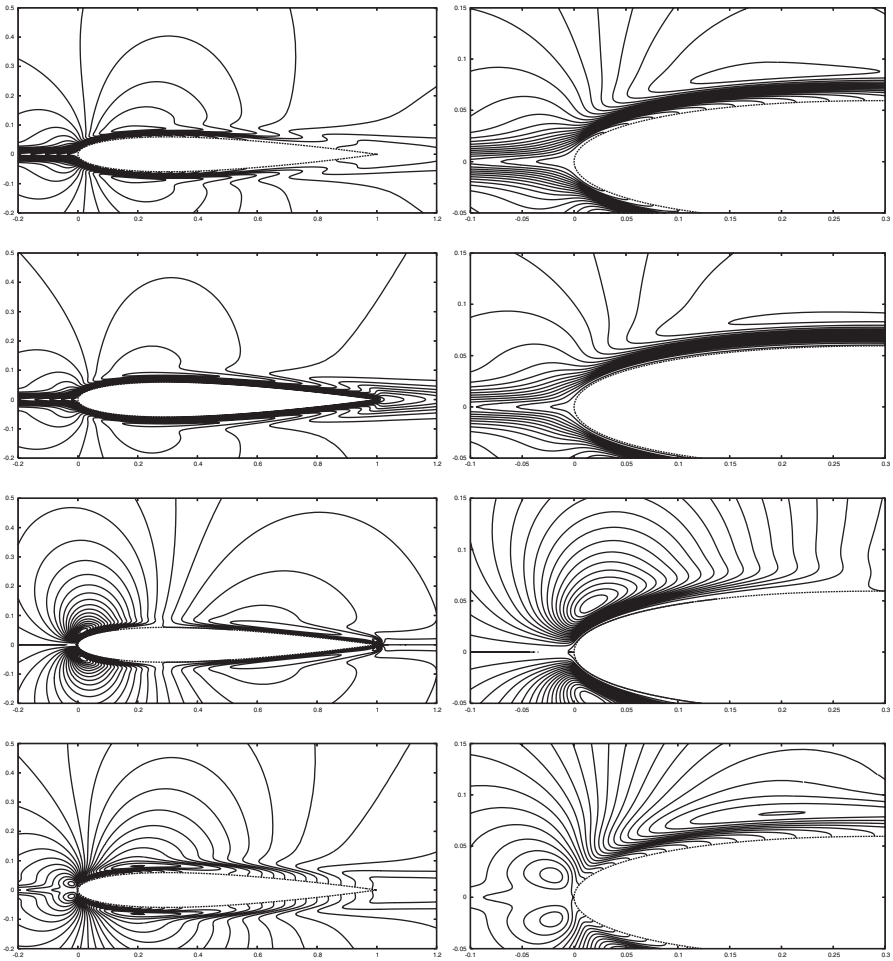


FIG. 7.3. *Adjoint consistent DG discretization of the* $M = 0.5, \alpha = 0^\circ, \mathrm{Re} = 5000$ *flow around the NACA*0012 *airfoil: ith row: Isolines of component* z_i, $i = 1, \ldots, 4$, *of the adjoint sol.* $\mathbf{z}$ *for* c_d.

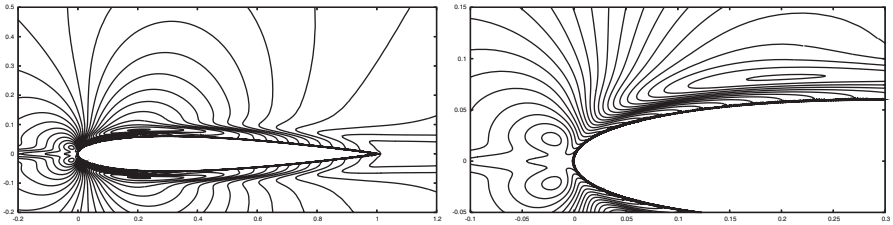


FIG. 7.4. *Original DG discretization* [16] *of the* $M = 0.5, \alpha = 0^\circ, \mathrm{Re} = 5000$ *flow around the NACA*0012 *airfoil: Isolines of component* z_4 *of the adjoint solution* $\mathbf{z}$ *for* c_d.

shows an $\mathcal{O}(h^3)$ convergence on average, whereas the convergence of the adjoint consistent discretization is $\mathcal{O}(h^6)$.

We note that independently Lu [23] demonstrated, for a $M = 0.5$ viscous compressible flow around the NACA0012 airfoil at $\mathrm{Re} = 5000$ and $\alpha = 2^\circ$ angle of attack, that the discrete adjoint solution to the adjoint consistent discretization based on the BR2 scheme [5] is smooth whereas that to the standard discretization is not. Fur-

TABLE 7.1

Error $J(\mathbf{u}) - J(\mathbf{u}_h)$ and rate of convergence of $\mathbf{u}_h \in \mathbf{V}_h^1$ measured in terms of c_d (see (6.3)) for (a) *the original SIPG formulation based on* (7.1) *(cf.* [16]*), for* (b) *the adjoint consistent discretization based on* (6.25) *and* (6.26)*, and for* (c) *the adjoint consistent discretization without the IP modification of the target functional, i.e., based on* (6.25) *and $\tilde{J}(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h))$.*

# Cells	# Dofs	(a) Error	rate	(b) Error	rate	(c) Error	rate
3072	49152	-3.164e-03	-	1.502e-03	-	-1.243e-03	-
12288	196608	8.048e-04	3.9	3.682e-05	40.8	6.994e-04	1.8
49152	786432	4.519e-04	1.8	-1.139e-06	32.3	4.795e-04	1.5

TABLE 7.2

Error $J(\mathbf{u}) - J(\mathbf{u}_h)$ and rate of convergence of $\mathbf{u}_h \in \mathbf{V}_h^2$ measured in terms of c_d (see (6.3)) for (a) *the original SIPG formulation based on* (7.1) *(cf.* [16]*), for* (b) *the adjoint consistent discretization based on* (6.25) *and* (6.26)*, and for* (c) *the adjoint consistent discretization without the IP modification of the target functional, i.e., based on* (6.25) *and $\tilde{J}(\mathbf{u}_h) = J(\mathbf{u}_\Gamma(\mathbf{u}_h))$.*

# Cells	# Dofs	(a) Error	rate	(b) Error	rate	(c) Error	rate
768	27648	-3.903e-02	-	5.565e-03	-	-2.054e-02	-
3072	110592	8.663e-04	45.1	6.234e-05	89.3	4.216e-04	48.7
12288	442368	4.987e-04	1.7	9.789e-07	63.7	4.139e-04	1.02

thermore, [23] showed optimal order of convergence results for the adjoint consistent discretization in comparison to reduced ones for the standard discretization which lacks adjoint consistency.

8. Conclusion. A discretization is adjoint consistent if the discrete adjoint problem is a consistent discretization of the continuous adjoint problem. In fact, adjoint consistency is the link between the so-called continuous adjoint approach (which discretizes the adjoint equations) and the discrete adjoint approach (which takes the adjoint of the discrete equations) in that the solutions to both approaches coincide. In particular, adjoint consistency is the key requirement for optimal order duality-based error estimates in L^2 as well as measured in terms of target functionals. Furthermore, adjoint consistency is closely related to the smoothness of the discrete adjoint solutions. However, in addition to the adjoint consistency of element and interior face terms, an adjoint consistent treatment of boundary terms as well as of target functionals is required for an adjoint consistent discontinuous Galerkin discretization.

In this article we have introduced a framework for analyzing consistency and adjoint consistency of discontinuous Galerkin discretizations of linear and nonlinear problems. This framework includes the derivation of the continuous adjoint problems and adjoint boundary conditions, provided the primal problem and the target functional satisfy a compatibility condition. It includes the derivation of the discrete adjoint problems and primal and adjoint residuals and a discussion of under which conditions the residuals vanish for the exact primal and adjoint solutions. In addition, we have introduced so-called consistent modifications of target functionals which allow us to modify (and possibly improve) computed target quantities without changing their exact values. We then analyzed the DG discretization of the linear advection equation, the interior penalty (IP)DG discretization of Poisson's equation, and the DG discretization of the compressible Euler equations. While recovering properties and conclusions drawn in [1, 11, 24], the outlined framework gives a unified analysis of these discretizations, including the definition of consistent modification of a target functional such as the so-called IP modification as well as a consistent modification of the force coefficients for inviscid compressible flows.

This framework has then been used to analyze the adjoint consistency property of the symmetric interior penalty discontinuous Galerkin discretization of the compressible Navier–Stokes equations. While the original formulation of the SIPG discretization [16] has been shown to be adjoint inconsistent, the analysis revealed that a special treatment of boundary terms as well as an IP modification of viscous force coefficients is required for recovering an adjoint consistent DG discretization of the compressible Navier–Stokes equations. Numerical experiments have confirmed that, in contrast to the original formulation in [16], the discrete adjoint solution to the adjoint consistent discretization is entirely smooth. Furthermore, numerical tests on globally refined meshes have shown that the adjoint consistent discretization is by a factor of 2–400 more accurate measured in terms of viscous force coefficients than the original formulation. Also, a significantly improved order of convergence has been observed.

REFERENCES

[1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal., 39 (2002), pp. 1749–1779.

[2] J. P. Aubin, *Approximation of Elliptic Boundary-Value Problems*, Pure Appl. Math., 26, Wiley, Interscience, New York, 1972.

[3] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys., 131 (1997), pp. 267–279.

[4] F. Bassi and S. Rebay, *High-order accurate discontinuous finite element solution of the* 2*d Euler equations*, J. Comput. Phys., 138 (1997), pp. 251–285.

[5] F. Bassi and S. Rebay, *GMRES discontinuous Galerkin solution of the compressible Navier-Stokes equations*, in Discontinuous Galerkin Methods, B. Cockburn, G. Karniadakis, and C.-W. Shu, eds., Lect. Notes Comput. Sci. Eng. 11, Springer-Verlag, Berlin, 1999, pp. 197–208.

[6] F. Brezzi, B. Cockburn, D. M. Marini, and E. Süli, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg., 195 (2006), pp. 3293–3310.

[7] F. Brezzi, L. D. Marini, and E. Süli, *Discontinuous galerkin methods for first-order hyperbolic problems*, Math. Models Methods Appl. Sci., 14 (2004), pp. 1893–1903.

[8] C. Castro, C. Lozano, F. Palacios, and E. Zuazua, *A systematic continuous adjoint approach to viscous aerodynamic design on unstructured grids*, AIAA, 2006-0051, 2006.

[9] N. Gauger, *Das Adjungiertenverfahren in der aerodynamischen Formoptimierung*, Technical report DLR-FB-2003-05 (ISSN 1434-8454), DLR, 2003.

[10] M. Giles and N. Pierce, *Adjoint equations in CFD: Duality, boundary conditions and solution behaviour*, AIAA, 97-1850, 1997.

[11] K. Harriman, D. Gavaghan, and E. Süli, *The Importance of Adjoint Consistency in the Approximation of Linear Functionals Using the Discontinuous Galerkin Finite Element Method*, Technical report, Oxford University Computing Laboratory, Oxford, 2004.

[12] K. Harriman, P. Houston, B. Senior, and E. Süli, *hp–Version discontinuous Galerkin methods with interior penalty for partial differential equations with nonnegative characteristic form*, in Recent Advances in Scientific Computing and Partial Differential Equations, Contemporary Mathematics 330, AMS, Providence, RI, 2003, pp. 89–119.

[13] R. Hartmann, *The role of the Jacobian in the adaptive Discontinuous Galerkin method for the compressible Euler equations*, in Analysis and Numerics for Conservation Laws, G. Warnecke, ed., Springer-Verlag, Berlin, 2005, pp. 301–316.

[14] R. Hartmann, *Derivation of an adjoint consistent discontinuous Galerkin discretization of the compressible Euler equations*, in Proceedings of the BAIL 2006 conference, G. Lube and G. Rapin, eds., 2006.

[15] R. Hartmann and P. Houston, *Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations*, J. Comput. Phys., 183 (2002), pp. 508–532.

[16] R. Hartmann and P. Houston, *Symmetric interior penalty DG methods for the compressible Navier–Stokes equations* I: *Method formulation*, Int. J. Numer. Anal. Model., 3 (2006), pp. 1–20.

[17] R. Hartmann and P. Houston, *Symmetric interior penalty DG methods for the compressible Navier–Stokes equations* II: *Goal–oriented a posteriori error estimation*, Int. J. Numer. Anal. Model., 3 (2006), pp. 141–162.
[18] P. Houston, R. Rannacher, and E. Süli, *A posteriori error analysis for stabilised finite element approximations of transport problems*, Comput. Methods Appl. Mech. Engrg., 190 (2000), pp. 1483–1508.
[19] P. Houston, C. Schwab, and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal., 39 (2006), pp. 2133–2163.
[20] A. Jameson, *Aerodynamic design via control theory*, J. Sci. Comput., 3 (1988), pp. 233–260.
[21] A. Jameson, N. Pierce, and L. Martinelli, *Optimum aerodynamic design using the Navier-Stokes equations*, Theoretical and Computational Fluid Dynamics, 10 (1998), pp. 213–237.
[22] J. L. Lions and E. Magenes, *Non-Homogeneous Boundary Value Problems and Applications*, Springer-Verlag, Berlin, 1972.
[23] J. Lu, *An a posteriori Error Control Framework for Adaptive Precision Optimization Using Discontinuous Galerkin Finite Element Method*, Ph.D. thesis, MIT, Cambridge, MA, 2005.
[24] J. Lu and D. L. Darmofal, *Dual-consistency analysis and error estimation for discontinuous Galerkin discretization: Application to first-order conservation laws*, IMA J. Numer. Anal., submitted.
[25] S. Prudhomme, F. Pascal, J. Oden, and A. Romkes, *Review of* a priori *error estimation for discontinuous Galerkin methods*, TICAM Report 00-27, University of Texas, 2000.